

The Impact of Community Cohesion on Crime

Usman Lawal Gulma

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
School of Geography
Centre for Spatial Analysis and Policy

September, 2018

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Academic Acknowledgements

I would like to thank the UK Data Service: Census Support for providing the Census 2011 statistics that were used in this research. I would also like thank the United Kingdom Boundary Outline and Reference Database for Educational and Research Study (UKBORDERS) for the outline maps used. I also thank Professor John Stillwell for permitting me to reproduce his work on Leeds Community Areas in this research. Finally, I would like to thank the West Yorkshire Police for providing their publicly available data through the data.police.uk website.

Acknowledgements

All praise is due to almighty Allah the Most Gracious and the Most Merciful for seeing me through to a successful completion of this research. I would like to thank my supervisors: Professor Alison Heppenstall, Dr Andrew Evans and Dr Nick Malleon, for their untiring and invaluable contribution throughout the duration of the research. Indeed, I am greatly indebted for all your support and expertise without which this research would not have been a success. I would also like to thank the members of research support group (RSG) in the person of Professor John Stillwell and Dr Andy Newing for their help.

My profound gratitude goes to the Tertiary Education Trust Fund (TETFUND) for funding the research project. I would also like to thank the School of Geography for providing me with an additional funding towards the completion of the research. I am also grateful to my office colleagues and all other persons too numerous to be mention here who have made my stay in Leeds a memorable one.

I would like to extend my sincere gratitude to my wife and children, the entire family of Late Tafida Lawal Gulma and particularly my mum for their prayers and continuous moral encouragement throughout the duration of this research. Finally, I would like to dedicate this work in memory of my late dad who, as Allah Has wished, could not be around to celebrate this achievement with me.

Abstract

Community cohesion generally acts to increase the safety of communities by increasing informal guardianship, and enhancing the work of formal crime prevention organisations. Understanding the dynamics of local social interactions is essential for community building. However, community cohesion is difficult to empirically quantify, because there are no obvious and direct indicators of community cohesion collected at population levels within official datasets. A potentially more promising alternative for estimating community cohesion is through the use of data from social media. Social media offers an opportunity for exploring networks of social interactions in a local community.

This research will use social media data to explore the impact of community cohesion on crime. Sentiment analysis of tweets can help to uncover patterns of community mood in different areas. Modelling of community engagement on Facebook is useful for understanding patterns of social interactions and the strength of social networks in local communities.

The central contribution of this thesis is the use of new metrics that estimate popularity, commitment and virality known as the *PCV indicators* for quantifying community cohesion on social media. These metrics, combined with diversity statistics constructed from “traditional” Census data, provide a better correlate of community cohesion and crime. To demonstrate the viability of this novel method for estimating the impact of community cohesion, a model of community engagement and burglary rates is constructed using Leeds community areas as an example. By examining the diversity of different community areas and strength of their social networks, from traditional and new data sources; it was found that stability and strong social media engagement in a local area are associated with lower burglary rates. The proposed new method can provide a better alternative for estimating community cohesion and its impact on crime. It is recommended that policy planning for resource allocation and community building needs to consider social structure and social networks in different communities.

Publication and Conferences

The following table presents peer reviewed paper publication and conference presentations from the thesis.

Peer-reviewed published paper	
Gulma U.L., Evans A., Heppenstall A. and Malleson N. (2018). Diversity and burglary: Do community differences matter? <i>Transactions in GIS</i> . DOI: 1111/tgis.12511	Part of the work in Chapter 5 of the thesis (see Appendix B)

Peer-reviewed conference presentations	
Gulma U.L., Evans A., Heppenstall A. and Malleson N. (2017). Diversity and burglary: Does community cohesion matter?	25 th GISRUK conference, University of Manchester 18 th – 21 st April
Gulma U.L., Evans A., Heppenstall A. and Malleson N. (2016). Diversity and crime in Leeds: Does community cohesion matter?	24 th GISRUK conference, University of Greenwich 30 th March – 1 st April

Table of Contents

List of Figures, Tables and Abbreviations

Abstract.....	iii
Table of Contents.....	v
Chapter 1 : Introduction.....	1
1.1 Background of the Study	1
1.1.1 Problem Statement	2
1.2 Aim and Objectives.....	3
1.2.1 Justification for the Study	4
1.2.2 Significance of the Study.....	5
1.2.3 Defining Community in the Context of Crime	5
1.3 Thesis Structure.....	6
Chapter 2 : Concept of Community Cohesion and Crime: A Literature Review	9
2.1 Introduction.....	9
2.2 Defining the Concept of Community Cohesion	9
2.3 Community Cohesion and Crime	10
2.4 Importance of Social Capital on Crime	13
2.5 Determinants of Social Capital in Relation to Crime	15
2.5.1 Age, Social Capital and Crime	18
2.5.2 Gender, Social Capital and Crime.....	19
2.5.3 Internet Usage and Social Capital.....	21
2.5.4 Online Membership and Social Capital	22
2.5.5 Participation in Community Activities, Social Capital and Crime	24
2.5.6 Participation in Election, Social Capital and Crime.....	25
2.5.7 Heterogeneity, Social Capital and Crime	26
2.5.8 Education, Social Capital and Crime.....	27
2.5.9 Length of Residence, Social Capital and Crime	28
2.5.10 Deprivation, Social Capital and Crime.....	29
2.6 Collective Efficacy and Crime	30

2.6.1 Measurement of Collective Efficacy on Crime.....	30
2.7 Weaknesses of Previous Studies	32
2.8 Theoretical Framework.....	33
2.8.1 Social Disorganisation Theory.....	33
2.8.2 Routine Activity Theory	35
2.9 Concluding Remarks	37
Chapter 3 : Traditional Methods of Modelling Urban Social Systems and Crime	38
3.1 Introduction.....	38
3.2 Research Methodology.....	38
3.3 Study Area.....	39
3.3.1 Geography	39
3.3.2 People.....	40
3.4 Data Sources	41
3.4.1 Census Data	42
3.4.2 Police Recorded Crime Data.....	47
3.5 Leeds Community Areas	49
3.6 Statistical Methods of Modelling Crime.....	51
3.6.1 Regression Models	51
3.7 Spatial Analyses	56
3.7.1 Choropleth Mapping.....	56
3.7.2 Hotspot Mapping	58
3.8 Spatial Statistical Methods.....	61
3.8.1 Spatial Association.....	61
3.9 Diversity Indices.....	63
3.10 Computational Models	65
3.11 Concluding Remarks	67
Chapter 4 : Non-Traditional Methods: Social Media for Community Cohesion and Crime	68
4.1 Introduction.....	68
4.2 Social Media and Social Cohesion	68

4.2.1 Facebook	72
4.2.2 Twitter	74
4.3 Social Media Community Cohesion and Crime: Action in Communities	78
4.4 Social Media Community Cohesion and Crime: Understanding Communities	80
4.5 Social Media and Law Enforcement	85
4.6 Social Media, Social Capital and Crime.....	86
4.7 Methods.....	89
4.7.1 Sentiment.....	89
4.7.2 Sentiment Analysis.....	90
4.7.3 Social Cohesion and Public Sentiment on Social Media	94
4.7.4 Twitter Sentiment and Community Structural Characteristics	95
4.7.5 Understanding Social Cohesion through Sentiment Analysis.....	96
4.7.6 Social Media Engagement	97
4.7.7 New Metrics of Community Engagement	99
4.8 Concluding Remarks	101
Chapter 5 : Traditional Modelling of Crime and Community Cohesion	102
5.1 Introduction.....	102
5.2 The Importance of Quantifying Urban Crime	102
5.3 Understanding Correlates of Crime in Urban Communities	103
5.3.1 Standard versus Adjusted Variables of Crime and Community Cohesion...	104
5.4 Exploring the Impact of Social Cohesion and Diversity on Crime	105
5.4.1 Social Capital as an Element of Social Cohesion.....	106
5.4.2 Diversity as an Element for Understanding Social Cohesion	106
5.5 Data and Method	109
5.5.1 Data	109
5.5.2 Method	109
5.6 Theoretical Justification for the Explanatory Variables	111
5.6.1 Age Distribution.....	111
5.6.2 Family Structure	113
5.6.3 Identity	115

5.6.4 Affluence and Wealth	118
5.6.5 Educational Attainment	120
5.6.6 Residential Instability	122
5.7 Results of the Traditional Multiple Regression Models	127
5.8 Concluding Remarks	131
Chapter 6 : Social Media Analysis for Understanding Community Cohesion	132
6.1 Introduction	132
6.3 Twitter Data	132
6.3.1 Keywords Method	132
6.3.2 Geographical Bounding Box Method.....	133
6.3.5 Twitter Data Pre-processing.....	134
6.3.6 Twitter Data Cleaning.....	135
6.4 Census Variables.....	138
6.5 Sentiment Analysis Procedure.....	138
6.5.1 Sentiment Classification, Accuracy and Validation.....	141
6.5.2 Analysis of AMT Validation Results.....	142
6.5.3 Twitter Sentiment and its Related Determinants	150
6.6 Sentiment Analysis Results	151
6.7 Twitter Sentiment and Diversity Characteristics of Communities.....	155
6.7.1 Twitter Sentiment and Ethnicity.....	155
6.7.3 Twitter Sentiment and Education	160
6.7.4 Twitter Sentiment and Length of Residence.....	162
6.7.5 Twitter Sentiment and Employment	164
6.8 Concluding Remarks	167
Chapter 7 : New Social Media Metrics for Quantifying Community Cohesion and Crime	169
7.1 Introduction.....	169
7.2 Understanding Community Engagement on Social Media.....	169
7.3 Facebook Pages and Groups	170
7.3.1 Facebook Pages	170

7.3.2 Facebook Groups.....	171
7.4 Facebook Data Collections.....	173
7.5 Quantifying Community Engagement on Facebook.....	179
7.5.1 Grouping Community Engagement Rates.....	181
7.5.2 Factors Affecting Community Engagement Rate	183
7.5.3 The Concept of Digital Divide.....	184
7.5.4 Influence of Facebook Interaction Metrics on Community Engagement	186
7.5.5 Engagement and Action in the Community Areas.....	187
7.6 Multiple Regression Models of Community Engagement and Burglary Crime Rates	189
7.6.1 Model Interpretation of Community Engagement and Burglary.....	198
7.7 Concluding Remarks	212
Chapter 8 : Community Cohesion and Crime: A New Leeds Community Areas	
Classification Profile	214
8.1 Introduction.....	214
8.2 Importance of Community Classification.....	214
8.2.1 Communities and Crime.....	216
8.2.2 Issues with Previous Classification Attempts	217
8.3 Cluster Analysis	218
8.3.1 Clustering Elements (Objects to Cluster)	219
8.3.2 Clustering Variables	219
8.3.3 Variable Standardisation	222
8.3.4 Measure of Proximity	223
8.3.5 Clustering Methods	224
8.3.6 Choosing Number of Clusters	228
8.3.7 Clustering Validation	235
8.4 Leeds Community Classification.....	239
8.4.1 Group 1	240
8.4.2 Group 2	242
8.4.3 Group 3	244

8.4.4 Group 4	247
8.5 Concluding Remarks	249
Chapter 9 : Conclusion.....	251
9.1 Introduction.....	251
9.2 Summary of Research Findings.....	251
9.3 Limitations of the Research	256
9.4 Recommendations for Future Research	257
9.5 Concluding Remarks	258
List of References	259
Appendix A.....	315
Appendix B.....	319

List of Figures

Figure 2.1: Crimes by age group. Source: Criminal Justice Statistics, 2014 (ONS, 2014c)	19
Figure 2.2: Social interaction between men and women. Source: ONS (2014e)	20
Figure 2.3: Crime by gender. Source: Criminal Justice Statistics, 2014 (ONS, 2014c)	21
Figure 2.4: Internet usage by age group in 2014. Source: ONS (2014d)	22
Figure 2.5: Percentage of UK households with internet. Source: ONS (2017c)	23
Figure 2.6: Participation in community activities by gender and age. Source: DCMS (2010)....	25
Figure 2.7: Participation in election by age groups. Source: TEC (2010).....	26
Figure 2.8: Eck crime triangle. Source: Eck and Weisburd (1995).....	37
Figure 3.1: Location map of Leeds LSOA boundaries. Source: Edina (2011)	40
Figure 3.2: Ethnic diversity of Leeds. Source: ONS (2011b).....	41
Figure 3.3: Typical UK Police data download website	48
Figure 3.4: Leeds 106 community area 2001 boundaries. Source: Stillwell and Phillips (2006)	50
Figure 3.5: Burglary crime rates in Leeds (2011-2015)	57
Figure 3.6: Kernel density estimation of burglary crime (2011-15) in Leeds.....	59
Figure 3.7: Kernel density estimation of burglary ranges (10%, 25%, 50% and 75%) in Leeds.	60
Figure 3.8: Spatial autocorrelation statistics for burglary crime in Leeds.....	62
Figure 4.1: Global social media active users. Source: Statistica.com (2016)	71
Figure 4.2: Typical Facebook account profile	73
Figure 4.3: Typical Facebook API development app.....	74
Figure 4.4: Twitter account profile.....	75
Figure 4.5: Typical Twitter developments API.....	76
Figure 4.6: West Yorkshire Police social media channels	86
Figure 4.7: Sentiment classification techniques. Source (Medhat <i>et al.</i> , 2014)	92
Figure 5.1: Age distribution variables (standard (a) and diversity (b)) in Leeds by LSOA	112
Figure 5.2: Family variables (standard (a) and diversity (b)) in Leeds by LSOA	114
Figure 5.3: Ethnicity variables (standard (a) and diversity (b)) in Leeds by LSOA	117
Figure 5.4: Distributions of affluence variables (standard (a) and diversity (b)) in Leeds by LSOA.....	119

Figure 5.5: Educational attainment variables (standard (a) and diversity (b)) in Leeds by LSOA	121
Figure 5.6: Length of residence variables (standard (a) and diversity (b)) in Leeds by LSOA .	123
Figure 5.7: Index of Multiple Deprivation in Leeds by LSOA.....	124
Figure 6.1: Histogram of the distribution of tweets in the community areas.....	137
Figure 6.2: Aggregated distribution of tweets.....	137
Figure 6.3: Sentiment analysis code. Source Breen (2011)	140
Figure 6.4: Distribution of ranked AMT classification	144
Figure 6.5 Distribution of the IRR analysis between binary and ternary classifications	148
Figure 6.6: Word cloud of frequent terms	151
Figure 6.7: Histogram of sentiment analysis score	152
Figure 6.8: Spatial distribution of positive Twitter sentiments in community areas.....	153
Figure 6.9: Spatial distribution of negative Twitter sentiments in community areas	154
Figure 6.10: Percentage shares of positive and negative sentiments in community areas	155
Figure 6.11: Scatterplot of sentiment scores and ethnic minority population.....	156
Figure 6.12: Scatterplot of sentiment scores and ethnic diversity.....	156
Figure 6.13: Distribution of Twitter sentiments and ethnic diversity in community areas	157
Figure 6.14: Scatterplot of sentiment scores and number of young persons (16-24).....	158
Figure 6.15: Scatterplot of sentiment scores and age diversity.....	159
Figure 6.16: Distribution of positive and negative Twitter sentiments and age diversity.....	160
Figure 6.17: Scatterplot of sentiment scores and persons with no educational qualification	161
Figure 6.18: Scatterplot of sentiment scores and education diversity.....	161
Figure 6.19: Distribution of positive and negative Twitter sentiments and education diversity	162
Figure 6.20: Scatterplot of sentiment scores and length of residence less than 2 years	163
Figure 6.21: Scatterplot of sentiment scores and length of residence diversity	163
Figure 6.22: Distribution of Twitter sentiments and length of residence diversity.....	164
Figure 6.23: Scatterplot of sentiment scores and employment diversity.....	165
Figure 6.24: Scatterplot of sentiment scores and employment diversity.....	166
Figure 6.25: Distributions of Twitter sentiments and employment diversity	167

Figure 7.1: Typical Facebook pages related to the word ‘Otley’ (Otley is a town in West Yorkshire).....	171
Figure 7.2: Typical types of Facebook groups in ‘Horsforth’ (a neighbourhood in Leeds).....	172
Figure 7.3: Distribution of Facebook pages and groups in different communities.....	174
Figure 7.4 Stacked graphs of the PCV metrics of engagement in different community areas ..	178
Figure 7.5: Frequency distribution of community areas engagement rate	180
Figure 7.6: Spatial distributions of engagement rates in community areas	182
Figure 7.7 Comparison between engagement rate and factors of social cohesion in selected community groups	184
Figure 7.8: Metrics of engagement in different community areas	186
Figure 7.9: Engagement rate in relation to crime rates in the community areas	187
Figure 7.10: Map of the distributions of burglary rates in Leeds community areas	190
Figure 7.11: Map of the distribution of age diversity in Leeds community areas	191
Figure 7.12: Map of the distribution of residential diversity in Leeds community areas	191
Figure 7.13: Map of the distribution of economically inactive population in Leeds community areas.....	192
Figure 7.14: Map of the distribution of virality in Leeds community areas	192
Figure 7.15: Plot of cross-validation (CV) score as a function of bandwidth.....	197
Figure 7.16: Map of the distribution of the local R^2 values of the GWR model	201
Figure 7.17 Distributions of the intercept coefficients	203
Figure 7.18: Map of coefficient of virality	204
Figure 7.19: Map of standard error of the virality parameter	204
Figure 7.20: Map of p-values of the virality parameter.....	205
Figure 7.21: Map of age diversity coefficient.....	206
Figure 7.22 Map of the standard error of age diversity parameter	207
Figure 7.23 Map of p-values of the age diversity parameter.....	207
Figure 7.24: Map of residential diversity coefficient	208
Figure 7.25: Standard error of residential diversity	209
Figure 7.26: p-values of residential diversity.....	209
Figure 7.27: Coefficient of economically inactive population	210

Figure 7.28 Standard error of economically inactive population	211
Figure 7.29 p-values of economically inactive population	211
Figure 8.1: Boxplots for standardised variables retained for clustering.....	223
Figure 8.2: Scree plot of the agglomeration schedule (left) and Elbow plot for K =10 cluster solutions (right)	230
Figure 8.3: Gap statistics plot	232
Figure 8.4: Average silhouette widths between 2 and 10 cluster solutions	233
Figure 8.5: Silhouette plots for candidate clusters (4, 5, 6 and 7)	234
Figure 8.6: Silhouette information for K=4 and K=6 cluster solutions compared.....	235
Figure 8.7: Graph plot of the comparison between RI and ARI index values	238
Figure 8.8: Combined radar charts for community groups	240
Figure 8.9: Silhouette width values of group 1 community areas	241
Figure 8.10: Boxplot of standardised z-scores of the variables in group 1.....	242
Figure 8.11: Radar charts profile of the variables defining group 1	242
Figure 8.12: Silhouette width values of group 2 community areas	243
Figure 8.13: Boxplot of standardised z-scores of the variables in group 2.....	244
Figure 8.14: Radar charts profile of the variables defining group 2.....	244
Figure 8.15: Silhouette width values of group 3 community areas	245
Figure 8.16: Boxplot of standardised z-scores of the variables in group 3.....	246
Figure 8.17: Radar charts profile of the variables defining group 3.....	246
Figure 8.18: Silhouette width values of group 4 community areas	247
Figure 8.19: Boxplot of standardised z-scores of the variables in group 4.....	248
Figure 8.20: Radar charts profile of the variables defining group 4.....	249
Figure 8.21: Final cluster map	250

List of Tables

Table 1.1: Thesis outline in relation to research objectives	7
Table 2.1: Social capital indicators	17
Table 2.2: Limitations of previous studies	32
Table 3.1: Census variables and codes.....	43
Table 3.2: Components used to measure different diversity metrics.....	65
Table 4.1: Popular social media networking platforms	70
Table 4.2: Twitter public APIs	76
Table 4.3: Social media metrics.....	77
Table 4.4: Summary of studies using social media (Twitter) for social interactions and crime ..	83
Table 4.5: Facebook page metrics used for quantifying engagement.....	99
Table 5.1: The core components of crime and community cohesion, and the variables used to represent them in the model.	108
Table 5.2: Pearson’s correlation coefficient for burglary and other dependent variables (standard and diversity)	126
Table 5.3: Model summary of stepwise regression.....	127
Table 5.4: Coefficients and tests of model performance	128
Table 6.1: Tweets before and after processing.....	135
Table 6.2: Correlations between Twitter sentiments, diversity and standard variables.....	141
Table 6.3: Variance analysis between human annotations.....	143
Table 6.4: Pearson’s correlation analysis of the relationships between the AMT validation results and sentiment algorithm results.	144
Table 6.5: Interpretation of Fleiss’s κ . Source: Landis and Koch (1977)	147
Table 6.6: Inter-rater reliability test between human annotators	148
Table 6.7: Ternary and binary comparisons between human average and algorithm scores	148
Table 7.1 Typical metrics available on a Facebook page	175
Table 7.2: Descriptive statistics of posts interaction and types of posts	177
Table 7.3: Community areas engagement rate categories.....	181
Table 7.4: Pearson’s correlation coefficient between burglary crime rates, PCV indicators and diversity and standard variables.....	194

Table 7.5: Independent variables included in the model	195
Table 7.6: Summary of global OLS regression model	198
Table 7.7: Summary of local GWR model	200
Table 8.1: Variables used in the cluster analysis of Leeds community areas.	219
Table 8.2: Pearson's correlation of the clustering variables.....	221
Table 8.3: AIC for K=10 cluster solutions	231
Table 8.4: Performance of validity indices between K=2 and K=9 solutions.....	238

Abbreviations

ABM	Agent-based Models
ACORN	A Classification of Residential Neighbourhoods
AIC	Akaike Information Criteria
AICC	Akaike Information Criterion Corrected
AMT	Amazon Mechanical Turk
API	Application Programming Interface
ARI	Adjusted Rand Index
ASB	Anti-Social Behaviour
BBC	British Broadcasting Commission
BCS	British Crime Survey
BIC	Bayesian Information Criterion
BME	Black and Minority Ethnic
CA	Community Area
CCTV	Closed Circuit Television
CLARA	Clustering Large Applications
CLARANS	Clustering Large Applications based on Randomised Search
CSEW	Crime Statistics for England Wales
CV	Cross Validation
CVI	Clustering Validation Indices
DCLG	Department for Community and Local Government
DCMS	Digital Culture Media and Sports
GAM	Geographical Analysis Machine
GIS	Geographic Information System

GPS	Global Positioning System
GWR	Geographically Weighted Regression
HTI	Human Intelligence Task
ICC	Intra-Class Correlation
ICT	Information and Communication Technology
IMD	Index of Multiple Deprivation
IRR	Inter-Rater Reliability
IRT	Item Response Theory
KDE	Kernel Density Estimation
KIC	Kullback Information Criteria
LSOA	Lower Super Output Area
MATLAB	Matrix Laboratory
MAUP	Modifiable Areal Unit Problem
MSM	Microsimulation Models
MSOA	Middle Super Output Area
NLP	Natural Language Processing
NMI	Normalised Mutual Information
OA	Output Area
OLS	Ordinary Least Squares
ONS	Office for National Statistics
PAM	Partition Around Medoids
PCV	Popularity Commitment and Virality
PHMC	Philadelphia Health Management Corporation
ProsDES	Programme Development and Evaluation System

RAT	Routine Activity Theory
RI	Rand Index
SA	Sentiment Analysis
SEE	Standard Error of Estimate
SM	Social Media
SPSS	Statistical Package for the Social Sciences
SSE	Sum of Squared Error
TEC	The Electoral Commission
TM	Text Mining
TPF	Troubled Families Programme
UK	United Kingdom
USA	United States of America
USDOJ	United States Department of Justice
VIF	Variance Inflation Factor
WYP	West Yorkshire Police

Chapter 1

Introduction

1.1 Background of the Study

Community crime prevention through the role of neighbourhood cohesion has been recognised in policing strategies (Skogan, 1989). Crime reduction strategies are perceived to a large extent to depend on the levels of social ties (cohesion) and informal social control that exist in a neighbourhood. For example, neighbourhood social networks have been found to have a mediating effects on crime (Kubrin and Weitzer, 2003). The ability of residents to protect their neighbourhoods against crime, as well as their courage to confront and question any suspicious character is a manifestation of cohesive communities (Kelling and Stewart, 1990; Sampson *et al.*, 1997; Kubrin and Weitzer, 2003). In the UK, for example, crime prevention as provided by the state puts an emphasis on political approaches through partnership (such as the Neighbourhood Watch) with the police, local government and local communities (Hope, 2001). This policy was underpinned by the Crime and Disorder Act of 1998 which has provided impetus to the establishment of such partnerships. Consequently, reduction of crime should be seen as a collective responsibility of law enforcement agencies with the participation of community members (such as the Neighbourhood Watch) (Hope, 1988). The community has a great role to play in crime prevention by ensuring the establishment of informal social control in their respective neighbourhoods. Social cohesion and trust arise in response to socially unacceptable activities within the neighbourhood; this has developed into social processes where people interact meaningfully to find solutions to their problems, especially related to crime (Pauwels and Hardyns, 2009).

However, though qualitative approaches could be used to explore the complexities of social relationships understanding social cohesion, especially in relation to crime, can be better explored using a quantitative approach. The emergence and availability of new administrative and commercial data sources, coupled with uncertainties surrounding traditional small-scale surveys relied upon in empirical sociology, has resulted in researchers engaging with new crowd sourced data to gain insights into social processes (Savage and Burrows, 2007, p.113). The role

of social media in society is now a matter of interest in everyday debate not only in the neighbourhood context but also in many academic fields (Adolf and Deicke, 2014). The last decade has witnessed a tremendous increase in both number of online social networking services such as Twitter and Facebook as well as the users of such services (Labatut *et al.*, 2014). For example, the number of Twitter users increased from 200 million in 2011 to about 500 million in 2012 (Richard, 2013). These statistics of Twitter users further increases to over one billion subscribers at the end of 2014, about 78% of whom are on mobiles with over 500 million tweets sent on a daily basis (Bennet, 2014).

This research uses a quantitative approach to explore the impact of social cohesion in mediating crime through spatial analysis and social media data applications. The contention is that the use of social media data might offer an opportunity to gain new insights into the relationship between community cohesion and crime. Similarly, the use of social media data may offer new opportunities in the field of academic research, especially in community cohesion studies which are hitherto carried out using traditional survey data.

1.1.1 Problem Statement

The crime reduction effort is a collective responsibility of all citizens and not that of authorities alone. Recent figures (released April, 2018) from Crime Statistics for England and Wales (CSEW) have shown a decrease in some crimes. For example, a total of 10.7 million incidents of crime against households were recorded in 2017 compared to 11.5 million recorded in the previous year (2016) (representing 3.6% decrease). While theft of personal property, household theft, violence and criminal damage decreased by 10%, 9%, 7% and 5%; robbery, vehicle offences and violence increased by 33%, 17% and 9% from the previous year respectively (ONS, 2017a). Despite this development, the rates of crime in Leeds which ranks 16.6 crimes/1000 population, is still high compared to other UK cities (such as Birmingham 16.0/1000, Bradford 15.6/1000 and Sheffield 13.8/1000) (Home Office, 2017). Community/police crime prevention partnership approaches are largely based on the premise that no single agency can deal with, or be responsible for, dealing with complex community safety and crime problems (Berry *et al.*, 2011b). Therefore, crime reduction is perceived to rely, to a large extent, on the levels of social ties and collective action of the members of the community (Kubrin and Weitzer, 2003). Additionally, neighbourhood social networks have been found to have mediating effects on

crime (Kubrin and Weitzer, 2003) and hence recognised in policing strategies (Skogan, 1989). Furthermore, previous researchers investigating the impact of social cohesion on crime have used various traditional surveys, undertaken periodically which make the information they provide rather lacking value (Traunmueller *et al.*, 2014). Additionally, predictive analysis based on past crime events, as characterised by traditional methods, may be erroneous (Bowers *et al.*, 2004).

The technological developments which have culminated in the emergence of social media are changing the way social capital (a factor of social cohesion) is perceived, especially relating to life-styles and crime prevention. For example, the availability of the internet has enabled maintenance of social networks especially on social media (Facebook, Twitter, LinkedIn, Myspace etc.) (Debenham, 2002; Babb, 2005). Facebook and Twitter provides one of the important channels among all social media networks for the exchange of information and strengthening of social ties across different communities (Gorman, 2013). The new data sets being produced from social media networks could contribute to understanding what the role of social cohesion is and the metrics to quantify it, especially, in relation to crime. Moreover, researchers, especially those concerned with social capital, have argued that quantitative approaches provide a better understanding of social capital (Walford *et al.*, 2010). Therefore, the present research aims to explore new quantitative data sources such as social media in order to gain new insights into the relationship between community cohesion and crime.

1.2 Aim and Objectives

The main aim of this research project is to explore the impact of social cohesion in crime prevention through spatial analysis and use of social media data.

The following specific objectives have been outlined:-

1. To critically review the literature on the concept of community cohesion and its impacts on crime.
2. To use a range of spatial analysis methods to perform a detailed analysis of the relationship between crime and urban community form.
3. To critically review the literature on the feasibility of using new 'big' data sources (such as social media) to explore crime and community cohesion.

4. To develop a neighbourhood area classification profile based on community cohesion in a range of geographical locations of the study area at community area level.
5. To extend understanding of the relationship between crime and community cohesion, based on insights from traditional and new data sources.

1.2.1 Justification for the Study

Partnership between the police and communities is believed to have a positive effect in mediating crime and therefore in maintaining order and increasing the safety of neighbourhoods (Anderson, 2008). Participation and collaboration with the local communities is not only required but necessary in order to effectively deal with the issues of crime. In England and Wales, the establishment of Crime and Disorder Act of 1998 has formalised Crime Reduction Partnerships. This has paved the way for the establishment of community driven programmes such as Neighbourhood Watch which encourages community members to exercise informal social control with a view of reducing crime rates and promoting safety in their respective neighbourhoods.

Similarly, in the US, the Weed and Seed Initiative was established in 1991, which sought to develop a multi-agency approach to controlling and preventing violent crime and drug trafficking in high crime areas (Berry *et al.*, 2011a). More recently, Safer Neighbourhoods involving interventions targeted at crime reduction with a partnership approach were also established (Berry *et al.*, 2011a).

In the context of social science and crime in particular, the importance of maintaining social cohesion is paramount (Pauwels and Hardyns, 2009). Nowadays, addressing neighbourhood crime and disorder has become the collective responsibility of the law enforcement agencies and the citizens at both individual and community levels (Roehl *et al.*, 2006; Jim *et al.*, 2006; McGarrell *et al.*, 2010). Additionally, a number of studies have also stressed the role of social cohesion in mediating neighbourhood crimes especially burglary and violence (Sampson *et al.*, 1997; Morenoff *et al.*, 2001; Silver and Miller, 2004). However, the major limitations of most of these studies was small sample size and heavy reliance on data obtained from traditional surveys prone to bias (Fanelli, 2009). The present study hopes to address this limitation by using social media data that provides opportunity for assembling and analysing large amounts of information, a factor lacking in traditional methods (Malleon and Andresen, 2015). It was argued that the use

of social media data may be a more accurate representation of true opinion or behaviour (Japoc *et al.*, 2015), than the traditional methods such as the surveys and opinion polls (Asur and Huberman, 2010) and may likely replace traditional and more expensive surveys (Stieglitz *et al.*, 2018), though some studies suggested the combination of both methods (DWP, 2014; Felt, 2016). Additionally, Asur and Huberman (2010) emphasised that social media data can be used to make quantitative predictions, build models to aggregate the opinions of the collective population and gain useful insights into their behaviour. Furthermore, previous research has shown that crime flourishes in disadvantaged areas, more so than in affluent areas (Skogan, 1989; Gartner, 2013; Thompson and Gartner, 2014), because most persistent offenders come from disadvantaged backgrounds (Wikström and Treiber, 2016). Here again, this proposal needs to be re-investigated, especially as neighbourhood characteristics are not static.

1.2.2 Significance of the Study

This research work was motivated by the emphasis on community participation in crime prevention and the paucity of data to quantify the concept of community cohesion, especially through social media. It is anticipated that the findings of this study will benefit society, especially the law enforcement agencies since social media plays an important role in crime prevention today (Ronoh *et al.*, 2017).

The research could also provide better understanding of the relationship between community cohesion and crime through the application of new “non-traditional” social media data sources. Hence the study would also extend empirical discussion in the context of criminology by proposing a new empirical approach for quantifying community cohesion and crime. Furthermore, this study will be beneficial to policy makers as it will provide insight for re-evaluating the role of community cohesion on crime control. Similarly, the research could also assist future researchers to uncover the potentialities offered new data sources for exploring the impact of community cohesion on crime.

1.2.3 Defining Community in the Context of Crime

Defining the term ‘community’ is very complex as different people have different views about their communities (MacQueen *et al.*, 2001). The concept of community, although disputed, also posits the notion of collectivity and commonality (Haq, 2006). However, MacQueen *et al.*

(2001) defines community as “group of people with diverse characteristics who are linked by social ties, share common perceptions and engaged in joint action in geographical locations or settings” (MacQueen *et al.*, 2001 p.1929). Similarly, Lee (1992) also defined community as a group of people who have shared common thoughts.

Therefore, in the context of crime, ‘community’ may be defined in terms of participation and involvement of members in collective action to prevent crimes in their neighbourhoods (Stenson, 1993). However, this does not suggest a fixed definition of community but rather a community in crime prevention and practice (Crawford, 1994). With the advent of the internet, the concept of community potentially transcends the geographical boundary as people can now gather virtually through social media platforms such as Twitter and Facebook and share common interests regardless of physical location (Ting, 2011). For example in the UK, social media platforms, particularly Facebook and Twitter are increasingly being used by law enforcement agencies to engage the public in crime prevention activities (Crump, 2011). Social media channels can be used to communicate to people quickly about locations of criminal events. For example, tweets can help the police to immediately alert the community and to request the general public for assistance with useful information that may lead to the arrest of the offenders (Glodava, 2015).

Therefore, for the purpose of this research project, the term ‘community’ is defined as one where people collectively come together (virtual and physical cohesion) to improve the security of their geographical areas with a view to reducing crime. Additionally, throughout this thesis the terms *community cohesion* and *social cohesion* may be used interchangeably.

1.3 Thesis Structure

This thesis will comprise nine chapters organised sequentially to reflect the development of the research bearing in mind the stated objectives outlined above. Table 1.1 outlines the chapter structure of the thesis in relation to the research objectives.

Chapter 2 is a critical review of literature on the concept of community cohesion and crime. Social capital and its importance in crime including its determinants are outlined. The effects of collective efficacy on social cohesion and by extension crime are discussed here. To ground the research into context, crime theories related to the research are discussed. This chapter aims to achieve objective 1 of the thesis.

Chapter 3 proceeds to describe and discuss quantitative methods of modelling urban social systems and data sources employed for analysis. Relevant datasets such as police recorded crime and census statistics required for the analyses and their sources are also highlighted. A range of spatial and statistical techniques, their relative merits, as well as application in community and crime research is examined. This chapter is designed to achieve objective 2 outlined in the thesis.

Table 1.1: Thesis outline in relation to research objectives

Chapter	Objective
<u>Chapter 2</u> : Concept of Community Cohesion and Crime: A Literature Review	1
<u>Chapter 3</u> : Traditional methods of Modelling Urban Social Systems and Crime	2
<u>Chapter 4</u> : Non-Traditional methods: Social Media for Community Cohesion and Crime	3
<u>Chapter 5</u> : Traditional Modelling of Crime and Community Cohesion	2 & 3
<u>Chapter 6</u> : Social Media Analysis for Understanding Community Cohesion	3
<u>Chapter 7</u> : New Social Media Metrics for Quantifying Community cohesion and Crime	4
<u>Chapter 8</u> : Community Cohesion and Crime: A New Leeds Community Areas Classification Profile	5

Chapter 4 reviews literature on social media, community cohesion and crime as it relates to action in communities and for understanding communities. The importance of social media in law enforcement activities is also discussed. Social media channels relevant to this research and the procedures used to access social media data: from Facebook and Twitter specifically are presented. Different methods and algorithms used for social media data analysis, such as to quantify sentiment and engagement are described. This chapter is set to achieve objective 3 of the thesis.

Having identified determinants of social capital and crime in chapter 2, and described the statistical methods in chapter 3, **chapter 5** moves on to develop metrics of diversity of urban community form and apply them in a multiple regression model to explore the relationship with burglary crime rates. Procedure and justification for variable selection into the model are described. This chapter relates to objectives 2 and 3 of the thesis.

Early in the research (chapter 4), the relationships between social media and community cohesion are examined. Building on insights from the regression model in chapter 5, **chapter 6** proceeds to use Twitter data to quantify public sentiment (mood) to explore relationships between of social cohesion and burglary crime rates using Leeds community areas as a spatial framework. This chapter aims to achieve objective 3 of the thesis.

Drawing on new insights about the relationship between community cohesion and crime based on analyses carried out in chapters 5, and 6 using traditional and new social media data sources, **Chapter 7** presents a set of new Facebook metrics (popularity, commitment and virality) for quantifying the relationship between engagement (community cohesion) and burglary crime rates in different local community areas. This chapter is designed to achieve objective 5 of the thesis.

Chapter 8 combines insights derived from chapters 5, 6 and 7 based on a new perception of community cohesion to classify a range of community areas in Leeds. This chapter is designed to achieve objective 4.

Finally, **Chapter 9** concludes the thesis by presenting a summary of the findings, limitations and recommendations for the future studies.

Chapter 2

Concept of Community Cohesion and Crime: A Literature Review

2.1 Introduction

The Cantle report highlighted that lack of mutual interactions (social cohesion) between groups in the same community as well as deprivation were the major issues limiting collective efficacy (informal social control) in a community; in such a situation, the community becomes less cohesive (Cantle, 2001). A decline in social cohesion in a community is often linked to increases in crime rates (Hirschfield and Bowers, 1997; Kawachi and Berkman, 2000; Forrest and Kearns, 2001; Goudriaan *et al.*, 2006). A literature review is usually meant to contextualise a field of study, to provide direction and support for new data gathering. In this chapter, we survey important literature on the concept of community cohesion and crime. The chapter begins with the definition of the concept (Section 2.2) and outlines the context of community cohesion and crime (Section 2.3); it then proceeds to describe the importance of social capital in crime (Section 2.4). Determinants of social capital in relation to crime are described in Section 2.5. The relationship between collective efficacy and crime is described in Section 2.6. The weaknesses of previous studies are outlined in Section 2.7 and a review of theoretical framework is presented in Section 2.8 while concluding remarks are presented in Section 2.9.

2.2 Defining the Concept of Community Cohesion

The concept of community cohesion and the issue of integration have been a matter of public discourse with a long history dating back to the 20th century when rapid urbanisation produced a strong social order alongside religious and diverse ethnic lines (Forrest and Kearns, 2001). In the UK, however, the concept of community cohesion came into the limelight in 2001 as a result of the disturbances that occurred in Bradford, Oldham and Burnley, where aggrieved ethnic minority groups protested against perceived marginalisation by the White majority populations.

The then British Home Secretary, David Blunkett, in a lecture delivered in New York 2003, argued that:

“True freedom enables individuals to participate actively in creating their own solutions to community problems; consciously engage in political discourse and influence its outcome; and become flourishing contributors to their own society ” (Blunkett, 2003 in Burnett, 2004, p.1).

Defining the concept of community cohesion is difficult as there is no universal definition for it (Hardyns and Pauwels, 2009). However, various authors and government agencies have made different attempts to define the concept. According to Cantle (2001), the focus of community cohesion is integration, distribution of wealth and social well-being of the people. The term community cohesion has been defined as interaction between different groups of people accepting one another with a view to living harmoniously together (Cooper and Innes, 2009). Socially cohesive areas according to Hirschfield and Bowers (1997) are communities where a sense of interaction and a sense of belonging exist. Forrest and Kearns (2001) have viewed community cohesion in communities in terms of common values, social order and social solidarity, integration and strong attachment.

Despite the lack of consensus on the appropriate definition of community cohesion, a broader working definition has been adopted for the UK. The Local Government Association (2002) defined cohesive communities as “where there is a common vision, sense of belonging, respect for diversity and similar life opportunities for all members of communities irrespective of their backgrounds” (Local Government Association, 2002, p.6).

2.3 Community Cohesion and Crime

Community crime prevention refers to actions intended to change the social conditions that are believed to sustain crime in a given residential community (Hope, 1995). This approach emphasises community involvement in crime prevention efforts in contrast to formal crime preventions by law enforcement agencies (Gill *et al.*, 2014). Crime does not only inflict fear on its victims but also impedes a community’s potential for establishing social ties among its members (Hartnagel, 1979; Taylor, 1995; Zani *et al.*, 2001; Garcia *et al.*, 2007). Other authors,

for example, Brantingham and Brantingham (1991) and Woldoff (2002), argued that deviant behaviour from within or outside the community encourages human interaction. Despite a wide range of strategies aimed at crime reduction, the effectiveness of the judicial system was argued to have limited impact in preventing crime (Rosenbaum, 1988; Sherman *et al.*, 1997; Homel, 2006). Therefore, as formal means of fighting crime become less effective, community crime prevention have emerged as an alternative way for reducing crime in neighbourhoods (Rosenbaum, 1988). However, reducing crime in urban neighbourhoods is linked to levels of informal social control defined as willingness of residents to intervene in local problems (Sampson and Groves, 1989; Bursik Jr and Grasmick, 1993; Sampson *et al.*, 1997; Sampson *et al.*, 2002; Silver and Miller, 2004; Burchfield, 2009; Warner *et al.*, 2010).

Gendrot (2001), in a comparative study of policies about crime and fear of crime in urban areas of France, USA and UK, noted that local authorities, police, educators and citizen participation are necessary for effective local crime prevention activities. He also stressed that the social capital of a neighbourhood can help to reduce social disorder in urban areas, indicating the place of collaboration for effective crime prevention efforts.

Previous researchers have used different traditional datasets to explore relationships between urban structural characteristics and neighbourhood crime. For example, Sampson and Groves (1989) used the British Crime Survey of 1982 and 1984. Employing socio-economic status, ethnic heterogeneity and residential mobility and family disruptions as dependent variables, they empirically tested the role of community structures, especially local social ties and participation, in reducing crime. They found that strong correlations exist between community structural characteristics and levels of crime rates and especially burglary and street crimes are directly reduced in cohesive communities. Building on Sampson and Groves (1989), Markowitz *et al.* (2001) also used the British Crime Survey of 1984, 1988 and 1992 to investigate the relationship between cohesion and disorder. As with the previous study, they found that low socio-economic status, ethnic heterogeneity, family disruptions, and residential instability are associated with increase in disorder, while cohesion reduces disorder in neighbourhoods.

Furthermore, Lee (2000) used cross-national and city level surveys from 27 countries to study the relationship between community cohesion and violent crime victimisation. He found that people who live in cohesive communities are less likely to experience violent crime.

Additionally, a community that has a higher level of informal social control is where members of such a community are more likely to intervene in public deviant activities. This concept is described as 'collective efficacy' (see Sampson *et al.*, 1997). He also noted that socio-economic, demographic, neighbourhood characteristics, life-styles and community cohesion reduce the likelihood of robbery and assault, suggesting that community cohesion may be important determinant of victimisation risk.

Wedlock (2006) used the Local Areas Boost to the 2003 Citizenship Survey to investigate the relationship between community cohesion and reported level of crime. She found that sense of community (a factor of community cohesion) and levels of multiple deprivation of an area are the key predictors of violent crime.

Cooper and Innes (2009) employed the British Crime Survey of 2007 and the Community Life Survey of Wales to examine the relationship between community cohesion and crime, looking at the variables of age and socio-economic status of the people and how these variables influence crime. They found that the community's level of cohesion depends to a large extent on its level of ties and that diversity and deprivation compromise cohesion. They also emphasised that community cohesion is important for understanding public perceptions of crime and that different area levels of community cohesion are linked to their social and economic context.

There has been a growing debate on the impact of community cohesion on crime. For example, Farkas and Jones (2007) argued that the greater the levels of community cohesion in an area, the greater the levels of social control which in effect check criminal activity. Similarly, Rosenbaum (1987) further observed that high levels of crimes are associated with social disorganisation in communities. Consequently, communities with strong network ties and social interactions are likely to have lower crime rates (Sampson and Groves, 1989). Kawachi and Berkman (2000) explained that the relationship between crime and community characteristics is better understood when viewed in terms of social capital as a factor of community cohesion.

Community cohesion generally acts to increase the safety of communities by reducing the socio-economic drivers of crime, through maintaining oversight of those potentially moving into criminal lifestyles (Lee, 2000), to increasing the oversight of potential sites of crime, and by reporting crimes when they occur. However, cohesion is a nuanced concept (there is considerable cohesion in communities ruled by criminal gangs) and cohesion is ill-represented

by standard socio-demographic variables (both middle and working class communities can experience a wide range of levels of cohesion). As such, cohesion is poorly captured in standard regression models of crime. In this research, the adjustment of standard regression variables is suggested. A set of 'adjusted' variables can better capture the range of loci in which social cohesion plays a part across the crime system. In addition, when these adjustments are made, these variables become more strongly predictive of crime than standard variables, suggesting the significant role social cohesion plays in the crime system and the significant role it plays as the link between standard regression variables and burglary crime rates.

2.4 Importance of Social Capital on Crime

The concept of social capital was pioneered by Durkheim (1893) who advocated the impact of social capital especially in addressing social problems (such as crime) in societies. The concept of social capital has no universal consensus definition. However, researchers (Coleman, 1990; Putnam, 1995) have attempted to define social capital as those features of the social organisations such as networks, trust and reciprocity, that facilitate mutual coexistence among the community. Social capital has also been defined as the chain of cooperative relationships between citizens that facilitate collective action towards the resolution of problems (Brehm and Rahn, 1997), the ability for people to work together for common purposes in groups and organisations (Fukuyama, 1996). Therefore, quality of life and the performance of social institutions are influenced by norms of networks and civic engagement between residents of working together to improve the conditions of their community (Putnam, 1995).

Social capital reflects the relationship between people and the outcomes of their relationships. Therefore, social capital is part of the daily social life and behaviour of people in the society (Sander and Teh, 2014). Social scientists have emphasised two main dimensions of social capital: social glue and social bridges. While social glue refers to the degree to which people take part in group life, social bridges on the other hand are the links between groups (Briggs and Wilson, 2005, p.152). Additionally, a reflection of high social capital can be seen in societies where people are committed, tolerant and help each other for their common good (ONS, 2015a). Therefore, Dodge and Kitchin (2007) argued that the overall importance of examining social capital rests in the belief that social networks have influence, especially in mediating crime and creating safer neighbourhoods.

Furthermore, Dodge and Kitchin (2007) suggest that the potential for social capital to make a positive contribution to outcomes in areas of social concern such as community safety has attracted the attention of policy makers and researchers. Studies have demonstrated that crime rates are closely linked to levels of social capital (Sampson and Raudenbush, 1999; Healy and Côté, 2001; Harper, 2001; Haezwindt, 2003; Akçomak and Ter Weel, 2012; Moore and Recker, 2013; Turcotte, 2015). Akçomak and Ter Weel (2012) maintain that social capital increases the risk and probability that the potential offenders might be caught in the process, which in turn reduces the crime rates in the neighbourhoods. For example, in societies where the levels of social capital and social bonds are strong, offenders are less likely to victimise people because they (offenders) tend to risk being apprehended as residents tend to look out for each other (Buonanno *et al.*, 2009). This notion was earlier highlighted by Putnam (1995) who stressed that levels of crime in a given society are related to the levels and strength of social connectedness and civic engagement which is a product of social capital existing in such societies. Additionally, Putnam (2000) also stressed that absence of social capital in the community may promote youths to engage in criminal activities (Putnam, 2000). Temkin and Rohe (1998) also argued that social capital can be used to predict neighbourhood stability.

There are a number of studies which seek to explore the relationship between social capital and crime. For example, Kennedy *et al.* (1998) explore the effect of social capital on fire arms violence using the US General Social Survey. They found that a decrease in social capital is associated with an increase in fire-arms violence. They suggested that increasing social capital and reduction of wealth disparity among people can have significant effect on crime, especially fire-arms and violence.

Additionally, Carcach and Huntley (2002) used Local Government areas data in mainland Australia to study the relationship between social capital and crime. They found that crime rates tend to be lower in local areas with high levels of participation (a factor of social capital) in local community oriented activities. Participation increases opportunities for social interaction, encouraging collective action towards local problems and enhancing public safety.

Kruger *et al.* (2007) attempted to explore the effect of social capital on assault using random sample of Census data of Genesee County in Michigan, USA. They found that neighbourhood social capital moderates the impact of assault injury rates on fear of crime. Deller and Deller

(2010), using Metropolitan US county data, maintain the influence of social capital in understanding patterns of rural crime pressures and that social capital can help mediate such pressures.

Additionally, Deller and Deller (2012) also emphasised that social capital has more significant impacts on property crime such as burglary and theft than on other crimes. Property crime takes place in public arenas where others may view the offence and hence the potential effect of social capital (Moore and Recker, 2013). Burglary, specifically, is an opportunity crime that flourishes in socially disorganised and less cohesive communities (Weisburd and Piquero, 2008). Disorganised neighbourhoods are more likely to be affected, because of weak social capital (Dunaway *et al.*, 2000). However, Moore and Recker (2013) argued that violent crimes often operate behind closed doors, where expected neighbourhood social capital may not be effective.

Akçomak and Ter Weel (2012) explore the effect of social capital in the Netherlands using data from 140 Dutch municipalities. They found out that variation in crime rate is associated with variation in social capital across cities. While the importance of social capital in mediating different types crime has been established, an effective way of measuring it still remains a challenging task among researchers (Harper, 2001; Babb, 2005).

Additionally, a wide range of literature has demonstrated the impact of social capital in mediating crime in society, however the data used are mostly US based (Sampson *et al.*, 1997; Kennedy *et al.*, 1998; Kruger *et al.*, 2007; Deller and Deller, 2010; 2012). Therefore the application of the findings needs to be validated in a UK situation. In this research, we used data from the UK in order to examine the importance of social capital sub-component of community cohesion, especially in mediating crime.

2.5 Determinants of Social Capital in Relation to Crime

Social capital is important for crime prevention but we do not know what it really is because of the paucity of appropriate data to empirically measure it (Kanazawa and Savage, 2009). Dodge and Kitchin (2007) emphasised that social capital data may help to give a better understanding of crime and community safety issues which may provide information that may enhance operational policing strategies (Dodge and Kitchin, 2007).

There has been much debate in the academic field, both within the UK and internationally, as to what social capital means and how best to measure it (Babb, 2005). The difficulty associated with defining social capital and its measurement has continued to make its quantification rather complex (Harper, 2001). Researchers have stressed that social capital is difficult to quantify, because no single indicator exists to measure it (Haezwindt, 2003; Plotkowiak, 2014; Appel *et al.*, 2014).

Measurement of social capital appears to be a problematic and complex task. However, Van Der Gaag and Snijders (2005) suggest that attempts to measure social capital should consider a range of indicators rather than focusing on a single variable. Therefore, to measure social capital, Woolcock and Narayan (2000) explained that contemporary researchers have to compile variables from a range of different sources.

Previous researchers (Coleman, 1988; Hall, 1999; Putnam, 2000) have used different indicators in attempting to measure social capital in relation to crime but the social capital indicators they have used merely focused on voluntary and civic participation which belong to a unique set of social constructs (Appel *et al.*, 2014). Because of this limitation, Moore and Recker (2013) argued that social capital has not been effectively defined or measured across previous studies especially with regards to crime.

Additionally, previous studies that have attempted to quantify the relationship between social capital and crime relied heavily on data from traditional sources, such as community surveys, to measure social capital where questions can easily be crafted and prone to bias; hence the reliability of those measures may be questioned (Deller and Deller, 2010). This shortcoming, coupled with validity issues, has remained a point of criticism (Appel *et al.*, 2014). To address this limitation, Appel *et al.* (2014) maintain that one measure that combines a number of constructs into a new measure of social capital is particularly internet social capital.

Technological developments such as the social media are changing the way social capital is perceived, especially relating to life-styles and crime prevention. For example, social networks are now being maintained through the internet, especially through social media such as Twitter and Facebook (Babb, 2005). This trend suggests that social media are potentially important for explaining social capital and metrics to quantify it specifically in relation to crime. Hence, Putnam (2000) argued that social capital is often better understood using quantitative

approaches. Therefore, social cohesion is only understandable when the context in which it is being assessed is determined; studying crime in this research is considered appropriate.

While it is apparent that social media enhances the formation of social capital (Ahn, 2012), it also helps in mediating crime (Akçomak and Ter Weel, 2012). However, researchers such as Moore and Recker (2013) have stressed the inadequacy of previous studies to effectively measure social capital. To address the limitations highlighted, this research uses sets of new metrics of social media, especially Facebook and Twitter, to construct a new meaning to social capital (a factor of community cohesion) through quantifying its importance in the crime system.

Previous studies have demonstrated that social capital is measured differently and that certain measures tend to have larger effects on crime than others (Deller and Deller, 2010; Moore and Recker, 2013). Therefore, in order to quantitatively measure the impact of social capital especially on crime and for the purpose of robustness, this research examines a range of variables (datasets) drawn and compiled from different data sources such as big data (social media) and UK Census data (socio-economic and demographic) to explore the effect of each variable on crime. For example, variables such as age, gender, ethnic diversity, length of residence, educational qualifications and deprivation, as well as social media usage (social networking), may have different effects on social capital and crime (Moore and Recker, 2013).

Table 2.1 shows social capital indicators used in previous studies.

Table 2.1: Social capital indicators

Social Capital (Variables)	Description	Reference
Age	Census 2011 KS102EW (ONS) Leeds LSOA	Levitt (1999), Johnston and Matthews (2004), Ponce <i>et al.</i> (2014), ONS (2015a)
Gender	Census 2011 QS104EW (ONS) Leeds LSOA	Tittle <i>et al.</i> (2003), Ponce <i>et al.</i> (2014), ONS (2015a)
Participation in clubs and associations	Lifestyle and social participation ST41 2012 (ONS) National	Putnam (1995), Wollebæk and Selle (2003), Johnston and Matthews (2004), Mandarano <i>et al.</i> (2010), Caruso (2011), Sagar <i>et al.</i> (2011), Ponce <i>et al.</i> (2014), ONS (2015a)

Social Capital (Variables)	Description	Reference
Internet usage (social networking)	e-society ST41 2014 (ONS) National	Wellman <i>et al.</i> (2001), ONS (2015a)
Online membership (Social media)	e-society ST41 2014 (ONS) National	Sajuria <i>et al.</i> (2014), Burke <i>et al.</i> (2011), Plotkowiak (2014), (Sander and Teh, 2014)
Heterogeneity	Census 2011 KS201EW (ONS) Leeds LSOA	Gilchrist and Kyprianou (2011), Birani and Lehmann (2013), Kindler <i>et al.</i> (2014)
Educational qualification	Census 2011 KS501EW (ONS) Leeds LSOA	Moretti (2005), Helliwell and Putnam (1999), Deller and Deller (2010), Machin <i>et al.</i> (2011), Caruso (2011)
Election turnout	Electoral Commission 2010 ST41 (ONS) National	Putnam (1995), Akçomak and Ter Weel (2012), Johnston and Matthews (2004), ONS (2015a)
Length of residence	Census 2011 QS803EW (ONS) Leeds LSOA	Yamamura (2011), Keene <i>et al.</i> (2013)
Deprivation	Index of Multiple Deprivation (IMD)	DCLG (2015), (Drukker <i>et al.</i> , 2003)

2.5.1 Age, Social Capital and Crime

People of different ages might have different levels of social capital. The levels of social interaction might also differ among different age groups since younger people are less likely to build social capital (particularly face-to-face) than older people (Johnston and Matthews, 2004). This is because young people (16-24) are less likely to participate in local community activities (Whiting and Harper, 2003). Ponce *et al.* (2014) argued that social capital increases with advancing age. Interaction and participation are important in social capital; in the UK, recent research has shown that elderly people have more interaction with neighbours than young persons (ONS, 2015a). This trend may be attributed to the fact the younger people tend to have more social capital through social media and interact more via the Internet, especially through

social media platforms. On the other hand, Tittle *et al.* (2003) argued that the age crime curve tends to increase from the adolescent years reaching maximum at adulthood and then sharply declines. This however depends on the type of crime.

Furthermore, Levitt (1999) stressed that population groups aged 16-24 and 25-39 are more heavily involved in criminal activities than those 65 and over (Figure 2.1). Similarly, Kanazawa (2003) maintains that, just as productivity declines with age, so also crime tends to reduce with growing age. However, Blonigen (2010) also argued that the links between crime and age is very difficult one but he agrees with Tittle *et al.* (2003) on the age-crime curve concept.

Therefore, given the limitation of previous studies that examine the relationship between age and crime as a result of normative changes (Blonigen, 2010) and as an outcome of self-control (Tittle *et al.*, 2003), the present study intends to explore other possibilities of link between age and crime.

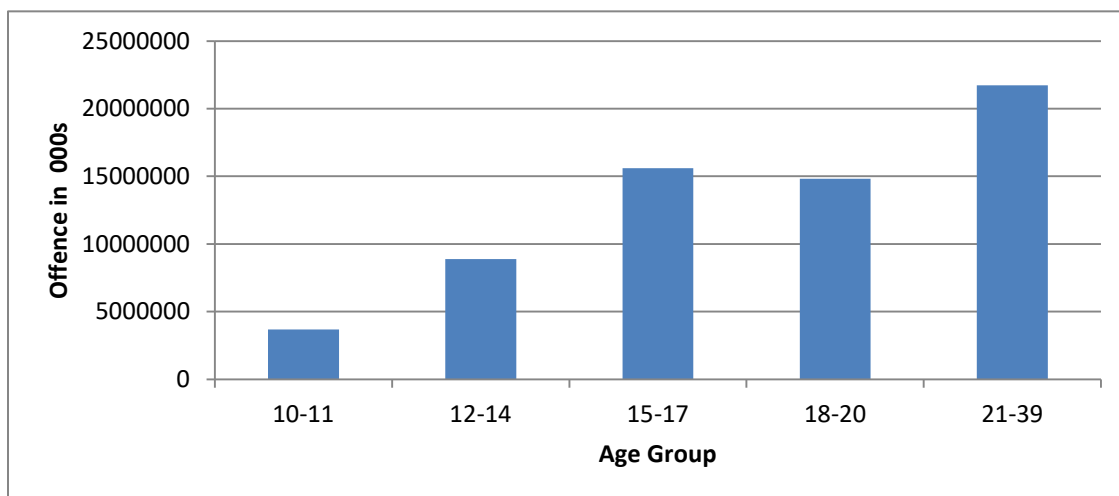


Figure 2.1: Crimes by age group. Source: Criminal Justice Statistics, 2014 (ONS, 2014c)

2.5.2 Gender, Social Capital and Crime

Social capital accumulation and its utilisation are related to and segmented along gender lines. Men are less likely to regularly and socially interact with neighbours than women across all ages (Ponce *et al.*, 2014; ONS, 2015a). However, in terms of social capital, Timberlake (2005) argued that women lag behind men due to their limited access to education, resources and networks which are essential for establishing social capital. Similarly, Migheli (2007) emphasised that women tend to have less trust than men therefore are less likely to invest in social capital. Van

(2006) also maintains that men are more effective in establishing social ties which is essential for creating social capital than women. On the other hand, Westermann *et al.* (2005) argued that the discrimination between men and women in terms of their involvement in social capital requires further investigation and empirical testing.

In terms of crime, differences exist between men and women. For example, research has shown that except for minor offences, males have a higher probability to commit crimes than females (Tittle *et al.*, 2003). Heidensohn and Gelsthorpe (2012) stressed that men tend to commit more crime than women. Statistics in England generally have indicated that males consistently constitute the vast the majority of offenders (about 77%) with females undertaking only 23% (ONS, 2014d). However, ONS (2014d) explains that the type of offences committed by females tends to be different and less serious compared to males. Figure 2.2 shows the comparison of interaction between men and women across different age groups. As noted earlier, social interactions increase with age and women tend to have more social interactions across the range of age groups. Figure 2.3 compares crime by gender. It was earlier noted that men tend to have a higher probability to commit crime than women.

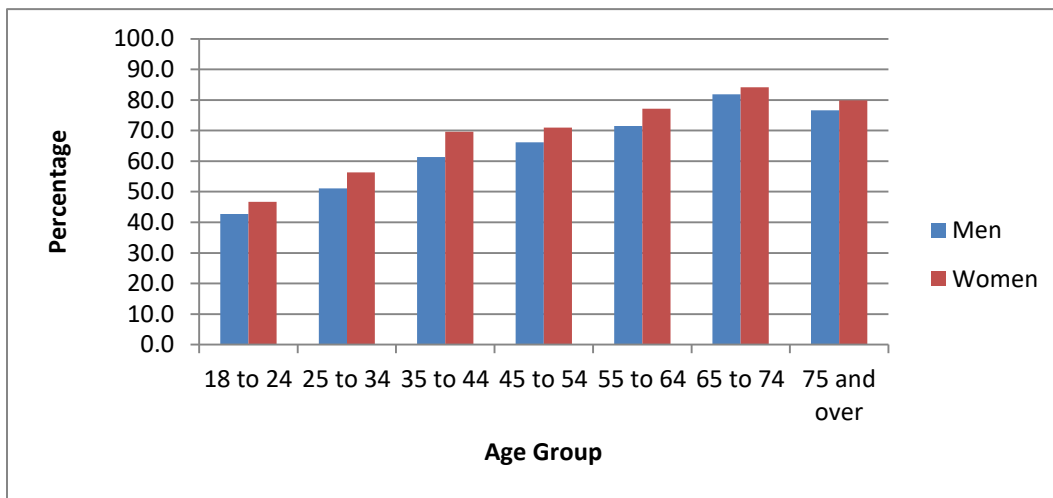


Figure 2.2: Social interaction between men and women. Source: ONS (2014e)

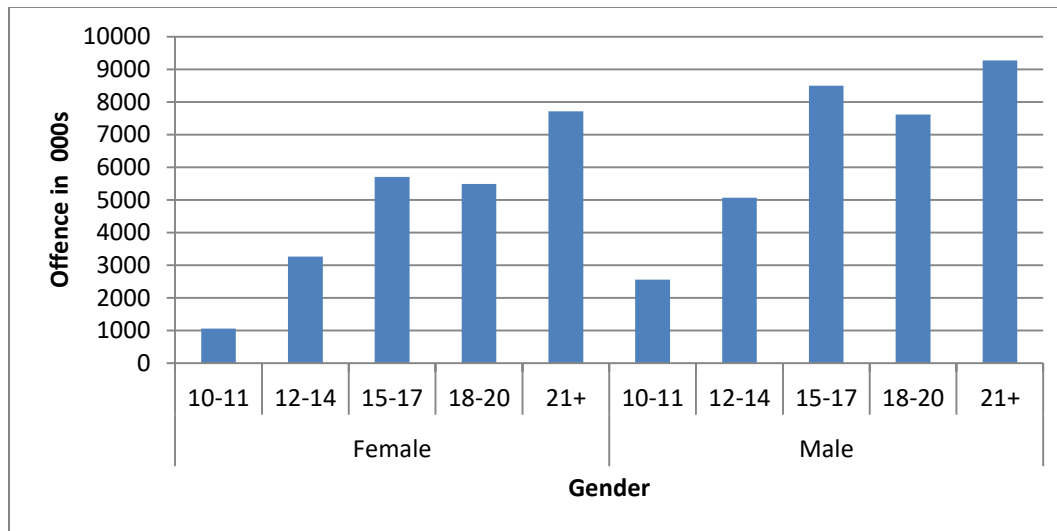


Figure 2.3: Crime by gender. Source: Criminal Justice Statistics, 2014 (ONS, 2014c)

2.5.3 Internet Usage and Social Capital

The internet provides a new means of interaction and socialisation that can enhance face-to-face relations. The use of the internet is more linked and associated with certain age groups than others. For example, younger people use the internet more frequently than older people for social networking, especially through social media platforms (Twitter and Facebook) to share information and common interest (ONS, 2015a). Johnston and Matthews (2004) argued that young people are less likely to have face-to-face social interactions than older people. Figure 2.4 describes the levels of internet usage by different categories of people.

Resnick (2001) and Wellman (2006) have maintained that internet use, especially through social media, greatly influences people's ability to form and maintain social capital, given that it provides the opportunity for interaction and information sharing in a variety of ways. Similarly, Ahn (2012) argues that social media network platforms may assist in cultivating social capital in both online and offline relationships.

Furthermore, Wellman *et al.* (2001) emphasised that rapidly expanding internet access has been a big hope in stimulating positive change by creating new online interaction and enhancing offline relationships. Additionally, Pénard and Poussing (2010) stressed that the internet may change the nature of an individual's social capital by enabling the accumulation of virtual social capital.

While previous studies have emphasised the possibility of the using internet as a platform especially through social media for creation of social capital (Resnick, 2001; Wellman, 2006), they fail to provide adequate explanation as to how these medium can be applied to quantify social capital and by extension its mediating effects on crime.

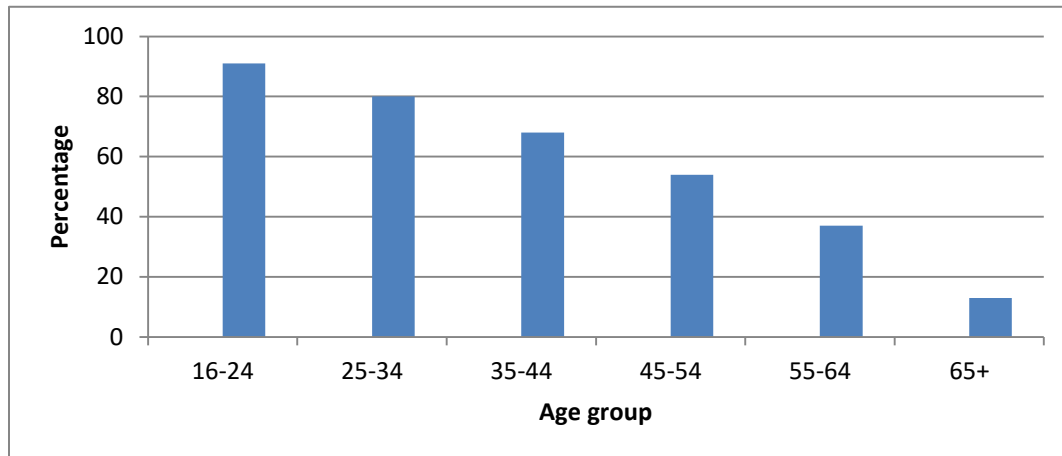


Figure 2.4: Internet usage by age group in 2014. Source: ONS (2014d)

2.5.4 Online Membership and Social Capital

Online community membership involves gathering of like-minded people brought together in cyberspace through shared interests known as hidden social capital (Phulari *et al.*, 2010). The influence of online community on social capital, especially through information sharing, has been highlighted by Chiu *et al.* (2006). Studies have demonstrated that social capital does not apply to traditional real life social networks alone, but can also be created online (Plotkowiak, 2014; Burke *et al.*, 2011). There is a growing evidence suggesting that social network sites (such as Twitter and Facebook) have the potential to generate social capital (Gainous and Wagner, 2014). Participants may use the sites to interact with people they already know or to meet with new people; in such situations, social bonds develop which motivate social capital.

Nahapiet and Ghoshal (1998) highlighted three dimensions of online social network interaction namely: structural, relational and cognitive. Structural social capital refers to patterns of communication between members while relational social capital indicates particular relationships that members have with one another and cognitive social capital refers to shared code or language used among members (Nahapiet and Ghoshal, 1998).

However, Ellison *et al.* (2007) argued that the internet has been linked to both increases and decreases of social capital. Nie (2001) stressed that online membership (internet use) might reduce face-to-face social interaction and inturn reduce social capital. On the other hand, Bargh and McKenna (2004) emphasised that online interactions enhance social capital and therefore has significant impact in social life. Additionally, Katz and Rice (2002) highlighted that online interactions through social network sites may be used to form social capital and that the internet has a positive impact on social interaction both online and offline. Similarly, Sander and Teh (2014) also explained that social networking creates social capital because, by belonging to online networks, the distance between people is diminished by sharing information in real time. Recently, Sajuria *et al.* (2014) used Twitter data from three events: the US Occupy movement in 2011, UK based IF Campaign of 2013 and the Chilean Presidential Elections 2013 to empirically test Putnam’s social capital theory. They found that online social interactions promote social capital; Twitter specifically has the potential for creating two forms of social capital: bonding and bridging beyond what the theory predicts. Bonding social capital refers to consolidation of existing ties within homogenous groups while bridging social capital relates to linking otherwise separate heterogeneous groups (Sajuria *et al.*, 2014). Figure 2.5 shows the growth in household’s access to internet in the UK. A steady increase in online activities by individuals indicates the potential of maintaining social capital through online interactions.

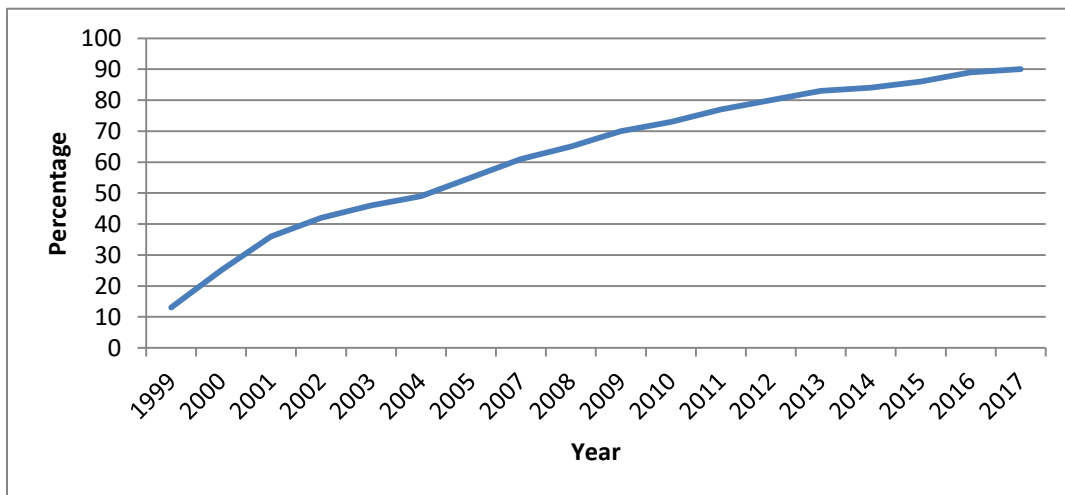


Figure 2.5: Percentage of UK households with internet. Source: ONS (2017b)

2.5.5 Participation in Community Activities, Social Capital and Crime

Participation in community activities such as clubs and associations and voluntary work has been recognised as a factor contributing to social capital. For example, Delaney and Keaney (2005) stressed that participation in sporting activities such as a local football or amateur dramatics can enhance creation of social capital in a society. In terms of participation in sports clubs and voluntary associations, research has demonstrated that men are more likely to be actively involved than women across all age groups (Wollebæk and Selle, 2003; Johnston and Matthews, 2004; Mandarano *et al.*, 2010; Ponce *et al.*, 2014; ONS, 2015a).

Furthermore, studies have indicated a relationship between participation in sports and crime control. Okayasu *et al.* (2010) have argued that social capital has been emphasised as a means of reducing community social problems such as crime, and participation in sports is considered as one of the catalysts in achieving that objective. For example, Caruso (2011) explores the impact of sports participation on crime in Italy using OLS regression. He found that an increase of 1% in sports participation significantly reduces the level of property crime by 0.3% and juvenile crime by 0.8% respectively. However, violent crime was found to be positively associated with sports participation, where an increase of 1% in sports participation increases the rate of violence by 0.4% (Caruso, 2011).

Sagar *et al.* (2011) examined the relationship between sports participation and crime in Britain using the participation of university students in sports. They found out that participation in sports negatively effects antisocial behaviour. However, Hastad *et al.* (1984) argued that there is no relationship between sports participation and deviant behaviour. Their findings indicated negative association between youth sports participation and deviant activities. Figure 2.6 indicates the levels of participation in clubs and associations for men and women cutting across different age groups. Therefore, sports participation may have an important role to play in both building social capital as well as addressing crime.

Additionally, Kang (2015) emphasised that participation in voluntary work such as neighbourhood watch scheme can promote a sense of community, social interaction, as well as safety in neighbourhoods.

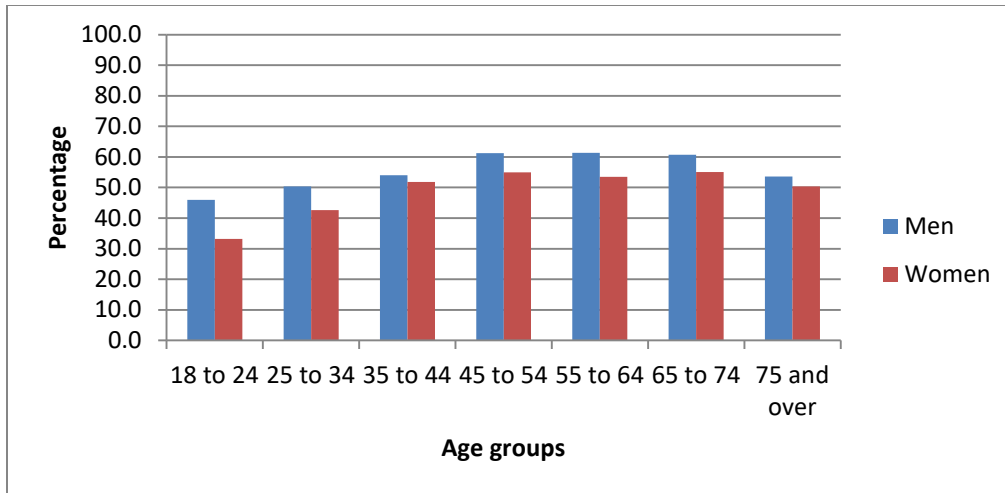


Figure 2.6: Participation in community activities by gender and age. Source: DCMS (2010)

2.5.6 Participation in Election, Social Capital and Crime

Participation in elections (voter turnout) is one of the important aspects of social capital in society. Research has shown that older people tend to participate more actively in political processes such as voting in an election than the younger ones (Johnston and Matthews, 2004; ONS, 2015a). In the UK, for example, statistics have shown that trends in voting has been consistent over the years (TEC, 2010). Figure 2.7 shows the percentage of turnout in the general elections by year and age group. This trend might be because younger people and (of both sexes) are less interested in politics than older people.

Coleman (2002) and Santangelo (2011) have all highlighted that voter turnout is associated with lower crime rates. Santangelo (2011) also stressed that the more people are connected with society, the less likely they are to commit crimes that affect other members of society as they feel greater responsibility for upholding common values.

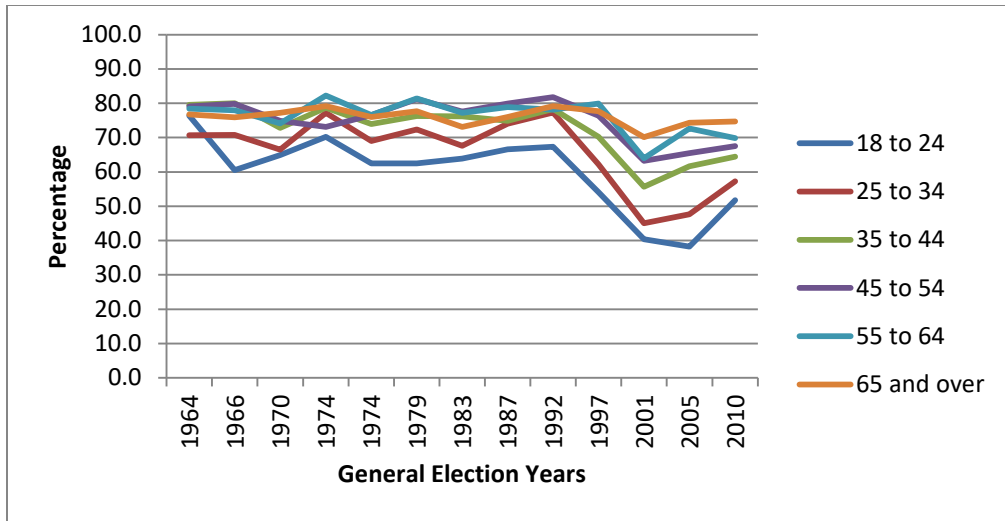


Figure 2.7: Participation in election by age groups. Source: TEC (2010)

2.5.7 Heterogeneity, Social Capital and Crime

There has been interest in recent times as to the effects of diversity in the creation of social capital by both academics and policy makers. Researchers have expressed different views about the impact of heterogeneity on social capital. For example, Birani and Lehmann (2013) emphasised that ethnicity is important in binding individuals together known as *ethnic social capital*, especially with regards to employment, guidance and safety. Identity plays an important role in the likelihood of people connecting and forming social relationships (Gilchrist and Kyprianou, 2011) and the integration of migrants into local neighbourhoods (Kindler *et al.*, 2014). However, Goulbourne and Solomos (2003) argued that understanding the relationship between ethnicity and social capital is complex especially in the UK context. On the other hand, Vermeulen *et al.* (2012) have argued that negative effects of diversity would probably be stronger on social networks of a heterogeneous (bridging) nature rather than a homogeneous (bonding) one. This is because the bridging heterogeneous ties are more vulnerable to decreasing levels of trust.

On the other hand, Gilchrist and Kyprianou (2011) argued that ethnic heterogeneity has less bearing on how people interact (social capital) virtually through the internet as facilitated by development in technologies such as social media. In this research, in order to address the controversy associated with social capital and diversity, we propose that social capital can be enhanced through social networks.

Ethnicity has a notorious relationship with crime, principally acting through socio-economic exclusion and disadvantage, but also acting through biases in reporting, the justice system and policing, the latter including complex relationships between race and prejudice, but most clearly expressed in the findings of the Stephen Lawrence Inquiry in the UK (Macpherson, 1999). Ethnicity might be important especially when considering victimisation and hate crime (Shepherd, 2006). In terms of victimisation, in the UK, statistics have indicated that minority groups tend to have higher risks of crime victimisation than Whites. While the White population has 6.8% of crime victimisation risk, minority groups have 11.1% (ONS, 2013).

Previous studies report that increases in proportion of immigrants positively correlates with the incidence of property crime. For example, Bianchi *et al.* (2012) explained that an increase of 1% in the immigrant population is associated with 0.1% increase in the total number of criminal offences. Previous studies in the UK have yet to empirically establish the link between increases in the number of immigrants in an area with incidence of property crime specifically. However, empirical evidence exists in the US on links between the number of immigrants and the occurrence of other types of crime such as motor vehicle theft and robbery (Bholowalia and Kumar, 2014). The divergent opinions expressed by different studies necessitate the need for re-examination of the relationship between social capital and diversity, especially as it affects crime.

2.5.8 Education, Social Capital and Crime

Education is one of the most important predictors of social engagement. The importance of education in social capital has been stressed by a number of scholars and policy makers. For example, while policy makers emphasise the importance of social capital as a medium that may help to reduce social exclusion (Catts and Ozga, 2005), Helliwell and Putnam (1999) argued that education can influence the behaviour of individuals and those around them to participate politically and socially. Additionally, Imandoust (2011) emphasised that the availability of more educational institutions can potentially boost the quality and quantity of social capital in a country. Therefore, people who are educated are more likely to socially connect than uneducated people.

In terms of crime, Moretti (2005) argued that educated people are less expected to commit crimes even though it is difficult to quantify. The theory of human capital suggests that skills and

qualifications determine wages, and the wider the distribution of qualifications, the wider the distribution of wages (Green *et al.*, 2006). Reynolds *et al.* (2001) stressed that the propensity of individuals to commit a crime is associated with their level of educational attainment. Furthermore, Stephen *et al.* (2010) also emphasised that increasing school leaving age can have significant effects on property crime by reducing the proportion of uneducated people in the society. Brilli and Tonello (2014) maintain that compulsory policy on education can reduce the rates of juvenile crimes especially in areas characterised by higher levels of social capital, than in areas with lower levels of social capital.

Therefore, the effect of education on crime has received the attention of researchers in the quantitative literature, arguing that less educated people are more likely to engage in criminal activities as compared to educated persons (Machin *et al.*, 2011). However, previous studies (e.g. Moretti, 2005; Brilli and Tonello, 2014) have used data outside the UK. To empirically validate the effects of education on social capital as well as crime, UK based data are required. Therefore this research employs UK data on educational attainment to explore the effects of education on social capital as well as crime.

2.5.9 Length of Residence, Social Capital and Crime

Residential stability in a neighbourhood is an important factor for the generation of social capital and place-based attachment. It is therefore expected that duration of residence is one determinant factor in this direction (Hooghe and De Vroome, 2016). Studies have demonstrated that the creation of social capital is associated with length of residence in an area. For example, Yamamura (2011) argued that personal interactions are built over time and tend to be more solid when people reside in a particular neighbourhood and are influenced by length of residence and home ownership. Similarly, Keene *et al.* (2013) also stressed that it takes time to create supportive social ties; therefore length of neighbourhood residency may be an important determinant of social integration.

Additionally, Oh (2003) emphasised that length of residence has a positive effect on friendships, social cohesion and trust. Therefore, people who reside in a particular area for a long time are more likely to have close friends and tend to interact more with people (Turcotte, 2015). On the other hand, Nisic and Petermann (2013) argued that length of residence only affects social capital in the early stage before it begins to consolidate after that phase and then tends to grow at

a slower rate thereafter. In contrast, residential instability in a neighbourhood is associated with weak social ties and low probability that residents connect with one another (Sampson *et al.*, 1997).

Furthermore, residents who have lived in an area for a long time have more opportunity for frequent contact and interactions with neighbours which helps in reducing local problems (such as crime) (Oh, 2003). Crime is also more likely to occur in transient neighbourhoods. For example, in the UK, the tendency to commit crime is related to length of residence; in other words, crime reduces as length of residence increases (Bell and Machin, 2011). Specifically, research has established the relationship between residential instability and violent crime (e.g. Boggess and Hipp, 2010). However, the relationship between residential instability and burglary is complex. Wisniewski *et al.* (2013) maintain that length of residence, age or gender has no effect on crime reporting.

Although a wide range of studies have attempted to explain the relationship between length of residence and social capital, only a few attempts have been made to relate length of residence to neighbourhood social problems such as crime. This research will explore the relationships between length of residence and crime.

2.5.10 Deprivation, Social Capital and Crime

Deprivation is considered to be a multi-dimensional problem embedded in range of domains such as income, health, education, housing, employment, disability and crime (DCLG, 2015). Furthermore, deprivation increases distance for social inclusion among neighbourhood residents and also decreases interactions (cohesion) among community members (Morenoff *et al.*, 2001; Takagi and Kawachi, 2014). Social capital within a community might be affected by income inequality. Pickett (2013) emphasised that income inequality undermines social capital and retards growth in the community. Cooper and Innes (2009) argued that diversity and deprivation tend to compromise a community's level of cohesion, while equality of income reduces divisions in a community which enhances social capital. Generally speaking, more deprived communities may also be lower in social capital (Drukker *et al.*, 2003)

The deprivation hypothesis suggests that deprived communities tend to have more crimes than affluent communities (Sampson and Wooldredge, 1987; Malczewski and Poetz, 2005). Disparity

of wealth can influence social capital (cohesion) and crime in a community. The feeling of wealth disparity between wealthy and poor people might lead some poor people within the communities to commit crimes against both poor and rich (Fajnzlber *et al.*, 2002; Rufrancos *et al.*, 2013). Additionally, researchers have found support for the relationship between property crime and income inequality (Witt *et al.*, 1998; Kelly, 2000; Demombynes and Özler, 2005; Reilly and Witt, 2008). However, little consensus exists across disciplines for the explanation of the relationship between income inequality and other types of crime (Rufrancos *et al.*, 2013).

2.6 Collective Efficacy and Crime

Collective efficacy is a neighbourhood concept referring to networks of informal social control necessary for establishing value systems reflective of prevailing social norms (Swatt *et al.*, 2013). Collective efficacy includes trust among residents and their willingness to intervene to achieve social control (Sampson, 2012; Swatt *et al.*, 2013). Similarly, Sampson *et al.* (1997) described collective efficacy as the ability of members of the community to work together to reduce delinquent behaviour of individual residents, as well as their desire to achieve safety in neighbourhoods (Bandura, 1997).

Furthermore, it has been stressed that collective efficacy of the residents has significant impact in mediating levels of crime in neighbourhoods (Browning, 2002; Gerell, 2014). Therefore the nature of interactions and trust among members of the community with a view to achieving common goals and values can be associated with their levels of crime (Swatt *et al.*, 2013). Residents of neighbourhoods where collective efficacy exists tend to watch out for each other, for example, monitoring of children's activities and youths in order to provide a sense of safety in the neighbourhood (Swatt *et al.*, 2013). However, it was argued that collective efficacy of a neighbourhood may develop through the capacity of residents to promote social ties and trust among themselves (Sampson *et al.*, 1997).

2.6.1 Measurement of Collective Efficacy on Crime

A number of studies have explained the effect of collective efficacy on different community processes, including neighbourhood crime rates (Sampson *et al.*, 1997; Browning *et al.*, 2004; Bruinsma *et al.*, 2013; Armstrong *et al.*, 2015; Yuan and McNeeley, 2017b). However, measuring informal social control is difficult, especially, because it is only displayed in specific

situations relating to violation of social order (Hipp and Wo, 2015) such as, youths hanging around on a street corner (Hipp, 2016). Researchers have used different statistical methods such as multilevel item response theory (IRT) to measure collective efficacy (Swatt *et al.*, 2013). IRT provides a technique for modelling multiple test scores and is widely used in social sciences (Loken and Rulison, 2010). Multilevel IRT was proposed by Kamata (1998) and is useful for data collected in hierarchical settings. Multilevel IRT modelling is advantageous in that more accurate results can be obtained from the relationship between predictor variables and parameters that affect it (Kamata, 1998).

Rukus and Warner (2013) have employed binomial regression to evaluate the impact of collective efficacy in property and violent crime in Florida using a nationwide survey. They found out that the collective efficacy of neighbourhoods has significantly reduced violent crimes. Browning (2002) applied a hierarchical linear model to explore the effect of collective efficacy on homicide in Chicago using census data, homicide data, and health and social life survey data. He found out that collective efficacy of the neighbourhoods regulates homicide rates.

On the other hand, Sutherland *et al.* (2013) have also used multilevel regression to examine the impact of collective efficacy on disadvantage and crime in London using a police public attitude survey and face-to-face interviews across 60,000 individuals. Their findings show a negative relationship between police-recorded violent crime and collective efficacy, but they found no statistical significant relationship between collective efficacy, neighbourhood structural characteristics and violence.

Recently, Hipp (2016) has used combined household data from three counties: North Carolina, Vance and Person, collected at three one-year time points, to explore the effect of collective efficacy on disorder in households. He used factor analysis and hierarchical linear modelling. He found that neighbourhood collective efficacy on perceived crime in a particular year has a negative effect on informal social control in the following year by increasing crime. He also found that efficacy against disorder can change over time. However, his study could not provide an explanation of possible reasons behind such occurrence. Furthermore, Gerell and Kronkvist (2016) attempted to measure collective efficacy using a community survey of residents in 2012 for the city of Malmo in Sweden (N=4,051), Census data and police reported data of violent crimes. They found a negative relationship between collective efficacy and violent crime, but

association between neighbourhood structural characteristics such as ethnic heterogeneity, disadvantage, residential instability and violent crime.

2.7 Weaknesses of Previous Studies

Limited empirical research on community cohesion and its impact on crime exists (Goodchild, 2013). Table 2.2 presents the gaps established from the review of previous studies.

Table 2.2: Limitations of previous studies

Limitation	Reference
Lack of consensus on universal definition of community cohesion and hence on the use of appropriate predictors of measurement	Pauwels and Hardyns (2009)
Lack of adequate sources of data to quantify social cohesion especially for small areas	Parker <i>et al.</i> (2007); Goodchild (2013)
Most of the research conducted is America based and application to UK situation has not been fully explored	Phillips and Lee (2011)
Previous researchers largely emphasise the effects of residential mobility on community, influence of subculture (diversity) on social cohesion and by extension on crime has largely been ignored.	Porter (1996); (Barnes, 2013); Openshaw (1991); Sampson and Groves (1989); Sheppard (1995); Schuurman (2000); Goss (1995); Curry (1997)
Over reliance on small sample data from traditional surveys	Bowers <i>et al.</i> (2004); Traunmueller <i>et al.</i> (2014)
No previous studies employ social media data to study community cohesion and crime	

2.8 Theoretical Framework

Knowledge of appropriate crime theories is vital in crime analysis (Eck *et al.*, 2005). Many theories have emerged over the years and they continue to be explored individually and in combination as researchers seek to find the best solutions to reduce the types and levels of crime in a society. Therefore, most crime prevention strategies are based on implicit knowledge of one or more theoretical understandings of crime. In this research, the criminological theories used as a framework include social disorganisation and routine activity. These theories are discussed in more detail in the next section.

2.8.1 Social Disorganisation Theory

Social disorganisation theory emerged from the 'Chicago School' in the 1920s. However the very first ecological study was believed to have emanated from the Cartographic School of Guerry (1833) in France and Quetelet (1835) in Belgium. Guerry's work was fascinating because he used statistical tables and figures to analyse his data on crime and suicide in different areas of France. The results of his study revealed that crimes and suicide are characterised by different variations across space, but remained stable when broken down into age, sex, seasonal and regional categories (Friendly, 2007). The last century however, witnessed new theoretical innovations, with advent the of Shaw and McKay's (1942) theory of social disorganisation (Chainey and Ratcliffe, 2013). Shaw and McKay (1942) applied Moriarty and Williams (1996) Concentric Zone Model to divide the city of Chicago into different zones (neighbourhoods) based on crime activities. They were able to map incidence of juvenile crime events in Chicago city based on the neighbourhood characteristics. The results of their study indicated that crimes (delinquent) were not evenly distributed in the city, and that socio-economic, residential instability and heterogeneity factors associated with different areas of the city tend to weaken neighbourhoods ability to institute social control, allowing delinquency to flourish (Shaw and McKay, 1942). Attempting to explain the influence of the variables used, Shaw and McKay (1942) suggested that the economic status of areas has an indirect influence on delinquency; while affluent areas provide an atmosphere for social control, deprived areas offer a conducive environment for delinquent behaviours of the diversity of the residents. They also argued that areas with large proportions of transient population tend to have higher rates of delinquency.

Furthermore, Shaw and McKay (1969) stressed that delinquency is linked to neighbourhood characteristics and absence of social control within the community resulting to social disorganisation. The work of Shaw and McKay continued to be relevant and forms the basis for criminological studies for many decades (Brantingham and Brantingham, 1981; Chainey and Ratcliffe, 2005). However, there have been criticisms of the work of Shaw and McKay, important among which is the application of their ideas especially to the UK context, especially when the ecological fallacy attributing characteristics to individuals based upon the characteristics of areas in which they reside (Shepherd, 2006).

Contemporary studies applying the social disorganisation theory of Shaw and McKay argued that economic disadvantage, ethnic differences and residential movements reduce informal social control and increases the probability of crime (Tewksbury and Mustaine, 2006; Sampson and Groves, 1989; Bursik Jr and Grasmick, 1993; Warner and Rountree, 1997; Markowitz *et al.*, 2001; Kubrin and Weitzer, 2003). Therefore, the underlying role of community ties (cohesion) and their subsequent effects on neighbourhood crime remain under tested (Warner and Rountree, 1997).

Researchers including Martin (2002) and Paulsen and Robinson (2004) also stressed that the spatial distribution of poverty, unemployment, ethnic diversity and neighbourhood social organisation are the major foci of social disorganisation theory. The effect of social disorganisation according to Paulsen and Robinson (2004) is the reduction of cohesion within the community which in turn increases crime rates and anti-social behaviour in those communities. However, more recently, studies have suggested a more complex relationship between poverty, residential stability, informal social control and crime rates and have advocated for measurement of social capital and collective efficacy in neighbourhood as indicators of informal social control (Martin, 2002).

While social disorganisation theory provides the basis upon which ecological factors can be related to crime (Sun *et al.*, 2004), ethnic diversity, family disharmony and economic deprivation tend to create disorganisation in neighbourhoods and increases in crime rates (Sampson and Groves, 1989). Therefore, Bursik Jr and Grasmick (1993) have argued that neighbourhoods are expected to share norms for collective efficacy in line with their neighbourhood expectations. However, the availability of data on such important variables of an organised society, which

include collective efficacy and cohesion, can be difficult to acquire (Paulsen and Robinson, 2004).

The focus of contemporary studies employing social disorganisation theory was multidimensional, examining the detailed inter-relationships between community cohesion, informal social control and collective efficacy, coupled with population characteristics linking the effects of these variables on overall crimes in a given area (Bellair, 1997; Rountree and Warner, 1999; Sampson and Groves, 1989; Markowitz *et al.*, 2001). However, the systemic model of community structures is very complex since the possibility of achieving formal and informal social interactions relates to the level of socialisation and family background (Bursik, 1986). The systemic model views the community system of friendship and kinship networks and informal interactions which are family based in the continued socialization processes. Modern social disorganisation theorists have attempted to explain the effect of community structural variables on crime through examining the community level of cohesion (Bursik Jr and Grasmick, 1993). Therefore, linking a community's ecological characteristics such as socioeconomic to crime has become a common design among researchers (Bursik, 1986; Sampson and Groves, 1989; Bellair, 1997).

Despite growing interest in criminology research and theoretical discussions related to the community approach to crime prevention, limited empirical tests exist to adequately explain the relationship between community cohesion and crime prevention especially in the UK context. The present study seeks to contribute to advance research in this area by using new social media data, especially Facebook and Twitter, to explore relationships between social cohesion and crime.

2.8.2 Routine Activity Theory

Routine Activity Theory (RAT) emanates from the work of Cohen and Felson (1979). The theory seeks to explain the context of crime in a broader perspective not just the behaviour of the offender. The theory examined criminal opportunity as a factor with three components: the offender himself, the target or prospective victim of the crime as well as the guardian who may watch or witness the crime. These three components must be there for any crime opportunity to take place, although convergence of these factors depends on the daily lifestyle (Felson, 1995; Chainey and Ratcliffe, 2005).

The theory further stresses that not all offenders have the technical experience to commit a crime and those targets are not always suitable as they may be well guarded in terms of their design. However, some guardians may not be capable (Short and Ditton, 1998). For example where CCTV cameras are installed, the offender may tend to avoid them unless under the influence of an intoxicant. The focus of researchers on crime prevention has shifted with the target and guardian being the main focus as the offender is seen as a constant or rather a dependable variable (Chainey and Ratcliffe, 2005). The target is the place or neighbourhood where the crime is committed and guardianship refers to informal social control.

Recent approaches to routine activity however, consider that informal social control of the offenders could be linked to the control theory (Hirschi and Stark, 1969). In this context, the social and moral upbringing received by children from their parents serves as an instrument of social control with parents becoming effective handlers for reshaping the perception of their children (Felson, 1995). Furthermore, Felson (1995) stressed that where there is an effective supervision by the guardian against suitable targets, then an offender may be discouraged to commit crime. Eck (1995) also emphasised that unlike the handler, a guardian can be a law enforcement officer such as a police man, a landlord or even a passer-by. Figure 2.8 shows the routine activity theory triangle known as Eck crime triangle. Similar to social disorganisation, in the routine activity perspective, place is the most important factor which determine a decrease or increase in the potential occurrence of crime. Therefore, these theories tend to overlap and complement one another in their attempt to provide an explanation of the occurrence of crime (Moriarty and Williams, 1996; Tewksbury and Mustaine, 2006; Gorman *et al.*, 2013).

The present research will employ social disorganisation theory and routine activity theory as underpinning theories to contextualise relationships between community cohesion and crime.



Figure 2.8: Eck crime triangle. Source: Eck and Weisburd (1995)

2.9 Concluding Remarks

The preceding sections have discussed the relationships between community cohesion and crime. The objective is to highlight the place of cohesion on crime and to set the context on which the thesis will be based. Section 2.1 has provided a brief introduction to the chapter. Section 2.2 followed with a discussion on different definitions of the concept of community cohesion, beginning with its historical evolution in the UK (Cantle, 2001). Section 2.3 explored the relationship between community cohesion and crime, referring to data and methods employed in previous studies to model relationships between community structures and crime. The major limitation of previous studies is that by using traditional data, cohesion is poorly captured by traditional regression models of crime, largely because cohesion is ill represented by standard socio-demographic variables. The importance of social capital is demonstrated in Section 2.4. Crime rates are closely linked to levels of social capital, though it is difficult to quantify (Harper, 2001) because of the paucity of data to measure it properly (Kanazawa and Savage, 2009; Appel *et al.*, 2014). Section 2.5 explores in detail the determinants of social capital in relation to crime. To set the ground for the research, Section 2.6 provides discussion on two underpinning theories: social disorganisations (Shaw and McKay, 1942) and routine activity (Cohen and Felson, 1979) which are used to provide context in building a model of community variables and crime (especially burglary). Chapter 3 follows with a discussion of traditional data sources and new data sources as well as approaches for modelling the urban social system.

Chapter 3

Traditional Methods of Modelling Urban Social Systems and Crime

3.1 Introduction

The concept of community cohesion in relation to crime and factors determining cohesion in a community as well as underpinning theories of crime were reviewed in chapter 2. This forms the basis for understanding the relationships between social cohesion and crime. This chapter proceeds to extend our understanding of these relationships through examining “traditional” methods used for modelling urban social systems and the data sources available. The chapter begins by describing the research methodology in Section 3.2, and then the geography of the study area is presented in Section 3.3. Section 3.4 examines the data sources used whilst Section 3.5 provides a description of Leeds community areas. Section 3.6 explores statistical methods of modelling the relationships between urban community structures and crime. Section 3.7 then describes the spatial data analysis methods while Section 3.8 provides a discussion on spatial statistical methods and Section 3.9 contains a discussion about diversity indices. Finally, Section 3.10 examines the computational model and Section 3.11 presents concluding remarks for the chapter.

3.2 Research Methodology

Research methods in social science can be divided broadly into qualitative and quantitative methods (Alasuutari *et al.*, 2008). Quantitative research is typically associated with positivist and objectivist principles, while qualitative research, on the other hand, is associated with interpretivist and constructionist principles (Bryman, 2008, p.13). While objectivism holds that research is about discovering objective truth, independent of the researchers own feelings, positivism view is also closely related to objectivism; it argues that social reality exists independent of the researcher and must be investigated through the scientific processes of inquiry (Gray, 2013). In contrast, constructivism views that truth exists through subjects (people’s) interactions with the world. In such a process, subjects construct their own meaning in

different ways even in relation to the same phenomena; constructivism is closely linked to interpretivist which asserts that social reality is different and therefore requires different kinds of method (Gray, 2013).

In social science, researchers often use quantitative methods to measure concepts such as crime and demographics (Neuman and Robson, 2004). Furthermore, Todorova (2013) stressed that the choice of the research method depends on the objectives, research questions and author's best judgement. This research will therefore employ a quantitative approach in line with the outlined objectives.

3.3 Study Area

3.3.1 Geography

The city of Leeds in West Yorkshire England is located at latitude 53° 57' 59"N and longitude 1° 33' 57"W about 190 miles (310 km) north of London. The city contains 33 *wards*, which are divided into 482 *lower super output areas* (LSOAs). The LSOAs have a minimum population of 1,000 (an average of 1,500) and a minimum resident household number of 400 (an average of 630) (ONS, 2011). According to the 2011 Census, Leeds' population was estimated at 751,485 (ONS, 2011), the third largest local authority in the UK after Greater London and Birmingham. Additionally, it contains some of the poorest wards in England (DCLG, 2015), but equally has wards containing some of the most affluent individuals in the country (BBC, 2003). The city also has relatively large number of crimes and characteristically different types of neighbourhood which makes it suitable for examining relationships between socio-economic and demographic diversity and burglary (Hirschfield *et al.*, 2013). Figure 3.1 shows the location map of Leeds.

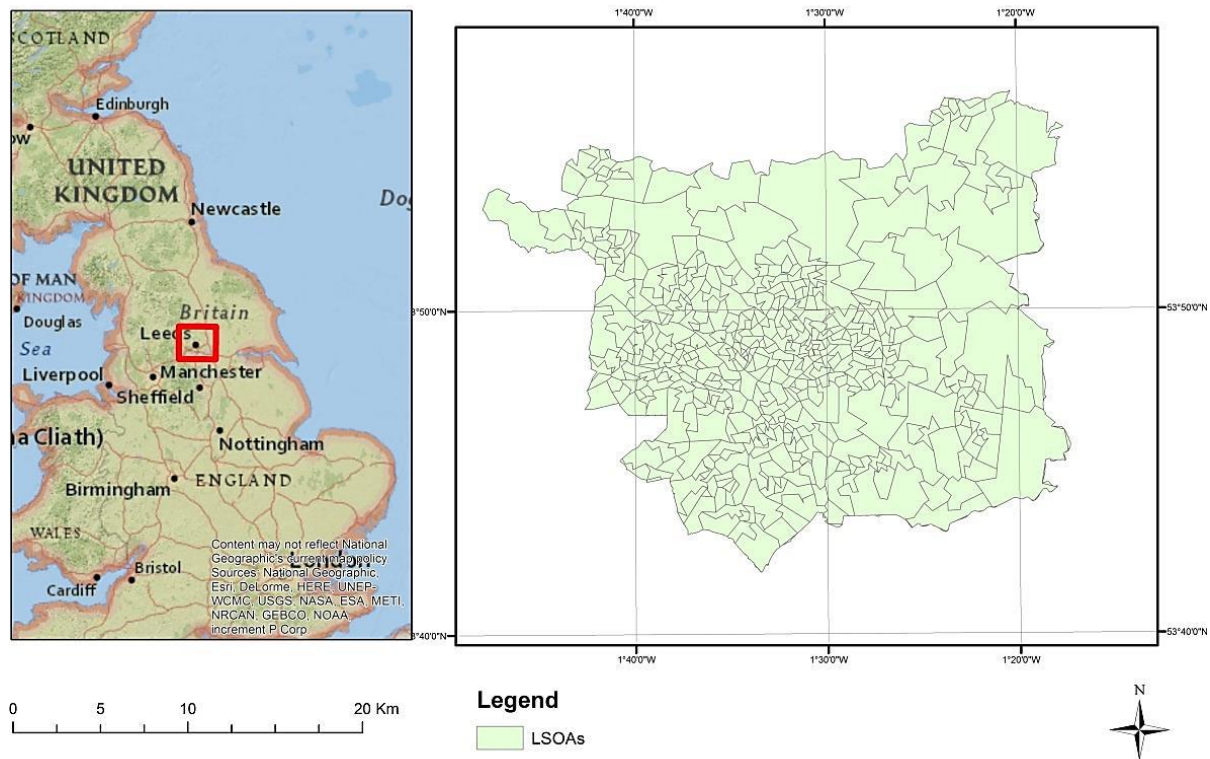


Figure 3.1: Location map of Leeds LSOA boundaries. Source: Edina (2011)

3.3.2 People

Leeds is a diverse city with over 75 ethnic groups with the minority ethnic population making about 11.6% of the total population according to the 2011 Census. About 85.1% of the population is comprised of Whites and the majority (81.1%) are of British origin. There are also 141,771 people that comprise the Black and Minority Ethnic (BME) communities (ONS, 2011). The ethnic diversity of people in Leeds is increasing, in the same manner as other European countries (Putnam, 2007). In 2001, the population of BME groups was 77,530 (about 10.8% of the resident population) and increased to 141,771 (representing 18.9% of the resident population) by 2011 (ONS, 2011). The majority of those who identified themselves as BME belong to the Pakistani, Indian, Chinese, Bangladeshi and African population. Figure 3.2 shows the ethnic diversity of Leeds. Areas including the City Centre, Hyde Park, Harehills, Chapeltown and Beeston are the most ethnically diverse places in Leeds.

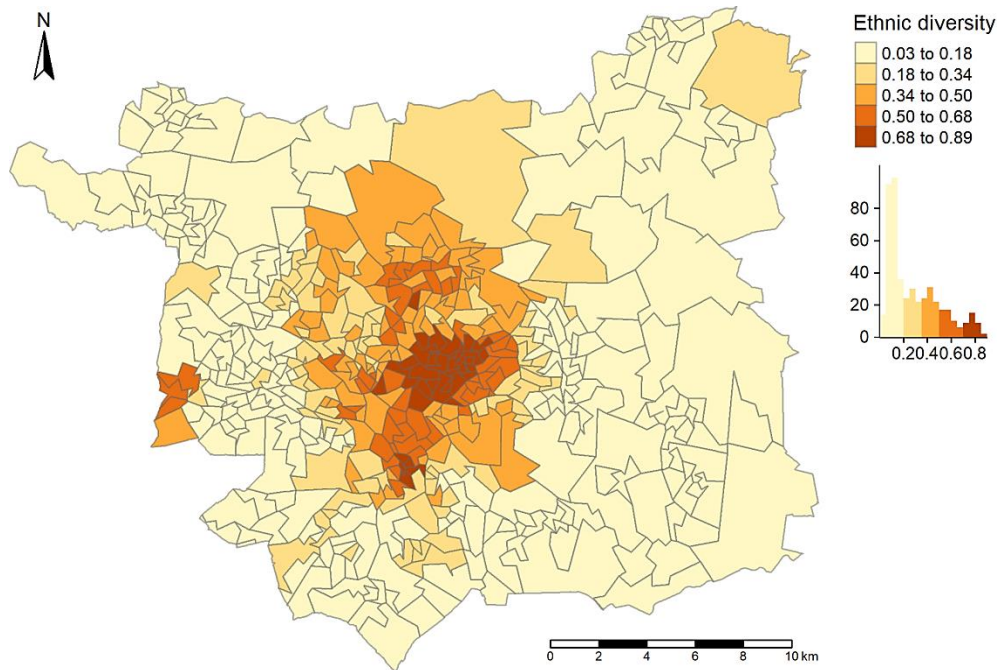


Figure 3.2: Ethnic diversity of Leeds. Source: ONS (2011)

Diversity has a complex relationship with social cohesion, and social cohesion has a complex relationship with crime. High diversity in the communities is acting to reduce social cohesion consequently increasing neighbourhood crime. There is certainly some evidence that diversity reduces social cohesion (Meer and Tolsma, 2014) and this seems especially true where diversity is found in conjunction with deprivation (Cooper and Innes, 2009). However, it is nevertheless important to note that diversity and cohesion levels are not always related in a simple manner and that cohesion has the opportunity to be affected both positively and negatively, and by more than just ethnic or economic diversity. It is also true that, overall, potential high diversity may have net positive impacts, in terms of multiculturalism and the disruption of embedded cultural processes, despite negative impacts in other areas. Data on standard Census variables are required to calculate diversity.

3.4 Data Sources

There are very few data sources available and used for community cohesion and crime research, most of which are national surveys carried out by the Home Office of the British Government. Researchers have used independent surveys and ethnography (Sampson and Groves, 1989;

Sampson and Raudenbush, 1999) but such instances are very few and constrained by resources. The following are datasets used most extensively by researchers.

3.4.1 Census Data

The Census of 2011 marks 210 years since the census first began in the UK. A census is a count of people and households and is used to set policies and estimate resources required to provide services such as education, housing, healthcare and transports (ONS, 2011). Census data are the most comprehensive and dependable socio-economic and demographic data available in the UK (Vickers and Rees, 2007). Typically, the Census involves counting people and their characteristics (Rees *et al.*, 2002). Every 10 years since 1801, a day is set aside to conduct a census and every effort is made to ensure that everyone is included. Censuses are the only surveys that provide a detailed picture of the entire population and are unique as everyone is included and the same questions asked across the entire country (ONS, 2011).

Between 1971 and 1991 digital Census data have been made available for research at different geographical scales (such as enumeration district). Output areas (OAs) began from 2001, making comparisons easier between different areas (Shepherd, 2006). In this research, aggregated data used were derived from UK 2011 Census, supplied by the UK Data Service¹ (Infuse) at lower super output area (LSOA) level.

3.4.1.1 Census Variables

As highlighted in Section 3.4.1, all variables used were derived from UK Census 2011. Although the relationship between community composition and crime is complex and multi-faceted, there are some core factors that regularly emerge as important determinants of crime rates. This section outlines the variables and justification for their inclusion in the later modelling work. Table 3.1 provides details of the Census variables used.

Age Distribution

Offenders are commonly drawn from younger age groups than the elderly people (Kongmuang, 2006). The age-crime curve tends to increase from the adolescent years reaching a maximum at adulthood and then sharply declining (Farrington, 1986; Gottfredson and Hirschi, 1990;

¹ infusecp.mimas.ac.uk

Sampson and Laub, 2003; McVie, 2005; Blonigen, 2010; McCall *et al.*, 2013; Sweeten *et al.*, 2013), although this varies by the type of crime (Tittle *et al.*, 2003). Crime, especially burglary is therefore likely to be affected by absolute numbers of young (aged 16-24) people. For example, according to Fagan *et al.* (2014), the incidence of crimes related to vehicles and drugs tends to be higher in early adulthood than in adolescence. While homicides tend to be committed by adults, theft-related offences, including burglaries, are more prevalent in the younger age groups than the elderly (Loeber *et al.*, 2012). Variables used in this category are from QS103EW 17-26 (Table 3.1).

However, crime may also be affected by age distributions. Younger people are less likely to build social cohesion (especially face-to-face) than older people (Johnston and Matthews, 2004; Takagi and Kawachi, 2014), but equally, different age distributions may have very different strengths of social control.

Table 3.1: Census variables and codes

Datasets	Variables description	Variable code
Age structure (QS103EW)	10-14	QS103EW0012-16
	15-19	QS103EW0017-21
	20-24	QS103EW0022-26
	25-29	QS103EW0027-31
	30-44	QS103EW0032-46
	45-59	QS103EW0047-61
	60-64	QS103EW0062-66
Household composition (QS112EW)	65-74	QS103EW0067-76
	Lone parent no dependent	QS112EW0018
	Lone parent 1 dependent child	QS112EW0023
	Lone parent 2 or more dependents	QS112EW0024
	Married couple no children	QS112EW0008
	Married couple 1 dependent child	QS112EW0009
Married couple 2 or more children	QS112EW0010	

Datasets	Variables description	Variable code
Ethnic groups (QS201EW)	White	QS201EW0002-0005
	Mixed	QS201EW0005-0009
	Asian	QS201EW0010-0014
	Black	QS201EW0015-0017
	Other	QS201EW0018-0019
Occupation (QS606EW001)	16-64 Managers/Directors	QS606EW0002
	16-64 Professionals	QS606EW0017
	16-64 Associate Professionals	QS606EW0038
	16-64 Administration and Secretariat	QS606EW0059
	16-64 Skilled Trades	QS606EW0068
	16-64 Caring Leisure and Services	QS606EW0086
	16-64 Customer Services	QS606EW0096
	16-64 Process Plants and Machines	QS606EW0104
	16-64 Elementary Occupation	QS606EW0114
Economic activity (QS602EW0011)	16-64 Economically inactive retired	QS602EW0012
	16-64 Economically inactive students	QS602EW0013
	16-64 Economically inactive looking after family	QS602EW0014
	16-64 Economically inactive long term sick or disabled	QS602EW0015
Qualification (QS501EW0001)	16-over no qualification	QS501EW0002
	16-over qualification level 1	QS501EW0003
	16-over qualification level 2	QS501EW0004
	16-over qualification level 3	QS501EW0005
	16-over qualification level 4	QS501EW0006
Length of residence (QS803EW)	Less than 2 years	QS803EW0003
	Less than 5 years	QS803EW0004
	More than 5 years	QS803EW0005
	10 years above	QS803EW0006
	Born in the UK	QS803EW0002

Source: ONS (2011)

Family Structure

Maginnis (1997) has argued that the children of some single parent families are more likely to have behavioural problems because they tend to lack economic resources and have lower parental input. In the UK, lone parents continue to suffer from inequalities of employment and housing, creating a gap between other household structure and lone parents (Berrington, 2014). Additionally, single parents are also most likely to be victims of crime due to social marginalization in terms of living conditions (Wikström and Wikström, 2001). Variables used in this category are from QS112EW23-24 (Table 3.1).

Identity

Identity (ethnicity) plays an important role in the likelihood that people will connect and form social relationships (Gilchrist and Kyprianou, 2011) and plays an important part in the integration of migrants into local neighbourhoods (Kindler *et al.*, 2014). Migrants especially from the black and minority ethnic populations (BME) often lack the wealth, social integration, or formal crime prevention connections to protect themselves (Sharp and Atherton, 2007). Because of these factors, the size of an immigrant population in an area positively correlates with the incidence of property crime (Bell and Machin, 2011). The ethnic minority variable is constructed from QS201EW0014-19 (see Table 3.1).

Affluence and Wealth

While there is a wide range of criminality across the socio-economic spectrum, for burglary specifically, the offenders in the vast majority are drawn from the poor and unemployed (Bursik Jr and Grasmick, 1993; Sariaslan *et al.*, 2013). Disparities of wealth within short distances encourage burglary and, in addition, disparity of wealth within a community can influence crime by weakening social cohesion (Fajnzlber *et al.*, 2002; Rufrancos *et al.*, 2013). Variables included in this category are from QS602EW0012-15 (Table 3.1).

Educational Attainment

Educational attainment has a great influence on individuals' social behaviour as well as on participation in community activities. The theory of human capital suggests that skills and qualifications determine wages, and the wider the distribution of qualifications the wider the distribution of wages (Green *et al.*, 2006). Reynolds *et al.* (2001) stressed that the propensity of

individuals to commit a crime is associated with their level of educational attainment and so we include lack of qualifications (QS501EW0002) as a standard variable (see Table 3.1).

Residential Instability

Residential stability in a neighbourhood is an important factor for generation of social capital and place-based attachment. It is therefore expected that duration of residence is one determinant factor in this direction (Thomas *et al.*, 2016). Studies have demonstrated that the creation of social ties is associated with the length of residence in an area. For example, Yamamura (2011) argued that personal interactions are built over time and tend to be more solid when people reside in a particular neighbourhood and are influenced by the length of residence and home ownership. The tendency to commit a crime is related to the length of residence. The variable included in this category is the proportion of those who reside in an area for less than two years (QS803EW0003, Table 3.1).

Housing Tenure

The housing tenure and characteristics of place are related to social connectedness. Tenure diversification can improve social interactions between renters and owner-occupiers (Wood, 2003; Caesar and Kopsch, 2018) generation of social capital by stimulating neighbourhood involvement, mutual trust and reciprocity between residents (Leviton-Reid and Matthew, 2018). Although policies on housing tenure are important for reducing social exclusions, in the UK, there has been little support of the policy for improving the social wellbeing (Graham *et al.*, 2009). Wood (2003) argued that in the UK context, housing diversification has merely segmented neighbourhoods, rather than tenure integration. Additionally, the residents of social housing tenure are isolated from the community, hence fewer opportunities for education and employment (Wood, 2003). Furthermore, areas of social isolation are likely to be higher in the rented sector with lower engagement between residents compared to areas of home ownership (Forrest and Kearns, 2001; Tersteeg and Pinkster, 2016) which undermines sense of community cohesion (Cooper and Innes, 2009).

In terms of crime, Bottoms and Wiles (1986) stressed that housing tenure is also related to offending and victimisation. Areas especially of social housing tenure, are more likely to have behavioural problems or to suffer victimisation (Wood, 2003) especially property crime (Farrall *et al.*, 2015). In a recent study, Hegerty (2017) found that the share of renters in an area

correlates with increases in thefts related offences in low income areas, but not high income areas.

In this research, housing tenure variable is not included in the model because it correlates with the variables of ethnicity, residential instability and family structure as in previous studies (e.g Cooper and Innes, 2009; Tersteeg and Pinkster, 2016).

Index of Multiple Deprivation

The deprivation hypothesis suggests that deprived communities tend to have more crimes than affluent communities (Sampson and Wooldredge, 1987; Malczewski and Poetz, 2005). Furthermore, deprivation widens the gap between the rich and poor which can reduce social cohesion (Morenoff *et al.*, 2001; Takagi and Kawachi, 2014). The Index of Multiple Deprivation (IMD) is a multi-dimensional metric that is measured, in England, through a combination of seven distinct domains: income; employment; education; health; crime; barriers to housing & services; and living environment (DCLG, 2015). Although, deprivation is often seen as a key indicator of social cohesion, as well as propensity to commit a crime such as burglary, UK deprivation statistics include crime and therefore it is inappropriate to use them in this research.

3.4.2 Police Recorded Crime Data

The police records of crimes are statistics that cover selected offences which have been reported to and recorded by the police. These records originate from police territorial offices across England and Wales and are then passed to the Office for National Statistics through the Home Office (ONS, 2014a). The police records provide the primary source of sub-national crime statistics and some of which are not covered by the crime Survey. While the police recorded crime figures cover a wider population and a broader set of offences than the Crime Survey of England and Wales (CSEW), crimes that are not brought to the attention of the police are not included. Hence police statistics based on recorded crime data have not met the required standard for designation as national statistics (ONS, 2014a). It is important to note that police recorded crime is driven by the trends in reporting. For instance, burglary crimes are on the increase in recent statistics, due to changes in reporting incidences (ONS, 2017a). Police records of crime have been employed, for example, to investigate public disorder in urban areas of Chicago (Sampson and Raudenbush, 1999), to explore the effects of community cohesion on levels of

recorded crimes in Merseyside (Hirschfield and Bowers, 1997) and to examine property and violent crimes in England and Wales (Han *et al.*, 2013). A typical UK police data website is shown in Figure 3.3.

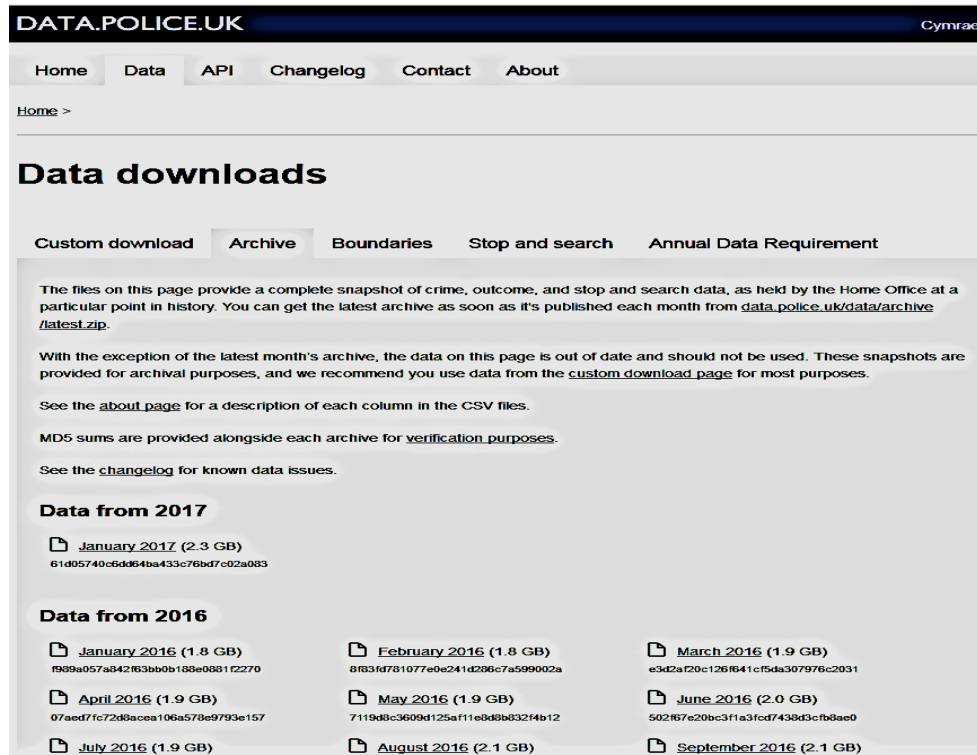


Figure 3.3: Typical UK Police data download website

Since 2011, data regarding individual police recorded crimes have been made public for research. However, to comply with UK Data Protection Act, 1998, which stipulates that victims of crime should be anonymised, geo-masking techniques have been introduced to reduce their spatial accuracy (Tompson *et al.*, 2015). Geo-masking is a technique used to provide privacy protection for individual address and information (Allshouse *et al.*, 2010). This should not be a problem as this research is concerned with the analysis of crime at neighbourhood levels. The crime data used in this research were obtained from the 'Police open public monthly data of reported crimes' provided by West Yorkshire Police at LSOA level.

3.5 Leeds Community Areas

Although, LSOA geography will be used for the initial traditional modelling, later in the rest of the thesis (chapters 6, 7 and 8), the LSOAs are aggregated into 106 Leeds community areas. Community areas are large enough and recognised by Leeds residents for meaningful social cohesion (Stillwell and Phillips, 2006) which is the focus of the present research. The exercise of neighbourhood mapping was initiated by Leeds City Council staff involving Leeds residents who helped to identify place-based communities. This exercise resulted in demarcating 100 community area boundaries, although this has been contested in some places (Shepherd, 2006). Nevertheless, the major problem with Leeds City Council boundaries is that they do not share a common boundary with small administrative boundaries such as Output Areas (OAs). To address this problem, new community boundaries were re-drawn by aggregating or splitting OA boundaries, so that OA is assigned to a community when 50% of its area fell within the community boundary. At the end of this exercise, an additional 6 community areas were added, bringing the number to 106 (Unsworth and Stillwell, 2004).

Ashby (2005) argued that communities tend to differ not only in socio-economic and demographic terms, but also in terms of their crime rates and victimisation. Since the importance of community cohesion is of increasing interest for neighbourhood safety, there is the need for detailed analysis of the socio-economic and demographic composition of different community areas of Leeds, with a view gaining insights into their perceived social cohesion especially relating to crime (Ashby, 2005). Additionally, Leeds' communities are changing demographically being the second most diverse city outside London, and crime, especially residential burglary, has been a particular problem, it is important to have a clear understanding of diversity of different areas (Leeds City Council, 2014). Figure 3.4 shows the 106 Leeds community area boundaries.

community name	community name
1 Adel	54 Hunslet / Stourton
2 Allerton Bywater & Gt and Lt Preston	55 Hyde Park
3 Ahwoodley & Wlton Moor	56 Klppax
4 Ardsley East / West (inc. Tingley)	57 Kirkstall
5 Armley	58 Woodhouse
6 Arthington & Pool	59 Lofthouse & Robin Hood
7 Bardsey & East Keswick	60 Manston
8 Barwick & Scholes	61 Meanwood
9 Beeston	62 Methley
10 Beeston Hill	63 Micklefield
11 Belle Isle	64 Middleton
12 Boston Spa	65 Moor Allerton
13 Bramham	66 Moortown
14 Bramhope	67 Morley North
15 Bramley	68 Morley South
16 Burley	69 New Farnley
17 Burley Lodge & Little Woodhouse	70 New Wortley
18 Burmantofts, Lincoln Green & Ebor Gardens	71 Oakwood
19 Calverley	72 Osmondthorpe
20 Chapel Allerton	73 Otley
21 Chapeltown	74 Dulton & Woodlesford
22 Churwell	75 Priesthorpe
23 City Centre	76 Pudsey
24 Collingham & Linton	77 Pudsey Lowtown
25 Colton	78 Rawdon
26 Cookridge	79 Richmond Hill
27 Cottingley	80 Rothwell
28 Crossgates	81 Roundhay
29 Drighlington	82 Sandfords, Ganners & Moor
30 East Bank	83 Thorner
31 Fairbank	84 Scott Hall & Miles Hill
32 Far Headingley	85 Seacroft North
33 Farnley	86 Seacroft South
34 Farsley & Rodley	87 Shadwell
35 Fearnville	88 Stanningley
36 Garforth East	89 Swardlffe
37 Garforth West	90 Swillington
38 Gildersome	91 Swinnow & Fairfields
39 Gipton North	92 Tinsill
40 Gipton South	93 Upper Armley
41 Gulseley	94 Upper Wortley
42 Halton / Whitkirk	95 West Park
43 Halton Moor	96 Wetherby
44 Harehills	97 Whinmoor
45 Harehills Triangle	98 Wortley
46 Harewood and District	99 Wythers
47 Hawksworth	100 Yeadon
48 Headingley	101 Ledston & Ledsham
49 Holbeck	102 Aberford
50 Holt Park	103 Scarcroft
51 Horsforth	104 Cross Green
52 Horsforth Newlaithes & Woodside	105 Little London
53 Horsforth West End	106 Ireland Wood

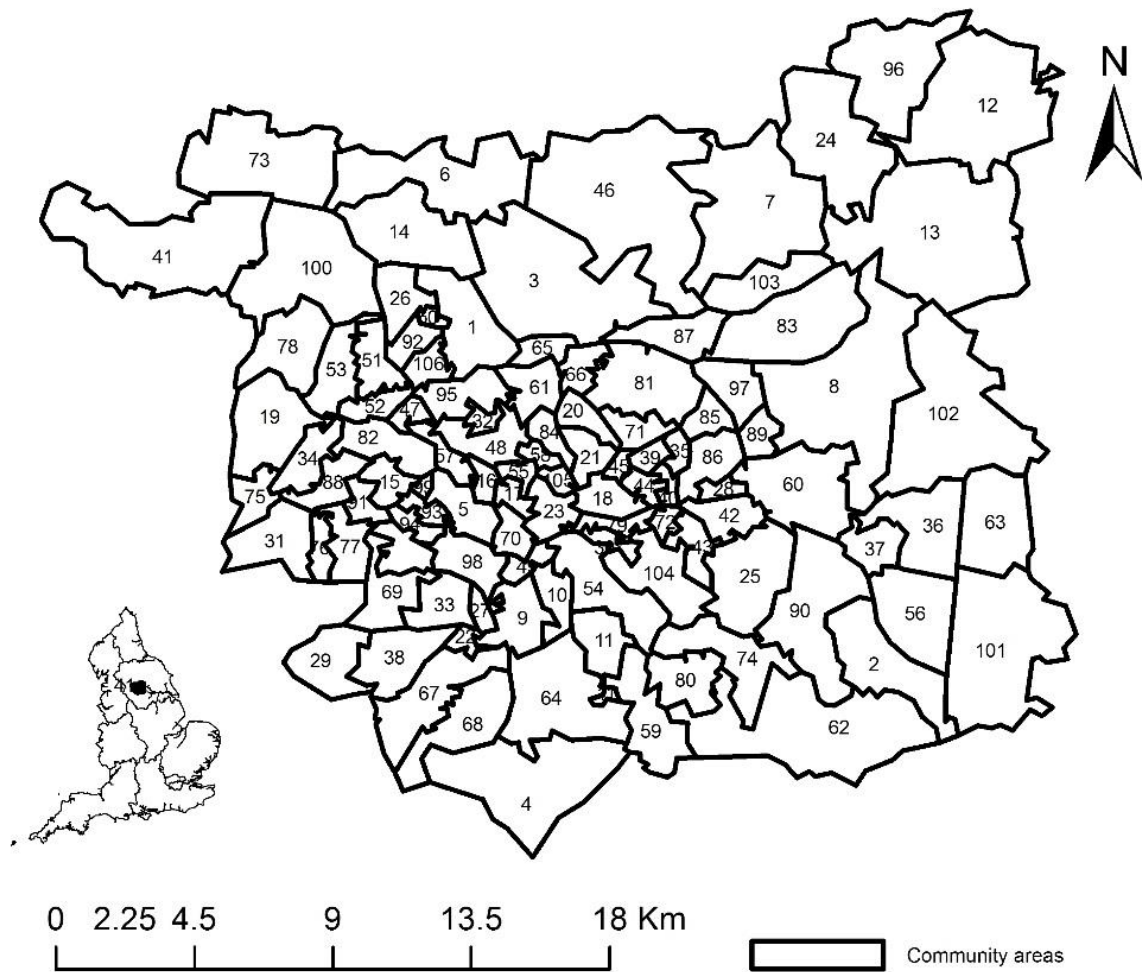


Figure 3.4: Leeds 106 community area 2001 boundaries. Source: Stillwell and Phillips (2006)

3.6 Statistical Methods of Modelling Crime

The analysis of data in order to model relationships between variables is practically impossible without analytical tools. Therefore, modelling crime and its associated determinants requires statistical techniques such as regression models in order to analyse data (O'Brien, 1992). Regression models are the most commonly applied statistical tools in criminal justice and empirical research for building models of crime and society (Kraak, 1999). The basis of regression analysis is not simply to understand relationships between variables, but also to understand the most important factors responsible for effecting changes (Weisburd and Britt, 2007). The following sections discuss regression models in more detail.

3.6.1 Regression Models

Regression models are statistical tools used for quantifying the relationships between variables. The researcher usually seeks to find the causal effect of one variable upon another (Sykes, 1993). Regression analysis is used to explain the relationship between the dependent variable (response or outcome) and independent variable (predictor or explanatory). When there is only one predictor variable it is called simple bivariate regression, but when there are more than one predictor it is referred to as multiple regression (Faraway, 2002). The goal of regression analysis is to estimate the association between one or more explanatory variables and a single outcome variable (Hoffmann, 2010). Weisberg (2005) has emphasised that predicting values of a response variable depends on which predictors are important.

Conceptually, the most efficient way of ensuring that the best model is achieved is by computing all the possible subsets of variables (Rawlings *et al.*, 1998). This is possible with a small number of variables but it becomes a major computing problem when there are many independent variables from which to choose.

Regression modelling has been employed in a number of studies especially for building models of the crime system. The most commonly used regression techniques include; multiple variate regressions (LaGrange, 1999; Markson *et al.*, 2010; Lane and Meeker, 2000; McGarrell *et al.*, 1997; Mukherjee *et al.*, 2014; Nathans *et al.*, 2012; Livingston *et al.*, 2014; Armstrong *et al.*, 2015), Poisson and negative binomial regression (Osgood, 2000; Campbell and Campbell; Poulsen and Kennedy, 2004; Efroymsen, 1960), logistic regression (Goudriaan *et al.*, 2006; Kim

et al., 2013; Kleck *et al.*, 2006) and multilevel regression models (Sampson *et al.*, 1997; Foster *et al.*, 2014; Sutherland *et al.*, 2013; Jennings *et al.*, 2014). This research develops a multiple regression model to quantify the relationship between crime and its related correlates because crime can be as a result of a combination of a number of factors not just a single factor (Weatherburn, 2001). Section 3.6.1.1 describes multiple regression modelling in more detail.

3.6.1.1 Multiple Regression Model

A multiple regression model is a technique for quantifying the relationship between a dependent variable and two or more independent variables (Ackerman and Rossmo, 2015). The assumption is that the simultaneous effects of explanatory variables added into the model will produce more variation in the dependent variable (Mennis, 2003). Multiple regression models are widely used in the field of social science (Peterson and Robbins, 2008), especially for building models of the urban social system (Ajimotokin *et al.*, 2015). The general specification of a multiple regression model takes the form:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_nX_n + \varepsilon \quad (3.1)$$

where Y is the value of the dependent variable, β_0 is the constant intercept, $\beta_1, \beta_2, \beta_3, \beta_n$ are the slope coefficients of the independent variables X_1, X_2, X_3, X_n , while ε is a standard error of component.

Stepwise Multiple Regression Model

More recently, attention has focussed on identifying the suitable subsets of variables without computing all possible subsets. Conceptually, the most efficient way of ensuring that the best model is achieved is by computing all the possible subsets of variables (Rawlings *et al.*, 1998). This is possible with a small number of variables however, it became a major computing problem when there are many independent variables from which to choose. However, finding the best subset model is not always a straight-forward procedure, as most statistical packages, such as Statistical Package for Social Science (SPSS), provide a collection of algorithms for this procedure referred to as stepwise multiple regression (Pitner *et al.*, 2012). These methods proposed by Whittingham *et al.* (2006) utilise the Ordinary Least Square (OLS) approach that the residual sums of squares cannot decrease when a variable is dropped from a model. Least

squares is a fitting technique applied to linear regression that provides a solution to finding the best fitting straight line through a set of points. Rawlings *et al.* (1998) stressed that stepwise regression reduces the amount of computing than is required by identifying and selecting subset models (not necessarily the best) and essentially stepwise regression is for variable selection (James *et al.*, 2013). The stepwise regression method involves two approaches: forward selection and backward elimination. Whereas forward selection adds variables one at a time in the model based on correlation with the dependent variable, the backward elimination starts with a full model and drops one variable at a time until the model contains only one variable (Rawlings *et al.*, 1998; James *et al.*, 2013). In both methods, the explanatory power of the variables determines their retention or removal from the regression model. Stepwise multiple regression procedure is given in Equations 3.2 , 3.3 and 3.4 based on Zhou and Jiang (2016):

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (3.2)$$

If X_1 is significant it is retained, otherwise dropped

$$Y = \beta_0 + \varepsilon \quad (3.3)$$

In the next iteration variables with a higher correlation with Y are added

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (3.4)$$

The procedure stops when all the variables have been checked for their significance and a final model is then identified.

However, the process of adding and dropping variables associated with stepwise regression has been criticised as it is possible to miss the optimal model, removing less significant predictors might increase the significance of other variables which may lead researchers to overstate the importance of the remaining variables (Rawlings *et al.*, 1998). Despite the limitations of stepwise multiple regression method, it is widely used in ecological studies (Pitner *et al.*, 2012) and it remains especially useful for model building with several potential variables (Wooldridge, 2012).

The stepwise regression method has been used in a number of studies to build models of crime correlates. For example, Piza (2012) employed the stepwise multiple regression technique using 1985 Japan population Census data, 1987 Employment Status Survey data and 1987 Japan

Statistical Yearbook data to explore the relationship between economic structure and crime. He found a significant positive relationship between unemployment rates and homicides and robberies.

Tseloni (2006) used the 1995 Survey of Orange County residents, California, to quantify the relationship between subcultural diversity and fear of crime and gangs. She included age, gender, ethnicity, education, income, home ownership and residential location as variables in a stepwise regression model. She found that the variables used have different explanatory power for predicting crime and fear of gangs. She also found that subcultural diversity (measured by age, education, race, income and home ownership) more strongly predict fear of crime. In a related study, Tsushima (1996) also employed stepwise regression method to examine the relationship between fear of crime and socio-demographic characteristics (age, gender, education and income, fear of victimisation (criminal threat), disorder (noise), satisfaction with law enforcement and neighbourhood safety (walking alone at night). Tsushima found that increased age lowered the level of fear of crime for women but not for men and that victimisation was more pronounced for women than men. She also found that higher educational attainment decreases fear of crime for urban dwellers but not rural residents. The limitations of this study is the use of cross-sectional data and non-control for neighbourhood differences which lead to a weaker relationship between fear of crime and associated independent variables. This research will address this limitation by controlling for neighbourhood differences using diversity statistics.

Wooldridge (2012) employed data from the Program Development and Evaluation System (ProsDES) on juvenile delinquency 1996-2003, 2000 US Census data, collective efficacy data from Philadelphia Health Management Corporation (PHMC) and crime data for the city of Philadelphia to quantify the effects of neighbourhood characteristics and spatial distribution of urban delinquency. Wooldridge used the stepwise regression method with variables as follows: participation, neighbourliness, improvement, belonging, renters, receiving benefits, and Hispanic and African-Americans. It was found that delinquency-related crimes such as violence (drugs, weapons and homicides) were concentrated in disadvantaged areas after controlling for poverty, an area with a high concentration of minority ethnic groups tended to exhibit high delinquency rates. The major limitation of this study was that no neighbourhood level factors were included to better understand the effects of interactions among individuals.

A common factor in these studies was the use of global OLS methods. OLS linear regression provides a global model to generate predictions or to model a dependent variable in terms of its relationships with a set of explanatory variables (Walker and Maddan, 2013). The global model parameters derived from the OLS method assumed that variables are constant over space (Charlton *et al.*, 2009). However, this assumption is invalid in most cases as spatial variation in relationships is not stationary (Erdogan *et al.*, 2013) and tend to vary across space known as *spatial heterogeneity*, hence the need for a localised approach such as Geographically Weighted Regression (GWR) (Charlton *et al.*, 2009).

Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a (local) modelling technique to estimate regression models with spatially varying relationships. The GWR model provides a useful method for addressing non-stationarity associated with variation in neighbourhood characteristics (Arnio and Baumer, 2012), especially where spatial heterogeneity is not adequately described by the stationary OLS models (Lu *et al.*, 2014). The basic idea of spatial regression modelling is provided by Tobler’s *First Law of Geography* which states that “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970, p.236). Additionally, spatial dependency in the data means that the value of the dependent variable in one spatial unit (community area in this context) is affected by the independent variables in nearby areas (Charlton *et al.*, 2009). Spatial dependence refers to the propensity in characteristics of nearby locations to have influence on each other (Anselin, 1989). However, the GWR approach employs a moving window weighting technique where local characteristics of the area are used for modelling (Charlton *et al.*, 2009). In other words, nearer observations have more influence in estimating the set of local parameters than observations further away (Fotheringham *et al.*, 1998). GWR method is fully described by Fotheringham *et al.* (2003). The typical GWR version of OLS is given as:

$$y_i(u) = \beta_{0i}(u) + \beta_{1i}(u) X_{1i} + \beta_{2i}(u) X_{2i} + \beta_{3i}(u) X_{3i} \dots + \beta_{ni}(u) X_{ni} + \varepsilon \quad (3.5)$$

$y_i(u)$ is independent variable at location i , (u) is a vector of two dimensional coordinates describing location i , $\beta_{0i}(u)$ is the intercept parameter at location i specific to that location, $\beta_{ni}(u)$ X_{ni} is the local regression coefficient for n th explanatory variable at location i .

In this research the GWR model will also be employed to explore spatial variation relationships.

3.7 Spatial Analyses

Spatial analysis has been defined by Bailey and Gatrell (1995) as the quantitative study of phenomena that are located in space. Fotheringham and Rogerson (2013) also defined spatial analysis as the general ability to manipulate spatial data into different forms and extract additional meaning as a result. The advent of Geographical Information Systems (GIS) in the recent past brought about the opportunity for handling spatial data (O'Sullivan and Unwin, 2010). Researchers are not only interested in the distribution of values observed in data but also in the spatial distribution of those entities over space. Since spatial data analysis is descriptive and exploratory, this is an important first step in all spatial analysis which is suitable for large and datasets (O'Sullivan and Unwin, 2010). GIS and spatial analysis, in particular, have proved to be important tools for studying criminal activity (Murray *et al.*, 2001). Additionally, crime analysis enables researchers to have a greater understanding of the complexities associated with criminal activities and factors contributing to their occurrence in different areas (Savage and Burrows, 2009). Spatial analysis methods commonly used by researchers are discussed in detail in the following sections.

3.7.1 Choropleth Mapping

Choropleth maps also called 'area symbol maps' are maps showing polygons for political or administrative convenience and extensively used for data presentation and visualisation (Schabenberger and Gotway, 2004). Choropleth maps are used to represent different events (e.g. crime and population) aggregating them according to geographical areas by assigning graduated shades and symbols (Poulsen and Kennedy, 2004). The density of colour shades in choropleth maps indicates increases in the associated events. For example, darker colours are used to show larger data values and lighter colours to show smaller values (Brewer, 1997). Choropleth mapping is one of the common techniques employed by researchers to represent aggregated events such as crime information (Savage and Burrows, 2009). Figure 3.5 shows an aggregated choropleth map of burglary crime rates in Leeds, 2011-2015.

Additionally, choropleth maps are useful for spatial data visualisation making it easier to describe spatial patterns and relationships (Kraak, 1999), and for making comparison across time

and space (Paynich and Hill, 2010). However, choropleth maps have their own shortcomings. For example, small incidences of crimes are likely to be overshadowed by a higher category such that areas with lower values are subsumed by areas with higher values (Chainey *et al.*, 2002). Furthermore, Harries (1999) also argued that areas or polygons are not the same in size, therefore, it would not be correct to say that areas with the same number of crime incidents are the same. Inconsistency in results can also arise from the use of different use geographical units termed the *modifiable areal unit problem* (MAUP) (Openshaw, 1979). For example, geographical boundaries such as census enumeration areas, administrative boundaries and political districts are all subject to modification resulting in different values of regression parameters being calibrated for different apatial systems (Wong, 2004) Despite their limitations, choropleth maps are widely used for spatial data exploration (Eck *et al.*, 2005).

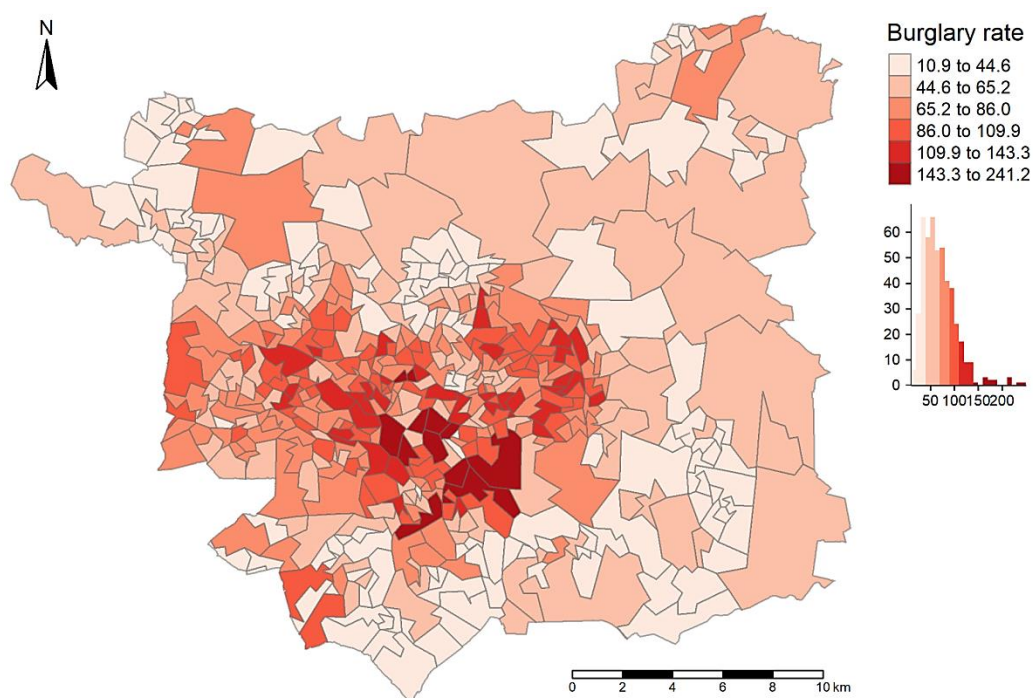


Figure 3.5: Burglary crime rates in Leeds (2011-2015)

A number of researchers have used choropleth maps in different studies including, for example, in the area of health to analyse the incidence of mortality (Brewer and Pickle, 2002) and to compare disparity in community health (Cromley and Cromley, 2009). Similarly, choropleth maps are employed in population studies (Mennis, 2003). A more common application of choropleth maps is in crime studies. For example, Murray *et al.* (2001) employed choropleth

maps to highlight the incidence of crimes in Brisbane, Australia. The results indicate the capabilities of the method for illustrating urban crimes.

Ackerman and Murray (2004) employ the choropleth mapping technique to examine the spatial patterns of neighbourhood crime in Lima Ohio, USA. They found that many smaller communities have above average crime rates for their sizes. Their study demonstrates the capability of choropleth mapping for spatial data exploration and visualisation.

Additionally, Brunson *et al.* (2007) also used choropleth maps when comparing different crime visualisation methods. Their conclusion was that each method has its merit for the crime analyst. Building on previous studies, in this research, choropleth maps will be used to visualise patterns and spatial distribution of different phenomena such as diversity and crime.

3.7.2 Hotspot Mapping

Chainey and Ratcliffe (2005) emphasised that crime does not occur at random but has an inherent geographical quality and therefore takes place at certain locations. There is no common definition of a crime hotspot. However, Eck *et al.* (2005) have stressed that areas of high clusters of crime are commonly referred to as hotspots. Therefore, hotspots can be referred to as areas that have above average number of criminal or disorder events. They can also be areas where people have a higher than average risk of victimisation (Eck *et al.*, 2005). Hotspot mapping is a popular analytical technique that enables visualisation of clusters of events such as crime, knowing which can help in the appropriate allocation of resources (Chainey *et al.*, 2008). There are different techniques such as spatial ellipses, grid thematic mapping and kernel density estimation used for identifying hotspots of crime (Van Patten *et al.*, 2009) depending on different types of crimes (Eck *et al.*, 2005). However, kernel density estimation (KDE) is one of the most widely used methods (Chainey *et al.*, 2008; O'Brien *et al.*, 2016). The KDE method takes into account incremental values of the mean in the statistical and spatial distribution of crime incidence (Barrantes and Sandoval, 2009). The concentration of certain crimes at some places could be attributed to the interaction between the victim and offender as well as the opportunity that avails itself (Brantingham and Brantingham, 1984; Cornish and Clarke, 1987; Chainey *et al.*, 2008). Figure 3.6 shows the hotspot of burglaries in Leeds using KDE.

Researchers have used hotspot maps to analyse crime events. For example, Bowers *et al.* (2004) used 1997 burglary data for south Liverpool to predict locations of burglary crime. They employed hotspot mapping techniques and moving averages in their analysis. The moving average is a time series calculation that takes the averages of subset data points of the full dataset. They found that the method is useful for a shift-by-shift deployment of police personnel but the limitation of their work is that they only used two months historical data which could affect the accuracy of their predictions.

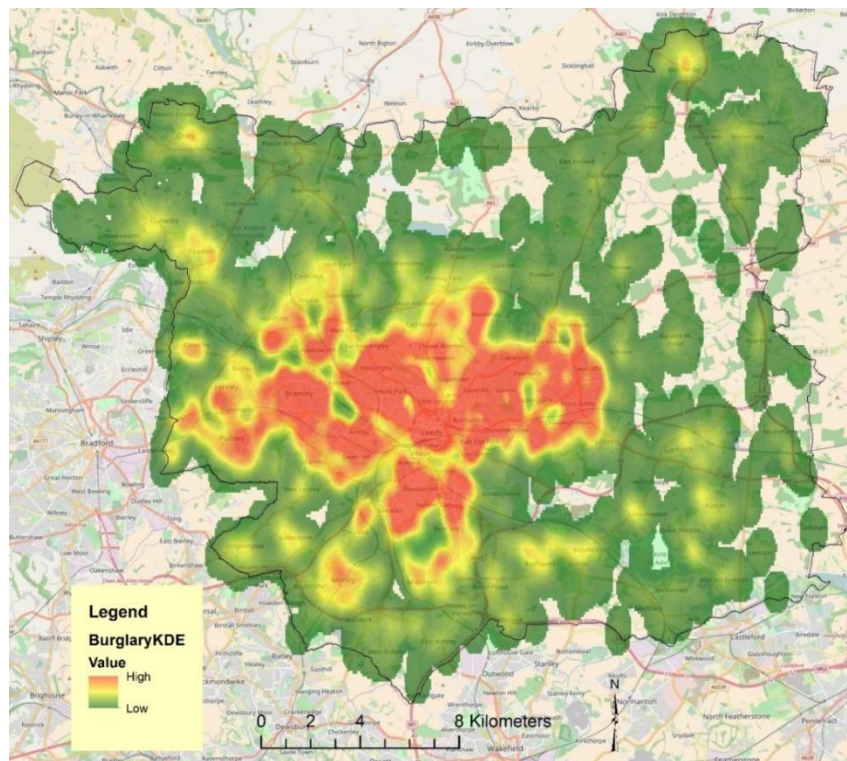


Figure 3.6: Kernel density estimation of burglary crime (2011-15) in Leeds

Piekut *et al.* (2012) employed KDE hotspot mapping techniques and scan statistics to visualise crime clusters in the city of Kyoto, Japan, using snatch-and-run offences data for 2003 and 2004. They found that the combined use of two statistical techniques revealed temporal inter-cluster associations, highlighting the importance of the combined approach for spatiotemporal exploratory data analysis of crime clusters.

Simpson (2004) used a hotspots optimisation tool, an application for spatial data mining for hotspots mapping. Simpson used historical data for 2004 to 2007 from a north-eastern city of

USA and found that the hotspot optimisation approach is capable of accurately mapping crime hotspots.

More recently, Malleon and Andresen (2015), used crowd-sourced data and local spatial statistics (Getis Ord and Geographical Analysis Machine) to explore the shift in crime hotspots in Leeds, UK. They found that the population-weighted hotspot of crime is different from when the population is estimated from social media data, as opposed to standard information on people's home locations. The main criticism of their work is that while hotspot maps might be used to show patterns, they are less likely to provide an explanation for the occurrence of crime, especially when the analysis is founded on the basis of noisy (unstructured) data; Eck *et al.* (2005) stressed that factors that give rise to the hotspots differ from place to place. The present research will address this problem by relating the diversity of different areas to hotspots of crimes in order to gain insights into possible links between them. Figure 3.7 shows KDE of burglaries for different spatial catchment range (10%, 25%, 50% and 75%) in Leeds.

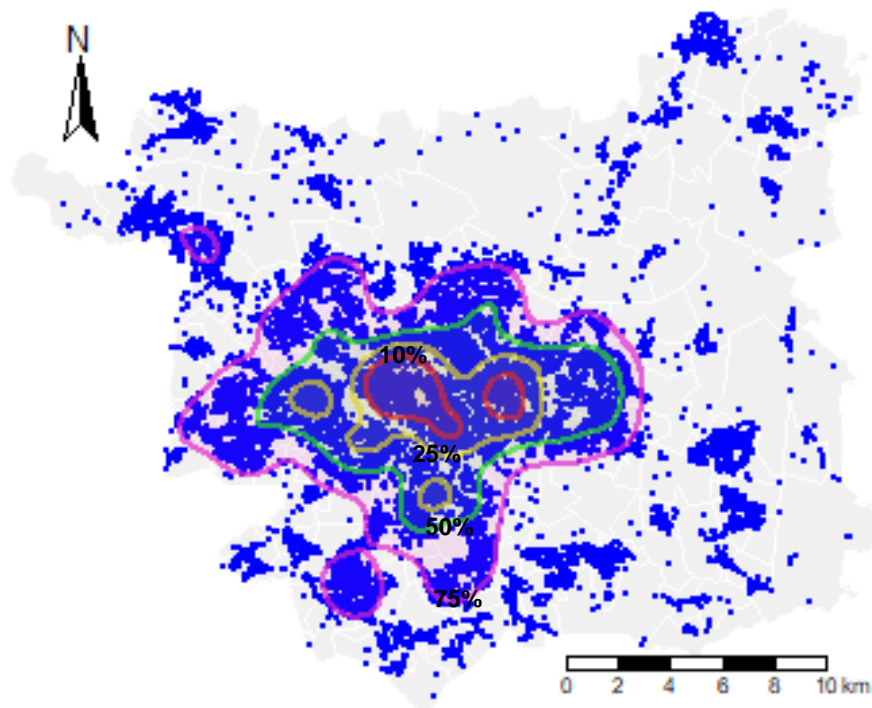


Figure 3.7: Kernel density estimation of burglary ranges (10%, 25%, 50% and 75%) in Leeds

3.8 Spatial Statistical Methods

Spatial statistical analysis is the application of statistical methods to query spatial data with a view to finding its suitability for using a statistical model to represent it (O'Sullivan and Unwin, 2010). Spatial techniques have been recognised as important tools used by researchers for analysing the occurrence events such as crime (Murray *et al.*, 2001) and widely used in social science for the analysis of spatial data (Schabenberger and Gotway, 2004). Several spatial statistical methods have been used to analyse spatial data, for example, spatial autocorrelation, OLS and GWR.

3.8.1 Spatial Association

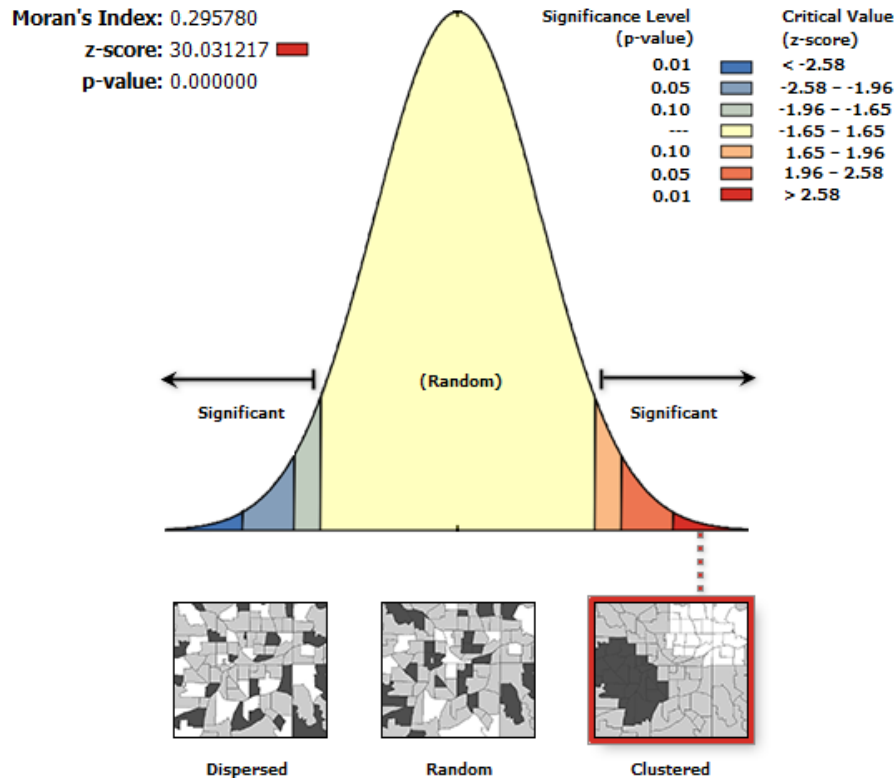
The spatial association and spatial dependency of geographic patterns are indicated by spatial autocorrelation. In modelling spatial data, when measurement at one location is influenced by the measurement at neighbouring locations then spatial autocorrelation is said to be present (Collins *et al.*, 2006). For example, the distribution of crime in city neighbourhoods can be affected by the factors of location and by social interaction (Collins *et al.*, 2006). The measurement of spatial autocorrelation enables examination of how social events such as crime occurred at different locations (Kubrin and Weitzer, 2003). Homogeneity of spatial patterns is an indication of a positive relationship while heterogeneity of spatial patterns indicates a negative relationship.

3.8.1.1 Modelling Spatial Processes

Anselin *et al.* (2000) stressed that modelling spatial processes are most appropriately examined by spatial autocorrelation. The most widely used global method of analysing clusters (spatial autocorrelation) is Moran's I (Levine, 2004; Eck *et al.*, 2005; O'Sullivan and Unwin, 2010; Ibrahim *et al.*, 2015) and to a lesser extent Geary's C (Getis and Ord, 1992). Moran's I is based on the covariance among the designated locations, while Geary's C takes into account numerical differences between associated locations (Getis, 1999; Eck *et al.*, 2005). Figure 3.8 shows the result for Moran's I spatial autocorrelation for burglary crimes in Leeds.

The common feature between Moran's I and Geary's C is that both of them are applied globally within the area under study. Therefore it is appropriate to use Geary's C (General G) in conjunction with Moran's I in order to reveal patterns not revealed by the use of Moran's I alone (Getis and Ord, 1992). Moran's I tends to provide some insights regarding the global level of

spatial autocorrelation in the occurrence of events (Murray *et al.*, 2001) but is not practicable in identifying local clusters with heterogeneous areas (Ord and Getis, 1995). However, despite its limitation, Moran's I is still to date the most widely used global method for investigating spatial autocorrelation (Ibrahim *et al.*, 2015).



Given the z-score of 30.031217256, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Figure 3.8: Spatial autocorrelation statistics for burglary crime in Leeds

3.8.1.2 Global Moran's I

The Spatial Autocorrelation (Global Moran's I) tool is an inferential statistic, which means that the results of the analysis are always interpreted within the context of its null hypothesis. For the Global Moran's I statistic, the null hypothesis states that the attribute being analysed is randomly distributed among the features in the study area or in other words, the spatial processes promoting the observed pattern of values is a factor of random chance. The Moran's I statistic for spatial autocorrelation based on Thuczak (2013) is:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \quad (3.6)$$

where \bar{x} is the mean of the x variable, w_{ij} are the elements of the weight matrix,

$$S_0 \text{ is the sum of the elements of the weight matrix: } S_0 = \sum_i \sum_j w_{ij} \quad (3.7)$$

where i and j are features being weighted

A number of researchers have used spatial autocorrelation in their analyses such as epidemiology and crime. For example, Cracolici and Uberti (2009) used spatial autocorrelation to investigate the spatial distribution of crime in Italy. Their findings highlighted that certain socio-economic variables have positive impacts on criminal activities. Andresen (2006) also used spatial autocorrelation to explore spatial aspects of crime activities in Vancouver, Canada employing social disorganisation theory and routine activity theory. His results show a positive effect of using the spatial autocorrelation technique over space. This research will also apply spatial autocorrelation technique in order to better understand patterns in the datasets.

3.9 Diversity Indices

A diversity index is a mathematical measure of the distribution of ethnic group diversity in an area (McIntosh, 1967). Therefore, diversity indices provide more insights about community composition than a simple analysis of data (numbers and percentages) (Bains, 2005). Although diversity is difficult to quantify, there is no consensus about which indices are more appropriate, partly because of the multitude of indices proposed (Morris *et al.*, 2014). A number of indices of diversity including Berger Parker, McIntosh, Camargo's, Smith and Wilson, Clarke and Warwick are used in different disciplines (see Magurran, 2004). However, the most widely used indices to quantify diversity are the Shannon (H) and Simpson (D) indices (Bains, 2005). While the Shannon (1948) index seeks to measure the proportion (dominance) of individuals from a group, the Simpson (1949) index represents the probability that two randomly chosen individuals belong to different ethnic groups.

Furthermore, when measuring diversity, richness (abundance) and evenness (distribution) are the paramount considerations. Richness is the number of different groups represented in the

community where each person of the population can be allocated to one group and each group is distinct. On the other hand, evenness refers to the equitable distribution of individuals among the different groups. While the Simpson's index has taken into account both considerations (richness and evenness) (Bains, 2005), the Shannon index only takes into account the richness (abundance) (Bains, 2005). This tends to make comparison across different communities nearly impossible (Barrantes and Sandoval, 2009) and introduces classification problems such that an individual can belong to any group (Hill *et al.*, 2003). Additionally, while the range in values of D irrespective of the number of groups is between 1 and 0, the bigger the value of D the more the diversity and does not take sample size into account, the range in value of H depends on the number of sample groups under consideration. This tends to make interpretation difficult (Magurran, 2004). Because of the limitation of the Shannon index, the Simpson index is more widely preferred (Bains, 2005) and is written as:

$$D_i = 1 - \frac{\sum_i n_i(n_i - 1)}{N(N - 1)} \quad (3.7)$$

Several researchers have used Simpson's diversity index in their analyses. For example, Rees and Butt (2004) used the Simpson index to measure diversity in England between 1981-2001. They found a rapid growth in the population of minority ethnic groups. Bains (2005) used UK Census data 2001 to quantify diversity in London using the Simpson index finding high concentrations of non-Whites in most London boroughs relative to the England and Wales average. Stillwell and Phillips (2006) examined the diversity of ethnic populations of Leeds (Asians specifically) using the Simpson index and data from 1991-2001 UK Census. They found evidence for the concentration of Asian population in inner Leeds, indicating greater potential for residential mobility from the heterogeneous multi-ethnic inner city areas.

Though the concept of diversity is multi-dimensional which includes demographic, socio-economic, health and physical disabilities, religion among others (Zanoni and Janssens, 2004; Jonsen *et al.*, 2011). Previous studies (Alesina and Ferrara, 2005; Stillwell and Phillips, 2006; Habyarimana *et al.*, 2007; Gijsberts *et al.*, 2012; Sturgis *et al.*, 2014; Jivraj and Simpson, 2015; Alesina *et al.*, 2016) have concentrated on ethnic diversity in their analysis. One study that attempted to include other variables beyond ethnicity is that by Piekut *et al.* (2012) which

employed Census data to quantify diversity of two cities Leeds in the UK and Warsaw, in Poland. They used clustering and Simpson’s index of diversity as methods in their analysis. They found a little overlap in community types across the two cities. In this research, the Simpson index will be used for constructing diversity statistics using different components representing crime and community cohesion (Table 3.2). Additionally, diversity statistics will be used for later regression modelling of crime and community cohesion.

Table 3.2: Components used to measure different diversity metrics

Diversity	Components included
Age	10-14, 15-19, 20-24, 25-29, 30-44, 45-59, 60-64, 65-74
Family structure	Lone parent no dependent, Lone parent one dependent child, Lone parent two or more dependent children, Married couple no children, Married couple one dependent child, Married couple two or more children
Ethnicity	All 18 ethnic groups included
Employment	16-64 Managers/Directors, 16-64 Professionals, 16-64 Associate Professionals, 16-64 Administration and Secretariat, 16-64 Skilled Trade, 16-64 Caring Leisure and Services, 16-64 Sales and Customer Services, 16-64 Process Plants and Machines, 16-64 Elementary Occupation
Education	16-over qualification level 1, 16-over qualification level 2, 16-over qualification level 3, 16-over qualification level 4
Residence length	Length of residence: Less than two years, Less than five years, More than five years, Ten years above, Born in the UK

3.10 Computational Models

Computational models are mathematical algorithms used to simulate a set of processes observed in the real world in order to gain an understanding and to predict the outcome of these processes (Hill *et al.*, 2001; Calder *et al.*, 2018). Computational models are increasingly used for research in different disciplines including social sciences (Poile and Safayeni, 2016). Brantingham and Brantingham (2004) argued that conventional statistical methods used in empirical crime research rely on extrapolation of past data. The central drawback of statistical crime models is that they are not able to capture the underlying processes that drive crime system (Malleon and

Birkin, 2012). However, due to the increasing complexity and dynamics underlying sociological systems, computational models are needed to overcome the limitations of conventional statistical methods (Brantingham *et al.*, 2005). Within the social sciences, two broad computational models including microsimulation and agent-based models are widely used for simulating geographical systems (Heppenstall *et al.*, 2011) and criminal justice systems (Stewart, 2014).

Microsimulation Models

Microsimulation modelling (MSM) is a micro-based methodology that uses micro units of analysis such as individuals and households and firms using surveys or administrative datasets (e.g census and survey data) (O'Donoghue, 2014). The idea behind MSM is to represent the constituent units with individual characteristics which comes from real data rather than average values (Bae *et al.*, 2016) and the use of these attributes for policy analysis at micro-level (Ballas *et al.*, 2005). MSM is widely applied within the social sciences to better understand how components of a dynamic system within a society are related (Groff and Birks, 2008). Previous researchers have experimented with MSM to explore crime (Kongmuang, 2006; Auerhahn, 2008; Hussain *et al.*, 2012; Stewart, 2014; Bailey, 2015). However, MSM has a number of issues related capture spatial variations in a range of variables; the matching processes in the estimation techniques; variations in household types in the surveys being reweighted and issues with similar household types showing different behaviours (Birkin and Clarke, 2012).

Agent-based Models

Agent-based models (ABM) are methods one approach for modelling complex social systems in which agents interact based on specified rules (Malleon *et al.*, 2014) to better understand and explore social phenomena (Eberlen *et al.*, 2017). Agents are entities that are capable of interacting with each other and with their environment (Malleon *et al.*, 2014). The interaction among agents and their environment often aggregate to create unexpected patterns, it is such emergent patterns (such as residential segregation and inequality) that are of interest to sociologists and policymakers (Bruch and Atwell, 2015). ABM has been used in the study of urban social systems, especially crime modelling (Malleon *et al.*, 2010; Malleon and Birkin, 2012; Gerritsen, 2015; Weisburd *et al.*, 2017; Groff *et al.*, 2018). Despite the increasing popularity of ABM in simulating complex social systems, there are drawbacks associated with ABMs. Modelling human social behaviour is complex (Malleon, 2010) and selecting adequate

number of parameters, features and behaviours to include in the model can be challenging (Eberlen *et al.*, 2017). Additionally, unlike conventional statistical models, ABM is computationally expensive technique (Malleon *et al.*, 2014).

The major difference between MSM and ABM is that MSM data-driven one directional interaction models that attempt to simulate the impact of the policy on individuals rather than policy on individuals; while ABM are theory-based dynamic interaction models between individual (agents) that decide within their environment (Crooks and Heppenstall, 2012; Bae *et al.*, 2016). However, of MSM and ABM techniques are conceptually linked, they create models based on bottom-up approach (Ballas *et al.*, 2019). These methods will not be used in the remainder of the thesis.

3.11 Concluding Remarks

This chapter has reviewed traditional methods of modelling spatial relationships in urban social systems. Different data sources to be used in regression modelling were also reviewed. As can be seen though, the relationship between community composition and crime is complex and multi-faceted; there are some core factors that regularly emerge as important determinants of crime rates. Here we justify the inclusion of variables that represent crime and community cohesion. These variables will be used in chapter 5 to calibrate models of crime and community cohesion so as to explore their significance.

As the data to be used for this research has spatial attributes, we reviewed spatial and statistical methods available for analysing spatial data. It has been suggested that standard regression variables can be adjusted to capture a range of places in which social cohesion plays a part across the crime system. To quantify this, we reviewed different indices used to measure diversity within a population, and have decided to adopt Simpson's diversity index in this research. This will provide justification for the inclusion of diversity statistics in "advanced" social media modelling to explore community cohesion and crime.

Chapter 4

Non-Traditional Methods: Social Media for Community Cohesion and Crime

4.1 Introduction

In chapter 3, “traditional” methods and data for modelling urban social systems and crime were reviewed. Chapter 4 proceeds to discuss “non-traditional” methods for modelling community cohesion and crime with new forms of social media. One of the objectives of this research is to investigate the feasibility of using new data sources to explore crime and community cohesion, especially because the limitations surrounding traditional data relied upon by social scientists have resulted in researchers seeking alternative data sources (Danneman and Heimann, 2014). Engaging with new forms of social media data can help provide insights into social processes such as social interactions (Savage and Burrows, 2007). Section 4.2 reviews relationships between social media and social cohesion with specific emphasis on Facebook and Twitter. Sources of social media data and methods of collection are also described. Section 4.3 will explain the relationship between social media and community cohesion with reference to action in communities. Section 4.4 will then move on to describe how the social media and community cohesion can be applied for understanding communities. Application of social media data by law enforcement will be described in Section 4.5 and 4.6 discuss social media and social capital in relation to crime. Section 4.7 will discuss methods of social media analysis and different algorithms used. Section 4.8 presents concluding remarks.

4.2 Social Media and Social Cohesion

Social interactions and social networks are now easier to capture through the social media, although the data, like any, has biases. Social media refers to networked communication platforms (e.g. Facebook and Twitter) that allow users to connect with other people and share user-generated content (Ellison and Boyd, 2013). Social media provides an enabling platform for the formation of social bonds (cohesion) among people as well as the exchange of information and creation of social networks (Cornelius *et al.*, 2009; Gorman, 2013; Wakamiya *et al.*, 2013; Angus *et al.*, 2015). Public participation in social media (Facebook and Twitter) has recently increased. The increase could be attributed to the internet becoming more accessible to people,

especially on mobile phones. The increasing use of the internet is changing the way communities are engaging and interacting in the public sphere (Bonsón *et al.*, 2015). For example, in the UK, the number of people using Twitter and Facebook for social networking on a daily basis has increased from 45% in 2011 and 61% in 2015 to 63% in 2016 across all age groups. Although, there is a strong bias towards the young (16-24 (93%) social networking; 25-34 (85%); 35-44 (72%)) but older generations are also engaging in social networking (45-54 (55%); 55-64 (44%), 65 and over (15%)) (ONS, 2015b; ONS, 2016b). Additionally, while SM is often viewed as biased towards the young, Facebook is of particular interest because it is increasingly becoming important for older people (Miller, 2016). Recent statistics on SM usage in the UK, show that older adults are closing the gap with those aged 45-54 (68%), 55-64 (51%) and 65 and over (27%) using SM on a daily basis, though the younger adults 16-24, 25-34 and 35-44 maintained their lead with 96%, 88% and 83% respectively (ONS, 2017d). The major reason for using social media was to find out what is happening in the local area (ONS, 2016c).

In terms of the relationship with off-line, Williams (2010) stressed that internet access can promote community building by providing an enabling environment for communication within and between communities. Furthermore, Sawyer and Chen (2012) emphasised that increasing access to the internet promotes interconnectedness and interdependence in a culturally diverse society. Steinfield *et al.* (2008) explain the increase in social media engagement by emphasising that social media networks are increasingly being used as part of an individual daily routine where people stay connected with the events in their communities. As a result, Hariche *et al.* (2011) emphasised that web resources can be used to foster social cohesion by creating a network of relationships between communities in a society. It is apparent that social media can strengthen local links within neighbourhoods, intertwining online and offline relationships (Harris and McCabe, 2017) and enhancing local social capital generation (Matthews, 2016). For example, a study of the relationship between frequent social media usage and overall social ties (Hampton *et al.*, 2011) suggests that social media offers opportunities for social networking for communities (Harris and McCabe, 2017).

In general, when some people think of social media like Facebook, Twitter, Myspace or YouTube, they think of specific enabling technologies especially for marketing and political activities (Baruah, 2012; Gao and Liu, 2015). However, the importance of social media has gone beyond simply a platform for promotion and marketing of products (Paquette, 2013) or even

eliciting individual votes in political campaigns (Jungherr, 2016). Today the emphasis is also beyond the individual, and their buying/voting power, with an emphasis on social interactions, local community support and participation, including in the area of crime reduction activities (Denef *et al.*, 2013), especially in the UK (Malleon and Andresen, 2015). Social media tools such as Twitter and Facebook have a great potential for community-based crime reduction efforts (Featherstone, 2013). Social media tools are useful for mobilising community support and encouraging community action, especially in solving local problems such as crime (Humphreys, 2010; Baruah, 2012). Table 4.1 shows the most popular social media platforms the date they were established and their key purpose.

Table 4.1: Popular social media networking platforms

Social media	Year established	Purpose
LinkedIn	2003	Professional networking
Myspace	2003	Music sharing
Skype	2003	Video call/messaging
Facebook	2004	Social connections/messaging/chat/business
Flicker	2004	Embed photo sharing
YouTube	2005	Video sharing
Reddit	2005	Social news/networking
Twitter	2006	Messaging (tweets)
Tumblr	2007	Video/audio/photo sharing
WhatsApp	2009	Messaging/video/audio/documents
Foursquare	2009	Local and discovery service
Pinterest	2010	Small businesses
Instagram	2010	Video/photo sharing
Snapchat	2011	Image messaging
Google+	2011	Small businesses

Source: Anthony (2016)

The emergence and use of social media channels such as Facebook and Twitter generate large-scale databases of human spatial activities, creating wide research interests (Tufekci, 2014). One of the ramifications of using social media platforms, especially Twitter and Facebook is their

potential for exploring the dynamics of human collective behaviours to uncover patterns and phenomena (Bentley *et al.*, 2014). Additionally, these data can be analysed to gain insights into human social behaviours (Tufekci, 2014). For example, analyses of Twitter data through text mining (sentiment) can be important in revealing patterns of public emotions. Facebook data analyses (engagement) are important in understanding patterns of social interactions in different areas which can help especially in planning for community building (Williams *et al.*, 2013), albeit with ethical issues around the reuse of public and private data. Figure 4.1 shows the number of global social media active users. However, the two most popular social media platforms (Facebook and Twitter) used for research are discussed in more detailed below (Section 4.2.1 and Section 4.2.2).

The relationship between social cohesion and crime is well researched (see section 2.3). Previous research has established the mediating effect of social cohesion on crime (Sampson *et al.*, 1997; Hirschfield and Bowers, 1997; Markowitz *et al.*, 2001; Forrest and Kearns, 2001; Wedlock, 2006; Uchida *et al.*, 2013). Gao and Liu (2015) emphasised that social media tools especially Facebook and Twitter, provide a useful medium for understanding dynamics of social interactions and exploring social problems such as crime. Hence in this research, analysis of social media data (Facebook and Twitter) might provide useful insights into the relationship between community cohesion and crime.

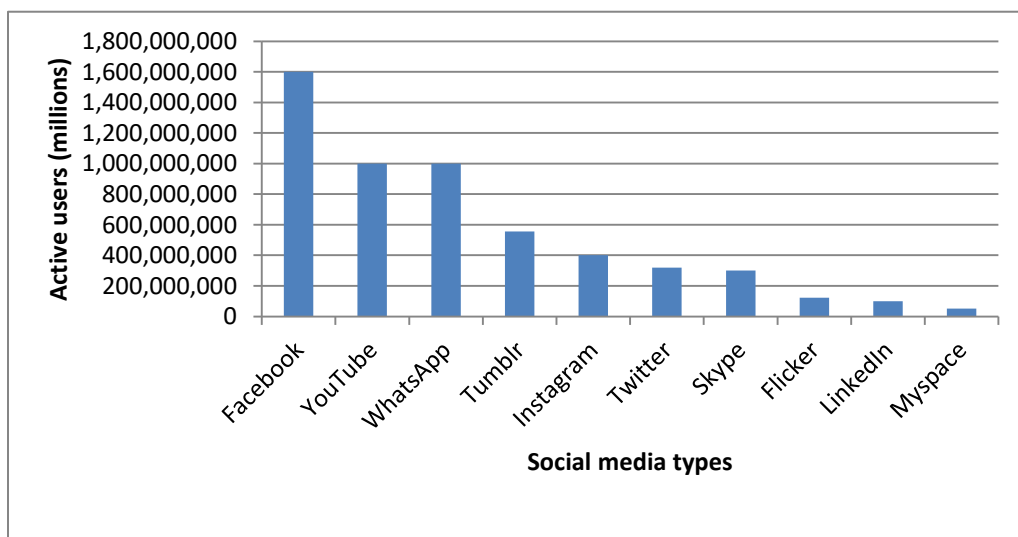


Figure 4.1: Global social media active users. Source: Statistica.com (2016)

4.2.1 Facebook

Established in 2004, Facebook is the fastest growing social media platform (Sachs *et al.*, 2011; Tess, 2013; Chaffey, 2017) and undoubtedly the most popular social networking site (Junco, 2015). The number of Facebook active users has grown from about 400 million users in 2009 to 1.23 billion in 2014 (Ryan *et al.*, 2014). As of March 2016, Facebook has 1.6 billion active users, 82.2% of them outside America and Canada (Facebook, 2016). To establish a connection, users are required to create a profile describing their activities, interests and values though not all these are always essential (Toma and Hancock, 2013); a friend request is then sent to other parties which must be accepted in order to establish links (Dwyer *et al.*, 2007).

Facebook *groups* allow users to connect based on common interest such as neighbourhood, school clubs, families, etc. Facebook groups are not public by default, meaning that only members can receive updates from them. A Facebook *page* is a public profile created for organisations, businesses or a community. Unlike a personal profile page, public Facebook pages do not have friends but “fans” which are people who have chosen to engage with the page. Facebook pages have functionalities that allow users to engage them, generating important data as a result that can facilitate quantifying the rate of public engagement on any post. A *post* is a message (user-generated content) in a form of a comment, picture or other media appearing on a user’s page. Facebook metrics are available as soon as the post is published to the public including; post clicks, reach, likes, comments and shares. The *clicks* can be described as “all in one” metrics that refer to the total number of clicks anywhere in a page post such as likes, comments, shares and views. The *reach* metrics refers the number of people who receive impressions (number of times a content is displayed on your News Feed) on a page post. News Feed is the constantly updating list of stories, status updates including stories, videos and links from people, pages and groups that you follow. The *like* functionality allows users to express their support about the preferred content, while the *comment* tool allows users to add voice to the content of a post, and the *share* function allows users to extend the content to others on their network (Facebook, 2017a).

In terms of access to data, unlike Twitter, Facebook does not allow automated access to data without consent, and most of the information is private by default unless otherwise set as public (Bechmann, 2014). Despite this limitation, considering its popularity, in this research Facebook

data will be used to explore social media engagement in different community areas of Leeds and the extent at which engagement on Facebook is related to demographics of areas. Figure 4.2 shows a typical Facebook profile.



Figure 4.2: Typical Facebook account profile

Developers API²

In order to obtain public Facebook data, users are required to create an app via Facebook application programme interface (API) available from the Facebook developers' website (www.developers.facebook.com) and obtain the necessary credentials such as an App ID and App Secret (Lanfear, 2016). Figure 4.3 shows a typical Facebook App that has been created. Having set up an App, users can start collecting data from public Facebook pages using page names or alternatively page IDs can be used. To ensure that data is collected from the correct pages, the name of the page can further be authenticated by checking its ID from the Facebook website³. In the R environment, a common method for collecting Facebook page data are uses an

² <https://developers.facebook.com>

³ <http://findmyfbid.com>

algorithm developed by Barbera *et al.* (2017). Typically, the Facebook metrics of engagement available for a particular Facebook page include “Likes”, “Comments” and “Shares” (Magno, 2016).



Figure 4.3: Typical Facebook API development app

4.2.2 Twitter

Twitter is a free social networking service created in 2006 that allows registered users to send short update messages called tweets over the web or on mobile devices (Vanam and Reznik, 2013). According to recent statistics, Twitter has 310 million active users with one billion tweets sent every month (Twitter, 2016). In contrast to Facebook, Twitter limits user-generated content (tweets) to 140 characters, allow users to follow whomever they wish, but those they follow do not have to follow them back (Procter *et al.*, 2013). Figure 4.4 shows a typical Twitter account profile.

Social media data are often difficult to obtain, with most of the social media sites restricting their data (Morstatter *et al.*, 2013; Sloan and Quan-Haase, 2017). However, Twitter is different in this respect as researchers can access Twitter data through two APIs, the REST API and the Streaming API (Figure 4.5), differing in the design and access method (Kumar *et al.*, 2014). This availability and ease of use makes Twitter the most popular information source for social

networks and opinion studies, and is widely used for research (O'Connor *et al.*, 2010), especially in social sciences (Tuomisto, 2010). Although studies have been carried out using other social media data, from sites such as Snapchat and YouTube, these works have not led to new insights or novel contributions to the extent that has occurred with Twitter (Aarts *et al.*, 2012). Table 4.2 shows details of the Twitter public APIs.

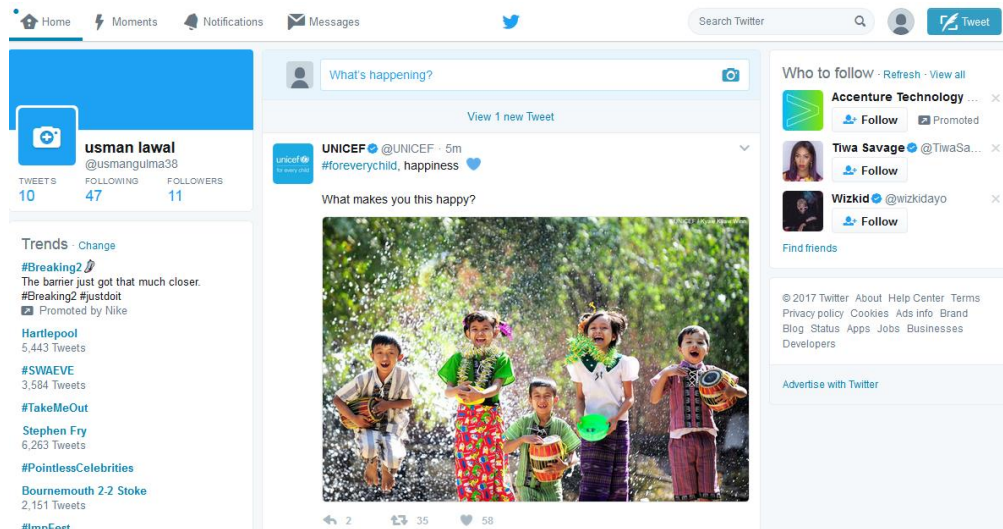


Figure 4.4: Twitter account profile

The APIs can be accessed only via an authenticated request with valid credentials (a consumer key, consumer secret, access token and access secret) provided by Twitter. Furthermore, access to the Twitter APIs is also limited to a specific number of requests within a time window, called the rate limit. These limits apply both at the individual user level and the application level (accessing tweets from multiple users). The Twitter APIs are discussed in more detail below based on Kumar *et al.* (2014). For example, Twitter restriction applies to requests within a rate limit of 15 individual user requests at 15 minutes per rate limit window (Twitter, 2017b).

Twitter offers relatively easy access to a 1% global random sample of tweet data via its Streaming API website⁴. Although users can overcome this limitation by accessing 100% of all public tweets through its “Firehose” data providers (e.g. GNIP and DataSift), the cost is prohibitive (Kumar *et al.*, 2014).

⁴ <https://dev.twitter.com/streaming/public>

Table 4.2: Twitter public APIs

API	URL	Data type	Public data Allowance	Search parameters
REST	https://dev.twitter.com/rest	Archive	1%	400 key words,
Streaming	https://dev.twitter.com/streaming	Continuous		25 bounding boxes, 5,000 user IDs

Rest API⁵

The REST API can be used to retrieve archive data which are not more than 7 days old. However, in this research, the Streaming API will be used, so we focus on this here.

The screenshot shows the 'usmangulma' developer application page. It includes navigation tabs for 'Details', 'Settings', 'Keys and Access Tokens', and 'Permissions'. The user profile shows the name 'usmangulma', ID 'lawal13101971', and a profile picture. The 'Organization' section is currently set to 'None'. The 'Application Settings' section lists various parameters: Access level (Read, write, and direct messages), Consumer Key (API Key) (n6Fw6xAgcW8DF2ie51GZij8r), Callback URL (None), Callback URL Locked (No), Sign in with Twitter (Yes), App-only authentication (https://api.twitter.com/oauth2/token), Request token URL (https://api.twitter.com/oauth/request_token), Authorize URL (https://api.twitter.com/oauth/authorize), and Access token URL (https://api.twitter.com/oauth/access_token).


Figure 4.5: Typical Twitter developments API

⁵ <https://dev.twitter.com/rest>

Streaming API⁶

The Streaming API provides access to a continuous stream of public information from Twitter. Once a request is made, the Streaming API provides a stream of updates based on the search request. The Streaming API limits the number of parameters which can be supplied in a request. Up to 400 keywords, 25 geographic bounding boxes and 5,000 user IDs can be provided in one request. This returns up to 1% of the total current volume of tweets published by Twitter that match the request at a given time. This research will therefore make use of Twitter's free 1% random sample data. Table 4.3 describes the metrics available from social media data. The Streaming API has three end points: public streams, user streams and site streams. Public streams contain public tweets on Twitter, while user streams provide streams from a single user; the site streams access tweets from multiple users. However, the public streams API is the most versatile streaming API (Kumar *et al.*, 2014). Therefore, in this research, the public stream API will be used for data scraping.

Table 4.3: Social media metrics

Facebook 	
Metrics	Description
Likes	Measures of popularity of a Facebook page posts. The function allows users to create and give feedback about a preferred content.
Comments	Measures of commitment on a particular post. It allows users to add their voice to the content of a post.
Shares	Measures the number of times a post is shared by other people to their own network. It is a measure of virality received by a particular content.
Clicks	Measures the number of times a particular content (likes, comments, shares, and video) is clicked.
Reach	Measures the number of users who saw a particular post on their newsfeed. It does not necessarily mean interacting with content.

⁶ <https://dev.twitter.com/streaming>

Metrics	Description
Tweets	Measures the number of times user generated content are posted on Twitter.
Retweets	Measures the number of times a content is shared by other users
Favourites/likes	Represents the number of most popular tweets. They are subjective, usually measures a positive sentiment.
Followers	Measures the number of people that subscribed to a particular updates from a particular user.
Mention	Measures the number of times a particular user name is mentioned in a content
Replies	Measures the number of times users respond directly to content.

Source: Jackson (2014)

4.3 Social Media Community Cohesion and Crime: Action in Communities

The flow of information about local community wellbeing is essential for neighbourliness, social cohesion and collective efficacy (willingness to intervene for the common good). Sampson *et al.* (1997) stressed that collective community action is linked to lower levels of crime (especially violent crime). One such way to achieve this process is through social media networks (Harris and Flouch, 2010). In many communities, social media platforms have emerged to play a central role in local community engagement (Harris and Flouch, 2010). Social media is changing the way we interact as individuals as well as a community, allowing people to organise and take action (Kawash, 2014). For example, in Brockley in south east London, the community members are using social media (Twitter specifically) for information sharing about local community events such as meetings, requests for help as well as for interactions between local residents, establishing ties with one another (Bingham-Hall and Law, 2015). In this way, community web resources can serve as a means of communication, supporting one another and improving local community participation.

Social cohesion literature emphasises the mobilisation of community support for solving local problems such as crime. Until recently, emphasis has been on collective efficacy (informal social control) Sampson *et al.* (1997) and partnerships between the police, local authority and the community (such as Neighbourhood Watch) (Hope, 2001) as provided by the Crime and Disorder Act of 1998. The Neighbourhood Watch program is a self-help community crime prevention initiative aimed at reducing crime and opportunities for crime. However, the recent social media revolution has enabled this partnership to be facilitated across the web. Laney (2013) argued that social media platforms are gradually changing the way people interact among themselves and even open more opportunities for people to express their views on government activities. Furthermore, online social media sources like Twitter and Facebook stimulate face-to-face connections which lead to other neighbourhood actions (Hugh and Kevin, 2011). Social science researchers have been exploring geo-located social media data (especially Twitter) to examine social interactions of people and to attempt an explanation on social dynamics within neighbourhood boundaries (Prasetyo *et al.*, 2013; Wakamiya *et al.*, 2013). The contention is that the use of social media data might offer a better opportunity to gain new insights into the relationship between community cohesion and crime.

Local authorities have also recognised the importance of social media for communication. For example, in the UK, a recent survey of social media usage by local councils indicates that 97% of the councils were using Twitter, 95% were using Facebook and 83% were using YouTube to communicate with the local community (James, 2017). Leeds City Council is using social media channels such as Twitter for community engagement. For example, the Twitter account (@YourCommunity) with 1,495 followers as at June, 2016, is one of the accounts used for communicating with different communities within the district so as to encourage local community participation, as well seeking their opinions on different issues affecting the city. However, because the adoption of social media as a means of communication by these organisations is a recent event, it is difficult to quantify its impacts.

One example of social media usage for community engagement is in Brisbane in Australia, where the city council has an official Twitter account (@brisbanecityqld). The city council is using social media to communicate with communities as well as responding to their enquiries by providing accurate information, building trust and encouraging community participation

(Howard, 2012). For example, during the Brisbane flood crises in 2011, the city council used its social media account to update people with accurate information on the situation.

On the other hand, West Yorkshire Police are also using social media channels like Twitter (e.g. @WYP_LeedsCity) to communicate with different neighbourhoods about safety issues. Through these channels, the police engage different neighbourhoods seeking help and encouraging local community participation, especially in crime prevention activities. Local community problems especially relating to neighbourhood safety, can be discussed directly (though publicly) with the relevant agencies for prompt action. Although using social media as a new form of community engagement is just beginning to develop, people are gradually starting to use them as media for discussing local activities as well as neighbourhood problems (such as crime), as well as responding to police appeals for assistance. For example, recently (21 July, 2016) the police posted an appeal for help about a missing teenager (Kerry Lund), and many people responded promptly by sharing the information across different neighbourhoods. As a result of this collective action the boy was found within a short time.

While organisations are using social media tools for communicating with communities, the communities themselves are also making use of social media channels to foster social cohesion. For example, Brockley in South London has a local community Twitter account (@BrockleyCentral) with 7,240 followers, a significant figure relative to a population of 17,000 (Bingham-Hall and Law, 2015). Using this channel, promotional events and requests for help are posted, allowing local residents to interact and establish social ties with one another (Bingham-Hall and Law, 2015).

4.4 Social Media Community Cohesion and Crime: Understanding Communities

Researchers have used social media data (Twitter in this context) to study crime in a geographical space (Malleon and Andresen, 2015; Corso, 2015; Luini *et al.*, 2015), disasters (Tapia *et al.*, 2013; Simon *et al.*, 2015) and to explore the social interactions of people (Huberman *et al.*, 2008; Macskassy, 2012; Wakamiya *et al.*, 2013). Additionally, social media tools have been employed to study humanitarian crises (Goldfine, 2011; Skuse and Brimacombe, 2014; Simon *et al.*, 2015). For example, Cooley and Jones (2013) attempted to employ Twitter data to study the humanitarian crisis resulting from the famine in Somalia using content analysis. They found that Twitter is increasingly used by aid agencies in relief efforts. The major

limitation of their study is the small sample size, limited in this case by the low levels of internet availability and accessibility. Internet failure is commonly experienced especially in developing countries, during crisis periods, making it difficult to employ such tools for communication (Mullaney, 2012). Small samples are difficult to use quantitatively, making interpretation of results difficult (Button *et al.*, 2013).

While studies using social media in humanitarian crises are common (Goldfine, 2011; Skuse and Brimacombe, 2014; Simon *et al.*, 2015), the application of social media to the study of social cohesion (especially crime) is limited (Aarts *et al.*, 2012). This research seeks to address this limitation through the application of social media in order to gain new insight into the relationship between community cohesion and crime. In this research Twitter data will be employed to analyse public sentiments with a view to establishing the link between social cohesion and crime.

The geolocation of tweets is important for understanding the role of geography in explaining or predicting different themes of interest. Approximately only 1% percent of tweets published by Twitter are geo-located (Kumar *et al.*, 2014) because only about 1.6% of Twitter users have global positioning system (GPS) functionality enabled on their device, chiefly for privacy reasons (Tuomisto, 2010). Swier *et al.* (2015) stressed that although the small proportion of geo-located tweets might not provide useful insights into the routine activity of people, a 1% sample can still be sizable compared to traditional surveys.

In terms of the use of Twitter data in the study of crime, Gerber (2014) used geotagged Twitter data and kernel density estimation to quantify and predict crime in Chicago, USA. He compared historical crime records with Twitter data within geographical areas of interest using hotspot maps and topic modelling. He found that Twitter data improves crime prediction, potentially enabling resource allocation, especially police personnel. The main limitation of his analysis was that he focussed on topic modelling of tweets and failed to address the issue of content analysis (contextual interpretation of document) of tweets. Sentiment analysis of tweets content, as in the present research, might provide better insights for planning, especially for crime reduction activities. The present research would address this problem by analysing the content of social media data (tweets) in order to gain more insights into the relationship between public emotion and demographic characteristics of different areas, and how they relate to crime.

In another study, Chaudhry (2015) employed geotagged Twitter data to examine hate crime using content analysis in six Canadian cities (Vancouver, Toronto, Winnipeg, Calgary, Edmond and Montreal). He argued that social media such as Twitter provides an opportunity to openly discuss issues relating to race discrimination in Canada which might be difficult in an offline situation. Therefore the contents of tweets (text) provide a rich data source for researchers to mine. He found that Twitter data can be used to explore racially motivated crime. The major limitation of his study is small sample size (he collected only 776 tweets). As mentioned previously, interpretation and generalisation of results based on small sample data can be difficult (Hackshaw, 2008; Button *et al.*, 2013).

Previous studies have concentrated on using geolocation of tweets published by Twitter to predict the occurrence of crime (Featherstone, 2013; Malleson and Andresen, 2015; Corso, 2015; Luini *et al.*, 2015). However, they have not analysed the textual content of tweets. Analysing the content of tweets might allow for better estimation of crime patterns (Williams *et al.*, 2016). For example, Sandhana and Sangareddy (2015) used Twitter data combined with climate data in the Karnataka district of India to quantify and predict the temporal variation in crime using sentiment analysis (opinion mining). They found a correlation between Twitter sentiment and seasonal changes in the occurrence of violent crime. The major shortcoming of their study is the absence of sentiment classification of text (tweets) content in terms of polarity (e.g. negative vs positive terms as in Jong, 2011; Rajan and Victor, 2014) or lexical or affinity classification (sad, happy, afraid, bored as in Cambria, 2013).

Most recently, Williams *et al.* (2016) used Twitter data, police recorded crime data and UK 2011 Census data to study crime and social disorder in London boroughs using linear random and fixed effects regression models. They argued that the analysis of tweet content might provide a better insight into crime patterns than simple geo-located data, which has limited explanatory power beyond routine activities. They also found that social disorder related tweets specifically correlate with actual police crime data. The major issue with their study is the use of randomised regression models, which, although useful for analysing large data sets (Clark and Linzer, 2015), are also prone to severe bias in estimating the effects of variables that are included (Allison, 2005). Instead, opinion mining (sentiment analysis) of tweet contents might provide more useful insights for exploring the dynamics of human collective behaviours (Gao *et al.*, 2013). Table 4.4 provides a summary of previous studies.

Table 4.4: Summary of studies using social media (Twitter) for social interactions and crime

Authors	Data used	Methods	Pros	Cons
Gerber (2014)	Geotagged Twitter data combined with historical crime data	Kernel density estimation (KDE) and Topic Modelling	<ul style="list-style-type: none"> • Twitter data improves crime prediction. • Can enhance planning for resource allocation. 	<ul style="list-style-type: none"> • Only focussed on topic modelling of tweets. • Failed to address issue of content of tweets. • Sentiment analysis might provide better insights especially for planning crime reduction activities.
Malleson and Andresen (2015)	Geotagged Twitter data	Geographical analysis techniques	<ul style="list-style-type: none"> • Describe shift in hotspot of crime from the city centre compared with police recorded crime data. • Hotspot maps are useful in describing patterns. 	<ul style="list-style-type: none"> • Hotspot maps might not provide insights for the factors responsible for the occurrence of crime clusters. • Factors that give rise to crime hotspots differ from place to place (Eck <i>et al.</i>, 2005)
Chaudhry (2015)	Geotagged Twitter data	Content analysis	<ul style="list-style-type: none"> • Twitter provides opportunity for open discussions relating to race discrimination. • Twitter data can be used to explore racially recorded crime. 	<ul style="list-style-type: none"> • Small sample size. • Interpretation and generalisation of results based on small sample can be difficult (Hackshaw, 2008; Button <i>et al.</i>, 2013)

Authors	Data used	Methods	Pros	Cons
Sandhana and Sangareddy (2015)	Twitter data combined with climate data	Sentiment analysis	<ul style="list-style-type: none"> • Describes correlation between tweets content, temporal changes and occurrence of violent crime. 	<ul style="list-style-type: none"> • Disregard sentiment classification of tweets e.g. (Cambria, 2013). • Archive weather data might not be valid beyond 2 weeks, making quantitative predictions unreliable (Epstein, 1988; Higgins, 2015).
Williams <i>et al.</i> (2016)	Twitter data police recorded crime data UK Census (2011)	Linear random and fixed efforts regression models	<ul style="list-style-type: none"> • Describes correlation between social disorder tweets and actual police crime data. • Twitter data might be use as alternative information source planning crime reduction activities. • Randomised regression models are useful for analysing large data set (Clark and Linzer, 2015). 	<ul style="list-style-type: none"> • Randomised models are prone to severe bias in estimating efforts of variables (Allison, 2005).

As can be seen from the literature reviewed, there is a growing body of research that leverages social media to study human social interactions and community cohesion, especially with regards to crime. Building on these previous studies, this research will also employ social media data (Facebook and Twitter in this context) in order to explore the relationship between community cohesion and crime.

4.5 Social Media and Law Enforcement

Social media tools such as Facebook and Twitter have been found to be useful in disseminating information about crime safety awareness and as a medium of interaction between the police and the public (Heverin and Zach, 2010; Vanam and Reznik, 2013; Trottier, 2015). The police use social media as an outreach platform to inform the general public about crime and safety. Figure 4.6 shows social media accounts used by West Yorkshire police. The number of people subscribing to these media is low but increasing across neighbourhoods. Furthermore, social media platforms enable real-time communication and reporting of events, and this information flow is relevant to law enforcement (Corso, 2015). For example, any member of a community with such facilities can use social media to send critical information about a crime or potential crime for timely action (Corso, 2015), albeit in a very public manner. Information collected from social media can be used to help identify criminals or criminal groups, and potentially to discover patterns and anticipate crime, as well as for resource allocation to combat crime (Hartle *et al.*, 2014; Trottier, 2015). For example, social media analyses (text mining and social network elucidations) are important in revealing patterns about public opinions and activities on social media which can help in planning police operations (Williams *et al.*, 2013).



Figure 4.6: West Yorkshire Police social media channels

However, while social media sites have become useful tools for the public and law enforcement agencies, criminals are also using those sites to coordinate criminal activity (USDOJ, 2013). Despite its potential for law enforcement and crime control, researchers have paid little attention to this area (Prichard *et al.*, 2015).

One particular example of social media (especially Twitter) engagement between the police and public is the UK 2011 riots. During the period of the riots, while the police saw a tremendous increase in followers, and they also engaged the public on a large scale (Denef *et al.*, 2013; Procter *et al.*, 2013). It has become important for the police to seek the cooperation of the public in order to successfully discharge their responsibility of maintaining peace and reducing crime in communities (Denef *et al.*, 2013).

4.6 Social Media, Social Capital and Crime

The evolution of technological developments such as the SM channels are changing the way social capital is perceived, especially relating to community life-styles and safety. There is a growing evidence suggesting that social network sites (e.g. Facebook) have the potential to generate social capital (see Section 2.4) (Gainous and Wagner, 2014). SM networks are

potentially useful for the generation and maintenance of social capital, facilitating public interactions and promoting face-to-face social bonds (Ellison *et al.*, 2007) which is essential for maintaining neighbourliness (Sampson *et al.*, 1997). Social capital reflects the relationship between people and the outcomes of their relationships (Coleman, 1990; Putnam, 1995). Social capital has been defined as the sum of resources, actual or virtual, that an individual or group benefits from belonging to a social network (Bourdieu and Wacquant, 1992). Social capital exists if people share resources and exchange information (Sander and Teh, 2014). Based on their study, Ellison *et al.* (2007) emphasised that when social capital declines, a community experiences increased social disorder, reduced participation in civic activities and more distrust among community members. Babb (2005) argued that face-to-face social capital is now being maintained through the Internet especially through the SM such as Facebook. In contrast to the advent of SM networks, social connections and communications were only maintained at the local level, or through telephone and postal services (Matthews, 2016). For example, Phulari *et al.* (2010) stressed that bridging and bonding social capital is built when communities of like-minded people increasingly converge online. The concept of bonding social capital takes place between socially homogeneous groups while bridging exists between socially heterogeneous groups (Putnam, 2000). However, researchers have stressed that social capital is difficult to quantify because it is a multidimensional perspective and no single indicator exists to measure it (see Section 2.5) (Haezwindt, 2003; Plotkowiak, 2014; Appel *et al.*, 2014). Previous studies that attempted to quantify social capital relied heavily on traditional data sources (e.g. surveys) which are limited by cost, sampling and questions can easily be manipulated and prone to bias, hence the reliability of those measures may be questioned (Deller and Deller, 2010; Szolnoki and Hoffmann, 2013). For example, the variables of trust, reciprocity and civic participation obtained from surveys were used as indicators of social capital in previous studies (Kawachi and Berkman, 2000; Healy, 2002).

Nevertheless, the advent of social network sites enabled the establishment and maintenance of social capital online. In recognition of the importance of social network sites for generation of social capital in the UK, Office for National Statistics has now recognised belonging to a social network site as one of the important variables for measuring social capital in a community (Veronique, 2014). An important component of belonging to a social media network site for social capital is that people can use the networks for accessing information or conveying social

support for the members of the community; and the more people are connected to the network the more social capital is increased (Ellison *et al.*, 2014). Social support can be in the form of goodwill messages like birthday wishes, encouragement, commiserating with the bereaved, advice and suggestions on how to deal with local community problems and an invitation for a community meeting. These are described as online reciprocity (Sander and Teh, 2014). Similarly, the exchange of information and resources requires trust; on social networks people do not discuss or share content with other people if they do not trust each other (Sander and Teh, 2014). Reciprocity and trust are essential components of social capital that facilitate interaction in the community (Vilares *et al.*, 2011) and Facebook specifically provides a good platform for enhancing those elements of social capital (Valenzuela *et al.*, 2009). Although Facebook as a medium for social capital depends on how users interacted with the content by liking, commenting and sharing; contents without interactions indicate lack of interest by the network (Ellison *et al.*, 2014). Putnam (2000) argued that social capital is often better understood using quantitative approaches and when the context in which it is being assessed is determined. In this research, Facebook is used for a number of reasons. Facebook is the most widely used SM network across all ages in the UK (Statistica.com, 2017), widely used in social science for understanding patterns of online social relationships (Wilson *et al.*, 2012; Kosinski *et al.*, 2015). Blank (2017) argued that the unlike Facebook, the unrepresentative characteristic of Twitter users suggests that Twitter data are not suitable for social science research. Facebook also enables creation and maintenance of social interaction quickly, enhance information dissemination with a wider number of users and enabling interactive feedbacks at low transaction cost (Ellison *et al.*, 2014). Facebook offers a guided interactions of messages via liking, commenting and sharing either of these actions strengthens social relations (Naseri, 2017), allowing people to create pages and groups for the purpose of social relationships (making new friends and maintaining existing relations with families and neighbours) (Ramadan, 2017). This active engagement is greater on Facebook than other SM platforms such as Instagram and Twitter (McClain, 2017). Although, other SM networks like Twitter also provide opportunity for social interaction by merely publishing text and links to stored media, Facebook offers full functionality for posting different multimedia messages (videos and photos) which makes it more versatile for local social interaction (Kwon *et al.*, 2014).

Furthermore, the exchange of information about local community wellbeing can also facilitate collective efficacy (a willingness to intervene for the common good). Sampson *et al.* (1997) stressed that collective community action is linked to lower levels of crime (especially violent crime). One such way to achieve this process is through SM networks (Harris and Flouch, 2010). SM channels also allow people to organise and make collective decisions (Kawash, 2014), share information on neighbourhood safety and allow residents to request for help when in need (Bingham-Hall and Law, 2015). Facebook specifically has been found to be an important channel for information sharing relating to community building and creating awareness neighbourhood crime (Hattingh, 2015; Sachdeva and Kumaraguru, 2015).

4.7 Methods

Previous sections have highlighted the importance of social media, especially Facebook and Twitter, as a media for social interaction and collaboration between the law enforcement and the public. As discussed, social media can have a useful impact on law enforcement, through both information dissemination and community cohesion building.

The review highlighted the limitations of previous studies, as well as loss of opportunities for more sophisticated use of text content of geo-located tweets. The next sections concentrate on specific techniques of analysis that have been used and may provide insights into social media data.

4.7.1 Sentiment

Sentiment is an expression of opinion or views about a particular event or situation as either positive or negative. For example, “good”, “wonderful” and “amazing” are positive sentiment words; “bad”, “poor” and “terrible” are negative sentiment words (Liu, 2012). Sentiment can be assessed based on polarity (e.g neutral, negative or positive), as well as on the basis of its strength positively or negatively (eg +1 to +5 or -1 to -5) (Thelwall *et al.*, 2012). Until recently, research on sentiment focussed on discussions linked with stock market reviews by investors (Bormann, 2013). However, with the growing interest in social media text mining, studies relating to sentiment detection using public posts for crime prediction are increasing (Wang *et al.*, 2012; Bolla, 2014; Jurek *et al.*, 2015). This shift provides new challenges as well as

opportunities for social scientists wanting to better understand the relationships between public social media interactions and their impacts on social behaviour especially with regards to crime (Felt, 2016). The opportunities arise as social scientists can now use social media channels to seek out and find public opinion about a particular event, which can enhance effective planning and decision making (Pang and Lee, 2008). However, the challenges relate to techniques used to accurately analyse large amounts of data being generated through social media in order to better make sense out of them (Liu, 2010b).

4.7.2 Sentiment Analysis

Sentiment analysis (SA) or opinion mining is a technique that uses Natural Language Processing (NLP) to extract mood from public messages posted on social media about particular issues, topics and events (Liu, 2012; Medhat *et al.*, 2014; Prichard *et al.*, 2015; Wilson *et al.*, 2016). Sentiment analysis is concerned with automatic extraction of sentiment related information from text (Thelwall *et al.*, 2012). It may also involve clustering techniques for the visualisation of sentiment clusters within a corpus (collection of text document) (Layton *et al.*, 2013). Social media, especially Twitter, provides rich textual data about public opinions, thoughts and behaviours. Twitter is particularly preferred because tweets are primarily public, and therefore broadly available for analysis, unlike Facebook posts that are restricted to friends (Deitrick *et al.*, 2013). Extracting public opinion from social media texts provides a challenging but yet a rich context for exploring computational models of natural language (O'Connor *et al.*, 2010). This analysis necessitates automated methods for the summary of opinions expressed in text documents such as social media, and SA emerged in response to this need (Liu, 2010a). The ability to extract, analyse and gain new insights from social media data is a practice that is widely adopted especially in social science (Smith *et al.*, 2009; Liu, 2012; Xiaojun *et al.*, 2015).

Furthermore, Tonidandel *et al.* (2015) has emphasised that SA is becoming an integral part of social listening completed by different organisations to understand public opinion on new products, as well as issues relating to the social well-being of groups. Nevertheless, despite its popularity, there is still plenty of potential for the technique. Bolla (2014) has stressed that SA of social media data, especially Twitter, might provide useful insights into crime for particular areas; for example, areas having more negative sentiments expressed on Twitter about a particular topic tend to correlate with the occurrence of high crimes, whereas areas having more

positive sentiments might experience low crimes (Bolla, 2014). Consequently, the SA analysis of social media content is potentially important especially for law enforcement agencies (Jurek *et al.*, 2015).

Recently, there has been growing interest in mining social media data such as Twitter in order to detect public sentiments (Singh and Kaur, 2015). Opinion mining is one way that ‘big data’ generated through communication tools such as social media can be utilised (Prichard *et al.*, 2015). However, analysing human language is a difficult process owing to various grammatical nuances, cultural variation, slang and misspellings common in social media posts (Tan *et al.*, 2014). NLP as a subject detection system for individual messages produces results that might not be without errors (O'Connor *et al.*, 2010). On the other hand, when SA is applied on aggregated data the margin of error is significantly reduced, hence increasing the reliability of results for decision making (Tan *et al.*, 2014). Additionally, Prichard *et al.* (2015) argue that as social media usage is heavily dominated by younger people, it is likely that their views might be over-represented such that the voices of other segments of the population might not be heard. This notwithstanding, results obtained from the SA of large amounts of social media data are extremely useful and unlikely to be as biased as suspected by Prichard and others (Mostafa, 2013). Arribas-Bel (2014) emphasised that the sample captured represents the population of interest who wish to volunteer information for quantitative analysis, hence dispelling the possibility of bias. Figure 4.7 shows a summary of sentiment classification techniques.

Despite its associated challenges, SA provides a promising method for measuring public opinions (Prichard *et al.*, 2015). SA converts unstructured text (tweets) into meaningful information so that the interpretation of text is simpler (Alessia *et al.*, 2015), making it a useful basis for decision making regarding marketing research as well as crime prediction and monitoring (Mostafa, 2013). Furthermore, unlike traditional methods for studying public opinion that require the recruitment of participants with an associated cost (Mathers *et al.*, 1998), bias and measurement errors (Roberts, 2007), SA allows the analysis of public opinions expressed on social media across geographical areas, which mitigates some of the spatial biases at least (Prichard *et al.*, 2015).

Given these advantages, the present research will employ sentiment analysis on social media data (Twitter specifically) to better understand the relationships between social cohesion and crime as expressed in public posts on social media.

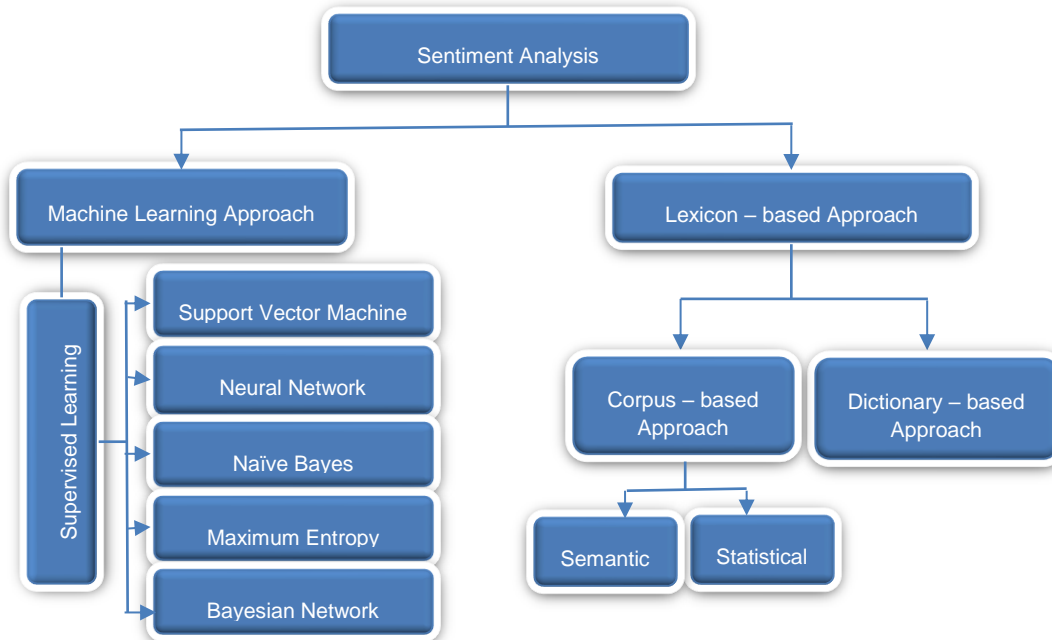


Figure 4.7: Sentiment classification techniques. Source (Medhat *et al.*, 2014)

Sentiment Classification Techniques

Sentiment classification algorithms are used for the classification of text materials as containing either negative sentiments or positive sentiments (Yazidi *et al.*, 2015). The process of sentiment classification can be at document, sentence or entity levels (Medhat *et al.*, 2014). Document-level SA aims to classify opinions in a document expressed as positive, negative or neutral. Sentence level SA aims to classify sentiment expressed in each sentence. A number of methods are used for measuring sentiments including lexicon based approaches and supervised machine learning methods (Gonçalves *et al.*, 2013; Medhat *et al.*, 2014).

Machine Learning Approach

Machine learning methods are based on supervised classification, which often relies on labelled data to train classifiers. There is a wide variety of machine learning algorithms, including support

vector machines, neural networks, maximum entropy methods and the Naïve Bayes algorithm. For detailed discussions on machine learning algorithms refer to Medhat *et al.* (2014). Naïve Bayes is an especially common framework used for the classification of textual data specifically (Go *et al.*, 2009).

Naïve Bayes Algorithm

The Naïve Bayes algorithm gives a sentiment score based on the number of times a positive or negative word occurs in a given text document (tweets in this case) (Ravindran and Garg, 2015). The algorithm allows the capture of uncertainty about a model by determining probabilities for the outcomes from the text document (Medhat *et al.*, 2014). It is also robust to noise input common in social media data such as Twitter (Hemalatha *et al.*, 2013). The main drawback of machine learning methods is their conditional reliance on labelled data; hence it is difficult to apply them to new data. Manual labelling of large datasets such as social media could be a daunting task or even prohibitive in terms of cost (Gonçalves *et al.*, 2013). Instead, researchers tend to employ a lexicon based approach.

Lexicon Based Approach

The lexicon based approaches (unsupervised) make use of a predefined list of words, each word being associated with a specific sentiment such as positive or negative (Taboada *et al.*, 2011). With this approach, a dictionary of positive and negative words is required, each with a positive or negative sentiment value assigned to it (Jurek *et al.*, 2015). For example, Hu and Liu (2004) and Liu (2010a) proposed a collection of 6,800 English words meant for sentiment orientation in textual data (positive, negative or neutral). These methods do not rely on labelled data (Dang *et al.*, 2010), so can be employed to analyse text at the document, sentence or entity levels without training (Zhang *et al.*, 2011). Additionally, lexicon methods are cost-effective (Korhonen, 2009); they can predict numerous words with high accuracy (Duygulu *et al.*, 2002). However, analysing human language expressed in social media posts (such as tweets) is a challenging process, owing to various grammatical nuances, cultural variation, sarcasm, slang and misspellings that are common in Twitter data (Tan *et al.*, 2014).

Despite its flaws, the lexicon approach is widely preferred for analysing social media content (Khan *et al.*, 2015). Researchers are increasingly using lexicon methods for SA of social media content (Berker, 2011; Palanisamy *et al.*, 2013; Jurek *et al.*, 2015; Asghar *et al.*, 2016). Building

on previous studies, the present research will also use a lexicon-based approach for sentiment analysis. There are two types of techniques used to classify sentiment orientation based on the lexicon approach: dictionary based and corpus-based techniques (Dang *et al.*, 2010).

Dictionary Based Technique

The dictionary-based technique uses lexical resources (e.g WordNet, Sentistrength, SentiWordNet) to determine the sentiment in words (Dang *et al.*, 2010). In this technique, the classifier iteratively adds newly found words to the seed list (collection of positive or negative words) and the process stops when no new words are found (Medhat *et al.*, 2014) The major limitation of the dictionary based approach is the inability to find opinion words with domain and context specific orientations (Medhat *et al.*, 2014)

Corpus Based Technique

The corpus-based technique aims to find the co-occurrence of patterns of words to determine their sentiments. For example, Breen's (2011) SA algorithm estimates sentiment by assigning an integer score (positive or negative) by subtracting the number of co-occurrences of negative words from that of positive words. However, if sentiment scores are the same for both sentiment classes or no matches were found from the sentiment word list, the text is classified as neutral (Wan and Gao, 2015).

4.7.3 Social Cohesion and Public Sentiment on Social Media

Social interactions and social networks, as well as the outcome of such connections are now easier to capture through social media. Social media provides an enabling platform for the formation of social bonds (cohesion) among people as well as the exchange of information and creation of new social networks among communities (Cornelius *et al.*, 2009; Gorman, 2013; Wakamiya *et al.*, 2013; Angus *et al.*, 2015; Mukaka, 2012). Laney (2013) argued that social media platforms are gradually changing the way people interact with one another and offering more opportunities for people to express their views, for example on government or local community activities or seeking for the like-minded.

While social media sites such as Facebook and Twitter reflect a number of social trends relating to communities, they can also highlight impacts of a diversity in social relationships between different groups of people; especially as these new forms of communication are increasingly

adopted by a number of culturally diverse users (Adolf and Deicke, 2014). The term diversity describes the level of variation in racial or ethnic composition, age, gender, religion, philosophy, physical abilities, socio-economic background and sexual orientation among a group (Goodin, 2014 p.7). Such diversity may hinder informal communication within neighbourhoods and affect the establishment of social interaction across groups (Browning *et al.*, 2008; Letki, 2008; Laurence, 2011). Younger people, for example, are less likely to build social cohesion (especially face-to-face) than older people (Johnston and Matthews, 2004; Takagi and Kawachi, 2014); identity plays an important role in the likelihood that people will connect and form social relationships (Gilchrist and Kyprianou, 2011); and social interactions are built over time and tend to be more solid when people reside in a particular neighbourhood and are influenced by length of residence (Yamamura, 2011; Keene *et al.*, 2013).

Hollander *et al.* (2016) stressed that social media data (Twitter in this context) provide rich textual information about public opinions, thoughts and behaviours and can be used for understanding social relationships in local communities. Although the concept of community is complex and multi-dimensional, the traditional view of a community is perceived to be a set of people with similar social relationships living within a geographic boundary (Gruzd *et al.*, 2011). However, the rise in social media, especially Facebook and Twitter in particular, has enabled the formation of social ties and building a sense of community beyond geographic boundaries (Gruzd *et al.*, 2011). Furthermore, previous studies have highlighted the relationships between the sense of community and social media usage (e.g Java *et al.*, 2007; Crowe, 2010; Poblete *et al.*, 2011; Scheepers *et al.*, 2014). The sense of community refers to an individual feeling of belonging, emotional connection, interaction and identification with groups of people (McMillan and Chavis, 1986).

4.7.4 Twitter Sentiment and Community Structural Characteristics

The basis for Twitter sentiment analysis is for classifying tweets based on their orientation as positive, negative and neutral (Syaifudin and Puspitasari, 2017; Ponnuru *et al.*, 2017). Sentiment analysis on Twitter data can help to uncover public opinions relating to topics of interest (Lansley and Longley, 2016; Wang *et al.*, 2017). Sentiment analysis is concerned with the automatic extraction of sentiment related information from text (Thelwall *et al.*, 2012); it may also involve clustering techniques for visualisation of sentiment clusters within a corpus

(collection of text documents) (Layton *et al.*, 2013). Wang *et al.* (2014) emphasised that the automated text analysis of social media (such as sentiment analysis) can be used to estimate community mood. This process can facilitate calls for collective community action especially relating to social behaviour in the community (Liu, 2012; Harris and McCabe, 2017).

In terms of the influence of different community structural characteristics on Twitter sentiment, Quercia *et al.* (2012) stressed that relationships exist between sentiment expressed on tweets and socio-economic well-being of communities. This factor was also found in previous studies, that differences in socio-economic and demographic factors of areas are related to people's emotions expressed on social media (Turney and Harknett, 2010; Moorhead *et al.*, 2013). However, the propensity of people to use social media (Twitter in this case) also varies by socio-demographic factors of different community areas (Williams *et al.*, 2016). Recent studies have found relationships between geographic locations (communities) where people live and sentiments expressed on social media (Gallegos *et al.*, 2016), and this relationship can be linked to different demographic and socio-economic characteristics of users as well as place-based influence (Venerandi *et al.*, 2015; Lansley and Longley, 2016). However, while Twitter sentiment analysis has been broadly studied in the social sciences, the focus has not been generally on the diversity of different communities. The present research will address this limitation by analysing Twitter data to quantify sentiments and linking them to diversity of different communities with a view highlighting their relationships.

4.7.5 Understanding Social Cohesion through Sentiment Analysis

As noted in Section 4.7.1.1, a number of algorithms have been proposed for sentiment classifications such as lexicon and machine learning (Medhat *et al.*, 2014; Ribeiro *et al.*, 2016). However, one of the simplest and widely used sentiment analysis methods is by classifying words as positive, negative or neutral using a lexicon based approach (Taboada *et al.*, 2011). Lexicon based approaches make use of predefined lists of words, each word being associated with a specific sentiment such as positive or negative feelings (Taboada *et al.*, 2011). Although, not all opinion expressed in a message (tweet) contains a negative or positive sentiment, owing to various grammatical nuances, cultural variation, slang and misspellings common in social media posts; people might simply be talking neutrally (Tan *et al.*, 2014). To provide a better understanding of different sentiment classifications, Koppel and Schler (2006) suggested the

inclusion of a neutral class, arguing that a neutral class provides a distinction between negative and positive sentiments as well as enhancing the accuracy of sentiment classification polarity. This does not mean the results obtained from sentiment analysis of Twitter data are free from bias. In this research, Twitter data are used for sentiment analysis because Twitter provides access to large amounts of public data free of charge. Bollen *et al.* (2011) argued that analysing a large Twitter dataset can potentially provide insights with respect to modelling social emotions (Bollen *et al.*, 2011).

Previous studies that apply sentiment analysis of social media data have concentrated on quantifying moods including those of customer satisfaction about particular brands and products (Agarwal *et al.*, 2011; Bormann, 2013; Fang and Zhan, 2015); predicting political elections (Tumasjan *et al.*, 2010; Chung and Mustafaraj, 2011; Bakliwal *et al.*, 2013) and estimating moods of people during disaster and crisis periods (Nagy and Stamberger, 2012; Schulz *et al.*, 2013; Caragea *et al.*, 2014). However, studies using sentiment analysis to quantify the relationships between social media and demographics to predict the outcomes of local social issues, such as those around community cohesion, which is of particular interest to policymakers, are limited (Hollander *et al.*, 2016). To address this limitation, in this research we use social media data (Twitter) to analyse public sentiment using Leeds community areas of 2001 as a spatial framework (Section 3.5). The objective is to uncover and predict the relationships between public sentiments on Twitter and diversity of local community areas, so as to potentially gain insights with regards to community cohesion. The 106 community areas (CAs) were created by aggregating 2,439 Census output areas (OAs) 2001 comprising Leeds district so as to represent places that have some meaning to residents of the city as distinctive communities (see Stillwell and Phillips (2006) for detailed description). Although Leeds district has grown by 5.1% in terms of population size between the 2001 and 2011 Censuses (Leeds City Council, 2013), the structure of the community areas has not changed significantly. Therefore, the use of 2001 community areas in this research is considered to be appropriate.

4.7.6 Social Media Engagement

The advent of social media has opened up unprecedented new opportunities for public engagement (Lee and Kwak, 2012), enabling generation of social capital in the process (Ellison *et al.*, 2007; Gil de Zúñiga *et al.*, 2012). There is strong recognition that Twitter can be a

powerful tool to engage the public (Honey and Herring, 2009). However, Facebook is one social media platform that can enhance local community engagement and is widely used for facilitating offline participation (Park *et al.*, 2009). Facebook engagement refers to the percentage of people that interacted with a post by liking, commenting or sharing (Facebook, 2017a). Table 4.5 shows Facebook metrics for quantifying engagement.

Public quantitative data provided on Facebook pages can be used to quantify community engagement (Bonsón and Ratkai, 2013). For example, Park *et al.* (2009) argued that for members to engage with a local community Facebook page they must have strong community attachment. Community attachment refers to extent and pattern of social participation and integration into the community (McCool and Martin, 1994).

Different Facebook metrics have been used in previous studies to estimate public engagement on social media. For example, Kaigo and Okura (2016) attempted to use metrics of followers to estimate the citizen engagement on a local government Facebook page in Tsukuba city of Japan. They found the relationship between the number of followers on a Facebook page and citizen engagement. However, the limitation of their study is that they use only one Facebook page which significantly affects generalisation of their findings. Additionally, the metrics of followers can no longer be used because of criticisms surrounding how the numbers were calculated and has since been phased out by Facebook (Agostino and Arnaboldi, 2016). Alternatively, Spiliopoulou *et al.* (2014) proposed the metrics of average reach to estimate public engagement with the British Museums Facebook page. They found that the metrics of reach are useful for measuring public interaction with the British museum. The limitation of their study is the use of a single metric to quantify user engagement. Peters *et al.* (2013) argued that the use of a single metric such as *reach* can rarely describe aspects of SM engagement; therefore a collection of metrics is required in order to adequately estimate engagement. Additionally, average post reach does not necessarily mean engagement. Engagement metrics indicate interaction beyond just simple views (Bendror, 2013). In contrast to Spiliopoulou *et al.* (2014), in this research, we will suggest the use of combined publicly available Facebook metrics of interaction to quantify engagement. The hypothesis is that the suggested metrics of engagement on Facebook can potentially be used to estimate levels of social cohesion of different community areas of Leeds.

Table 4.5: Facebook page metrics used for quantifying engagement

Metrics	Engagement
Posts	
Likes (P)	E= P + C + V/ Fans
Comments (C)	
Shares (V)	

P-popularity, C-commitment, V-virality

In this research, to quantify social media engagement of a community Facebook page the following equation is used:

$$E_i = \frac{n_{pcv}}{F_i} \quad (4.1)$$

where n_{pcv} is the total number of likes, comments and shares, and F_i is number of fans following Facebook page in a community.

4.7.7 New Metrics of Community Engagement

Community engagement on social media (Facebook in this context) can potentially be quantified by a set of metrics that show popularity, commitment and virality. While popularity (P) is measured by the “likes”, commitment (C) refers to the number of “comments” and virality (V) is measured by the “shares” (Bonsón and Ratkai, 2013; Haro-de-Rosario *et al.*, 2016). Collectively these metrics are known as the PCV indicators. The extent of engagement and participation of community members on Facebook posts can potentially help to estimate the strength of social activity of a community (Agostino and Arnaboldi, 2016) and is potentially an indication of social cohesion among members (Ellison and Boyd, 2013). Despite the importance of Facebook metrics for quantifying levels of community activity and engagement, previous studies that attempted to use Facebook metrics to explore public engagements mainly concentrated on specific areas such as marketing (Coulter *et al.*, 2012; Tsimonis and Dimitriadis, 2014), local government (Bonsón *et al.*, 2015; Sivarajah *et al.*, 2015; Gandía *et al.*, 2016), predicting political elections (Kristensen *et al.*, 2017) and education (Roblyer *et al.*, 2010; Wise *et al.*, 2011; Junco, 2015). Additionally, little evidence exist on the adoption of Facebook metrics to quantify community engagement

(Agostino and Arnaboldi, 2016). This work is the first academic attempt using Facebook metrics of engagement to quantify community cohesion.

It might be thought that a new Facebook account will have less activity than a well-established one. Bendror (2013) emphasised that a large number of posts on a Facebook page does not necessarily imply engagement but really depends on how compelling a post is to the intended users. However, the proposed new PCV metrics are independent of the number of posts and length of establishment of a particular profile but how fans interact with such posts, therefore appropriate for measuring community engagement (Bonsón *et al.*, 2014). There has been some criticism of the metrics of engagement in terms how they are weighted. For example, Ernest and Bernad (2015) argued that the metrics of popularity (i.e. 'likes') are likely to be higher than those measuring commitment ('comments') and virality ('shares') because it takes less effort for the users to click the like button than by sharing a post or putting their thoughts in writing by using the comment button. Stockley *et al.* (2013) have also emphasised that the metrics of likes are the shallowest indicators of engagement because the users are merely expressing agreement on the content of a post, while comments show a greater degree of engagement because the users want their voice to be heard. Overall, the metrics of shares demonstrate the greatest degree of engagement because users are expressing a desire for others on their networks to see posts and interact with them (Stockley *et al.*, 2013). Facebook places different weights on engagement metrics giving shares the most weight and likes the least but it is unknown how the Facebook algorithm computes these weights (Kim and Yang, 2017). However, engagement on Facebook is driven by the combination of all metrics such as likes, comments and shares (Lev-On and Steinfeld, 2015). In this research, these metrics will be used based on literature.

Despite the criticisms of PCV metrics of engagement, they are widely used as indicators of the degree of interaction on a community Facebook page (Boldt *et al.*, 2016; Bhattacharya *et al.*, 2017). However, the drawback of these measures of interaction is the assumption that all the metrics contribute equally to engagement. Additionally, there is lack of an appropriate method for normalizing them in the literature.

In this research, therefore, the new metrics based on PCV are proposed to potentially quantify social cohesion and crime in different local community areas of Leeds, UK. Agostino and Arnaboldi (2016) argued that the PCV metrics can be used as a proxy for measuring public

engagement with a Facebook page posts. The advantages of the suggested new metrics data derived from Facebook are that the proposed metrics of post interaction such as likes, comments and shares are publicly available as soon as posts are published. Additionally, the new metrics might potentially provide an alternative to traditional survey data (such as the Community Life Survey of England) for estimating social cohesion in a community. Facebook is used because it is the fastest growing social media platform (Sachs *et al.*, 2011; Tess, 2013; Chaffey, 2017) and undoubtedly the most widely used social networking site used for social interaction in different communities and is also popular across all categories of users both young and old (Junco, 2012). Although, Facebook does not allow automated access to data without consent where an account is set as private by default (Bechmann, 2014). Nevertheless, community Facebook pages are usually public; therefore accessing basic data (likes, comments and shares) does not require consent (Facebook, 2017b).

4.8 Concluding Remarks

This chapter discussed “non-traditional” social media approaches for modelling social cohesion in relation to crime. Section 4.2 reviewed relationships between social media and social cohesion. It was shown that social media platforms especially Facebook and Twitter, can be used for building and maintaining social cohesion, and increasing access to the Internet makes this processes easier. Section 4.3 explained how relationships between social media and community cohesion can be used to mobilise for action in communities. Furthermore, as shown in Section 4.4, social media channels can potentially be used for understanding communities, while Section 4.5 described how law enforcement agencies are using social media channels for crime preventions. Different analyses methods of social media data such as sentiment and engagement were discussed in Section 4.6. These techniques will be used later in chapters 6 and 7 for social media modelling in relation to community cohesion and crime.

Chapter 5

Traditional Modelling of Crime and Community Cohesion

5.1 Introduction

This chapter employs 2011 Census variables identified as important indicators of crime and community cohesion (in chapter 3) along with aggregated police recorded crime data (burglary) for a 5 year period (2011- 2015), for modelling the relationships between burglary rate and neighbourhoods structural characteristics. Section 3.4 provided a review of Census variables used and police crime data were described in Section 3.4.2. In this research, burglary rate is calculated *per 1,000 population* in order to reduce the biases created by using areas with differences in population size.

The chapter begins with a review of the importance of quantifying urban crime in Section 5.2 and proceeds to describe the correlates of crime in urban communities (Section 5.3). Section 5.4 explores the impact of social cohesion and diversity on crime. The theoretical justification for the explanatory variables used is provided in Section 5.5 while Section 5.6 describes the method and data used for the analysis. The results are discussed in Section 5.7 and concluding remarks are given in Section 5.8.

5.2 The Importance of Quantifying Urban Crime

Measurement of crime is necessary for any quantitative assessment of crime policy change (Ludwig and Marshall, 2015). Knowledge about how the incidence of crime is distributed over space can also enhance the effectiveness of police operations and collective community programmes such as Neighbourhood Watch (Brunsdon *et al.*, 2007). For example, local knowledge of where crime is clustered will increase the capacity of the police to employ prevention measures, thereby improving the safety of communities (Moore and Trojanowicz, 1988; Bruce and Santos, 2011). Therefore, urban and regional planners, policy makers and policing agencies have all recognised the importance of better understanding the patterns and dynamics of crime (Murray *et al.*, 2001).

5.3 Understanding Correlates of Crime in Urban Communities

A common method of understanding crime is through correlation with socioeconomic factors in the areas where it occurs, and this is an important component of so-called environmental criminology (Meera and Jayakumar, 1995; Alves *et al.*, 2013). Socioeconomic factors such as wealth disparity, educational attainment, proportion of young people and deprivation are commonly found to correlate with crime rates in urban areas (Bandyopadhyay *et al.*, 2010). Such variables act as proxies for, or direct measures of, the underlying causes of crime in a system that relate to the offender drivers, victim lifestyles and environment-related opportunities. However, the accuracy and representativeness of variables that act as proxies vary considerably and many variables, such as metrics of multiple deprivations, are widely regarded as ‘catch-all’ variables that encompass a wide variety of underlying factors.

Despite these complexities, many of the crime theories used to predict the locations of crimes are much simpler. Researchers employ the social disorganisation theory of Shaw and McKay (1942) to explain variation in crime rate in different neighbourhoods. This theory posits that high levels of ethnic heterogeneity, residential instability and socio-economic disadvantage undermine social cohesion, which in turn, increase delinquency and crime rates (Shaw and McKay, 1942). Sampson and Groves (1989), in their empirical extension of Shaw and McKay’s theory, argued that disorganisation in a community lowers the ability of residents to work together towards problem-solving, while collective efficacy among neighbourhood residents mitigates crime rates (Sampson *et al.*, 1997). Additionally, Kristjánsson (2007) emphasised that weak networks of social ties decrease informal social control in the community, which increases deviant behaviour. Crime theories used in this research have been described in Section 2.9.

This research focuses on residential burglary for a number of reasons. Firstly, burglary is an opportunity crime that flourishes in socially disorganised and less cohesive communities (Weisburd and Piquero, 2008; Chamberlain and Boggess, 2016; Roth, 2018) and can be traumatising to victims (Wollinger, 2017). It is likely that disorganised neighbourhoods might have higher burglary crime rates because of weak social cohesion than affluent areas where strong social connectedness facilitates the ability of residents to be on the lookout for criminal behaviour (Dunaway *et al.*, 2000). Additionally, previous research has demonstrated that social capital (participation in community-oriented policing in this context) in a community is linked to

lower levels of residential burglary (Martin, 2002). Routine Activities Theory (RAT) (Cohen and Felson, 1979) is regularly used by scholars to explain the occurrence of opportunity crimes such as burglary. This is based on the premise that a crime requires the simultaneous presence of three elements: motivated offenders, suitable targets and absence of capable guardians.

Secondly, Crime Statistics for England and Wales (CSEW) have shown a reduction in theft offences from person and property (Flatley, 2015; ONS, 2016a). Despite this development, Leeds city ranks third in burglary after London and Birmingham (Yorkshire Evening Post, 2018). Additionally, recent statistics from CSEW have indicated a rise in recorded burglary offences by 438,971 (9% increase) between 2016 and 2017 (ONS, 2017a). These developments necessitate further investigation of the relationships between rates of burglary and community characteristics in Leeds.

5.3.1 Standard versus Adjusted Variables of Crime and Community Cohesion

Of particular interest in this research is the role of community cohesion in the crime system. Community cohesion generally acts to increase the safety of communities: by reducing the socio-economic drivers of crime; through maintaining oversight of those potentially moving into criminal lifestyles (Lee, 2000); by increasing the oversight of potential sites of crime; and by reporting crimes when they occur. However, cohesion is a nuanced concept (there is considerable cohesion in communities ruled by criminal gangs) and cohesion is ill-represented by standard socio-demographic variables (both middle and working class communities can experience a wide range of levels of cohesion). As such, cohesion is poorly captured in standard regression models of crime. Section 2.3 reviewed the importance of community cohesion on crime.

This research suggests that standard regression variables can be adjusted to better capture a range of loci in which social cohesion plays a part across the crime system. When these adjustments are made, these variables become more strongly predictive of crime than standard treatments, suggesting the significant role social cohesion plays in the crime system and the significant part it plays as the link between standard regression variables and crime rates. The relationship between burglary and a series of standard socioeconomic variables will be examined. It is argued that such variables, while acting to represent components of the crime system, capture the effects of social cohesion acting within those components in a weak manner. In contrast, a new set of

alternative representations of these variables, centred on diversity statistics are generated. For example, rather than looking at the percentage of a specific age group, the diversity of ages within a community is considered. These variables are included within a stepwise regression model, along with the more standard variables, to empirically investigate their worth.

5.4 Exploring the Impact of Social Cohesion and Diversity on Crime

Social cohesion is a nuanced property of communities and best examined with a particular viewpoint in mind. In the case of crime, we are interested in those elements of cohesion which reduce offender drivers towards crime and decrease the opportunities for crime, not least by increasing informal guardianship and enhancing the work of the police and other formal crime prevention organisations. Although it is largely not the focus of the present research dealing with crime prevention, we might also include the effect of cohesion on the recovery after crime; often the most important part of the crime event for victims.

In terms of crime, we know that an important role of social cohesion is social interaction between community members which increases familiarity and promotes the exchange of information (Pearson *et al.*, 2015). This can be achieved, for example, by encouraging community-based ties and participation in communal activities (Forrest and Kearns, 2001). One result of improved communication is that social norms can be more easily and accurately promulgated, while familiarity allows social control to be more easily enacted by a community (Blau and Blau, 1982). As noted by Lee (2000), children brought up in cohesive communities are less likely to be involved in offending because of higher levels of informal guardianship. In addition, informal social control in cohesive communities helps to reduce more general delinquent behaviour which could *lead* to offending (Ross *et al.*, 2011), not least through the construction of a generally weakened social coherence in an area and a decay in environmental quality (Hovel, 2014; Berkhin, 2006; Mohit *et al.*, 2016). A second result of enhanced information and familiarity is that they also increase the likelihood that offenders will be observed as unusual, and that people observing such offenders will either disturb them or engage with formal crime prevention organisations such as the police or informally through local neighbourhood voluntary crime prevention schemes such as the Neighbourhood Watch. Overall, cohesive communities are likely to have enhanced collective efficacy with regards to crime (Sampson *et al.*, 1997; Browning, 2002; Browning *et al.*, 2008), and, for example, burglary and

street crimes are directly reduced in more cohesive communities (Sampson and Groves, 1989; Wedlock, 2006; Clarke and Hope, 2012).

5.4.1 Social Capital as an Element of Social Cohesion

One potential element of cohesion is social capital. Social capital reflects the relationship between people and the outcomes of their relationships (Coleman, 1990; Putnam, 1995). Social capital is part of the daily social life of a community and a key element of the behaviour of its people (Sander and Teh, 2014). Akçomak and Ter Weel (2012) also maintain that social capital decreases the risk of victimisation and increases the probability that potential offenders might be caught or deterred, which in turn reduces the crime rate. The importance of social capital on crime has been discussed in Section 2.4.

Unfortunately, social capital is notoriously difficult to quantify, and, while social media, communications and transport data go some way to allowing for the study of social networks and their strengths, elements of social interconnection are not directly represented in traditional datasets collected at the population level by governments.

5.4.2 Diversity as an Element for Understanding Social Cohesion

Given the difficulty in measuring social capital, within a diverse urban community such as Leeds, the lack of direct measures of cohesion within government datasets, and the wide range of manners in which cohesion acts, an alternative is to examine the diversity of communities as a proxy for these relationships (or their absence). Overall then, although it is likely that social cohesion and capital are indicators of a number of components of crime and crime prevention systems, it seems clear that diversity is likely to be a higher-level indicator and a driver that is not only relatively simple to quantify but which also influences a wide variety of social elements that play out in the crime propensity in both areas of cohesion and day to day social interactions. Diversity has the advantage that it can be assessed using standard population-level government datasets, but the disadvantage is that it is only a loose proxy for (weak) social coherence. The term diversity describes the level of variation in racial or ethnic composition, age, gender, religion, philosophy, physical abilities, socioeconomic background and sexual orientation among a group (Goodin, 2014 p.7). In the context of the UK and US, socio-economic and demographic diversity has been linked to decreased social cohesion and the variation of crime in

neighbourhoods (Sampson and Groves, 1989; Bursik Jr and Grasmick, 1993). Diversity may hinder informal communication within neighbourhoods and tends to affect the establishment of social interaction across groups (Browning *et al.*, 2008; Letki, 2008; Laurence, 2011). In The Netherlands for example, Meer and Tolsma (2014) argued that heterogeneity in a community, low levels of trust and meaningful interactions tend to undermine intra-neighbourhood social cohesion. Employing the Metropolitan Police Public Attitude Survey (METPAS) of London, it has been argued that ethnically diverse communities especially with large proportion of transient populations are often characterised by distrust, low levels of social cohesion and disputes (Sturgis *et al.*, 2014) which negatively affect individual behaviours (Mellgren, 2011). Furthermore, recent research into the spatial distribution of neighbourhood crime in Japan, consistently shows that areas characterised by ethnic diversity, wealth diversity and age diversity (calculated at individual level with surveys) have high rates of crime (Takagi and Kawachi, 2014).

In a study of community integration in Berlin neighbourhoods, Gruner (2010) employs Bourdieu's concept of 'habitus' (socialised norms that guide behaviour) to explain the distribution pattern of neighbourhoods in terms of socioeconomic and demographic structure. She found that the patterns and distribution of neighbourhoods is associated with different cultural norms and unwillingness of minority groups to integrate, describing it as self-segregation. Bourdieu's theory of habitus postulates the effects of physical embodiment of cultural capital; individuals' who grow up in similar conditions develop similar habitus (Bourdieu, 1989). People with similar habitus feel attracted by and are more comfortable with each other (Bourdieu, 1989). The theory is extended to the study of social problems such as crime and perceived problems associated with migration flows in urban neighbourhoods (e.g. Shammass and Sandberg, 2015). For example, growing up in a socially disorganised and crime-ridden neighbourhood might greatly influence the behaviour of people especially the young (O'Connor, 2004), thereby facilitating delinquency (Johnston, 2016). This is especially pertinent to the local context within which this research is set. In Leeds, the Family First initiative under the Trouble Families Programme (TFP) is working in partnership with local communities, West Yorkshire Police and Jobcentre Plus, helping families (especially the less privileged) to address multiple problems such as crime and unemployment (Leeds City Council, 2016). The TFP is a national programme in England established in 2011 and funded by the Department for

Communities and Local Government (DCLG) to provide opportunities and support towards helping families with multiple problems including youth offending, antisocial behaviour, domestic abuse and joblessness (Fletcher *et al.*, 2012; Hayden and Jenkins, 2014; Bate, 2017).

The intuitive conclusion from this set of studies is that high diversity in the communities is acting to reduce social cohesion consequently increasing neighbourhood crime. There is certainly some evidence that diversity reduces social cohesion (Meer and Tolsma, 2014) and this seems especially true where diversity is found in conjunction with deprivation (Cooper and Innes, 2009). However, it is nevertheless important to note that diversity and cohesion levels are not always related in a simple manner and that cohesion has the opportunity to be affected both positively and negatively, and by more than just ethnic or economic diversity. It is also true that, overall, potential high diversity may have net positive impacts in terms of multiculturalism and the disruption of embedded cultural processes, despite negative impacts in other areas.

To better understand these relationships, this research, will model a number of socio-demographic factors and compare the impacts of standard variables representing those factors directly (for example, the number of young people) with variables representing their diversity (for example, age diversity). Table 5.1 shows the variables that are ultimately used in the model. Theoretical justification for their inclusion is provided in Section 3.4.1.1 while Census variables provided in Table 3.1 are used to investigate the possible relationship between burglary and other variables (standard and diverse).

Table 5.1: The core components of crime and community cohesion, and the variables used to represent them in the model.

Component	Standard Variable (%)	Diversity Variable
Age distribution	Young persons (Age 16-24)	Age diversity
Family structure	Lone parents	Diversity of family structure
Identity	Ethnic minority population	Ethnic diversity
Affluence / wealth	Age 16-64 economically inactive	Diversity of employment type
Educational attainment	Age 16 over no qualification	Diversity of educational attainment
Residential instability	Resident less than 2 years	Length of residence diversity

Source: ONS (2011)

5.5 Data and Method

5.5.1 Data

The crime data used for this research were obtained from the ‘police open public monthly data of reported crimes’ <https://data.police.uk>, a portal that provides for a customised crime data downloads for all police force in England and Wales. In this case West Yorkshire Police for the period 2011-2015 aggregated (N = 51,800) using *burglary rate per 1000 population* for the city of Leeds. Burglary is chosen because it is relatively well reported by the public and relatively well recorded (ONS, 2017a). The geographical neighbourhoods of analysis used in this research are the 482 lower super output areas (LSOAs) of Leeds (see Section 3.4.1). The spatial distributions of independent variables (standard and diversity) are shown in Figure 5.1 to Figure 5.6 respectively.

5.5.2 Method

In this research, a new approach is proposed to explore the impact of diversity/social cohesion on burglary crime in neighbourhoods of the Leeds district using regression techniques. Diversity statistics were compared with non-diversity (standard) variables to examine which one of them performs better when regressed against burglary crime rates. Although, as discussed (Section 5.4), diversity has a complex relationship with social cohesion, this approach has the potential to better represent the influence of social cohesion in the crime system, and to reveal those social themes where cohesion plays out most significantly against burglary crime.

Researchers have used a number of methods to measure diversity (Morris *et al.*, 2014), and it is likely that different diversity metrics will reveal different elements of community cohesion (see Section 5.3.1). Nevertheless, this initial study concentrates on diversity indices which report the probability that two individuals taken at random are different (described in Section 3.9). Such diversity indices therefore uses equivalent classes weighted on the same scale irrespective of the total community size, with each class within the community having members that share common attributes (Jost, 2006).

In this research, the categories used to measure diversity were determined by census data availability and were computed using Equation 3.7. The variables included and their categories are provided in Table 3.2. Utilising the variables in Table 5.1, a selected set of correlates with crime was constructed. Identifying the best model fit requires an iterative process that examines different combinations of explanatory variables. Exploratory regression analysis is important for selecting the best explanatory variables for a given model (Braun and Oswald, 2011). Exploratory regression builds ordinary least square (OLS) models using all possible combinations of explanatory variables and assesses which models pass the OLS checks (Rosenshein *et al.*, 2011). This process is useful for ensuring that only variables with highest significance are retained. Here, to test the strength of the relationship between the variables and crime, we utilise stepwise linear regression. Stepwise methods are commonly used to select the best variables in a regression model, especially multiple regression with many predictors such in this study (Wooldridge, 2012; Sinha *et al.*, 2015). However, the process of adding and dropping variables associated with stepwise regression has been criticised given that it is possible to miss the optimal model, removing less significant predictors increases the significance of others which may lead researchers to overstate the importance of the remaining variables (Rawlings *et al.*, 1998). Despite the limitations of stepwise multiple regression method, it is widely used in different ecological studies (Meera and Jayakumar, 1995; Raftery *et al.*, 1997; Collins *et al.*, 2007; Pitner *et al.*, 2012; Caplan *et al.*, 2015).

Optimal models are a balance of correlation against parsimony. Although such balances are largely subjective and centred around use-cases, traditionally scree graphs have been used to help in the decision making as there is often a natural kink in the graph of, for example, R-squared versus numbers of model components, which indicates considerable decreasing explanatory power being provided by additional components (Culotta, 2014; Mahmud *et al.*, 2012).

Prior to building the model, the dependent variable (burglary rate) and independent variables (standard and diverse) were correlated to investigate their relationship; and to test for possible multicollinearity (Table 5.2). Multicollinearity is present when there is a high degree of correlation among the independent variables. This can significantly affect model performance and reliability (Wang, 1996). There is no standard rule for filtering out variables based on the issue. However, we can make a judgement by checking metrics such as the variance inflation factor (VIF). VIF is used to describe how much multicollinearity exists in regression analysis

which is an indication of variable redundancy. Any variable with VIF value greater than 7.5 is considered as redundant and is usually dropped from the analysis (Charlton *et al.*, 2009). In this research, correlation between the independent variables above Pearson's r .80 is regarded as very significant. Economically inactive population correlates with percentage young persons aged 16-24 (.88); age diversity correlates with percentage of young persons aged 16-24 (.89); ethnic minority correlates with ethnic diversity (.94) and length of residence diversity (.93); and ethnic diversity correlates with length of residence diversity (.97). Therefore, the following variables were removed to avoid redundancy: length of residence diversity, ethnic minority and percentage of young persons aged 16-24. Additionally, the retained variables were found to have a stronger correlation with the dependent variable which is preferable in a linear regression model (see Table 5.2).

5.6 Theoretical Justification for the Explanatory Variables

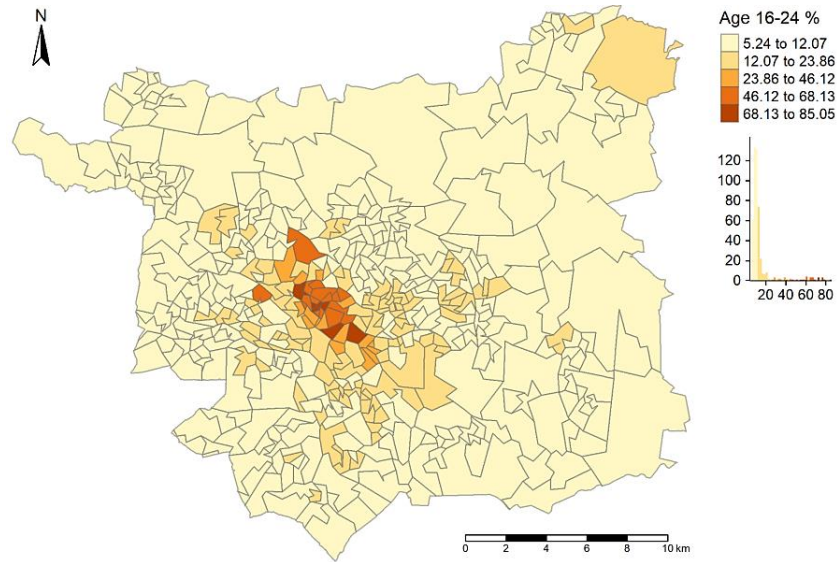
Although the relationship between community composition and burglary crime is complex and multi-faceted, there are some core factors that regularly emerge as important determinants of crime rates. Section 3.4.1 described the most common factors; it is from these that the variables used in the traditional modelling work are derived. In each case, the standard variable is represented and then the diversity variable.

5.6.1 Age Distribution

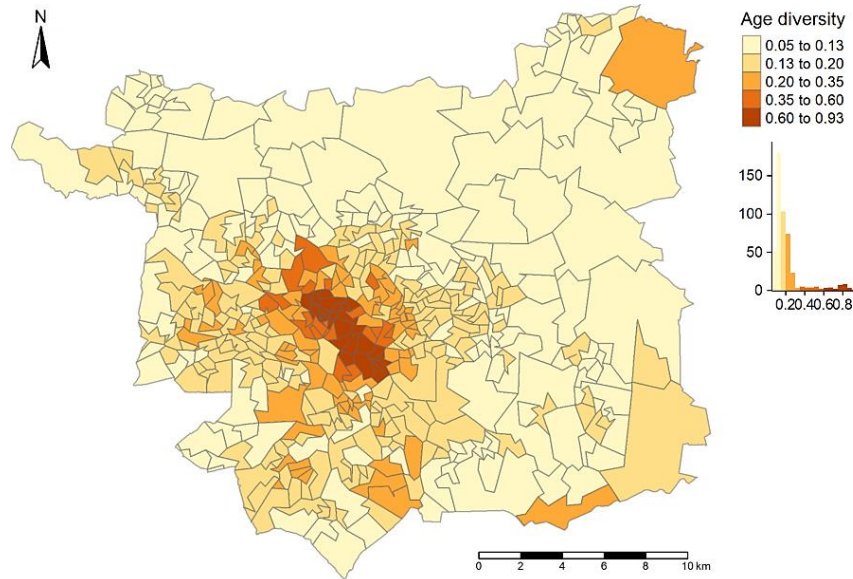
Offenders are commonly drawn from younger age groups (Kongmuang, 2006). The age-crime curve tends to increase from the adolescent years reaching a maximum at adulthood and then sharply declining (Farrington, 1986; Gottfredson and Hirschi, 1990; Sampson and Laub, 2003; McVie, 2005; Blonigen, 2010; McCall *et al.*, 2013; Sweeten *et al.*, 2013), although this varies by the type of crime (Tittle *et al.*, 2003). Burglary is therefore likely to be affected by absolute numbers of young (aged 16-24) people. For example, according to Fagan *et al.* (2014), the incidence of crimes related to vehicles and drugs tend to be higher in early adulthood than in adolescence. While homicides tend to be committed by adults, theft-related offences including burglaries are more prevalent in the younger age groups than the elderly (Palguna *et al.*, 2015).

However, crime may also be affected by age distributions. A mixed population may put more or fewer offenders near more or fewer victims, but will also affect social cohesion. Younger people

are less likely to build social cohesion (especially face-to-face) than older people (Johnston and Matthews, 2004; Takagi and Kawachi, 2014), but equally, different age distributions may have very different strengths of social control. We therefore include age diversity within population (calculated using equation 3.8 and the age components in Table 3.2), but with an acceptance that different measures of diversity may draw out very different relationships. Figure 5.1 shows age distribution and standard and diversity variables in Leeds.



(a) Young persons aged 16-24 (percentage), LSOA



(b) Age diversity by LSOA

Figure 5.1: Age distribution variables (standard (a) and diversity (b)) in Leeds by LSOA

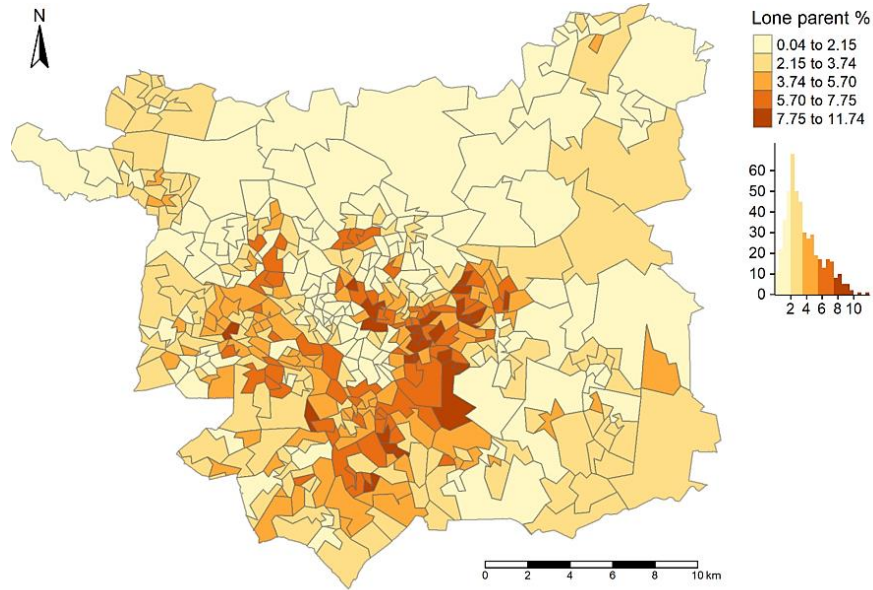
It can be seen from Figure 5.1 (a) that the city centre has the highest concentration of the young population (up to 85.05%); it also has the highest diversity (0.93) of age in Leeds (Figure 5.1b). These areas have a higher proportion of students notably in Headingley, Hyde Park and Burley & Little Woodhouse respectively and burglary rates of 241.2 per 1000 population (Figure 3.5) are the highest in the city and present suitable targets for burglary victimisation. In affluent areas such as Otley, Methley, East Keswick, Collingham and Linton the rates are between 10.9 and 65.2 per 1000 population. These areas also tend to have a lower percentage (5.2 to 12.09) of young population and lower age diversity (0.05 to 0.13) respectively. In terms of correlation with burglary rates (Table 5.2), the age diversity variable has a relatively stronger correlation ($r = 0.36$, $p < 0.01$) with burglary rates than the percentage of young population variable ($r = 0.27$, $p < 0.01$) both appeared to be strongly correlated ($r = 0.89$).

5.6.2 Family Structure

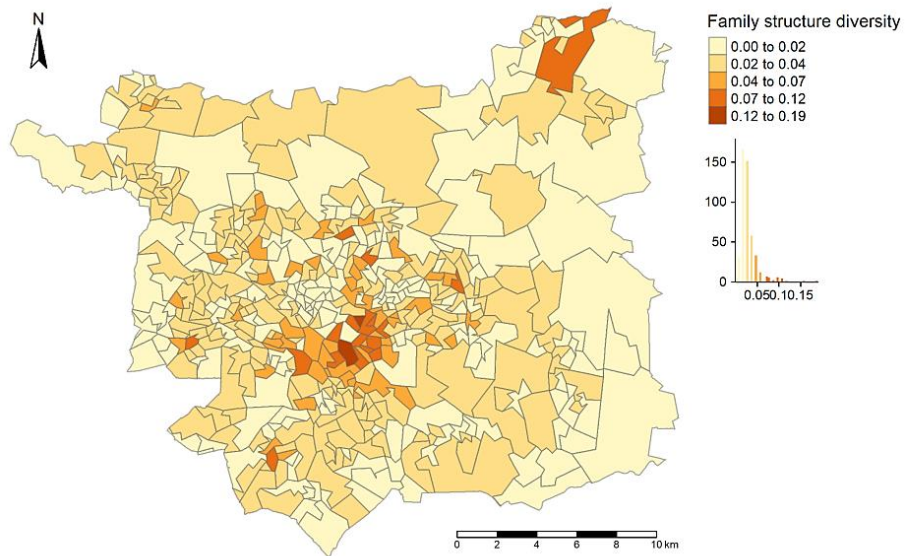
Maginnis (1997) has argued that the children of some single-parent families are more likely to have behavioural problems, because they tend to lack economic resources and have lower parental input. In the UK, lone parents continue to suffer from inequalities of employment and housing, creating a gap between couples and lone parents (Berrington, 2014). Additionally, single parents are also most likely to be victims of crime due to social marginalization in terms of living conditions (Wikström and Wikström, 2001). Given this, we include lone parents as an indicator from the traditional literature.

In terms of diversity, it seems likely that the distribution of family structures constitutes an important determining factor in social cohesion among community residents. However, this is likely to revolve around households with or without children. For example, two-parent families with children tend to form social groups within the community that are distinct from single-parent families (Sampson and Wooldredge, 1987; Kanazawa, 2003). Community support within lone parenting groups is undoubtedly strong in some areas, but is likely to be more geographically variable. Given that the determinants are largely the presence or absence of children, and the presence or absence of lone parents bearing in mind that the number of children is largely random in most populations, and ethnically controlled otherwise, and that not having children encompasses populations that are both very young and very old, and little else (Rees and Butt, 2004), we offer diversity of family structure (calculated using equation 3.8 and the family

structure components in Table 3.2) as a variable in the model based on these factors. The spatial distribution of family structure variable in Leeds is presented in Figure 5.2.



(a) Lone parent households (percentage), LSOA



(b) Family structure diversity by LSOA

Figure 5.2: Family variables (standard (a) and diversity (b)) in Leeds by LSOA

The distributions of family variables (Figure 5.2 (a and b)) exhibit less obvious spatial patterns but it can be seen that a relatively higher percentage (11.74) of lone parent families concentrate in areas like Cross Green and Burmantofts, while areas such as Wetherby and Holbeck have a relatively higher family structure diversity (0.19). Both areas are associated with higher burglary rates of between 109.9 and 143.3 per 1000 population (Figure 3.5). The relationship between burglary rate and family variables is weak but significant. For example, the correlation between family structure diversity and burglary is 0.26 while for the percentage of lone parent variable the correlation is 0.16; the two family variables are not correlated (0.12) (Table 5.2).

5.6.3 Identity

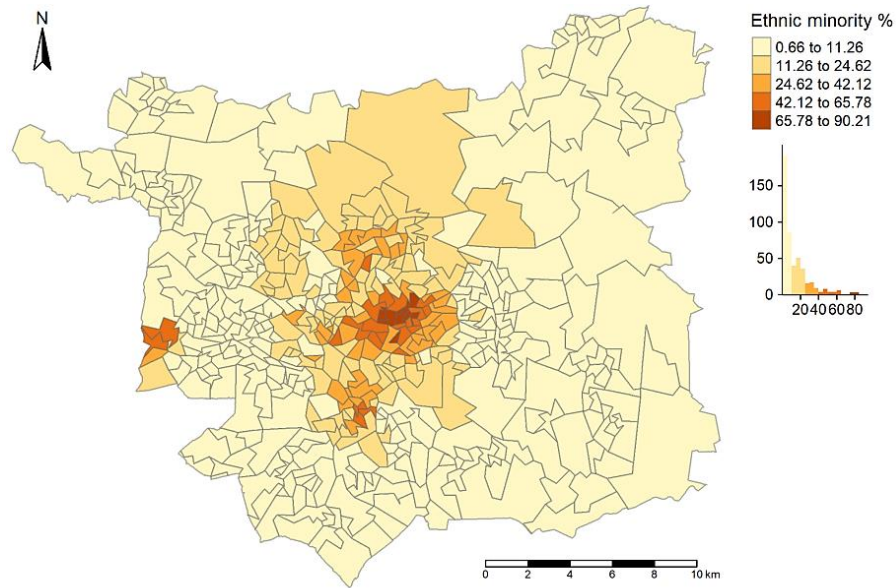
Identity plays an important role in the likelihood that people will connect and form social relationships (Gilchrist and Kyprianou, 2011) and plays an important part in the integration of migrants into local neighbourhoods (Kindler *et al.*, 2014). Migrants especially from the black and minority ethnic (BME) populations often lack the wealth, social integration, or formal crime prevention connections to protect themselves (Sharp and Atherton, 2007). Because of these factors, the size of an immigrant population in an area positively correlates with the incidence of property crime (Bell and Machin, 2011). In a study of the relationship between immigrants and crime across Italian provinces, Bianchi *et al.* (2012) found an increase of 1% immigrant population to be associated with 0.1% increase in the total number of criminal offences. Empirical evidence from the US also demonstrates links between size of an immigrant population and occurrence of other types such as motor vehicle theft and robbery (Bholowalia and Kumar, 2014). However, previous studies in the UK have yet to empirically establish the link between increases in the size of immigrants in an area with incidence of property crime specifically.

Ethnicity has a notorious relationship with crime, principally acting through socio-economic exclusion and disadvantage. For example, figures from the police stop and search by ethnicity has shown an increase of arrest by 7% (68% to 75%) on White suspects compared to BME that has reduced by 5% (17% to 12%) between 2009 and 2014 (ONS, 2015c). However, in terms of convictions for criminal offences based on recorded ethnic backgrounds (by the police), the ONS study has found that defendants being from a BME background are more likely to be sent to prison compared to those from the White background (Kathryn, 2016). There are biases in

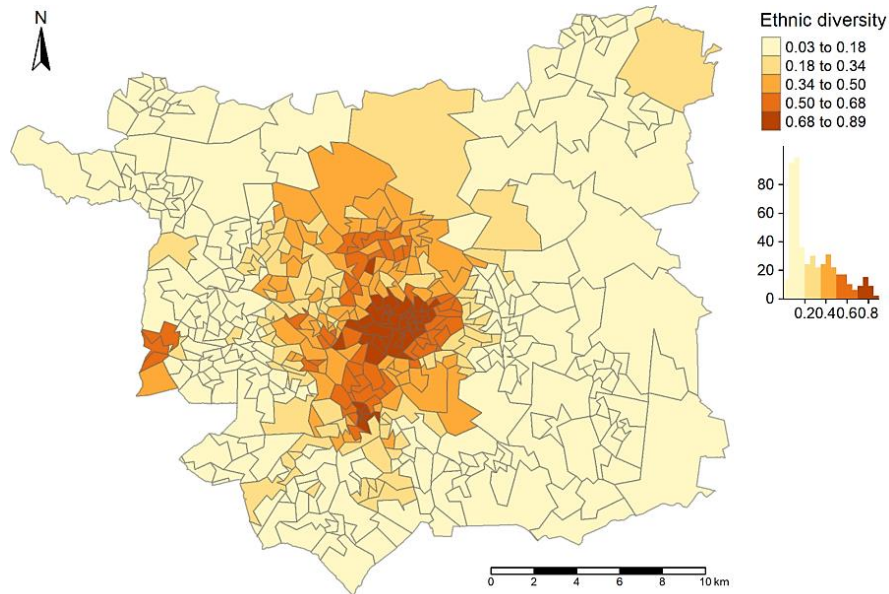
reporting, the justice system, and policing, the latter including complex relationships between race and prejudice, most clearly expressed in the findings of the Stephen Lawrence Inquiry in the UK (Macpherson, 1999). Here we deal with the relationship between residential ethnicity and reported crime, ignoring, unfortunately, these complex and important nuances, the statistical treatments for which are largely unresolved.

Additionally, in Leeds the number of students is important because of the presence of large residential educational institutions (especially the two universities). Students are more likely to fall victims of crime especially burglary, because multiple occupancy homes are attractive to burglars, and because students are less likely to be at home (Shepherd, 2006; Kongmuang, 2006). Furthermore, student residences are attractive to burglars because students are more likely to possess valuable items especially electronic gadgets (e.g. DVDs, laptops, Ipads and mobile phones) and be less careful about security of their personal belongings (Barberet and Fisher, 2009). Additionally, some students reside in poor accommodation that lacks security surveillance devices such as closed circuit television (CCTV) and may not be adequately patrolled (Masike and Mofokeng, 2014). There is no equivalent diversity statistic for student population.

In terms of diversity, ethnic diversity might be important in negative ways across broad types of the crime system including offending and victimisation, the latter including, obviously, hate crime (Shepherd, 2006). Moreover, Vermeulen *et al.* (2012) have argued that the negative effects of ethnic diversity on social networks would probably be stronger in heterogeneous communities than in more homogenous neighbourhoods, because in a more diverse society lack of interpersonal trust, as well as differences in interests and needs between groups weakens networks of social interaction, therefore limiting chances for social integration. Though counter-arguments can be made in areas where everyone is essentially part of a minority population, the nuances of tension and disadvantage in communities of multiple ethnicities and country of origin are likely to be complex. We therefore include diversity of ethnicity (calculated using equation 3.8 and the ethnicity components in Table 3.2) to capture the complex elements of offending and victimhood associated with ethnicity in mixed communities. Figure 5.3 shows the distribution of ethnicity variable in Leeds.



(a) Ethnic minority (percentage), by LSOA



(b) Ethnic diversity by LSOA

Figure 5.3: Ethnicity variables (standard (a) and diversity (b)) in Leeds by LSOA

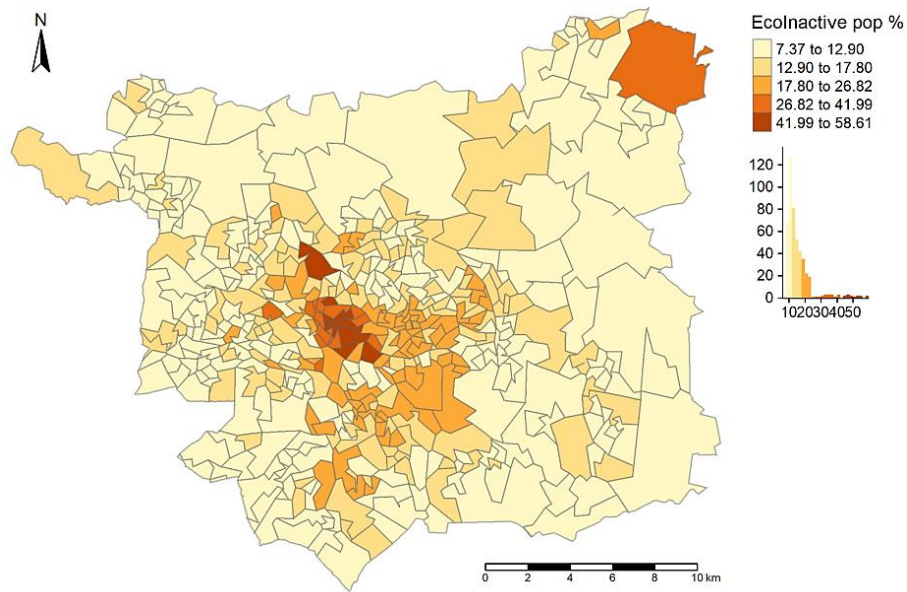
It is clear from Figure 5.3 (a) that higher concentration of the ethnic minority population (90.21 %) is located around inner city suburbs especially, Harehills, Chapeltown and Gipton. The ethnic diversity (0.89) of these areas is also the highest in Leeds (Figure 5.3b), meaning that there is a likelihood that two individuals chosen at random to be of different ethnic groups; and tend to

have higher burglary rates of 109.9 per 1000 population (Figure 3.5). However, areas with a lower percentage of minority population (0.66 to 11) and lower ethnic diversity (0.03 to 0.18) such as Methley, Ardsley East & West are characterised with a lower burglary rates of 10.9 to 44.6 per 1000 population. The correlation between ethnic diversity and burglary rate ($r = 0.32$, $p < 0.01$) is relatively stronger than the percentage of ethnic minority population ($r = 0.19$, $p < 0.01$) and both variables are strongly correlated (.94) (Table 5.2).

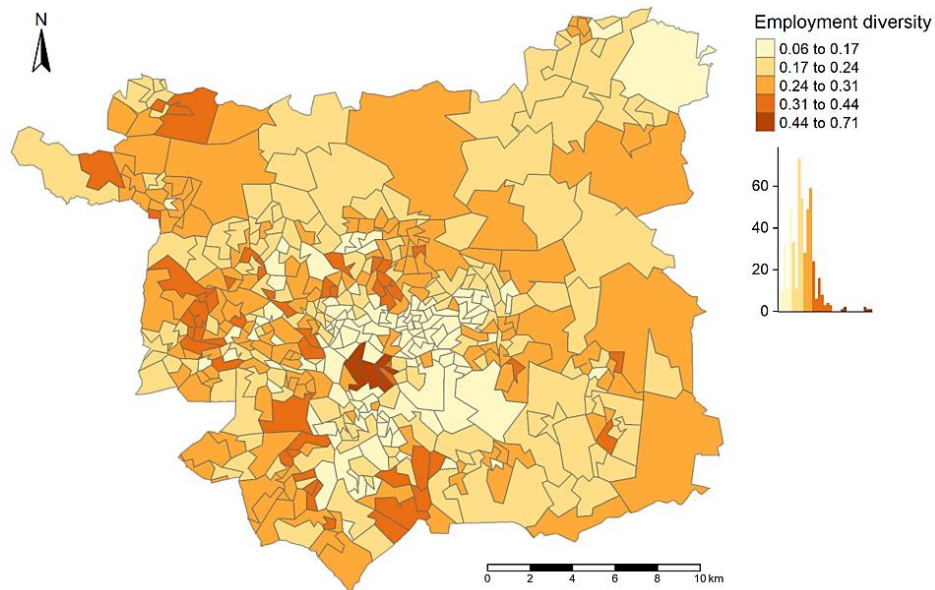
5.6.4 Affluence and Wealth

While there is a wide range of criminality across the socio-economic spectrum, for burglary, the offenders in the vast majority are drawn from the poor and unemployed (Bursik Jr and Grasmick, 1993; Sariaslan *et al.*, 2013). Given this relationship, the level of economic status in those age categories could be used as a key indicator.

Wealth diversity within a community may act to increase crime. Given that most burglars only travel a short distance to commit crimes (Ashby, 2005), there is some evidence that disparities of wealth within short distances encourage burglary, and in addition, disparity of wealth within a community can influence crime by weakening social cohesion (Fajnzlber *et al.*, 2002; Rufrancos *et al.*, 2013). Equally, low wealth diversity can enhance social cohesion (Cooper and Innes (2009). Although the picture is complicated across other types of crime (Rufrancos *et al.*, 2013), researchers have found support for the relationship between property crime and income inequality (Witt *et al.*, 1998; Kelly, 2000; Demombynes and Özler, 2005; Reilly and Witt, 2008). We therefore include (in the absence of income data in the UK census) diversity of employment (calculated using equation 3.8 and different employment components in Table 3.2) type in our assessment. The spatial distribution of the affluence variables in Leeds is shown in Figure 5.4.



(a) Economically inactive population (percentage), LSOA



(a) Distribution of employment diversity by LSOA

Figure 5.4: Distributions of affluence variables (standard (a) and diversity (b)) in Leeds by LSOA

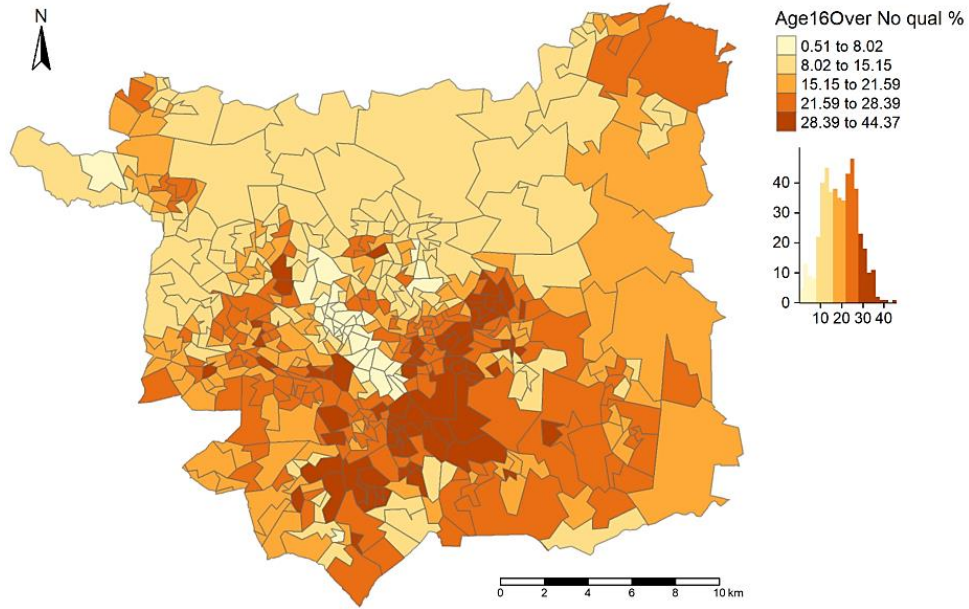
The City Centre and West Park have the highest percentage of economically inactive population (up to 58.61) and Boston Spa (up to 41.99) and the lowest percentage is found in areas like Harewood and Bramhope (7.37 to 12.90) (Figure 5.4 (a)). The distribution of employment

diversity (Figure 5.4 (b)) is complex and does not exhibit clear pattern, but the City Centre has the highest values of between 0.44 and 0.71; however, the general pattern tends to reflect higher values of employment diversity to the affluent areas. Correlation between the variables of affluence/wealth and burglary rate (Table 5.2) shows that diversity of employment has a non-significant negative correlation ($r = -0.03$) while the percentage of economically inactive population has a significant positive correlation with burglary ($r = 0.28$, $p < 0.01$), correlation between the two variables of affluence is weak (-0.51).

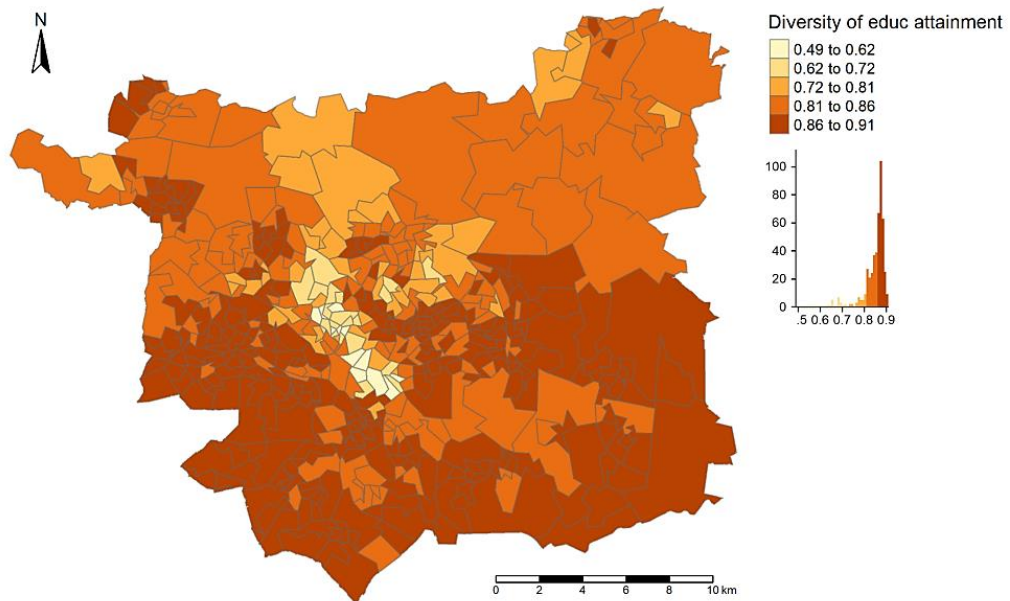
5.6.5 Educational Attainment

Educational attainment has a great influence on individuals' social behaviour as well as on participation in community activities. The theory of human capital suggests that skills and qualification determine wages, and the wider the distribution of qualifications, the wider the distribution of wages (Green *et al.*, 2006). Reynolds *et al.* (2001) stressed that the propensity of individuals to commit a crime is associated with their level of educational attainment and so we include lack of qualifications as a standard variable.

However, there is also likely to be an indirect relationship between social cohesion, educational inequality and crime. Sabates *et al.* (2008) stressed that educational inequality is associated with violent crime. While it is unclear whether such a relationship acts at the intra-area scale independent of any effect on wealth, we include a diversity statistic centred on education (calculated using equation 3.8 and different components of educational attainment in Table 3.2) to test the potential relationship. Figure 5.5 shows the spatial distribution of educational attainment in Leeds.



(a) Percentage of people with no qualification by LSOA



(b) Educational attainment by LSOA

Figure 5.5: Educational attainment variables (standard (a) and diversity (b)) in Leeds by LSOA

As clearly seen from the distribution of population aged 16-64 with no educational qualification (Figure 5.5 (a)), a higher percentage (28.39 to 44.37) of this category of people concentrate in areas such as Hunslet/Stourton and a relatively lower percentage (8.02 to 15.15) are found in

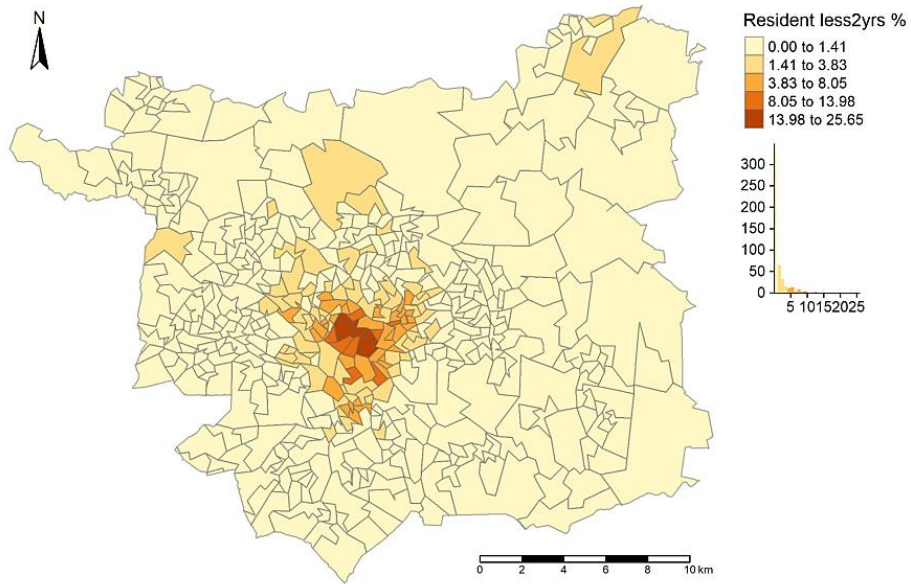
areas such as Thorner. The spatial distribution of educational attainment (Figure 5.5 (b)) shows a higher diversity values (0.86 to 0.91) in areas like Drighlington and Gildersome and a lower diversity values (0.49 to 0.62) in areas such as Hyde Park and Burley Lodge. The correlation between the percentage of population without educational qualification and burglary rate (Table 5.2) is not significant ($r = 0.08$); however, a relatively significant but weak relationship is found between diversity of educational attainment and rate of burglary ($r = -0.24$, $p < 0.01$) and both variables are weakly correlated (.57).

5.6.6 Residential Instability

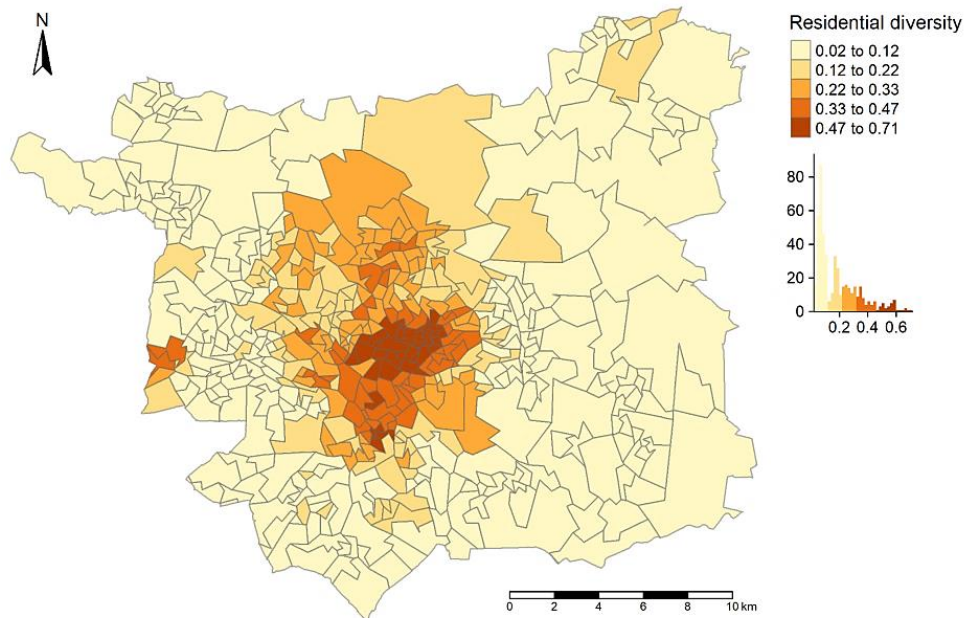
Residential stability in a neighbourhood is an important factor for generation of social capital and place-based attachment. It is therefore expected that duration of residence is one determinant factor in this direction (Thomas *et al.*, 2016). Studies have demonstrated that the creation of social ties is associated with the length of residence in an area. For example, Yamamura (2011), argued that personal interactions are built over time and tend to be more solid when people reside in a particular neighbourhood and are influenced by length of residence and home ownership. Similarly, Keene *et al.* (2013) also stressed that it takes time to create supportive social ties, therefore length of neighbourhood residency may be an important determinant of social integration. Additionally, Oh (2003) also emphasised that length of residence has a positive effect on friendships, social cohesion and trust which also enhances the probability of working together to solve local problems. In contrast, residential instability in a neighbourhood is associated with weak social ties and a low probability of residents connecting (Sampson *et al.*, 1997). These communities include large proportions of students and other migrants on a short stay.

Crime is also more likely to occur in transient neighbourhoods. For example, in the UK, the tendency to commit crime is related to length of residence, in other words, crime reduces as length of residence increases (Bell and Machin, 2011). Specifically, research has established the relationship between residential instability and violent crime (e.g. Boggess and Hipp, 2010). However, the relationship between residential instability and burglary is likely to be complex (Markowitz *et al.*, 2001; Martin, 2002). Given this, we include length of residence less than two years as a standard variable and diversity of length of residence (calculated using equation 3.8

and different components of length of residence in Table 3.2) as proxy for residential instability (Figure 5.6).



(a) Population resident less than 2 years (percentage), by LSOA



(b) Length of residence diversity by LSOA

Figure 5.6: Length of residence variables (standard (a) and diversity (b)) in Leeds by LSOA

For the percentage of length of residence less than 2 years variable, highest values (13.98 to 25.65) are mainly located in the City Centre (Figure 5.6 (a)). Similarly, the diversity equivalent variable also indicates a higher diversity value (0.47 to 0.7) in the City Centre (Figure 5.6 (b)); burglary rate in these areas is between 143.3 and 241.2 per 1000 population the highest in Leeds. Correlation between burglary rate and these variables is relatively strong, for length of residence diversity ($r = 0.31$, $p < 0.01$) and the percentage of population resident less than 2 years ($r = 0.30$, $p < 0.01$) while correlation between the stability variables is relatively strong (.74) (Table 5.2).

Index of Multiple Deprivation

Finally, the deprivation hypothesis suggests that deprived communities tend to have more crimes than affluent communities (Sampson and Wooldredge, 1987; Malczewski and Poetz, 2005). Furthermore, deprivation widens the gap between the rich and poor which can reduce social cohesion (Morenoff *et al.*, 2001; Takagi and Kawachi, 2014). The Index of Multiple Deprivation (IMD) is a multi-dimensional metric that is measured, in England, through a combination of seven distinct domains: income; employment; education; health; crime; barriers to housing & services; and living environment (DCLG, 2015). Although, deprivation is often seen as a key indicator of social cohesion as well as propensity to commit a crime such as burglary, the UK deprivation statistics include crime and therefore it is inappropriate to use them in this analysis. Figure 5.7 shows the distribution of IMD in Leeds.

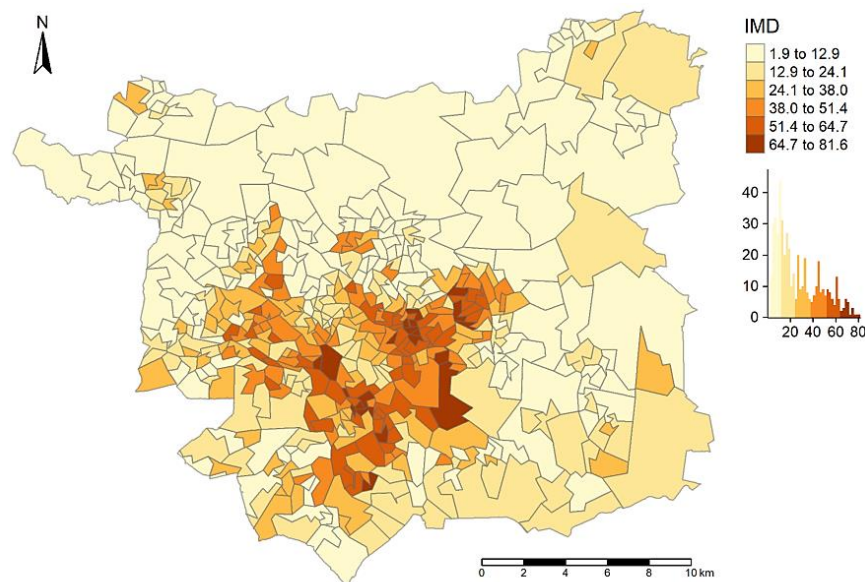


Figure 5.7: Index of Multiple Deprivation in Leeds by LSOA

From Figure 5.7, the most deprived areas of Leeds appear to be concentrated around the City Centre, notably areas with a higher concentration of minority populations. For example, Harehills and Burley lodge have a multiple deprivation index of between 64.7 and 81.6 compared to less deprived areas such as Yeadon with a multiple deprivation index of between 1.9 and 12.9.

Table 5.2: Pearson's correlation coefficient for burglary and other dependent variables (standard and diversity)

Variables	Burglary rate	Eth div	Age div	Empl div	Edu div	Res div	Familyst div	Age16-24%	Age16-64 EcoInactive%	Lone parent%	Ethnic minority%	Lres2yrs%
Eth div	0.32**											
Age div	0.36**	0.39										
Empl div	-0.03	-0.26	0.14									
Edu div	-0.24**	-0.12	-0.75	-0.27								
Res div	0.31**	0.97	0.44	-0.23	-0.15							
Familyst div	0.26**	0.23	0.16	0.21	-0.16	0.3						
Age16-24 %	0.27**	0.27	0.89	-0.16	-0.7	0.3	-0.11					
Age16-64Eco Inactive %	0.28**	0.43	0.73	-0.51	-0.49	0.46	-0.09	0.88				
Loneparent %	0.16**	0.24	-0.2	-0.5	0.47	0.2	0.12	-0.21	0.05			
Ethnic Minority %	0.19**	0.94	0.27	-0.32	0.01	0.93	0.14	0.19	0.38	0.23		
Lres2yrs %	0.30**	0.63	0.68	-0.07	-0.37	0.74	0.29	0.54	0.58	-0.06	0.58	
Age16over Noqual %	0.08	-0.02	-0.45	-0.52	0.57	-0.03	0.23	-0.4	-0.13	0.71	0.03	-0.22

5.7 Results of the Traditional Multiple Regression Models

The results of the analyses (Table 5.3) summarise the stepwise regression models used to assess the relative importance of each variable in the model. The statistics reported are Pearson's product moment correlation (R), which shows the correlation with the dependent variable for each model. R-squared reports the percentage of variation in rate of burglary crime explained by the variables used in the model. Adjusted R-squared is the fraction by which the square of the standard error of the regression is less than the variance of the dependent variable. It increases only if the variables improve the model. It is usually used to evaluate which model performs better, where a model with a smaller standard error of estimate is likely to produce a higher adjusted R-squared (Kongmuang, 2006).

Table 5.3: Model summary of stepwise regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.358	.128	.126	32.061
2	.447	.200	.196	30.746
3	.469	.220	.215	30.384
4	.481	.232	.225	30.191
5	.492	.242	.234	30.010

Table 5.4 presents the coefficients of the model. The elements reported are standardized and unstandardized coefficients, standard error, t-statistics, significance and collinearity (*VIF*) tests. In regression analysis, standardized coefficients are estimates standardized so that the variance of the dependent variable produced by the change in the independent is between -1 and 1; while unstandardized coefficients expressed values of the relationship in raw values (Landis, 2014). The standard error is a measure of the accuracy of predictions obtained from difference between the observed and predicted values, smaller values better indicate how closer observations are to the fitted regression line (Bakliwal *et al.*, 2013). The stepwise regression results indicate that the parameters are within acceptable standards for regression modelling. An important guide for understanding this are the t-statistics (Dunn, 1989). The t-statistic is the estimated coefficient divided by its own standard error. Significant t-statistics should be approximately 1.96 in magnitude, corresponding to a p-value less than 0.05 or 95% confidence level (Coe, 2002b). The result obtained from the stepwise regression in this analysis indicates that the values of the t-statistics for all variables in the models were greater

than 1.96, meaning that all variables are statistically significant. In stepwise regression models, variables that are not statistically significant are iteratively dropped and model variables that are statistically significant are retained. Additionally, the variables show no multicollinearity problem with VIF values less than 5. In this analysis, model 5 has been identified as the most reliable and is given in a regression Equation (5.3) by computing the values of unstandardized coefficients (B):

Table 5.4: Coefficients and tests of model performance

Model		Unstandardized Coefficients		Standardized Coefficients		Sig.	Collinearity statistics VIF
		B	Std. Error	Beta	t		
1	(Constant)	53.775	2.336		23.018	.000	
	Age_div	74.289	8.851	.358	8.393	.000	1.000
2	(Constant)	23.531	5.130		4.587	.000	
	Age_div	102.110	9.490	.492	10.759	.000	1.250
	Age16overNoqual%	1.273	.194	.300	6.554	.000	1.250
3	(Constant)	22.092	5.086		4.344	.000	
	Age_div	86.719	10.341	.418	8.386	.000	1.520
	Age16overNoqual%	1.145	.195	.270	5.867	.000	1.294
	Eth_div	24.609	6.964	.157	3.534	.000	1.216
4	(Constant)	101.265	30.075		3.367	.001	
	Age_div	60.622	14.180	.292	4.275	.000	2.894
	Age16overNoqual%	1.344	.208	.316	6.469	.000	1.483
	Eth_div	28.885	7.102	.185	4.067	.000	1.281
	Edu_div	-93.151	34.882	-.181	-2.670	.008	2.850
5	(Constant)	106.442	29.962		3.553	.000	
	Age_div	85.607	17.060	.412	5.018	.000	4.240
	Age16overNoqual%	1.509	.216	.355	6.985	.000	1.625
	Eth_div	32.582	7.202	.208	4.524	.000	1.333
	Edu_div	-97.983	34.723	-.190	-2.822	.005	2.858
	Age16-64EcoInactive%	-.675	.260	-.163	-2.600	.010	2.482

Dependent variable: Burglary rate

The most notable result of this analysis is the almost complete exclusion of standard variables in preference for diversity statistics. This not only highlights the importance of diversity in the crime system, with a concomitant suspicion that this acts through community cohesion,

but also highlights that standard statistics are probably, in part, representing community cohesion, and are being excluded here simply because the new metrics are more stronger correlates of community cohesion. In short, this analysis is likely to highlight the importance of community cohesion within the crime system.

It is a common practice to assess the appropriateness of a model using the coefficient of determination, although it is not an absolute indicator of goodness of fit (Reisinger, 1997), and a low effect size does not mean the model is inefficient (Weisburd and Piquero, 2008; Martin, 2014). Although the analysis explained approximately 23% of the variation of burglary rate, this is good compared to other studies: Zhao *et al.* (2015), Karyda (2015), Hino *et al.* (2016), Lai *et al.* (2016) and Boateng (2016) having their models explaining 21%, 10%, 14%, 25% and 12% only. Crime (especially burglary), is difficult to understand, predict and model (Malleon and Birkin, 2012). The percentage of variation of the dependent variable explained in a model can sometimes be misleading, as small effect size can produce a better and more meaningful outcomes than larger ones (Lieberman, 1985). However, this depends on the unit of analysis, type of crime and underpinning theory (Weisburd and Piquero, 2008). The following is the optimal model of burglary rate using LSOAs in Leeds.

$$\text{Burglary rate} = 106.442 + 85.606 * \text{Agediv} + 1.509 * \%16\text{overNoqual} + 32.582 * \text{Ethdiv} - 97.983 * \text{Edudiv} - .675 * \% \text{Age}16\text{-}64\text{EcoInactive} \quad (5.3)$$

The model development process identified age diversity, percentage of population with no educational qualification, ethnic diversity, diversity of educational attainment and percentage of economically inactive population as the most important correlates of crime from those input. Educational diversity and economically inactive population and diversity of employment negatively correlated with burglary rates.

In this research, diversity of age is the most important variable when regressed against burglary rate and this remains consistent throughout the models (see model coefficients in Table 5.4). Age diversity plays to studies that have shown that offenders are commonly drawn from younger age groups than elderly people (Farrington, 1986; Gottfredson and Hirschi, 1990; Sampson and Laub, 2003; McVie, 2005; Blonigen, 2010; McCall *et al.*, 2013; Sweeten *et al.*, 2013) but that it is likely that a wide age range puts young offenders in close proximity with older victims with, potentially, more to steal. Equally, however, we know that the young are also targets for crime (Mahmud *et al.*, 2014), and it makes some sense that the

broader the range of population characteristics in an area the more likely that there will be suitable target criteria for burglars making decisions about risk (Bernasco and Nieuwbeerta, 2005).

Educational attainment has great influence on individuals' social behaviour as well as income. Education attainment determines wages (Green *et al.*, 2006) as well as the propensity of individuals to commit a crime (Reynolds *et al.*, 2001). Educational attainment increases returns through legitimate means; it also raises the opportunity cost of illegal behaviour (Machin *et al.*, 2011). In this study, however, *diversity* of educational attainment negatively correlates with burglary crime, meaning that the smaller the diversity of educational attainment, the more propensity there is to commit crime in an area. This requires further research, but immediate hypotheses are that low diversity of educational attainment is correlated with deprivation, and/or that low diversity areas including student residential areas, which in Leeds are very homogeneous communities with a high level of victimisation.

Research has shown that ethnically heterogeneous communities are often characterised by distrust, low levels of social cohesion and disputes (Sturgis *et al.*, 2014) which negatively affect individual behaviours (Mellgren, 2011). Recent studies into the spatial distribution of neighbourhood crime consistently show that areas which are characterised by ethnic diversity have high rates of crime (Gartner, 2013; Takagi and Kawachi, 2014; Hooghe and De Vroome, 2016). In this study, we have also found strong support for this relationship between ethnic diversity and rates of burglary crime.

In this research, we found a significant negative correlation between economically inactive population and burglary crime which might be seen as counterintuitive. Previous studies have found support for relationships between income inequality and property crime (Witt *et al.*, 1998; Kelly, 2000; Demombynes and Özler, 2005; Reilly and Witt, 2008). However, correlation between economic inactivity with burglary crime does not necessarily imply causation as this relationship might only suggest that unemployment might contribute to offending elsewhere. Recent statistics in the UK show that economically inactive people are twice as likely to be victims of burglary crime than those who are economically active, considering this category of population comprise of students, those who are retired and people with long-term health challenges (ONS, 2014b), so clearly the relationship for Leeds needs further investigation.

5.8 Concluding Remarks

This chapter employed traditional regression modelling (stepwise) to explore the impact of diversity and community cohesion on burglary rate in the Leeds district, UK using the standard and diversity indices. The analysis demonstrates that diversity based statistics are a better correlate with burglary than most standard metrics, highlighting the importance of diversity in the crime system, and suggesting the importance of social cohesion in preventing crime. This suggest that standard statistics go some way, normally, to capture the influence of social cohesion, but that this is better captured through diversity statistics.

The variables used in this study have provided useful insights into the relationship between neighbourhood social context (diversity) and the spatial variability of burglary crime in Leeds. The most important predictor for modelling burglary crime rates in this analysis is age diversity. However, other predictors such as ethnic diversity, distribution of educational attainment, proportion of those with no qualification and economically inactive population also made a valuable contribution to the models. Notably, economically inactive population had a slight negative relationship with crime, and this needs further investigation.

It seems likely that community cohesion is an important factor in establishing social control and collective efficacy in the neighbourhoods with regards to crime. In this analysis a simple diversity statistic is used to highlight the possibilities for investigating this relationship. The insights gained from the analysis in this chapter especially using the adjusted (diversity) and the standard variables will be employed in chapter 6 to explore with new forms of data sources such as social media (Twitter) the relationship between sentiments and community cohesion.

Chapter 6

Social Media Analysis for Understanding Community Cohesion

6.1 Introduction

“Non-traditional” social media methods for quantifying community cohesion were reviewed in chapter 4. As highlighted, social media data, especially Facebook and Twitter, have a wide range of potential applications, one of which is for understanding social cohesion in different communities. The sentiment of Tweet content and posts from Facebook pages can be analysed to gain insights into the dynamics of human collective social behaviour and to uncover patterns of social relationship. This chapter will extend our understanding of these relationships through social media data analysis, especially sentiment, with a view of highlighting the potential relevance of these new forms of data to explore social phenomena, which is one of the objectives of the present research. Section 6.2 begins by discussing the importance of social cohesion and public sentiment on social media. Section 6.3 describes the Twitter data collection procedure, while Census variables included are highlighted in Section 6.4. Sentiment Analysis is described in Section 6.5 while results of the analysis are presented in Section 6.6. Section 6.7 follows with a discussion of findings from the analysis and finally, concluding remarks are given in Section 6.8. Chapter 7 will further extend our understanding of the relationships between community cohesion and social media through engagement analysis of posts from the community Facebook pages.

6.3 Twitter Data

Twitter APIs used for data scraping is described in Section 4.2.2. The streaming API used in this research provides two different methods of data collection from Twitter: keywords and geographical bounding box. These methods are described in Sections 6.3.1 and 6.3.2.

6.3.1 Keywords Method

Keywords search method is commonly used for data scraping on Twitter (Lewis *et al.*, 2013; Gerlitz and Rieder, 2013). Keywords search refers to a method for extracting tweets that match specific terms or criteria on Twitter (Marujo *et al.*, 2015). For example, a query #leeds will find tweets containing the hashtag “leeds”; while a query @saferleeds will search for tweets mentioning Twitter account @saferleeds; and a query “community cohesion” will

return tweets containing the exact phrase “community cohesion” (Twitter, 2017c). However, the criticism of the keywords search method is that the terms requested are already predetermined by the researcher, hence introducing elements of bias in data collection (González-Bailón *et al.*, 2012) and therefore not considered as representative of a random sample of events (Zhang *et al.*, 2016). Instead, in this research, data were collected within the bounding box of Leeds (the smallest square encompassing Leeds Metropolitan District) avoiding any specific search terms.

6.3.2 Geographical Bounding Box Method

The geographical bounding box method is used for requesting a sample of tweets as allowed by Twitter posted within a specifically delineated boundary (defined set of coordinates). This type of query method will typically return tweets with a geolocation (Culotta, 2014). However, data collected using this method may also include tweets without geolocations because while the original tweet may have a geolocation, the retweet may not (Twitter, 2017d). The benefit of this method is that it can potentially increase the proportion of geolocated tweets captured from within an area of interest, reducing bias and increasing the value of the data. Although this method of data collection might capture tweets from both usually resident inhabitants and temporary communities (those commuting to Leeds from elsewhere for work) (Mahmud *et al.*, 2012; 2014). According to Malleson and Andresen (2016), cities (Leeds included) attract large numbers of people during the day who return to their homes at night. The ambient population is the average number of the expected population present at some point in time at a particular spatial scale (Andresen, 2011). Such populations are highly dynamic and tend to exhibit strong spatial fluctuations (Malleson and Andresen, 2015). Additionally, data collected in this way can potentially produce more reliable results. Based on this method, tweets data published in the English language (N=5,000,000) were collected over a period of nine months (December 2015 to September 2016) at a different time, this is to ensure that data collected reflects different moods. Moreover, the advantage of collecting data at different points is to increase the generality, minimising the tendency of data exhibiting a similar general mood when collected all in one go (Sampson *et al.*, 2002). For example, tweets collected during a festive period like Christmas might have significant portion referring to positive sentiments.

However, because we are interested in tweets with a geolocation, we apply a filtering method to remove duplicates and extract geolocated tweets from unique users that fall within the

geographical boundaries of Leeds community areas (N=63,375). This sample is large enough as it accounts for about 9% of Leeds population as is in 2011, and is therefore used in this research.

6.3.4. Twitter Metadata

Twitter data has associated metadata including the tweet text itself, the username and numerical ID of the sender, the geo-location of the sender at the time of tweeting (where this functionality is enabled on a device), timestamps (digital record of time when the tweet is posted) retweets, mentions and replies (Bruns and Stieglitz, 2014). Retweets preceded by RT refer to the number of times a user retweeted the tweet; mentions refer to tweets that mention your handle (e.g. @handle) and replies are the number of times a user replied to the tweet (Twitter, 2017a).

Of all the metadata provided by Twitter, geolocation is the most unpredictable because it has to do with the privacy of the user. The geolocation metadata became available on Twitter in 2009, allowing users to willingly provide location information of where each tweet is published (Leetaru *et al.*, 2013). There are two geolocation options provided by Twitter: place-based and exact location. Place-based location allows users to manually specify a city or neighbourhood from a predefined list supported from the Twitter menu; regular updating may be required depending on user movement from one country to another. This type of geolocation is primarily used while tweeting from a fixed location device (e.g. desktop). On the other hand, the exact location provides a set of geographic coordinates for the user at the time of tweeting and no action is required by the user provided that GPS functionality is enabled on the mobile device (Twitter, 2017a).

6.3.5 Twitter Data Pre-processing

Detecting sentiment is difficult as tweets are unstructured and contain colloquial language (Schulz *et al.*, 2013). Data collected from social media sources, especially Twitter, contained unwanted noises, such as, web links (e.g <http://twitter.com>); stop words (e.g “to”, “is”, “are” “we”, “this”, “that”); symbols (“@”, “#”); and non English letters (“Ã”, “Ì”, “Õ”, “Ù”, “É”), as well as numbers and blank spaces. When sentiment analysis (SA) is applied on raw tweets it may result in poor performance because it is difficult for the algorithm to analyse all noises contained in tweets. Therefore, preprocessing involving data cleaning is needed in order to achieve a satisfactory outcome from the analysis of public tweets (Sahayak *et al.*, 2015; Kharde and Sonawane, 2016). Table 6.1 shows an example of raw tweets and processed

tweets. The text mining (TM) package, a natural language processing package for the English language developed by Feinerer and Hornik (2015), is used for data preprocessing and cleaning. The TM package algorithm is useful for converting tweets into a vector corpus (a collection of text documents), removal of stopwords and blank spaces and punctuation as well as word stemming. Stemming is a standardisation technique for reducing a word down to its roots in order to avoid multiple words referring to the same concept but in different grammatical contexts being identified as referring to different things. For example, the words “trust”, “trusting”, “trusted” may be stemmed to “trus”. The purpose is to make stem words as one when counting frequency. All words were also converted into lower case. We used a modified code developed by Zhao (2015) widely used for data mining using R, a language and environment for statistical computing (version 3.2.5; (R Core Team, 2013)). Prior to conducting SA on the Twitter data, it is a common practice to explore the corpus in order to visualise frequently occurring terms contained in the corpus. One method of visualizing tweets data is by using wordcloud and bar plots. To construct a wordcloud in R, we used a Wordcloud package developed by Fellows *et al.* (2015).

Table 6.1: Tweets before and after processing

Raw Tweets	Processed Tweets
Getting into my stouts now. Strong chocolate taste with a good kick... (Naughty & Nice) https://t.co/qdFEqqAsK0 #photo	getting stouts now strong chocolate taste good kick naughty amp nice photo
@GillianA We Love you over here you know xxx	gilliana love know xxx
Hometime :) (@ Leeds Bradford International Airport (LBA) - @lbiairport in Leeds_WestYorkshire) https://t.co/6c5WBIk1jd	hometime leeds bradford international airport lba lbairport leeds west yorkshire
Just posted a photo @ Otley Town Centre https://t.co/4spvt3AQlc	just posted photo otley town centre
Really looking forward to the second part of my #90daychallenge. It's all about muscle gain https://t.co/9VEUBAPyw5	Really looking forward second part day challenge muscle gain

6.3.6 Twitter Data Cleaning

Before SA is conducted, it is important to ensure that the data to be used meet the required quality in line with the set-out objective. Data pre-processing is used for reducing the complexity of data by transforming them into a more readable format in order to adapt the

data to the requirement of different algorithms (García *et al.*, 2016). Furthermore, cleaning is necessary to ensure that the data are free from elements that can cause bias in the results. Data cleaning is a process for detecting and removing errors and inconsistencies from the data in order to improve the data quality (Rahm and Do, 2000). However, identifying relevant tweets from a large dataset is a challenging process (Ghosh and Guha, 2013). For example, the presence of “bots” is common in tweets. A bot is internet software that can perform automated repetitive tasks such as tweeting and retweeting multiple times from the same user (Morstatter *et al.*, 2016). We included tweets with a social context in the analysis. Social context in tweets refers to the Twitter messages that have a likelihood of being retweeted because their content is engaging (Tao *et al.*, 2012). In this research, data cleaning was performed in the following steps:

- The first step in data cleaning was to query the dataset in order to identify and collect tweets from unique users with geolocation. Where a tweet falls within the bounding box of Leeds community area it is retained, otherwise they are removed from this step. Tweet geolocation is important in order to study patterns of human activities in different places (Holbrook *et al.*, 2016).
- The second step involves removal of duplicate tweets. Duplicate tweets are identical content sent from an identical user account (Tao *et al.*, 2013). Duplication increases the chances for the data to skew which can adversely affects the reliability of results (Gao *et al.*, 2017). In this research, we used an R code to scan and remove duplicate tweets that contained exactly similar words from the dataset.
- The third step was to identify the presence of bots. These are unwanted tweets contained in the data that have no relevance in the analysis. Bot detection is an important task in social media data cleaning, as their presence can lead to false conclusions about the populations under study (Morstatter *et al.*, 2016). In general, there is no simple set of rules for assessing whether an account is human or bot; however, a high volume of activity, especially in the same location, indicates a likelihood that the account is a bot (Varol *et al.*, 2017). It is therefore expected that bots will post tweet or retweet more frequently than humans, bots heavily retweeting existing content rather than posting new tweets (Gilani *et al.*, 2017). In this context, for example, bot tweets from a meteorological station, traffic and those with pornographic contents were detected in the dataset by calculating frequencies of tweets from each twitter user. We then manually checked the frequency of tweeting

from each user account. In this way, we set a threshold of 25 posts from any single user account, beyond which the account is regarded as suspicious bots and they are removed in order not to bias the analysis. These were then aggregated into groups. Figure 6.1 shows the histogram distribution of the first 100 accounts from the data and Figure 6.2 shows the aggregated distribution of tweets in the community areas.

- Step four checked for multiple tweets from the same user within the same location as they may tend to dominate the analysis and those with ambiguous coordinates. Ambiguity of geographical coordinates can affect accurate location of users (Huck *et al.*, 2015). Multiple tweets within the same location from the same user are removed, unless if they have different texts content.

The cleaning process ends when irrelevant contents are removed from the dataset; otherwise, the process is then repeated.

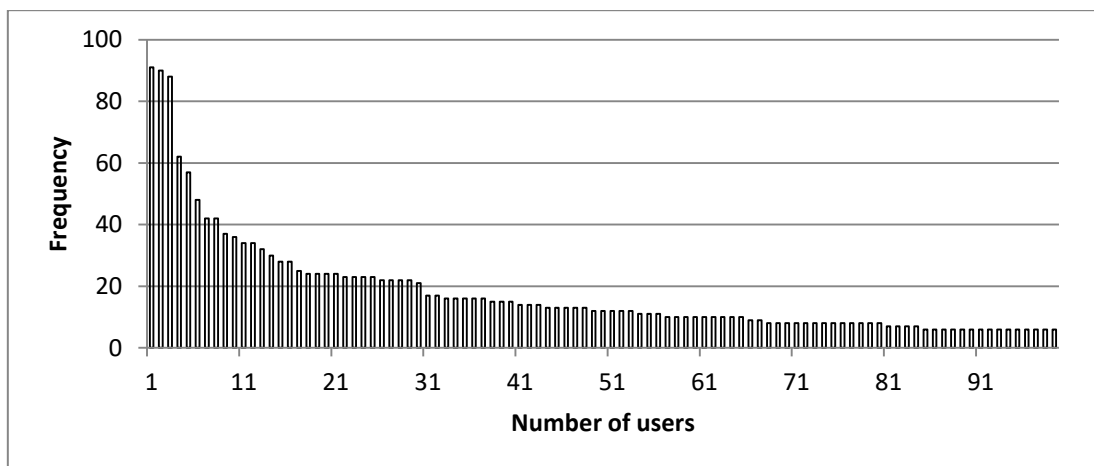


Figure 6.1: Histogram of the distribution of tweets in the community areas

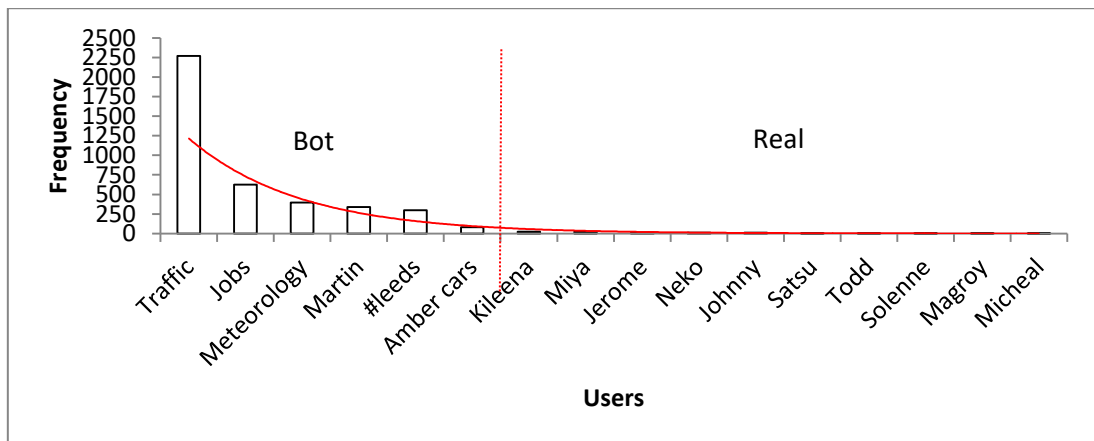


Figure 6.2: Aggregated distribution of tweets

6.4 Census Variables

Socio-demographic variables used in this research were mainly derived from the UK Census 2011. Variables selected for this analysis are age distribution, ethnicity, employment, educational attainment and length of residence as in chapter 5. These variables were chosen because of their potential as important indicators of social cohesion in the literature (e.g. Leventhal and Brooks-Gunn, 2000; Sampson *et al.*, 2002) as well as their influence on social media usage (Mitchell *et al.*, 2013; Lerman *et al.*, 2016; Lansley and Longley, 2016; Murthy *et al.*, 2016). Section 3.4.1 provides a detailed description of the variables while Table 3.2 provide the components included to measure diversity. Vargo and Hopp (2017) emphasised that census datasets can be combined with social media data to explain a social phenomenon. All variables were collected at lower super output area (LSOA) level and aggregated into 106 Leeds community areas (CAs).

In this research, in order to explore the relationships between Twitter sentiments and the demographic characteristics of areas, we compare SA results with the diversity (ethnic, age, employment, education and length of residence) of different areas. Equation 3.8 is used to construct diversity indices. Section 3.9 discussed diversity indices in more detail.

6.5 Sentiment Analysis Procedure

The SA algorithm used in this study is based on an unsupervised lexicon approach (see Section 4.7.1.1), typically used for classifying polarity in a collection of text documents as positive or negative (Taboada *et al.*, 2011). Lexicon methods, unlike human annotation, are cost effective and fast (Korhonen, 2009); they can also predict sentiment orientation of numerous words with high accuracy compared to a supervised approach (Duygulu *et al.*, 2002; Yang *et al.*, 2016) and are widely used for sentiment classifications (Kiritchenko *et al.*, 2014). This approach requires a dictionary of positive and negative words, each with a positive or negative sentiment value assigned to it (Jurek *et al.*, 2015). Although lexicon methods are sensitive to a domain-specific context, where a single word could have different meanings depending on where it is used (Hamilton *et al.*, 2016), they are sometimes difficult to deal with sarcastic sentences (Liu, 2012) but are widely used in SA research (Medhat *et al.*, 2014). In this research, a collection of 6,800 English words based on Hu and Liu (2004) were used for analysing sentiment orientation (positive and negative) in textual data. These methods do not rely on labelled data (Dang *et al.*, 2010), so can be employed to analyse text at the document, sentence or entity levels (Zhang *et al.*, 2011). When SA is applied on

aggregated data, the margin of error is significantly reduced, hence increasing reliability of results for decision making (Tan *et al.*, 2014).

In this analysis, we used aggregated Twitter data (document level) using the Breen (2011) sentiment analysis algorithm for classifying public sentiment. The algorithm scans the document and assigns an integer score (positive or negative) to each word. Then the sentiment value for each tweet is obtained by subtracting the number of occurrences of negative words from that of positive words in a text (Figure 6.3). In addition, superlatives are taken into account. For example, a tweet containing the word “good” would be assigned a positive score, “very good” would be assigned a high positive score, while a tweet containing the word “bad” would be assigned a negative score and so on (Hollander *et al.*, 2016). A neutral class provides a distinction between negative and positive sentiments as well as enhancing the accuracy of SA polarity (Koppel and Schler, 2006).

- If the score is greater than 0, this means the sentence has an overall positive opinion.
- If the score is less than 0, this means the sentence has an overall negative opinion.
- If the score is equal to 0, this means the sentence is considered as neutral.

```
# The first step is to scan positive and negative words into R environment
# The second step is to install the required packages (plyr and stringr) to manipulate text strings
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
  require(plyr)
  # plyr package will handle a list or a vector as an "l"
  # here we want a simple array of scores back,
  # so we use "l" + "a" + "ply" = lapply:
  require(stringr)
  # stringr package provides a function to manipulate characters inside a vector
  scores = lapply(sentences, function(sentence, pos.words, neg.words) {
# The third step is to clean up sentences with gsub() function to remove punctuation, symbols and character:
    sentence = gsub('[[:punct:]]', "", sentence)
    sentence = gsub('[[:cntrl:]]', "", sentence)
  })
}
```

```

        sentence = gsub("\\d+', ", sentence)
# The forth step is to convert text to lower case:
        sentence = tolower(sentence)
# split into words. str_split is in the stringr package
        word.list = str_split(sentence, "\\s+')
# sometimes a list() is one level of hierarchy too much
        words = unlist(word.list)
# The fifth step is to compare our words to the dictionaries of positive and negative words
        pos.matches = match(words, pos.words)
        neg.matches = match(words, neg.words)
# match() returns the position of the matched term or NA, we just want a TRUE/FALSE:
        pos.matches = !is.na(pos.matches)
        neg.matches = !is.na(neg.matches)
# True/False is treated as 1/0 by using the sum() function
        score = sum(pos.matches) - sum(neg.matches)
        return(score)
# The sixth step is to return a data frame of text containing sentiment scores
        scores.df = data.frame(score=scores, text=sentences)
        return(scores.df)
}

```

Figure 6.3: Sentiment analysis code. Source Breen (2011)

In order to provide a better understanding of the potential relationships between Twitter sentiments and diversity variables, we used correlation analysis. Pearson's correlation coefficient is the most popular measure (Jinyuan *et al.*, 2016). Pearson's correlation is a statistical measure used to assess the strength of association between two variables (Mukaka, 2012). However, Pearson's correlation is appropriate when both variables being studied are normally distributed (linear); when one or both variables are skewed (non-linear), Spearman's correlation is used instead (Mukaka, 2012). Spearman's correlation is a statistical measure of a non-linear relationship that addresses the limitation associated with Pearson's

correlation (as the case with Twitter in this research) (Jinyuan *et al.*, 2016). Table 6.2 shows the correlations between Twitter sentiment scores, socio-economic and demographic variables (diversity and standard).

Table 6.2: Correlations between Twitter sentiments, diversity and standard variables

Variables	Twitter sentiment	Ethnic diversity	Age diversity	Educational diversity	Length of residence diversity	Employment diversity
Twitter sentiment		.066	.341**	-.296**	.080	.228*
Significance		.500	.000	.002	.415	.019
		Ethnic minority	Age 16-24	16 over no qualification	Length of residence less < 2yrs	Age 16-64 economically inactive
Twitter sentiment		.110	-.052	.049	.111	-.017
Significance		.263	.599	.621	.259	.861

*correlation significant at 0.05 level (95% confidence)

**correlation significant at 0.01 level (99% confidence)

As can be seen in Table 6.2, there is a significant but moderate correlation between sentiment expressed in Twitter and diversity of age and education. A weak but significant correlation was also observed between ethnic diversity and length of residence in an area with Twitter sentiment. However, correlation between Twitter sentiment and standard variables were not significant.

6.5.1 Sentiment Classification, Accuracy and Validation

The performance or accuracy of sentiment classification is important for the reliability of the overall SA results. One way of evaluating the accuracy of the SA algorithm is to compare it with two or more different algorithms (Pang and Lee, 2004; Sokolova and Lapalme, 2009). Alternatively, we can employ a group of human annotators to manually check a sample of text documents previously classified using an algorithm (Taboada *et al.*, 2011); if more people give the same response on a given text, the probability of accuracy rises (Augustyniak *et al.*, 2015). Although algorithms can achieve some level of accuracy depending on the datasets, validation using human annotation, although sometimes expensive and slow, is

considered to be more accurate and hence more reliable (Rao, 2016). In this research, we employ Amazon Mechanical Turk (AMT), a commercial Human Intelligence Task (HTI) platform that employs 5 different human annotators to validate each tweet. AMT is widely used by researchers for natural language annotation and validation (Snow *et al.*, 2008; Akkaya *et al.*, 2010; Taboada *et al.*, 2011; Botchan, 2012; Shashidhar *et al.*, 2015; Burmania *et al.*, 2016). Overall the results obtained from AMT have been recognised to be of high quality, accurate and reliable (Buhrmester *et al.*, 2011).

6.5.2 Analysis of AMT Validation Results

Section 6.3.2 described Twitter data used in this research. To validate the results of the sentiment algorithm classification, a sample data (about 10%) representing 5,209 tweets were submitted to AMT. There is no acceptable standard sample size for data validation. However, adequate sample size is needed for validation to ensure that the research results are to the acceptable standard as smaller sample size can produce imprecise estimates (Malhotra and Indrayan, 2010), and as the sample size increases prediction improves (Figueroa *et al.*, 2012). The task workers were requested to judge sentiment orientation of tweets from positive to strongly positive and from negative to strongly negative and neutral. The AMT online system has sentiment ratings of +2 to -2 based on which five judgements were given on each tweet.

Manual sentiment annotation has its own challenges, especially where sentences are expressing sarcasm or where sentences are expressing differing sentiment towards multiple entities; such that workers are unsure of how to label them and thus produce inconsistent classifications (Mohammad and Turney, 2013; Mohammad, 2016). Additionally, individual task workers often exhibit high variance in annotation accuracy (Tang and Lease, 2011). To address this problem, researchers have suggested detailed analysis of manual annotation in order to ensure that random and erroneous classifications do not affect the quality of the results being validated, since the workers are not necessarily experts in natural language processing (Callison-Burch, 2009). Moreover, workers had no knowledge of the domain context other than the text content; and that human annotators are also not absolutely perfect in their judgements which can result in different scores (Marge *et al.*, 2010).

6.5.2.1 Analysis of Scores Variability

Accuracy measurement is an important aspect of data validation. Similar to Snow *et al.* (2008), in this research, the quality and reliability of human annotations were assessed using variance analysis. Variance analysis is a measure of the distribution of the extent to a set of

values in the data are spread from the mean (Kruskal and Wallis, 1952). A variance of 0 indicates that all the values are identical, the smaller the variance the closer the results are from the mean and from one another. A higher variance score (>50%) is usually problematic and is an indication of how widespread the annotator’s scores are from one another. Though variance analysis has no specific unit of measurement, in this research we used percentages for easy interpretation such that smaller values indicate closer agreement in the judgements produced by the task workers and larger values indicate poor agreement. Table 6.3 summarises the variance analysis between human annotations.

Table 6.3: Variance analysis between human annotations

Workers variance %	Frequency	Percentage	Interpretation	Cumulative %
0	626	12.03	Perfect agreement	
10-20	1315	25.24	Good agreement	69.74
21-30	1240	23.80	Moderate agreement	(3634 tweets)
31-40	452	8.67	Moderate agreement	
41-50	0	0	Nil	
51-60	506	9.71	Slight agreement	30.26
61-70	494	9.48	Slight agreement	(1575 tweets)
71-80	73	1.40	Poor agreement	
81-90	0	0	Nil	
91-100	150	2.87	Poor agreement	
>100	352	6.75	Poor agreement	

Table 6.3 shows that the task workers have a perfect agreement at 0 percent variance (ratio, 5:0) and a good agreement (ratio, 4:1) with variance range between 10 to 20 percent. Similarly, the results show a moderate agreement (ratio, 3:2) at variance range between 21 to 40 percent and a slight agreement (ratio, 2:3) between 51 to 70 percent variance. However, at higher variance ranges (ratio, 2:1:1:1 to 1:1:1:1:1) 71 and over the agreement was poor. Cumulatively, the workers tend to largely agree on 3,632 tweets (about 69.74%) and disagree in their classifications on 1,575 tweets (about 30.26%). Agreement here refers to cases where 3 or more annotators gave a similar judgement while disagreement refers to where only 2 annotators gave a similar judgement and where they gave different judgements on tweets.

6.5.2.2 Human Annotation Ranking

In order to make comparisons simpler, the average scores of AMT workers were ranked. Ranking enables data sorting into ordered form so as to make interpretation easier (Coe, 2002a). In this analysis, human average annotation scores greater than 0.449 were classified as positive and scores more negative than -0.499 negative; scores between -0.499 and 0.499 were categorised as neutral. Figure 6.4 shows the distribution of AMT classification on positive, negative and neutral ranking.

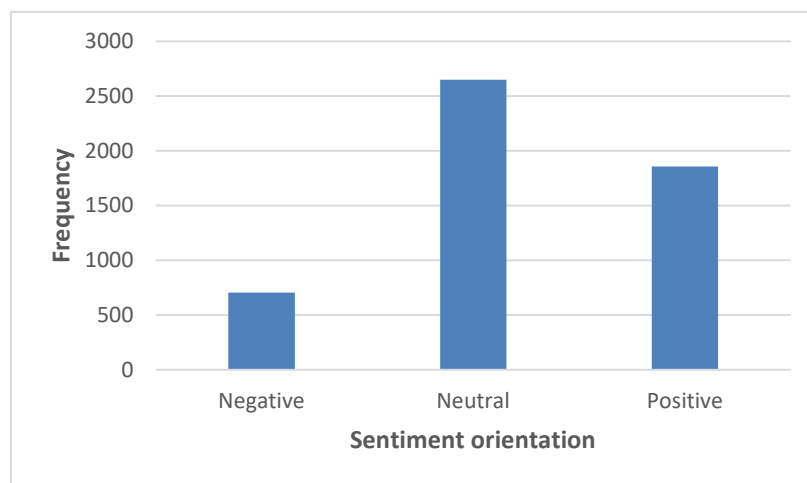


Figure 6.4: Distribution of ranked AMT classification

As seen in Figure 6.4, the majority of tweets 2648 (50.84%) have been classified as neutral sentiments, 705 (13.53%) were negative and 1856 (35.63%) were judged as positive based on rankings of AMT average scores. The process of determining sentiment expressed in a tweet is not an easy task and depends on the subjective judgement of human annotators; often the annotators disagree among themselves (Mozetič *et al.*, 2016).

6.5.2.3 Correlation Analysis

Pearson's correlation analysis was performed to measure the relationship between human annotations and the algorithm classification. Table 6.4 presents Pearson's correlation analysis of the relationships between the AMT validation results and sentiment algorithm results.

Table 6.4: Pearson's correlation analysis of the relationships between the AMT validation results and sentiment algorithm results.

	Algorithm	Human avg.	Answer1	Answer2	Answer3	Answer4
Algorithm						
Human avg.	.629**					
Answer1	.464**	.740**				
Answer2	.415**	.703**	.384**			
Answer3	.452**	.720**	.399**	.370**		
Answer4	.477**	.733**	.430**	.381**	.405**	
Answer5	.484**	.741**	.455**	.400**	.426**	.467**

**correlation significant at 0.01 level (99% confidence)

Correlation analysis (Table 6.4) shows a strong significant relationship between human average sentiment classification and algorithm classification ($r = .629$, $p < 0.001$). A fairly strong significant correlation also exists between the five individual workers annotations and the algorithm scores. For example, the scores of the fifth and fourth annotators show a stronger correlation with the algorithm score ($r = .484$, $p < 0.001$ and $r = .477$, $p < 0.001$) respectively. However, a higher correlation between the individual workers and human average score is as a result of multicollinearity (see Section 5.6.2).

6.5.2.4 Intraclass Correlation Coefficient

In the sentiment analysis literature, estimating reliability and accuracy of human annotation is implemented by quantifying the degree of agreement that exists between the majority of annotators called *inter-rater reliability* (IRR) (Remus *et al.*, 2010; Neviarouskaya *et al.*, 2011; Sameki *et al.*, 2016; Benoit *et al.*, 2016; Tosti-Kharas and Conley, 2016). IRR is a measure of consistency among observation ratings provided by multiple annotators widely evaluated using intraclass correlation coefficient (ICC) (Hallgren, 2012). ICC is a method used to measure the degree of correlation and agreement between raters (Koo and Li, 2016). The range of ICC is between 0 and 1. ICC is usually expressed as Cronbach's alpha α , the value of 0 indicating random agreement and 1 indicating perfect agreement while negative ICC indicates disagreement between raters' scores (Hallgren, 2012). Cronbach's alpha (α) provides a measure of the internal consistency of annotation scores (Cronbach, 1951) as follows:

$$a = \frac{n}{n-1} \left(1 - \frac{\sum_i V_i}{V_t} \right) \quad (6.2)$$

where α is a measure of consistency in agreement, V_i is the variance of scores in item i , V_t is the variance of test scores and n is the number of items.

The acceptable values of alpha should be .7 and above (Santos, 1999; Tavakol and Dennick, 2011; Kehm *et al.*, 2015). However, alpha values less than .5 are indicating poor agreement, alpha values between .5 and .69 are indicating moderate agreement and alpha values .9 and above are indicating excellent agreement (Koo and Li, 2016).

Kappa Index

Cohen's (1960) Kappa (κ) index is also widely used statistical measure for assessing IRR for categorical values. However, Cohen's κ is appropriate where reliability measurement is applied between two annotators (Fleiss, 1971). Additionally, it is difficult to compare κ values with values from other reliability indices such as the Cronbach's alpha (Perreault and Leigh, 1989). To address the limitation of Cohen's κ , Fleiss's κ was proposed. Fleiss (1971) provides an extension to Cohen's κ to that can be used to quantify the degree of reliability between multiple number of annotators, which is appropriate in this research. Fleiss's κ is define as:

$$k = \frac{Po - Pe}{1 - Pe} \quad (6.2)$$

where κ is the chance-corrected agreement, Po is the observed proportion of rater agreement, Pe is the proportion of rater agreement expected by chance and 1 is the maximum value of rater agreement.

Therefore, in this research both Cronbach's α and Fleiss's κ are used to measure the degree of IRR. Landis and Koch (1977) provide a useful benchmark for interpreting Fleiss κ and is adopted in this research (Table 6.5).

Table 6.5: Interpretation of Fleiss's κ . Source: Landis and Koch (1977)

Fleiss κ	Interpretation
< 0	Poor agreement
0.01 – 0.2	Slight agreement
0.21 – 0.4	Fair agreement
0.41 – 0.6	Moderate agreement
0.61 – 0.8	Substantial agreement
0.81 – 1.0	Almost perfect agreement

6.5.2.5 Exclusion Criteria

As in Tosti-Kharas and Conley (2016), in this analysis exclusion criteria were used such that tweets labelled as neutral by the workers were excluded from the analysis. This is because our data are binary (i.e positive and negative) and did not include a neutral class. Neutral tweets do not contain any sentiment polarity (positive or negative), and a neutral class is considered to be objective and less informative so they can be filtered out in order to improve the performance of subjective binary classification (Pang and Lee, 2004) although this has been criticised by Koppel and Schler (2006). Therefore including them to measure the performance of the algorithm classification against human annotation is not appropriate. Additionally, the inclusion of a neutral class could also potentially bias the accuracy and reliability of the results. Furthermore, a recent study has found that the accuracy of binary classification increases when a neutral class is removed from the analysis and decreases otherwise (Bouazizi and Ohtsuki, 2017).

6.5.2.6 Inter-rater Reliability Results

IRR results based on workers scores are presented in Table 6.6. These analyses were performed before and after exclusion criteria were applied to the distribution of the data in order to measure the quality of inter-annotator results and to make comparisons between human annotations and algorithm classification simpler. One of the aims of this research is to explore how social media can be used to explore community cohesion. Therefore the objective of sentiment analysis is to use tweets classified as positive and negative so that we can potentially use them to quantify social cohesion in the communities. Prior to applying exclusion criteria, accuracy and reliability was assessed between the manually classified data

and algorithm classification in order to evaluate the extent to which the results will differ after the exclusion criteria is applied.

Table 6.6: Inter-rater reliability test between human annotators

Inter-annotator			
Fleiss (κ)	Interpretation	Cronbach's (α)	Interpretation
.236	Fair agreement	.775	Good agreement

Table 6.7 show IRR results to quantify the average scores of human annotators and algorithm classification as ternary (positive, negative and neutral) and binary (positive and negative). This is done in order to quantify the performance of automated classification against human ratings. Figure 6.5 shows comparative distribution between the two measurement classifications (binary and ternary).

Table 6.7: Ternary and binary comparisons between human average and algorithm scores

<u>ICC human average vs algorithm</u>	
Cronbach's (α)	Interpretation
.645 (ternary)	Moderate agreement
.769 (binary)	Good agreement



Figure 6.5 Distribution of the IRR analysis between binary and ternary classifications

As seen in Figure 6.5, IRR analysis between ternary and binary classifications indicates that if sentiment scores are classified into binary (positive and negative), the accuracy of validation increases compared to where scores are classified into ternary (positive, negative and neutral). This suggests that excluding the neutral class for IIR assessment can potentially improve the overall accuracy of measurement (see Bouazizi and Ohtsuki, 2017).

In this analysis we used different measures of IRR to assess the accuracy between human sentiment annotation and algorithm classification widely used in sentiment analysis research (Remus *et al.*, 2010; Neviarouskaya *et al.*, 2011; Sameki *et al.*, 2016; Benoit *et al.*, 2016; Tosti-Kharas and Conley, 2016). Specifically, when variance analysis was performed to explore the distribution of human annotation, then the following indices were computed: Pearson's correlation coefficient (r), a measure of linear relationship between two sets of data; Fleiss's κ , a measure of reliability; and intraclass coefficient; and Cronbach's α a measure of degree of internal consistency between multiple raters. Overall, the accuracy of our algorithm compared to human annotation indicates good agreement for binary analysis ($\alpha = .769$) and moderate agreement for ternary analysis ($\alpha = .645$) slightly below the benchmark of .7. Nevertheless, the validation analysis indicates that the results of our sentiment algorithm is good and is potentially useful for further analysis.

However, while different measures of reliability analysis can be used to assess the accuracy of sentiment including human and automated classifications, they have some limitations. Similar to challenges associated with sentiment classifications as highlighted in previous studies (e.g Marge *et al.*, 2010; Mohammad and Turney, 2013; Mohammad, 2016; Mozetič *et al.*, 2016), so also do the quantitative measures such as ICC used to assess their accuracy and reliability (Shrout and Fleiss, 1979). For example, one limitation of ICC as a measure of reliability is that it is dependent on the context in which it is being assessed (Shrout and Fleiss, 1979). Additionally, the degree of heterogeneity of the samples can also affect the accuracy of ICC such that a higher ICC value does not necessarily mean high reliability of results based on some analytical goals (Atkinson and Nevill, 1998; Weir, 2005; Vaz *et al.*, 2013).

Furthermore, ICC is a ratio index of within and between subject variability, where agreement between group of subjects is being assessed in repeated iterations; however, information about individual change or error in scores are not reported (Šerbetar, 2015). Moreover, different types of ICC's *absolute agreement* and *consistency* can produce different results when applied to the same data; therefore ICC values should be interpreted with caution (Koo

and Li, 2016). Despite this limitation ICC is widely used in the social sciences (Vaz *et al.*, 2013).

6.5.3 Twitter Sentiment and its Related Determinants

A number of studies have found relationships between public sentiment on Twitter, socio-economic and demographic factors of different places such as age, ethnicity, employment and education (Mitchell *et al.*, 2013; Bertrand *et al.*, 2013). Gallegos *et al.* (2016) also stressed that geographic locations (places) where people live tend to correlate with sentiments expressed on Twitter. Additionally, where people are happier tend to show more positive tweets, an indication of how places influence human emotions (Gallegos *et al.*, 2016). Furthermore, places where people are more affluent and educated can influence positive sentiments (Volkova and Bachrach, 2015; Lerman *et al.*, 2016); and that areas with high proportions of older people are less likely to share negative emotions than areas with high proportions of younger people (Jalonen, 2017).

One study that used the reported last name of social media users in the US to estimate ethnicity found that 75% users are from the White ethnicity (Sadah *et al.*, 2016). The perception of ethnic diversity is that areas with a higher concentration of immigrants are less likely to have a lower Twitter sentiment score (Hooghe and De Vroome, 2015). In a recent study, Guimarães *et al.* (2017) found that the age distribution of users can influence the way a person expresses sentiment on Twitter. Another study found that users with higher income are less likely to post emotional (negative) content on Twitter (Preoțiu-Pietro *et al.*, 2015), therefore, lower socio-economic neighbourhoods are more likely to use informal language in their conversations on Twitter (Lansley and Longley, 2016). Following previous studies, this research will be testing the following hypotheses that:

- The perception that ethnic diversity in different community areas is related to sentiment expressed on Twitter.
- The age distribution of different communities is related to Twitter sentiment.
- The educational attainment of people has an influence on Twitter sentiment.
- Residential stability in different communities is related in some way with sentiment expressed on Twitter.
- Affluence/wealth of different community areas to has an influence on Twitter sentiment.

To further explore the nature of sentiment classifications (positive and negative) expressed in different community areas, positive and negative sentiment shares in each area is calculated as:

$$S_{p_i} = \frac{ps_i}{N_i} * 100 \quad (6.4)$$

$$S_{n_i} = \frac{ns_i}{N_i} * 100 \quad (6.5)$$

where ps_i and ns_i are the proportions of positive and negative sentiments in an area i , and N_i is the total sentiment scores for area i .

6.6 Sentiment Analysis Results

The result of exploratory analysis (Figure 6.6) shows the word cloud of most mentioned terms contained in the tweets. For example, the mention of places such Leeds, Kirkstall, Otley and Pudsey is an indication of a sense of belonging, that people are identifying with their communities while tweeting. Similarly, words like amazing, happy, love and great are positive words frequently mentioned. Visualisation of tweets provides an exploratory analysis of the content of the tweets in order to explore what people are talking about (Jussila *et al.*, 2013). This can be achieved using visually attractive options such as word clouds (Stojanovski *et al.*, 2014). Font size is used to indicate the frequency, the larger the font size the more frequently the word is used in the collection of document tweets (Miley and Read, 2012).



Figure 6.6: Word cloud of frequent terms

Figure 6.7 shows a histogram of the distribution of the the SA score. The number of tweets and their corresponding sentiment score relating to the whole document are all given. The sentiment scores range from -1 to -7 (negative), 0 (neutral) and 1 to 7 (positive). The length of a sentence (tweet) is important in assigning a sentiment score when it contains words that can increase the orientation to change negatively or positively. Below is an example of some tweets and how the sentiment score is assigned to them:

- Huge[0] thanks[1] to[0] the[0] crews[0] who[0] work[0] across[0] the[0] Kippax[0] amp[0] Methley[0] Ward[0] You[0] keep[0] our[0] villages[0] looking[0] fantastic[2] all[0] year[0] round[0] in[0] all[0] weathers[0] We[0] really[0] do[0] appreciate[1] it[0] (+4, positive).
- Stolen[-1] in[0] a[0] farm[0] burglary[-1] in[0] the[0] Otley[0] area[0] Stihl[0] chainsaw[0] Honda[0] TRX[0] quad[0] Suzuki[0] Picture[0] of[0] the[0] quad[0] attached[0] is[0] of[0] a[0] similar[0] quad[0] Crime[-1] number[0] Anyone[0] with[0] any[0] information[0] please[1] call[0] or[0] Crimestoppers[-1] (-3, negative).
- Photos[0] from[0] our[0] community[0] engagement[0] day[0] at[0] Kippax[0] yesterday[0] (0, neutral)

If a document has a more positive sentiment score than negative sentiment score, it is regarded as having positive content, and negative otherwise (Alistair and Diana, 2005). The results of this analysis indicate that 13.25 % of the tweets are classified as having negative sentiments, 27.36 % as positive and 59.38 % as neutral Figure 6.7.

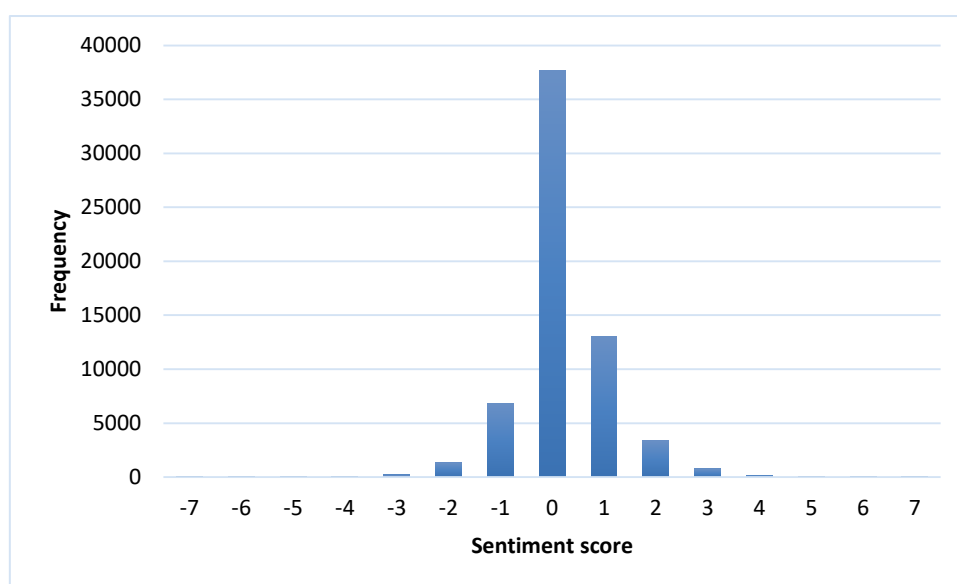


Figure 6.7: Histogram of sentiment analysis score

Figure 6.8 and Figure 6.9 shows the spatial distribution of Twitter sentiment orientation (positive and negative) share in different community areas in Leeds. Holt Park and Hawksworth community areas have higher positive sentiments of 100% and 94%; while Cottingley and Gipton South community areas have higher negative sentiments 83% and 80% respectively. The higher positive sentiment expressed on Twitter in different places is likely because places, where people are happy with their local community, are more likely to comment positively through tweets. The result also indicates that the spatial distribution of Twitter sentiments is mixed and complex.

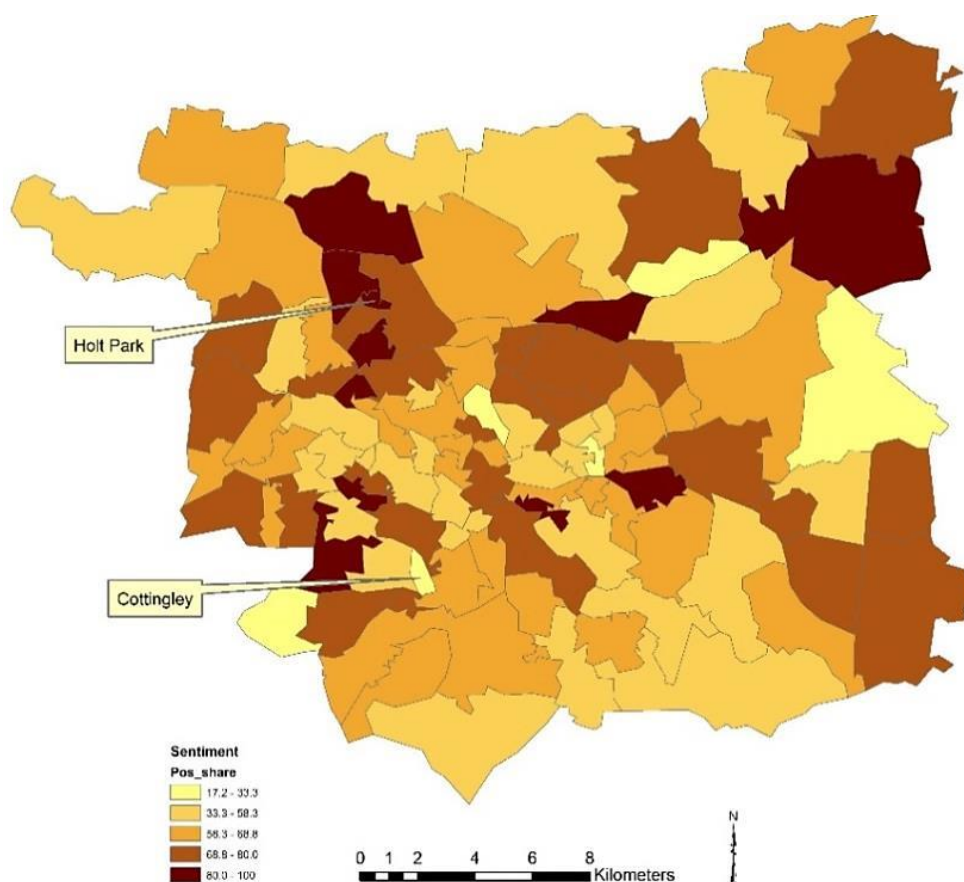


Figure 6.8: Spatial distribution of positive Twitter sentiments in community areas

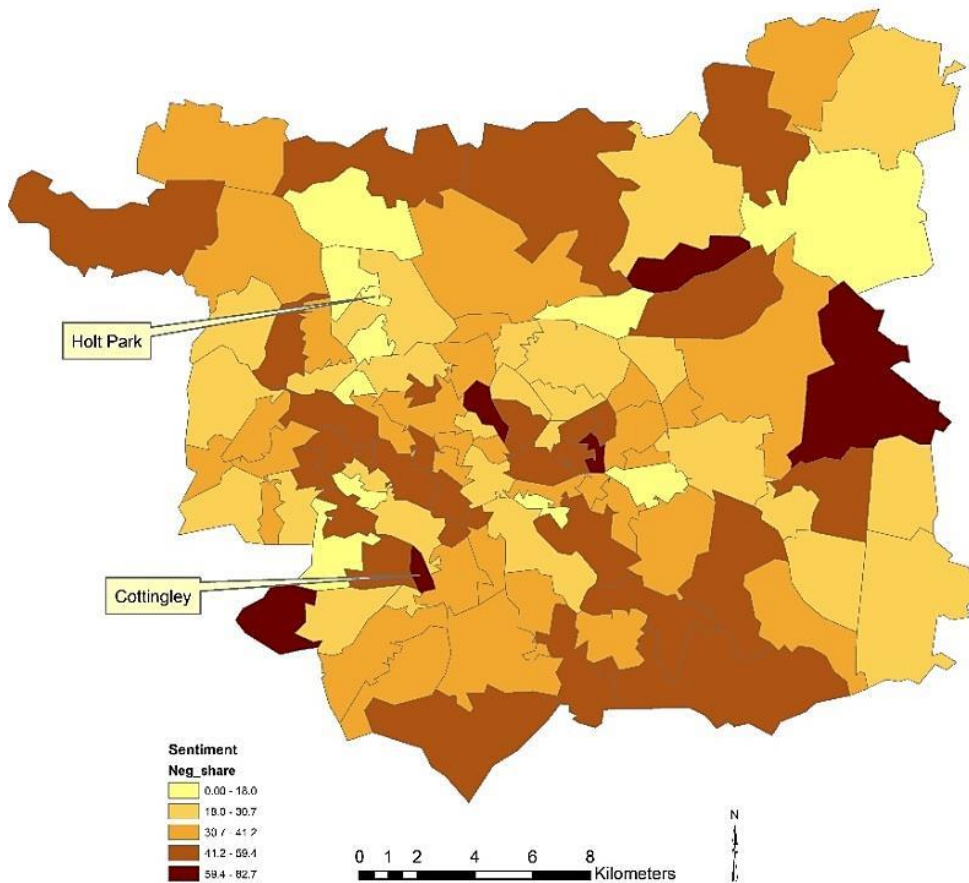


Figure 6.9: Spatial distribution of negative Twitter sentiments in community areas

As seen in Figure 6.8, Holt Park, Bramham, Cookridge, Bramham, New Farnley, Shadwell and Halton & Whitkirk community areas have high numbers of positive sentiments. On the other hand, as can be see in Figure 6.9, Cottingley, Little London, Scarcroft, Scott Hall & Miles Hill, Drighlington and Gipton South community areas have high numbers of negative sentiments. However, despite the complex nature of sentiment expressions, there is an indication of the relationships between demographic attributes of different locations and Twitter sentiment orientation. Similar to the work of Gallegos *et al.* (2016) and Yang *et al.* (2016) that found relationships between Twitter sentiment and demographic areas, in this research we also found a relationship between sentiments expressed on Twitter and different community areas.

Figure 6.10 shows a graph of the percentage share of positive and negative sentiments across Leeds community areas. Holt Park community area has the highest (100%) positive sentiments and lowest negative sentiments, while Cottingley community area has the highest negative (82.8 %) sentiments and lowest positive sentiments respectively.

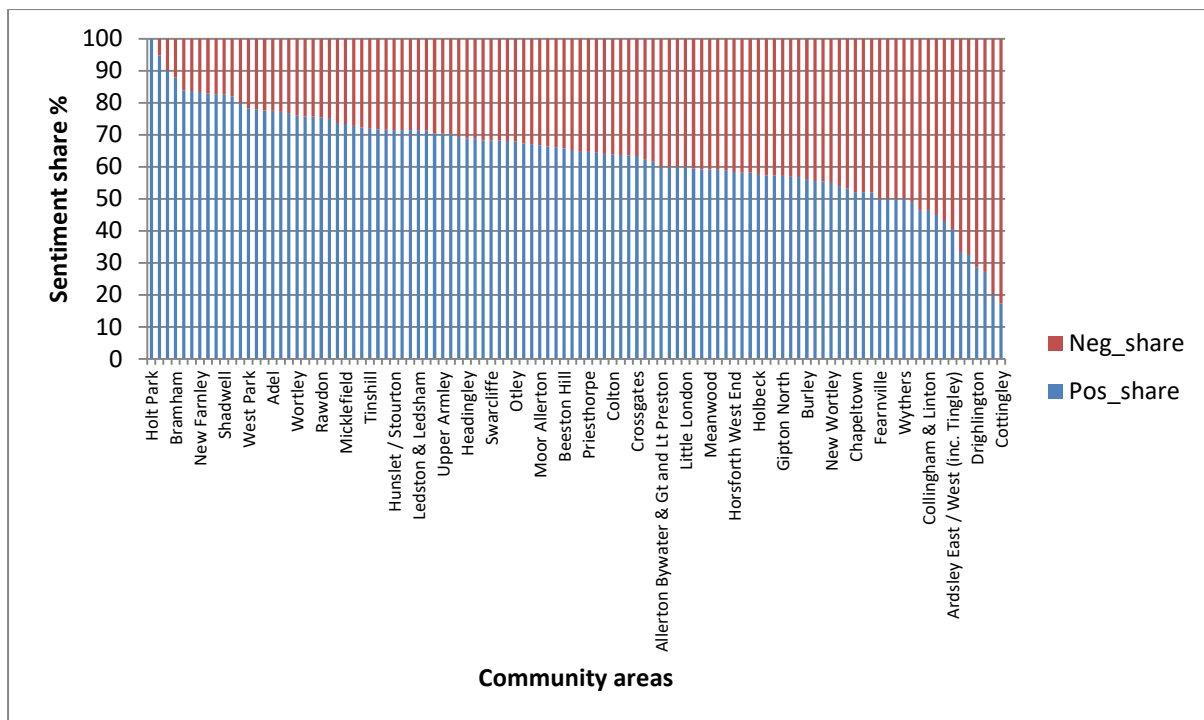


Figure 6.10: Percentage shares of positive and negative sentiments in community areas

6.7 Twitter Sentiment and Diversity Characteristics of Communities

On the scatterplots of the relationships between Twitter sentiment and socio-economic and demographic factors are shown in Sections 6.7.1 to 6.7.5 respectively. Comparisons were made between the standard and diversity variables to examine how each of them is related to sentiment expressed on Twitter.

6.7.1 Twitter Sentiment and Ethnicity

Previous studies have emphasised that socio-demographic composition of different community areas is an important determinant of Twitter sentiments (Turney and Harknett, 2010; Moorhead *et al.*, 2013). The correlations (Figure 6.11 and Figure 6.12) show no relationships between Twitter sentiments and ethnicity variables (standard and diversity). The relationship is weak and not significant, for ethnic diversity $r = .066$ and ethnic minority population $r = -.110$ respectively. On social cohesion, previous studies have demonstrated that ethnic diversity is potentially an important factor in the likelihood that people will connect and form social relationships in different communities (Forrest and Kearns, 2001; Easterly *et al.*, 2006; Letki, 2008; Saggat *et al.*, 2012); therefore it is unlikely that ethnically heterogeneous communities will establish meaningful social ties because members do not trust each other (Gijsberts *et al.*, 2012; Dinesen and Sønderskov, 2015; Tselios *et al.*, 2016).

Social media can potentially reflect this social trend relating to communities (Adolf and Deicke, 2014). Gilchrist and Kyprianou (2011) argued that ethnic heterogeneity has less bearing on how people interact through the internet as facilitated by development in technologies such as social media.

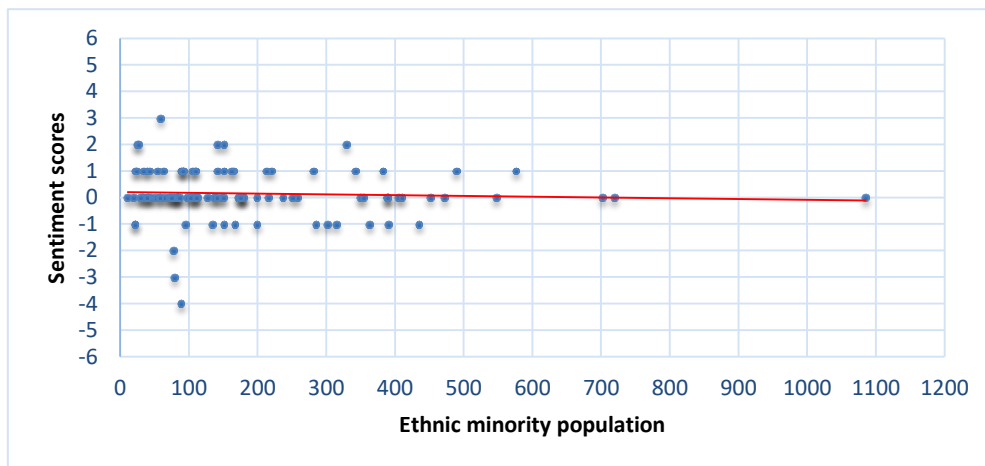


Figure 6.11: Scatterplot of sentiment scores and ethnic minority population

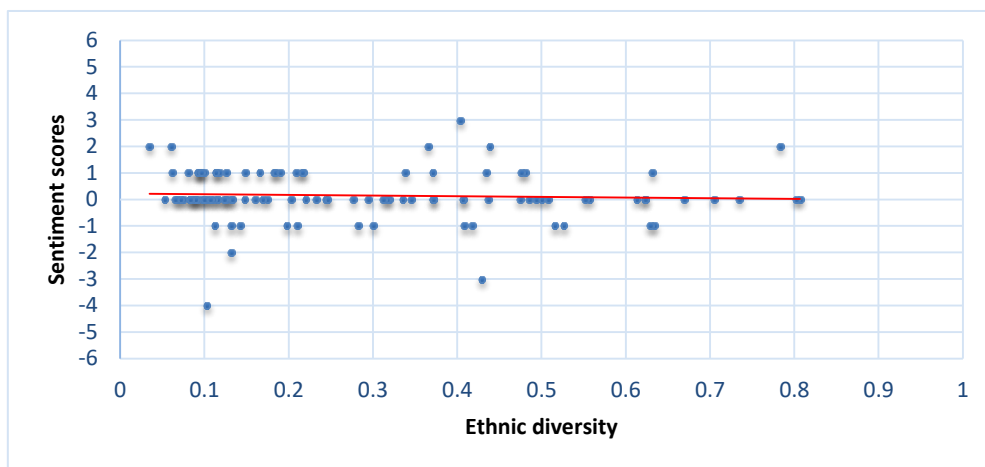


Figure 6.12: Scatterplot of sentiment scores and ethnic diversity

The results obtained in this research are consistent with previous studies that found relationships between the nature of sentiment orientation on Twitter and demographic attributes of different areas (e.g Mitchell *et al.*, 2013; Bertrand *et al.*, 2013). For example, in ethnically more homogeneous community areas of Leeds (Figure 6.13) with a low diversity indices such as Allerton Bywater & Great and Little Preston (0.03), Kippax (0.05), Bramham (0.06), Ledston & Ledsham (0.07), Halton/Whitkirk (0.08) and Otley (0.09) were found to have a relatively larger share of positive Twitter sentiment about 60%, 69%, 88%, 71%, 83%

and 68% respectively; a likely indication of strong social networks in those local community areas. Previous studies have found that ethnically homogeneous communities tend to have more trust with each other and are more likely to have strong social interactions and build cohesion (Laurence and Bentley, 2016). In contrast, it was also found that more ethnically heterogeneous community areas like Scott Hall & Miles Hills (0.6), Gipton South (0.33) and Cottingley (0.40) with higher diversity indices tend to have a higher negative share of Twitter sentiments 73%, 83% and 80% respectively; implying that people are commenting more negatively on Twitter, a likelihood for weak social networks in those areas (Figure 6.13). Previous studies have demonstrated that it is less likely for ethnically heterogeneous communities to trust each other and to establish social ties (Putnam, 2007; Gijsberts *et al.*, 2012; Dinesen and Sønderskov, 2015; Tselios *et al.*, 2016), because community integration is difficult as a result of differences in norms (Laurence and Bentley, 2016). Consistent with previous literature, in this research, we also found support for the relationship between ethnic diversity and Twitter sentiment orientation (Lathia *et al.*, 2012; Quercia *et al.*, 2012; Venerandi *et al.*, 2015; Lerman *et al.*, 2016; Gallegos *et al.*, 2016), though correlation analysis between overall sentiment scores and variables of ethnicity indicate no relationship.

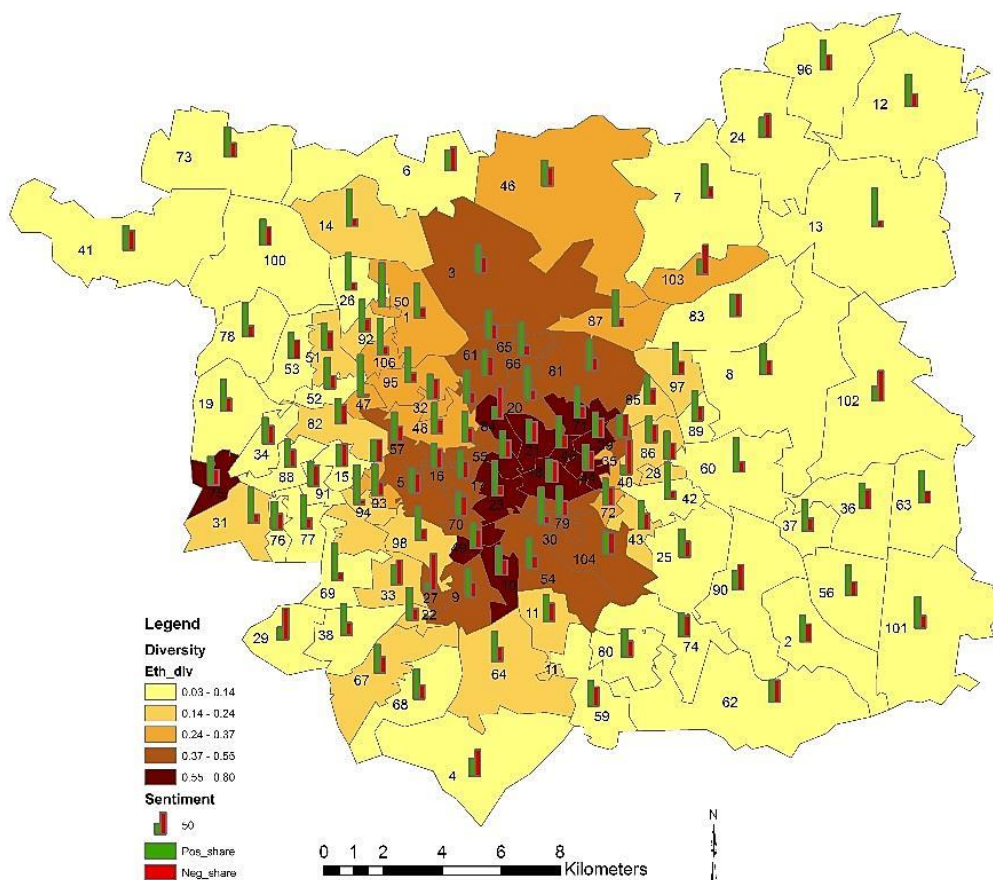


Figure 6.13: Distribution of Twitter sentiments and ethnic diversity in community areas

However, while social media can reflect a number of social trends such as social networks (Adolf and Deicke, 2014), this relationship can also be complex in some instances. For example, in some community areas including Swillington (0.06), Aberford (0.08), Collingham and Linton (0.09) as well as Drighlington (0.09) with low ethnic diversity tend to show higher negative sentiments on Twitter of about 57%, 68%, 53% and 71% respectively, indicating a complex relationship between Twitter sentiment and ethnicity.

6.7.2 Twitter Sentiment and Age Distribution

The correlation between sentiments expressed on Twitter and age diversity is shown in Figure 6.14 and Figure 6.15. The analysis indicates different relationships in the age component. While age diversity shows a significant relationship with Twitter sentiment ($r = .341$, $p\text{-value} < .001$), there was no relationship with the number of young people (aged 16-24) ($r = -.052$) indicating that this relationship is not significant. Previous studies have demonstrated that different age distributions may have very different strengths of social relationships, younger people are less likely to build social cohesion (especially face-to-face) than older people (Johnston and Matthews, 2004; Takagi and Kawachi, 2014).

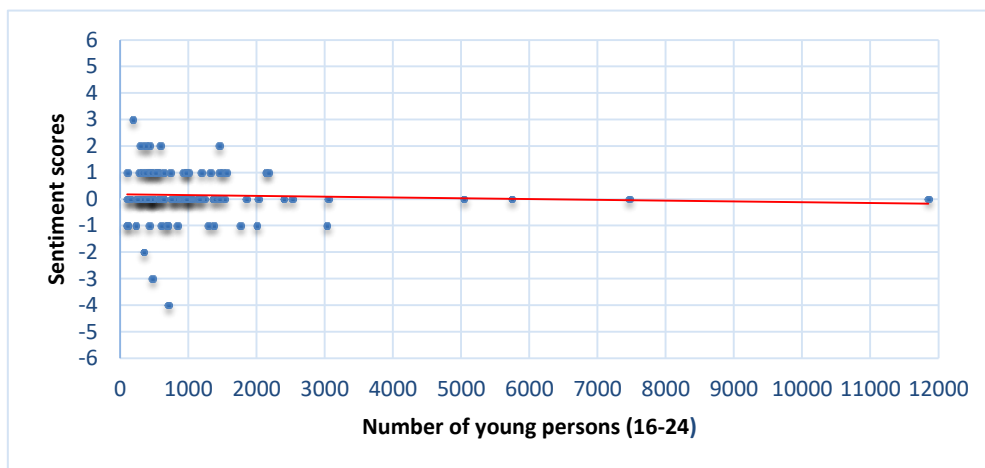


Figure 6.14: Scatterplot of sentiment scores and number of young persons (16-24)

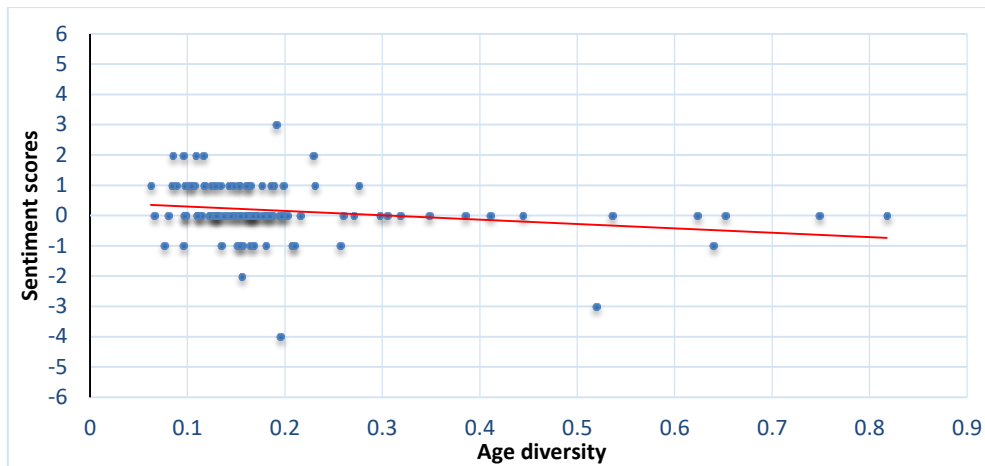


Figure 6.15: Scatterplot of sentiment scores and age diversity

Social media was primarily thought to be the domain of the young (aged 16-24) because they are brought up in the digital era and are considered the first adopters (Bolton *et al.*, 2013). It is more likely that their voices might be over-represented on these channels (Prichard *et al.*, 2015). However, recent statistics in the UK show an increase in adoption of social media by users over 65 years from 15% in 2015 to 23% in 2016 (ONS, 2016b). One reason for this increase is the potential rise in the use of social media channels especially, Twitter and Facebook for social networking among communities (Steinfeld *et al.*, 2008; Teng and Joo, 2017). In this research, we found support for the relationship between age diversity and distribution of sentiment on Twitter (Figure 6.16) but not with the young age group (16-24). For example, community areas with a lower age diversity such as Bramhope (0.06), Cookridge (0.08) and Bramham tend to have a higher positive share of sentiments on Twitter about 83% 84% and 88% respectively; compared to community areas with relatively higher age diversity such as Burley (0.51) and Holbeck (0.41) that have 56% and 57% positive sentiments. The finding suggests that social media is increasingly used in areas with a relatively higher proportion of older people and that their voices are also represented on social media, a factor that recent studies also found (ONS, 2016b; Teng and Joo, 2017). The results also indicate that areas with a higher proportion of older people are more likely to express positive emotions on Twitter than areas with a higher proportion of younger populations (Jalonen, 2017). However, this study also contradicts the findings of Bolton *et al.* (2013) that social media is primarily a domain of the young. It is more likely that the distribution of age within a population is an important factor for social networking and sentiment expression on Twitter as well as maintaining social cohesion. As Hudson *et al.*

(2007) pointed out, age-related division is one factor working against social cohesion in communities.

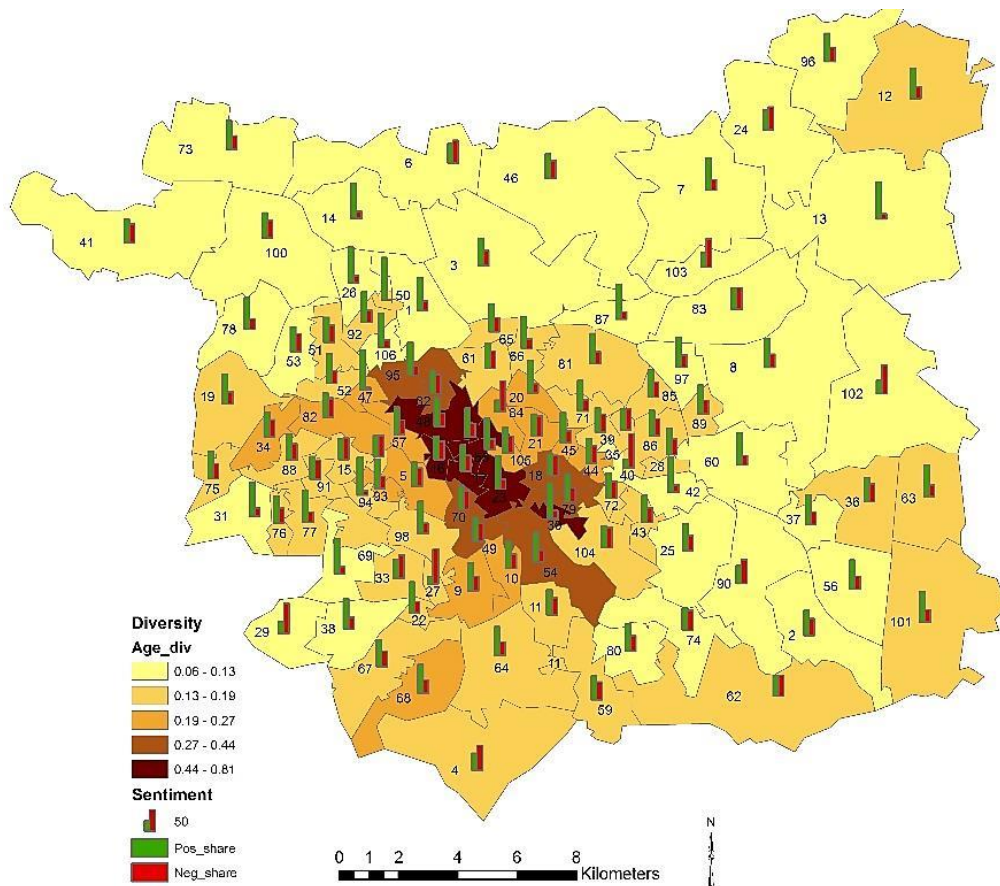


Figure 6.16: Distribution of positive and negative Twitter sentiments and age diversity

6.7.3 Twitter Sentiment and Education

Educational attainment has a great influence on individuals' social behaviour as well as on participation in community activities and has a wider impact on social cohesion (Green *et al.*, 2006; McNab, 2009; Greaves *et al.*, 2013; Gamache-O'Leary and Grant, 2017). Additionally, higher educational attainment can influence social media usage (De la Torre-Díez *et al.*, 2012). Therefore, educated people are more likely to express different sentiments on Twitter (Greene *et al.*, 2011; Park *et al.*, 2011). Figure 6.17 shows the scatterplots of the relationship between Twitter sentiment and diversity of educational attainment; while the relationship between Twitter sentiment and proportion of those with no educational qualification (standard variable) is shown in Figure 6.18 respectively. Correlation analysis shows no relationship between Twitter sentiment and population of those with no educational qualification ($r = .049$); while correlation between Twitter sentiment and diversity of educational attainment is weak but significant ($r = -.296, p < .002$).

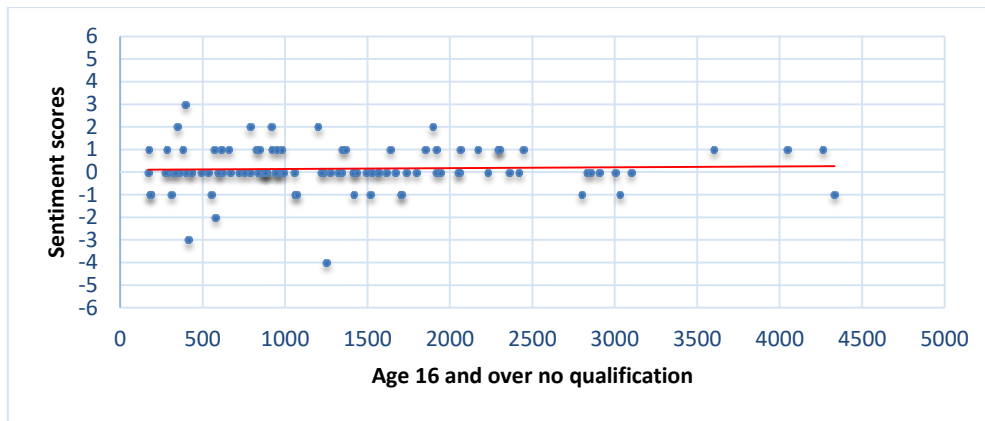


Figure 6.17: Scatterplot of sentiment scores and persons with no educational qualification

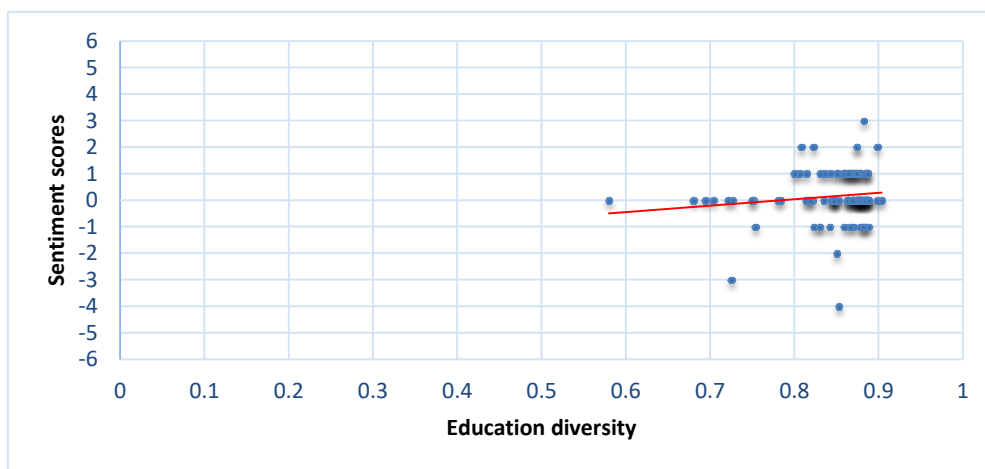


Figure 6.18: Scatterplot of sentiment scores and education diversity

With regard to the relationship between educational attainment and Twitter sentiments, the results of this research suggest links between the diversity of educational attainment and social media usage supporting previous studies (e.g. De la Torre-Díez *et al.*, 2012). Furthermore, like in previous studies, for example, Greene *et al.* (2011) and Park *et al.* (2011) that have found relationships between sentiments expressed on Twitter and an individual level of education. In this study, areas with a higher diversity of educational attainment were found to show a higher positive share of sentiment expressed on Twitter (Figure 6.19). Community areas in this category include: Holt Park (0.88), Hawksworth (0.87) Little London (0.89) and Halton/Whitkirk (0.88) diversity indices tend to have positive sentiment share of about 100%, 95% and 83% respectively; while areas with relatively lower education diversity such as Burley Lodge and Little Woodhouse (0.72), Far Headingley (0.70) and Hyde Park (0.57) tend to have lower positive Twitter sentiment share of 62%, 55% and 68% respectively. Studies have found links between educational attainment and participation in community activities which also have wider impacts on social cohesion (Green *et al.*, 2006;

McNab, 2009; Greaves *et al.*, 2013; Gamache-O'Leary and Grant, 2017). However, we also found that Little London community area with higher education diversity (0.89) has a lower positive Twitter sentiment. It is likely that the heterogeneous nature of this community area to influence emotion expressed on Twitter.

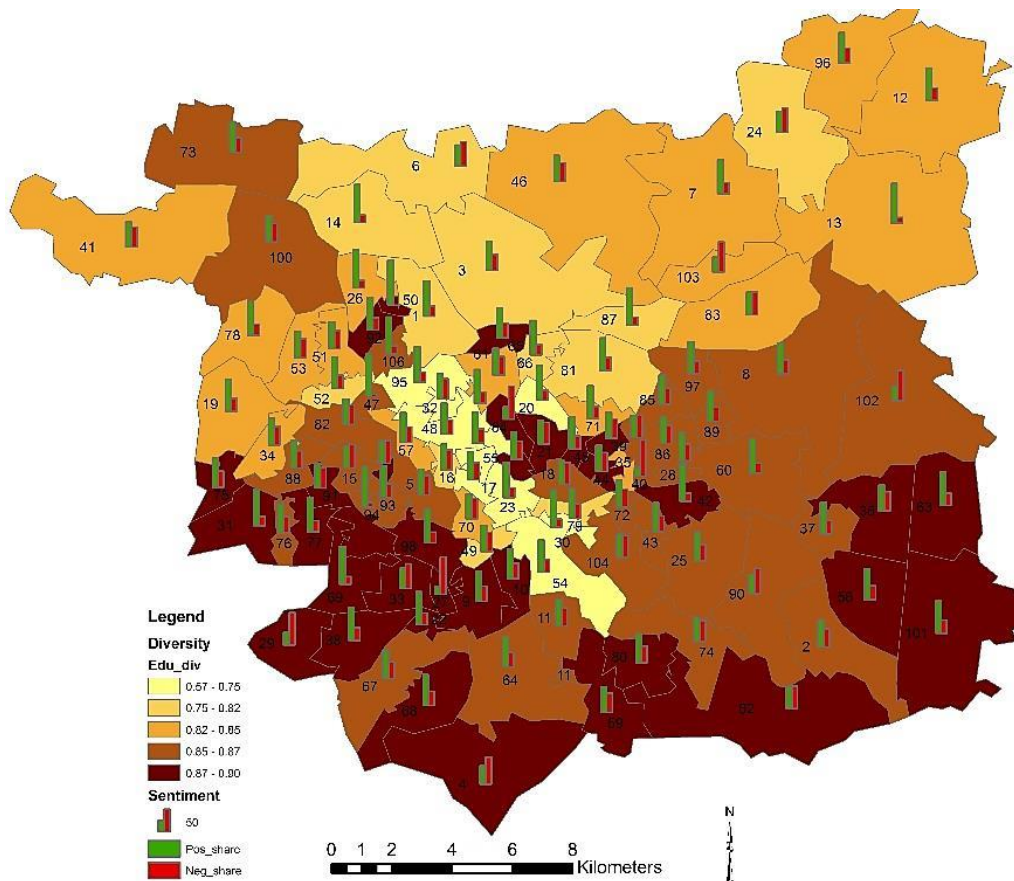


Figure 6.19: Distribution of positive and negative Twitter sentiments and education diversity

6.7.4 Twitter Sentiment and Length of Residence

Figure 6.20 and Figure 6.21 shows the correlation between length of residence variables (standard and diversity) and Twitter sentiment. Spearman's correlation shows weak relationships between Twitter sentiment and duration of residence. The results of this analysis reveal that diversity of length of residence is not significant ($r = .080$). The correlation is also not significant with the population of those that lived for less than two years in an area ($r = -.111$). Studies on social cohesion demonstrate that social bonds are built over time especially when people reside in a particular neighbourhood for a longer period of time (Yamamura, 2011; Keene *et al.*, 2013). Residential stability is important for building social cohesion in communities (Thomas *et al.*, 2016). It is likely that Twitter sentiments do not reflect this connection.

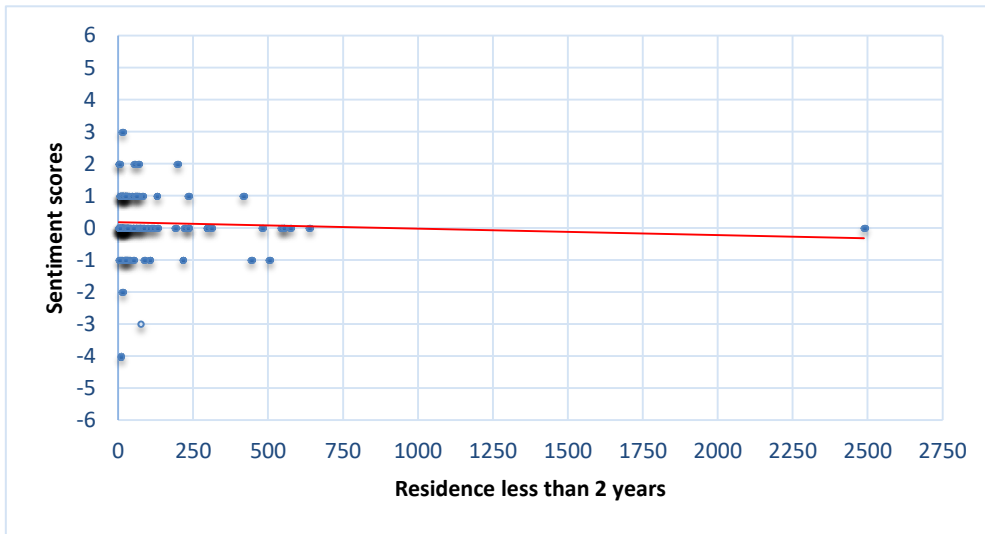


Figure 6.20: Scatterplot of sentiment scores and length of residence less than 2 years

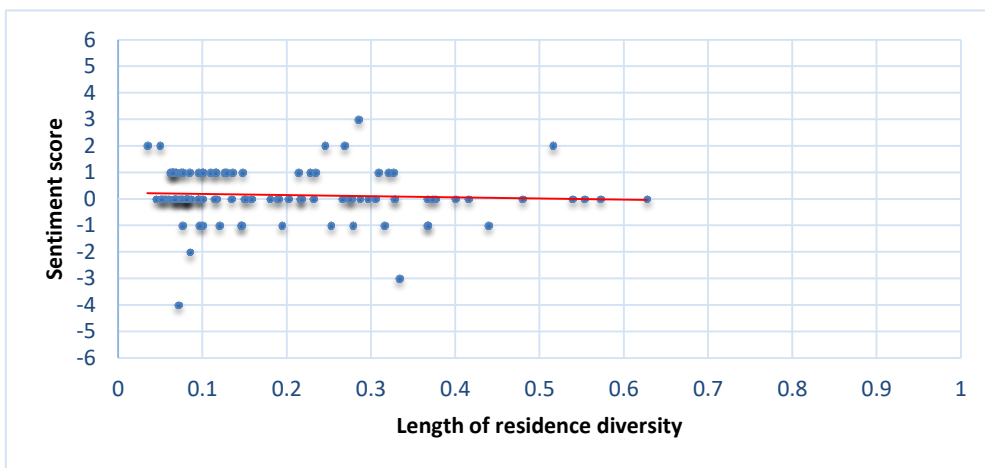


Figure 6.21: Scatterplot of sentiment scores and length of residence diversity

In this research, no statistically significant relationship was found between Twitter sentiment and diversity of length of residence (Figure 6.22). Nevertheless, interestingly it was also found that some community areas with lower diversity of length of residence such as New Farnley (0.07) and Halton/Whitkirk (0.06) tend to have a higher positive Twitter sentiments about 83% and 82%; however, in residentially less stable community areas like Scott Hall & Miles Hill (0.36) and Chapeltown (0.53) tend to show slightly lower positive Twitter sentiments of 27% and 52% respectively; suggesting that residential stability is likely to be important for social media interactions in some way. A number of studies have demonstrated the importance of residential stability for the creation of social ties in different communities

(e.g. Yamamura, 2011; Keene *et al.*, 2013). Migrant neighbourhoods are considered to be less committed to local neighbourhood activities because of their social mobility characteristics (Wu and Logan, 2016). Therefore, it is more likely that residentially stable areas to have a better established bonds of social relationships than residentially unstable areas (Sampson, 1988; Turney and Harknett, 2010), because length of residence in an area increases the potential for creating friendships and strengthens social ties with neighbours (Anton and Lawrence, 2014). online social media networking can also support the maintenance of existing social ties (Ellison *et al.*, 2007). In the context of social media and Twitter specifically, Gallegos *et al.* (2016) stressed that correlations exist between geographic locations (places) where people live and sentiments expressed on social media. However, this does not mean that the residential status of people can influence the nature of sentiments expressed on Twitter.

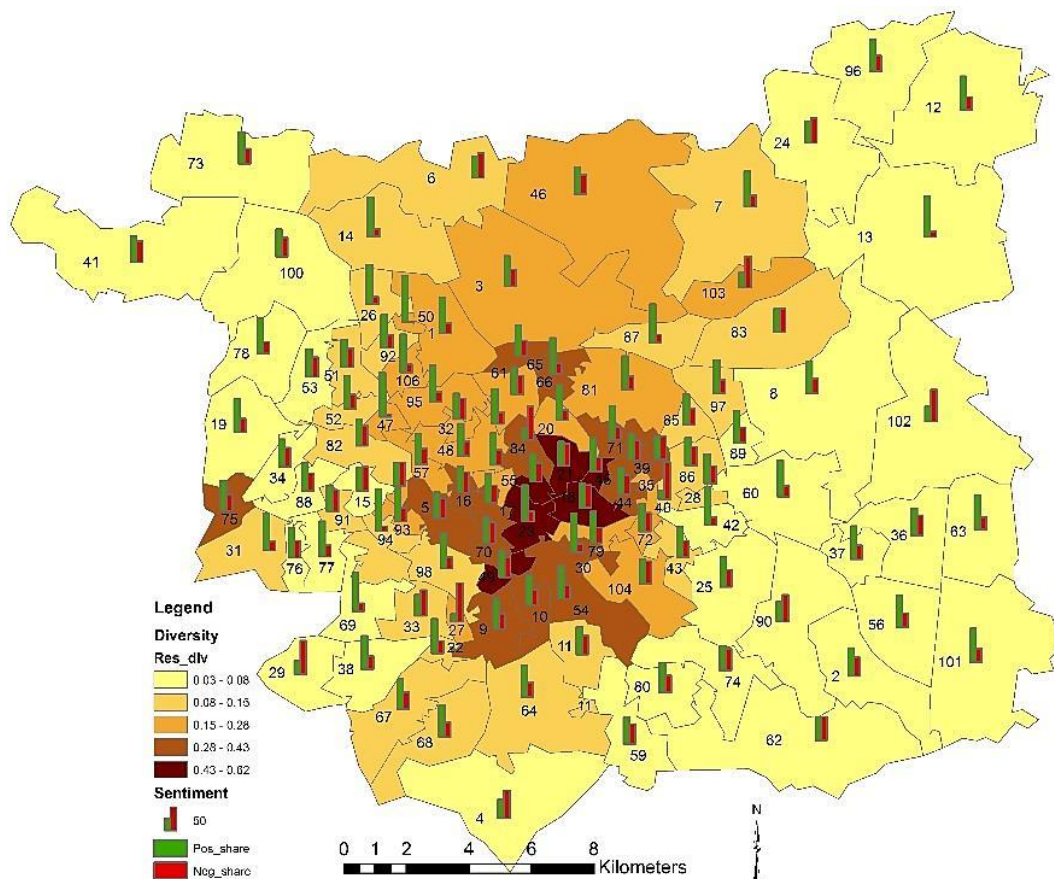


Figure 6.22: Distribution of Twitter sentiments and length of residence diversity

6.7.5 Twitter Sentiment and Employment

In terms of social cohesion, studies have demonstrated that more affluent communities tend to be happier and are likely to be more cohesive than less affluent communities (Delhey and

Dragolov, 2016; Ejrnaes and Greve, 2017). While the nature of our jobs determines who we are in society, unemployment can cause depression with profound effects on the social stability of families as well as decreasing potential for participation in communal activities (Norton and de Haan, 2012). Income deprivation tends to widen the gap between the rich and poor which can reduce social cohesion (Morenoff *et al.*, 2001). Social media can also reflect this relationship, as highlighted in recent studies (e.g. Lerman *et al.*, 2016; Greenwood *et al.*, 2016) that have shown that relationships exist between Twitter sentiment and affluence. Figure 6.23 and Figure 6.24 show correlations for this relationship. For the economically inactive population, the relationship is not significant ($r = -.017$) while correlation between Twitter sentiment and diversity of employment is significant $r = .228$, $p\text{-value} < .019$. According to Volkova and Bachrach (2015), affluence/wealth can influence Twitter sentiment, it is likely that diversity of employment in an area is important for this relationship.

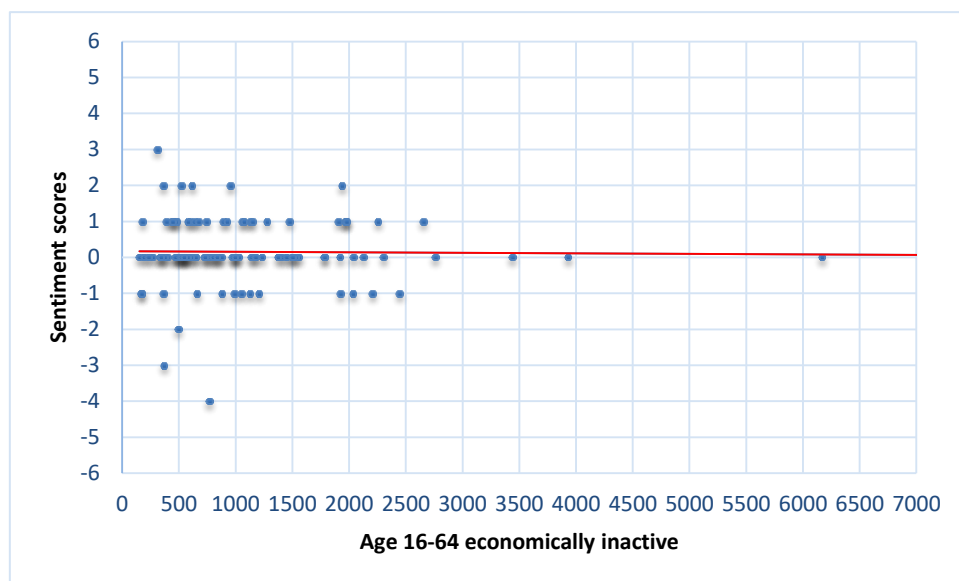


Figure 6.23: Scatterplot of sentiment scores and employment diversity

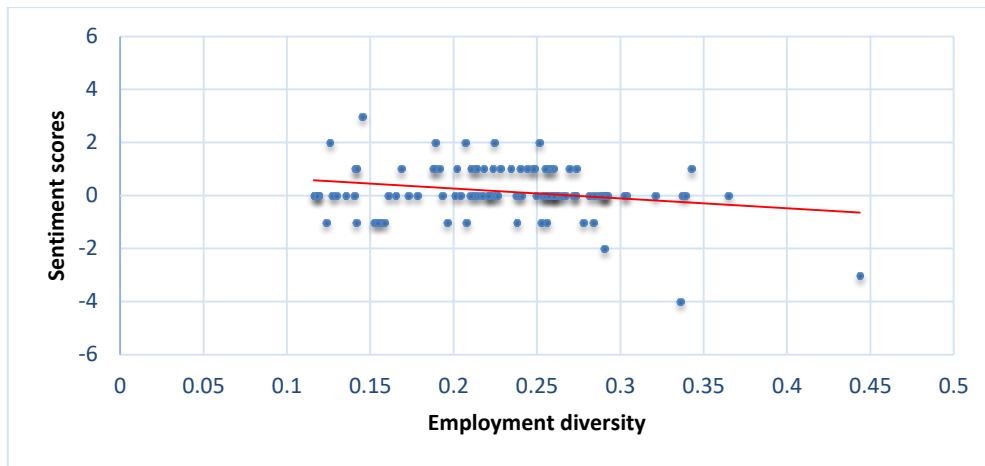


Figure 6.24: Scatterplot of sentiment scores and employment diversity

In this research, diversity of employment types is used as a proxy for affluence/wealth in the absence of income data in the UK Census. The result of this analysis indicates a relationship between diversity of employment and Twitter sentiment, but not with the economically inactive population. For example, while community areas with relatively high employment diversity such as East Bank (0.44) and Churwell (0.32) has about 83% and 74% positive sentiments; some areas of lower employment diversity like Woodhouse (0.1) and Upper Wortley (0.1) also tend to have higher positive 75% and 90% respectively; this suggests a complex relationship between Twitter sentiment and affluence (Figure 6.25). Our findings, however, are in contrast to that of Lerman *et al.* (2016) who used US Census data where mean household income is available as a variable to quantify relationships between Twitter emotion and socio-economic factors of places. Although the results of this research support the findings of Greenwood *et al.* (2016) who indicated that those who are employed are more likely to use social media, especially Twitter, than the unemployed. Recent studies have demonstrated that more affluent communities tend to be socially more cohesive than the less affluent communities (Delhey and Dragolov, 2016; Ejrnæs and Greve, 2017). It is more likely that affluence can influence sentiments on Twitter.

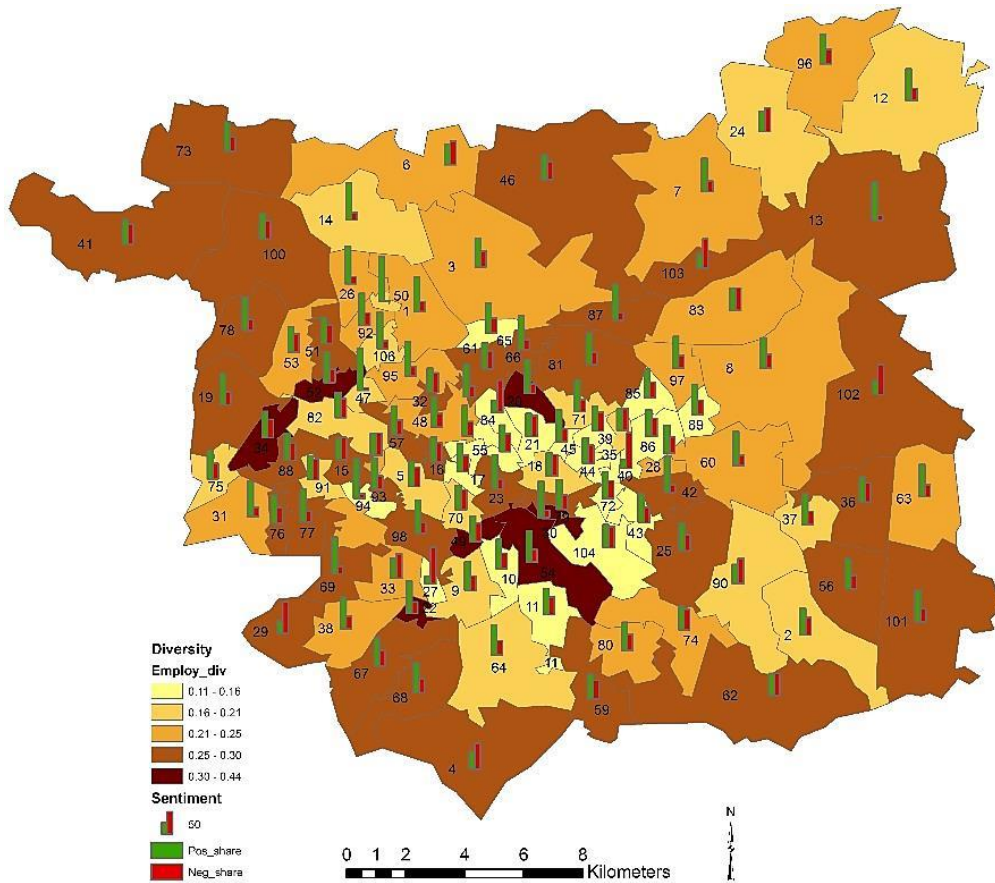


Figure 6.25: Distributions of Twitter sentiments and employment diversity

6.8 Concluding Remarks

This chapter has explored with new social media data (Twitter specifically) how public sentiments on Twitter along with socio-demographic attributes of different community areas derived from Census datasets can be used for understanding community social ties (cohesion). As indicated in the results, Twitter sentiment analysis is potentially useful for understanding the relationships between socio-economic and demographic characteristics of different areas and how they can potentially influence Twitter emotions. Section 6.7 discussed in detail the relationships between Twitter sentiments and diversity of areas and highlighted the important role social media networking especially Twitter can play to better understand social cohesion in different areas. The findings of this research demonstrate that new social media data are potentially useful for uncovering underlying patterns of social relationships in local communities, especially when they are used in conjunction with more trusted “traditional” Census data. However, with the acceptance that social media data obtained from Twitter users may not be a strong representation of the general population. This notwithstanding, however, innovating with new forms of social media data through

sentiment analysis of public comments on Twitter can potentially be useful for exploring community cohesion which is one of the objectives of this research. Chapter 7 will further extend our understanding of how we can leverage social media using Facebook data to quantify local community engagements. Finally, insights gained from the social media analyses in chapters 6 and 7 will be combined in later regression modelling work in chapter 8 with a view of quantifying their relationship with community cohesion and crime.

Chapter 7

New Social Media Metrics for Quantifying Community Cohesion and Crime

7.1 Introduction

From the outset, an objective of this research has been to investigate the feasibility of the new social media data sources, especially Facebook and Twitter, to explore degrees of community cohesion. Chapter 6 has provided insight on how sentiment analysis of Twitter data can be used to explore social cohesion in relation to the diversity of different community areas. This chapter will build on the previous chapter using publicly available Facebook metrics in order to extend our understanding of community cohesion and crime. Section 7.2 describes the importance of social media for community engagement; Section 7.3 describes Facebook pages and groups. Section 7.4 describes Facebook data collection procedures and Section 7.5 will explore how community engagement can be quantified on Facebook. In Section 7.6 multiple regression models of community engagement on Facebook and crime will be explored using the new metrics of interaction and, finally, concluding remarks are given in Section 7.7.

7.2 Understanding Community Engagement on Social Media

The emergence and development of information and communication technologies (ICTs) and the increasing and diverse uses of the internet are changing the way communities are engaging and interacting in the public sphere (Bonsón *et al.*, 2015). This can be particularly seen with the evolution of social media platforms (Agostino and Arnaboldi, 2016). Social media (SM) refers to networked communication platforms (e.g. Facebook and Twitter) that allow users to connect with other people and share user-generated content (Ellison and Boyd, 2013). The term “engagement” refers to the participation of community members in social activities aimed at community building on SM channels, especially Facebook (Gil de Zúñiga *et al.*, 2012). One common feature of SM engagement is that it is an important source of community information and education (Bonsón *et al.*, 2014). It can also facilitate real-time public interactions making collective decisions faster and easier (Agostino and Arnaboldi, 2016), as well as promoting community building and encouraging off-line social relationships (Williams, 2010). Facebook in particular can be a potentially powerful tool in this regard (Warner *et al.*, 2010).

Furthermore, SM tools such as Facebook and Twitter also have a great potential for community-based crime reduction efforts (Featherstone, 2013). For example, SM can be useful for mobilising community support and encouraging community action, especially in solving local problems such as crime (Humphreys, 2010; Baruah, 2012; Kawash, 2014). In this way, social media channels, especially Facebook and Twitter, can serve as a means of communication, supporting one another and improving local community participation (Maurer and Wiegmann, 2011). However, Harris and Flouch (2012) argued that establishing well-developed social media networks in disadvantaged areas is difficult, because residents of some disadvantaged areas face challenges with inadequate internet access which can sometimes be related to the economic status of the people which in turn affect the development of neighbourhood social media networks (McCabe *et al.*, 2013; Townsend *et al.*, 2015).

Government agencies such as the police and local councils have also recognised the importance of SM engagement with the community so as to better understand their needs as well as promoting public safety (Zavattaro and Sementelli, 2014). Additionally, SM provides a new way of interacting between the police and public, building confidence and relationships as well gathering information and intelligence from local communities (Foundation, 2014). A study on the police use of SM has shown an increasing adoption of these channels, especially Facebook and Twitter, by the police in the UK for engaging with communities (Fernandez *et al.*, 2017). While SM is widely accepted by communities and organisations for enhancing engagement, its application to the study of community cohesion (especially on crime) is limited (Aarts *et al.*, 2012; Wallace *et al.*, 2017). This research has employed Facebook metrics to quantify community engagement in order to gain insight into community cohesion and to explore crime. Detailed reviews of SM, community cohesion and crime are provided in Sections 4.3 and 4.4.

7.3 Facebook Pages and Groups

7.3.1 Facebook Pages

A Facebook *page* is a public profile created for the purpose of interacting with the public (e.g. community, business and organisation). Unlike a personal profile, a page does not have “friends” but “fans” usually displayed on a page profile. *Fans* are people who choose to follow or like a particular Facebook page and receive updates from them (Hicks, 2010). There are six categories of Facebook pages: *Local Business* (e.g. bar, hotel, museum, book

store and banks) *or Place* refers to Facebook pages used for promoting business for customers to make a physical visit; *Company* (e.g. travel and insurance company) *or Organisation or Institution* (e.g. publishing and Church) refers to Facebook pages used where an organisation has multiple locations; *Brand or Product* (e.g. electronics, clothing and furniture) are Facebook pages intended for products selling through multiple outlets; *Artist, Band or Public Figure* are Facebook pages used for promoting specific persons, musical bands or political figures; Facebook pages commonly used for sports, movies and radio stations are categorized as *Entertainment* pages; and *Cause or Community* pages (e.g. community or neighbourhood, education and religious organisation) are particularly for non-profit outfits (Harry, 2016). Furthermore, a community Facebook page enables users to connect and share content dedicated to a topic of interest and enhancing social interactions and collective community actions (Smith, 2010). Figure 7.1 shows different categories of pages available on FB website.

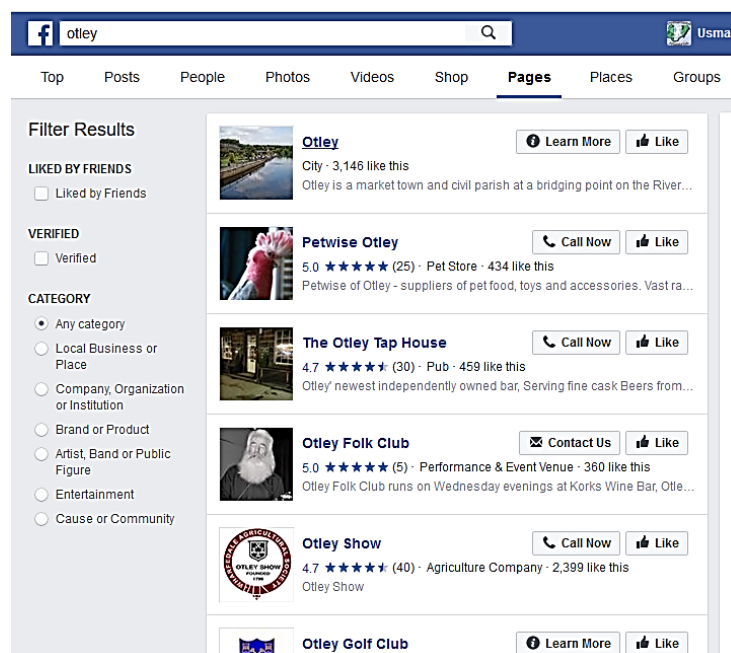


Figure 7.1: Typical Facebook pages related to the word ‘Otley’ (Otley is a town in West Yorkshire)

7.3.2 Facebook Groups

Different from Facebook pages, the *Groups* option is also available to users. Groups allow Facebook users to organise around a common cause, discuss and share content based on different interests (Parsons, 2013). Unlike Facebook pages, Groups can be public or closed and vary across different categories such as neighbourhood, community, friends, buying and

selling and sports club supporters (Hicks, 2010). Figure 7.2 shows the type of Facebook groups.

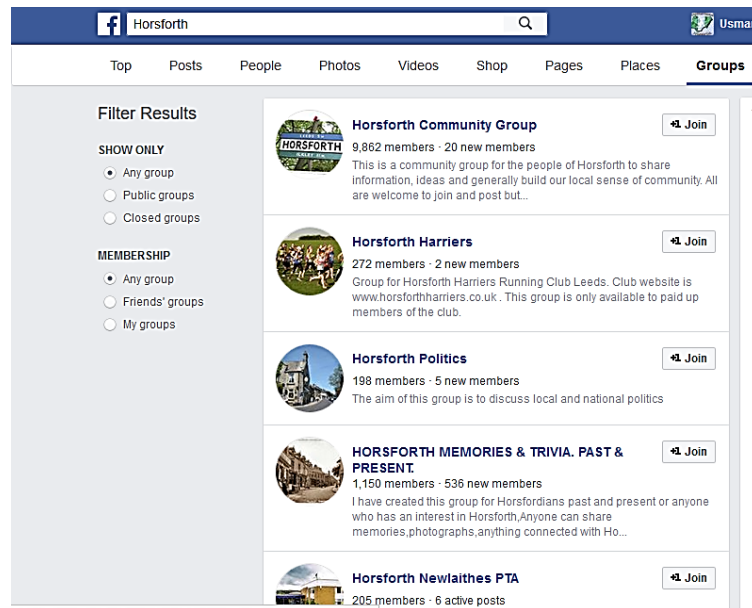


Figure 7.2: Typical types of Facebook groups in ‘Horsforth’ (a neighbourhood in Leeds)

In this research, community Facebook pages will be used to explore social media engagement as a proxy for community cohesion. Pages are chosen because they are widely used in different community areas (Parsons, 2013) and have a greater potential of reaching many people than Groups established for specific interests (Bonnand, 2015). However, some communities do not maintain Pages, and in such situations, Groups will be used instead. Choosing a Facebook Group that represents a community interest is a difficult task because a particular community can have different Facebook Groups. For example, since the last search (20/09/2017) on community Facebook Groups in Horsforth (Figure 7.2), there is now a ‘New Horsforth Community Group’. Therefore, Horsforth residents could use either or both. Additionally, privacy concerns sometimes make access difficult as most of them are “closed” and yet the few “public” ones are used for buying and selling of items or connecting with school mates (Section 7.5 provides statistics on groups).

In this research, a number of criteria were employed for selecting Facebook Groups. The first criterion for selecting a community group is to check whether it is public, meaning that data for that particular group is publicly accessible by default. Facebook has made it easier to find public groups by allowing users to filter these groups in searching for results on Leeds Community Areas. Once privacy level has been established, the second criterion is to check the name of the group. For example, ‘Mickelfield community’, ‘Crossgates yesterday, today

and tomorrow’, ‘Moortown memories’ and ‘Thorner community’ are all groups that attempt to represent the community as a whole rather than those formed for a specific interest. The third criterion involves the number of members. In this regard, a community group with the largest numbers of members is usually preferred but where this is not the case, because of privacy issues the second largest group is selected. The community area group with the minimum number of members is Halton Moor (144), while Crossgates community group has the maximum membership (6,018) respectively.

7.4 Facebook Data Collections

Facebook allows anonymous access to publicly available data on public pages and groups by obtaining the necessary authorisation via the Application Programming Interface (API) from the Developers website⁷. The first procedure involves searching for the names of 106 different community areas in Leeds on Facebook search facility via an existing account. The search returns about 15,000 pages and further refinement was conducted to retain only geo-located pages (n = 8,122) and groups (n = 973) that are within community boundaries. The majority of the group accounts 595 (61.2%) assessed were found to be closed, while the remaining 378 (38.8%) were public. In all, 67 community areas were found using Pages while 27 are using Groups for community interaction and 12 community areas do not have either a page or group. Figure 7.3 shows the distribution of Facebook pages and groups in different communities (Figure 3.4 provides interpretation for the community codes). A total of 94 community areas representing 89% of the total community areas in Leeds are using Facebook for community engagement. The sample is considered to be large enough for analyses in the present research. A sample size of about 69% is recommended in social science because the larger the sample size, the lower the error margin expected and the more the accuracy of the results (Tomczak *et al.*, 2014). The error margin is a statistical term used to express the confidence interval in a research project (Gilliland and Melfi, 2010). Furthermore, because the number of the community areas covered cut across every part of Leeds District spatially, it is believed that there could be potentially similar patterns in characteristics as a result of socio-demographic similarities of areas (from the perspective view of the *First Law of Geography*, near things are more related than distant things (Tobler, 1970). Therefore, they can potentially be used as a representation of the entire community areas (those community areas with no Facebook presence).

⁷ <https://developers.facebook.com/>

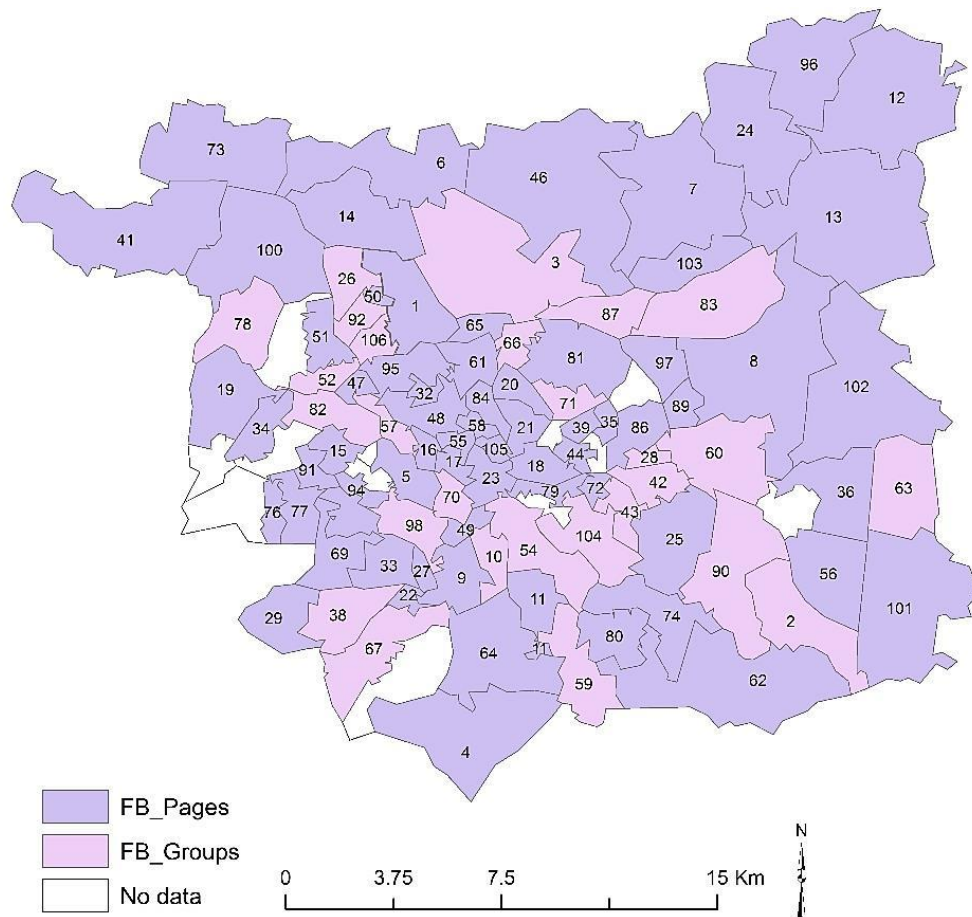


Figure 7.3: Distribution of Facebook pages and groups in different communities

In order to ensure that all pages used are under the appropriate category, we used the filter tool from the search facility to select public Facebook pages that fall under the category of “community”; and then obtain an identification (ID) number for each page using an online facility for checking ID numbers of Facebook pages and groups⁸. An ID is a unique number assigned to every Facebook user account which is used to identify it. A similar procedure was also used to obtain the group’s user IDs. Having obtained the ID numbers for each community page and group, we then proceed to collect their timeline data using the Rfacebook package, an algorithm based on Barbera *et al.* (2017), typically used for scraping data from Facebook in R environment version 3.2.5 (R Core Team, 2016). Table 7.1 shows the metrics available on a typical Facebook page.

⁸ <https://findmyfbid.com/>

Table 7.1 Typical metrics available on a Facebook page

Message	Type	Likes count	Comments count	Shares count
If anyone knows the Clown with the exploding car thats driving round whinmoor to nt , please send him my way so i can stick his head up his own arse , you're not big and defo not clever . we have your number .	status	23	23	1
Hello all. I'm at the end of my tether, and I'm after some help please! My property has a Pulsacoil 2000 thermal store (boiler) and it constantly is breaking. My landlord is struggling to fix it as it's such a specialist piece of equipment. Would anyone be able to recommend an electrician/plumber that might have experience of these systems, that I could contact please? Thankyou in advance.	status	2	4	0
Thursday 21st September 2017 marks the 50th anniversary of the Crossgates Arndale Shopping Centre. ???? https://www.facebook.com/yourcrossgates/posts/1443772859025858	photo	13	4	3
Please come and support our sister units in the Caribbean rebuild their lives...Four of Girlguiding's five branches in the Caribbean have been hit by Hurricane Irma, three of them directly. The levels of damage and the loss of life across the islands are unprecedented.	photo	7	0	0
A trip down memory lane. I've been scanning a lot of negatives recently (that have revealed some forgotten gems) and this is one of them. Taken circa May 1983 at Crossgates station. Almost everything has changed here - how many can you spot? :-)	photo	29	13	0

Hi! I'm wondering if anyone can help? My son has broken his iPad tonight and is devastated! He uses it a lot for studying and this year is his gcse year. Does anyone have one they no longer want or need? If so, please let me know how much! Thanks in advance!! Xxx	status	4	12	2
Looking for a cheap but reliable electrician in or around Crossgates to fit a new light switch in a bedroom (old dimmer switch broken - just need replacing with new normal switch) and a new double plug socket in a kitchen (one side doesn't work) tia.	status	1	4	0
Does anyone know of any child minders that can do morning school run in the wykebeck area any info would be hugely appreciated	status	3	6	1
In Crossgates:	link	5	9	17
If this is your skill and attitude to driving/parking then you shouldn't have a licence !! Seen 2 people do this today in crossgates.... Your putting other people in danger by parking half way across the road!!	photo	13	32	0

The UK template for SM suggests a minimum of 2 and a maximum of 10 posts per day for each official account which gives a total of 3,600 posts per account per year (Williams, 2009). Though not all pages return such a number of posts, some communities tend to exceed that total. Facebook timeline posts data were collected for different community areas between 28th August to 14th November 2017 (N = 85,000). Additionally, total number of posts activities were 988,965 comprising of total likes 557,148 (56.33%); total comments 311,054 (31.45%) and total shares 120,763 (12.21%) respectively. Furthermore, a total number of individual users (N=87,737) interacted with various posts from different community pages. Figure 7.4 shows the stacked graph of the PCV metrics of engagement in different community areas. Table 7.2 shows descriptive statistics of posts by interaction and type of posts.

Table 7.2: Descriptive statistics of posts interaction and types of posts

Type	Count	Percentage	Description
Post likes	59,688	49.16	Number of posts liked
Post comments	39,974	32.92	Number of posts with comments
Post shares	21,738	17.90	Number of posts shared
Events	2,734	3.21	Posts related to community events
Links	12,831	15.09	Posts containing web links
Photo	35,638	41.92	Posts with images
Status	29,928	35.21	Text related information
Video	2,028	2.38	Motion picture related posts
Others	1,841	2.16	Any other posts

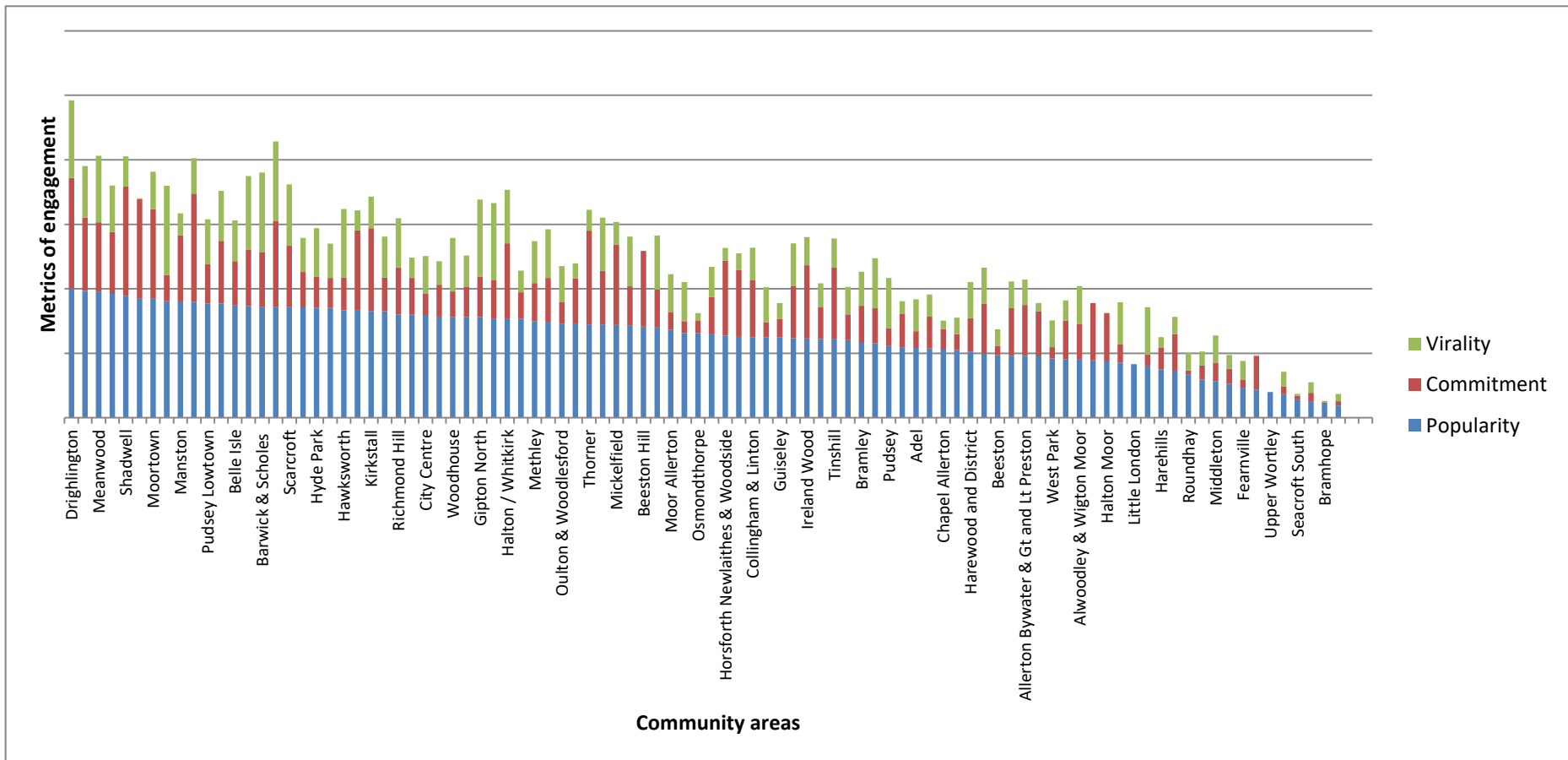


Figure 7.4 Stacked graphs of the PCV metrics of engagement in different community areas

7.5 Quantifying Community Engagement on Facebook

In this research, publicly available data provided on Facebook pages (e.g. PCV metrics) were used to quantify community engagement rate similar to previous studies (e.g. Bonsón and Ratkai, 2013). Facebook engagement rate is a measure of how users (fans) interact with Facebook posts by liking, commenting and sharing (Hoffelmeyer, 2016; Facebook, 2017a). The engagement rate on a Facebook page is related to the number of fans and posts. Facebook pages and groups with a larger number of fans are likely to have more interactions with posts producing a higher engagement rate. On the other hand, where a larger number of fans exhibit lower post interaction, the engagement rate also declines. In this research, the number of fans on a community Facebook account is used to measure engagement rate because they determine the level of interaction on posts as used in previous studies (e.g. Cvijikj and Michahelles, 2013; Schultz, 2017; Dolan *et al.*, 2017). Equation 7.1 is used to quantify the engagement rate on a community Facebook page which is employed as a proxy for social cohesion. Engagement rate E (as measured on Facebook), in the community area, i , can be written follows:

$$E_i = \frac{n_{pcv}}{F_i} \quad (7.1)$$

where n_{pcv} is the total number of likes, comments and shares, and F_i is total number fans in that community. A frequency distribution for the engagement rate in different community areas is presented in Figure 7.5

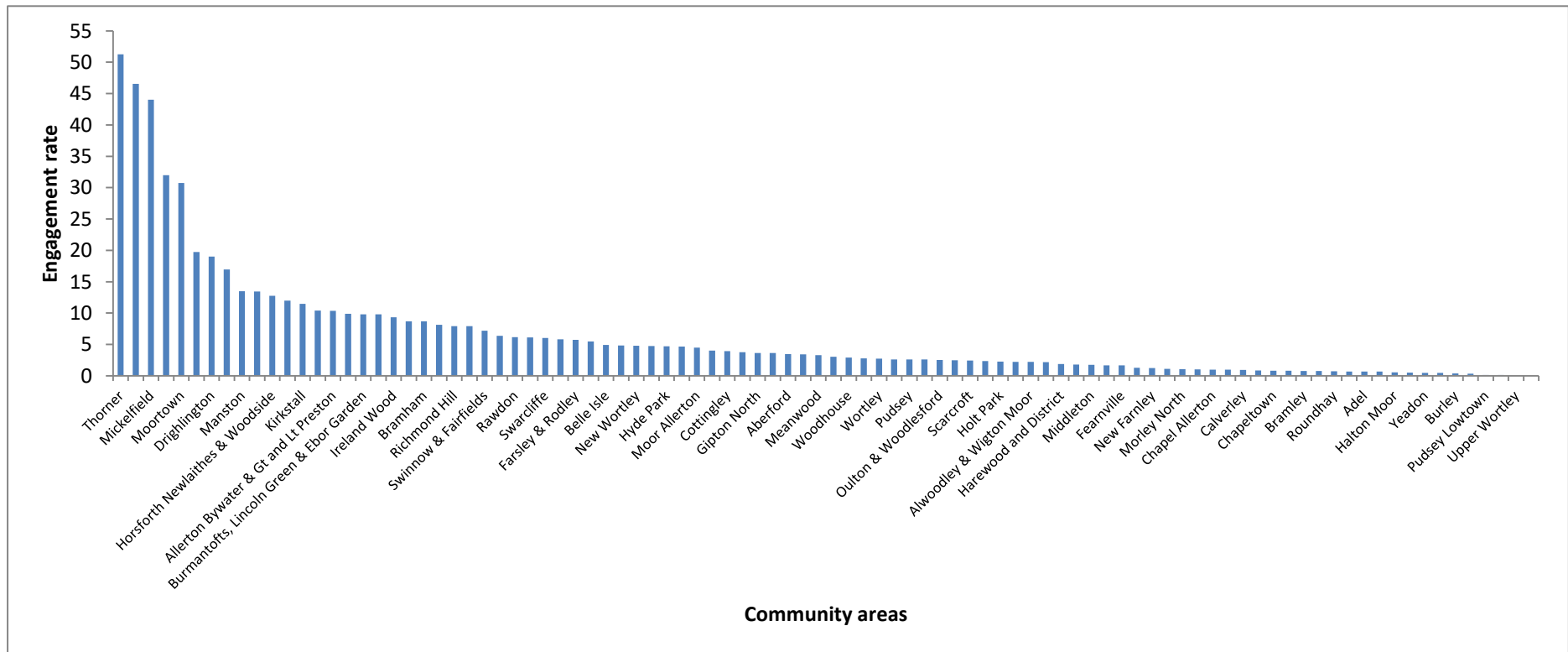


Figure 7.5: Frequency distribution of community areas engagement rate

7.5.1 Grouping Community Engagement Rates

In order to group different community areas based on their social interaction on Facebook pages, the range of engagement rates were categorised into quartiles. A quartile is a statistical method for categorising quantitative data for easy description (Whitley and Ball, 2001). Though there are different ways of grouping numerical data, quartiles were used to simplify the classification of different community areas into a more manageable size and to make referencing easier. In this research, the data were first grouped into quartiles with each quartile representing 25% in data distribution and the results were ranked using Excel. Table 7.3 shows how community areas were categorised into different groups. The spatial distribution of engagement rates in different community areas is presented in Figure 7.6.

Table 7.3: Community areas engagement rate categories

Engagement rate	Community Areas	Percentage	Description	Group
0.06 – 1.08	24	25.53	Low	4
1.11 - 3.03	23	24.47	Moderate	3
3.29 – 7.18	23	24.47	High	2
7.93 – 51.27	24	25.53	Very high	1

Table 7.3 shows that group 1 which comprises of 24 community areas accounted for very high engagement rate (between 7.93 to 51.27), group 2 with 23 community areas have high engagement rate (between 3.29 and 7.18); while group 3 comprising of community areas tend to have moderate engagement rate (between 1.11 to 3.03) and group 4 with low engagement rate (between 1.08 and 0.06) comprises of 24 community areas.

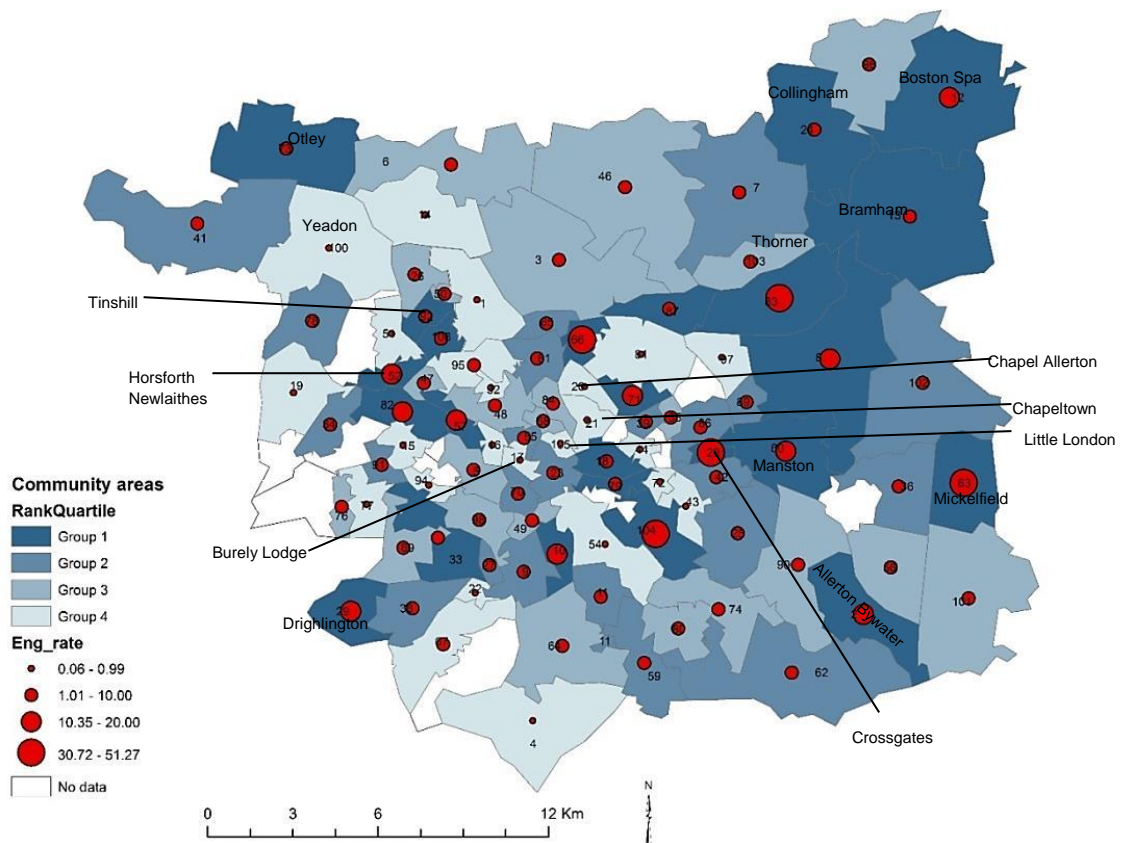


Figure 7.6: Spatial distributions of engagement rates in community areas

As seen in Figure 7.6 there is a considerable difference in average engagement rates between different community groups (the darker the colour, the higher the engagement rates). A higher engagement rate is found in communities in *group 1* including Thorner, Crossgates and Mickelfield; having 51.2, 46.5 and 44.0 engagement rates respectively. These areas tend to have lower diversity indices. For example, residential, ethnic and age diversity is less than 0.2, while employment diversity about 0.3 suggesting that communities in this group tend to be more stable and are likely to have a high level of social interaction. Community areas in *group 2* which include Boston Spa, Drighlington, Manston, Horsforth Newlaithes and Allerton Bywater & Great and Little Preston have engagement rates of 19.7, 19.0, 13.4, 12.7 and 10.3 respectively. This group tends to have similar diversity characteristics as in group 1, except that in group 2, employment diversity is slightly higher (0.4) than group 1. Similarly, *group 3* which comprises of community areas including Tinshill that has a 9.8 engagement rate, while Collingham, Bramham and Otley have engagement rates of 8.7, 8.6 and 8.1 respectively. Community areas in this group have lower engagement rates compared to group 1 and 2. While the age and residential diversity are characteristically similar as in the previous groups, ethnic diversity is slightly higher (about 0.3) than in those groups.

Additionally, employment diversity in this group is also lower than that of group 3. However, community areas belonging to *group 4* including Chapel Allerton, Burley Lodge and Little Woodhouse have 0.98 engagement rates each, while Little London and Chapeltown also have 0.8 engagement rates respectively. Except for Yeadon community area, group 4 community areas which have the lowest engagement rate and highest diversity indices in Leeds. For example, residential diversity is between 0.3 and 0.8, ethnic diversity is between 0.5 and 0.8 and age diversity is between 0.2 and 0.6 respectively (Full list of the community areas Facebook engagement metrics can be found in the Appendix B). Previous studies have established that diversity within a community can affect the establishment of social interactions between residents (Browning *et al.*, 2008; Letki, 2008; Laurence, 2011). Section 5.4.2 provides detailed discussions on the effects of diversity on social cohesion.

7.5.2 Factors Affecting Community Engagement Rate

Variation in the engagement rates between different community areas can be attributed to a number of socio-economic and demographic factors. Firstly, residential stability enables residents to know themselves better and easily organise themselves to establish a common social media platform like a Facebook page or group; to exchange information, discuss local problems affecting them and maintain local social relationships. In contrast to residentially less stable areas where people are transient, organising people to have a common focus especially on a social media channel is difficult. Secondly, ethnic diversity in a community area can affect the way people relate and trust each other because of difference in norms (Sturgis *et al.*, 2014). Therefore, it is less likely for heterogeneous and disadvantaged areas to establish a common presence on social media (McCabe *et al.*, 2013; Park, 2017). Nevertheless, some areas where a community social media profile such as Facebook is available residents do not show much concern to interact with it. On the other hand, in homogeneous communities trust and belonging enhances local social interaction between residents which can enable the establishment of social media presence relatively easier.

Thirdly, some residents of deprived areas continue to face challenges with access to the internet which can greatly affect the way they create and interact with social media channels such as Facebook. Different from deprived communities, in affluent areas, internet availability and accessibility promotes the creation of social media presence and potentially encouraging interaction on those channels. Fourthly, educational attainment also can be potentially an important factor affecting community engagement on Facebook. Community

areas where residents are more educated are likely to maintain social media interaction to keep in touch with local community activities (Perrin, 2015). Whereas, areas with less educated residents are less likely to value the importance of having a community social media presence let alone engaging with them. Figure 7.7 shows comparisons between engagement rate and social cohesion factors in different community groups highlighted in Section 7.6.1.

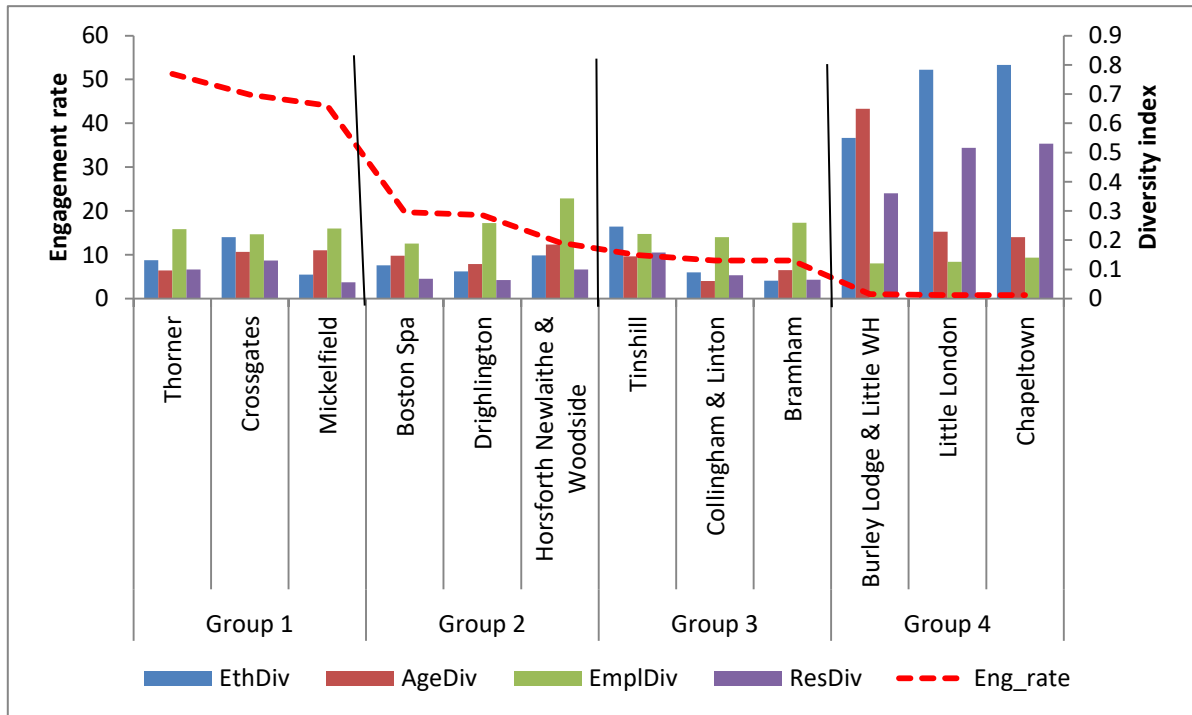


Figure 7.7 Comparison between engagement rate and factors of social cohesion in selected community groups

As shown in Figure 7.7 it is clear that engagement rate has some potential relationship with diversity characteristics of different community areas. In general, engagement rate reduces as the diversity increases and vice versa.

7.5.3 The Concept of Digital Divide

The concept of digital divide and digital inequality has attracted the attention of policy makers and academics in different disciplines including social science (Philip *et al.*, 2017). Digital divide has been defined as a gap between people, households, social groups and geographic areas in their Internet access and utilisation (Philip *et al.*, 2017; Pearce and Rice, 2017). DiMaggio *et al.* (2004) described the concept of digital divide as inequalities in access to Internet, quality of technical connections and diversity of users. The Internet was initially perceived to provide a virtual space where inequalities could dissolve and individuals could interact with others across social, temporal and spatial boundaries (Pearce and Rice, 2017).

Digital divide is a complex and multidimensional issue which requires taking into account different economic, social and technological factors. The economic factors might be because individuals and communities do not have access to computers or an up-to-date internet service due to income challenges; social factors were originally attributed to internet accessibility though nowadays people have access but lack the skills necessary to utilise ICT; and geographical factors exist between rural-urban or between developed and developing countries (Várallyai *et al.*, 2015). Geographic disparities in access can exist due to physical distance from network suppliers (broadband speed); large disparities exist between urban and rural areas. Within the UK context, there are urban areas with populations that are digitally engaged yet materially deprived, or inversely, urban areas with low digital engagement but also low levels of material deprivation. These are part of the wider complex relationships that influences digital divide (Riddlesden and Singleton, 2014). Recent statistics in the UK show that households access to the internet is increasing from 9% in 1998 to 90% in 2017 and social networking is the second most important Internet activity performed after email (ONS, 2017c). Additionally, of the 10% households without the internet access, 64% reported internet is not useful to them while 20% felt that they lack skills and 12% reported that they access the internet elsewhere (ONS, 2017c). The internet can now be accessed using different devices (smart phones, tablets and computers) either via the public or private connections (broadband and Wi-Fi) and 3G and 4G networks are all available (Philip *et al.*, 2017). There are differences of access between geographic locations (rural/urban) (Haight *et al.*, 2014). Furthermore, the overall broadband speeds in rural areas are consistently slower and erratic than those in the urban areas (Philip *et al.*, 2017).

Digital inequality has continued to elude the economically disadvantaged segment of the population. Lack of access to the internet can significantly undermine efforts to obtain current news. Understanding individual's online engagement and the range of activities that users perform is important for assessing how those who are connected are taking the advantage provided by the internet (Haight *et al.*, 2014). Consistently, studies have linked internet access to the factors of education, income and age (Haight *et al.*, 2014; Estacio *et al.*, 2017; Dutton and Reisdorf, 2017). Race and ethnicity are also important determinants of digital divide (Robinson *et al.*, 2015). However there is a gradual shift to the notion of digital skills with individuals having the required skills potentially benefiting more than those who do not have the skills and knowledge (Van Deursen and Mossberger, 2018). Therefore social support in the form of digital literacy to assist those who face challenges of learning new

technologies, especially the older adults, could potentially reduce the digital divide gap and increases participation in social networking among people (Tsai *et al.*, 2017).

7.5.4 Influence of Facebook Interaction Metrics on Community Engagement

To evaluate how each of the new PCV metrics contributes to engagement in different community areas, we divide the sum of each metric by the number of communities in those categories to obtain the mean contribution of the metrics. Figure 7.8 shows how different metrics contributed to engagement rates in different community groups.

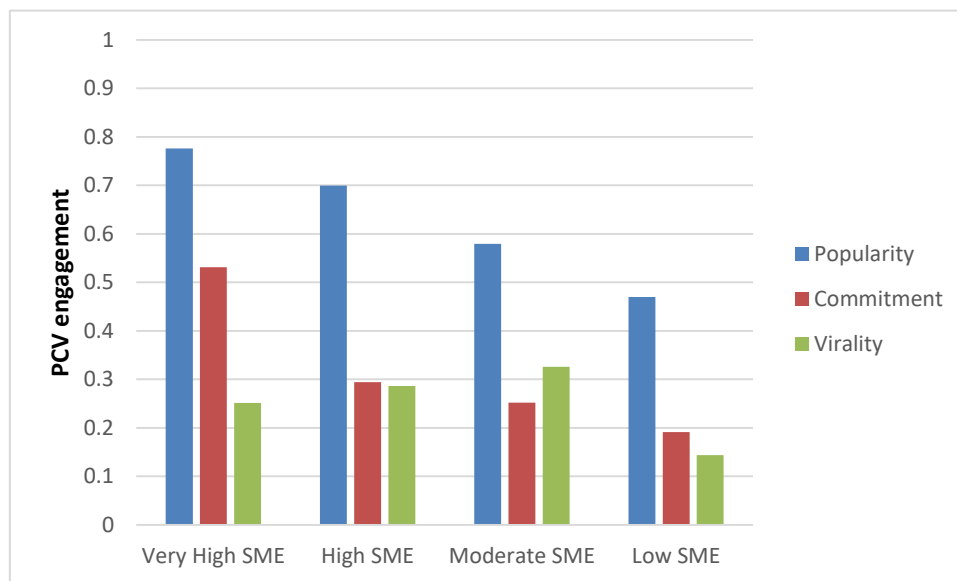


Figure 7.8: Metrics of engagement in different community areas

Figure 7.8 shows the proportion of the contributions of PCV metrics to engagement rates in different community area groups. It is clear that the metric of popularity (P) contributed more to engagement (between 0.46 and 0.77) across different community areas; followed by the metrics of commitment (C) (between 0.19% and 0.53%) and metrics of virality (V) (between 0.14% and 0.32%) respectively. Communities with higher engagement rates also tend to have higher percentages of popularity and commitment compared to areas with moderate and low engagement, indicating the importance of the metrics for engagement in those areas. Previous studies (see above) also found that the metrics of popularity (i.e. ‘likes’) are likely to be higher than those measuring commitment (‘comments’) and virality (‘shares’). However, that does not mean the metric of popularity is the highest indicator of engagement.

7.5.5 Engagement and Action in the Community Areas

Engagement on a community Facebook channel can be related to crime rates of different community areas. The ability of cohesive communities to employ online social networks such as Facebook can potentially support efforts in crime reduction in the neighbourhoods, as information about residents concern about safety is being discussed online for collective action (Hattingh, 2015). Furthermore, social ties in a community have been found to be associated with lower levels of neighbourhood crime especially burglary and violence (Yuan and McNeeley, 2017a). Previous studies have attempted to use SM to study crime patterns at different spatial scales albeit using Twitter data (Gerber, 2014; Malleson and Andresen, 2015; Williams *et al.*, 2016). For example, Bendler *et al.* (2014) emphasised that online social networks such as Twitter and Facebook can be used for facilitating a virtual neighbourhood watch and supporting the police in crime prevention (see Section 4.3 for detailed discussion). Figure 7.9 shows Facebook engagement rate in relation to crime rates in Leeds community areas.

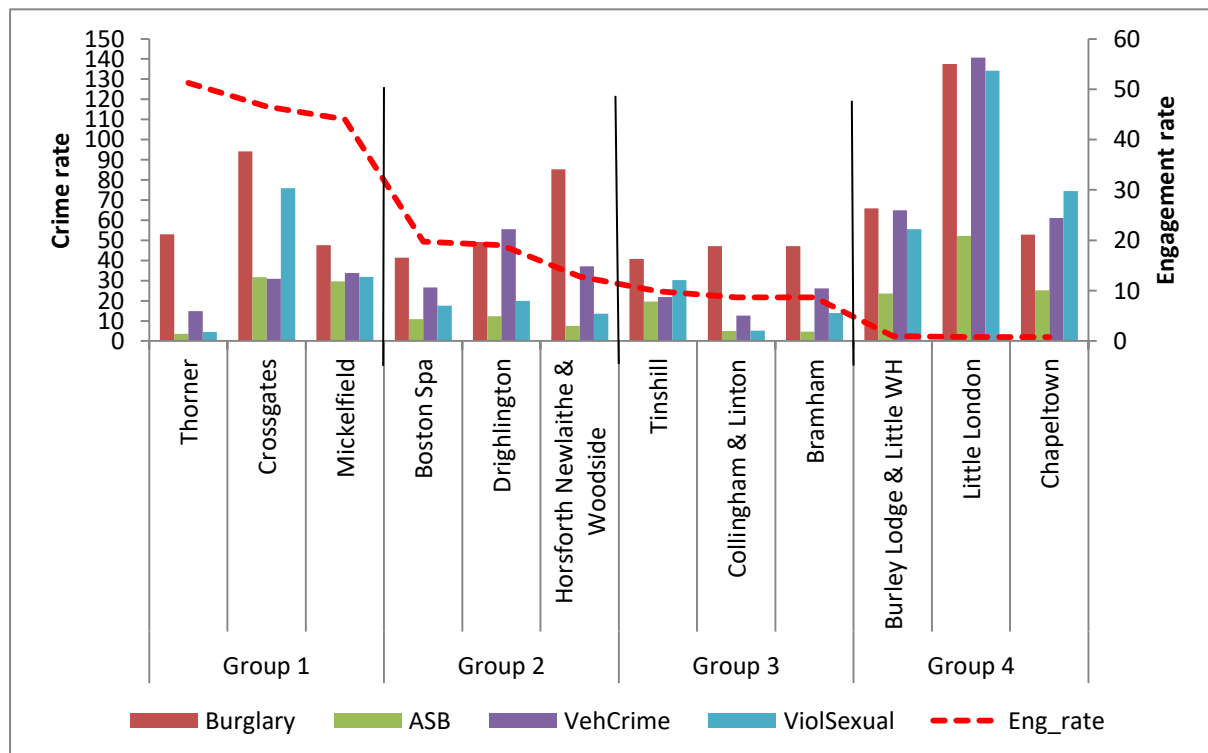


Figure 7.9: Engagement rate in relation to crime rates in the community areas

As seen in Figure 7.9, the engagement rate tends to decrease with increasing crime rates in most community areas. Recent police statistics⁹ as of December 2017 shows that crime rates per 1000 population in Leeds for burglary (12.83), vehicle crime (11.23), violence and sexual offences (34.95) and ASB (25.56) respectively. In group 1 community areas (very high engagement rate) tend to have the lowest crime rates. For example, Thorner community area has the lowest rates of anti-social behaviour (ASB) and sexual and violence crimes (3.52 and 4.58) in Leeds. Though ASB rate is slightly higher than the city average in areas like Mickelfield (29.58) and Crossgates (31.68), this is likely because of the proximity to less privileged areas like Gipton South and Harehills. Previous studies have found a positive relationship between ASB and children raised in deprived areas which is likely to affect more affluent surrounding neighbourhoods (Odgers *et al.*, 2015). Overall burglary crime rates are between 47.64 in Mickelfield and 94.08 in Crossgates which is higher than the city average. Routine activities of residents combined with interconnectivity, mobility and interaction have significant implications for crime victimisation especially burglary (Cohen and Felson, 1979). Additionally, the cost of crime profitability increases the likelihood of burglars to target affluent areas than their deprived counterparts (Ward *et al.*, 2014) and the risk of burglary victimisation is more likely to be higher where affluent neighbourhoods are surrounded by deprived neighbourhoods (Mburu and Bakillah, 2016).

Crime rates in group 2 community areas with high engagement are not far from being similar to those in group 1 except that vehicle crime rates were relatively higher in group 2 than in group 1. For example, vehicle crime rates were between 26.66 in Boston Spa and 55.56 in Drighlington; compared to 14.79 in Thorner and 33.81 in Mickelfield community areas from group 1.

Community areas comprising group 3 with moderate engagement tend to have the lowest crime rates compared with the community areas in group 1 and group 2. In this group, ASB rates were between 4.60 in Bramham and 19.59 in Tinshill community areas; violence and sexual related offences were between 5.23 in Collingham and 30.26 in Tinshill community areas; while vehicle rates were between 12.60 in Collingham and 26.10 in Bramham community areas respectively. All these crime rates are lower than the city averages. However, as in group 1 and 2, burglary crime rates were relatively higher than the city average but slightly lower than those in group 1 and 2.

⁹ https://www.police.uk/west-yorkshire/LDT_W/performance/compare-your-area/

Crime rates are distinctively different in group 4 community areas with low engagement. In this group except for ASB rates in Burley Lodge and Little Woodhouse community area that were slightly below the city average, all the other crime rates were higher than the city averages. Specifically, burglary crime rates were between 52.88 in Chapeltown and 137.45 in Little London community areas; violence and sexual offence crime rates were between 55.51 in Burley Lodge and Little Woodhouse and 134.20 in Little London community areas; while vehicle crime rates were between 61.09 in Chapeltown to 140.72 in Little London community areas respectively. These relationships will be quantified further in the following sections using regression models so as to explore more insights.

7.6 Multiple Regression Models of Community Engagement and Burglary Crime Rates

In this research, multiple regression models (global and local) were employed to quantify the relationship between community engagement and recorded burglary crime rates (dependent variable). The purpose of the model calibration is to explore whether the inclusion of SM variable (virality) can potentially improve the traditional regression models of burglary crime and community cohesion performed in chapter 5. The model uses the combination of standard and adjusted (diversity indices) variables constructed from the Census statistics, and PCV indicators generated from Facebook. Diversity statistics and their standard equivalent are used to control for socio-economic and demographic effects in different community areas (Section 5.6 provides detailed description of the variables).

The spatial distribution of burglary rates in Leeds (Figure 7.10) shows that community areas in the city centre tend to have a higher burglary rate compared to the community areas further away from the city centre. The spatial distributions of the independent variables are mapped in Figure 7.11 to Figure 7.14.

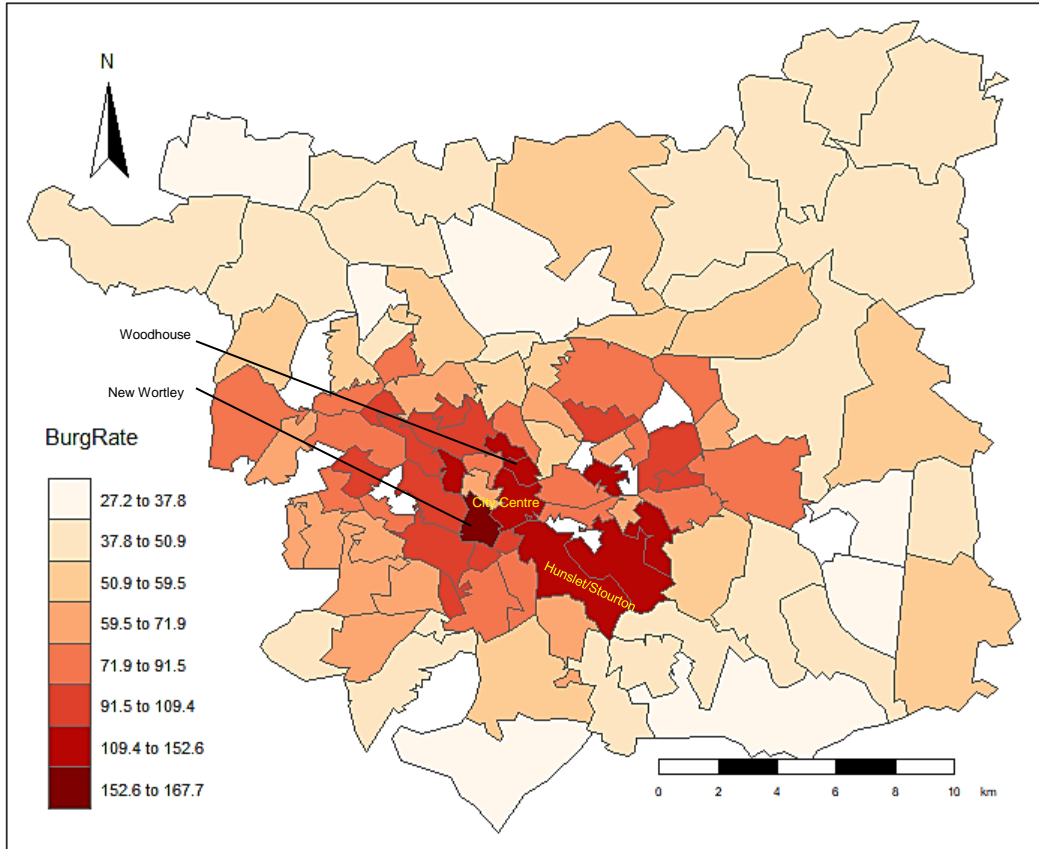


Figure 7.10: Map of the distributions of burglary rates in Leeds community areas

For the distribution of the age diversity variable (Figure 7.11), higher values can be seen in the City Centre, Hyde Park, Woodhouse, Headingley, Hunselet & Stourton and surrounding areas. The distributions of residential diversity variable (Figure 7.12) indicate higher values in the city centre and towards the north, with relatively lower values around community areas on the fringes of the district reflecting the degree of residential stability. The distribution of the economically inactive variable (Figure 7.13) reflects the pattern of the population that makes up the variable (i.e. students, retired and people living with long-term illness). The spatial distributions of the virality variable (Figure 7.14) tend to show patterns that indicate the degree of virality of information across different community areas.

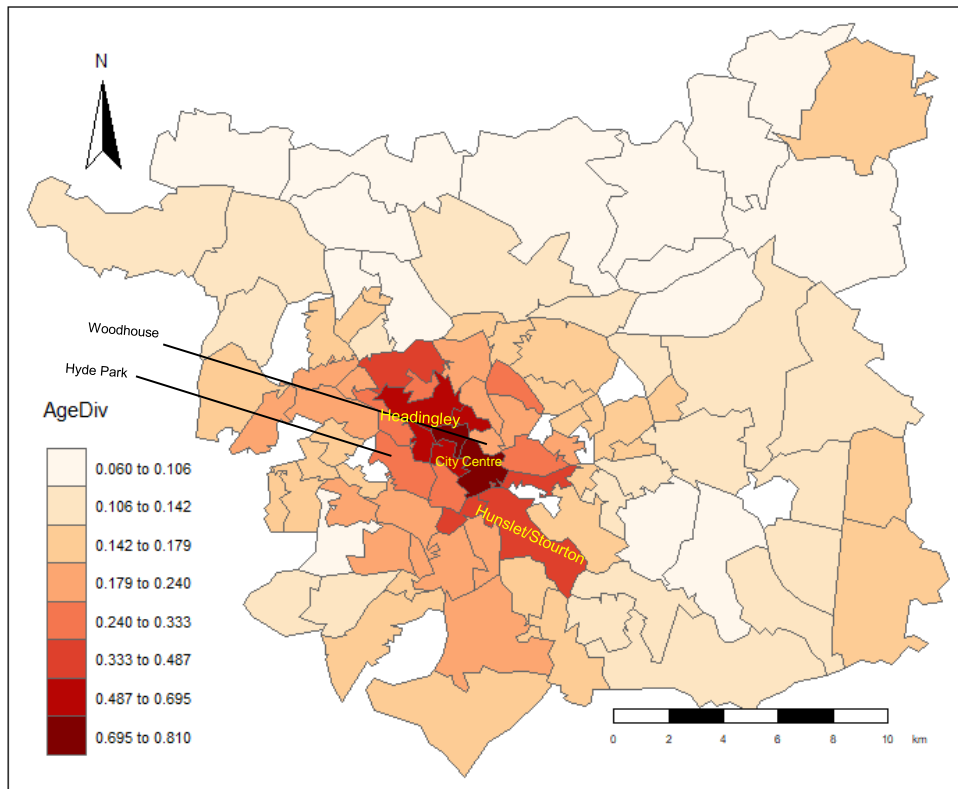


Figure 7.11: Map of the distribution of age diversity in Leeds community areas

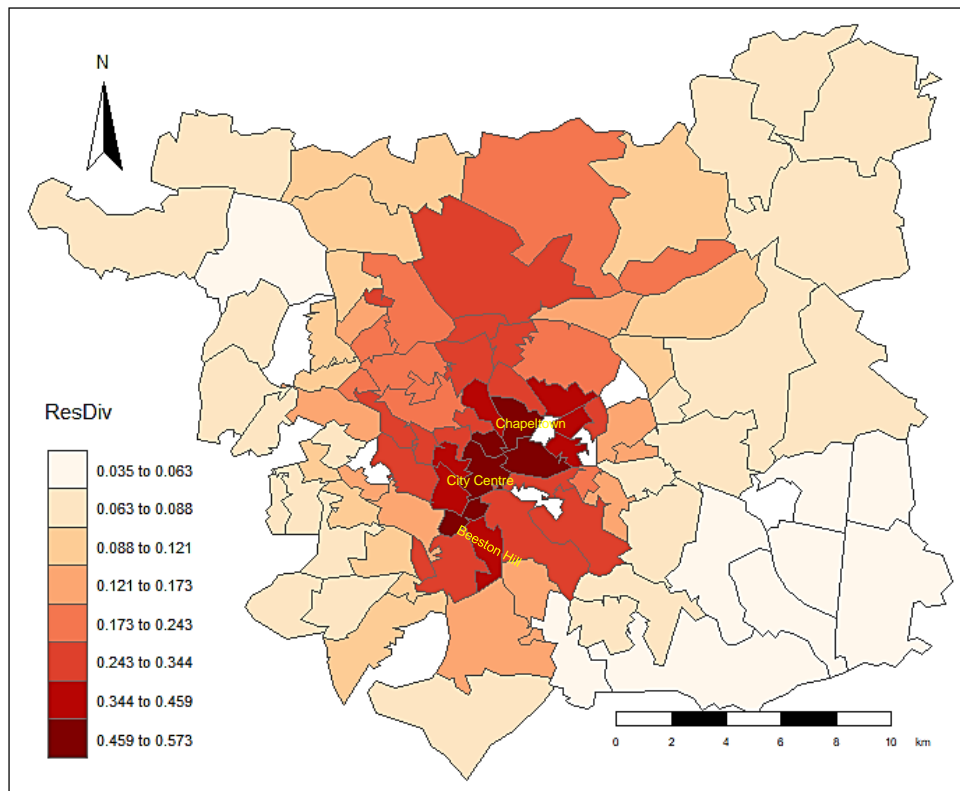


Figure 7.12: Map of the distribution of residential diversity in Leeds community areas

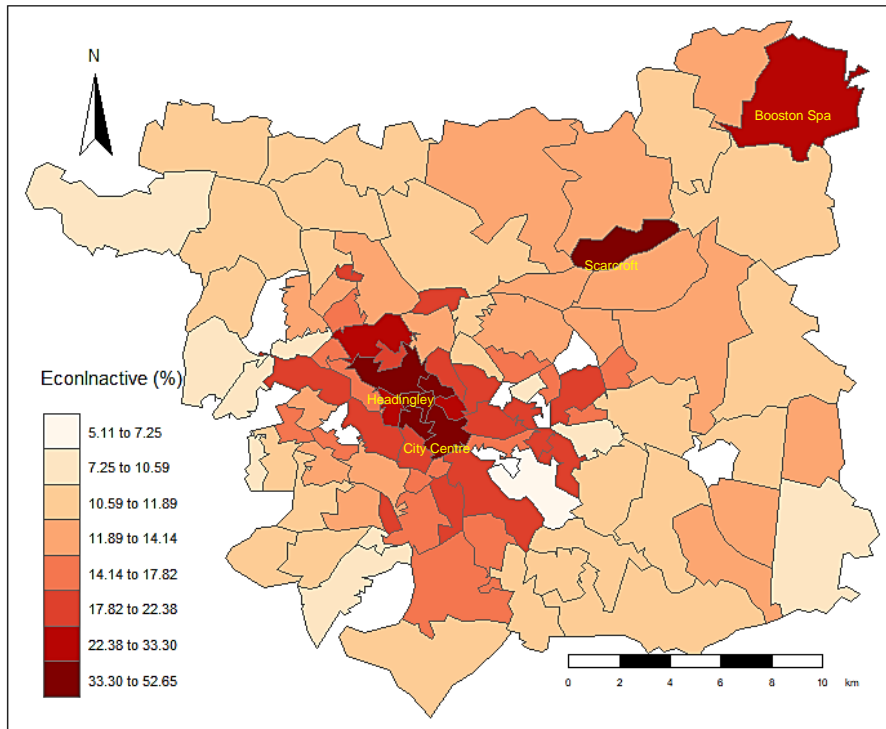


Figure 7.13: Map of the distribution of economically inactive population in Leeds community areas

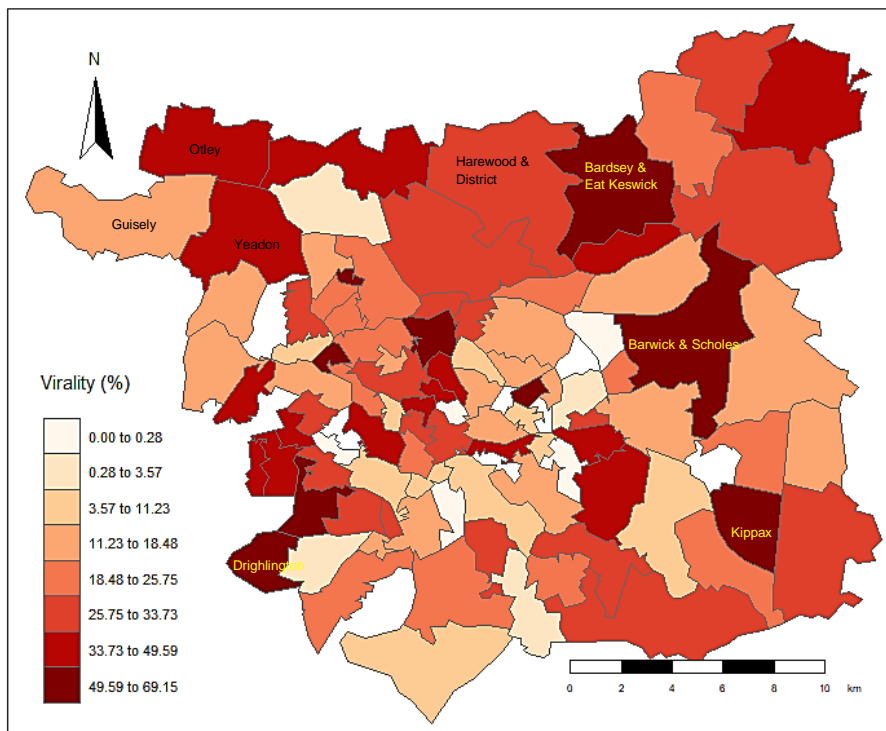


Figure 7.14: Map of the distribution of virality in Leeds community areas

In this research, the PCV indicators are used as a proxy for the strength of social networks in different community areas. Additionally, the metrics are widely used in previous studies to quantify engagement on Facebook pages (e.g Cvijikj and Michahelles, 2013; Theiss *et al.*, 2016). For example, Hattingh (2015) attempted to use Facebook data to explore the impact of community watch on crime in Gauteng, South Africa. He found that information sharing related to community building and awareness on neighbourhood safety on a Facebook page can significantly help in reducing neighbourhood crime especially burglary. The major limitation of his study is that he analysed only 247 posts from a single community Facebook group which affected generalisation of the findings. To address this limitation in this research, PVC interaction metrics on 85,000 posts generated from 94 community Facebook accounts are used in the analysis.

Prior to model construction, an exploratory analysis was performed using Pearson's correlation coefficient (r) to assess the relationship between the dependent (burglary crime rates) and independent variables (PCV indicators, standard and diversity statistics) to be used in the model. Variables were also checked for possible presence of multicollinearity. Multicollinearity is present when there is a high degree of correlation between independent variables leading to redundancy. High degree of correlation between independent variables is not desirable because it can significantly affect model performance and reliability (Wang, 1996). There is no standard rule for filtering out variables based on the issue. In this research, correlations greater than $r = .80$ between independent variables are regarded as very significant; therefore one such correlated variable will be dropped from the analysis. On the other hand, variables that have a higher correlation with the dependent are usually preferred for inclusion in the model. Table 7.4 presents Pearson's correlation coefficient between burglary crime rates, PCV indicators and standard and diversity indices.

Table 7.4: Pearson's correlation coefficient between burglary crime rates, PCV indicators and diversity and standard variables

Variables	Burglary	P	C	V	Ethnic div	Age div	Employment div	Education div	Residential div	Ethnic minority	Age 16-24	Age16-64Inactive	Age16-64Noqual
Burglary													
P	-.157												
C	-.171	.639**											
V	-.249*	.610**	.255*										
Ethnic div	.544**	-.192	-.287**	-.131									
Age div	-.507**	-.043	-.298**	-.031	.480**								
Employment div	-.216*	.063	.175	.031	-.388**	-.112							
Education div	-.227*	.018	-.210*	.027	-.235*	-.761**	-.087						
Residential div	.569**	-.182	-.301**	-.133	.981**	-.534**	-.360**	-.263*					
Ethnic minority	.288*	-.252*	-.335**	-.149	.682**	.485**	-.465**	-.276**	.679**				
Age16-24	.250*	-.037	-.263*	.001	.267**	.792**	-.662**	-.662**	.308**	.583**			
Age16-64Inactive	.266*	-.093	-.300**	-.033	.343**	.721**	-.155	-.561**	.387**	.726**	.954**		
Age16-64Noqual	-.013	-.248*	-.180	-.149	.050	-.073	-.072	.272**	.057	.497**	.166	.386**	
ResidenceLess2yrs	.339**	-.030	-.210*	-.037	.487**	.682**	-.037	-.465**	.579**	.569**	.669**	.733**	.094

** Correlation is significant at the 0.01 level (2-tailed), * Correlation is significant at the 0.05 level (2-tailed).

Table 7.4 shows Pearson’s correlation coefficient between burglary crime rates (dependent) and the independent variables (PCV indicators, standard and diversity statistics). Multicollinearity was detected between ethnic diversity and residential diversity ($r = .981$, p -value < 0.01) variables; and between young population (16-24) and proportion of those aged 16-64 who are economically inactive ($r = .954$, p -value < 0.01) meaning that in each case one of the variables will not be included in the model. In order to decide on which of the variables will be dropped from the model, their correlations with the dependent (burglary crime rates) were assessed. Residential diversity and proportion of economically inactive population variables were found to have a higher correlation with burglary crime rates ($r = .569$, p -value < 0.01 and $r = .266$, p -value < 0.05) respectively. Consequently, ethnic diversity and age 16-24 variables were dropped from the analysis. Table 7.5 presents independent variables included in the model based on their correlations with the dependent variable.

Table 7.5: Independent variables included in the model

Variables	Description
Post likes (P)	Percentage of the total post that have been liked (number of post with likes in an area/total number of post in that area *100)
Post comments (C)	Percentage of the total post that have been commented on (number of post with comments in an area/total number of post in that area *100)
Post shares (V)	Percentage of the total post that have been shared (number of post shared in an area/total number of post in that area *100)
Standard	Ethnic minority, Length of residence less than 2 years, Age 16-64 economically inactive and Age 16-64 no qualification
Diversity	Age, Employment, Education and Length of residence

P - popularity, C - commitment, V- virality

Global Regression Model

In this research, global Ordinary Least Square (OLS) multiple regression models (stepwise) were employed to model the relationships between burglary and community structural variables. The OLS linear regression provides a global model to generate predictions or to model a dependent variable in terms of its relationships to a set of explanatory variables and

is probably the most popular procedure used in criminal justice and criminal research (Walker and Maddan, 2013). Though the global model parameters derived from the OLS assumed that variables are constant over space (Charlton *et al.*, 2009) this assumption does not always hold as spatial variations in relationships is not stationary (Erdogan *et al.*, 2013). They tend to vary across space known as *spatial heterogeneity*, hence the need for a localised model such as Geographically Weighted Regression (GWR) (Charlton *et al.*, 2009). GWR is a more robust alternative in exploring the spatial variation of relationships across space (Arnio and Baumer, 2012). A typical global OLS is provided in Equation written in the form of Equation 3.1.

Local Regression Model

A detailed discussion on GWR modelling technique is provided in Section 3.6.1.1. In this analysis, the *Gwmodel* package based on Lu *et al.* (2014) was employed in R statistical environment (R Core Team, 2017). The package provides functions for bandwidth selection (*bw.gwr*), GWR model calibration (*gwr.basic*) with different statistical outputs (such as coefficients, local R^2 , and standard error) and tests for model significance (*gwr.t.adjust*) including p-values, adjusted p-values and t-statistics. The outputs from the package also include the global regression results which make comparison between the global OLS and local GWR models much simpler. To calibrate the GWR model, the same parameters used in the global model were used (Equation 3.5).

For model calibration in GWR, the choice of bandwidth is an important step. A bandwidth is a distance search window over which a localised model is controlled (Lu *et al.*, 2014). Bandwidths are locally chosen by a data-driven method based on minimization of a local cross-validation (CV) criterion (Vieu, 1991; Arlot and Celisse, 2010). A CV score is essentially the estimated squared production errors (Fotheringham *et al.*, 1998). There are different types of kernel which can be used to estimate a bandwidth in a local regression such as Gaussian, Exponential, Tri-cube, Bi-square and Box-car functions. Gaussian and exponential kernels are continuous (uses all data); weight of points is between 0 and 1 which decreases according to the shape of kernel. The Tri-cube, Box-car and Bi-square kernels are discontinuous (uses nearest neighbour or a fixed distance). The main difference between them is that the continuous kernels uses a *fixed* bandwidth approach, while the discontinuous kernels uses an *adaptive* bandwidth approach which can be specified beforehand (Lu *et al.*, 2014).

In this research, the *adaptive bandwidth* approach based on a bi-square kernel function was used. The adaptive bandwidth is widely preferred because it allows the same number of sample points to be used on each estimation (Charlton *et al.*, 2009). It also provides an intermediate weighting between the box-car and Gaussian functions. The approach can be affected by discontinuity where fewer data points are considered other than the one specified. On the other hand, the advantage of using the bi-square kernel function is that it gives a fractional decaying weight according to the proximity of the data to each point for a fixed or specified nearest neighbour (Lu *et al.*, 2014). Here, an optimum bandwidth was obtained using 8 nearest neighbours (about 8% of data points) as illustrated by Figure 7.15.

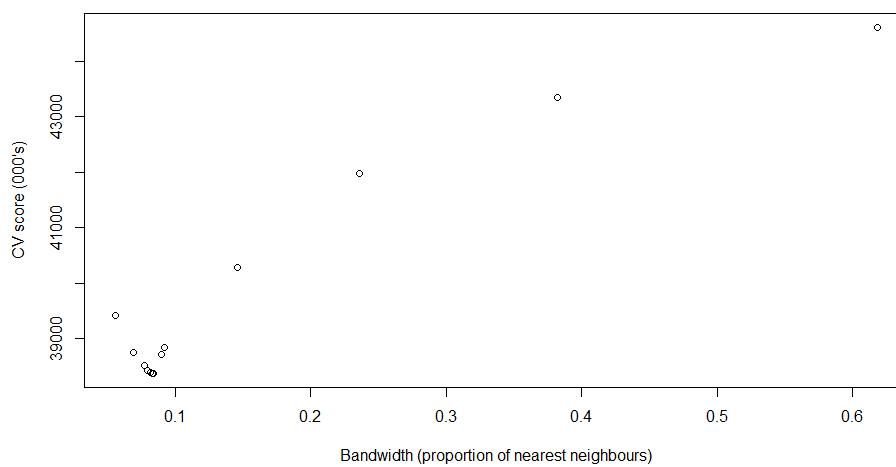


Figure 7.15: Plot of cross-validation (CV) score as a function of bandwidth

Furthermore, the outputs of a GWR model can be mapped for exploratory purposes and to guide decision making. These outputs include local R^2 and standard error. Local R^2 shows how well GWR performed in modelling the relationships between the dependent and independent variables across different locations; while the standard error is used to determine the accuracy of predictions. A smaller standard error value is preferred because it indicates a better model fit. Additionally, Fotheringham *et al.* (1998) suggested mapping of t-values statistics for each parameter in order to assess the strength of the relationships across space. In this analysis, the main outputs from GWR such as the coefficients of the parameters, local R^2 , standard errors and local t-values for each were all mapped in order to examine local variations across Leeds community areas.

7.6.1 Model Interpretation Of Community Engagement and Burglary

Global Model

A Global OLS and local GWR models were employed to quantify the relationship between community engagement (used as a proxy of community cohesion) and burglary crime rates in Leeds. Table 7.6 presents a summary of the global regression model that can be used to assess the contribution of each variable based on its relative importance. The statistics reported are the standard error, t-value and p-value respectively. One of the important guides for evaluating the explanatory power of the global model is significant t-statistics (Dunn, 1989). The t-statistic is the estimated coefficient divided by its own standard error. Significant t-statistics should be greater than 1.96 in magnitude, corresponding to a p-value of less than 0.05. The standard error of estimate is a measure of the accuracy of prediction in regression modelling (McDonald, 2009).

Additionally, for a conventional global model, the goodness of fit can be assessed using the adjusted r-squared (AR^2) (Hurvich *et al.*, 1998; Kongmuang, 2006). AR^2 is the fraction by which the square of the standard error of the regression is less than the variance of the dependent variable (Faraway, 2002). It increases only if the variables are contributing to the model. Overall, the AR^2 value for the global OLS model is 43%; an improvement of about 20% compared to the traditional model of burglary (in chapter 5) where only 23% of the variation in the dependent was explained by the model. This suggests that the inclusion of an additional variable from the PCV indicators has significantly contributed to improving the performance of the previous model.

Table 7.6: Summary of global OLS regression model

	Estimate	Std.Error	t value	Sig.
(Intercept)	52.148	5.470	9.534	.000
ResDiv	80.353	19.388	4.144	.000
AgeDiv	100.489	24.422	4.115	.000
Age 16-64 EconInactive	-0.007	.003	-2.612	.011
Virality (V)	-0.325	.134	-2.431	.017

As shown Table 7.6, length of residence diversity and age diversity were very positively correlated with burglary and statistically significant (p-value <0.001); economically inactive

population and the variable post sharing (V) were negatively correlated with burglary and statistically significant (p-value <0.05) respectively. The values of t-statistics were all greater than 2 indicating stronger relationships between the dependent and independent variables.

Residential instability reduces the potential for social interaction in a community (Thomas *et al.*, 2016) and increases the tendency for crime victimisation (Bell and Machin, 2011). It is likely that the residential stability increases community building and collective action especially for reducing crime.

In terms of relationships between age diversity and crime, previous studies have demonstrated that offenders are commonly drawn from younger age groups than elderly people (Farrington, 1986; Gottfredson and Hirschi, 1990; Sampson and Laub, 2003; McVie, 2005; Blonigen, 2010; McCall *et al.*, 2013; Sweeten *et al.*, 2013). Equally, however, we know that young are also targets for crime (Mahmud *et al.*, 2014).

Significant negative correlation found in this research between economically inactive population and burglary crime which might be seen as counterintuitive. Previous studies have found support for relationships between income inequality and property crime (Witt *et al.*, 1998; Kelly, 2000; Demombynes and Özler, 2005; Reilly and Witt, 2008). However, correlation between economic inactivity with burglary crime does not necessarily imply causation as this relationship might only suggest that unemployment might contribute to offending elsewhere. Recent statistics in the UK show that economically inactive people are likely to be twice as likely to be victims of burglary crime than those who are economically active, considering this category of population comprise of students, those who are retired and people with long-term health challenges (ONS, 2014c; ONS, 2016a), so clearly the relationship for Leeds needs further investigation.

A negative correlation between the virality (V) indicator of engagement and burglary indicates that information sharing about community building and community safety is associated with a decrease in burglary victimisation. and information sharing relating to community building can create awareness about neighbourhood crime which can lead to community action (Hattingh, 2015; Sachdeva and Kumaraguru, 2015). Previous studies have found that the metric of virality (shares) shows the greatest degree of engagement because users are expressing a desire to circulate the information on their networks to increase interaction (Stockley *et al.*, 2013).

Local Model

Table 7.7 presents the results of the local GWR model. Unlike the global model that uses the AR^2 to evaluate the goodness of fit, the GWR model uses the Akaike Information Criteria Corrected (AICc) to assess the goodness of fit. AICc provides a measure of information distance between the model which has actually been fitted and the unknown ‘true’ model (Charlton *et al.*, 2009). An AICc value of less than 3 means that the models being compared are equivalent. Though this measure is relative (values can be large or small), it is the difference in the values between the models being compared that is important. In this research, the global model produced an AICc value of 846.367 compared to the GWR model with 838.512, indicating an improvement to the local model. Additionally, the AR^2 value increases to 0.63 (63%) in the GWR model compared with 0.43 (43%); an improvement of 20% further suggests that GWR model performs better than the OLS model.

Table 7.7: Summary of local GWR model

	Min.	1st Qu.	Median	3rd Qu.	Max.
(Intercept)	43.143	54.372	59.708	71.884	87.909
ResDiv	14.506	43.732	59.734	88.517	117.903
AgeDiv	33.012	72.674	84.231	96.411	136.719
Age 16-64 EconInactive	-0.010	-0.007	-0.006	-0.005	-0.002
Virality (V)	-0.667	-0.524	-0.386	-0.151	-0.012

From Table 7.7 it is clear that variation in the parameter estimates exists in Leeds community areas. For example, the coefficients of age diversity range between a minimum value of 33.01 and a maximum value of 136.72 meaning that one unit change in age diversity (index) is associated with an increase in burglary rates between 33.01/1000 population in some areas and 136.72/1000 population in the other areas respectively. The inter-quartile range for age diversity is between 72.67 (1st quartile) and 88.51 (3rd quartile) respectively. Compared to the age diversity, the coefficient of virality (sharing) of information about crime and neighbourhood safety is associated with decreases in burglary rates across contrasting community areas in Leeds. The interquartile coefficients range of between -0.52 and -0.15 (1st quartile and 3rd quartile), indicate that 1 unit change in virality (percentage) resulting in a decrease in burglary rates by the corresponding values. Additionally, the global model

(Table 7.6) coefficients lie between the interquartile ranges of the GWR model. Charlton *et al.* (2009) emphasised that mapping such spatial variations can provide a better understanding of the relationships being quantified (community cohesion and burglary crime rates in this context). In this analysis, the GWR model local R^2 values and intercept coefficients showing spatial variations across Leeds are shown in Figure 7.16 and Figure 7.17 respectively.

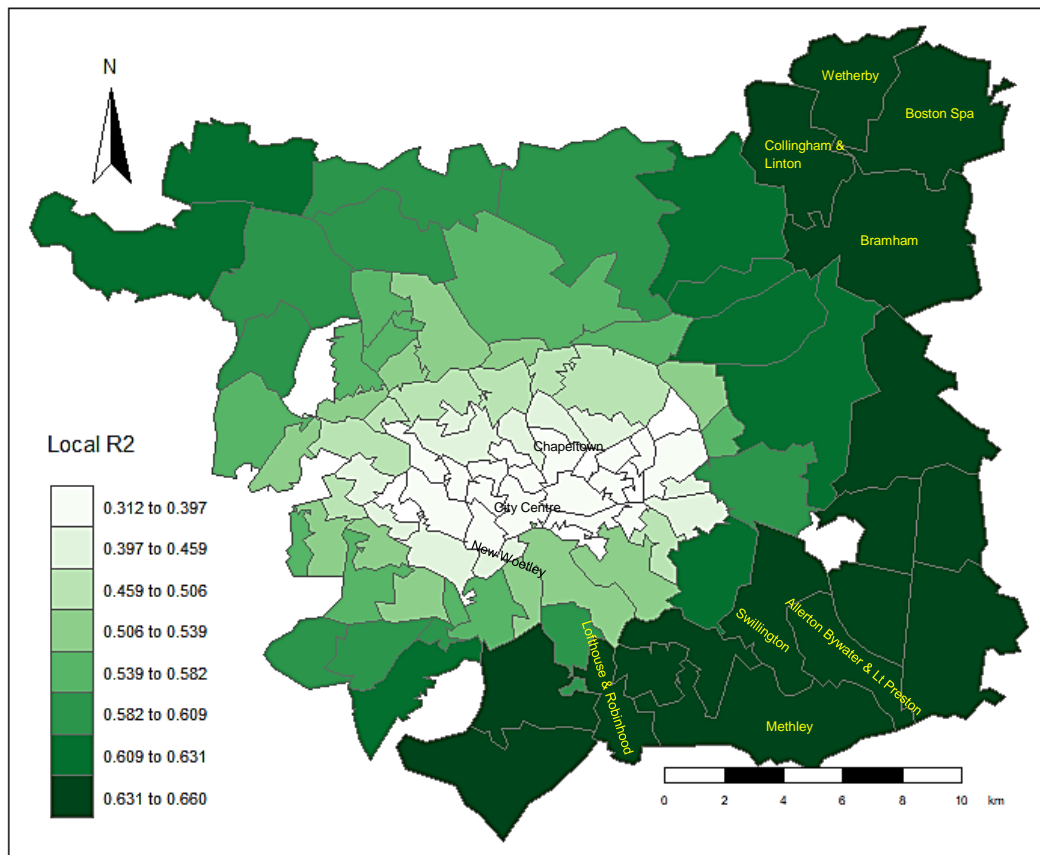


Figure 7.16: Map of the distribution of the local R^2 values of the GWR model

It can be seen from the map of the local R^2 values (Figure 7.16) that the GWR model performed best in the northeast and southern parts of Leeds. These community areas include Wetherby, Boston Spa, Bramham and Collingham & Linton in the northeast; Methley, Lofthouse & Robin Hood, Allerton Bywater & Little Preston and Swillington in the south with local R^2 values between 63% and 66% respectively. In the central parts of Leeds community areas (including City Centre, New Wortley, Burley Lodge & Little Woodhouse, Hyde Park, Woodhouse, Little London and Chapeltown) the model performs relatively lower compared to other areas (local R^2 values between 31% and 45%) respectively. It is likely that

certain key indicators relating to community engagement necessary for mediating burglary might be missing in those areas.

The intercept coefficients of the GWR model (Figure 7.17), show a clear variation in the distribution of burglary rates in different community areas of Leeds. Higher values are found in community areas around the city centre decreasing outwards. The community areas in this category mainly constitute areas with a higher proportion of *student residences*, including Little London, Hyde Park and Headingley (79.68 to 83.44); Burley Lodge & Little Woodhouse (74.28 to 79.68) respectively. Student areas are more likely to be victimised because of their routine activity (Barberet and Fisher, 2009) and possession of attractive valuable items such as electronics and mobile phones (Shepherd, 2006; Kongmuang, 2006). However, higher rates of burglary in *disadvantaged* community areas such as Chapeltown, Harehills and Burley might be associated with lower levels of community engagement characterised by socially disorganised areas. Crime (burglary in particular) thrives in disadvantaged areas especially with lower levels of social capital, because residents are likely not look out for each other (see Section 2.4). Different from the city centre scenario, in the *affluent* areas especially in the northwest (such as Arthington & Pool, Yeadon, Bramhope, Otley and Guiseley) and towards the south community areas including Gildersome, Morley North and Middleton the rates of burglary crime decreases to between 43.52 and 52.85. Similarly, a relatively lower burglary rates (52.85 to 59.71) were also found in the northeast and southwest areas respectively. It is likely that stronger community engagements in those areas are acting to mediate the levels of victimisation in those areas. In general, burglary rates vary across contrasting community areas in Leeds as evident from the spatial distribution of the intercept coefficient (Figure 7.17). Variation in the rates of burglary is characterised by community differences between the student areas, city centre, affluent and deprived areas in Leeds; suggesting differences in the degree of community engagement, a factor found by previous studies (e.g. Sampson and Groves, 1989).

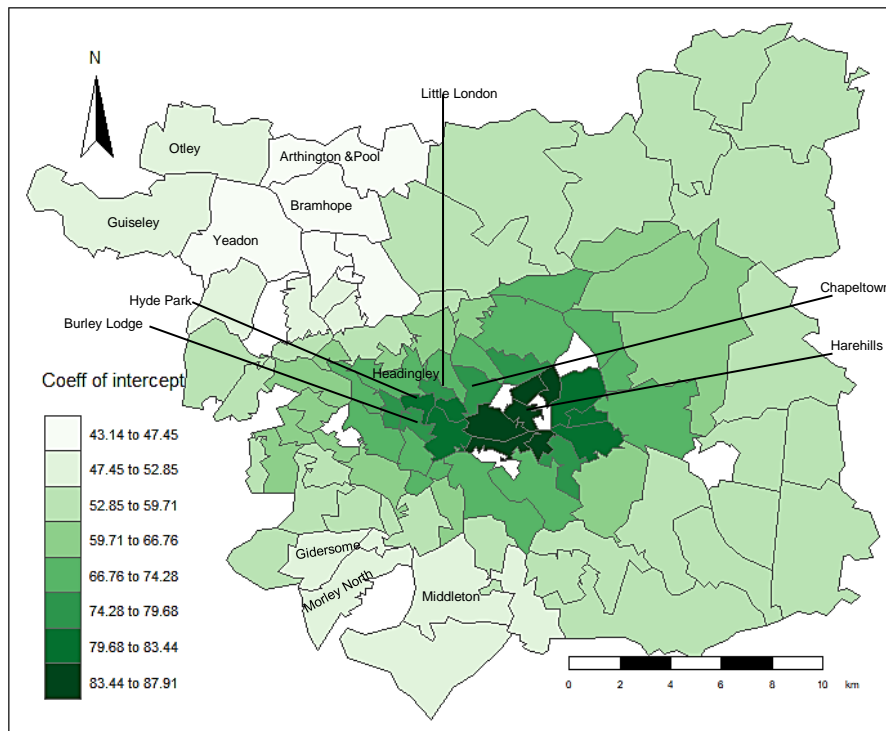


Figure 7.17 Distributions of the intercept coefficients

The variation in the coefficients of the parameter estimates, for virality (V), age diversity, residential diversity and economically inactive population; along with corresponding standard error of estimates and p-values in different community areas are mapped in Figure 7.18 to Figure 7.29 respectively.

The distribution of the virality parameter (Figure 7.18) shows clearly that in the northwest and southwest community areas the relationship is less negative than areas around the city centre. Additionally, there is indication from the patterns of the coefficient of the parameter that the effect of virality of information is associated with reduced rates of burglaries across different community areas. A possible explanation for this occurrence might be because networks of social engagement on SM especially Facebook are contributing to creation of awareness about crime through information sharing and allowing people to organise and take collective action (see above). However, it is also likely that the greater effect of virality of information in reducing burglary rates to be higher in areas with a higher burglary rates than areas with a lower rates of burglary, suggesting a greater need for enhancing community engagement. The lower standard error values (Figure 7.19) in those areas also suggest that the results can be reliable. The p-values range between 0.002 and 0.014 of the parameter (Figure 7.20) further indicate a significant relationship across many community areas of Leeds.

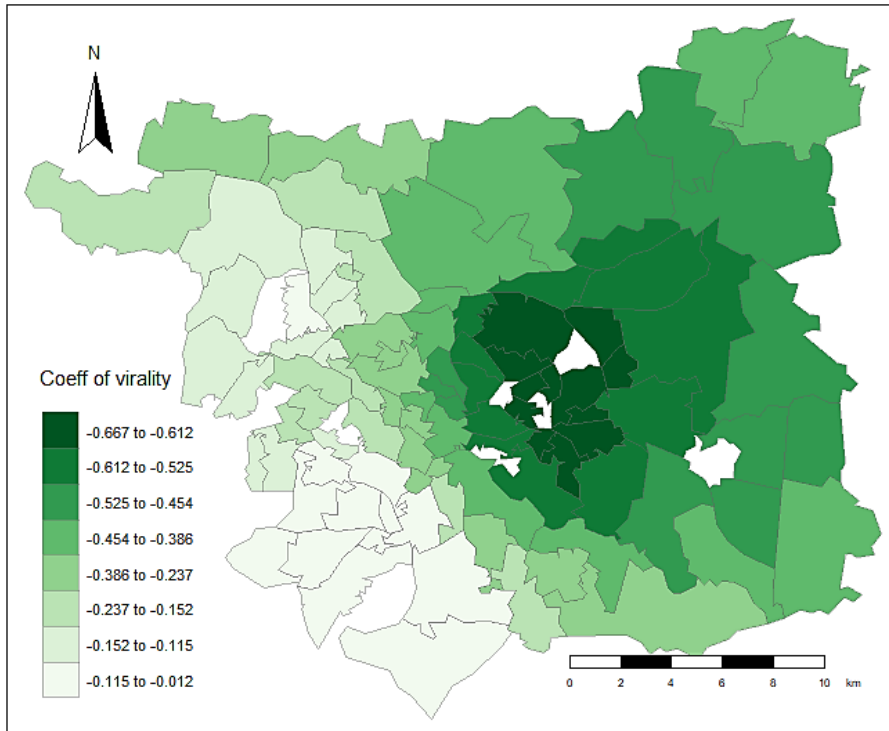


Figure 7.18: Map of coefficient of virality

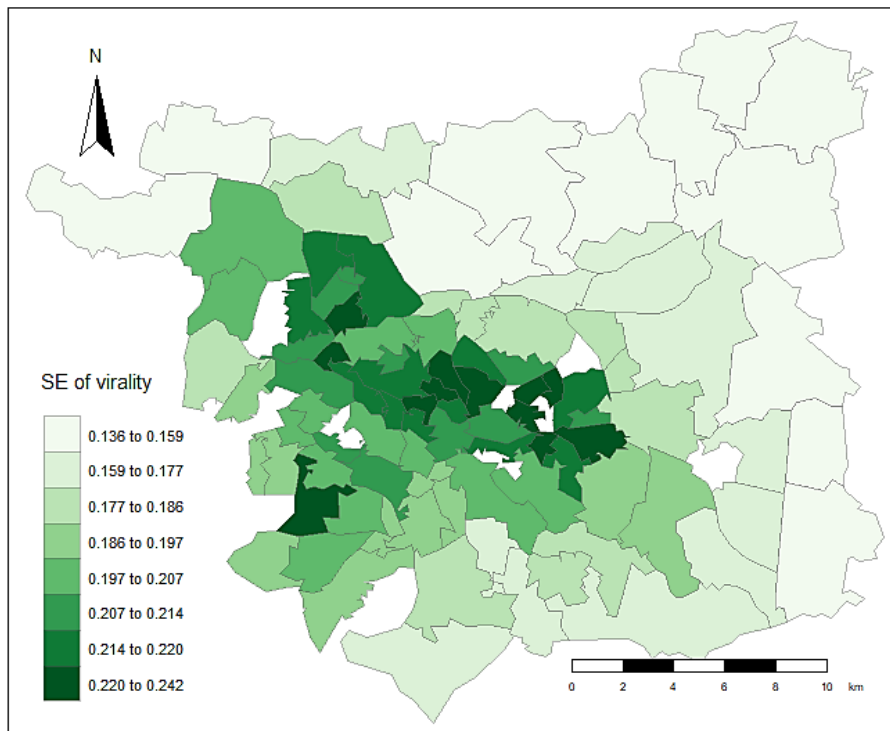


Figure 7.19: Map of standard error of the virality parameter

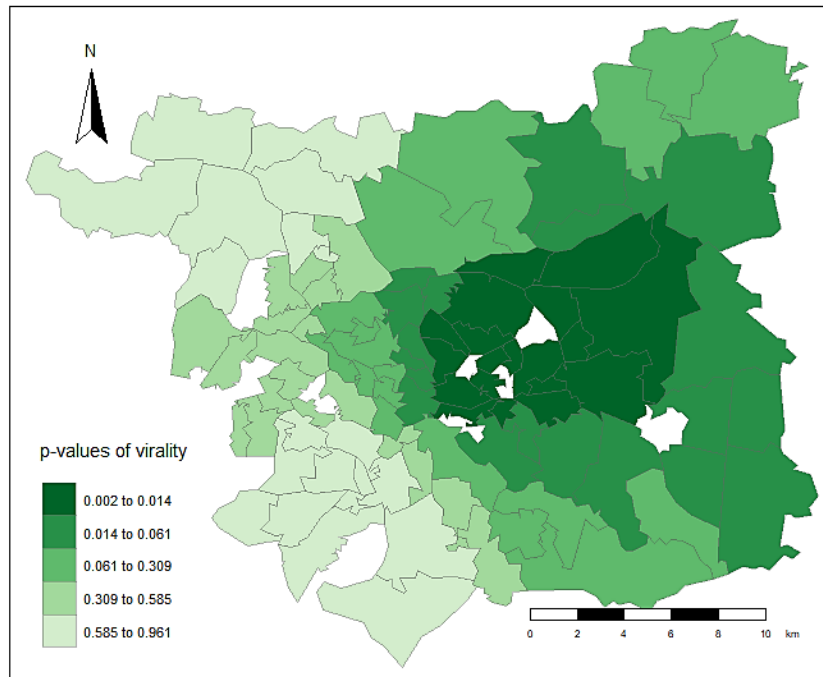


Figure 7.20: Map of p-values of the virality parameter

In terms of the effect of the coefficient of age diversity (Figure 7.21), it is clear that the communities in the northwest (such as Guiseley, Yeadon, Rawdon and Otley) tend to have larger effect (109.0 to 136.7). A potential explanation to this relationship might be that the effect of differences in age on burglary is more likely to be higher in areas with lower age diversity. It is also likely that a wide age range puts young offenders in close proximity with older victims with, potentially, more to steal. Additionally, it makes some sense that the broader the range of population characteristics in an area the more likely that there will be suitable target criteria for burglars making decisions about risk (Bernasco and Nieuwbeerta, 2005). The coefficient of age diversity has the largest variability in both global and local models (Table 7.6 and Table 7.7) respectively; suggesting the importance of age distributions in crime especially burglary. The spatial patterns of the parameter also indicate a higher likelihood that affluent areas might be victimised more than deprived areas. The relationship between age and burglary is described in Section 5.5.1.

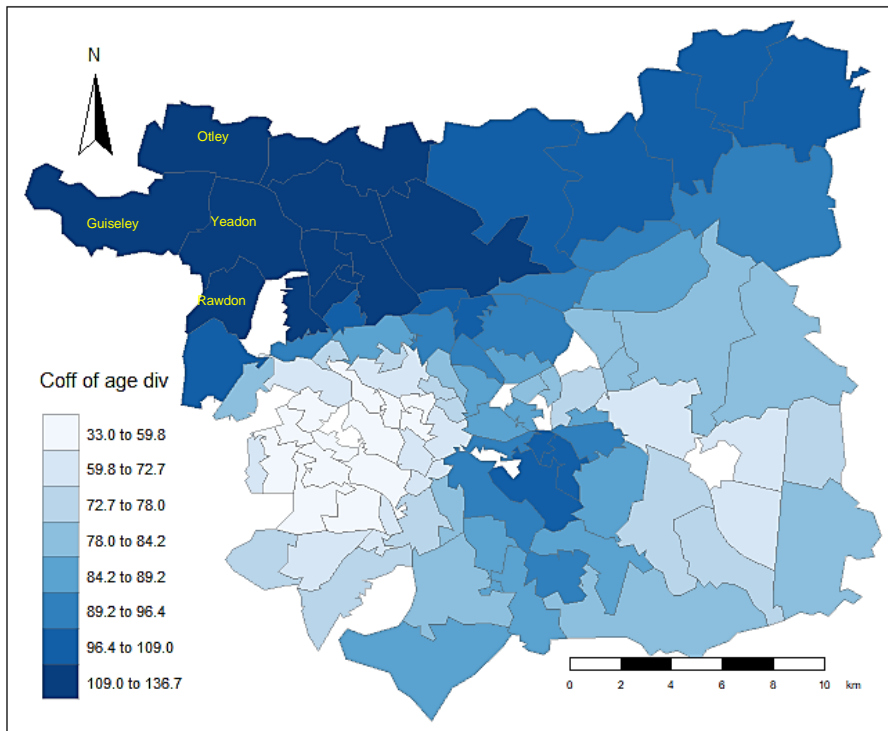


Figure 7.21: Map of age diversity coefficient

Furthermore, lower values indicated by the map of the standard error of the age diversity parameter (Figure 7.22) provides more confidence in the GWR model results, although, relatively higher standard error values were observed in areas like Yeadon and Rawdon indicating that the model might be missing some key parameters in those locations. Nevertheless, the distributions of the local p-value statistics (Figure 7.23) show significant relationships in the northeast community areas, indicating reliability of the model performance.

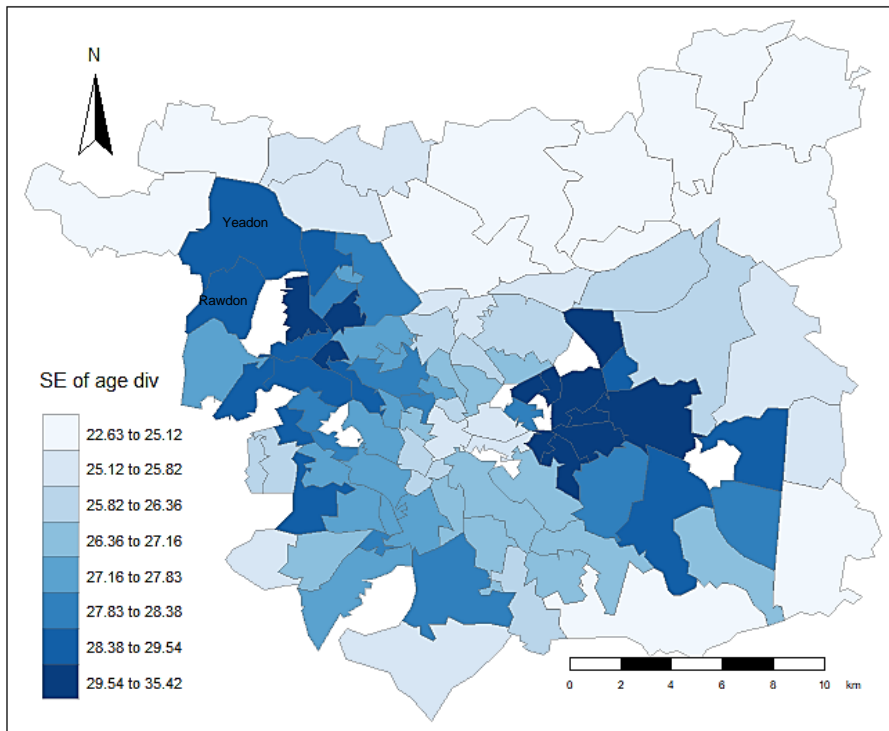


Figure 7.22 Map of the standard error of age diversity parameter

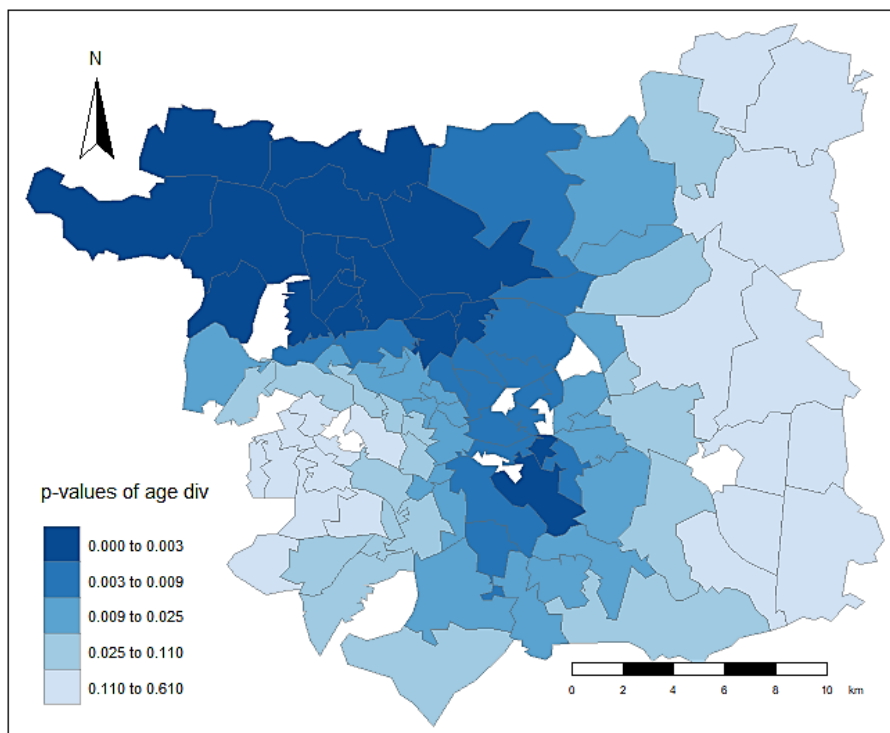


Figure 7.23 Map of p-values of the age diversity parameter

From the coefficient of residential diversity shown in Figure 7.24, it can be seen that the parameter has a larger effect in community areas in the southern edge of Leeds. These areas (including Methley, Rothwell and Ardsley East & West) are largely homogeneous and residentially stable. The interpretation of the relationship between residential instability and burglary is rather a complex one (Markowitz *et al.*, 2001; Martin, 2002). It is likely that the cost of crime profitability increases the likelihood of burglars to target residentially stable communities than their unstable counterparts (Ward *et al.*, 2014) and the risk of burglary victimisation is more likely to be higher where affluent neighbourhoods were surrounded by deprived neighbourhoods (Mburu and Bakillah, 2016). On the other hand, it is more likely that an increase in the residential diversity in an area to disrupt the bonds of social networks thereby increasing the risk of victimisation. Residential instability in a neighbourhood is associated with weak social ties and a low probability of residents connecting (Sampson *et al.*, 1997).

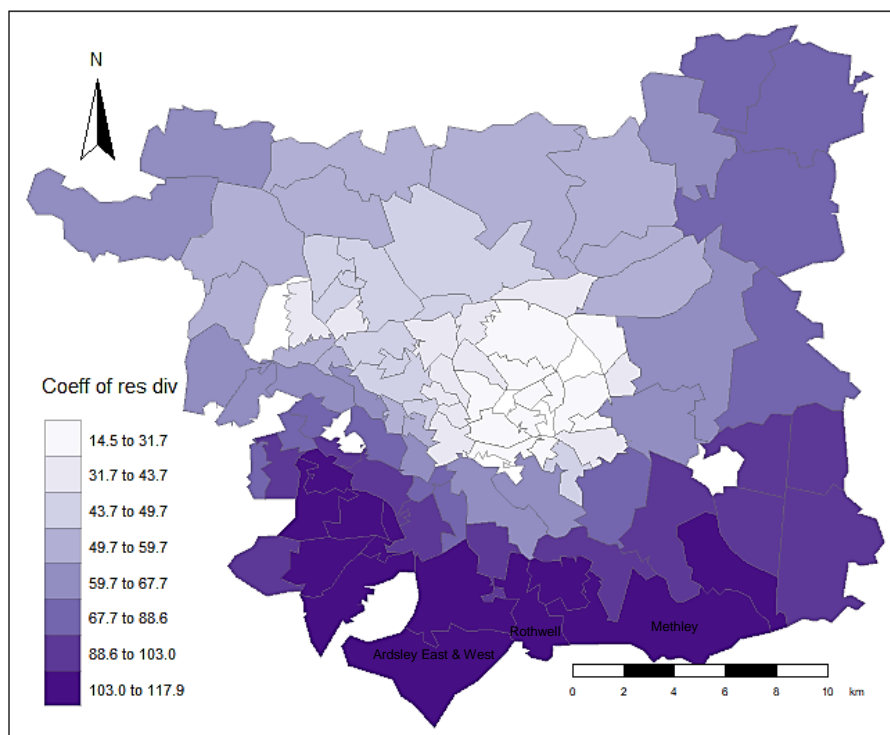


Figure 7.24: Map of residential diversity coefficient

It is reasonable to suggest that local variations in the relationship between community cohesion and burglary can be attributed to contextual variations in different areas (Malczewski and Poetz, 2005), though this relationship might not necessarily be picked up by the GWR model. In order to assess the performance of the model, the standard error statistics were all mapped (Figure 7.25); lower standard error values can be seen clearly in the north

and the south of Leeds. However, the distribution of the local p-values (Figure 7.26) shows those communities in the south were more significant, indicating the reliability of the model in those areas.

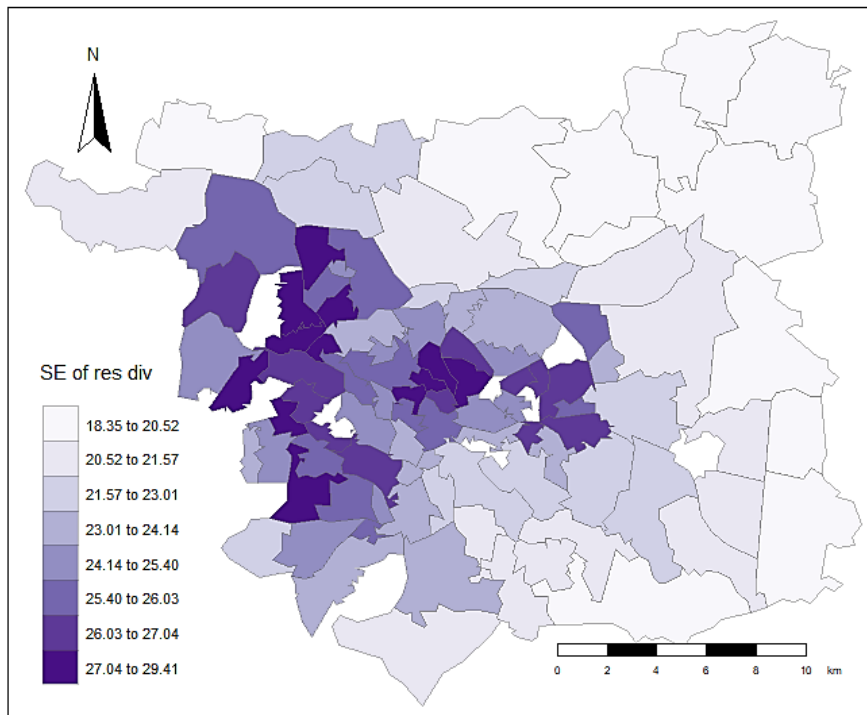


Figure 7.25: Standard error of residential diversity

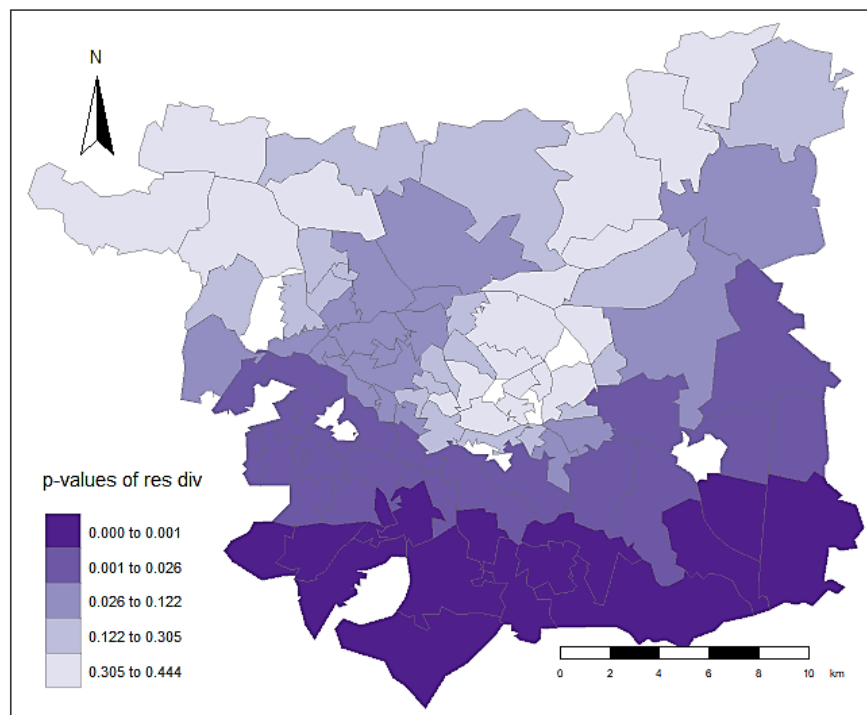


Figure 7.26: p-values of residential diversity

In terms of the spatial variation in the coefficient of the economically inactive population (Figure 7.27), shows more influence on community areas in the south (such as Belle Isle, Lofthouse & Robin Hood and Rothwell) and northeast including Guiseley, Otley and Yeadon, though the variability in the coefficient of the parameter is not large across Leeds. However, in the UK, statistics have consistently shows that economically inactive population are more likely to be victims of burglary crime than those who are economically active (ONS, 2014b; ONS, 2016a). The standard error statistics for this relationship shown in Figure 7.28 indicate that the values are relatively small across Leeds. Additionally, the local p-values were more significant in the south and northeast communities, suggesting reliability in the model results in those locations. The map of local p-values of the economically inactive variable shown in Figure 7.29 also indicates a statistically significant (p-values) relationship towards areas to the northeast (p-values <0.05) and southern (p-values <0.01) parts of Leeds respectively. This further provides confidence in the results of the local GWR model in these areas.

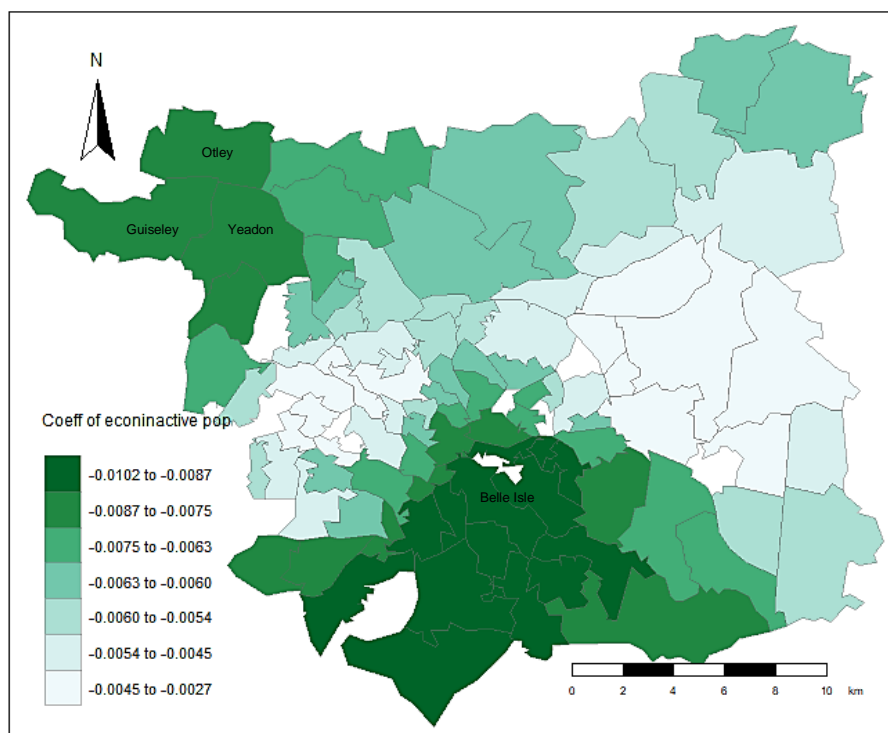


Figure 7.27: Coefficient of economically inactive population

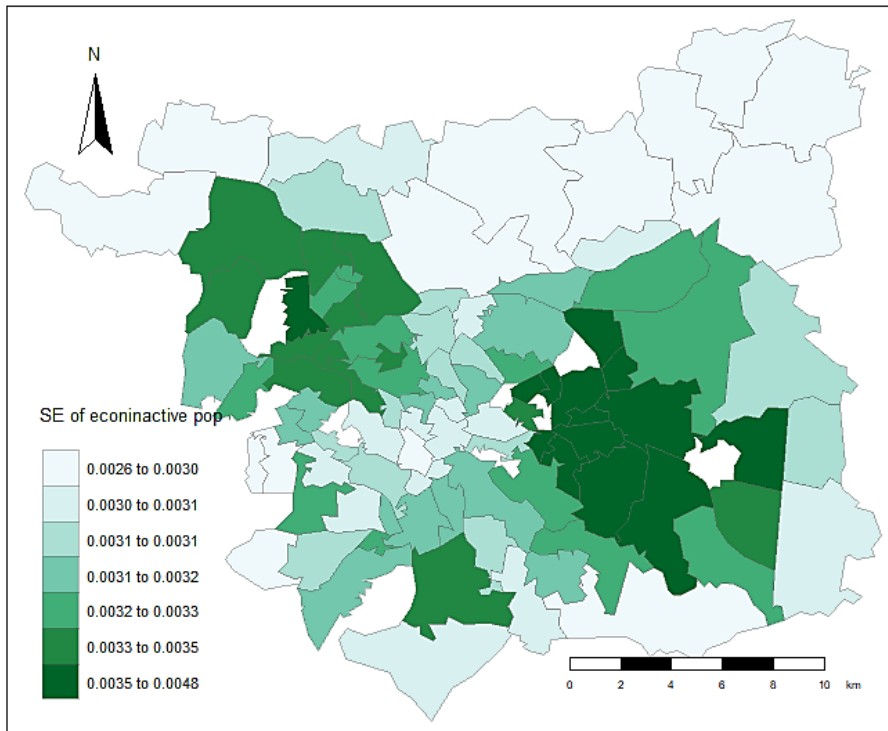


Figure 7.28 Standard error of economically inactive population

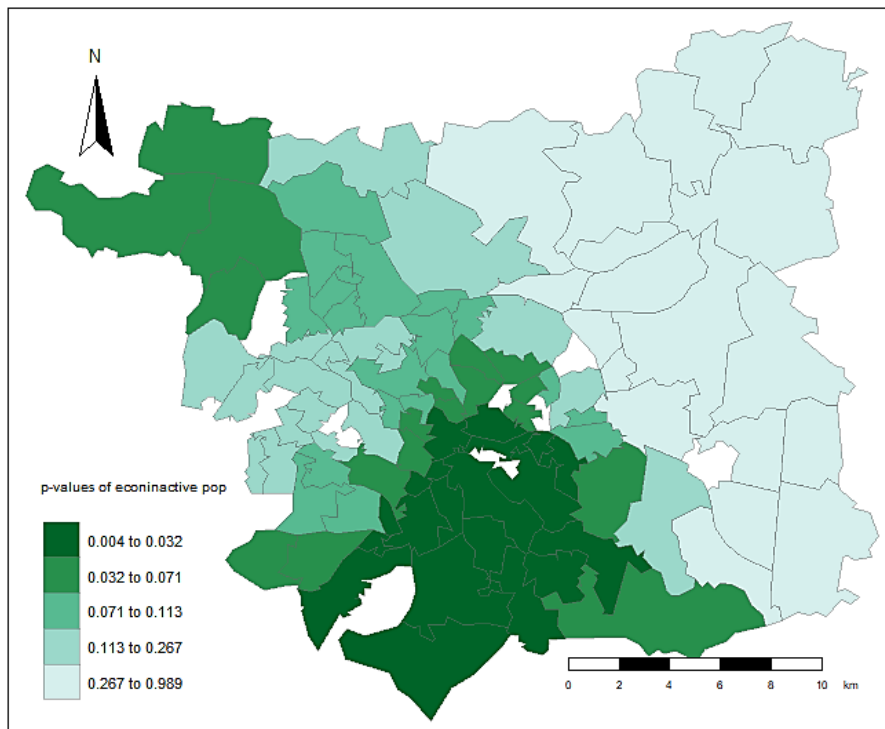


Figure 7.29 p-values of economically inactive population

7.7 Concluding Remarks

This chapter has extended the findings from chapter 6 on investigating the feasibility of using new data sources such as social media to explore community cohesion and crime which forms an important part of this research. Consequently, the chapter has employed data from 94 Leeds community areas Facebook pages and groups to quantify engagement rates. The chapter began by describing the importance of social media engagement on social media especially Facebook; then proceeds to explain how Facebook can be used for the generation of social capital in the community areas. It was also highlighted that post content can influence interaction on Facebook and potentially increases engagement rate on a community page. It can be assumed that higher engagement rate on a community Facebook page is potentially a strong indication of social cohesion in the local area, which could potentially be useful for predicting off-line community cohesion. Although the engagement rate is limited to the number of fans who choose to follow updates on community Facebook pages and groups, sampling bias is likely. Additionally, like any social media, the numbers of fans are subject to change over time which can affect the post interaction metrics. Meaning that engagement rate is also subject to change as the interaction behaviour of users changes.

However, the new metrics are potentially useful for highlighting insights into social engagement in different community areas, therefore appropriate for adoption as a proxy for exploring community cohesion. This is especially important owing to lack of data to quantify community cohesion from the official government data. Additionally, the new metrics are potentially useful alongside the traditional variables of regression for modelling crime (burglary specifically), as demonstrated in this research.

An intuitive finding from the regression models (global OLS and local GWR) indicates the virality of information about community safety is associated with decreases in burglary rates across Leeds community areas, highlighting the key contribution of the new metric of post share (V) in quantifying community cohesion and burglary rates. Additionally, the inclusion of the metric of engagement has greatly improved the performance of the traditional regression model in chapter 5. Although the PCV indicators are collectively useful for measuring community engagement, it can be argued that the metric of *virality* is the most important variable contributing in modelling burglary rates. The findings also suggest the need for an increasing SM network engagement in the community areas especially, where higher crime rates are prevalent. While facilitating collective community building, SM networks can also enhance community action by creating awareness about community safety,

therefore complimenting the law enforcement efforts in crime reduction. Finally, the insights derived from the analysis in this chapter and previous chapters (5 and 6) will be used in chapter 8 for constructing a new Leeds community areas classification profile, so as to advance understanding of the relationship between community cohesion and crime.

Chapter 8

Community Cohesion and Crime: A New Leeds Community Areas Classification Profile

8.1 Introduction

The main aim of this research is to explore the relationship between community cohesion and crime through the use of new social media data sources. In the previous chapters (6 and 7), social media data were used to explore community cohesion and crime in contrasting community areas of Leeds. Specifically, chapter 6 employed a *sentiment analysis* method on Twitter data to quantify community cohesion in different areas. Chapter 7 involved *engagement analysis* using Facebook data for modelling community cohesion and burglary rates. Earlier in chapter 5, diversity statistics were constructed and used in a *traditional regression* model of community cohesion and crime (burglary) in Leeds. This chapter brings together insights derived from the analyses of traditional and social media data used in the previous chapters to develop a new understanding of the relationship between community cohesion and crime for Leeds community areas. The objective of any classification is to group together areas with similar characteristics as much as possible (Kaufman and Rousseeuw, 1990). Section 8.2 explains the importance of community classification. The cluster analysis approach taken, as well as the methods employed for the clustering, are described in Section 8.3. An interpretation of the classification is presented in Section 8.4, with Section 8.5 providing concluding remarks.

8.2 Importance of Community Classification

Spatial classification of neighbourhoods has long been an area of interest in geography and urban sociology (Singleton and Longley, 2009). Social scientists have developed a tradition of mapping variations in urban neighbourhoods in order to better understand patterns and characteristics (Spielman and Thill, 2008). The spatial location and composition of different areas is an integral part that characterises urban community form. Urban community form refers to the different characteristics of neighbourhoods in terms of their socio-economic and demographic structure (Arundel and Ronald, 2017). The structure and composition of urban community is an important factor determining the potential for social interactions between residents (Delmelle *et al.*, 2013), generation of social capital (Putnam, 1995), greater chances of building a cohesive community (Cabrera *et al.*, 2017) and has been shown to influence the

social behaviour of members, especially the young (Leventhal and Brooks-Gunn, 2000; Maiese, 2003). While the concept of classification also includes the environmental characteristics of a given area and its land uses (Dempsey *et al.*, 2010), this aspect of classification is beyond the scope of this research. Environmental characteristics include soils, climate and vegetation; while land use includes the transportation network, residential and commercial activities.

A better understanding of urban community setting is important for policy planning, especially with regards to community building and increasing neighbourhood safety and well-being of society (Power, 2004; Amin, 2008). Local community classifications can also provide an opportunity for a detailed analysis of their characteristics and enhance better understanding of the dynamics of their social relationships (Clark, 2007). For example, Warner and Rountree (1997) argued that community structural characteristics such as residential instability, heterogeneity and socio-economic disadvantage limit the ability of residents to establish meaningful social ties. In such situations, collective community action, especially for controlling delinquent behaviour, is difficult (Sampson and Groves, 1989; Bursik Jr and Grasmick, 1993). Consequently, a change in demographic settings of a community is likely to have implications for social composition (Rees *et al.*, 2012) and the extent of social relationships that determine cohesion, especially in areas with a large diverse population (Nandi and Platt, 2014).

Previous classifications have concentrated on grouping neighbourhoods together with a view to identifying areas for the marketing of products/services and such attempts have given rise to commercial geodemographic classifications like A Classification of Residential Neighbourhoods (ACORN), MOSAIC from Experian and CAMEO from CallCredit (Brunsdon *et al.*, 2016). More recently, an open source geodemographic classification has been developed entirely from 2001 Census statistics (Vickers and Rees, 2007) and another based on 2011 Census data (Gale *et al.*, 2018). Unlike earlier classifications, more recent studies have attempted to classify the spatial pattern of urban community areas towards social applications (Brunsdon *et al.*, 2016). For example, in Leeds, Stillwell and Phillips (2006) employed geodemographic classification of ethnic compositions. A similar classification was also implemented based on profiling neighbourhoods for community safety, albeit to identify areas based on similarity of crime (Shepherd, 2006). However, research into the geography of community groups, especially relating to their social cohesion, is limited (Smith *et al.*, 2011). This research provided the first academic attempt to classify Leeds' community areas based

on community cohesion and crime. This new classification will have policy implications for resource allocation and community building.

8.2.1 Communities and Crime

Studies into community structures and their effects on crime originate from social disorganisation theory, Shaw and McKay (1942) argued that communities characterised by deprivation, residential instability and ethnic heterogeneity are prone to crime due to lack of social control; these communities are regarded as *disorganised communities*. The concept of the theory has been extensively used in a number of studies (e.g. Sampson and Groves, 1989; Sampson *et al.*, 1997; Veysey and Messner, 1999; Choi and Choi, 2012; Teasdale *et al.*, 2012; Livingston *et al.*, 2014; Armstrong *et al.*, 2015). In these socially disorganised communities characterised by weak social ties, realisation of common goals especially to solve local problems (such as crime) is difficult (Kubrin and Weitzer, 2003; Saggat *et al.*, 2012), and the capacity of members to entrench social control (ability to constrain violation of norms) in their communities decreases (Pickett, 2013), meaning crime is likely to flourish (Bellair, 1997; Markowitz *et al.*, 2001; Kubrin and Weitzer, 2003). Studies have found support for the relationship between social disorganisation and crime (e.g. Sampson and Groves, 1989; Takagi and Kawachi, 2014). It was argued that communities having a wider interaction are likely to establish stronger social ties and to have a greater potential of being cohesive, with the ability for informal social control (Warner and Rountree, 1997).

However, understanding the complexities of different features existing in the community areas, especially with regard to the degree of social cohesion, is a difficult process (Hickman *et al.*, 2008). Piekut *et al.* (2012) stressed that communities could have differing qualities identified in two ways: from one viewpoint, a spatial region could be made of smaller groups which are heterogeneous both internally (referring to ethnic differences) and externally (referring to socio-economic differences); in contrast, groups could be homogeneous inside, yet unique from each other (Piekut *et al.*, 2012). Additionally, Ashby (2005) argued that communities tend to differ not only socio-economic and demographic terms but also in terms of their crime rates and victimisation. As the importance of understanding community cohesion is of increasing interest for neighbourhood safety (Ashby, 2005), coupled with recent increases in crime rates in West Yorkshire (Flatley, 2017), there is the need for a detailed analysis of the socio-economic and demographic composition of different community areas of Leeds, with a view of gaining insight into their perceived social cohesion

especially relating to crime. Leeds urban communities are changing demographically; Leeds is currently the home to over 140 ethnic groups, making it the second most diverse city outside London (Leeds City Council, 2017). Crime has been a particular problem (Leeds City Council, 2014); it is important to have a clear understanding of the populations in different areas in order to guide policies appropriate for community building (Leeds City Council, 2014). For example, Leeds Metropolitan District comprises of 106 community areas (Stillwell and Phillips, 2006), therefore, when areas are clustered into few groups based on similarity of their properties, can enhance our understanding of the dynamics in the areas that contribute towards crime (Vickers and Rees, 2007).

8.2.2 Issues with Previous Classification Attempts

The underpinning principle of neighbourhood classification is to group areas that are most likely to contain similar structural characteristics, even though they may be geographically distinct (Leventhal, 2016). Additionally, the classification of an area into a given group may be used to describe the likely characteristics and behaviour patterns of its residents (Leventhal, 2016). Previous attempts to classify areas used data mainly from Census statistics (Vickers and Rees, 2007; Lansley *et al.*, 2015). Census data are the most comprehensive and dependable socio-economic and demographic data available in the UK and commonly employed for area classification (Vickers and Rees, 2007). Although useful, however, the analysis of Census data collected on a decennial basis alone rarely provides measures of the relationship between urban community form and crime (Sampson and Groves, 1989). Additionally, the use of standard socio-demographic variables (e.g. age, ethnicity, employment and education) as associated with previous attempts rarely captures the nuanced concept of community cohesion; this might limit better understanding of the dynamic composition of spatial urban communities (Piekut *et al.*, 2012). Furthermore, the advent of new data sources ('big data') from social media has profound implications on our understanding of the dynamics of urban neighbourhoods (see Section 4.4) (Adnan *et al.*, 2013). In order to better understand the socio-demographic diversity of urban communities, especially in the context of community cohesion and crime, the combination of multidimensional variables such as those from social media (i.e. Facebook and Twitter) and crime need to be included in the analysis (Piekut *et al.*, 2012). The combination of new social media sources with contextual variables from Census statistics can lead to the generation of new community typologies (Leventhal, 2016). However, classification is criticised due to the ecological fallacies (Openshaw, 1984) which arises when results of an analysis based on areal

level aggregate statistics (such as output areas) are assumed to apply at the individual level (Tranmer and Steel, 1998). However, in this research, classification is performed at community area level so as to simplify description of different areas based on their characteristics.

In this research, in order to create a new typology for Leeds' community areas, a range of multidimensional variables will be employed in the clustering analysis in order to classify areas into groups based on the similarity of their characteristics, so as to explore the relationship between social cohesion and crime in different community areas of Leeds.

8.3 Cluster Analysis

Cluster analysis is the process of classifying objects into homogeneous groups (clusters) from datasets in which the number of groups and characteristics are unknown (Kaufman and Rousseeuw, 1990; Mirkin, 2012). Clustering is a common technique for statistical data analysis used in different fields including social sciences (Bijuraj, 2013). As classification is aimed at grouping objects with similar characteristics which are distinct from objects in different groups, it is important as it reveals new insights about an area (Kaufman and Rousseeuw, 1990). The multifaceted nature of classification building means that no solution can be regarded as perfect, especially with different choices of variables, and the clustering algorithm or number of clusters is liable to produce a different solution each time (Dennett and Stillwell, 2011 p165). A number of criteria have been proposed for cluster analysis by different authors. For example, Halkidi *et al.* (2001) suggested four steps for clustering including feature selection, the clustering algorithm, validation and interpretation. Feature selection involves more accurate selection of features (data pre-processing) on which clustering is to be performed; the clustering algorithm refers to the choice of an algorithm and number of clusters expected from the data should be decided by the researcher; while validation refers to the evaluation of cluster results using appropriate techniques; and interpretation is necessary in order to draw the right conclusion. However, a more comprehensive approach to cluster analysis for area classification is provided by Milligan (1996), extended by Everitt *et al.* (2001) and used by Vickers and Rees (2007) and is considered more appropriate in this research. The seven steps are outlined in Section 8.3.1 to Section 8.3.7 .

8.3.1 Clustering Elements (Objects to Cluster)

Clustering elements refer to geographic (neighbourhoods) coverage where the cluster analysis will be based. For example, in this research, Leeds' community areas will be used as neighbourhoods for the cluster analysis. Community areas are chosen because they are recognised by Leeds residents as representing areas large enough for meaningful social cohesion (Stillwell and Phillips, 2006). A detailed description of Leeds community areas is provided in Section 3.5.

8.3.2 Clustering Variables

These refer to attributes representing areas to be clustered. Variable selection is an important process in clustering and classification, however, determining which variables are 'important' can be difficult (Andrews and McNicholas, 2014). There should be a good reason for a variable to be included in the analysis and noisy variables that are likely to distort significant patterns should be dropped from the analysis. In clustering, noisy variables are strongly correlated with each other and their distribution is almost similar in every cluster (Fraiman *et al.*, 2008). In this study, combinations of multidimensional variables from a range of data sources are used for the cluster analysis. Table 8.1 presents the datasets used and their sources. The variables were chosen based on literature that highlights their significance as indicators of community cohesion. A detailed theoretical discussion on Census and crime variables is provided in Section 5.5 while Section 4.2 has presented a review on social media and community cohesion.

Table 8.1: Variables used in the cluster analysis of Leeds community areas.

Variables	Components included
Census	
Age diversity	Age groups:10-14, 15, 16-17, 18-19, 20-20, 25-29, 30-44, 45-59, 60-64, 65-74
Ethnic diversity	All 18 ethnic groups identified in the 2011 Census.
Education diversity	Age:16-over qualification level 1, 16-over qualification level 2, 16-over qualification level 3, 16-over qualification level 4
Employment diversity	Age:16-64 Managers/Directors, 16-64 Professionals, 16-64 Associate Professionals, 16-64 Administration and Secretariat, 16-64

Variables	Components included
	Skilled Trade, 16-64 Caring Leisure and Services, 16-64 Customer Services, 16-64 Process Plants and Machines, 16-64 Elementary Occupation
Residential length diversity	Length of residence: Less than two years, Less than five years, More than five years, Ten years above, Born in the UK
Crime	
Burglary	Rate per 1000 population
Vehicle	Rate per 1000 population
ASB	Rate per 1000 population
Violence and sexual	Rate per 1000 population
Social media	
Positive sentiment	Percentage of positive tweets in an area
Negative sentiment	Percentage of negative tweets in an area
Engagement rate	Total likes + Total comments + Total shares/number of Fans on a community Facebook profile.

Correlation diagnostics performed on the variables indicates that some variables were highly correlated (Table 8.2). A strong correlation between variables is undesirable because this can result in data redundancy (Vickers and Rees, 2007). In this analysis, consideration is given to ensure that variables are not highly correlated such that each variable contributes a unique dimension to the classification. For example, a strong correlation was found between ethnic diversity and length of residence diversity ($r = 0.98$); between positive sentiment and negative sentiment ($r = -1$); and between antisocial behaviour (ASB) and violent offences ($r = 0.95$) respectively. However, a higher positive or negative correlation between variables may not necessarily justify their removal (Shepherd, 2006). Vickers and Rees (2007) suggested that variables that are likely to predict the value of other variables can be retained in the analysis because they are likely to predict the behaviour of the classification. In this research, the correlated variables are retained because both are important indicators of social cohesion in the community areas. Removing either of these may distort the behaviour of other variables in the analysis.

Table 8.2: Pearson's correlation of the clustering variables

	Eng_rate	BurgRate	ASB	Veh_Crime	Viol_sexual	EthDiv	ResDiv	AgeDiv	EmplDiv	EduDiv	Psentiment
Eng_rate											
BurgRate	-0.02										
ASB	-0.07	0.56**									
Veh_Crime	-0.15	0.61**	0.7								
Viol_sexual	-0.1	0.62**	0.95**	0.68**							
EthDiv	-0.09	0.54**	0.45**	0.45**	0.57**						
ResDiv	-0.09	0.57**	0.53**	0.52**	0.65**	0.98**					
AgeDiv	-0.12	0.51**	0.45**	0.39**	0.46**	0.48**	0.53**				
EmplDiv	-0.01	-0.22*	-0.33*	-0.08	-0.34**	-0.39*	-0.36*	-0.11			
EduDiv	0.09	-0.23*	-0.07	-0.15	-0.07	-0.24*	-0.26*	-0.76**	-0.09		
Psentiment	-0.03	-0.01	-0.01	-0.09	-0.01	-0.05	-0.06	0.07	0.11	-0.15	
Nsentiment	0.03	0.01	0.01	0.09	0.01	0.05	0.06	-0.07	-0.11	0.15	-1

** Correlation is significant at the 0.01 level (2-tailed), * Correlation is significant at the 0.05 level (2-tailed).

8.3.3 Variable Standardisation

Standardisation is a method for the adjustment of means of the data applied in order to control for extraneous variables commonly used in geodemographic research (Kalton, 1968; Lane and Nelder, 1982). Examples of variable standardisation methods include range standardisation (often referred to as min:max standardisation) and standardised z-scores. Standardisation is required in most clustering processes because of the variability of measurements used in different variables. The type of standardisation method to be employed should be decided by the type of data because, unlike regression analysis, cluster analysis has no mechanism for detecting relevant and irrelevant variables (Cornish, 2007). Additionally, it is necessary to ensure that each variable is equally represented in the cluster analysis. Variables with a wider range than others tend to dominate the analysis (Vickers and Rees, 2007) and for this reason variable standardisation is recommended (Cornish, 2007). Furthermore, Milligan (1996) stressed that variable standardisation reduces the effect of outliers in the datasets. The most common data standardisation method used in the literature is the *z-score* (Yim and Ramdeen, 2015). A *z-score* is the statistical measurement of a score's relationship (standard deviation) to the mean, having a mean of 0 and variance of -1 and +1 (Abdi and Williams, 2010). In this research, *z-score* is used to standardise the variables prior to cluster analysis. *z-scores* are computed as:

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (8.1)$$

where Z_i is the standardised value of the variable in area i , X_i is the original value of the variable for area i , μ is the mean and σ is the standard deviation of the variable.

Figure 8.1 shows the boxplots of standardised variables constructed in R environment.

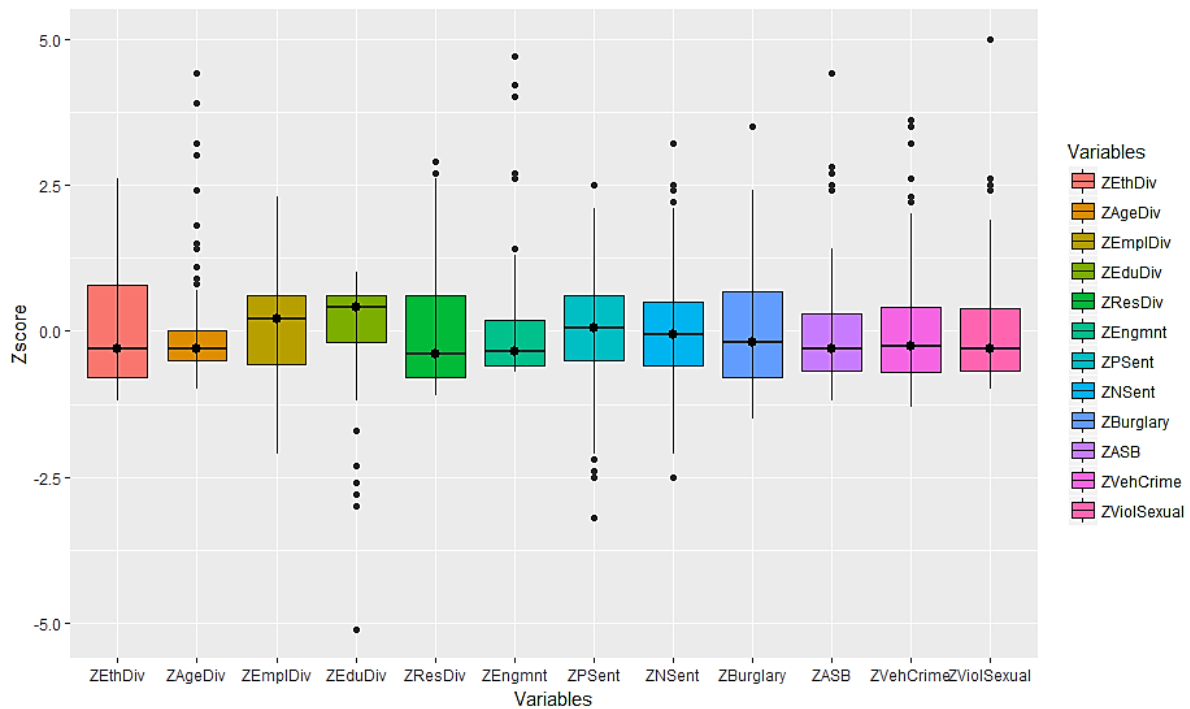


Figure 8.1: Boxplots for standardised variables retained for clustering

It can be seen from Figure 8.1 that except for the ethnic diversity and employment diversity variables, the rest of the variables were characterised by outliers that has been reduced by standardisation. Standardising enabled scaling of the variables so that equal weight is assigned to them with a median z-score between +1 and -1.

8.3.4 Measure of Proximity

Since the basis for clustering is grouping similar objects, a measure that can determine whether two objects are similar or dissimilar is required (Rokach and Maimon, 2005). Depending on the data type, different distance measures of similarity or dissimilarity such as Euclidean, Manhattan and Minkowski distances are commonly used (Irani *et al.*, 2016). Euclidean is simply the ordinary distance from two points, considered a standard metric for geometrical problems; Manhattan distance is a metric that calculates the absolute difference between coordinates of a pair of data objects; and Minkowski distance is a generalised metric that can be used for ordinal variables (Irani *et al.*, 2016). Many clustering algorithms such as partitioning (e.g., K-means) and divisive (e.g., hierarchical) use distance between points to cluster similar data points together, their efficiency influences the performance of clustering (Shirkhorshidi *et al.*, 2015). Additionally, for continuous data (such as that used in this research), Euclidean distance is widely recommended (Shirkhorshidi *et al.*, 2015). Euclidean

distance as a measure of dissimilarity reports larger values as less similar. It is computed by summing the square root of the differences between data points. Equation 8.2 for Euclidean distance is given based on Madhulatha (2012):

$$d = \sqrt{\sum_{j=1}^n (x_j - y_j)^2} \quad (8.2)$$

where n is the number of variables, and x_j and y_j are variables respectively.

8.3.5 Clustering Methods

A number of clustering methods exist in the literature. The choice of clustering algorithms should be decided with knowledge of their robustness to different kinds of data and the type of clusters suspected to be present (Jain *et al.*, 1999). The multifaceted and heuristic nature of different classification building algorithms means that no solution can be regarded as perfect (Dennett and Stillwell, 2011). Clustering algorithms can be broadly grouped into two types: *partitional* and *hierarchical* (Ng and Han, 1994; Kaufman and Rousseeuw, 2005). Partitional algorithms determine all clusters at a time, whereas hierarchical algorithms find successful clusters using previously established clusters (Madhulatha, 2012).

8.3.5.1 Hierarchical Methods

Hierarchical clustering is devoted to representing data in the form of hierarchy over the entity set (Mirkin, 2012). There are two approaches to building a hierarchical cluster namely *agglomerative* and *divisive*. The agglomerative approach builds a hierarchy of clusters in the bottom-up fashion, starting from the smaller clusters and sequentially merging according to similarity. The agglomerative clustering procedure is a widely used hierarchical method. Unlike agglomerative methods, the divisive approach builds a hierarchy of clusters in the bottom-top fashion by splitting larger clusters into smaller ones starting from the entire dataset (Sarstedt and Mooi, 2014).

The advantage of hierarchical methods is that their output can be graphically displayed which is far more informative than the unstructured clusters produced by partitional methods (Rani and Rohil, 2013). However, the hierarchical algorithms suffer from the lack of back-tracking

capability; once clusters are formed they cannot be disbanded (Rokach and Maimon, 2005; Everitt *et al.*, 2011). Similarly, once objects are separated they will never be re-grouped into the same cluster (Ng and Han, 1994). Additionally, hierarchical methods do not handle large datasets and are sensitive to outliers (Fahad *et al.*, 2014). These set of drawbacks associated with hierarchical methods makes them unsuitable for this research.

8.3.5.2 Partitional Methods

Partitional methods relocate objects by moving them from one cluster to another starting by an initial partitioning (Rokach and Maimon, 2005). There are different types of partitional algorithms which include K-means, K-medoids: partition around medoids (PAM), clustering large applications (CLARA) and clustering large applications based on randomised search (CLARANS). The basic idea of partitioning algorithms is to find structures that minimise a certain error criterion which measures the distance of each object to its cluster centre value. This criterion is known as the sum of squared error (SSE), which measures the squared Euclidean distance of objects to other centres (Rokach and Maimon, 2005). The simplest and most widely used algorithm employing the SSE criterion is the K-means algorithm.

K-means

K-means is an iterative relocation algorithm, where objects are assigned and re-assigned to clusters which have less within-cluster variation (squared distance from each observation), preferred because this allows subjects to be moved from one cluster to another unlike in hierarchical cluster analysis (Cornish, 2007). K-means clustering was first applied by MacQueen (1967), though originally proposed by Stuart Lloyd in 1957 as a clustering technique, but was only published in 1982 (Lloyd, 1982). K-means clustering is a technique within unsupervised learning that can be employed to better understand the structure of the underlying data (Beil *et al.*, 2002), is relatively simple and fast, with the capability of handling large datasets (Zhang *et al.*, 2008). However, it requires a researcher to determine the appropriate number of clusters (K) and the order in which they are stored can also affect the partition (Kaufman and Rousseeuw, 1990). Furthermore, K-means is sensitive to outliers (Rokach and Maimon, 2005) and each run produces different results (Pavan *et al.*, 2012). The K-means clustering algorithm has been widely used in previous studies (Vickers and Rees, 2007; Piekut *et al.*, 2012; Lansley *et al.*, 2015; Mohit *et al.*, 2016). Equation 8.3 presents the K-means algorithm based on Faber (1994):

$$E = \sum_{i=1}^k \sum_{j=1}^{n_i} \|X_{ij} - Z_i\|^2 \quad (8.3)$$

where E is the sum of all the variances, x_{ij} is the j th point in the i th cluster, z_i is the reference point of the i th and n_i is the number of points in that cluster. The notation $\|x_{ij} - z_i\|^2$ stands for the distance between x_{ij} and z_i .

In this research, an exploratory analysis of the datasets was experimented with using SPSS in order to guide the selection of the optimum number of clusters. Though there is an issue with conducting K-means in SPSS this is to do with randomisation of the initial seeds. One way around this is to specify the starting seeds so these do not change each time the algorithm is re-run.

Partition Around Medoids (PAM)

PAM is similar to the K-means algorithm; it differs from the latter mainly in its representation of different clusters (Kaufman and Rousseeuw, 1990). In the PAM algorithm, each cluster is represented by the most centric object (medoids) in the cluster rather than the cluster mean. It begins by selecting an object as the medoid of each cluster, then each of the non-selected objects is grouped with the medoid to which it is most similar; the algorithm then swaps medoids with other non-selected objects until all objects qualify as medoids (Halkidi *et al.*, 2001) The K-medoid method is more robust than K-means in the handling of noise and presence of outliers (such as in this research), because a medoid is less influenced by outliers (Rokach and Maimon, 2005; Berkhin, 2006). However, PAM has a higher computational cost than the K-means method, with both methods requiring the researcher to specify the number of K clusters from the outset. PAM has been shown to perform well when used on datasets (few hundreds) in not more than 5 clusters (Ng and Han, 1994).

Furthermore, visualisation of the cluster solution is crucial for verifying the goodness of any clustering algorithm (Halkidi *et al.*, 2001). In this direction, the PAM algorithm has the advantage of graphical display known as *silhouette plots*. Silhouette plots show which object lies within their cluster, and which ones are merely somewhere in between clusters (Kaufman and Rousseeuw, 1990). They can also be used to aid selection of an appropriate number of clusters and validation of cluster analysis. Additionally, silhouettes offer the advantage that

they only depend on the actual partition of objects and not on the clustering algorithm that was used to obtain them; therefore, could be used to improve cluster analysis (e.g. move objects with a negative value to their nearest neighbour) (Rousseeuw, 1987: P59). Silhouette values close to 1 indicate that objects are well placed in their cluster and values closer to 0 indicate that the fit was not good while negative values indicate that observations are probably placed in a wrong cluster. Equation 8.4 for computing silhouettes is given based on Rousseeuw (1987):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (8.4)$$

where $a(i)$ is the average dissimilarity of (i) to all other objects in the same cluster, $b(i)$ is the lowest average dissimilarity of (i) to the other objects in all other clusters in the whole solution.

Clustering can be performed using different software and how the initial clusters are seeded differs between the algorithms (programs) used. For example, Vickers and Rees (2007) used SPSS to performed K-means clustering and Dennett and Stillwell (2011) used MATLAB in their clustering analysis. Dennett and Stillwell (2011) argued that SPSS is not flexible in the assignment of initial cluster centroids, such that different clusters solutions are produced using k-means algorithm in SPSS on the same dataset. While these programs are useful in different ways for clustering, they are limited in graphical output and diagnostic information that can guide the decision on the best possible cluster solutions. In this research, clustering was implemented using cluster and factoextra packages in R statistical environment (R Core Team, 2017). The *cluster* package Maechler *et al.* (2017) provides functions for the implementation of partitioning algorithms such as K-means and PAM clustering, while *factoextra* package of Kassambara and Mundt (2017) provides useful functions for an elegant visualisation of partitioning methods. Shepherd (2006) successfully implemented a similar approach in the R environment to create a neighbourhood area classification for community safety in Leeds.

Clustering Large Application (CLARA)

Similar to the PAM algorithm, clustering large applications (CLARA) was also developed by Kaufman and Rousseeuw (1990). Unlike PAM that finds representative objects (medoids) of

the entire dataset, CLARA draws multiple samples from the dataset then applies PAM on the samples and then outputs the best clustering groups from the sample. Therefore the algorithm minimises the computational cost associated with PAM which makes it suitable for clustering tasks on a large number of datasets (Kuang and Zhang, 2017). However, this method is constrained by a quality issue as using sampling techniques in clustering may not represent the initial dataset (Andritsos, 2002). Additionally, CLARA cannot find the best clustering especially where the sample drawn does not belong to the best medoids (Boomiya and Phil, 2008). These limitations make the application of CLARA unsuitable in the present research.

Clustering Large Applications Based on Randomised Search (CLARANS)

The CLARA algorithm was proposed by Ng and Han (1994). While CLARA compares samples in the dataset, CLARANS applies a random search to generate neighbours (Berkhin, 2006). This has the advantage of not confining a search to a localised area (Ng and Han, 1994). The complexity of CLARANS is in terms of the multiple numbers of points drawn while searching for the best medoids (Berkhin, 2006). Furthermore, both CLARA and CLARANS are based on the clustering criterion of PAM (i.e. distance from the medoids) (Andritsos, 2002). Consequently, after careful review different clustering methods available, in this research, a final decision was arrived to proceed with implementation of the PAM approach. This is because it was found to be the most appropriate method for the task ahead.

8.3.6 Choosing Number of Clusters

This is the most difficult decision to be made in cluster analysis. There are different procedures suggested for selecting the most appropriate number of clusters (K), however, no standard rules apply in the literature (Tibshirani *et al.*, 2001). A suitable number should be decided upon to reflect patterns within the dataset and research objective (Yim and Ramdeen, 2015). However, a number of methods have been proposed such as the elbow method, Akaike information criteria (AIC), Gap statistics and average silhouette width (Chiang and Mirkin, 2010). The *elbow* method is based upon the idea that one should choose the number of clusters such that adding another cluster does not give a better modelling of the data (Bholowalia and Kumar, 2014); *AIC* is a model selection criteria that is applied in a multivariate analysis such as clustering to assess the relative quality of a model (Akaike, 1981); *Gap* statistics is a measure of within cluster dispersion relative to the distribution of objects in the dataset (Tibshirani *et al.*, 2001); and *Silhouette width* is a measure of similarity (cohesion) of objects in their cluster and dissimilarity (separation) between clusters (Kaufman

and Rousseeuw, 1990). In this research, all these methods have been experimented with and evaluated before deciding on the optimal number of clusters to be used in the final partition. The optimal value of k was evaluated within the range of 2 to 10 clusters. The idea behind the choice of 10 cluster range is that the optimum number of clusters is likely to fall between these ranges, considering the number of variables in the dataset.

Elbow and Scree Plot Methods

The elbow and scree plot are methods implemented to guide the selection of an optimal number of clusters. The agglomeration schedule in the hierarchical method can be a starting point to explore a possible number of clusters in a dataset. The agglomeration schedule displays how clusters are progressively formed and the coefficient at each stage represents the distance between the objects being combined which can be observed by a scree plot (Yim and Ramdeen, 2015). A larger increase in coefficient value indicates an increase in dissimilarity and at that stage, it would be ideal to stop the clustering process. A scree plot is simply a line graph (a visual representation of the agglomeration schedule) which can be plotted in Microsoft Excel because SPSS does not produce scree plot in its output (Yim and Ramdeen, 2015).

Similarly, the rationale behind the elbow method is that increasing the number clusters at some point does not contribute to clustering; therefore at such point distortion begin to set in (Kodinariya and Makwana, 2013). However, the elbow method can also be ambiguous because sometimes there is no elbow or several elbows which makes it difficult to choose an optimal number of clusters (Kodinariya and Makwana, 2013). In this research, the K-means elbow plot was implemented in the R statistical environment. Figure 8.2 shows the scree plot of the coefficients of agglomeration performed using the hierarchical method (left) and elbow plot for the K-means clustering procedure (right).

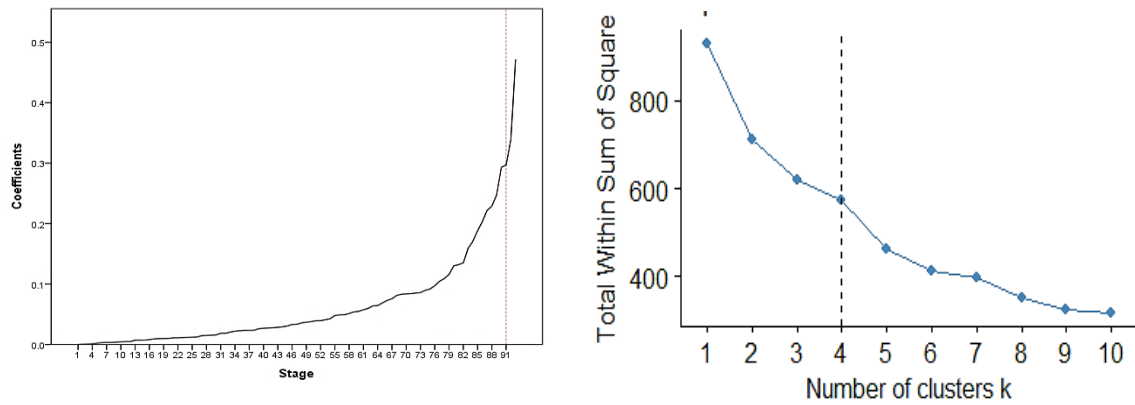


Figure 8.2: Scree plot of the agglomeration schedule (left) and Elbow plot for K =10 cluster solutions (right)

From Figure 8.2 (scree plot) it can be seen that the largest increase in dissimilarity occurred at stage 91 out of 94 stages (possible number of clusters are counted from right to left), while the elbow plot shows a break at four cluster solution. Both figures suggest that a four cluster solution is likely to be the most appropriate for the dataset in this research.

Akaike Information Criteria (AIC)

As highlighted in Section 8.3.6, AIC is used in this research in order to guide the best number of clusters to be implemented. Although a number of information criteria are used in the literature for determining the number of clusters, such as Bayesian Information Criterion (BIC) and Kullback Information Criterion (KIC) (Akogul and Erisoglu, 2017), AIC is widely used method for determining the number of clusters in different disciplines including social sciences (Snipes and Taylor, 2014). BIC is less favourable because it tends to overestimate the number of clusters (Mirkin, 2011). These information criteria are likely to produce different results for estimating the number of clusters in a dataset (Akogul and Erisoglu, 2017). In this research, AIC was assessed using SPSS within 10 different cluster options as shown in Table 8.3. The best clustering solution is one with the lowest AIC value.

Table 8.3: AIC for K=10 cluster solutions

Auto-Clustering

Number of Clusters	Akaike's Information Criterion (AIC)	AIC Change ^a	Ratio of AIC Changes ^b	Ratio of Distance Measures ^c
1	686.545			
2	596.360	-90.185	1.000	1.879
3	567.066	-29.294	.325	1.463
4	559.707	-7.359	.082	1.367
5	565.069	5.362	-.059	1.121
6	574.179	9.109	-.101	1.176
7	587.920	13.742	-.152	1.037
8	602.591	14.670	-.163	1.299
9	623.097	20.507	-.227	1.310
10	648.220	25.123	-.279	1.001

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Table 8.3 shows the AIC values for 10 cluster solutions. From the results, it can be seen that the four cluster model has the lowest values. A model with lowest AIC value is usually considered as the best performing model (Snipes and Tailor, 2014). In this research, the AIC method suggests four clusters as the most appropriate for the dataset.

Gap Statistics

The Gap statistics is based on the idea that comparing the change in within cluster dispersion with that expected under appropriate reference null distribution (Tibshirani *et al.*, 2001). The procedure starts by creating a reference data that represents the observed data; the reference data is then compared with observed data in order to identify the value of K from the observed data with least noise (Pedersen and Kulkarni, 2006). Gap methods can be applied to any clustering algorithm such as hierarchical or K-means. However, the gap method has been criticised for overestimating the number of clusters (Dudoit and Fridlyand, 2002). In this research, the Gap method was implemented in R environment using the K-means output. Figure 8.3 shows the plot of Gap statistics for 10 cluster solutions.

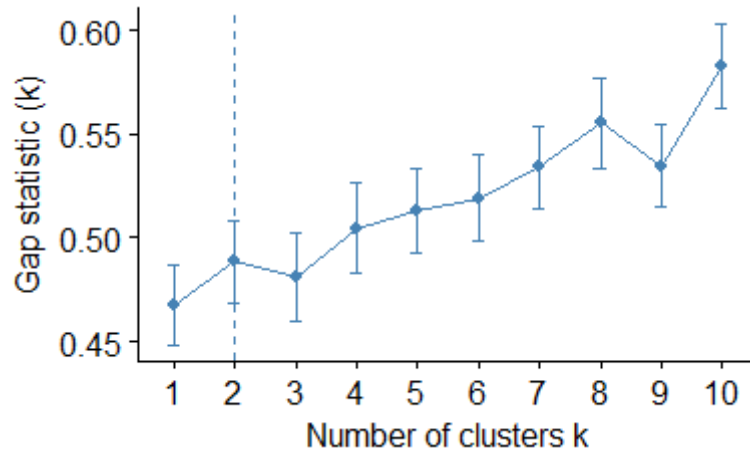


Figure 8.3: Gap statistics plot

From Figure 8.3, the gap method estimate of the number of clusters is 2. The gap function starts to rise again after 4, 6, 8 clusters and 10, suggesting that there are other candidate clusters. Tibshirani *et al.* (2001) suggested examining the entire gap curve rather than simply the position of its maximum in order to explore other potential clustering options available.

Average silhouette width

The silhouette technique provides graphical output for visualising how well clustered objects lie in their clusters. The average silhouette plots are useful for assessing the potential number of clusters in a dataset. The silhouette procedure is well described in Kaufman and Rousseeuw (1990). Average silhouette approach is widely used in a number of studies to guide the selection of an optimal number of clusters (e.g. Shepherd, 2006; Dennett and Stillwell, 2011). Building on the previous studies, in this research, average silhouette width is used in order to determine the appropriate number of clusters for community classification in Leeds. Similar to previous methods, a range of partitions were evaluated in order to arrive at a decision on the optimal number of clusters as shown in Figure 8.4.

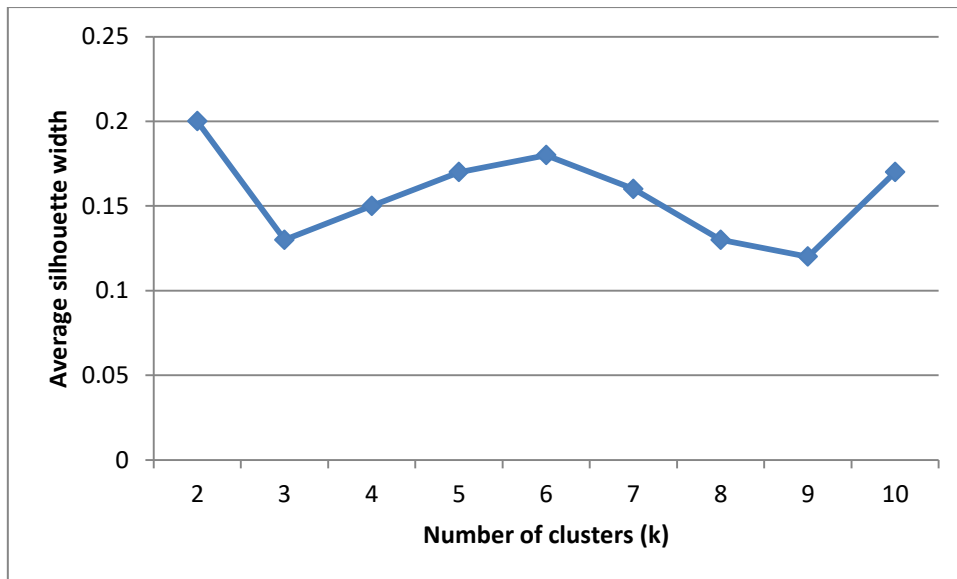


Figure 8.4: Average silhouette widths between 2 and 10 cluster solutions

It can be seen from Figure 8.4 that the $K = 2$ cluster solution has the largest average silhouette width (2.0). There is an indication that as the number of clusters increases, the average silhouette widths decreases and then increases again. Following Shepherd (2006), in this research, consideration is given to the selection of clusters that will reflect true community differences in Leeds in terms of social cohesion and crime rates. In this regard, choosing few clusters (e.g. 2) is rather inappropriate considering the number of Leeds community areas and the complex nature of estimating the degree of community cohesion in different areas. Additionally, fewer groups are undesirable because they will represent broader generalisations (Dennett and Stillwell, 2011). Likewise, selecting many clusters would rather complicate interpretation. Therefore, considering the task at hand, the 4, 5, 6 and 7 cluster solutions could be potential candidates for selection in this research. Although 5, 6 and 7 cluster solutions have slightly higher average silhouette width values (0.17, 0.18 and 0.16) than the four cluster solution (the difference is very negligible); they represent too many clusters that will be difficult to interpret in the context of the present research. Additionally, classifying the degree of community cohesion and crime of an area beyond the range of four groups might not be meaningful. In the light of the foregoing considerations, the four cluster solution with an average silhouette value of 0.15 could be a potential alternative. Figure 8.5 shows silhouette plots the candidate clusters.

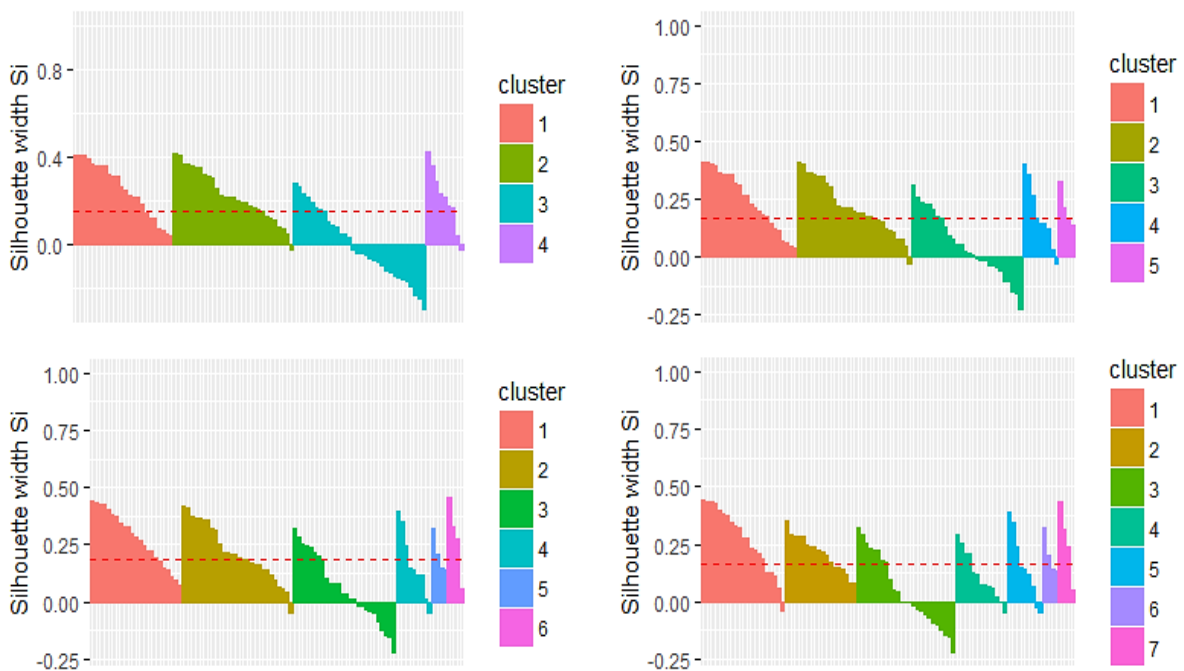


Figure 8.5: Silhouette plots for candidate clusters (4, 5, 6 and 7)

An average silhouette value of 0.15 in the four cluster solution (Figure 8.5) produced in this research is relatively good compared to previous studies. For example, Shepherd (2006) used the average silhouette method to evaluate cluster solutions and successfully implemented classification of Leeds neighbourhoods with an average silhouette value of 0.11. Similarly, Dennett and Stillwell (2011) also employed a similar method in their study to classify internal migration for Britain with an average silhouette width of 0.13. Therefore, increasing the number of clusters does not seem to make any significant improvement beyond four clusters, but rather creating subset partitions from the existing clusters. This is evident in 5, 6 and 7 cluster solutions as seen in Figure 8.5. Rousseeuw (1987) stressed that when K is set too high, natural clusters will be artificially divided in order to conform to the specified number of groups. This has been demonstrated here, where the first three clusters seem to be retained and subsequent clusters appeared to be created from the fourth cluster.

Furthermore, in order to assess whether to proceed with the four cluster solution for our classification, a comparison was made with the six cluster solution that has the largest average silhouette width (Figure 8.4). In both clustering solutions, the object with largest silhouette value (0.42) is Headingley community area. Additionally, examining the values of average silhouette widths for the two different clustering solutions shows that they both have

some similarities (Figure 8.6). Again this makes it more difficult to decide on which clustering solution to choose from the two. However, in a situation of this kind where similarities exist between two different cluster solutions, it was suggested that a better solution will likely be one with less negative values (Dennett and Stillwell, 2011). In this research, the four cluster solution has slightly fewer communities (20) with negative silhouette values compared to the six cluster solution that has 22 communities with negative silhouette values. Negative values indicate that objects do not conform well in the assigned cluster.

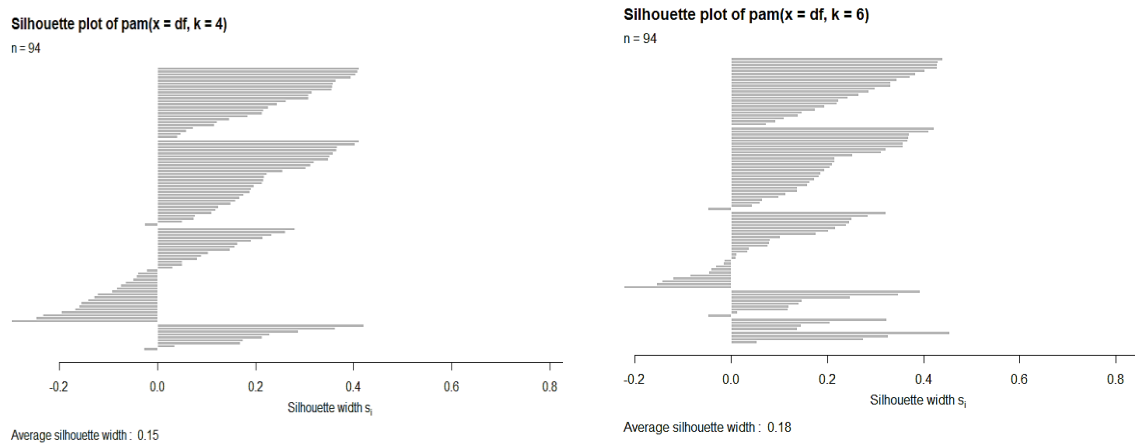


Figure 8.6: Silhouette information for K=4 and K=6 cluster solutions compared.

Taking into consideration all evidence from different methods for the selection of the optimal number of clusters explored, therefore in this research, a four cluster solution is deemed more appropriate. Despite a slightly negligible differences in the average silhouette width between the cluster solutions.

8.3.7 Clustering Validation

Clustering validation is a process of evaluating the goodness of clustering results and is an essential step for the success of any clustering application (Liu *et al.*, 2010). It is always a good practice in cluster analysis to re-run the analysis with different algorithms in order to see if similar solutions could be found. Charrad *et al.* (2014) stressed that cluster validation is necessary because different algorithms produce different clusters, even for the same algorithm selection of different parameters or order of data can greatly affect the clustering solutions. Also, the validity of the solution should be tested to ensure that clusters are representative of the variables included in the analysis. Generally, clustering validity can be evaluated in two ways; internal criteria and external criteria. The internal criteria are based on

the quantities and features inherent in the dataset; on the other hand, external criteria involve the structure of the resulting algorithm applied to the dataset (Halkidi *et al.*, 2001).

Internal Validation

Linoff and Berry (2011) proposed two approaches to internal clustering evaluations namely *compactness* and *separation*. Compactness measures how closely related the objects in a cluster are, while separation measures how distinct or well separated a cluster is from other clusters (Liu *et al.*, 2010). Clustering validation indices (CVIs) can be used to measure the quality of clustering based on compactness and separation of the clusters (Hämäläinen *et al.*, 2017). A number of CVI techniques have been proposed in the literature and a comprehensive review of the CVIs validation techniques was provided in (Arbelaitz *et al.*, 2013). For the internal clustering validation, the Dunn index (Dunn, 1974) and the Silhouette index (Kaufman and Rousseeuw, 1990) are commonly employed. However, the Dunn index is less preferred because it is computationally expensive and sensitive to noisy data (Saitta *et al.*, 2007). Additionally, unlike the Silhouette index, the Dunn index has no graphical output for clustering solutions. Therefore, in this research, Silhouette index, a widely used internal validation technique was employed for assessing the quality of clusters (Liu *et al.*, 2010; Thinsungnoena *et al.*, 2015).

External Validation

The quality of the clustering solution can be evaluated externally by comparing similarities between two partitions (Desgraupes, 2013). External validation mainly quantifies how good a partitioning is obtained with respect to prior class labelled information available. The Rand index (RI), F-measure, normalised mutual information (NMI) and purity are some common examples of external validity indices used in the literature (Alok *et al.*, 2014). However, RI is widely used measure for external cluster validation (Santos and Embrechts, 2009; Van Craenendonck and Blockeel, 2015). RI is a measure of similarity between two different clustering methods of the same set of datasets (Rand, 1971). The measure assesses how much each pair of data points is assigned in each clustering. RI values range between 0 and 1, with values closer to 1 indicating a perfect match and values towards 0 indicate no match. RI of 2 partitions (x and y) is given as:

$$RI = \frac{a + d}{a + b + c + d} \quad (8.5)$$

where a and c are the number of pairs of objects that are in the same cluster in x and in the same cluster in y; b and d are the number of pairs of objects that are in a different cluster in x and in a different cluster in y.

In this research, the results from K-means and PAM algorithms were used for the clustering validation in order to compare their similarity using both measures of validity. The Fossil package by Vavrek (2011), a package for analysing geographical datasets in R, was employed for computing RI and ARI respectively.

However, a problem with RI is that the expected value of random clustering does not take a constant value (such as 0). Therefore, to address this limitation, an Adjusted Rand Index (ARI) was proposed (Hubert and Arabie, 1985). ARI is a correction of the RI that measures the similarity between two classifications of the same objects by the proportion of agreements between the two partitions, taking account of random chance that will cause some objects to occupy the same cluster. The correction is obtained by subtracting from the RI its expected value. ARI can be written as:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}} \quad (8.6)$$

where n_{ij} is the number of objects common in the two partitions, a_i is the sum of values in row i and b_j is the sum of values in column j ; and

where $\sum_{ij} \binom{n_{ij}}{2} = \text{index}$, $[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2} = \text{expected index}$ and $\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] = \text{max index}$.

The ARI equation can also be written in a simplified format as:

$$ARI = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} \quad (8.7)$$

Table 8.4 presents the performance of measures of the validity indices of the clustering algorithms used in this research.

Table 8.4: Performance of validity indices between K=2 and K=9 solutions

K	RI	ARI
2	0.67	0.34
3	0.64	0.29
4	0.78	0.44
5	0.76	0.38
6	0.77	0.38
7	0.78	0.27
8	0.84	0.41
9	0.85	0.35

Table 8.4 shows the comparison between RI and ARI external validity indices. As highlighted, the 4 cluster has the largest ARI (0.44), indicating a fairly strong agreement between the clustering algorithms. Additionally, the RI value of 0.78 obtained indicates a strong match between the two clustering solutions. RI is usually higher than the ARI, since RI lies between 0 and 1 the expected value of RI (though not a constant value) must be greater than or equals to 0; while the expected value of ARI has a value of 0 and a maximum value of 1 (Santos and Embrechts, 2009 p.2). Figure 8.7 shows the graph plot of the comparison between RI and ARI index values as a function of the number of clusters.

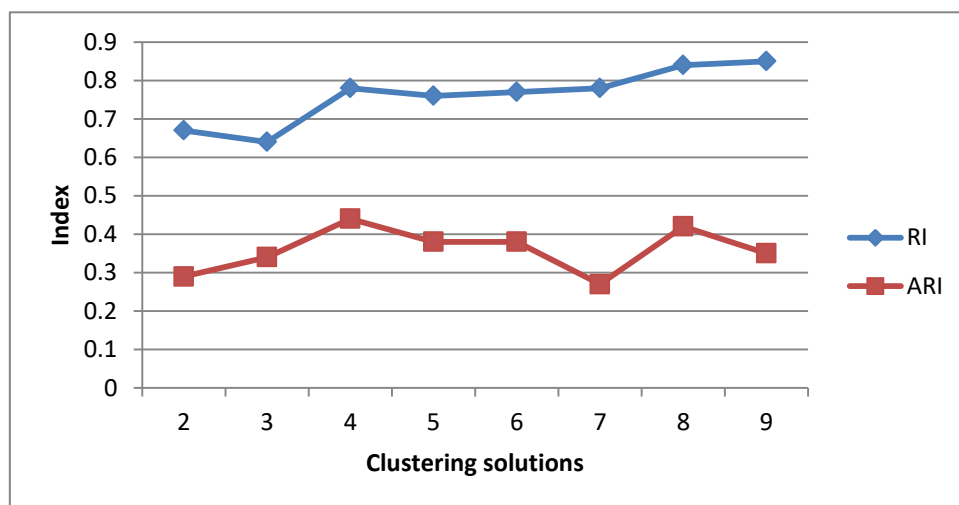


Figure 8.7: Graph plot of the comparison between RI and ARI index values

8.4 Leeds Community Classification

The final step in the classification involves an interpretation of the typology produced. This task, however, is sensitive and needs to be handled with caution, especially where names are assigned to groups. The process of creating a name that best describes a cluster can be difficult and using a contentious name can be offensive (Vickers and Rees, 2007). The name is expected to describe the general area/population but cannot be expected to describe each individual. Additionally, the names assigned to clusters should as much as possible reflect the characteristics of the communities in that cluster. In this research, the typology of the clusters is intended to reflect the degree of community cohesion of different areas as they relate to crime. However, as highlighted by Vickers and Rees (2007) and for the avoidance of naming clusters that may not be overly well received within some community areas; clusters will simply be called *groups* (similar to Shepherd (2006)).

Furthermore, the geography of different classifications can be mapped for visualisation. The influence of different variables in a particular cluster can also provide a useful guide for describing the characteristics of that cluster. One way of achieving that is by visualising the variables using a radar chart. Radar charts consist of spikes representing the different variables and how these compare to the global average. They have the advantage of displaying multidimensional data without truncating the data (Kalonia *et al.*, 2013). Radar charts are widely used for visualising classifications (Vickers and Rees, 2007; Adnan, 2011; Alexiou *et al.*, 2016). In this research, while clusters are mapped to explore their spatial distributions, radar charts are also used for visualising the influence of clustering variables in each group (Figure 8.8). Since the variables were transformed to z-scores, the mean of each variable is 0. Groups are defined by variables with the highest and lowest standardised variables. Additionally, boxplots are also provided to show the distribution of the data for each variable.

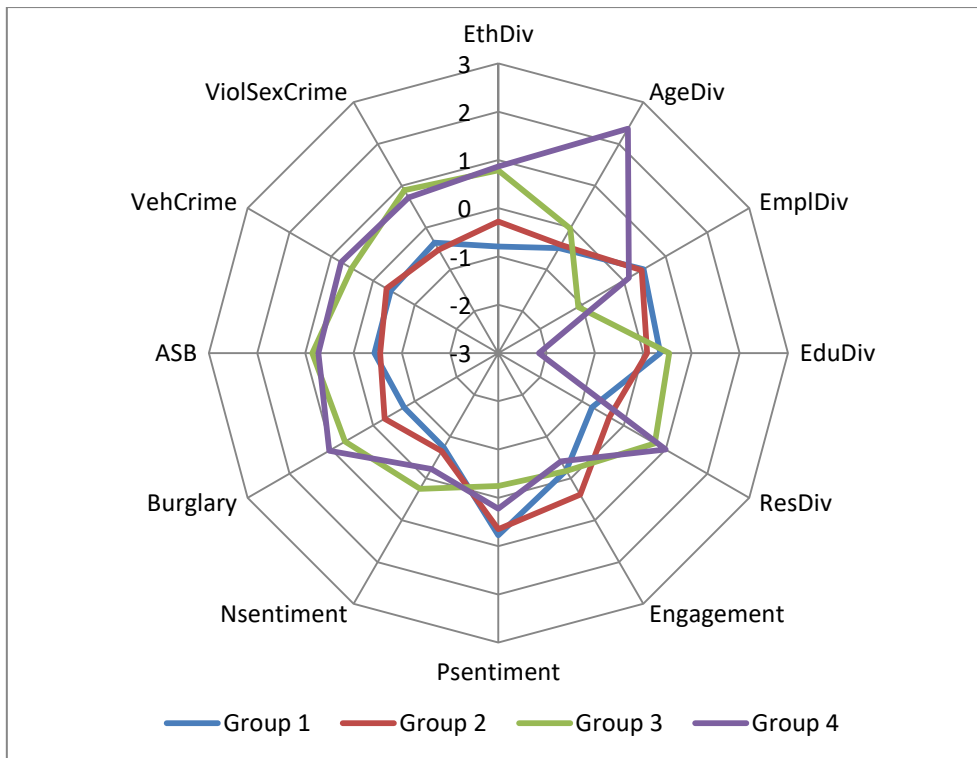


Figure 8.8: Combined radar charts for community groups

8.4.1 Group 1

Group 1 comprises of 24 community areas which are largely located around the fringes of Leeds district (Figure 8.9). The silhouette width value of most of the community areas is between 0.25 and 0.41, indicating a strong similarity of the members. Additionally, the average silhouette width value of 0.25 obtained for this group suggests that it is well defined, though, a few community areas like Pudsey, Morley North Rothwell tend to be weakly placed in the group. Similarly, Colton community area has a negative silhouette value (-0.02) suggesting non-conformity in the cluster which is likely due to random chance (see Section 8.3.7). The boxplots of standardised (z-scores) variables of the cluster are shown in Figure 8.10. The few outliers observed in some variables that defined the group in results from the areas that were not strongly placed.

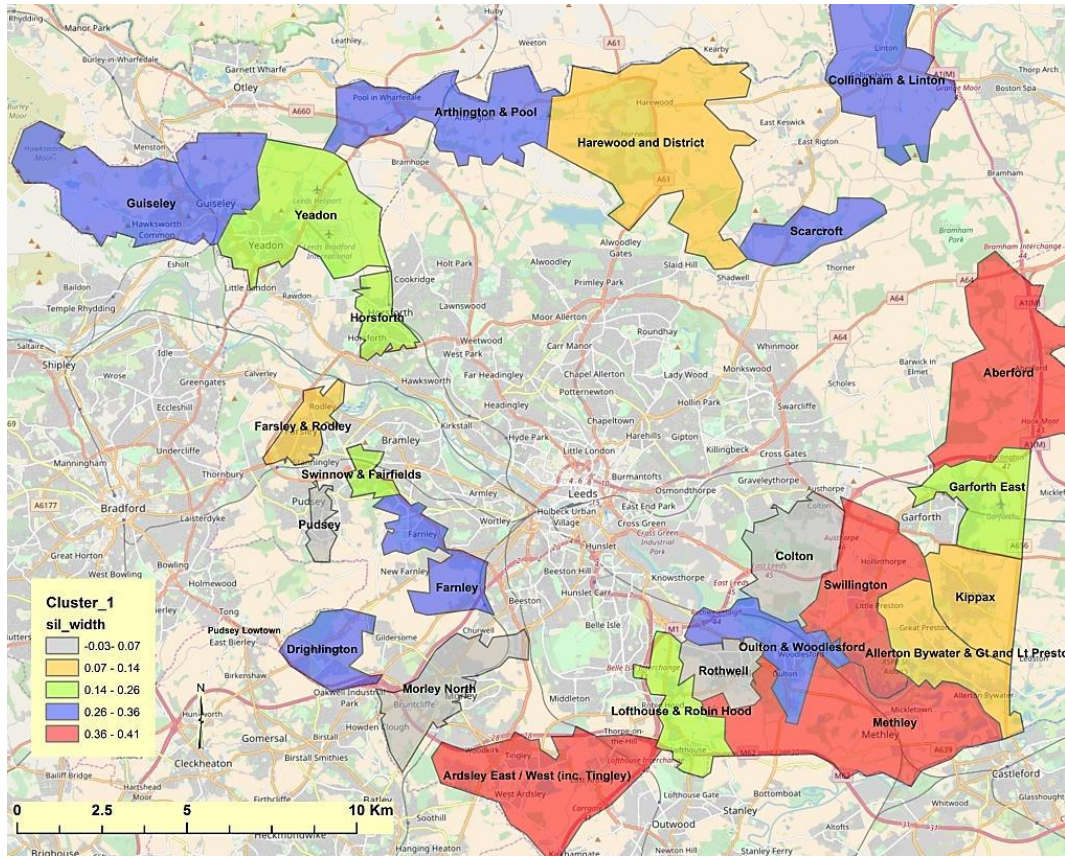


Figure 8.9: Silhouette width values of group 1 community areas

The cluster is characteristically homogeneous as indicated by their lowest ethnic diversity and age diversity (Figure 8.11). The communities are likely to be affluent, owing to a higher proportion of educational attainment and employment status of the residents. Lower residential instability in the areas also suggests that most of the residents are likely to own their homes. The bond of social networks is relatively high, meaning that the people are likely to express relatively higher positive sentiments. A higher positive sentiment is likely because people are happier in their community areas (see Section 6.5.3).

Crime rates are generally lower in this group compared to all other areas of the district. *Burglary*, *ASB*, *vehicle* crime and violence were below the mean suggesting a strong social capital of the residents.

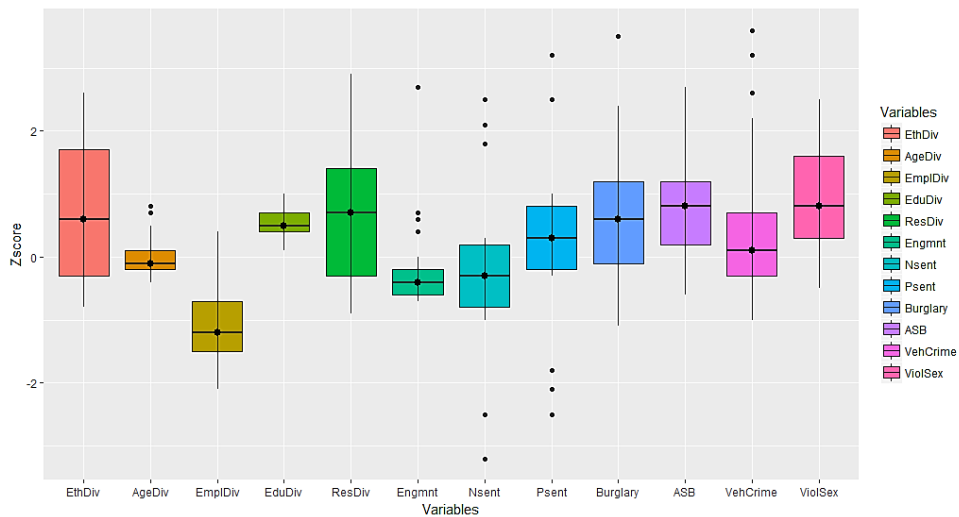


Figure 8.10: Boxplot of standardised z-scores of the variables in group 1

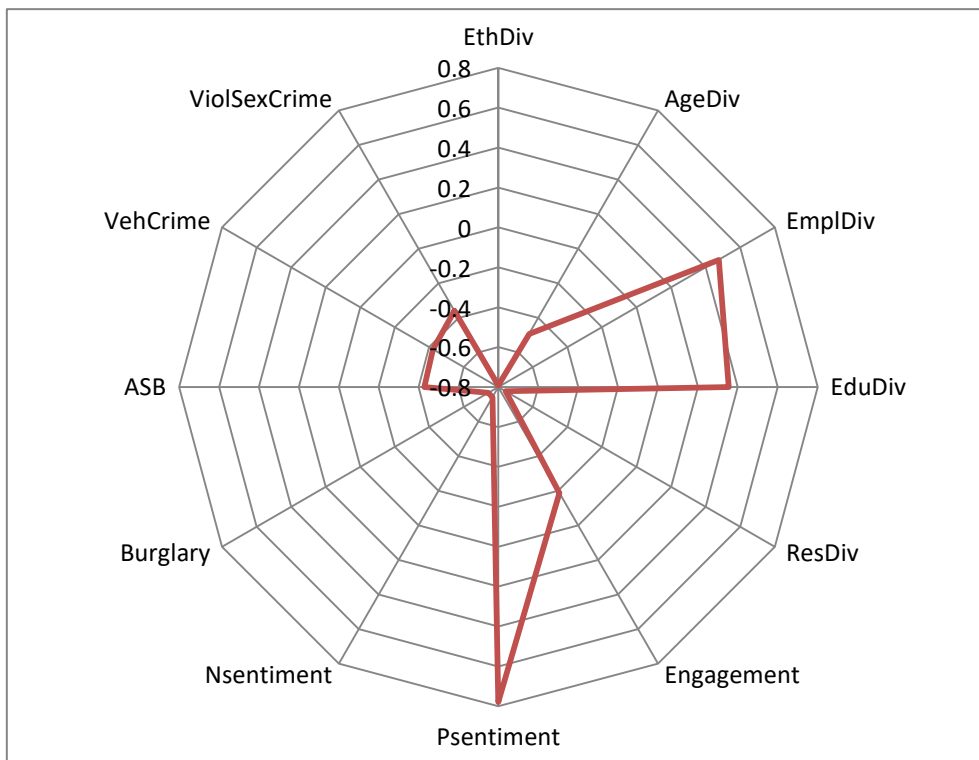


Figure 8.11: Radar charts profile of the variables defining group 1

8.4.2 Group 2

Group 2 is the largest cluster comprising of 34 community areas is mainly spread in the north including Otley, Wetherby and Boston Spa with a few community areas in the central, south-east and south-west (Figure 8.12). According to the average silhouette width value (0.22), the cluster is also strongly defined. Shadwell community area has the largest silhouette width

value (0.41) in the group. Despite a strong definition of the cluster, six community areas including Kirkstall, Meanwood, Roundhay and Wortley have shown non-conformity with the cluster characteristics having a negative silhouette value respectively.

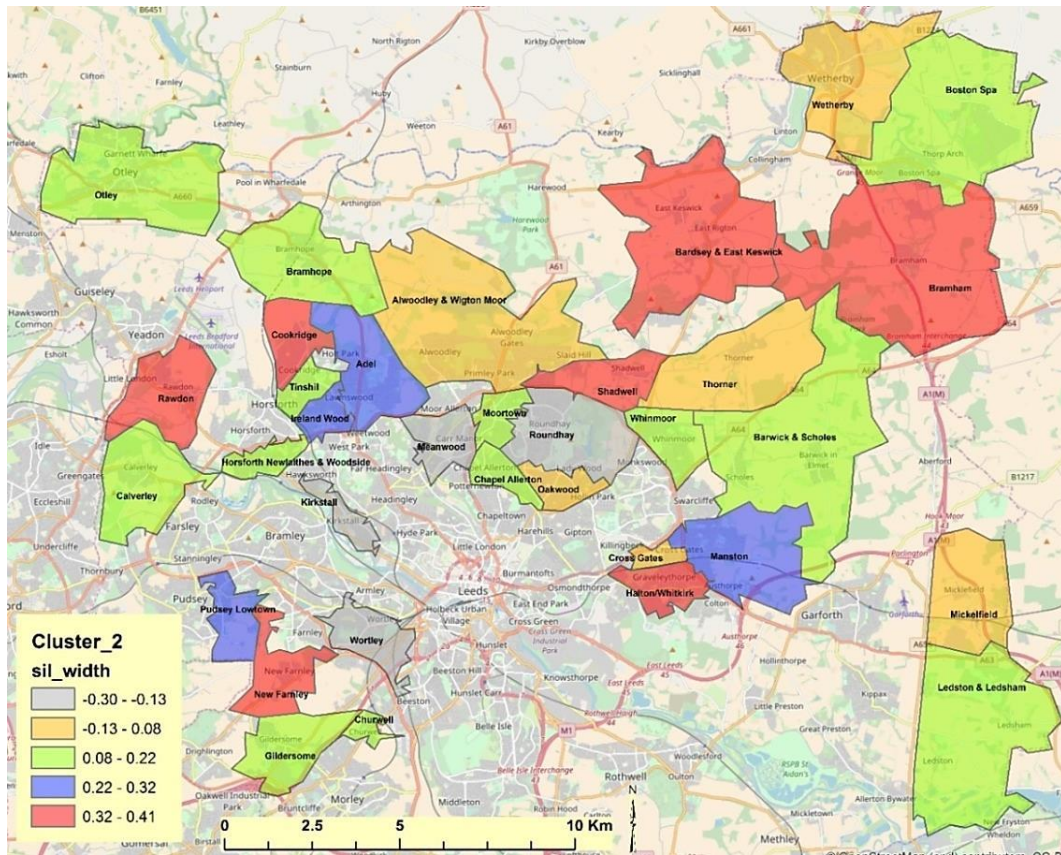


Figure 8.12: Silhouette width values of group 2 community areas

Figure 8.13 shows the boxplots of standardised (z-scores) of the variables that defined group 2. The cluster is characteristically similar to group 1, except that the z-score values are slightly higher in this cluster area. Community areas in this group are relatively homogeneous, stable and are likely to be affluent (Figure 8.14). Additionally, the population is likely to comprise of educated people in the top level of employment, though slightly lower than those in group 1. Furthermore, the presence of high social networks (engagement) in the cluster suggests that the residents are likely to work together for their common good. This is evident by the higher proportion of positive emotions in the community areas. There is the likelihood that the people are happier in their interactions.

Crime rates, especially, *ASB*, *vehicle-related* offences and violence are similar to those in group 1 (z-scores below the mean). *Burglary* is slightly higher in this group compared to group 1, especially, in areas with negative silhouette values in the cluster like Wortley and Kirkstall.

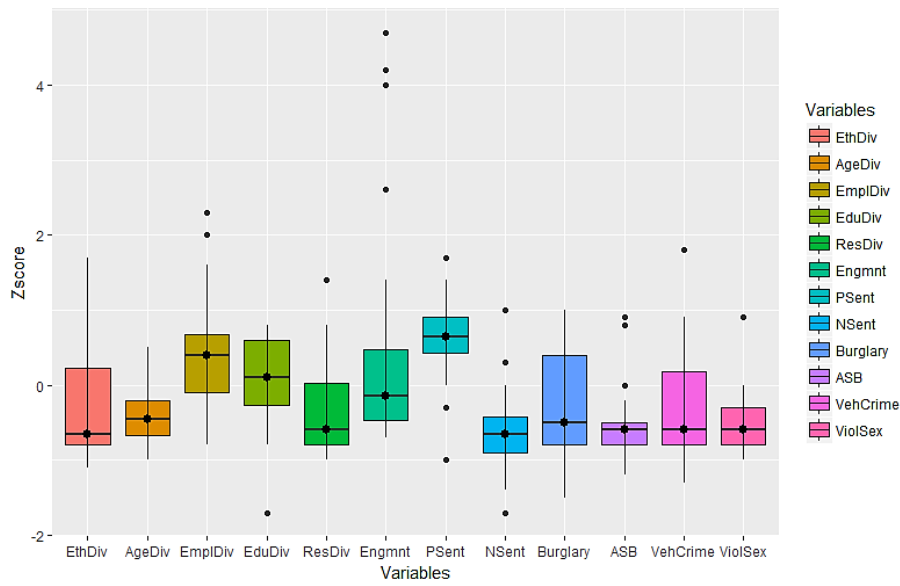


Figure 8.13: Boxplot of standardised z-scores of the variables in group 2

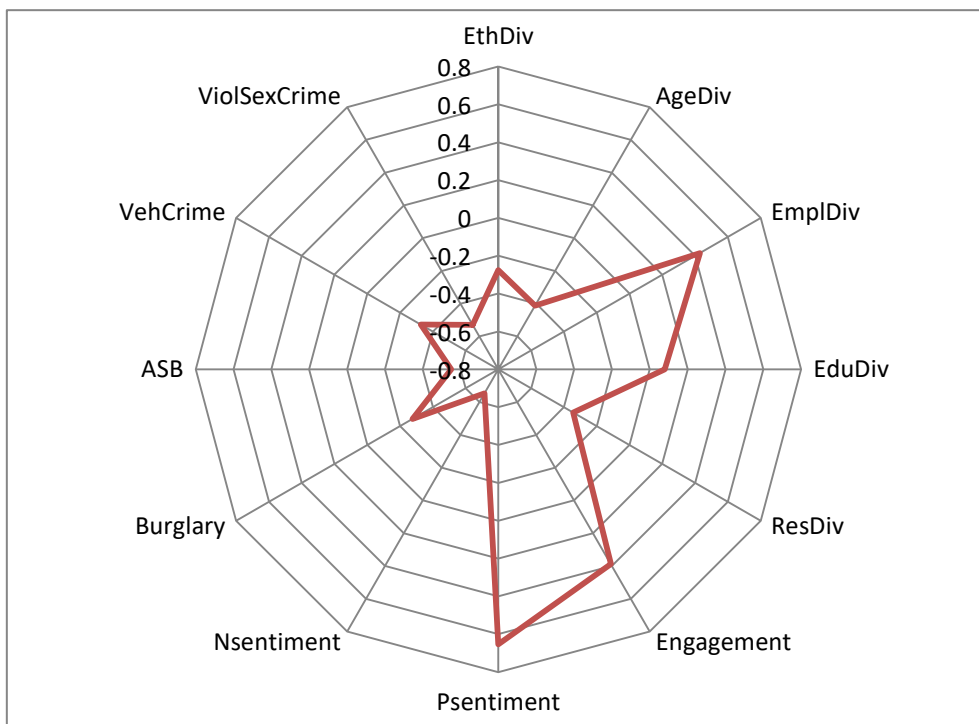


Figure 8.14: Radar charts profile of the variables defining group 2

8.4.3 Group 3

Group 3 is the second largest with 25 community areas located on the east-central and west-central parts of the district (Figure 8.15). Holt Park is the only community area located in the north. The largest silhouette width value (0.28) is Harehills. The mean silhouette width value

of -0.01 indicates a poorly defined cluster. Additionally, 10 community areas of the cluster have negative silhouette values, suggesting non-conformity with the cluster. One potential explanation of this occurrence might be the proximity of these areas to neighbouring community areas in group 1. Figure 8.16 shows the boxplots of the standardised variables in the cluster. The cluster is characteristically multicultural comprising of a higher proportion of ethnic minority, especially in areas like Chapeltown, Little London, Gipton North and Beeston (Figure 8.15). Additionally, the cluster is one characterised by relatively high proportion (around the city average) of young population who are likely to be in the lower cadre of employment. The educational attainment of the residents is slightly above the city likely in the range of level 3 and 4, depicting the socio-economic status of the areas. Higher residential instability in this cluster reflects the immigration status of the population (consisting of a higher proportion of renters). Social networks are lower than the city average, a likely indication of weak social interaction in most of the community areas. The higher proportion of negative emotion is potentially indicating lack of satisfaction in the areas.

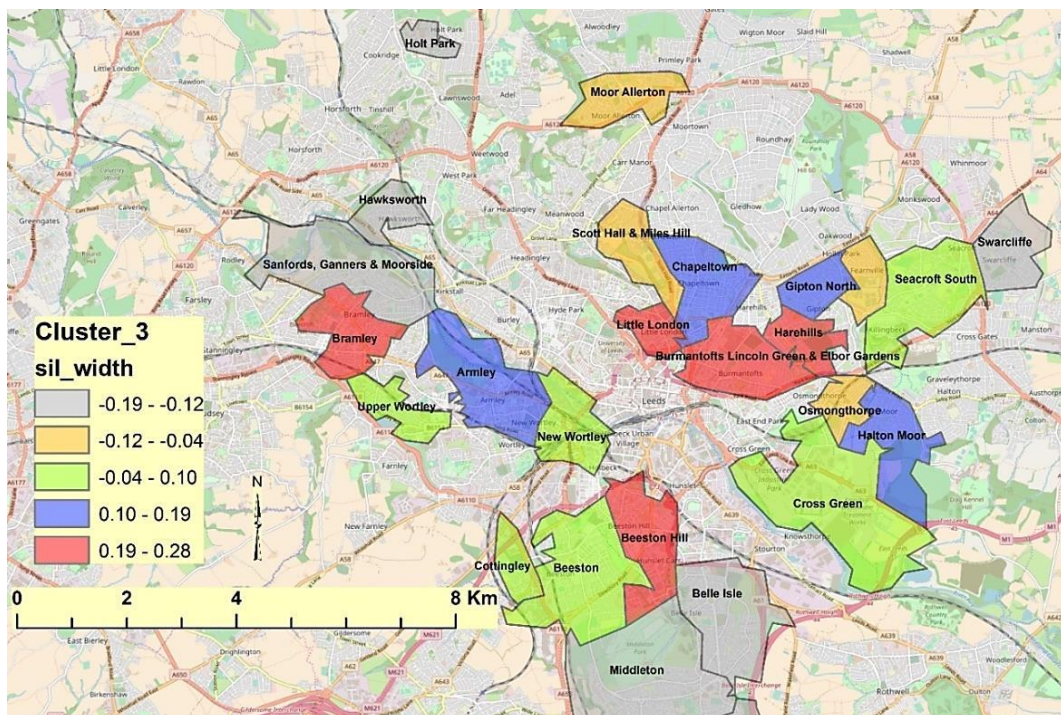


Figure 8.15: Silhouette width values of group 3 community areas

Crime rates in this cluster are by far above the city average (Figure 8.17). *Burglary, violence, ASB* and *vehicle-related* crime rates are higher in New Wortley, Little London and Harehills

community areas, a likely indication of the difficulty of establishing social norms in these areas.

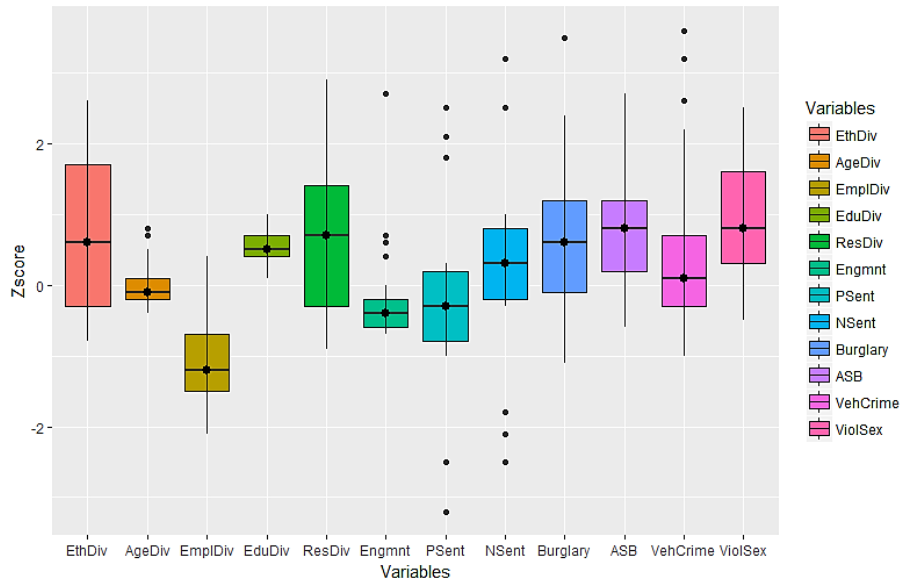


Figure 8.16: Boxplot of standardised z-scores of the variables in group 3

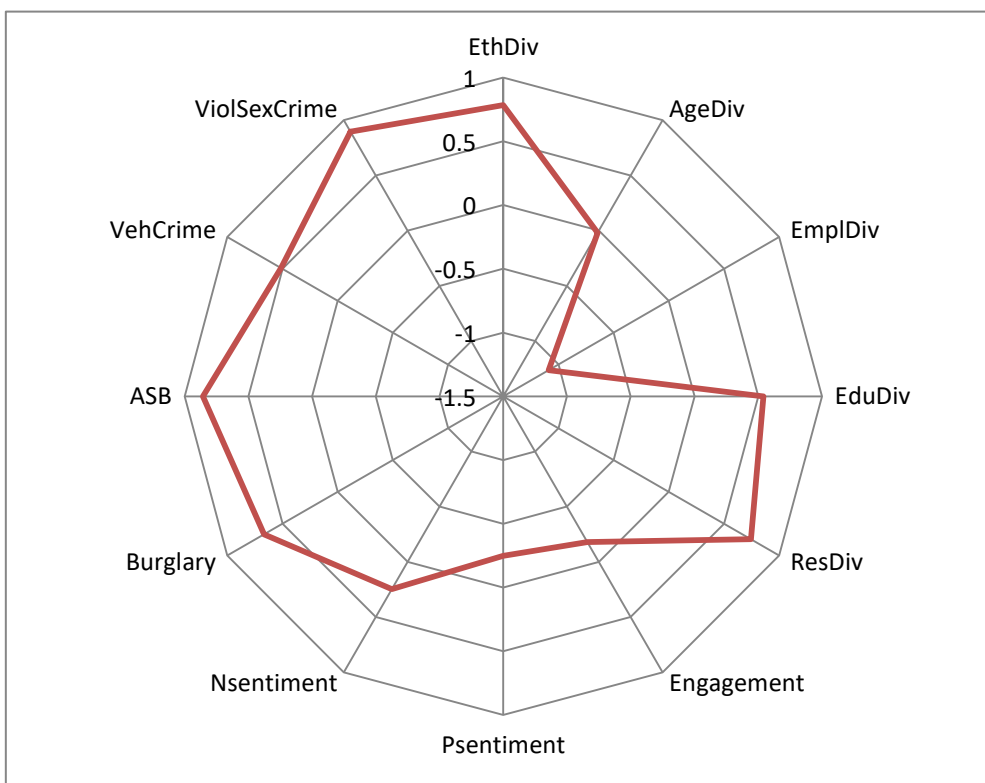


Figure 8.17: Radar charts profile of the variables defining group 3

8.4.4 Group 4

Group 4 is the smallest with 11 community areas mainly concentrated at the city centre (Figure 8.18). The mean silhouette value of 0.21 indicates a reliably defined cluster. Headingley community area has the largest silhouette width (0.42) in this cluster. There are three community areas (West Park, Richmond Hill and Hunslet) with negative silhouette values, indicating non-conformity in the cluster. Boxplots of the standardised (z-scores) variables for this cluster is shown in Figure 8.19.

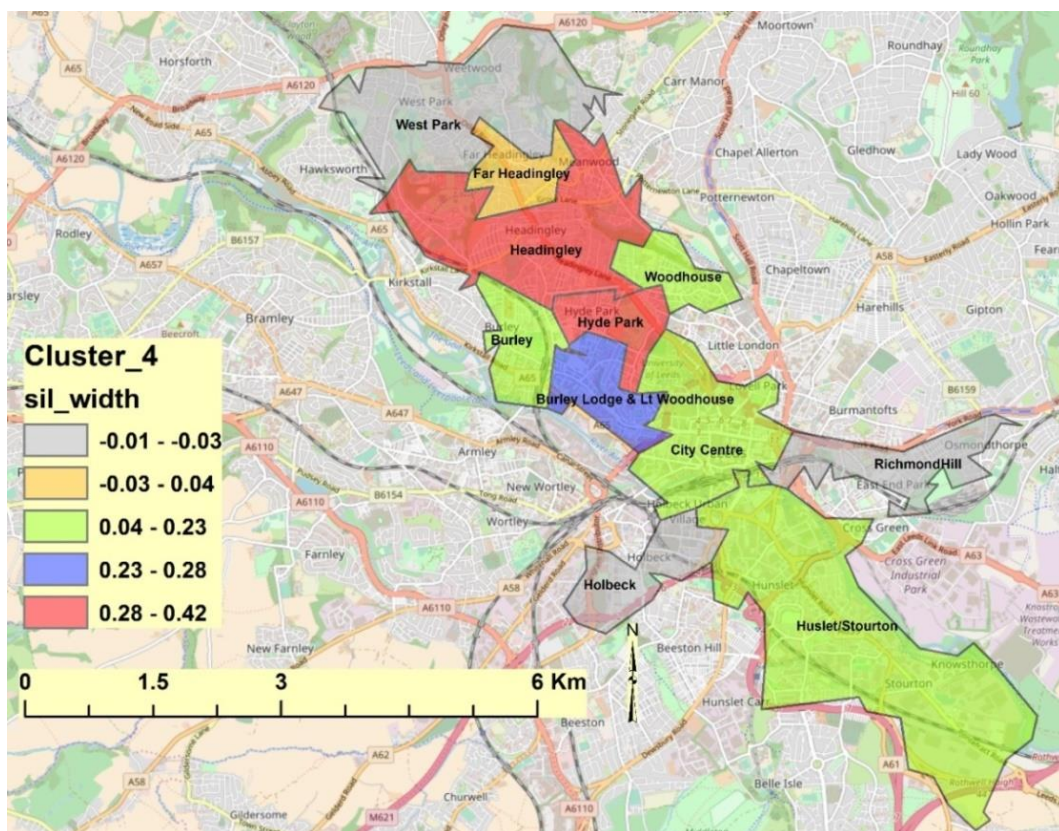


Figure 8.18: Silhouette width values of group 4 community areas

The cluster is characterised by heterogeneous community areas largely comprising of a higher proportion of the young population. The diversity of educational attainment in this cluster is the lowest in the city (Figure 8.20). This is likely because of the higher concentration of student population from the two universities, especially in Headingley, Hyde Park and Burley Lodge & Little Woodhouse community areas respectively. Employment is slightly above the city average in some areas like Hunslet/Stourton community areas. The cluster is also characterised by highest proportion of transient people, typical of a student community. Social networks are below the mean, a likely indication of weak bonds of social

interaction. Positive emotion is slightly above the mean in some areas such as Richmond Hill. The characteristics of this cluster are similar to those of group 3.

Crime rates in this cluster are all above the city average very similar to those in group 3, except that *ASB* is relatively higher in group 3. Additionally, rates of *burglary* and *vehicle-related theft* are highest in this cluster compared to all other groups. A likely explanation to the higher rates of burglary and vehicle crime in this cluster might be as a result of the presence of student residences and commercial activities with large parking spaces, especially in the City Centre and Hunslet/Stourton community areas.

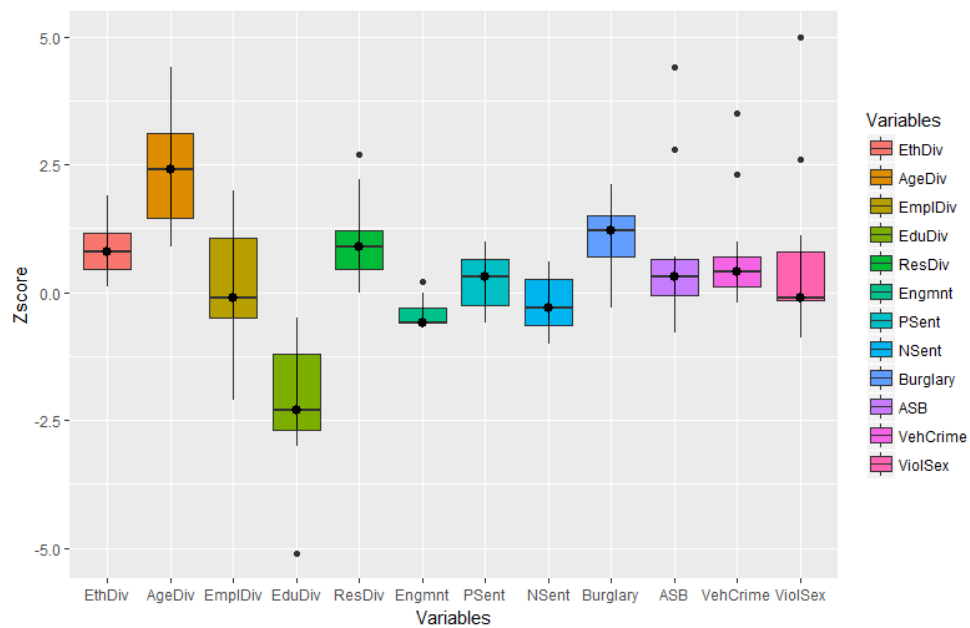


Figure 8.19: Boxplot of standardised z-scores of the variables in group 4

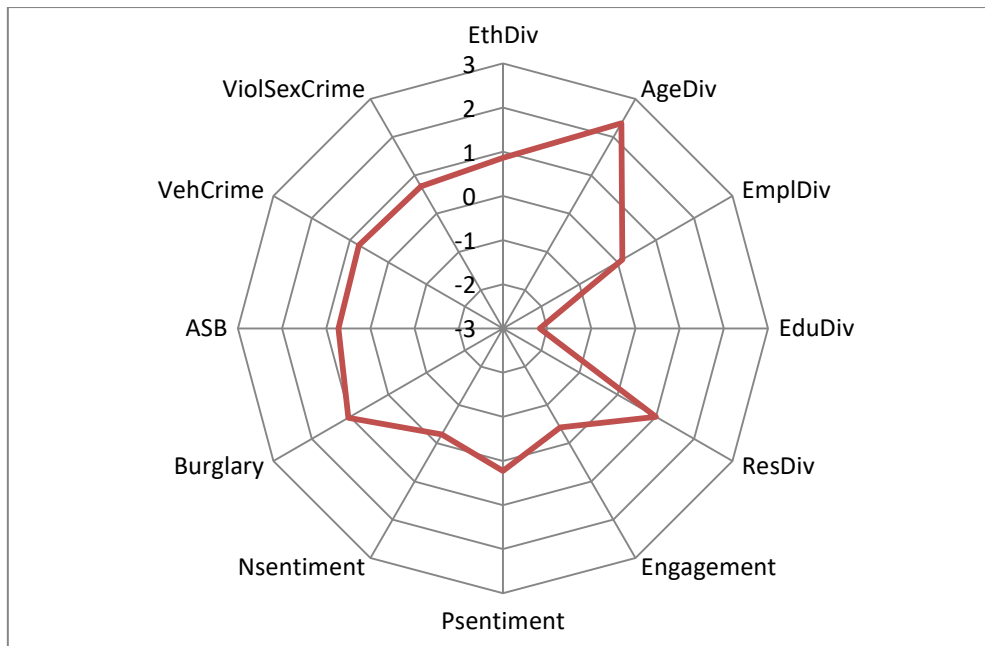


Figure 8.20: Radar charts profile of the variables defining group 4

8.5 Concluding Remarks

This chapter has presented a new classification profile of Leeds community areas based on the degree of community cohesion and crime. At the beginning of the chapter, the importance of community classification was highlighted in Section 8.2. The relationship between communities and crime was described in Section 8.2.1, and issues with previous classifications were also discussed (in Section 8.2.2). The clustering procedure was reviewed and the method implemented based on the literature (Section 8.3). The results were also validated using the appropriate validation methods (Section 8.3.7).

In this research, considerations were made in constructing a classification that will reflect true community differences in Leeds in terms of social cohesion and crime rates. In this regard, considering the number of Leeds community areas and the complex nature of estimating the degree of community cohesion in different areas, appropriate steps were taken in choosing an optimal number of clusters. The final classification profile (Figure 8.21) consists of four clusters (groups) that have distinctive characteristics, though with some similarities in other areas. This classification has highlighted the different characteristics that exist in different community areas, which also reflects the degree of their social relationships. Though classification of this is complex, no matter how well designed some issues might still arise. Interestingly, the new typology produced can potentially be useful for understanding the impact of community building on crime rates in Leeds. Additionally, the insight highlighted

from the classification profile also reflects the findings of regression models performed in chapter 5 and 7 of this thesis; demonstrating the relationship between community cohesion and burglary rates in Leeds. The next chapter (9) will conclude the thesis by providing a summary, limitations and recommendations for future research.

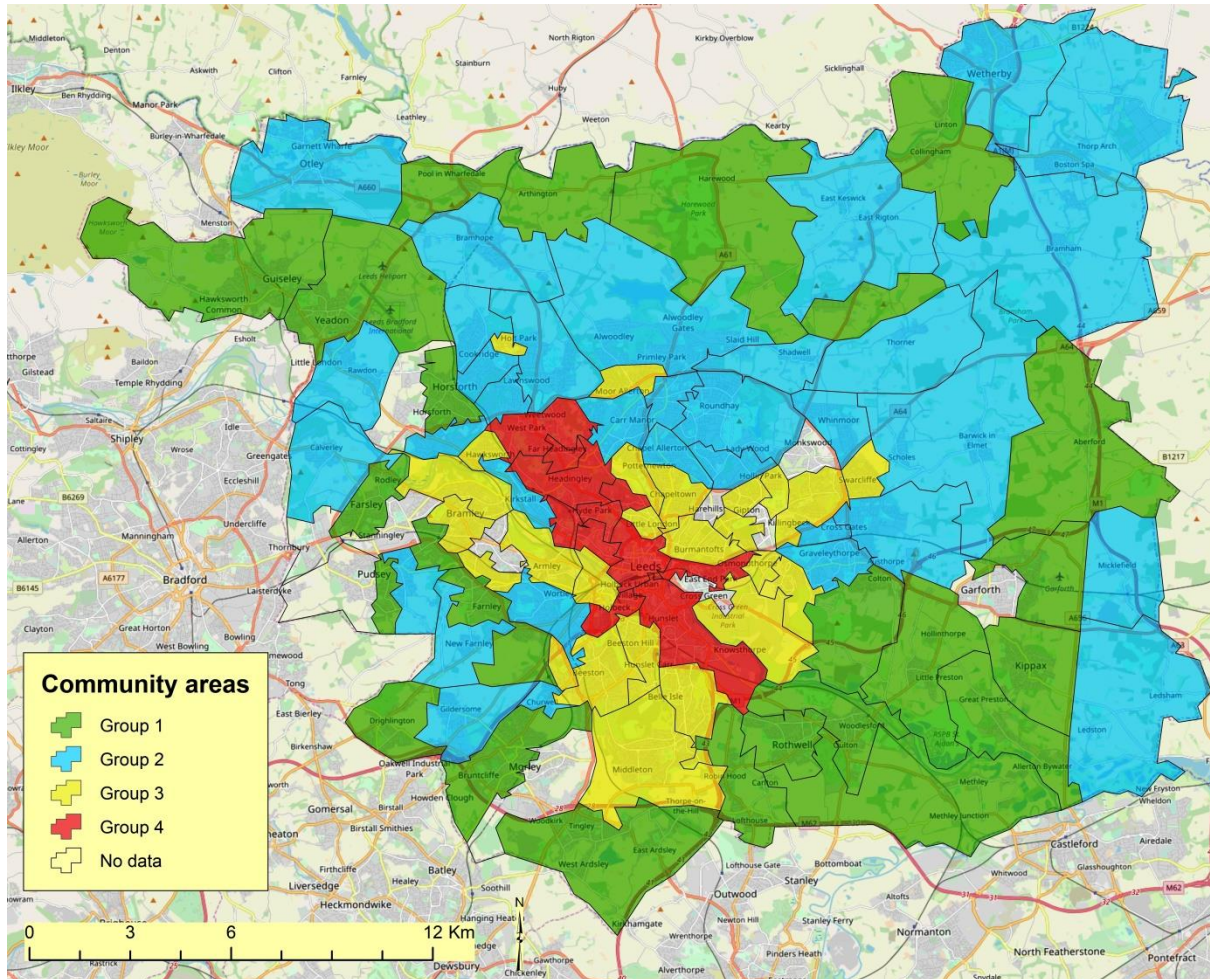


Figure 8.21: Final cluster map

Chapter 9

Conclusion

9.1 Introduction

The work within this thesis has successfully achieved the aim and objectives outlined in Section 1.2: *to explore the relationship between community cohesion and crime*. This chapter concludes the thesis by summarising the main research findings in Section 9.2, and assesses the extent to which the research objectives have been met. The limitations of the research are presented in Section 9.3. Recommendations for future research are given in Section 9.4. Concluding remarks are provided in Section 9.5.

9.2 Summary of Research Findings

This thesis presents a novel approach to using new ‘big’ data sources from social media (Facebook and Twitter) for exploring the relationship between community cohesion and crime. This approach in particular is useful in highlighting new ways of quantifying the relationship between community cohesion and crime. Through this novel approach, the research has successfully explored the relation between community cohesion and crime based on insights from traditional and new data sources. In this section, the outlined objectives of the research (Section 1.2) will be evaluated to demonstrate how these objectives were met from within the body of the research presented within this thesis.

Objective 1: *to critically review the literature on the concept of community cohesion and its impact on crime*. This objective was achieved through a critical review of literature related to the concept of community cohesion and crime in chapter 2. The chapter highlighted the relationship between community structural characteristics and crime. The chapter argued that although community cohesion generally acts to increase the safety of communities by reducing the socio-economic drivers of crime, cohesion is poorly captured by standard regression models of crime. Social capital and its importance in crime including its determinants were discussed. To set the context of the research landscape within this area, the effects of collective efficacy on social cohesion, and by extension crime, are discussed here. Additionally, the weaknesses of previous studies were highlighted in Table 2.2. To put the research in context, crime theories related to the research such as social disorganisation theory and routine activity theory were also reviewed. The main finding from the literature

review is that community cohesion generally acts to increase the safety of communities by reducing the socio-economic drivers of crime, and by reporting crimes when they occur. Consequently, cohesive communities are likely to have lower crime rates. Diversity and deprivation have been identified as the key factors that compromise a community's level of cohesion. The literature review provides the frame of the research within the thesis.

Objective 2: *to use a range of spatial analysis methods to perform a detailed analysis of the relationship between crime and urban community form.* The requirement of this objective was met in chapters 3 and 5 respectively. Chapter 3 reviewed and discussed a range of quantitative methods of modelling urban social systems, their relative merits and demerits as well as their application to community and crime research. Traditional data sources such as police crime records and Census statistics employed for the analysis were also reviewed. This was in order to identify and select the most appropriate variables that reflect aspects of community cohesion and crime. Regression models are the most widely used methods in social sciences for quantifying the relationship between urban social systems such as community cohesion and crime in the context of this research. Standard regression variables that were identified as potential determinants of community cohesion and crime and their adjusted equivalents (Table 3.2), were employed in traditional multiple regression models (chapter 5) to explore the influence that each variable has. The diversity statistics were found to be more statistically significant correlates of burglary; this is because these variables capture a range of loci in which social cohesion plays a part across the crime system more effectively than standard treatments. This indicates the important part that social cohesion plays in reducing crime in the local area.

Objective 3: *to critically review the literature on the feasibility of using new 'big' data sources from the social media to explore crime and community cohesion.* Chapters 4 and 6 fulfilled the requirements of this objective. In chapter 4, literature on how social media is being used for social interaction was reviewed. Specifically, an emphasis was given to the most popular social media platforms (Facebook and Twitter) which were used for exploring social cohesion and crime as they relate to action in communities and for understanding communities. The importance of social media in law enforcement activities was also discussed. Social media channels relevant to this research and the procedures used to access Facebook and Twitter data were presented. Additionally, methods and algorithms used for social media data analysis such as to quantify sentiment classification of tweets and explore engagement from Facebook posts were reviewed. The review of social media in this chapter

is about understanding the basis of how to use the data and getting around issues associated with these data. Using social media data for exploring community cohesion is an area of novelty for the thesis and presents the first academic attempt in this area.

The findings from the literature review on social media usage for the creation and maintenance of social relationships in chapter 4 were extended in chapter 6. Methods used for Twitter data collection were critically reviewed. Selection of the *bounding box* method in this research, and avoidance of any specific search term is also novel approach to reducing sampling bias in data collection compared to the *keyword* search method commonly used in which the data collected are predetermined. However, data collected using this method will also include tweets without geolocations because although the original tweet may have a geolocation, the retweet may not (Twitter, 2017d). In this research, the attributes of location are important for understanding patterns of users. If tweets are not geolocated, they are considered not to be of value in the context of the task at hand. As sentiment analysis is prone to errors when applied on raw tweets, the rigorous process of Twitter data cleaning, especially for detecting and removing bots was described. The accuracy of the sentiment classification algorithm used in this thesis was validated at the Amazon Mechanical Turk (AMT) using manual human coding (see Section 6.5.2). Determining the sentiment expressed in a tweet is challenging and is reliant on the subjective judgement of human annotators, often the annotators disagree among themselves. However, a key finding from this validation indicates a strong correlation ($r = .629$) between human (manual coding) and algorithm classifications and the reliability scores were fair on the Fleiss's scale of measurement ($\kappa = .236$) and good on the Cronbach's scale ($\alpha = .775$). This indicates that the internal consistency between the human annotators and algorithm is up to a standard acceptable level alpha of .70 and above based on literature (e.g. Santos, 1999; Tavakol and Dennick, 2011). Uncovering the relationship between sentiment and diversity is proved to be more of a challenging task. This is an area that requires further investigation using additional variables. Similarly, sentiments in tweets were found to be related to some diversity variables (age, education and employment), though in a weak manner. Notwithstanding, the findings have highlighted new insights into the relationship between sentiments and demographic characteristics of different areas.

Objective 4: *to develop neighbourhood area classification profiles based on community cohesion in range of geographical locations of Leeds at community area level.* This objective was attained in chapter 8. The chapter combines insights derived from chapters 5, 6 and 7

based on a new perception of community cohesion to classify a range of community areas within Leeds. The chapter critically reviewed issues with previous classification attempts, and identified several different clustering methods before adopting the partitional methods. The PAM (K-medoid) method is more robust in the handling of noise and presence of outliers associated with some variables used in this thesis. It is therefore most appropriate method for the types of data and the research questions. Additionally, the PAM algorithm has the advantage of a graphical display of *silhouette plots* which can be used for cluster evaluation (see Section 8.3.7). The final classification produced four community groups based on their potential degree of social interaction and crime and most of the community areas conform well to the clusters in which they are assigned. *Group 1* community areas potentially have a stronger bond of social cohesion and the lowest crime rates below the city average, social networks of are also stronger and positive sentiment is above the city average. Community areas in *Group 2* are potentially second in terms of social cohesion, crime rates are also below the city mean but slightly higher than those in group 1. Following the first two groups in terms of community integration is *Group 3* with community areas that tend to have higher crime rates usually above the city average but slightly lower than those in *Group 4*; social cohesion is potentially weak in these areas and negative sentiment is relatively higher. Group 4 community areas are characteristically more diverse with the highest crime rates and are likely the least cohesive. Social networks are also below the city average. Overall, the classification indicates that different community settings are associated with different crime rates based on their potential levels of community cohesion.

Objective 5: *to extend understanding of the relationship between crime and community cohesion based on insights from traditional and new data sources.* The requirement of this objective was met by chapter 7. The chapter draws on the literature reviewed in chapters 3 and 4, insights on the relationship between community cohesion and crime based on analyses carried out in chapters 5 and 6 using traditional and new social media data sources. It then employed the new metrics of engagement popularity, commitment and virality (PCV) generated from Facebook combined with traditional Census statistics (standard and diversity equivalent) variables in a multiple regression models (global and local) to explore the relationship between community cohesion and burglary in Leeds. The introduction of the V-engagement indicator and the use of a more robust local regression model (GWR) contributed to an improvement in the model performance. For example, while the *traditional OLS* model (in chapter 5) explains only 23% in the variation of burglary rates, *non-traditional global*

OLS produced 43% compared to the *local GWR* which explains 63% in variation of burglary rates in Leeds (see Section 7.7.1). Generally, the model performs better with high local *r*-squared values around community areas on the fringes of the district. Though there might be edge some effects in the northeast and southern community areas bordering with rural areas. *GWR* is more appropriate for spatial modelling and it captures the effects of spatial heterogeneity better. This is also novel as the model has accurately captured to a large extent the contrasting variations in community make-up in Leeds.

The distribution of the coefficients also highlighted interesting patterns in the local variations in the relationship between community cohesion and rates of burglary in contrasting community areas of Leeds district. For example, the spatial distribution of the *V* coefficient shows that greater effect of virality of information in reducing burglary rates to be higher in areas with a higher burglary rates (such as city centre, student areas and disadvantaged neighbourhoods) than areas with a lower rates of burglary (such as affluent suburbs), suggesting a greater need for the creation of social networks of community engagement towards creating awareness on neighbourhood safety which can lead to community action. The effect of differences in age distribution on burglary is more likely to be higher in areas with lower age diversity, this suggests strengthening the collective efficacy of members in order to control the socially unacceptable (antisocial) behaviours amongst the community members especially, the young which could lead to offending.

The relationship between residential diversity and burglary is complex because while residential instability is associated with increases in burglary rates which can cause people to move or change their residences. It can also disrupt existing bonds of social networks in otherwise more residentially stable communities. Consequently increasing the risk of victimisation which indicate the need for closer monitoring of increasing population churn in different areas.

The economically inactive population are more likely to be victims of burglary crime than those who are economically active as UK statistics have consistently shown (ONS, 2014b; ONS, 2016a). The vulnerability situation of these segments of the population that include full-time students, retired and people living with long-term illness is a major constraint for them to establish meaningful community cohesion so as to reduce burglary rates in their areas.

9.3 Limitations of the Research

This thesis is innovative in its use of new social media data sources that extend our understanding of the relationship between community cohesion and crime. New data sources from social media such as Facebook have profound implications for community building (Ellison *et al.*, 2007), understanding the dynamics of local community engagement (Matthews, 2015; Matthews, 2016) and collective community action (Harris and McCabe, 2017). Data from these sources, combined with the traditional variables (diversity statistics) constructed from the Census 2011 data (chapter 5), has further advanced our understanding of the impact of community cohesion especially on rates of burglary in Leeds. The new metrics (PCV) of engagement proposed in chapter 7 are potentially useful for highlighting insights into social engagement in different community areas, therefore appropriate for adoption as a proxy for exploring community cohesion. This is especially important owing to lack of statistics to quantify community cohesion from the official government datasets. However, a major drawback of data obtained from social media (Facebook and Twitter) is that they may not be a strong representation of the general population and therefore sampling bias is likely. It is also likely that the problem of the *digital divide* (see Section 7.6.3) can also affect the data collection such that some community areas might not be well represented. The results involving the use of social media data needs to be interpreted with caution, as the science behind these methods is still evolving.

Moreover, sentiment classification of tweets is a challenging task and sarcastic sentences are often difficult to analyse, which can affect the performance of sentiment algorithms. The algorithm used in this research is not an exception to this problem which resulted in large proportion of tweets classified as neutral sentiment. Additionally, like any social media, the number of fans on Facebook pages is subject to change which can affect the post interactions metrics over time in different areas, though changes in these metrics are not being considered over time. Thus, it is possible for the engagement rate to change as the interaction behaviour of users' changes. Similarly, some community areas do not maintain a Facebook page, while some do not post regularly. These problems limit the measuring of engagement rate and by reducing the amount of data available across the entire district.

Mapping the coefficients from the local modelling technique (GWR) used in chapter 7, has provided more insights into the spatial heterogeneity of the relationship between different parameters and burglary rates across contrasting community areas of Leeds. The GWR model has also provided a useful way of ameliorating the effect of the ecological fallacy that may

otherwise arise as a result of the modifiable areal unit problem (MAUP). The model results are important for policy planning aimed at community building and for increasing safety in different community areas. However, high standard error statistics and non-significant local p-values of the parameters in some areas indicate that the model might be missing some key variables. This is an area for future research.

Furthermore, the partitioning clustering algorithms used in chapter 8 enabled a new classification of Leeds community areas to be successfully developed. Considerations were made in constructing a classification that will reflect true community differences in Leeds in terms of social cohesion and crime rates. The majority of the community areas conform well in the clusters in which they were assigned. However, the process of building a classification of this kind (using multidimensional variables) might not necessarily be appropriate for everyone. Despite this limitation, the new typology produced in this research has extended our understanding of the impact of community building on crime rates in different community areas in Leeds.

9.4 Recommendations for Future Research

Having summarised the research findings and achievements as well as highlighting the limitations of the research, there are several directions from which the work undertaken in this thesis can be extended. Firstly, the standard regression variables used in this thesis for constructing the diversity statistics to capture range loci in which social cohesion plays a part across burglary rates can be applied to model different crime types.

Secondly, the algorithm used for sentiment analysis in this thesis employed a dictionary of 6,800 positive and negative words Hu and Liu (2004), which in its simplest form is capable of analysing the sentiment polarity in tweets. There is the potential for extending the performance of sentiment classification through improving the algorithm, so as to capture the sarcastic nature of tweets, reducing the proportion of neutral (unclassified) tweets and increasing the accuracy of the prediction. Additionally, topic modelling can be made used to improve selection of tweets referring to localities. Topic modelling approach group document into topic and then associate words into topics which can then be used for identifying and filtering of tweets referring to community/place.

Thirdly, the likely issues with sampling biases associated with social media data might be improved in the future by inferring the demographic characteristics of the users (such as ethnicity, age and gender) then comparing with conventional geodemographic data sources

such as Census statistics for each spatial unit (such as LSOA) (e.g Longley *et al.*, 2015). Reweighting of the sample data can also be performed and necessary adjustment made to improve the accuracy of the model prediction (e.g Culotta, 2014). Although there is concern that confidentiality of the personal information of users can be compromised, these issues do not apply to this research.

Finally, the predictive power of the new PCV metrics of engagement proposed in this thesis could be improved through normalisation. Facebook is used because, at the time of writing this thesis, it is the largest and most widely used social media channel for social engagement in different community areas. Similar metrics from new and emerging data sources (big data) that measure social engagement can also be explored. For example, Foursquare data is an emerging data source that captures elements of social networking that can potentially be used as a proxy to quantify the degree of social interaction in a community.

9.5 Concluding Remarks

This work is novel and exciting. The thesis has explored the use of *new social media for quantifying the relationship between community cohesion and residential burglary* in contrasting community areas of Leeds district. The research work has contributed in different ways in extending our understanding of the relationship between social cohesion and crime. The central contribution of this thesis is the use of new metrics that estimate popularity, commitment and virality known as the *PCV indicators* for quantifying community cohesion on social media. These metrics, combined with diversity statistics constructed from “traditional” Census data, provide a better correlate of community cohesion and crime. Variations in burglary rates are attributed to the differences in the structural characteristics of the local areas, which relates to the degree of their social interaction. Although the novel approaches employed in this research can be improved because social cohesion is a nuanced concept, it is hoped that the findings of the research will help in policy planning aimed at community building.

List of References

- Aarts, O., Van Maanen, P. P., Ouboter, T. and Schraagen, J. M. (2012). Online social behavior in twitter: A literature review. *12th International Conference on Data Mining Workshops (ICDMW)*, 739-746.
- Abdi, H. and Williams, L. (2010). Normalizing data. *Encyclopedia of research design*, 935-938.
- Ackerman, J. M. and Rossmo, D. K. (2015). How far to travel? A multilevel analysis of the residence-to-crime distance. *Journal of Quantitative Criminology*, 31, 237-262.
- Ackerman, W. V. and Murray, A. T. (2004). Assessing spatial patterns of crime in Lima, Ohio. *Cities*, 21, 423-437.
- Adnan, M. (2011). *Towards real-time geodemographic information systems: design, analysis and evaluation*. PhD thesis, University College London.
- Adnan, M., Lansley, G. and Longley, P. A. (2013). A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London. *Proceedings of the 21st Conference on GIS Research UK (GISRUK)*, 1-6.
- Adolf, M. and Deicke, D. (2014). New modes of integration: Individuality and sociality in digital networks. *First Monday*, 20, 1.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011). Sentiment analysis of twitter data. *Proceedings of the workshop on languages in social media*, 30-38.
- Agostino, D. and Arnaboldi, M. (2016). A measurement framework for assessing the contribution of social media to public engagement: An empirical analysis on Facebook. *Public Management Review*, 18, 1289-1307.
- Ahn, J. (2012). Teenagers' experiences with social network sites: Relationships to bridging and bonding social capital. *The Information Society*, 28, 99-109.
- Ajimotokin, S., Haskins, A. and Wade, Z. (2015). The effects of unemployment on crime rates in the US. Research papers, Ivan Allen College, Georgia.
- Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of econometrics*, 16, 3-14.
- Akçomak, I. S. and Ter Weel, B. (2012). The impact of social capital on crime: Evidence from the Netherlands. *Regional Science and Urban Economics*, 42, 323-340.
- Akkaya, C., Conrad, A., Wiebe, J. and Mihalcea, R. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. *Proceedings of the NAACL HLT*, 195-203.
- Akogul, S. and Erisoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. *Entropy*, 19, 452.
- Alasuutari, P., Bickman, L. and Brannen, J. (2008). *The SAGE Handbook of Social Research Methods*, Sage.
- Alesina, A. and Ferrara, E. L. (2005). Ethnic diversity and economic performance. *Journal of economic literature*, 43, 762-800.

- Alesina, A., Harnoss, J. and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, 21, 101-138.
- Alessia, D., Ferri, F., Grifoni, P. and Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125.
- Alexiou, A., Singleton, A. and Longley, P. A. (2016). A classification of multidimensional open data for urban morphology. *Built Environment*, 42, 382-395.
- Alistair, K. and Diana, I. (2005). Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of FINEXIN*.
- Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*, Sas Institute.
- Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., Serre, M. L. and Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto international*, 25, 443-452.
- Alok, A. K., Saha, S. and Ekbal, A. (2014). Development of an external cluster validity index using probabilistic approach and min-max distance. *IJCISIM*, 6, 494-504.
- Alves, L. G., Ribeiro, H. V., Lenzi, E. K. and Mendes, R. S. (2013). Distance to the scaling law: a useful approach for unveiling relationships between crime and urban metrics. *PLoS One*, 8, e69580.
- Amin, A. (2008). Collective culture and urban public space. *City*, 12, 5-24.
- Anderson, C. (2008). The end of theory. *Wired magazine*, 16, 16-07.
- Andresen, M. A. (2006). Crime measures and the spatial analysis of criminal activity. *British Journal of Criminology*, 46, 258-285.
- Andresen, M. A. (2011). The ambient population and crime analysis. *The Professional Geographer*, 63, 193-212.
- Andrews, J. L. and McNicholas, P. D. (2014). Variable selection for clustering and classification. *Journal of Classification*, 31, 136-153.
- Andritsos, P. (2002). Data clustering techniques. *Rapport technique, University of Toronto. Department of Computer Science*.
- Angus, M., Kevin, H. and Gooweon, J. (2015). Community Action and Social Media. Available from: <http://www.birmingham.ac.uk/generic/tsrc/documents/tsrc/news/2015/CASM-position-paper-feb-2015-FINAL.pdf> [Accessed 04/07/2015].
- Anselin, L. (1989). What is special about spatial data? Alternative perspectives on spatial data analysis. Symposium on spatial statistics, past, present and future, Department of Geography, Syracuse University, (89-4).
- Anselin, L., Cohen, J., Cook, D., Gorr, W. and Tita, G. (2000). Spatial analyses of crime. *Criminal justice*, 4, 213-262.
- Anthony, M. (2016). Popular social media sites. Available from: <https://smallbiztrends.com/2016/05/popular-social-media-sites.html>. [Accessed 05/05/2017].

- Anton, C. E. and Lawrence, C. (2014). Home is where the heart is: The effect of place of residence on place attachment and community participation. *Journal of Environmental Psychology*, 40, 451-461.
- Appel, L., Dadlani, P., Dwyer, M., Hampton, K., Kitzie, V., Matni, Z. A., Moore, P. and Teodoro, R. (2014). Testing the validity of social capital measures in the study of information and communication technologies. *Information, Communication & Society*, 17, 398-416.
- Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46, 243-256.
- Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, 40-79.
- Armstrong, T. A., Katz, C. M. and Schnebly, S. M. (2015). The relationship between citizen perceptions of collective efficacy and neighborhood violent crime. *Crime & Delinquency*, 61, 121-142.
- Arnio, A. N. and Baumer, E. P. (2012). Demography, foreclosure, and crime: Assessing spatial heterogeneity in contemporary models of neighborhood crime rates. *Demographic Research*, 26, 449-488.
- Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49, 45-53.
- Arundel, R. and Ronald, R. (2017). The role of urban form in sustainability of community: The case of Amsterdam. *Environment and Planning B: Urban Analytics and City Science*, 44, 33-53.
- Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R. and Kundi, F. M. (2016). SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, 5, 1.
- Ashby, D. I. (2005). Policing neighbourhoods: Exploring the geographies of crime, policing and performance assessment. *Policing & Society*, 15, 413-447.
- Asur, S. and Huberman, B. (2010). Predicting the future with social media. *International conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, IEEE/WIC/ACM, 1, 492-499.
- Atkinson, G. and Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports medicine*, 26, 217-238.
- Auerhahn, K. (2008). Using simulation modeling to evaluate sentencing reform in California: choosing the future. *Journal of Experimental Criminology*, 4, 241-266.
- Augustyniak, Ł., Szymański, P., Kajdanowicz, T. and Tuligłowicz, W. (2015). Comprehensive study on lexicon-based ensemble classification sentiment analysis. *Entropy*, 18, 4.
- Babb, P. (2005). Measurement of social capital in the UK. *Statistics, Knowledge and Policy*, 532.
- Bae, J. W., Paik, E., Kim, K., Singh, K. and Sajjad, M. (2016). Combining microsimulation and agent-based model for micro-level population dynamics. *Procedia Computer Science*, 80, 507-517.

- Bailey, S. L. (2015). A Microsimulation Model to Assess the Impact of Prevention Efforts to Combat Sex Trafficking out of Five Eastern European States. *Journal of Human Trafficking*, 1, 167-186.
- Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*, Longman Scientific & Technical Essex.
- Bains, B. (2005). Data Management and Analysis Group: Ethnic Diversity Indices [Online]. <http://legacy.london.gov.uk/gla/publications/factsandfigures/dmag-briefing-2005-12.pdf>
- Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L. and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. 49-58.
- Ballas, D., Broomhead, T. and Jones, P. M. 2019. Spatial Microsimulation and Agent-Based Modelling. *The Practice of Spatial Analysis*. Springer.
- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B. and Rossiter, D. (2005). SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place*, 11, 13-34.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*, Macmillan.
- Bandyopadhyay, S., Bhattacharya, S. and Han, L. (2010). Determinants of Violent and Property Crimes in England and Wales: A Panel Data Analysis. Available at SSRN 1691801.
- Barbera, P., Micheal, P. and Andrew, G. (2017). Rfacebook. R package version 0.6.11. Available from: <https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf>. [Accessed 23/05/2017].
- Barberet, R. and Fisher, B. S. (2009). Can security beget insecurity? Security and crime prevention awareness and fear of burglary among university students in the East Midlands. *Security Journal*, 22, 3-23.
- Bargh, J. A. and McKenna, K. Y. (2004). The Internet and social life. *Annu. Rev. Psychol.*, 55, 573-590.
- Barnes, T. J. (2013). Big Data, little history. *Dialogues in human geography*, 3, 297-302.
- Barrantes, G. and Sandoval, L. (2009). Conceptual and statistical problems associated with the use of diversity indices in ecology. *Revista de biología tropical*, 57, 451-460.
- Baruah, T. D. (2012). Effectiveness of Social media as a tool of communication and its potential for technology enabled connections: A micro level study. *International Journal of Scientific and Research Publications*, 2, 1-10.
- Bate, A. (2017). The troubled families programme (England). Available from: <http://dera.ioe.ac.uk/28802/1/CBP-7585.pdf>. [Accessed 10/04/2018].
- BBC (2003). Northern tops 'real' rich league. Carole Mitchell report. Available from: <http://news.bbc.co.uk/1/hi/business/3025321.stm#text>. [Accessed 17/03/2016].
- Bechmann, A. (2014). Non-informed consent cultures: Privacy policies and app contracts on Facebook. *Journal of Media Business Studies*, 11, 21-38.
- Beil, F., Ester, M. and Xu, X. (2002). Frequent term-based text clustering. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 436-442.

- Bell, B. and Machin, S. (2011). The impact of migration on crime and victimisation. A report for the Migration Advisory Committee. *London: UK*.
- Bellair, P. E. (1997). Social interaction and community crime: Examining the importance of neighbor networks. *Criminology*, 35, 677-704.
- Bendler, J., Brandt, T., Wagner, S. and Neumann, D. (2014). Investigating crime-to-twitter relationships in urban environments-facilitating a virtual neighborhood watch. *Twenty Second European Conference on Information Systems, Tel Aviv*.
- Bendror, Y. (2013). Interpreting Facebook page insights: Reach vs engagement. Available from: <https://ymarketingmatters.com/interpreting-facebook-insights-reach-vs-engagement/> . [Accessed 15/02/2017].
- Bennet, S. (2014). Media Stats [Online]. <http://www.adweek.com/socialtimes/social-media-statistics-2014/499230>. [Accessed 03/07/2015].
- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M. and Mikhaylov, S. (2016). Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review*, 110, 278-295.
- Bentley, R. A., O'Brien, M. J. and Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences*, 37, 63-76.
- Berker, M. (2011). *Using genetic algorithms with lexical chains for automatic text summarization*. Bogaziçi University.
- Berkhin, P. 2006. A survey of clustering data mining techniques. *Grouping multidimensional data*. Springer.
- Bernasco, W. and Nieuwbeerta, P. (2005). How do residential burglars select target areas? A new approach to the analysis of criminal location choice. *British Journal of Criminology*, 45, 296-315.
- Berrington, A. (2014). The changing demography of lone parenthood in the UK. ESRC Centre for Population Change. Working paper No. 48.
- Berry, G., Briggs, P., Erol, R. and Van Staden, L. (2011a). The Effectiveness of Partnership Working in a Crime and Disorder Context. *A rapid evidence assessment*, 1.
- Berry, G., Briggs, P., Erol, R. and van Staden, L. (2011b). The effectiveness of partnership working in a crime and disorder context: A rapid evidence assessment. *H. Office (Ed.), Research Report*, 52.
- Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A. and Bar-Yam, Y. (2013). Sentiment in New York city: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*.
- Bhattacharya, S., Srinivasan, P. and Polgreen, P. (2017). Social media engagement analysis of US Federal health agencies on Facebook. *BMC medical informatics and decision making*, 17, 49.
- Bholowalia, P. and Kumar, A. (2014). EBK-Means: A Clustering technique based on elbow method and K-Means in WSN. *International Journal of Computer Applications*, 105.
- Bianchi, M., Buonanno, P. and Pinotti, P. (2012). Do immigrants cause crime? *Journal of the European Economic Association*, 10, 1318-1347.

- Bijuraj, L. (2013). Clustering and its applications. *Proceedings of National Conference on New Horizons in IT-NCNHIT*, 169.
- Bingham-Hall, J. and Law, S. (2015). Connected or informed?: Local Twitter networking in a London neighbourhood. *Big Data & Society*, 2, 2053951715597457.
- Birani, A. and Lehmann, W. (2013). Ethnicity as social capital: an examination of first-generation, ethnic-minority students at a Canadian university. *International Studies in Sociology of Education*, 23, 281-297.
- Birkin, M. and Clarke, G. (2012). The enhancement of spatial microsimulation models using geodemographics. *The Annals of Regional Science*, 49, 515-532.
- Blank, G. (2017). The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, 35, 679-697.
- Blau, J. R. and Blau, P. M. (1982). The cost of inequality: Metropolitan structure and violent crime. *American Sociological Review*, 114-129.
- Blonigen, D. M. (2010). Explaining the relationship between age and crime: Contributions from the developmental literature on personality. *Clinical Psychology Review*, 30, 89-100.
- Blunkett, D. (2003). Security and Justice, Mutuality and Individual Rights. Home Office. Available from: <http://webarchive.nationalarchives.gov.uk/20130128103514/http://homeoffice.gov.uk/docs/johnjayspeech.html>. [Accessed 28/03/2016].
- Boateng, F. D. (2016). Fearfulness in the community empirical assessments of influential factors. *Journal of interpersonal violence*, 0886260516642295.
- Bogges, L. N. and Hipp, J. R. (2010). Violent crime, residential instability and mobility: Does the relationship differ in minority neighborhoods? *Journal of Quantitative Criminology*, 26, 351-370.
- Boldt, L. C., Vinayagamorthy, V., Winder, F., Schnittger, M., Ekran, M., Mukkamala, R. R., Lassen, N. B., Flesch, B., Hussain, A. and Vatrappu, R. (2016). Forecasting Nike's sales using Facebook data. IEEE International Conference on Big Data. 2447-2456.
- Bolla, R. A. (2014). *Crime pattern detection using online social media*. PhD Thesis, Missouri University of Science and Technology.
- Bollen, J., Mao, H. and Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11, 450-453.
- Bolton, R. N., Parasuraman, A., Hoefnagels, A., Migchels, N., Kabadayi, S., Gruber, T., Komarova Loureiro, Y. and Solnet, D. (2013). Understanding generation Y and their use of social media: A review and research agenda. *Journal of Service Management*, 24, 245-267.
- Bonnand, S. (2015). *Innovative solutions for building community in academic libraries*, IGI Global.
- Bonsón, E. and Ratkai, M. (2013). A set of metrics to assess stakeholder engagement and social legitimacy on a corporate Facebook page. *Online Information Review*, 37, 787-803.

- Bonsón, E., Royo, S. and Ratkai, M. (2014). Facebook Practices in Western European Municipalities An Empirical Analysis of Activity and Citizens' Engagement. *Administration & Society*, 0095399714544945.
- Bonsón, E., Royo, S. and Ratkai, M. (2015). Citizens' engagement on local governments' Facebook sites. An empirical analysis: The impact of different media and content types in Western Europe. *Government Information Quarterly*, 32, 52-62.
- Boomija, M. and Phil, M. (2008). Comparison of partition based clustering algorithms. *Journal of Computer Applications*, 1, 18-21.
- Bormann, S.-K. (2013). Sentiment indices on financial markets: What do they measure? *Kiel Institute for the World Economy, Economics Discussion Paper*, 58, 2013.
- Botchan, N. (2012). *Recognizing Meaning in the Crowd: Building Word Sense Inventories on Amazon Mechanical Turk.*, Master Dissertation, University of Brandeis, Massachusetts, USA.
- Bottoms, A. and Wiles, P. (1986). Housing tenure and residential community crime carriers in Britain. *Crime and Justice*, 8, 101-162.
- Bouazizi, M. and Ohtsuki, T. (2017). A pattern-based approach for multi-class sentiment analysis in Twitter. *IEEE Access*.
- Bourdieu, P. (1989). Social space and symbolic power. *Sociological Theory*, 7, 14-25.
- Bourdieu, P. and Wacquant, L. J. (1992). *An invitation to reflexive sociology*, University of Chicago press.
- Bowers, K. J., Johnson, S. D. and Pease, K. (2004). Prospective hot-spotting the future of crime mapping? *British Journal of Criminology*, 44, 641-658.
- Brantingham, P. J. and Brantingham, P. L. (1984). *Patterns in crime*, Macmillan New York.
- Brantingham, P. L. and Brantingham, P. J. (1991). *Environmental Criminology*, Prospect Heights IL, Waveland Press.
- Brantingham, P. L. and Brantingham, P. J. (2004). Computer simulation as a tool for environmental criminologists. *Security Journal*, 17, 21-30.
- Brantingham, P. L., Glasser, U., Kinney, B., Singh, K. and Vajihollahi, M. (Year) Published. A computational model for simulating spatial aspects of crime in urban environments. *Systems, Man and Cybernetics*, 2005 IEEE International Conference on, 2005. IEEE, 3667-3674.
- Brantingham, A. and Brantingham, P. L. (1981). *Introduction: Dimension of crime*, Beverly Hills, Sage Publication.
- Braun, M. T. and Oswald, F. L. (2011). Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behavior Research Methods*, 43, 331-339.
- Breen, J. (2011). R by example: mining Twitter for consumer attitudes towards airlines. *Boston Predictive Analytics Meetup Presentation*.
- Brehm, J. and Rahn, W. (1997). Individual-level evidence for the causes and consequences of social capital. *American journal of political science*, 999-1023.

- Brewer, C. A. (1997). Spectral schemes: Controversial color use on maps. *Cartography and Geographic Information Systems*, 24, 203-220.
- Brewer, C. A. and Pickle, L. (2002). Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92, 662-681.
- Briggs, X. N. D. S. and Wilson, W. J. (2005). *The geography of opportunity: Race and housing choice in metropolitan America*, Brookings Institution Press.
- Brilli, Y. and Tonello, M. (2014). Rethinking the crime reducing effect of education: the role of social capital and organized crime. Working paper no. 19, European University Institute.
- Browning, C. R. (2002). The span of collective efficacy: Extending social disorganization theory to partner violence. *Journal of Marriage and Family*, 64, 833-850.
- Browning, C. R., Burrington, L. A., Leventhal, T. and Brooks-Gunn, J. (2008). Neighborhood structural inequality, collective efficacy, and sexual risk behavior among urban youth. *Journal of Health and Social Behavior*, 49, 269-285.
- Browning, C. R., Dietz, R. D. and Feinberg, S. L. (2004). The paradox of social organization: Networks, collective efficacy, and violent crime in urban neighborhoods. *Social Forces*, 83, 503-534.
- Bruce, C. and Santos, R. B. (2011). Crime Pattern Definitions for Tactical Analysis: White paper international association of crime analysts. Available from: http://www.iaca.net/Publications/Whitepapers/iacawp_2011_01_crime_patterns.pdf. [Accessed 02/02/2018].
- Bruch, E. and Atwell, J. (2015). Agent-based models in empirical social research. *Sociological methods & research*, 44, 186-221.
- Bruinsma, G. J., Pauwels, L. J., Weerman, F. M. and Bernasco, W. (2013). Social disorganization, social capital, collective efficacy and the spatial distribution of crime and offenders: An empirical test of six neighbourhood models for a Dutch city. *British Journal of Criminology*, 53, 942-963.
- Bruns, A. and Stieglitz, S. (2014). Metrics for understanding communication on Twitter. *Twitter and Society*, 89, 69-82.
- Brunsdon, C., Charlton, M. and Rigby, J. E. (2016). An open source geodemographic classification of small areas in the Republic of Ireland. *Applied Spatial Analysis and Policy*, 1-22.
- Brunsdon, C., Corcoran, J. and Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers Environment and Urban Systems*, 31, 52-75.
- Bryman, A. (2008). The end of the paradigm wars. *The Sage handbook of social research methods*, 13-25.
- Buhrmester, M., Kwang, T. and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3-5.
- Buonanno, P., Montolio, D. and Vanin, P. (2009). Does social capital reduce crime? *Journal of Law and Economics*, 52, 145-170.

- Burchfield, K. B. (2009). Attachment as a source of informal social control in urban neighborhoods. *Journal of Criminal Justice*, 37, 45-54.
- Burke, M., Kraut, R. and Marlow, C. (2011). Social capital on Facebook: Differentiating uses and users. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 571-580.
- Burmania, A., Parthasarathy, S. and Busso, C. (2016). Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7, 374-388.
- Bursik Jr, R. J. and Grasmick, H. G. (1993). Economic deprivation and neighborhood crime rates, 1960-1980. *Law & Society Review*, 27, 263.
- Bursik, R. J. (1986). Ecological stability and the dynamics of delinquency. *Crime and Justice-a Review of Research*, 8, 35-66.
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365-376.
- Cabrera, J. F., Scholz, S., Hobor, G. and Lizardo, O. (2017). Integrating “standard” residents into “non-standard” communities: a longitudinal analysis of social capital in a new urbanist development. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 10, 63-76.
- Caesar, C. and Kopsch, F. (2018). Municipal land allocations: a key for understanding tenure and social mix patterns in Stockholm. *European Planning Studies*, 1-19.
- Calder, M., Craig, C., Culley, D., de Cani, R., Donnelly, C. A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N. and Hargrove, C. (2018). Computational modelling for decision-making: where, why, what, who and how. *Royal Society open science*, 5, 172096.
- Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 286-295.
- Cambria, E. (2013). An introduction to concept-level sentiment analysis. *Mexican International Conference on Artificial Intelligence*, 478-483.
- Campbell, D. and Campbell, S. 2008. Introduction to regression and data analysis. *Stat Lab Workshop Series*, pp1-15.
- Cantle, T. (2001). Community cohesion: A report of the independent review team. Home Office, London.
- Caplan, J. M., Kennedy, L. W., Barnum, J. D. and Piza, E. L. (2015). Risk terrain modeling for spatial risk assessment. *Cityscape*, 17, 7.
- Caragea, C., Squicciarini, A. C., Stehle, S., Neppalli, K. and Tapia, A. H. (2014). Mapping moods: Geo-mapped sentiment analysis during hurricane sandy. *ISCRAM*.
- Carcach, C. and Huntley, C. (2002). Community participation and regional crime. *Trends & Issues in Crime and Criminal Justice*, 1.
- Caruso, R. (2011). Crime and sport participation: Evidence from Italian regions over the period 1997–2003. *The Journal of Socio-Economics*, 40, 455-463.

- Catts, R. and Ozga, J. (2005). *What is Social Capital and how Might it be Used in Scotland's Schools?*, Centre for Educational Sociology.
- Chaffey, D. (2017). Global social media research summary. Available from: <http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>. [Accessed 01/03/2017].
- Chainey, S. and Ratcliffe, J. (2005). *GIS and Crime Mapping*, Wiley.
- Chainey, S. and Ratcliffe, J. (2013). *GIS and crime mapping*, John Wiley & Sons.
- Chainey, S., Reid, S. and Stuart, N. (2002). *When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime in Chainey, S. and Ratcliffe, J. 2005. GIS and Crime Mapping*, Taylor and Francis, London, England.
- Chainey, S., Tompson, L. and Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21, 4-28.
- Chamberlain, A. W. and Boggess, L. N. (2016). Relative difference and burglary location: Can ecological characteristics of a burglar's home neighborhood predict offense location? *Journal of Research in Crime and Delinquency*, 53, 872-906.
- Charlton, M., Fotheringham, S. and Brunsdon, C. (2009). Geographically weighted regression. *White paper. National Centre for Geocomputation. National University of Ireland Maynooth*.
- Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A. and Charrad, M. M. (2014). Package 'NbClust'. *J. Stat. Soft*, 61, 1-36.
- Chaudhry, I. (2015). # Hashtagging hate: Using Twitter to track racism online. *First Monday*, 20.
- Chiang, M. M.-T. and Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of classification*, 27, 3-40.
- Chiu, C.-M., Hsu, M.-H. and Wang, E. T. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems*, 42, 1872-1888.
- Choi, C. G. and Choi, S. O. (2012). Collaborative partnerships and crime in disorganized communities. *Public Administration Review*, 72, 228-238.
- Chung, J. E. and Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? . *25 AAAI conference on Artificial Intelligence*, 11, 1770-1771.
- Clark, A. (2007). Understanding communities: A review of networks. Available from: http://eprints.ncrm.ac.uk/469/1/0907_understanding_community.pdf. [Accessed 13/05/2016].
- Clark, T. S. and Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, 3, 2, 399-408.
- Clarke, R. and Hope, T. (2012). *Coping with burglary: Research perspectives on policy*, Springer Science & Business Media.
- Coe, R. (2002a). Analyzing ranking and rating data from participatory on-farm trials. in *Mauricio R.B. and Jane R. (eds.). Quantitative Analysis of Data from Participatory Methods in Plant Breeding. CIMMYT, Mexico*, 44-65.

- Coe, R. (2002b). It's the effect size, stupid. *Paper presented at the British Educational Research Association annual conference. University of Exeter, England.*
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, L. E. and Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American sociological review*, 588-608.
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, S95-S120.
- Coleman, J. S. (1990). *Foundations of social theory*, Harvard University Press.
- Coleman, S. (2002). A test for the effect of conformity on crime rates using voter turnout. *The Sociological Quarterly*, 43, 257-276.
- Collins, K., Babyak, C. and Molone, J. (2006). Treatment of spatial autocorrelation in geocoded crime data. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 2864-2871.
- Collins, K., Babyak, C. and Moloney, J. (2007). Spatial Modeling of Geocoded Crime Data. Upcoming Methodology Branch Working Paper, Statistics Canada, Ottawa, ON.
- Cooley, S. and Jones, A. (2013). A forgotten tweet: Somalia and social media. *Ecquid Novi: African Journalism Studies*, 34, 68-82.
- Cooper, H. and Innes, M. (2009). The Causes and Consequences of Community Cohesion in Wales: A Secondary Analysis. *Cardiff: UPSI, Cardiff University*, 36, 64-5.
- Cornelius, I., Komito, L. and Bates, J. (2009). Virtually local: social media and community among Polish nationals in Dublin. *Aslib Proceedings*, 61, 232-244.
- Cornish, D. B. and Clarke, R. V. (1987). Understanding crime displacement: An application of rational choice theory. *Criminology*, 25, 933-948.
- Cornish, R. (2007). Statistics: Cluster analysis. *Mathematics Learning Support Centre*.
- Corso, A. J. (2015). Toward Predictive Crime Analysis via Social Media, Big Data, and GIS Spatial Correlation. *iConference 2015 Proceedings*.
- Coulter, K. S., Gummerus, J., Liljander, V., Weman, E. and Pihlström, M. (2012). Customer engagement in a Facebook brand community. *Management Research Review*, 35, 857-877.
- Cracolici, M. F. and Uberti, T. E. (2009). Geographical distribution of crime in Italian provinces: a spatial econometric analysis. *Jahrbuch für Regionalwissenschaft*, 29, 1-28.
- Crawford, A. (1994). Appeals to community and crime prevention. *Crime, Law and Social Change*, 22, 97-126.
- Cromley, R. G. and Cromley, E. K. (2009). Choropleth map legend design for visualizing community health disparities. *International Journal of Health Geographics*, 8, 52.

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Crooks, A. T. and Heppenstall, A. J. 2012. Introduction to agent-based modelling. *Agent-based models of geographical systems*. Springer.
- Crowe, J. (2010). Community attachment and satisfaction: The role of a community's social network structure. *Journal of Community Psychology*, 38, 622-644.
- Crump, J. (2011). What are the police doing on Twitter? Social media, the police and the public. *Policy & Internet*, 3, 1-27.
- Culotta, A. (2014). Reducing sampling bias in social media data for county health inference. *Joint Statistical Meetings Proceedings*, 1-12.
- Curry, M. R. (1997). The digital individual and the private realm. *Annals of the Association of American Geographers*, 87, 681-699.
- Cvijikj, I. P. and Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining*, 3, 843-861.
- Dang, Y., Zhang, Y. and Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25, 46-53.
- Danneman, N. and Heimann, R. (2014). *Social Media Mining with R*, Packt Publishing Ltd.
- DCLG (2015). English indices of deprivation. Available from: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>. [Accessed 07/03/2016].
- De la Torre-Díez, I., Díaz-Pernas, F. J. and Antón-Rodríguez, M. (2012). A content analysis of chronic diseases social groups on Facebook and Twitter. *Telemedicine and e-Health*, 18, 404-408.
- Debenham, J. (2002). Understanding geodemographic classification: Creating the building blocks for an extension. White Paper 02, School of Geography, University of Leeds.
- Deitrick, W., Valyou, B., Jones, W., Timian, J. and Hu, W. (2013). Enhancing sentiment analysis on twitter using community detection. *Communications and Network*, 5, 192-197.
- Delaney, L. and Keaney, E. (2005). Sport and social capital in the United Kingdom: Statistical evidence from national and international survey data. *Dublin: Economic and Social Research Institute and Institute for Public Policy Research*, 32.
- Delhey, J. and Dragolov, G. (2016). Happier together. Social cohesion and subjective well-being in Europe. *International Journal of Psychology*, 51, 163-176.
- Deller, S. and Deller, M. (2012). Spatial heterogeneity, social capital, and rural larceny and burglary. *Rural Sociology*, 77, 225-253.
- Deller, S. C. and Deller, M. A. (2010). Rural crime and social capital. *Growth and Change*, 41, 221-275.
- Delmelle, E. C., Haslauer, E. and Prinz, T. (2013). Social satisfaction, commuting and neighborhoods. *Journal of Transport Geography*, 30, 110-116.

- Demombynes, G. and Özler, B. (2005). Crime and local inequality in South Africa. *Journal of Development Economics*, 76, 265-292.
- Dempsey, N., Brown, C., Raman, S., Porta, S., Jenks, M., Jones, C. and Bramley, G. (2010). *Elements of urban form*, Springer.
- Denef, S., Bayerl, P. S. and Kaptein, N. A. (2013). Social media and the police: tweeting practices of British police forces during the August 2011 riots. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3471-3480.
- Dennett, A. and Stillwell, J. (2011). A new area classification for understanding internal migration in Britain. *Population Trends*, 145, 146-171.
- Desgraupes, B. (2013). Clustering indices. *University of Paris Ouest-Lab ModalX*, 1, 34.
- DiMaggio, P., Hargittai, E., Celeste, C. and Shafer, S. (2004). *Digital inequality: From unequal access to differentiated use in Kathryn N. Social Inequality*, New York, Russell Sage Foundation.
- Dinesen, P. T. and Sønderskov, K. M. (2015). Ethnic Diversity and Social Trust Evidence from the Micro-Context. *American Sociological Review*, 0003122415577989.
- Dodge, M. and Kitchin, R. (2007). " Outlines of a world coming into existence": Pervasive computing and the ethics of forgetting. *Environment and Planning B*, 34, 431-445.
- Dolan, R., Dolan, R., Conduit, J., Conduit, J., Fahy, J., Fahy, J., Goodman, S. and Goodman, S. (2017). Social media: communication strategies, engagement and future research directions. *International Journal of Wine Business Research*, 29, 2-19.
- Drukker, M., Kaplan, C., Feron, F. and Van Os, J. (2003). Children's health-related quality of life, neighbourhood socio-economic deprivation and social capital. A contextual analysis. *Social Science & Medicine*, 57, 825-841.
- Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3, research0036. 1.
- Dunaway, R. G., Cullen, F. T., Burton, V. S. and Evans, T. D. (2000). The myth of social class and crime revisited: An examination of class and adult criminality. *Criminology*, 38, 589-632.
- Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 95-104.
- Dunn, R. (1989). A dynamic approach to two-variable color mapping. *The American Statistician*, 43, 245-252.
- Durkheim, E. (1893). *The Division of Labor in Society*, New York, Free Press.
- Dutton, W. H. and Reisdorf, B. C. (2017). Cultural divides and digital inequalities: attitudes shaping Internet and social media divides. *Information, Communication & Society*, 1-21.
- Duygulu, P., Barnard, K., de Freitas, J. F. and Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *European conference on computer vision*, 97-112.

- DWP (2014). The Use of Social Media for Research and Analysis: A Feasibility Study. Department of Works and Pension. The National Archives. Kew, London TW9 4DU. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/387591/use-of-social-media-for-research-and-analysis.pdf. [Accessed 04/07/2015].
- Dwyer, C., Hiltz, S. and Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. *AMCIS 2007 proceedings*, 339.
- Easterly, W., Ritzen, J. and Woolcock, M. (2006). Social cohesion, institutions, and growth. *Economics & Politics*, 18, 103-120.
- Eberlen, J., Scholz, G. and Gagliolo, M. (2017). Simulate this! An Introduction to Agent-Based Models and their Power to Improve your Research Practice. *International Review of Social Psychology*, 30.
- Eck, J., Chainey, S., Cameron, J. and Wilson, R. (2005). Mapping crime: Understanding hotspots. National Institute of Justice, Washington DC.
- Eck, J. E. and Weisburd, D. (1995). Crime places in crime theory. *Crime and Place, Crime Prevention Studies*, 4, 1-33.
- Efroymsen, M. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, 1, 191-203.
- Ejrnæs, A. and Greve, B. (2017). Your position in society matters for how happy you are. *International Journal of Social Welfare*, 26, 206-217.
- Ellison, N. B. and Boyd, D. (2013). Sociality through social network sites. *The Oxford Handbook of Internet Studies*, 151-172.
- Ellison, N. B., Steinfield, C. and Lampe, C. (2007). The benefits of Facebook “friends:” Social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication*, 12, 1143-1168.
- Ellison, N. B., Vitak, J., Gray, R. and Lampe, C. (2014). Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19, 855-870.
- Epstein, E. S. (1988). Long-range weather prediction: limits of predictability and beyond. *Weather and Forecasting*, 3, 69-75.
- Erdogan, S., Yalçın, M. and Dereli, M. A. (2013). Exploratory spatial analysis of crimes against property in Turkey. *Crime, Law and Social Change*, 59, 63-78.
- Ernest, E. and Bernad, R. (2015). Investigating public universities facebook Pages: Extent of users engagement. *IJALIS* 3(2), PP. 31-36.
- Estacio, E. V., Whittle, R. and Protheroe, J. (2017). The digital divide: Examining socio-demographic factors associated with health literacy, access and use of internet to seek health information. *Journal of health psychology*, 1359105317695429.
- Everitt, B., Landau, S. and Leese, M. (2001). Cluster analysis 4th Edition, Arnold. London.
- Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011). Hierarchical clustering. *Cluster Analysis, 5th Edition*, 71-110.

- Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, 22.
- Facebook (2016). Facebook statistics. Available from: <http://newsroom.fb.com/company-info/>. [Accessed 05/06/2016].
- Facebook (2017a). How is engagement measured? Available from <https://www.facebook.com/help/794890670645072>. [Accessed 21/01/2018].
- Facebook (2017b). What is public information? Available from: <https://www.facebook.com/help/203805466323736>. [Accessed 22/01/2018].
- Fagan, A. A., Wright, E. M. and Pinchevsky, G. M. (2014). The protective effects of neighborhood collective efficacy on adolescent substance use and violence following exposure to violence. *Journal of Jouth and Adolescence*, 43, 1498.
- Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S. and Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2, 267-279.
- Fajnzlber, P., Lederman, D. and Loayza, N. (2002). Inequality and violent crime. *Journal of Law & Economy*, 45, 1.
- Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, 4, e5738.
- Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2, 1.
- Faraway, J. J. (2002). Practical regression and ANOVA using R. University of Bath.
- Farkas, M. A. and Jones, R. S. (2007). Community Partners: 'Doing doors' as a community crime prevention strategy. *Criminal Justice Studies*, 20, 295-312.
- Farrall, S., Hay, C., Jennings, W. and Gray, E. (2015). Thatcherite Ideology, Housing Tenure and Crime: The Socio-spatial Consequences of the Right to Buy for Domestic Property Crime. *British Journal of Criminology*, 56, 1235-1252.
- Farrington, D. P. (1986). Age and crime. *Crime and Justice*, 189-250.
- Featherstone, C. (2013). The relevance of Social Media as it applies in South Africa to crime prediction. *IST-Africa Conference and Exhibition (IST-Africa)*, 2013, 1-7.
- Feinerer, I. and Hornik, K. (2015). Text mining package. Version 0.6-2.
- Fellows, I., Fellows, M. I. and Rcpp, L. (2015). Package 'wordcloud'.
- Felson, M. (1995). Those who discourage crime. *Crime and Place*, 4, 53-66.
- Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, 3, 2053951716645828.
- Fernandez, M., Dickinson, T. and Alani, H. (2017). An analysis of UK Policing Engagement via Social Media. *International Conference on Social Informatics*, 289-304.
- Figueroa, R. L., Zeng-Treitler, Q., Kandula, S. and Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12, 8.
- Flatley, J. (2015). Crime in England and Wales, Year Ending March 2015.

- Flatley, J. (2017). Crime in England and Wales: year ending Mar 2017. Office of National Statistics.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378.
- Fletcher, A., Gardner, F., McKee, M. and Bonell, C. (2012). The British government's troubled families programme. *BMJ (Clinical research ed)*, 344, e3403.
- Forrest, R. and Kearns, A. (2001). Social cohesion, social capital and the neighbourhood. *Urban Studies*, 38, 2125-2143.
- Foster, S., Giles-Corti, B. and Knuiiman, M. (2014). Does fear of crime discourage walkers? A social-ecological exploration of fear as a deterrent to walking. *Environment and Behavior*, 46, 698-717.
- Fotheringham, A. S., Brunson, C. and Charlton, M. (2003). *Geographically weighted regression: The analysis of spatially varying relationships*, John Wiley & Sons.
- Fotheringham, A. S., Charlton, M. E. and Brunson, C. (1998). Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, 30, 1905-1927.
- Fotheringham, S. and Rogerson, P. (2013). *Spatial Analysis and GIS*, Florida, USA, CRC Press.
- Foundation, T. P. (2014). Police Use of Social Media. Available from: <http://www.police-foundation.org.uk/publications/briefings/police-use-of-social-media>. [Accessed 13/02/2017].
- Fraiman, R., Justel, A. and Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, 103, 1294-1303.
- Friendly, M. (2007). A. M. Guerry's "Moral Statistics of France": Challenges for multivariable spatial analysis. *Statistical Science*, 368-399.
- Fukuyama, F. (1996). *Trust: The social virtues and the creation of prosperity*, Free Press, New York.
- Gainous, J. and Wagner, K. M. (2014). *Tweeting to power: The social media revolution in American politics*, USA, Oxford University Press.
- Gale, C. G., Singleton, A., Bates, A. G. and Longley, P. A. (2018). Creating an open geodemographic classification of the UK using 2011 census data. In Stillwell J. (ed.). *The Routledge Handbook of Census Records, Methods and Applications*. Routledge, London, pp 213-229.
- Gallegos, L., Lerman, K., Huang, A. and Garcia, D. (2016). Geography of Emotion: Where in a City are People Happier? *Proceedings of the 25th International Conference Companion on World Wide Web*, 569-574.
- Gamache-O'Leary, V. and Grant, G. (2017). Social media in healthcare. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 3774-3783.
- Gandía, J. L., Marrahí, L. and Huguet, D. (2016). Digital transparency and Web 2.0 in Spanish city councils. *Government Information Quarterly*, 33, 28-39.

- Gao, C. and Liu, J. (2015). Uncovering spatiotemporal characteristics of human online behaviors during extreme events. *PloS one*, 10, e0138673.
- Gao, S., Liu, Y., Wang, Y. and Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17, 463-481.
- Gao, Y., Zhou, Y., Zhou, B., Shi, L. and Zhang, J. (2017). Handling data skew in MapReduce cluster by using partition tuning. *Journal of Healthcare Engineering*, 2017.
- Garcia, R. M., Taylor, R. B. and Lawton, B. A. (2007). Impacts of violent crime and neighborhood structure on trusting your neighbors. *Justice Quarterly*, 24, 679-704.
- García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. and Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1, 9.
- Gartner, R. (2013). Neighbourhood Change and the Spatial Distribution of Violent Crime. Available from: <http://neighbourhoodchange.ca/wp-content/uploads/2014/04/2de90c57bfeb615f986fbdd82d49c6c8.pdf>. [Accessed 04/03/2016].
- Gendrot, S. B. (2001). The politics of urban crime. *Urban Studies*, 915-928 in Gumus, E. 2004. Crime in urban areas: An empirical investigation. *Akdeniz IIBF Dergisi*, 4, 98-109.
- Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115-125.
- Gerell, M. (2014). Collective efficacy, neighborhood and geographical units of analysis: Findings from a case Study of Swedish residential neighborhoods. *European Journal on Criminal Policy and Research*, 1-22.
- Gerell, M. and Kronkvist, K. (2016). Violent crime, collective efficacy and city-centre effects in Malmö. *British Journal of Criminology*, azw074.
- Gerlitz, C. and Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal*, 16.
- Gerritsen, C. (2015). Agent-based modelling as a research tool for criminological research. *Crime science*, 4, 2.
- Getis, A. (1999). Spatial statistics. *Geographical Information Systems*, 1, 239-251.
- Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24, 189-206.
- Ghosh, D. and Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, 40, 90-102.
- Gijsberts, M., Van Der Meer, T. and Dagevos, J. (2012). 'Hunkering down' in multi-ethnic neighbourhoods? The effects of ethnic diversity on dimensions of social cohesion. *European Sociological Review*, 28, 527-537.
- Gil de Zúñiga, H., Jung, N. and Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, 17, 319-336.

- Gilani, Z., Crowcroft, J., Farahbakhsh, R. and Tyson, G. (2017). The Implications of Twitterbot Generated Data Traffic on Networked Systems. *Proceedings of the SIGCOMM Posters and Demos*, 51-53.
- Gilchrist, A. and Kyprianou, P. (2011). Social networks, poverty and ethnicity. *Programme paper, Joseph Rowntree Foundation, York*.
- Gill, C., Weisburd, D., Telep, C. W., Vitter, Z. and Bennett, T. (2014). Community-oriented policing to reduce crime, disorder and fear and increase satisfaction and legitimacy among citizens: a systematic review. *Journal of Experimental Criminology*, 10, 399-428.
- Gilliland, D. and Melfi, V. (2010). A note on confidence interval estimation and margin of error. *Journal of Statistics Education*, 18, 1-8.
- Glodava, K. M. (2015). *Using Twitter Data as a Community Policing Mechanism of Criminal Activity in Washington DC. Master Dissertation, George Mason University, Virginia, USA.*
- Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1, 12.
- Goldfine, E. (2011). *Best practices: The use of social media throughout emergency & disaster relief*. American University Washington, DC.
- Gonçalves, P., Araújo, M., Benevenuto, F. and Cha, M. (2013). Comparing and combining sentiment analysis methods. *Proceedings of the first ACM conference on Online social networks*, 27-38.
- González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J. and Moreno, Y. (2012). Assessing the bias in communication networks sampled from twitter. *arXiv preprint arXiv:1212.1684*.
- Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in human geography*, 3, 280-284.
- Goodin, R. E. (2014). *Just the facts 101: The Oxford handbook of political science (1st edition)*, Cram 101 publishing.
- Gorman, D. M., Gruenewald, P. J. and Waller, L. A. (2013). Linking places to problems: geospatial theories of neighborhoods, alcohol and crime. *GeoJournal*, 78, 417-428.
- Gorman, S. P. (2013). The danger of a big data episteme and the need to evolve geographic information systems. *Dialogues in Human Geography*, 3, 285-291.
- Goss, J. (1995). " We know who you are and we know where you live": The instrumental rationality of geodemographic systems. *Economic Geography*, 171-198.
- Gottfredson, M. R. and Hirschi, T. (1990). *A general theory of crime*, Stanford University Press.
- Goudriaan, H., Wittebrood, K. and Nieuwbeerta, P. (2006). Neighbourhood characteristics and reporting crime effects of social cohesion, confidence in police effectiveness and socio-economic disadvantage. *British journal of criminology*, 46, 719-742.
- Goulbourne, H. and Solomos, J. (2003). Families, ethnicity and social capital. *Social Policy and Society*, 2, 329-338.

- Graham, E., Manley, D., Hiscock, R., Boyle, P. and Doherty, J. (2009). Mixing housing tenures: Is it good for social well-being? *Urban studies*, 46, 139-165.
- Gray, D. E. (2013). *Doing research in the real world*, Sage, London.
- Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. and Donaldson, L. (2013). Harnessing the cloud of patient experience: Using social media to detect poor quality healthcare. *BMJ Qual Saf*, 22, 251-255.
- Green, A., Preston, J. and Janmaat, G. (2006). *Education, equality and social cohesion-A comparative analysis*, Palgrave.
- Greene, J. A., Choudhry, N. K., Kilabuk, E. and Shrank, W. H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of General Internal Medicine*, 26, 287-292.
- Greenwood, S., Perrin, A. and Duggan, M. (2016). Social media update 2016. Pew Research Center. Available from: http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/11/10132827/PI_2016.11.11_Social-Media-Update_FINAL.pdf. [Accessed 21/08/2017].
- Groff, E. and Birks, D. (2008). Simulating crime prevention strategies: A look at the possibilities. *Policing: A journal of Policy and Practice*, 2, 175-184.
- Groff, E. R., Johnson, S. D. and Thornton, A. (2018). State of the Art in Agent-Based Modeling of Urban Crime: An Overview. *Journal of Quantitative Criminology*.
- Gruner, S. (2010). 'The others don't want...'. small-scale segregation: Hegemonic public discourses and racial boundaries in German neighbourhoods. *Journal of Ethnic and Migration Studies*, 36, 275-292.
- Gruzd, A., Wellman, B. and Takhteyev, Y. (2011). Imagining Twitter as an imagined community. *American Behavioral Scientist*, 55, 1294-1318.
- Guerry, A. M. (1833). *Essai sur la statistique morale de la France*, Crochard, Paris, FR.
- Guimarães, R. G., Rosa, R. L., De Gaetano, D., Rodríguez, D. Z. and Bressan, G. (2017). Age groups classification in social network using deep learning. *IEEE Access*, 5, 10805-10816.
- Habyarimana, J., Humphreys, M., Posner, D. N. and Weinstein, J. M. (2007). Why does ethnic diversity undermine public goods provision? *American Political Science Review*, 101, 709-725.
- Hackshaw, A. (2008). Small studies: strengths and limitations. *European Respiratory Journal*, 32, 1141-1143.
- Haezwindt, P. (2003). Social trends: The role of social capital. *Social Trends* No. 23.
- Haight, M., Quan-Haase, A. and Corbett, B. A. (2014). Revisiting the digital divide in Canada: The impact of demographic factors on access to the internet, level of online activity, and social networking site usage. *Information, Communication & Society*, 17, 503-519.
- Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17, 107-145.
- Hall, P. A. (1999). Social capital in Britain. *British Journal of Political Science*, 29, 417-461.

- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8, 23.
- Hämäläinen, J., Jauhiainen, S. and Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms*, 10, 105.
- Hamilton, W. L., Clark, K., Leskovec, J. and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.
- Hampton, K. N., Lee, C.-j. and Her, E. J. (2011). How new media affords network diversity: Direct and mediated access to social capital through participation in local social settings. *New Media & Society*, 13, 1031-1049.
- Han, L., Bandyopadhyay, S. and Bhattacharya, S. (2013). Determinants of violent and property crimes in England and Wales: A panel data analysis. *Applied Economics*, 45, 4820-4830.
- Haq, J. M. (2006). The borders and boundaries of community: social cohesion and responses to domestic and racial violence. PhD Thesis, Newcastle University, Newcastle uponTyre.
- Hardyns, W. and Pauwels, L. (2009). The geography of social cohesion and crime at the municipality level. *Contemporary Issues in the Empirical Study of Crime*, 157-178.
- Hariche, A. C., Loiseau, E. and Mac Erlaine, R. (2011). Web-enabled social cohesion: Harnessing participation. *Proceedings of the International Conference on Social Cohesion and Development*, 1-32.
- Haro-de-Rosario, A., Sáez-Martín, A. and del Carmen Caba-Pérez, M. (2016). Using social media to enhance citizen engagement with local government: Twitter or Facebook? *New Media & Society*, 1461444816645652.
- Harper, R. (2001). Social capital: A review of the literature. Social Analysis and Reporting Division, Office for National Statistics.
- Harries, K. A. (1999). Mapping crime: Principle and practice. Washington, DC: National Institute of Justice (NCJ 178919).
- Harris, K. and Flouch, H. (2010). The Online Neighbourhood Networks Study: A study of the social impact of citizen-run online neighbourhood networks and the implications for local authorities, Section 1: Social capital and cohesion. The Networked Neighbourhoods Group. London.
- Harris, K. and Flouch, H. (2012). Online neighbourhood networks in low income areas. Networked Neighbourhoods. <http://networkedneighbourhoods.com/wp-content/uploads/2012/10/OnlineNbhdNetsLowIncomeAreas20122.pdf>. [Access 15/02/2017].
- Harris, K. and McCabe, A. (2017). Community Action and Social Media. TRSC Working paper, University of Birmingham.
- Harry, K. (2016). Facebook Page Types Explained. Available from: <http://gosmallbiz.com/facebook-business-page-types-explained/>. [Accessed 18/10/2017].
- Hartle, F., Parker, M. and Wydra, C. (2014). The digital case file: The future of fighting crime with big data. *Issues in Information Systems*, 15.

- Hartnagel, T. F. (1979). The perception and fear of crime: Implications for neighborhood cohesion, social activity, and community affect. *Social Forces*, 58, 176-193.
- Hastad, D. N., Segrave, J. O., Pangrazi, R. and Peterson, G. (1984). Youth sport participation and deviant behavior. *Sociology of Sport Journal*, 1, 366-373.
- Hattingh, M. (2015). The use of Facebook by a Community Policing Forum to combat crime. *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists*, 19.
- Hayden, C. and Jenkins, C. (2014). 'Troubled families' programme in England: 'Wicked problems' and policy-based evidence. *Policy Studies*, 35, 631-649.
- Healy, T. (2002). The measurement of social capital at international level. *Social Capital: The Challenge of International Measurement Series of the Organisation for Economic Co-operation and Development (OECD)*. Paris: OECD. URL: <http://www.oecd.org/dataoecd/1/60/2380281.pdf>.
- Healy, T. and Côté, S. (2001). *The Well-being of nations: The role of human and social capital*. Education and Skills, ERIC.
- Hegerty, S. W. (2017). Crime, housing tenure, and economic deprivation: Evidence from Milwaukee, Wisconsin. *Journal of Urban Affairs*, 39, 1103-1121.
- Heidensohn, F. and Gelsthorpe, L. (2012). *Gender and crime: Oxford handbook of criminology*, Oxford University Press.
- Helliwell, J. F. and Putnam, R. D. (1999). Education and social capital. National Bureau of Economic Research.
- Hemalatha, I., Varma, D. G. S. and Govardhan, A. (2013). Sentiment analysis tool using machine learning algorithms. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2, 105-109.
- Heppenstall, A. J., Crooks, A. T., See, L. M. and Batty, M. (2011). *Agent-based models of geographical systems*, Springer Science & Business Media.
- Heverin, T. and Zach, L. (2010). Twitter for city police department information sharing. *Proceedings of the American Society for Information Science and Technology*, 47, 1-7.
- Hickman, M., Crowley, H. and Mai, N. (2008). Immigration and social cohesion in the UK. York Joseph Rowntree Foundation.
- Hicks, M. (2010). What is the difference between Facebook Page and Group? Available from: <https://www.facebook.com/notes/facebook/facebook-tips-whats-the-difference-between-a-facebook-page-and-group/324706977130/> [Accessed 17/08/2017].
- Higgins, S. (2015). *Limitations to seasonal weather prediction and crop forecasting due to nonlinearity and model inadequacy*. London School of Economics and Political Science (University of London).
- Hill, L. L., Crosier, S. J., Smith, T. R. and Goodchild, M. (2001). A content standard for computational models. *D-Lib Magazine*, 7, 1082-9873.

- Hill, T. C., Walsh, K. A., Harris, J. A. and Moffett, B. F. (2003). Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology*, 43, 1-11.
- Hino, K., Uesugi, M. and Asami, Y. (2016). Official crime rates and residents' sense of security across neighborhoods in Tokyo, Japan. *Urban Affairs Review*, 1078087416667619.
- Hipp, J. R. (2016). Collective efficacy: How is it conceptualized, how is it measured, and does it really matter for understanding perceived neighborhood crime and disorder? *Journal of Criminal Justice*, 46, 32-44.
- Hipp, J. R. and Wo, J. C. (2015). Collective efficacy and crime. *International Encyclopedia of the Social & Behavioral Sciences*, 4, 169-173.
- Hirschfield, A., Birkin, M., Brunson, C., Malleon, N. and Newton, A. (2013). How places influence crime: The impact of surrounding areas on neighbourhood burglary rates in a British city. *Urban Studies*, 0042098013492232.
- Hirschfield, A. and Bowers, K. J. (1997). The effect of social cohesion on levels of recorded crime in disadvantaged areas. *Urban Studies*, 34, 1275-1295.
- Hirschi, T. and Stark, R. (1969). Hellfire and delinquency. *Social Problems*, 17, 202-213.
- Hoffelmeyer, K. (2016). How Twitter, Facebook and Instagram measure engagement. Available from: <http://mediashift.org/2016/06/how-twitter-facebook-instagram-measure-engagement/> [Accessed 09/11/2017].
- Hoffmann, J. P. (2010). *Linear regression analysis: Applications and assumptions*, Brigham Young University, Provo, Utah, USA.
- Holbrook, E., Kaur, G., Bond, J., Imbriani, J., Nsoesie, E. and Grant, C. (2016). Tweet Geolocation Error Estimation. *International Conference on GIScience Short Paper Proceedings*, 1, 130-133.
- Hollander, J. B., Graves, E., Renski, H., Foster-Karim, C., Wiley, A. and Das, D. (2016). *A short history of social media sentiment analysis*, Springer.
- Home Office (2017). Police Recorded Crime Community Safety Partnership Open Data Tables. Available from: <https://www.gov.uk/government/statistics/police-recorded-crime-open-data-tables>. [Accessed 13/07/2018].
- Homel, R. (2006). Can police prevent crime? Bryett & Lewis. *Unpeeling tradition: Contemporary Policing*. Sydney: Macmillan Australia, 18-22.
- Honey, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, 1-10.
- Hooghe, M. and De Vroome, T. (2015). The perception of ethnic diversity and anti-immigrant sentiments: A multilevel analysis of local communities in Belgium. *Ethnic and Racial Studies*, 38, 38-56.
- Hooghe, M. and De Vroome, T. (2016). The relation between ethnic diversity and fear of crime: An analysis of police records and survey data in Belgian communities. *International Journal of Intercultural Relations*, 50, 66-75.

- Hope, T. (1988). Support for neighbourhood watch: a British Crime Survey analysis. *Communities and Crime Reduction*, 146-161.
- Hope, T. (1995). Community crime prevention. *Crime and Justice*, 21-89.
- Hope, T. (2001). Community crime prevention in Britain: A strategic overview. *Criminology and Criminal Justice*, 1, 421-439.
- Hovel, A. (2014). *Crime, Income Inequality, and Density at the Neighborhood Level*. Theses Saint John's University, Minnesota, USA.
- Howard, A. (2012). Connecting with communities: How local government is using social media to engage with citizens. Available from: http://www.governanceinstitute.edu.au/magma/media/upload/ckeditor/files/1353548699_Connecting_Communities_ANZSIG-ACELG_August_2012.pdf. [Accessed 09/07/2016].
- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168-177.
- Huberman, B. A., Romero, D. M. and Wu, F. (2008). Social networks that matter: Twitter under the microscope. Available at SSRN 1313405.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193-218.
- Huck, J., Whyatt, D. and Coulton, P. (2015). Visualizing patterns in spatially ambiguous point data. *Journal of Spatial Information Science*, 2015, 47-66.
- Hudson, M., Phillips, J., Ray, K. and Barnes, H. (2007). Social cohesion in diverse communities. Joseph Rowntree Foundation, York.
- Hugh, F. and Kevin, H. (2011). Councils and Online Neighbourhood Networks: Report of the second Networked Neighbourhoods survey of council officers and elected members. Available from: <http://networkedneighbourhoods.com/wp-content/uploads/2011/11/2011-Online-Nhood-Networks-final.pdf>. [Accessed 04/07/2015].
- Humphreys, L. (2010). Mobile social networks and urban public space. *New Media & Society*, 12, 763-778.
- Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 271-293.
- Hussain, K. Z., Durairaj, M. and Farzana, G. R. J. (2012). Application of data mining techniques for analyzing violent criminal behavior by simulation model. *IJCSITS*, 2, 1-5.
- Ibrahim, S., Hamisu, I. and Lawal, U. (2015). Spatial pattern of tuberculosis prevalence in Nigeria: A comparative analysis of spatial autocorrelation indices. *American Journal of Geographic Information System*, 4, 87-94.
- Imandoust, S. (2011). Relationship between education and social capital. *International Journal of Humanities and Social Science*, 1, 52-57.

- Irani, J., Pise, N. and Phatak, M. (2016). Clustering techniques and the similarity measures used in clustering: A survey. *International Journal of Computer Applications*, 134.
- Jackson, D. (2014). All Social Media Metrics that Matter. Available from: <http://sproutsocial.com/insights/social-media-metrics-that-matter/>. [Accessed 10/04/2017]
- Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31, 264-323.
- Jalonen, H. (2017). "A good bell is heard from far, a bad one still further": A Socio-demography of disclosing negative emotions in social media. *The Journal of Social Media in Society*, 6, 69-109.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*, New York, Springer.
- James, N. (2017). A Review on Local Government Social Media Usage. Available from: <https://www.bdo.co.uk/en-gb/insights/industries/public-sector/a-review-of-social-media-usage>. [Accessed 04/01/2018].
- Japac, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, 79, 839-880.
- Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56-65.
- Jennings, J. M., Milam, A. J., Greiner, A., Furr-Holden, C. D. M., Curriero, F. C. and Thornton, R. J. (2014). Neighborhood alcohol outlets and the association with violent crime in one Mid-Atlantic city: the implications for zoning policy. *Journal of Urban Health*, 91, 62-71.
- Jim, J., Ngo Mitchell, F. and Kent, D. R. (2006). Community-oriented policing in a retail shopping center. *Policing: An International Journal of Police Strategies & Management*, 29, 146-157.
- Jinyuan, L., Wan, T., Guanqin, C., Yin, L. and Changyong, F. (2016). Correlation and agreement: Overview and clarification of competing concepts and measures. *Shanghai Archives of Psychiatry*, 28, 115.
- Jivraj, S. and Simpson, L. (2015). How has ethnic diversity grown? *Ethnic Identity and Inequalities in Britain: The Dynamics of Diversity*, 19.
- Johnston, R. and Matthews, J. S. (2004). Social capital, age, and participation. *Paper, Youth Participation Workshop of annual meeting of Canadian Political Science Association. Winnipeg.*
- Johnston, T. (2016). Synthesizing structure and agency: A developmental framework of Bourdieu's constructivist structuralism theory. *Journal of Theoretical & Philosophical Criminology*, 8, 1.
- Jong, J. (2011). Predicting rating with sentiment analysis. Stanford Univ., Stanford, CA.

- Jonsen, K., Maznevski, M. L. and Schneider, S. C. (2011). Special Review Article: Diversity and its not so diverse literature: An international perspective. *International Journal of Cross Cultural Management*, 11, 35-62.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113, 363-375.
- Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers & Education*, 58, 162-171.
- Junco, R. (2015). Student class standing, Facebook use, and academic performance. *Journal of Applied Developmental Psychology*, 36, 18-29.
- Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*.
- Jurek, A., Mulvenna, M. D. and Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4, 1.
- Jussila, J., Huhtamäki, J., Kärkkäinen, H. and Still, K. (2013). Information visualization of Twitter data for co-organizing conferences. *Proceedings of International Conference on Making Sense of Converging Media*, 139.
- Kaigo, M. and Okura, S. (2016). Exploring fluctuations in citizen engagement on a local government Facebook page in Japan. *Telematics and Informatics*, 33, 584-595.
- Kalonia, C., Kumru, O. S., Kim, J. H., Middaugh, C. R. and Volkin, D. B. (2013). Radar chart array analysis to visualize effects of formulation variables on IgG1 particle formation as measured by multiple analytical techniques. *Journal of Pharmaceutical Sciences*, 102, 4256-4267.
- Kalton, G. (1968). Standardization: A technique to control for extraneous variables. *Applied Statistics*, 118-136.
- Kamata, A. (1998). One-parameter hierarchical generalized linear logistic model: An application of HGLM to IRT. *annual meeting of the American Educational Research Association, San Diego, CA*.
- Kanazawa, S. (2003). Why productivity fades with age: The crime–genius connection. *Journal of Research in Personality*, 37, 257-272.
- Kanazawa, S. and Savage, J. (2009). Why nobody seems to know what exactly social capital is. *Journal of Social, Evolutionary, and Cultural Psychology*, 3, 118.
- Kang, J. H. (2015). Participation in the community social control, the neighborhood watch groups individual-and neighborhood-related factors. *Crime & Delinquency*, 61, 188-212.
- Karyda, M. (2015). *The effect of crime in the community on becoming not in education, employment or training (NEET) at 18-19 years in England*. UCL Institute of Education.
- Kassambara, A. and Mundt, F. (2017). Package ‘factoextra,’ Version 1.0.5. *R topics documented*, 75.
- Kathryn, H. (2016). Associations between police recorded ethnic background and being sentenced to prison in England and Wales. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/479874/analysis-of-ethnicity-and-custodial-sentences.pdf. [02/03/2018].

- Katz, J. E. and Rice, R. E. (2002). *Social consequences of Internet use: Access, involvement, and interaction*, MIT Press, Cambridge, MA.
- Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 68-125.
- Kaufman, L. and Rousseeuw, P. J. (2005). *Finding groups in data*. Hoboken, John Wiley & Sons, Inc, NJ.
- Kawachi, I. and Berkman, L. (2000). Social cohesion, social capital, and health. *Social Epidemiology*, 174-190.
- Kawash, J. (2014). *Online social media analysis and visualization*, Switzerland, Springer.
- Keene, D., Bader, M. and Ailshire, J. (2013). Length of residence and social integration: The contingent effects of neighborhood poverty. *Health & Place*, 21, 171-178.
- Kehm, R., Davey, C. S. and Nanney, M. S. (2015). The role of family and community involvement in the development and implementation of school nutrition and physical activity policy. *Journal of School Health*, 85, 90-99.
- Kelling, G. L. and Stewart, J. K. (1990). Neighborhoods and police: The maintenance of civil authority. *Criminal Law Forum*, 1, 459-476.
- Kelly, M. (2000). Inequality and crime. *Review of Economics and Statistics*, 82, 530-539.
- Kennedy, B. P., Kawachi, I., Prothrow-Stith, D., Lochner, K. and Gupta, V. (1998). Social capital, income inequality, and firearm violent crime. *Social Science & Medicine*, 47, 7-17.
- Khan, A. Z., Atique, M. and Thakare, V. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCE)*, 89.
- Kharde, V. and Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *arXiv preprint arXiv:1601.06971*.
- Kim, C. and Yang, S.-U. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review*, 43, 441-449.
- Kim, H., Grogan-Kaylor, A., Han, Y., Maurizi, L. and Delva, J. (2013). The association of neighborhood characteristics and domestic violence in Santiago, Chile. *Journal of Urban Health*, 90, 41-55.
- Kindler, M., Ratcheva, V. and Piechowska, M. (2014). Social networks, social capital and migrant integration at local level-European literature review. *King Desk Research Paper*.
- Kiritchenko, S., Zhu, X. and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50, 723-762.
- Kleck, G., Tark, J. and Bellows, J. J. (2006). What methods are most frequently used in research in criminology and criminal justice? *Journal of Criminal Justice*, 34, 147-152.

- Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in K-means clustering. *International Journal*, 1, 90-95.
- Kongmuang, C. (2006). *Modelling crime: A spatial microsimulation approach*. PhD thesis, The University of Leeds.
- Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155-163.
- Koppel, M. and Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22, 100-109.
- Korhonen, A. (2009). Automatic lexical classification--balancing between machine learning and linguistics. *PACLIC*, 19-28.
- Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V. and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70, 543.
- Kraak, M.-J. (1999). Visualising spatial distributions. *Chapter*, 11, 157-173.
- Kristensen, J. B., Albrechtsen, T., Dahlgaard, E., Jensen, M., Skovrind, M. and Bornakke, T. (2017). Parsimonious data: How a single Facebook like predicts voting behaviour in multiparty systems. *arXiv preprint arXiv:1704.01143*.
- Kristjánsson, Á. L. (2007). On social equality and perceptions of insecurity: A comparison study between two European countries. *European Journal of Criminology*, 4, 59-86.
- Kruger, D. J., Hutchison, P., Monroe, M. G., Reischl, T. and Morrel-Samuels, S. (2007). Assault injury rates, social capital, and fear of neighborhood crime. *Journal of Community Psychology*, 35, 483-498.
- Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583-621.
- Kuang, L. and Zhang, L. (2017). A scheduling algorithm based on Clara clustering. *AIP Conference Proceedings*, 1864, 0200161-7.
- Kubrin, C. E. and Weitzer, R. (2003). New directions in social disorganization theory. *Journal of Research in Crime and Delinquency*, 40, 374-402.
- Kumar, S., Morstatter, F. and Liu, H. (2014). *Twitter data analytics*, New York, Springer.
- Kwon, S. J., Park, E. and Kim, K. J. (2014). What drives successful social networking services? A comparative analysis of user acceptance of Facebook and Twitter. *The Social Science Journal*, 51, 534-544.
- Labatut, V., Dugué, N. and Perez, A. (2014). Identifying the community roles of social capitalists in the twitter network. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 371-374.
- LaGrange, T. C. (1999). The impact of neighborhoods, schools, and malls on the spatial distribution of property damage. *Journal of Research in Crime and Delinquency*, 36, 393-422.

- Lai, Y.-L., Ren, L. and Greenleaf, R. (2016). Residence-based fear of crime a routine activities approach. *International Journal of Offender Therapy and Comparative Criminology*, 0306624X15625054.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.
- Landis, R. S. (2014). Standardized regression coefficients. *Wiley StatsRef: Statistics Reference Online*.
- Lane, J. and Meeker, J. W. (2000). Subcultural diversity and the fear of crime and gangs. *Crime & Delinquency*, 46, 497-521.
- Lane, P. W. and Nelder, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, 613-621.
- Laney, D. (2013). 3D Data management: Controlling data volume, velocity and variety. Meta Group. Available from: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. [Accessed 08/06/2015].
- Lanfear, C. (2016). Faceboob with RFacebook and SocialMediaLab. Available from: https://rstudio-pubs-static.s3.amazonaws.com/164365_68da649570ad4079ac5e80099698c246.html.
- Lansley, G. and Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85-96.
- Lansley, G., Wei, Y. and Rains, T. (2015). Creating an Output Area Classification of Cultural and Ethnic Heritage to Assist the Planning of Ethnic Origin Foods in Supermarkets in England and Wales.
- Lathia, N., Quercia, D. and Crowcroft, J. (2012). The hidden image of the city: sensing community well-being from urban mobility. *International Conference on Pervasive Computing*, 91-98.
- Laurence, J. (2011). The effect of ethnic diversity and community disadvantage on social cohesion: A multi-level analysis of social capital and interethnic relations in UK communities. *European Sociological Review*, 27, 70-89.
- Laurence, J. and Bentley, L. (2016). Does ethnic diversity have a negative effect on attitudes towards the community? A longitudinal analysis of the causal claims within the ethnic diversity and social cohesion debate. *European Sociological Review*, 32, 54-67.
- Layton, R., Watters, P. and Dazeley, R. (2013). Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19, 95-120.
- Lee, B. (1992). Colonialization and community: Implications for First Nations development. *Community Development Journal*, 27, 211-219.
- Lee, G. and Kwak, Y. H. (2012). An open government maturity model for social media-based public engagement. *Government Information Quarterly*, 29, 492-503.
- Lee, M. R. (2000). Community cohesion and violent predatory victimization: A theoretical extension and cross-national test of opportunity theory. *Social Forces*, 79, 683-706.

- Leeds City Council (2013). Leeds Population. Available from: <http://www.leeds.gov.uk/council/Pages/Leeds-population.aspx>. [Accessed 26/06/2016].
- Leeds City Council (2014). Equality and Diversity Annual Update. Available from: <http://www.leeds.gov.uk/docs/Equality%20and%20diversity%20update%202014.pdf>. [Accessed 24/01/2017].
- Leeds City Council (2016). Family First. Retrieved from: <https://www.leeds.gov.uk/docs/16%20-%20Families%20First%20-%20March%202016.pdf>. [Accessed 02/03/2018].
- Leeds City Council (2017). Equality and Diversity Annual Update. Available from: <https://www.leeds.gov.uk/docs/Equality%20and%20diversity%20update%202017.pdf>. [Accessed 10/04/2018].
- Leetaru, K., Wang, S., Cao, G., Padmanabhan, A. and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18.
- Lerman, K., Arora, M., Gallegos, L., Kumaraguru, P. and Garcia, D. (2016). Emotions, Demographics and Sociability in Twitter Interactions. *Tenth International AAAI Conference on Web and Social Media*.
- Letki, N. (2008). Does diversity erode social cohesion? Social capital and race in British neighbourhoods. *Political Studies*, 56, 99-126.
- Lev-On, A. and Steinfeld, N. (2015). Local engagement online: Municipal Facebook pages as hubs of interaction. *Government Information Quarterly*, 32, 299-307.
- Leventhal, B. (2016). *Geodemographics for marketers: Using location analysis for research and marketing*, Kogan Page Publishers.
- Leventhal, T. and Brooks-Gunn, J. (2000). The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin*, 126, 309.
- Levine, N. (2004). CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0). *Houston (TX): Ned Levine & Associates/Washington, DC: National Institute of Justice*.
- Leviten-Reid, C. and Matthew, R. A. (2018). Housing Tenure and Neighbourhood Social Capital. *Housing, Theory and Society*, 35, 300-328.
- Levitt, S. D. (1999). The limited role of changing age structure in explaining aggregated crime rates. *Criminology*, 37, 581-598.
- Lewis, S. C., Zamith, R. and Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, 57, 34-52.
- Lieberson, S. (1985). *Making it count: The improvement of social research and theory*, USA, University of California Press.
- Linoff, G. S. and Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*, John Wiley & Sons.
- Liu, B. (2010a). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2, 627-666.

- Liu, B. (2010b). Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems*, 25, 76-80.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5, 1-167.
- Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010). Understanding of internal clustering validation measures. *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 911-916.
- Livingston, M., Kearns, A. and Bannister, J. (2014). Neighbourhood structures and crime: the influence of tenure mix and other structural factors upon local crime rates. *Housing Studies*, 29, 1-25.
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28, 129-137.
- Local Government Association (2002). *Guidance on community cohesion*, Local Government Association, London.
- Loeber, R., Menting, B., Lynam, D. R., Moffitt, T. E., Stouthamer-Loeber, M., Stallings, R., Farrington, D. P. and Pardini, D. (2012). Findings from the Pittsburgh Youth Study: Cognitive impulsivity and intelligence as predictors of the age-crime curve. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51, 1136-1149.
- Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63, 509-525.
- Longley, P. A., Adnan, M. and Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47, 465-484.
- Lu, B., Harris, P., Charlton, M. and Brunson, C. (2014). The GWmodel R package: Further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, 17, 85-101.
- Ludwig, A. and Marshall, M. (2015). Using crime data in academic research: Issues of comparability and integrity. *Records Management Journal*, 25, 228-247.
- Luini, L. P., Cardellicchio, D., Felletti, F. and Marucci, F. S. (2015). Socio-Spatial Intelligence: social media and spatial cognition for territorial behavioral analysis. *Cognitive Processing*, 16, 299-303.
- Machin, S., Marie, O. and Vujić, S. (2011). The crime reducing effect of education. *The Economic Journal*, 121, 463-484.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1, 281-297.
- MacQueen, K. M., McLellan, E., Metzger, D. S., Kegeles, S., Strauss, R. P., Scotti, R., Blanchard, L. and Trotter, R. T. (2001). What is community? An evidence-based definition for participatory public health. *American Journal of Public Health*, 91, 1929-1938.
- Macskassy, S. A. (2012). On the Study of Social Interactions in Twitter. *ICWSM*.

- Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0. 6. .
- Maginnis, R. (1997). Single-Parent Families Cause Juvenile Crime. From Juvenile Crime: Opposing Viewpoints, P 62-66, 1997, AE Sadler, ed.--See NCJ-167319.
- Magno, L. (2016). Using Facebook Metrics to Measure Student Engagement in Moodle. *IJODEL*, Vol. 2, No. 2.
- Magurran, A. E. (2004). *Measuring biological diversity*, Oxford, Blackwell.
- Mahmud, J., Nichols, J. and Drews, C. (2012). Where is this Tweet from? Inferring home locations of Twitter users. *ICWSM*, 12, 511-514.
- Mahmud, J., Nichols, J. and Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5, 47.
- Maiese, M. (2003). Social Structural Change. [Online]: <http://www.beyondintractability.org/essay/social-structural-changes>. [Accessed 21/01/2017].
- Malczewski, J. and Poetz, A. (2005). Residential burglaries and neighborhood socioeconomic context in London, Ontario: Global and local regression analysis. *The Professional Geographer*, 57, 516-529.
- Malhotra, R. K. and Indrayan, A. (2010). A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian Journal of Ophthalmology*, 58, 519.
- Malleson, N. (2010). *Agent Based Modelling of Buglary*. Phd, University of Leeds.
- Malleson, N. and Andresen, M. A. (2015). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42, 112-121.
- Malleson, N. and Andresen, M. A. (2016). Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*, 46, 52-63.
- Malleson, N. and Birkin, M. (2012). Analysis of crime patterns through the integration of an agent-based model and a population microsimulation. *Computers, Environment and Urban Systems*, 36, 551-561.
- Malleson, N., Heppenstall, A. and See, L. (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, environment and urban systems*, 34, 236-250.
- Malleson, N., See, L., Evans, A. and Heppenstall, A. 2014. Optimising an agent-based model to explore the behaviour of simulated burglars. *Theories and Simulations of Complex Social Systems*. Springer.
- Mandarano, L., Meenar, M. and Steins, C. (2010). Building social capital in the digital age of civic engagement. *Journal of Planning Literature*, 25, 123-135.
- Marge, M., Banerjee, S. and Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 5270-5273.

- Markowitz, F. E., Bellair, P. E., Liska, A. E. and Liu, J. (2001). Extending social disorganization theory: modeling the relationships between cohesion, disorder, and fear. *Criminology*, 39, 293-319.
- Markson, L., Woodhams, J. and Bond, J. W. (2010). Linking serial residential burglary: Comparing the utility of modus operandi behaviours, geographical proximity, and temporal proximity. *Journal of Investigative Psychology and Offender Profiling*, 7, 91-107.
- Martin, D. (2002). Spatial patterns in residential burglary Assessing the effect of neighborhood social capital. *Journal of Contemporary Criminal Justice*, 18, 132-146.
- Martin, K. (2014). Can regression model with small r-squared be useful? Available from: <http://www.theanalysisfactor.com/small-r-squared/>. [Accessed 18/12/2016].
- Marujo, L., Ling, W., Trancoso, I., Dyer, C., Black, A. W., Gershman, A., de Matos, D. M., da Silva Neto, J. P. and Carbonell, J. G. (2015). Automatic Keyword Extraction on Twitter. *ACL* (2), 637-643.
- Masike, L. G. and Mofokeng, J. (2014). Safety of students in residences of a university in the Tshwane Metropolitan. *Acta Criminologica: Southern African Journal of Criminology*, 2014, 64-80.
- Mathers, N., Fox, N. J. and Hunn, A. (1998). *Surveys and questionnaires*, NHS Executive, Trent.
- Matthews, P. (2015). Neighbourhood belonging, social class and social media—Providing ladders to the cloud. *Housing Studies*, 30, 22-39.
- Matthews, P. (2016). Social media, community development and social capital. *Community Development Journal*, 51, 419-435.
- Maurer, C. and Wiegmann, R. (2011). Effectiveness of advertising on social network sites: a case study on Facebook. *ENTER*, 485-498.
- Mburu, L. W. and Bakillah, M. (2016). Modeling spatial interactions between areas to assess the burglary risk. *ISPRS International Journal of Geo-Information*, 5, 47.
- McCabe, A., Gilchrist, A., Harris, K., Afridi, A. and Kyprianou, P. (2013). Making the links: Poverty, ethnicity and social networks. Perth, UK: University of Birmingham, Joseph Rowntree Foundation.
- McCall, P. L., Land, K. C., Dollar, C. B. and Parker, K. F. (2013). The age structure-crime rate relationship: Solving a long-standing puzzle. *Journal of Quantitative Criminology*, 29, 167-190.
- McClain, C. R. (2017). Practices and promises of Facebook for science outreach: Becoming a “Nerd of Trust”. *PLoS biology*, 15, e2002020.
- McCool, S. F. and Martin, S. R. (1994). Community attachment and attitudes toward tourism development. *Journal of Travel research*, 32, 29-34.
- McDonald, J. H. (2009). *Handbook of biological statistics*, Baltimore, Sparky House.
- McGarrell, E. F., Corsaro, N., Hipple, N. K. and Bynum, T. S. (2010). Project safe neighborhoods and violent crime trends in US cities: Assessing violent crime impact. *Journal of Quantitative Criminology*, 26, 165-190.
- McGarrell, E. F., Giacomazzi, A. L. and Thurman, Q. C. (1997). Neighborhood disorder, integration, and the fear of crime. *Justice Quarterly*, 14, 479-500.

- McIntosh, R. P. (1967). An index of diversity and the relation of certain concepts to diversity. *Ecology*, 48, 392-404.
- McMillan, D. W. and Chavis, D. M. (1986). Sense of community: A definition and theory. *Journal of Community Psychology*, 14, 6-23.
- McNab, C. (2009). What social media offers to health professionals and citizens. *Bulletin of the World Health Organization*, 87, 566-566.
- McVie, S. (2005). Patterns of deviance underlying the age-crime curve: The long term evidence. *British Society of Criminology e-journal*, 7, 1-15.
- Medhat, W., Hassan, A. and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5, 1093-1113.
- Meer, T. v. d. and Tolsma, J. (2014). Ethnic diversity and its effects on social cohesion. *Annual Review of Sociology*, 40, 459-478.
- Meera, A. K. and Jayakumar, M. D. (1995). Determinants of crime in a developing country: a regression model. *Applied Economics*, 27, 455-460.
- Mellgren, C. (2011). What's neighbourhood got to do with it?: the influence of neighbourhood context on crime and reactions to crime. Malmö University, Sweden.
- Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55, 31-42.
- Migheli, M. (2007). Trust, Gender and Social Capital: Experimental Evidence from Three Western European Countries. Available at SSRN 976380.
- Miley, F. and Read, A. (2012). Using word clouds to develop proactive learners. *Journal of the Scholarship of Teaching and Learning*, 11, 91-110.
- Miller, D. (2016). *Social Media in an English Village*, London, UCL Press.
- Milligan, G. (1996). Clustering validation: Results and implications for applied analyses. In Dan Vickers and Phil Rees 2007 Creating the UK National Statistics for Output Area Classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 2, 379-403.
- Mirkin, B. (2011). Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 252-260.
- Mirkin, B. (2012). *Clustering: A data recovery approach*, Chapman & Hall, CRC Press.
- Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S. and Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS One*, 8, e64417.
- Mohammad, S. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. *WASSA@ NAACL-HLT*, 174-179.
- Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29, 436-465.
- Mohit, K., Singh, B. R. and Gurpreet, S. (2016). K-means demographic based crowd aware movie recommendation system. *Indian Journal of Science and Technology*, 9, 23, 1-4.

- Moore, M. D. and Recker, N. L. (2013). Social capital, type of crime, and social control. *Crime & Delinquency*, 0011128713510082.
- Moore, M. H. and Trojanowicz, R. C. (1988). *Policing and the Fear of Crime*, US Department of Justice, National Institute of Justice Washington, DC.
- Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A. and Hoving, C. (2013). A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research*, 15, 4: e85.
- Morenoff, J. D., Sampson, R. J. and Raudenbush, S. W. (2001). Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. *Criminology*, 39, 517-558.
- Moretti, E. (2005). Does education reduce participation in criminal activities. *symposium on "The Social Costs of Inadequate Education"*. Columbia University Teachers College.
- Moriarty, L. J. and Williams, J. E. (1996). Examining the relationship between routine activities theory and social disorganization: An analysis of property crime victimization. *American Journal of Criminal Justice*, 21, 43-59.
- Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E. and Prati, D. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution*, 4, 3514-3524.
- Morstatter, F., Pfeffer, J., Liu, H. and Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.
- Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M. and Liu, H. (2016). A new approach to bot detection: Striking the balance between precision and recall. *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, 533-540.
- Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, 40, 4241-4251.
- Mozetič, I., Grčar, M. and Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS One*, 11, e0155036.
- Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24, 69-71.
- Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D. W., Slotow, R. and Hamer, M. (2014). Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution*, 5, 320-328.
- Mullaney, M. J. (2012). Optimizing social media in humanitarian crisis responses. *The Macalester Review*, 2, 3.
- Murray, A. T., McGuffog, I., Western, J. S. and Mullins, P. (2001). Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment. *British Journal of Criminology*, 41, 309-329.

- Murthy, D., Gross, A. and Pensavalle, A. (2016). Urban social media demographics: An exploration of Twitter use in major American cities. *Journal of Computer-Mediated Communication*, 21, 33-49.
- Nagy, A. and Stamberger, J. (2012). Crowd sentiment detection during disasters and crises. *Proceedings of the 9th International ISCRAM Conference*, 1-9.
- Nahapiet, J. and Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, 23, 242-266.
- Nandi, A. and Platt, L. (2014). Britishness and identity assimilation among the UK's minority and majority ethnic groups. ISER Working Paper Series.
- Naseri, S. (2017). Online Social Network Sites and Social Capital: A Case of Facebook. *International Journal of Applied Sociology*, 7, 13-19.
- Nathans, L. L., Oswald, F. L. and Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, 17, 9, 1-19.
- Neuman, W. L. and Robson, K. (2004). *Basics of social research: Qualitative and quantitative approaches*, Boston, Pearson.
- Neviarouskaya, A., Prendinger, H. and Ishizuka, M. (2011). SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2, 22-36.
- Ng, R. T. and Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. *Proceedings of VLDB*, 144-155.
- Nie, N. H. (2001). Sociability, interpersonal relations, and the internet reconciling conflicting findings. *American Behavioral Scientist*, 45, 420-435.
- Nisic, N. and Petermann, S. (2013). New city= new friends? The restructuring of social resources after relocation. *Comparative Population Studies*, 38, 1, 199-226.
- Norton, A. and de Haan, A. (2012). Social cohesion: Theoretical debates and practical applications with respect to jobs. World Bank Development Report. Available from: https://openknowledge.worldbank.org/bitstream/handle/10986/12147/WDR2013_bp_Social_Cohesion_Norton.pdf. [Accessed 21/08/2012].
- O'Brien, L. (1992). *Introducing quantitative geography: measurement, methods, and generalised linear models*, New York, Taylor & Francis.
- O'Connor, A. (2004). The sociology of youth subcultures. *Peace Review*, 16, 409-414.
- O'Connor, B., Balasubramanyan, R., Routledge, B. R. and Smith, N. A. (2010). From Tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11, 1.2.
- O'Donoghue, C. (2014). *Handbook of microsimulation modelling*, Emerald Group Publishing, pp.1 - 21.
- O'Sullivan, D. and Unwin, D. J. (2010). *Geographic Information Analysis*, John Wiley & Sons, New Jersey.
- O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D. and O'Brien, J. P. (2016). A fast and objective multidimensional kernel density estimation method: fastKDE. *Computational Statistics & Data Analysis*, 101, 148-160.

- Odgers, C. L., Donley, S., Caspi, A., Bates, C. J. and Moffitt, T. E. (2015). Living alongside more affluent neighbors predicts greater involvement in antisocial behavior among low-income boys. *Journal of Child Psychology and Psychiatry*, 56, 1055-1064.
- Oh, J. H. (2003). Assessing the social bonds of elderly neighbors: The roles of length of residence, crime victimization, and perceived disorder. *Sociological Inquiry*, 73, 490-510.
- Okayasu, I., Kawahara, Y. and Nogawa, H. (2010). The relationship between community sport clubs and social capital in Japan: A comparative study between the comprehensive community sport clubs and the traditional community sports clubs. *International Review for the Sociology of Sport*, 45, 163-186.
- ONS (2011). Census aggregate data. Available from: <https://census.ukdataservice.ac.uk/get-data/aggregate-data>. [Accessed 05/06/2015].
- ONS (2013). Statistics on Race and the Criminal Justice System. Ministry of Justice. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/269399/Race-and-cjs-2012.pdf. [Accessed 14/03/2015].
- ONS (2014a). Crime in England and Wales. Home Office, UK Government.
- ONS (2014b). Crime Statistics: Focus on Property Crime. Available from: <http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/crime-stats/crime-statistics/focus-on-property-crime--2013-14/index.html>. [Accessed 15/10/2016].
- ONS (2014c). Criminal justice system statistics quarterly. Available from: <https://www.gov.uk/government/statistics/criminal-justice-system-statistics-quarterly-december-2014>. [Accessed 19/06/2017].
- ONS (2014d). Statistics on Women and the Criminal Justice System. A Ministry of Justice publication, Home Office, UK Government. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/380090/women-cjs-2013.pdf. [Accessed 09/8/16].
- ONS (2015a). Inequalities in social capital by age and sex. Available from: http://www.ons.gov.uk/ons/dcp171766_410190.pdf. [Access 21/07/2015].
- ONS (2015b). Internet users and non users in the UK by geographical location. Available from: <https://www.gov.uk/government/statistics/internet-users-in-the-uk-2015>. [Accessed 12/07/2017].
- ONS (2015c). Statistics on Race and the Criminal Justice System. Ministry of Justice. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/480250/bulletin.pdf. [Accessed 14/09/2017].
- ONS (2016a). Crime Statistics: Focus on Property Crime. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/focusonpropertycrime/yearendingmarch2016#characteristics-associated-with-being-a-victim-of-property-crime>. [Accessed 04/01/2018].
- ONS (2016b). Social Media Usage by Age Groups. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/ho>

meinternetandsocialmediausage/datasets/internetaccesshouseholdsandindividualsreferencetables. [Accessed 09/11/2017].

ONS (2016c). Taking Part: Focus on Social Media. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/519678/Social_media_-_FINAL.pdf. [Accessed 04/01/2018].

ONS (2017a). Criminal justice system statistics quarterly: December 2017. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/criminenglandandwales/yearendingdecember2017>. [Accessed 17/05/2018].

ONS (2017b). Internet access – households and individuals.

ONS (2017c). Internet access – households and individuals. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2017>. [Access 25/02/2018].

ONS (2017d). Social Media Usage by Age Groups. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage>. [Accessed 13/04/2018].

Openshaw, S. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. *Spatistica Applications in the Spatial Sciences*, 127-144.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A*, 16, 17-31.

Openshaw, S. (1991). A view on the GIS crisis in geography, or, using GIS to put Humpty-Dumpty back together again. *Environment and Planning A*, 23, 621-628.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27, 286-306.

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16, 21-43.

Palanisamy, P., Yadav, V. and Elchuri, H. (2013). Serendio: Simple and Practical lexicon based approach to Sentiment Analysis. *proceedings of Second Joint Conference on Lexical and Computational Semantics*, 543-548.

Palguna, D. S., Joshi, V., Chakaravarthy, V. T., Kothari, R. and Subramaniam, L. V. (2015). Analysis of Sampling Algorithms for Twitter. *IJCAI*, 967-973.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135.

Paquette, H. (2013). Social Media as a Marketing Tool: A Literature Review. Paper series, University of Rhode Island.

Park, G. (2017). Screens are a game changer: How environments influence social capital in the digital era. *Cogent Social Sciences*, 3, 1372028.

- Park, H., Rodgers, S. and Stemmler, J. (2011). Health organizations' use of Facebook for health advertising and promotion. *Journal of Interactive Advertising*, 12, 62-77.
- Park, N., Kee, K. F. and Valenzuela, S. (2009). Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *CyberPsychology & Behavior*, 12, 729-733.
- Parker, S., Uprichard, E. and Burrows, R. (2007). Class places and place classes geodemographics and the spatialization of class. *Information, Communication & Society*, 10, 902-921.
- Parsons, A. (2013). Using social media to reach consumers: A content analysis of official Facebook pages. *Academy of Marketing Studies Journal*, 17, 27.
- Paulsen, D. J. and Robinson, M. B. (2004). *Spatial aspects of crime: Theory and practice*, Allyn & Bacon.
- Pauwels, L. and Hardyns, W. (2009). Measuring community (dis) organizational processes through key informant analysis. *European Journal of Criminology*, 6, 401-417.
- Pavan, K. K., Rao, A. A., Rao, A. and Sridhar, G. (2012). Robust seed selection algorithm for k-means type algorithms. *arXiv preprint arXiv:1202.1585*.
- Paynich, R. and Hill, B. (2010). *Fundamentals of crime mapping in Identifying high crime areas: ICCA 2013*, Jones & Bartlett Learning.
- Pearce, K. E. and Rice, R. E. (2017). Somewhat separate and unequal: digital divides, social networking sites, and capital-enhancing activities. *Social Media+ Society*, 3, 2056305117716272.
- Pearson, A. L., Breetzke, G. and Ivory, V. (2015). The effect of neighborhood recorded crime on fear: does neighborhood social context matter? *American Journal of Community Psychology*, 56, 170-179.
- Pedersen, T. and Kulkarni, A. (2006). Automatic cluster stopping with criterion functions and the gap statistic. *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations*, 276-279.
- Pénard, T. and Poussing, N. (2010). Internet use and social capital: The strength of virtual ties. *Journal of Economic Issues*, 44, 569-595.
- Perreault, W. D. and Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, 26, 135.
- Perrin, A. (2015). *Social Media Usage: 2005-2015: 65% of Adults Now Use Social Networking Sites--a Nearly Tenfold Jump in the Past Decade*. [Access 17/08/2017. http://www.pewinternet.org/files/2015/10/PI_2015-10-08_Social-Networking-Usage-2005-2015_FINAL.pdf, Pew Research Trust.
- Peters, K., Chen, Y., Kaplan, A. M., Ognibeni, B. and Pauwels, K. (2013). Social media metrics—A framework and guidelines for managing social media. *Journal of interactive marketing*, 27, 281-298.
- Peterson, M. and Robbins, B. (2008). Using the MMPI-A to predict recidivism in adjudicated minors. *Applied Psychology in Criminal Justice*, 4, 2, 172-181.

- Philip, L., Cottrill, C., Farrington, J., Williams, F. and Ashmore, F. (2017). The digital divide: Patterns, policy and scenarios for connecting the 'final few' in rural communities across Great Britain. *Journal of Rural Studies*, 54, 386-398.
- Phillips, P. and Lee, I. (2011). Crime analysis through spatial areal aggregated density patterns. *Geoinformatica*, 15, 49-74.
- Phulari, S., Khamitkar, S., Deshmukh, N., Bhalchandra, P., Lokhande, S. and Shinde, A. (2010). Understanding formulation of social capital in online social network sites (SNS). *arXiv preprint arXiv:1002.1201*.
- Pickett, K. (2013). Reducing inequality: An essential step for development and wellbeing. *Progressive Economy*, 1, 39-43.
- Piekut, A., Rees, P., Valentine, G. and Kupiszewski, M. (2012). Multidimensional diversity in two European cities: thinking beyond ethnicity. *Environment and Planning A*, 44, 2988-3009.
- Pitner, R. O., Yu, M. and Brown, E. (2012). Making neighborhoods safer: Examining predictors of residents' concerns about neighborhood safety. *Journal of Environmental Psychology*, 32, 43-49.
- Piza, E. L. (2012). *Using poisson and negative binomial regression models to measure the influence of risk on crime incident counts*, Rutgers Center on Public Security, Newark, NJ, USA.
- Plotkowiak, T. (2014). *The Influence of Social Capital on Information Diffusion in Twitter's Interest-Based Social Networks*. University of St. Gallen.
- Poblete, B., Garcia, R., Mendoza, M. and Jaimes, A. (2011). Do all birds tweet the same?: characterizing twitter around the world. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 1025-1030.
- Poile, C. and Safayeni, F. (2016). Using computational modeling for building theory: A double edged sword. *Journal of Artificial Societies and Social Simulation*, 19.
- Ponce, M. S. H., Rosas, R. P. E. and Lorca, M. B. F. (2014). Social capital, social participation and life satisfaction among Chilean older adults. *Revista de Saude Publica*, 48, 5, 739-749.
- Ponnuru, K. R., Gupta, R. and Trivedi, S. K. (2017). *Sentiment Analysis with Social Media Analytics, Methods, Process, and Applications*, IGI Global.
- Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*, Princeton University Press.
- Poulsen, E. and Kennedy, L. W. (2004). Using dasymetric mapping for spatially aggregated crime data. *Journal of Quantitative Criminology*, 20, 243-262.
- Power, A. (2004). Sustainable communities and sustainable development: a review of the sustainable communities plan. ESRC report 23. Available: <http://eprints.lse.ac.uk/28313/1/CASereport23.pdf>.
- Prasetyo, P. K., Gao, M., Lim, E.-P. and Scollon, C. N. 2013. Social sensing for urban crisis management: The case of singapore haze. *Social Informatics*. Springer.
- Preoțiu-Pietro, D., Lampos, V. and Aletras, N. (2015). An analysis of the user occupational class through Twitter content.

- Prichard, J., Watters, P., Krone, T., Spiranovic, C. and Cockburn, H. (2015). Social media sentiment analysis: A new empirical tool for assessing public opinion on crime. *Current Issues Crim. Just.*, 27, 217.
- Procter, R., Vis, F. and Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16, 197-214.
- Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, 6, 65-78.
- Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*, Simon and Schuster, New York.
- Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century the 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, 30, 2, 137-174.
- Quercia, D., Ellis, J., Capra, L. and Crowcroft, J. (2012). Tracking gross community happiness from tweets. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 965-968.
- Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés ou essai de physique sociale*, Bachelier, Paris.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179-191.
- Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23, 3-13.
- Rajan, A. P. and Victor, S. (2014). Web sentiment analysis for scoring positive or negative words using Tweeter data. *International Journal of Computer Applications*, 96.
- Ramadan, R. (2017). Questioning the role of Facebook in maintaining Syrian social capital during the Syrian crisis. *Heliyon*, 3, e00483.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846-850.
- Rani, Y. and Rohil, H. (2013). A study of hierarchical clustering algorithm. *International Journal of Information & Computation Technology*, 3, 10, 1115-1122.
- Rao, N. R. 2016. Chapter 9. Social Media: An Enabler in Developing Business. *Social Media Listening and Monitoring for Business Applications*.
- Ravindran, S. K. and Garg, V. (2015). *Mastering social media mining with R*, Packt Publishing Ltd, Mumbai.

- Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). *Applied regression analysis: a research tool*, Springer Science & Business Media, New York.
- Rees, P. and Butt, F. (2004). Ethnic change and diversity in England, 1981–2001. *Area*, 36, 174-186.
- Rees, P., Martin, D. and Williamson, P. (2002). Census data resources in the United Kingdom. In *The Census Data System* (eds P. Rees, D. Martin and P. Williamson), pp. 1-24. Wiley, Chichester.
- Rees, P., Wohland, P., Norman, P. and Boden, P. (2012). Ethnic population projections for the UK, 2001–2051. *Journal of Population Research*, 29, 45-89.
- Reilly, B. and Witt, R. (2008). Domestic burglaries and the real price of audio-visual goods: Some time series evidence for Britain. *Economics Letters*, 100, 96-100.
- Reisinger, H. (1997). The impact of research designs on R2 in linear regression models: an exploratory meta-analysis. *Journal of Empirical Generalisations in Marketing Science*, 2, 1-12.
- Remus, R., Quasthoff, U. and Heyer, G. (2010). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. *LREC*.
- Resnick, P. (2001). Beyond bowling together: Sociotechnical capital. *HCI in the New Millennium*, 247-272.
- Reynolds, A. J., Temple, J. A., Robertson, D. L. and Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *Jama*, 285, 2339-2346.
- Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A. and Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5, 1-29.
- Richard, H. (2013). Twitter in numbers [Online]. <http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html>. [Accessed 04/07/2015].
- Riddlesden, D. and Singleton, A. D. (2014). Broadband speed equity: A new digital divide? *Applied Geography*, 52, 25-33.
- Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review no.8 ESRC.
- Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M. and Stern, M. J. (2015). Digital inequalities and why they matter. *Information, Communication & Society*, 18, 569-582.
- Roblyer, M. D., McDaniel, M., Webb, M., Herman, J. and Witty, J. V. (2010). Findings on Facebook in higher education: A comparison of college faculty and student uses and perceptions of social networking sites. *The Internet and Higher Education*, 13, 3, 134-140.
- Roehl, J., Rosenbaum, D. P., Costello, S. K., Coldren, J., Schuck, A., Kunard, L. and Forde, D. (2006). Strategic Approaches to Community Safety Initiative (SACSI) in 10 US Cities: The Building Blocks for Project Safe Neighborhoods. *Washington, DC: National Institute of Justice*.

- Rokach, L. and Maimon, O. 2005. Clustering methods. *Data mining and knowledge discovery handbook*. Springer, Boston MA.
- Ronoh, L., Karie, N. and Rabah, K. (2017). Using SNA Centrality Metrics to Detect Suspicious Social Media Users to Aid Law Enforcement Agencies in Kenya. *Mara Research Journal of Computer Science & Security-ISSN 2518-8453*, 2, 1-19.
- Rosenbaum, D. P. (1987). The theory and research behind neighborhood watch: Is it a sound fear and crime reduction strategy? *Crime & Delinquency*, 33, 103-134.
- Rosenbaum, D. P. (1988). Community crime prevention: A review and synthesis of the literature. *Justice Quarterly*, 5, 323-395.
- Rosenshein, L., Scott, L. and Pratt, M. (2011). Exploratory Regression: A tool for modeling complex phenomena. ESRI ArcUser.
- Ross, A., Duckworth, K., Smith, D. J., Wyness, G. and Schoon, I. (2011). Prevention and Reduction: A review of strategies for intervening early to prevent or reduce youth crime and anti-social behaviour. *Centre for Analysis of Youth Transitions, Department of Education, UK*.
- Roth, J. J. (2018). Property Crime Clearance in Small Jurisdictions: Police and Community Factors. *Criminal Justice Review*, 0734016817752434.
- Rountree, P. W. and Warner, B. D. (1999). Social ties and crime: Is the relationship gendered. *Criminology*, 37, 789-814.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
- Rufrancos, H. G., Power, M., Pickett, K. E. and Wilkinson, R. (2013). Sociology and Criminology-Open Access.
- Rukus, J. and Warner, M. E. (2013). Crime rates and collective efficacy: The role of family friendly planning. *Cities*, 31, 37-46.
- Ryan, T., Chester, A., Reece, J. and Xenos, S. (2014). The uses and abuses of Facebook: A review of Facebook addiction. *Journal of Behavioral Addictions*, 3, 3, 133-148.
- Sabates, R., Feinstein, L. and Shingal, A. (2008). Educational Inequality and Juvenile Crime: An Area-based Analysis. Research report No. 26. Available from: <http://webarchive.nationalarchives.gov.uk/20130401151715/http://www.education.gov.uk/publications/eOrderingDownload/WBL26.pdf>. [Accessed 02/03/18].
- Sachdeva, N. and Kumaraguru, P. (2015). Social networks for police and residents in India: exploring online communication for crime prevention. *Proceedings of the 16th Annual International Conference on Digital Government Research*, 256-265.
- Sachs, D. E., Eckel, E. J. and Langan, K. A. (2011). Striking a balance: Effective use of Facebook in an academic library. *Internet Reference Services Quarterly*, 16, 35-54.
- Sadah, S. A., Shahbazi, M., Wiley, M. T. and Hristidis, V. (2016). Demographic-based content analysis of web-based health-related social media. *Journal of Medical Internet Research*, 18, 8: e148.

- Sagar, S. S., Boardley, I. D. and Kavussanu, M. (2011). Fear of failure and student athletes' interpersonal antisocial behaviour in education and sport. *British Journal of Educational Psychology*, 81, 3, 391-408.
- Saggar, S., Somerville, W., Ford, R. and Sobolewska, M. (2012). The impacts of migration on social cohesion and integration. *Final report to the Migration Advisory Committee, Home Office, London*.
- Sahayak, V., Shete, V. and Pathan, A. (2015). Sentiment Analysis on Twitter Data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE) Issue, 1*.
- Saitta, S., Raphael, B. and Smith, I. F. (2007). A bounded index for cluster validity. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 174-187.
- Sajuria, J., Hudson, D., Dasandi, N. and Theocharis, Y. (2014). Tweeting Alone? An Analysis of Bridging and Bonding Social Capital in Online Networks. *American Politics Research*, 1532673X14557942.
- Sameki, M., Gentil, M., Mays, K. K., Guo, L. and Betke, M. (2016). Dynamic Allocation of Crowd Contributions for Sentiment Analysis during the 2016 US Presidential Election. *arXiv preprint arXiv:1608.08953*.
- Sampson, R. J. (1988). Local friendship ties and community attachment in mass society: A multilevel systemic model. *American Sociological Review*, 766-779.
- Sampson, R. J. (2012). *Great American city: Chicago and the enduring neighborhood effect*, University of Chicago Press, Chicago.
- Sampson, R. J. and Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American journal of sociology*, 774-802.
- Sampson, R. J. and Laub, J. H. (2003). Life-Course Desisters? Trajectories Of Crime Among Delinquent Boys Followed To Age 70. *Criminology*, 41, 555-592.
- Sampson, R. J., Morenoff, J. D. and Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual Review of Sociology*, 28, 443-478.
- Sampson, R. J. and Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban Neighborhoods 1. *American Journal of Sociology*, 105, 3, 603-651.
- Sampson, R. J., Raudenbush, S. W. and Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.
- Sampson, R. J. and Wooldredge, J. D. (1987). Linking the micro-and macro-level dimensions of lifestyle-routine activity and opportunity models of predatory victimization. *Journal of Quantitative Criminology*, 3, 371-393.
- Sander, T. and Teh, P. L. (2014). A concept to measure social capital in social network sites. *International Journal of Future Computer and Communication*, 3, 105-107.
- Sandhana, C. and Sangareddy, B. (2015). Survey on predicting crime using twitter sentiment and weather data. National Conference on Emerging Trends and Advances. In Information Technology Department of CS & E, AIT, Chikmagalur, January 29-30, 2016.

- Santangelo, T. M. (2011). *Does voting really matter? The effect of voting turnout rates on crime.*, PhD thesis, Georgetown University, Washington DC.
- Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. *International Conference on Artificial Neural Networks*, 175-184.
- Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension*, 37, 3, 1-5.
- Sariaslan, A., Långström, N., D'Onofrio, B., Hallqvist, J., Franck, J. and Lichtenstein, P. (2013). The impact of neighbourhood deprivation on adolescent violent criminality and substance misuse: a longitudinal, quasi-experimental study of the total Swedish population. *International Journal of Epidemiology*, 42, 4, 1057-1066.
- Sarstedt, M. and Mooi, E. (2014). *Cluster analysis*, Springer, Berlin.
- Savage, M. and Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41, 885-899.
- Savage, M. and Burrows, R. (2009). Some further reflections on the coming crisis of empirical sociology. *Sociology*, 43, 762-772.
- Sawyer, R. and Chen, G.-M. (2012). The impact of social media on intercultural adaptation. *Intercultural Communication Studies*, 21,2, 151-169.
- Schabenberger, O. and Gotway, C. A. (2004). *Statistical methods for spatial data analysis*, Chapman and Hall, CRC Press, Florida.
- Scheepers, H., Scheepers, R., Stockdale, R. and Nurdin, N. (2014). The dependent variable in social media use. *Journal of Computer Information Systems*, 54, 25-34.
- Schultz, C. D. (2017). Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages? *Electronic Commerce Research and Applications*, 26, 23-34.
- Schulz, A., Thanh, T. D., Paulheim, H. and Schweizer, I. (2013). A fine-grained sentiment analysis approach for detecting crisis related microposts. *ISCRAM*.
- Schuurman, N. (2000). Trouble in the heartland: GIS and its critics in the 1990s. *Progress in Human Geography*, 24, 569-590.
- Šerbetar, I. (2015). Establishing Some Measures of Absolute and Relative Reliability of a Motor Test. *Hrvatski časopis za odgoj i obrazovanje*, 17, 37-48.
- Shammas, V. L. and Sandberg, S. (2015). Habitus, capital, and conflict: Bringing Bourdieusian field theory to criminology. *Criminology and Criminal Justice* 16(2), pp195-213.
- Shannon, C. (1948). A mathematical theory of communication, bell System technical Journal 27: 379-423 and 623-656. *Mathematical Reviews (MathSciNet): MR10, 133e*.
- Sharp, D. and Atherton, S. (2007). To serve and protect? The experiences of policing in the community of young people from black and other ethnic minority groups. *British Journal of Criminology*, 47, 746-763.
- Shashidhar, V., Pandey, N. and Aggarwal, V. (2015). Spoken english grading: Machine learning with crowd intelligence. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2089-2097.

- Shaw, C. R. and McKay, H. D. (1942). *Juvenile delinquency and urban areas*, University of Chicago Press, IL, US.
- Shaw, C. R. and McKay, H. D. (1969). *Juvenile delinquency and urban areas. Revised edition*, Chicago University Press, Chicago.
- Shepherd, P. J. (2006). *Neighbourhood profiling and classification for community safety*. Phd, University of Leeds.
- Sheppard, E. (1995). GIS and society: towards a research agenda. *Cartography and Geographic Information Systems*, 22, 5-16.
- Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P. and Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising: A report to the United States Congress*, US Department of Justice, Office of Justice Programs Washington, DC.
- Shirkhorshidi, A. S., Aghabozorgi, S. and Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS One*, 10, 12, e0144059.
- Short, E. and Ditton, J. (1998). Seen and now heard: Talking to the targets of open street CCTV. *The British Journal of Criminology*, 38, 404-428.
- Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86, 420.
- Silver, E. and Miller, L. L. (2004). Sources of informal Social Control in Chicago neighbourhoods. *Criminology*, 42, 551-584.
- Simon, T., Goldberg, A. and Adini, B. (2015). Socializing in emergencies—A review of the use of social media in emergency situations. *International Journal of Information Management*, 35, 609-619.
- Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688.
- Simpson, L. (2004). Statistics of racial segregation: measures, evidence and policy. *Urban Studies*, 41, 661-681.
- Singh, R. and Kaur, R. (2015). Sentiment Analysis on Social Media and Online Review. *International Journal of Computer Applications*, 121, 20, 44-48.
- Singleton, A. D. and Longley, P. A. (2009). Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29, 289-298.
- Sinha, A., Malo, P. and Kuosmanen, T. (2015). A Multiobjective Exploratory Procedure for Regression Model Selection. *Journal of Computational and Graphical Statistics*, 24, 154-182.
- Sivarajah, U., Irani, Z. and Weerakkody, V. (2015). Evaluating the use and impact of Web 2.0 technologies in local government. *Government Information Quarterly*, 32, 473-487.
- Skogan, W. G. (1989). Communities, crime, and neighborhood organization. *Crime & Delinquency*, 35, 437-457.
- Skuse, A. and Brimacombe, T. (2014). Social networking, social media and complex emergencies: issues paper.

- Sloan, L. and Quan-Haase, A. (2017). *The SAGE handbook of social media research methods*, Sage, London.
- Smith, D. P., Edwards, R. and Caballero, C. (2011). The geographies of mixed-ethnicity families. *Environment and Planning A*, 43, 1455-1476.
- Smith, M. (2010). Facebook Community Pages: What Your Business Need to Know. Available from: <http://www.socialmediaexaminer.com/facebook-community-pages-what-your-business-needs-to-know/>. [Accessed 09/11/2017].
- Smith, M., Hansen, D. L. and Gleave, E. (2009). Analyzing enterprise social media networks. *Computational Science and Engineering, 2009. CSE'09. International Conference on*, 4, 705-710.
- Snipes, M. and Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3, 3-9.
- Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the conference on empirical methods in natural language processing*, 254-263.
- Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45, 427-437.
- Spielman, S. E. and Thill, J.-C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, 32, 110-122.
- Spiliopoulou, A.-Y., Mahony, S., Routsis, V. and Kamposiori, C. (2014). Cultural institutions in the digital age: British Museum's use of Facebook Insights. *Participations*, 11, 286-303.
- Statistica.com (2016). Global social networks ranked by number of users. Available from: <http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. [Accessed 31/07/2016].
- Statistica.com (2017). Facebook users in the United Kingdom by age group and gender. Available from: <http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. [Accessed 31/07/2016].
- Steinfeld, C., Ellison, N. B. and Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29, 434-445.
- Stenson, K. (1993). Community policing as a governmental technology: Community policing as a governmental technology. *Economy and Society*, 22, 373-389.
- Stephen, M., Olivier, M. and Sunčica, V. (2010). The Crime Reducing Effect of Education.
- Stewart, A. 2014. *Microsimulation Models of the Criminal Justice System. Encyclopedia of Criminology and Criminal Justice*. Springer.
- Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C. (2018). Social media analytics—Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39, 156-168.

- Stillwell, J. and Phillips, D. (2006). Diversity and change: understanding the ethnic geographies of Leeds. *Journal of Ethnic and Migration Studies*, 32, 1131-1152.
- Stockley, P. B., Charles Wankel, D., Ann Voss, K. and Kumar, A. (2013). The value of social media: are universities successfully engaging their audience? *Journal of Applied Research in Higher Education*, 5, 156-172.
- Stojanovski, D., Dimitrovski, I. and Madjarov, G. (2014). TweetViz: Twitter Data Visualization. *Proceedings of the Data Mining and Data Warehouses*.
- Sturgis, P., Brunton-Smith, I., Kuha, J. and Jackson, J. (2014). Ethnic diversity, segregation and the social cohesion of neighbourhoods in London. *Ethnic and Racial Studies*, 37, 1286-1309.
- Sun, I. Y., Triplett, R. and Gainey, R. R. (2004). Neighborhood characteristics and crime: A test of Sampson and Groves' model of social disorganization. *W. Criminology Rev.*, 5, 1.
- Sutherland, A., Brunton-Smith, I. and Jackson, J. (2013). Collective efficacy, deprivation and violence in London. *British Journal of Criminology*, 53, 405-420.
- Swatt, M. L., Varano, S. P., Uchida, C. D. and Solomon, S. E. (2013). Fear of crime, incivilities, and collective efficacy in four Miami neighborhoods. *Journal of Criminal Justice*, 41, 1-11.
- Sweeten, G., Piquero, A. R. and Steinberg, L. (2013). Age and the explanation of crime, revisited. *Journal of Youth and Adolescence*, 42, 921-938.
- Swier, N., Komarniczky, B. and Clapperton, B. (2015). *Using geolocated Twitter traces to infer residence and mobility*. GSS Methodology Series No 41, Office for National Statistics, London.
- Syaifudin, Y. W. and Puspitasari, D. (2017). Twitter Data Mining for Sentiment Analysis on Peoples Feedback Against Government Public Policy. *MATTER: International Journal of Science and Technology*, 3.
- Sykes, A. O. (1993). An introduction to regression analysis. Working paper no. 20. University of Chicago.
- Szolnoki, G. and Hoffmann, D. (2013). Online, face-to-face and telephone surveys—Comparing different sampling methods in wine consumer research. *Wine Economics and Policy*, 2, 57-66.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37, 267-307.
- Takagi, D. and Kawachi, I. (2014). Neighborhood social heterogeneity and crime victimization in Japan: Moderating effects of social networks. *Asian Journal of Criminology*, 9, 4, 271-284.
- Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C. and He, X. (2014). Interpreting the public sentiment variations on twitter. *Ieee Transactions on Knowledge and Data Engineering*, 26, 5, 1158-1170.
- Tang, W. and Lease, M. (2011). Semi-supervised consensus labeling for crowdsourcing. *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*, 1-6.

- Tao, K., Abel, F., Hauff, C. and Houben, G.-J. (2012). What makes a tweet relevant for a topic? # MSM, 49-56.
- Tao, K., Abel, F., Hauff, C., Houben, G.-J. and Gadiraju, U. (2013). Groundhog day: near-duplicate detection on twitter. *Proceedings of the 22nd International Conference on World Wide Web*, 1273-1284.
- Tapia, A. H., Moore, K. A. and Johnson, N. J. (2013). Beyond the trustworthy tweet: A deeper understanding of microblogged data use by disaster response and humanitarian relief organizations. *Proceedings of the 10th International ISCRAM Conference*, 770-778.
- Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53.
- Taylor, R. B. (1995). The impact of crime on communities. *The Annals of the American Academy of Political and Social Science*, 539, 28-45.
- Teasdale, B., Clark, L. M. and Hinkle, J. C. (2012). Subprime lending foreclosures, crime, and neighborhood disorganization: Beyond internal dynamics. *American Journal of Criminal Justice*, 37, 163-178.
- TEC (2010). Election turnout. The Electoral Commission. Available from: <https://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/uk-general-elections/2010-uk-general-election-results>. [Accessed 28/07/2016].
- Temkin, K. and Rohe, W. M. (1998). Social capital and neighborhood stability: An empirical investigation. *Housing Policy Debate*, 9, 61-88.
- Teng, C. E. and Joo, T. M. (2017). Analyzing the Usage of Social Media: A Study on Elderly in Malaysia. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 11, 713-719.
- Tersteeg, A. K. and Pinkster, F. M. (2016). "Us Up Here and Them Down There" How Design, Management, and Neighborhood Facilities Shape Social Distance in a Mixed-Tenure Housing Development. *Urban Affairs Review*, 52, 751-779.
- Tess, P. A. (2013). The role of social media in higher education classes (real and virtual)—A literature review. *Computers in Human Behavior*, 29, A60-A68.
- Tewksbury, R. and Mustaine, E. E. (2006). Where to find sex offenders: An examination of residential locations and neighborhood conditions. *Criminal Justice Studies*, 19, 61-75.
- Theiss, S. K., Burke, R. M., Cory, J. L. and Fairley, T. L. (2016). Getting beyond impressions: an evaluation of engagement with breast cancer-related Facebook content. *Mhealth*, 2, 41, 1-7.
- Thelwall, M., Buckley, K. and Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63, 163-173.
- Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K. and Kerdprasopb, N. (2015). The clustering validity with silhouette and sum of squared errors. *Learning*, 3, 7.

- Thomas, M. J., Stillwell, J. C. and Gould, M. I. (2016). Modelling the duration of residence and plans for future residential relocation: a multilevel analysis. *Transactions of the Institute of British Geographers*, 41, 297-312.
- Thompson, S. K. and Gartner, R. (2014). The Spatial Distribution and Social Context of Homicide in Toronto's Neighborhoods. *Journal of Research in Crime and Delinquency*, 51, 1, 88-118.
- Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411-423.
- Timberlake, S. (2005). Social capital and gender in the workplace. *Journal of Management Development*, 24, 34-44.
- Ting, I. (2011). *Social network mining, analysis, and research trends: Techniques and applications*, IGI Global, Hershey, USA.
- Tittle, C. R., Ward, D. A. and Grasmick, H. G. (2003). Gender, age, and crime/deviance: A challenge to self-control theory. *Journal of Research in Crime and Delinquency*, 40, 426-453.
- Thuczak, A. (2013). The analysis of the phenomenon of spatial autocorrelation of indices of agricultural output. *Metody Ilościowe w Badaniach Ekonomicznych*, 14, 251-260.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 234-240.
- Todorova, M. R. (2013). *The valence of electronic word-of-mouth (eWOM) and choice of online opinion platform*. Masters Dissertation, Copenhagen Business School.
- Toma, C. L. and Hancock, J. T. (2013). Self-affirmation underlies Facebook use. *Personality and Social Psychology Bulletin*, 39, 321-331.
- Tomczak, M., Tomczak, E., Kleka, P. and Lew, R. (2014). Using power analysis to estimate appropriate sample size. *Trends in Sport Sciences*, 21, 4, 195-206.
- Tompson, L., Johnson, S., Ashby, M., Perkins, C. and Edwards, P. (2015). UK open source crime data: accuracy and possibilities for research. *Cartography and Geographic Information Science*, 42, 97-111.
- Tonidandel, S., King, E. and Cortina, J. (2015). *Big data at work: The data science revolution and organizational psychology*, Routledge, New York.
- Tosti-Kharas, J. and Conley, C. (2016). Coding Psychological Constructs in Text Using Mechanical Turk: A Reliable, Accurate, and Efficient Alternative. *Frontiers in Psychology*, 7.
- Townsend, L., Wallace, C. and Fairhurst, G. (2015). 'Stuck out here': The critical role of broadband for remote rural places. *Scottish Geographical Journal*, 131, 171-180.
- Tranmer, M. and Steel, D. G. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, 30, 817-831.
- Traunmueller, M., Quattrone, G. and Capra, L. (2014). *Mining mobile phone data to investigate urban crime theories at scale*, Springer, Barcelona, Spain.

- Trottier, D. (2015). Open source intelligence, social media and law enforcement: Visions, constraints and critiques. *European Journal of Cultural Studies*, 18, 530-547.
- Tsai, H.-y. S., Shillair, R. and Cotten, S. R. (2017). Social Support and “Playing Around” An Examination of How Older Adults Acquire Digital Literacy With Tablet Computers. *Journal of Applied Gerontology*, 36, 29-55.
- Tselios, V., McCann, P. and van Dijk, J. (2016). Understanding the gap between reality and expectation: Local social engagement and ethnic concentration. *Urban Studies*, 0042098016650395.
- Tseloni, A. (2006). Multilevel modelling of the number of property crimes: Household and area effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 2, 205-233.
- Tsimonis, G. and Dimitriadis, S. (2014). Brand strategies in social media. *Marketing Intelligence & Planning*, 32, 328-344.
- Tsushima, M. (1996). Economic structure and crime: the case of Japan. *The Journal of Socio-Economics*, 25, 497-515.
- Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsn*, 10, 178-185.
- Tuomisto, H. (2010). A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia*, 164, 853-860.
- Turcotte, M. (2015). Trends in Social Capital in Canada: Results from General Social Survey. Available from: <https://www150.statcan.gc.ca/n1/en/pub/89-652-x/89-652-x2015002-eng.pdf?st=hjwewEou>. [Accessed 21/08/2016].
- Turney, K. and Harknett, K. (2010). Neighborhood disadvantage, residential stability, and perceptions of instrumental support among new mothers. *Journal of Family Issues*, 31, 499-524.
- Twitter (2016). Twitter usage: Company facts. Available from: <https://about.twitter.com/company>. [Accessed 28/05/2016].
- Twitter. (2017a). Adding Locations to your tweets. Available from: <https://support.twitter.com/articles/78525#>. [Accessed 10/08/2017].
- Twitter. (2017b). API Rate Limits. Available from: <https://dev.twitter.com/rest/public/rate-limiting>. [Accessed 10/08/2017].
- Twitter. (2017c). How to Build a Query. Available from: <https://dev.twitter.com/rest/public/search>. [Access 17/08/2017].
- Twitter. (2017d). Streaming API request parameters. Available from: <https://dev.twitter.com/streaming/overview/request-parameters>. [Accessed 05/09/2017].
- Uchida, C. D., Swatt, M. L., Solomon, S. E. and Varano, S. (2013). *Neighborhoods and crime: collective efficacy and social cohesion in Miami-Dade county*, Final Report submitted to the National Institute of Justice.

- Unsworth, R. and Stillwell, J. C. H. (2004). *Twenty-first century Leeds: Geographies of a regional city*, Leeds University Press.
- USDOJ (2013). Developing a Policy on the Use of Social Media in Intelligence and Investigative Activities. US Department of Justice. Available from: <https://it.ojp.gov/documents/d/Developing%20a%20Policy%20on%20the%20Use%20of%20Social%20Media%20in%20Intelligence%20and%20Inves....pdf>. [Accessed 23/05/2016].
- Valenzuela, S., Park, N. and Kee, K. F. (2009). Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication*, 14, 875-901.
- Van Craenendonck, T. and Blockeel, H. (2015). Using internal validity measures to compare clustering algorithms. *Benelearn 2015 Poster presentations (online)*, 1-8.
- Van Der Gaag, M. and Snijders, T. A. (2005). The Resource Generator: social capital quantification with concrete items. *Social Networks*, 27, 1-29.
- Van Deursen, A. J. and Mossberger, K. (2018). Any thing for anyone? A new digital divide in internet-of-things skills. *Policy & internet*, 122-140.
- Van, E. (2006). Gender differences in the creation of different types of social capital: A multilevel study. *Social Networks*, 28, 1, 24-37.
- Van Patten, I. T., McKeldin-Coner, J. and Cox, D. (2009). A microspatial analysis of robbery: prospective hot spotting in a small city. *Crime Mapping: A Journal of Research and Practice*, 1, 1, 7-32.
- Vanam, R. and Reznik, Y. A. (2013). Perceptual pre-processing filter for user-adaptive coding and delivery of visual information. *PCS*, 426-429.
- Várallyai, L., Herdon, M. and Botos, S. (2015). Statistical analyses of digital divide factors. *Procedia Economics and Finance*, 19, 364-372.
- Vargo, C. J. and Hopp, T. (2017). Socioeconomic Status, Social Capital, and Partisan Polarity as Predictors of Political Incivility on Twitter: A Congressional District-Level Analysis. *Social Science Computer Review*, 35, 10-32.
- Varol, O., Ferrara, E., Davis, C. A., Menczer, F. and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.
- Vavrek, M. J. (2011). Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14, 1T.
- Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R. and Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS One*, 8, e73990.
- Venerandi, A., Quattrone, G., Capra, L., Quercia, D. and Saez-Trumper, D. (2015). Measuring urban deprivation from user generated content. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 254-264.
- Vermeulen, F., Tillie, J. and van de Walle, R. (2012). Different effects of ethnic diversity on social capital: density of foundations and leisure associations in Amsterdam neighbourhoods. *Urban Studies*, 49, 337-352.

- Veronique, S. (2014). Measuring Social Capital. ONS Framework Report. Available from: http://webarchive.nationalarchives.gov.uk/20160107115718/http://www.ons.gov.uk/ons/dcp171766_371693.pdf. [Accessed 07/01/2018].
- Veysey, B. M. and Messner, S. F. (1999). Further testing of social disorganization theory: An elaboration of Sampson and Groves's "community structure and crime". *Journal of Research in Crime and Delinquency*, 36, 156-174.
- Vickers, D. and Rees, P. (2007). Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 379-403.
- Vieu, P. (1991). Nonparametric regression: optimal local bandwidth choice. *Journal of the Royal Statistical Society. Series B (Methodological)*, 453-464.
- Vilares, I., Dam, G. and Kording, K. (2011). Trust and reciprocity: Are effort and money equivalent? *PLoS One*, 6, e17113.
- Volkova, S. and Bachrach, Y. (2015). On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18, 726-736.
- Wakamiya, S., Lee, R. and Sumiya, K. (2013). *Social-urban neighborhood search based on crowd footprints network*, Springer, Kyoto, Japan.
- Walford, G., Tucker, E. and Viswanathan, M. (2010). *The SAGE handbook of measurement*, Sage Publications, London.
- Walker, J. and Maddan, S. (2013). *Statistics in criminology and criminal justice: Analysis and interpretation, Forth edition* Jones & Bartlett Learning, Burlington, Massachusetts.
- Wallace, C., Vincent, K., Luguzan, C., Townsend, L. and Beel, D. (2017). Information technology and social cohesion: a tale of two villages. *Journal of Rural Studies*, 54, 426-434.
- Wan, Y. and Gao, Q. (2015). An ensemble sentiment classification system of twitter data for airline services analysis. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 1318-1325.
- Wang, G. C. (1996). How to handle multicollinearity in regression modeling. *The Journal of Business Forecasting*, 15, 23.
- Wang, N., Kosinski, M., Stillwell, D. and Rust, J. (2014). Can well-being be measured using Facebook status updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research*, 115, 483-491.
- Wang, X., Gerber, M. S. and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 231-238.
- Wang, Z., Bai, G., Chowdhury, S., Xu, Q. and Seow, Z. L. (2017). TwiInsight: Discovering Topics and Sentiments from Social Media Datasets. *arXiv preprint arXiv:1705.08094*.
- Ward, J. T., Nobles, M. R., Youstin, T. J. and Cook, C. L. (2014). Placing the neighborhood accessibility–burglary link in social-structural context. *Crime & Delinquency*, 60, 739-763.

- Warner, B. D., Beck, E. and Ohmer, M. L. (2010). Linking informal social control and restorative justice: Moving social disorganization theory beyond community policing. *Contemporary Justice Review*, 13, 355-369.
- Warner, B. D. and Rountree, P. W. (1997). Local social ties in a community and crime model: Questioning the systemic nature of informal social control. *Social Problems*, 44, 520-536.
- Weatherburn, D. (2001). What causes crime? *BOCSAR NSW Crime and Justice Bulletins*, 11.
- Wedlock, E. (2006). *Crime and cohesive communities*, Home Office, London.
- Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19, 1, 231.
- Weisberg, S. (2005). *Applied linear regression*, John Wiley & Sons, USA.
- Weisburd, D., Braga, A. A., Groff, E. R. and Wooditch, A. (2017). Can hot spots policing reduce crime in urban areas? An agent-based simulation. *Criminology*, 55, 137-173.
- Weisburd, D. and Britt, C. (2007). *Statistics in criminal justice*, Springer Science & Business Media, New York.
- Weisburd, D. and Piquero, A. R. (2008). How well do criminologists explain crime? Statistical modeling in published studies. *Crime and Justice*, 37, 453-502.
- Wellman, B. (2006). *Networks in the global village: Life in contemporary communities*, Tailor & Frances Group, London.
- Wellman, B., Haase, A. Q., Witte, J. and Hampton, K. (2001). Does the Internet increase, decrease, or supplement social capital? Social networks, participation, and community commitment. *American Behavioral Scientist*, 45, 3, 436-455.
- Westermann, O., Ashby, J. and Pretty, J. (2005). Gender and social capital: The importance of gender differences for the maturity and effectiveness of natural resource management groups. *World Development*, 33, 1783-1799.
- Whiting, E. and Harper, R. (2003). Young people and social capital. *Office for National Statistics, London*.
- Whitley, E. and Ball, J. (2001). Statistics review 1: Presenting and summarising data. *Critical Care*, 6, 66.
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B. and Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, 75, 5, 1182-1189.
- Wikström, P.-O. H. and Treiber, K. (2016). Social disadvantage and crime: a criminological puzzle. *American Behavioral Scientist*, 60, 1232-1259.
- Wikström, S. and Wikström, P.-O. H. (2001). Why are Single Parents More often Threatened with Violence? A Question of Ecological Vulnerability? *International Review of Victimology*, 8, 183-198.
- Williams, J. (2010). Resources in unexpected places: Social cohesion and successful community internet. Community Informatics conference.

- Williams, M. L., Burnap, P. and Sloan, L. (2016). Crime sensing with big data: the affordances and limitations of using open source communications to estimate crime patterns. *British Journal of Criminology*, 10.1093/bjc/azw031.
- Williams, M. L., Edwards, A., Housley, W., Burnap, P., Rana, O., Avis, N., Morgan, J. and Sloan, L. (2013). Policing cyber-neighbourhoods: tension monitoring and social media networks. *Policing and Society*, 23, 4, 461-481.
- Williams, N. (2009). Template Twitter Strategy for Government Departments. Cabinet Office. Available from: <http://webarchive.nationalarchives.gov.uk/20090810000341/http://blogs.cabinetoffice.gov.uk/digitalengagement/post/2009/07/21/Template-Twitter-strategy-for-Government-Departments.aspx>. [Accessed 15/2/2017]
- Wilson, R. E., Gosling, S. D. and Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, 7, 3, 203-220.
- Wilson, T., Lovelace, R. and Evans, A. J. (2016). A path toward the use of trail users' tweets to assess effectiveness of the environmental stewardship scheme: An exploratory analysis of the pennine way national trail. *Applied Spatial Analysis and Policy*, 1-29.
- Wise, L., Skues, J. and Williams, B. (2011). Facebook in higher education promotes social but not academic engagement. *Changing demands, changing directions. Proceedings ascilite Hobart*, 1332-1342.
- Wisniewski, E., Bologeorges, S., Johnson, T. and Henry, D. B. (2013). The geography of citizen crime reporting. *American Journal of Community Psychology*, 52, 3, 324-332.
- Witt, R., Clarke, A. and Fielding, N. (1998). Crime, earnings inequality and unemployment in England and Wales. *Applied Economics Letters*, 5, 265-267.
- Woldoff, R. A. (2002). The effects of local stressors on neighborhood attachment. *Social Forces*, 81, 87-116.
- Wollebæk, D. and Selle, P. (2003). Participation and social capital formation: Norway in a comparative perspective. *Scandinavian Political Studies*, 26, 1, 67-91.
- Wollinger, G. R. (2017). Choice behavior after burglary victimization: Moving, safety precautions, and passivity. *European Journal of Criminology*, 14, 3, 329-343.
- Wong, D. W. (2004). *The modifiable areal unit problem (MAUP)*, Springer, Dordrecht, Netherlands.
- Wood, M. (2003). A balancing act? Tenure diversification in Australia and the UK. *Urban Policy and Research*, 21, 45-56.
- Woolcock, M. and Narayan, D. (2000). Social capital: Implications for development theory, research, and policy. *The World Bank Research Observer*, 15, 2, 225-249.
- Wooldridge, J. (2012). Chapter 3: Multiple regression analysis: Estimation pp73-112. *Introductory Econometrics, A Modern Approach (5th Ed.)*. USA.
- Wu, F. and Logan, J. (2016). Do rural migrants 'float' in urban China? Neighbouring and neighbourhood sentiment in Beijing. *Urban Studies*, 53, 2973-2990.
- Xiaojun, W., Leroy, W., Xu, C., He, W., Shen, J., Tian, X., Li, Y., Akula, V., Yan, G. and Tao, R. (2015). Gaining competitive intelligence from social media data:

- Evidence from two largest retail chains in the world. *Industrial Management & Data Systems*, 115, 1622-1636.
- Yamamura, E. (2011). How do neighbors influence investment in social capital? Homeownership and length of residence. *International Advances in Economic Research*, 17, 451-464.
- Yang, X., Zhang, Z., Zhang, Z., Mo, Y., Li, L., Yu, L. and Zhu, P. (2016). Automatic construction and global optimization of a multisentiment lexicon. *Computational Intelligence and Neuroscience*, 2016, 1-8.
- Yazidi, A., Hammer, H. L., Bai, A. and Engelstad, P. (2015). On Enhancing the Label Propagation Algorithm for Sentiment Analysis Using Active Learning with an Artificial Oracle. *International Conference on Artificial Intelligence and Soft Computing*, 799-810.
- Yim, O. and Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, 11, 1, 8-21.
- Yorkshire Evening Post. (2018). The UK's burglary claims hotspots. [Online]: <https://www.yorkshireeveningpost.co.uk/read-this/revealed-the-uks-burglary-claims-hotspots>. [Accessed 16/04/18].
- Yuan, Y. and McNeeley, S. (2017a). Social ties, collective efficacy, and crime-specific fear in Seattle neighborhoods. *Victims & Offenders*, 12, 1, 90-112.
- Yuan, Y. and McNeeley, S. (2017b). Social ties, collective efficacy, and crime-specific fear in Seattle neighborhoods. *Victims & offenders*, 12, 90-112.
- Zani, B., Cicognani, E. and Albanesi, C. (2001). Adolescents' sense of community and feeling of unsafety in the urban environment. *Journal of Community & Applied Social Psychology*, 11, 6, 475-489.
- Zanoni, P. and Janssens, M. (2004). Deconstructing difference: The rhetoric of human resource managers' diversity discourses. *Organization Studies*, 25, 1, 55-74.
- Zavattaro, S. M. and Sementelli, A. J. (2014). A critical examination of social media adoption in government: Introducing omnipresence. *Government Information Quarterly*, 31, 257-264.
- Zhang, H., Hill, S. and Rothschild, D. (2016). Geolocated Twitter Panels to Study the Impact of Events. *2016 AAAI Spring Symposium Series*.
- Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. and Liu, B. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. Hewlett-Packard Labs Technical Report HPL.
- Zhang, Z., Zhang, J. and Xue, H. (2008). Improved K-means clustering algorithm. *Image and Signal Processing, 2008. CISP'08. Congress on*, 5, 169-172.
- Zhao, J. S., Lawton, B. and Longmire, D. (2015). An examination of the micro-level crime-fear of crime link. *Crime & Delinquency*, 61, 19-44.
- Zhao, Y. (2015). *R and data mining: Examples and case studies*, Elsevier, New York.
- Zhou, X. and Jiang, T. (2016). Metamodel selection based on stepwise regression. *Structural and Multidisciplinary Optimization*, 54, 641-657.

Appendix A

Community Areas Facebook Engagement Metrics

Code	Community	Year Estab	FB Type	Total post	Post liked	Post comnt	Post shared	Fans	Total likes	Total comments	Total shares
56	Drighlington	2016	P	511	510	440	308	889	12661	3319	939
49	Boston Spa	2014	P	1220	1201	692	488	2023	34288	3298	2338
55	Meanwood	2015	P	192	188	103	99	1074	1951	336	1251
82	Arthington & Pool	2014	P	25	24	12	9	77	122	43	37
64	Shadwell	2014	G	188	178	159	44	346	1446	1210	88
42	Gildersome	2014	G	178	165	137	1	176	624	10	6
32	Moortown	2014	G	7024	6480	4866	2051	5373	119070	37243	8795
61	Bardsey and East Keswick	2016	P	94	85	19	65	144	451	43	298
38	Manston	2013	G	937	846	481	158	559	5320	1477	736
85	Crossgates	2013	G	8924	8035	7470	2450	6018	127872	130456	21765
88	Pudsey Lowtown	2012	P	26	23	8	9	1355	155	21	12
75	Colton	2012	P	199	176	97	77	360	1639	294	278
40	Belle Isle	2014	P	177	154	61	56	275	521	703	136
91	Kippax	2014	P	135	117	59	77	972	814	373	443
57	Barwick & Scholes	2011	P	907	778	388	558	1549	10351	1601	4152
87	New Farnley	2014	P	42	36	28	26	331	266	69	75
84	Scarcroft & Miles Hill	2015	P	21	18	10	10	79	151	21	22
33	Bramham	2014	P	320	274	88	84	335	2487	214	213
48	Hyde Park	2013	P	1101	939	263	416	1645	5646	467	1616
59	Farnley	2015	P	322	274	75	86	154	1136	149	224
50	Hawksworth	2012	P	113	94	29	60	272	529	44	132
28	Cross Green	2010	G	1023	849	638	159	257	5614	2024	584
73	Kirkstall	2015	G	1330	1101	855	324	1708	12135	4087	3345
27	Cottingley	2015	P	304	251	79	97	343	978	169	207
51	Richmond Hill	2012	P	2297	1839	840	875	2818	14824	2335	5202
41	Far Headingley	2014	P	190	152	54	30	1854	710	98	45
39	City Centre	2012	P	1882	1488	319	554	1453	6104	792	2405
65	Aberford	2016	P	93	73	23	17	147	413	40	55

14	Woodhouse	2012	P	774	606	152	321	1111	2090	261	905
60	Garforth East	2014	P	550	430	129	133	596	2051	425	404
71	Gipton North	2014	p	136	106	43	81	349	707	119	449
74	Holt Park	2012	P	570	439	168	342	1925	2415	780	1148
63	Halton and Whitkirk	2015	G	1250	962	729	517	4033	6011	5316	7825
10	Burmantofts	2013	P	335	257	69	56	112	882	124	94
7	Methley	2012	P	646	484	191	209	966	3668	528	1433
4	Farsley	2013	P	3246	2417	1101	1222	4763	16367	4283	6628
83	Oulton and Woodlesford	2014	P	96	70	16	27	347	334	70	476
26	Sandford, Ganners & Moorside	2015	G	438	318	155	51	185	1677	448	91
36	Thorner	2013	G	4195	3038	3047	677	1044	31138	18786	3607
89	Yeadon	2014	P	29	21	12	12	512	126	34	77
23	Mickelfield	2014	G	4860	3500	3017	858	1138	22822	22703	4558
77	Otley	2014	P	1804	1287	552	694	2511	13700	1583	5141
45	Beeston Hill	2011	G	92	65	54	0	25	154	172	10
92	Armley	2014	P	186	131	54	78	1181	723	346	445
16	Moor Allerton	2010	P	44	30	6	13	33	123	7	18
1	Burley Lodge & Little WH	2013	P	326	214	30	99	772	546	43	174
18	Osmondthorpe	2014	P	32	21	3	2	57	38	3	2
15	Swarcliffe	2014	P	419	273	119	98	223	791	353	203
30	Horsforth Newlathie & Woodside	2013	G	465	296	270	46	248	1809	1250	110
25	Oakwood	2010	G	5024	3153	2607	648	1862	16060	13797	1686
70	Collingham & Linton	2013	P	1295	810	570	327	1544	8606	2314	2520
3	Headingley	2011	P	1617	1010	189	439	1864	2867	308	989
62	Guiselley	2015	P	1901	1183	278	231	1940	5546	967	1304
67	Ledston & Ledsham	2011	P	125	77	51	41	241	295	195	176
21	Ireland Wood	2014	G	3367	2068	1916	732	3003	9865	13857	4357
8	New Wortley	2009	G	792	484	193	148	438	1300	488	317
35	Tinshill	2014	G	3905	2378	2166	886	3264	11029	15430	5850
54	Rothwell	2015	P	776	467	155	165	1787	2356	776	786

81	Bramley	2016	P	98	57	28	26	707	295	121	120
11	Swinnow & Fairfields	2012	P	1294	747	351	504	696	2069	1203	1728
86	Pudsey	2012	P	1282	717	173	499	1379	2017	278	1284
58	Swillington	2016	G	208	113	55	20	194	309	221	58
53	Adel	2008	P	61	33	8	15	174	67	17	37
17	Churwell	2012	P	93	50	23	16	227	89	40	26
2	Chapel Allerton	2015	P	299	158	48	19	574	436	92	40
46	Calverley	2015	P	196	103	24	25	457	302	45	89
47	Harewood and District	2014	P	210	108	54	59	333	263	98	267
93	Horsforth	2013	P	178	89	68	50	1813	596	208	468
66	Beeston	2015	P	99	48	7	13	95	244	35	48
90	Morley North	2011	G	387	186	143	80	1208	480	442	387
94	Allerton Bywater	2013	G	3371	1617	1333	664	2055	7861	7804	5611
52	Ardsley East/West	2010	P	63	30	22	4	249	58	51	19
78	West Park	2012	P	139	64	12	29	208	147	21	44
19	Rawdon	2016	G	1469	667	439	231	874	3032	1517	843
22	Alwoodley	2011	G	244	110	67	72	312	331	147	216
79	Whinmoor	2015	P	9	4	4	0	171	13	10	0
68	Halton Moor	2011	G	32	14	12	0	144	24	56	0
13	Wetherby	2014	P	676	289	95	222	647	936	269	393
34	Little London	2014	P	12	5	0	0	15	6	4	2
9	Scott Hall	2011	P	643	256	59	237	691	663	156	421
29	Harehills	2013	P	24	9	4	2	62	14	4	2
72	Cookridge	2015	G	717	258	207	96	615	480	664	301
20	Roundhay	2012	P	411	137	14	56	262	151	13	25
44	Burley	2015	P	37	11	4	4	84	18	5	8
12	Middleton	2013	P	894	254	125	192	705	477	271	504
6	Wortley	2015	G	1492	391	177	160	581	818	396	373
24	Fearnville	2013	P	415	97	26	60	358	375	58	160
80	Lofthouse & Robinhood	2011	G	105	23	27	1	23	40	65	2
69	Upper Wortley	2015	P	5	1	0	0	53	1	0	4
31	Chapelton	2011	P	137	25	8	16	110	32	12	44

37	Seacroft South	2011	P	211	29	7	3	17	48	10	6
43	Hunslet/Stourton	2017	G	109	14	7	9	896	29	9	23
76	Bramhope	2016	P	100	11	1	1	26	17	1	4
5	Holbeck	2013	P	210	20	7	12	62	36	9	24

Appendix B

The following peer-reviewed paper (Gulma et al., 2018), co-authored by the writer of this thesis has been published in *Transaction in GIS* journal DOI: 1111/tgis.12511. Part of the research has been adapted from chapter 5 of the thesis.

RESEARCH ARTICLE

Diversity and burglary: Do community differences matter?

Usman L. Gulma  | Andy Evans | Alison Heppenstall | Nick Malleon

School of Geography, University of Leeds,
Leeds, UK

Correspondence

Usman L. Gulma, School of Geography,
University of Leeds, University Road, Leeds
LS2 9JT, UK.

Email: gyulg@leeds.ac.uk

Funding

This study was funded by the Tertiary
Education Trust Fund (TETFUND) 2014/2015
grant as part of the PhD research project

Abstract

Diversity within a population has been linked to levels of both social cohesion and crime. Neighborhood crimes are the result of a complex set of factors, one of which is weak community cohesion. This article seeks to explore the impacts of diversity on burglary crime in a range of neighborhoods, using Leeds, UK as a case study. We propose a new approach to quantifying the correlates of burglary in urban areas through the use of diversity metrics. This approach is useful in unveiling the relationship between burglary and diversity in urban communities. Specifically, we employ stepwise multiple regression models to quantify the relationships between a number of neighborhood diversity variables and burglary crime rates. The results of the analyses show that the variables that represent diversity were more significant when regressed against burglary crime rates than standard socio-demographic data traditionally used in crime studies, which do not generally use diversity variables. The findings of this study highlight the importance of neighborhood cohesion in the crime system, and the key place for diversity statistics in quantifying the relationships between neighborhood diversities and burglary. The study highlights the importance of policy planning aimed at encouraging community building in promoting neighborhood safety.

1 | INTRODUCTION

The measurement of crime is necessary for any quantitative assessment of crime policy change (Ludwig & Marshall, 2015). Knowledge of how crime patterns are distributed over space can also enhance the effectiveness of police operations and collective community programs such as Neighbourhood Watch in the UK (Brunsdon, Corcoran,

& Higgs, 2007). For example, local knowledge of where crime is clustered will increase the capacity of the police to employ prevention measures, thereby improving the safety of communities (Bruce & Santos, 2011; Moore & Trojanowicz, 1988). Therefore, urban and regional planners, policy makers, and policing agencies have all recognized the importance of better understanding the dynamics of crime (Murray, McGuffog, Western, & Mullins, 2001). A common method of understanding both short-term crime and its longer-term drivers is through correlation with socio-economic and demographic factors in the areas where it occurs, an important component of environmental criminology (Andresen, 2014). Socio-economic and demographic factors such as wealth disparity, education attainment, proportion of young people, and deprivation are commonly found to correlate with crime rates in urban areas (Bandyopadhyay, Bhattacharya, & Han, 2010). Such variables act as proxies for, or direct measures of, the underlying causes of crime in a system that links offender drivers, victim lifestyles, and environment-related opportunities. However, the accuracy and representativeness of variables that act as proxies vary considerably, and many variables, such as metrics of multiple deprivation, could be seen as “catch-all” variables that encompass a wide variety of underlying factors. Additionally, traditional operationalizations of socio-demographic variables used when exploring explanations for neighborhood variation in crime rates do not measure diversity, and traditional measures of diversity in the crime context do not cover diversity across those various socio-demographic dimensions. The term “diversity” describes the level of variety in racial or ethnic composition, age, gender, religion, philosophy, physical abilities, socio-economic background, and sexual orientation among a group (Goodin, 2014, p. 7).

In this article, we will suggest that the treatment of standard regression variables can be adjusted to better capture a range of loci in which diversity plays a part across the crime system. In addition, we show that when these adjustments are made, these variables become more strongly predictive of crime than standard treatments, suggesting the significant part diversity plays in the crime system and the significant part it plays as the link between standard regression variables and crime rates. We will examine the relationship between crime and a series of standard socio-economic and demographic variables. We argue that such variables, while acting to represent components of the crime system, capture the effects of social cohesion acting within those components in a weak manner. In contrast, we generate a new set of alternative representations of these variables, centered on diversity statistics. For example, rather than looking at the percentage of a specific age group, we look at the diversity of ages within a community. We then include these statistics within a stepwise regression, along with the more standard metrics, to show their worth. As with most statistical treatments, ours is only a proxy for the real factors in the system, which, as we shall see, is multifaceted and sometimes contradictory. Nevertheless, we see that, empirically, diversity statistics must capture some elements of these complex relationships better than standard treatments. We consider diversity in a more general sense of variation within a population (such as age, education, employment, and family) beyond ethnic diversity, as is commonly used in this sense.

The article is organized as follows: Section 2 draws out some of these complexities from the literature on the relationship between diversity and crime (Section 2.1) and discusses the theoretical justification for the explanatory variables (Section 2.2); Section 3 describes the data, diversity statistics, and analysis approach; Section 4 presents the results of the analysis; while Section 5 discusses the findings and Section 6 provides some conclusions.

2 | LITERATURE REVIEW AND EXPLANATORY VARIABLES

2.1 | Exploring the relationship between diversity and crime

In the context of the UK and the USA, socio-economic and demographic diversity has been linked to decreased social cohesion and the variation of crime in neighborhoods (Sampson & Groves, 1989; Bursik Jr & Grasmick, 1993). Diversity may hinder informal communication within neighborhoods and tends to negatively affect the establishment of social interactions across groups (Browning, Burrington, Leventhal, & Brooks-Gunn, 2008; Laurence, 2011; Letki, 2008). In the Netherlands, for example, Meer and Tolsma (2014) have found that heterogeneity in a community leads to low levels of trust and meaningful interactions and tends to undermine intra-neighborhood

social cohesion. Employing the Metropolitan Police Public Attitude Survey (METPAS) of London, studies have found that ethnically diverse communities—especially with large transient populations—are often characterized by distrust, low levels of social cohesion, and high levels of dispute (Sturgis, Brunton-Smith, Kuha, & Jackson, 2014), with potential negative consequences for the individual as well as the community at large (Mellgren, 2011). Recent research in Japan consistently shows that areas characterized by ethnic diversity, wealth diversity, and age diversity (calculated at individual level with surveys) have high rates of crime (Takagi & Kawachi, 2014). However, this relationship needs to be investigated in the UK context.

Researchers have employed the social disorganization theory of Shaw and McKay (1942) to explain the variation in crime rate in different neighborhoods. Social disorganization theory posits that high levels of ethnic heterogeneity, residential instability, and socio-economic disadvantage undermine social networks, which in turn increases delinquency and crime rates (Shaw & McKay, 1942). Sampson and Groves (1989), in their empirical extension of Shaw and McKay's theory, showed that disorganization in a community lowers the ability of residents to work together toward problem solving, while collective efficacy among neighborhood residents mitigates crime rates (Sampson, Raudenbush, & Earls, 1997). Additionally, Kristjánsson (2007) stressed that weak networks of social ties decrease informal social control in the community, which increases deviant behavior.

Burglary, specifically, is a crime that thrives in socially disorganized and less cohesive communities (Weisburd & Piquero, 2008). It is likely that disorganized neighborhoods tend to have higher burglary crime rates because of weaker social cohesion than affluent areas where strong social connectedness facilitates the ability of residents to be on the lookout for criminal behavior (Dunaway, Cullen, Burton, & Evans, 2000). Routine activities theory (RAT) (Cohen & Felson, 1979) is regularly used by scholars to explain the occurrence of crimes such as burglary. This is based on the premise that a crime requires the simultaneous presence of three elements: motivated offenders, suitable targets, and the absence of capable guardians. A recent study shows that neighborhoods with diverse characteristics (occupation, education, income, ethnic, and residential instability), with low social cohesion and capable guardianship, may experience higher levels of burglary rates (Louderback & Sen Roy, 2018). Thus, social disorganization theory and RAT may help to explain the occurrence of neighborhood crime (Eck & Weisburd, 1995).

In a study of community integration in Berlin neighborhoods, Gruner (2010) employs Bourdieu's concept of "habitus" (socialized norms that guide behavior) to explain the distribution patterns of neighborhoods in terms of socio-economic and demographic structure. He found that the patterns and distributions of neighborhoods are associated with different cultural norms and the unwillingness of minority groups to integrate, describing it as self-segregation. Bourdieu's theory of habitus postulates the effects of physical embodiment of cultural capital: individuals who grow up in similar conditions develop similar habitus (Bourdieu, 1989). People with similar habitus feel attracted by and are more comfortable with each other (Bourdieu, 1989). The theory is extended to the study of social problems such as crime and perceived problems associated with migration flows in urban neighborhoods (e.g. Shammass & Sandberg, 2015). For example, growing up in a socially disorganized and crime-ridden neighborhood might greatly influence the behavior of people, especially the young (O'Connor, 2004), thereby facilitating delinquency (Tricia, 2016). This is especially pertinent to the local context in which this research is set. The conclusion from this set of studies is that high diversity in communities is acting to reduce social cohesion, consequently increasing neighborhood crime. There is some evidence that diversity reduces social cohesion (Meer & Tolsma, 2014) and this seems especially true where diversity is found in conjunction with deprivation (Cooper & Innes, 2009). However, it is important to note that diversity and cohesion levels are not always related in a simple manner, and that cohesion has the opportunity to be affected both positively and negatively, and by more than just ethnic or economic diversity (Ariely, 2014). Potentially, high diversity may have net positive impacts in terms of multiculturalism and the disruption of embedded cultural processes, despite negative impacts in other areas.

This article will model a number of socio-demographic factors and compare the impacts of standard variables representing those factors directly (for example, the proportion of young people) with variables representing their diversity (for example, age diversity). As study site, we focus on Leeds, UK, a city of approximately 750,000 people situated in the north of England (ONS, 2011).

2.2 | Theoretical justification for the explanatory variables

Although the relationship between community composition and crime is complex and multifaceted, there are some core factors that regularly emerge as important determinants of crime rates. This section will outline the most common factors used to explain variations in neighborhood crime rates; it is from these that the variables used in the later modeling work are derived. In each case, we will cover the more traditional variable, and then the diversity variable. The specific diversity equations used will be covered later.

2.2.1 | Age distribution

An age distribution is defined as the proportional numbers of persons in each age category in a given population. Previous studies have indicated that offenders are commonly drawn from younger age groups (Kongmuang, 2006). The age-crime curve tends to increase from the adolescent years, reaching a maximum at adulthood and then sharply declining (Blonigen, 2010; Farrington, 1986; Gottfredson & Hirschi, 1990; McCall, Land, Dollar, & Parker, 2013; McVie, 2005; Sampson & Laub, 2003; Sweeten, Piquero, & Steinberg, 2013), although this varies by the type of crime (Tittle, Ward, & Grasmick, 2003). Burglary is therefore likely to be affected by the absolute proportion of young (age 16–24) people. For example, according to Fagan and Western (2005), the incidence of crimes related to vehicles and drugs tends to be higher in early adulthood than in adolescence. Further, while homicides tend to be committed by adults, theft-related offences including burglaries are more prevalent in the younger age groups than the elderly (Loeber et al., 2012).

However, crime may also be affected by age distributions. A mixed population may put more or fewer offenders near more or fewer victims, but will also affect social cohesion. Younger people are less likely to build social cohesion (especially face to face) than older people (Johnston & Matthews, 2004; Takagi & Kawachi, 2014). Although offending is skewed toward the young because older adults have less opportunities for crime (Feldmeyer & Steffensmeier, 2007), the challenge is in accurately measuring the age effect on crime. Previous studies relied on raw numbers and proportions and did not use age standardization techniques (Hirschi & Gottfredson, 1983). Additionally, the tendency to commit crimes can change over time, regardless of age (Piquero, Farrington, & Blumstein, 2003). Recent comparative studies that used crime data from Taiwan and the USA found a considerable divergence from the age effect on crime (Steffensmeier, Zhong, & Lu, 2017). We therefore include population age diversity as a variable to investigate its relationship with crime, but with a prediction that different measures of diversity will identify different relationships. In this study, it is hypothesized that age diversity would be positively associated with burglary rates.

2.2.2 | Family structure

Family structure refers to whether the family unit includes children or not, both parents or a single parent. The family is generally regarded as an important social institution that shapes behavior, especially that of children (Nam, 2004). Maginnis (1997) has argued that the children of some single-parent families are more likely to have behavioral problems, because they tend to lack economic support and have lower parental input (Cheung & Park, 2016). In the UK, single parents continue to suffer from inequalities of employment and housing, creating a gap between couples and lone parents (Berrington, 2014). Additionally, single parents are also more likely to be victims of crime due to social marginalization in terms of living conditions (Wikström & Wikström, 2001). Given this, we include the proportion of single parents with children as an indicator from the traditional literature.

In terms of diversity, it seems likely that the distribution of family structures constitutes an important determining factor in social cohesion among community residents. For example, two-parent families with children tend to form social groups within the community that are distinct from single-parent families (Kanazawa, 2003; Sampson & Wooldredge, 1987). Community support within single-parenting groups is undoubtedly strong in some

areas, but is likely to be more geographically variable. The determinants of community support are largely the presence or absence of children, and the presence or absence of single parents, bearing in mind that the number of children is largely random in most populations (Umberson, Pudrovska, & Reczek, 2010) and ethnically controlled otherwise (Lee & McLanahan, 2015). Not having children encompasses populations that are both very young and very old, and little else (Rees & Butt, 2004). According to Tasgin and Morash (2016), different family characteristics (e.g. economically disadvantaged families and families with parents who have limited education) can have a negative impact on children's upbringing and behavior. Additionally, family indifference (lack of interest in a child's behavior), as characterized especially in communities with a diverse family structure, has been found to be a major cause of delinquency (Baek, Roberts, & Higgins, 2018; Bobbio, Lorenzino, & Arbach, 2016). Furthermore, family diversity may have a differential impact on urban crime rates, which suggests the need for including measures of family structure beyond traditionally used variables (such as the percentage of single parents) in urban crime studies (Parker & Johns, 2002). We offer diversity of family structure as a variable in the model based on these factors, with the hypothesis that it would be positively correlated with burglary rates.

2.2.3 | Ethnic identity

Ethnic identity is defined as the extent to which an individual identifies with an ethnic category (Chandra, 2006). Identity plays an important role in the likelihood that people will connect and form social relationships (Gilchrist & Kyprianou, 2011), and plays an important part in the integration of migrants into local neighborhoods (Kindler, Ratcheva, & Piechowska, 2014). Migrants—especially from black and minority ethnic (BME) populations—often lack the wealth, social integration, or formal crime-prevention connections to protect themselves (Sharp & Atherton, 2007). Because of these factors, the size of an immigrant population in an area positively correlates with the incidence of property crime (Bell & Machin, 2011). Empirical evidence from the US also demonstrates links between the size of an immigrant population and the occurrence of motor vehicle theft and robbery (Bholowalia & Kumar, 2014). However, previous studies in the UK have yet to empirically establish the link between increases in the size of the immigrant population in an area with the incidence of property crime specifically (Papadopoulos, 2014).

Ethnicity has a well-established relationship with crime (Tonry, 1997; Piquero & Brame, 2008; Unnever, 2018), principally acting through socio-economic exclusion and disadvantage. There are also biases in reporting, the justice system, and policing, the latter including complex relationships between race and prejudice, most clearly expressed in the findings of the Stephen Lawrence Inquiry in the UK (Macpherson, 1999). These issues seem entrenched. For example, police figures show that stop and search of white suspects increased by seven percentage points between 2009 and 2014 (from 68 to 75% of stops) and decreased by five percentage points for black suspects (from 17 to 12% of stops) (ONS, 2015). However, research has found that defendants from a BME background are more likely to be sent to prison compared to those from a white background (Kathryn, 2016). In this study we address the relationship between residential ethnicity and reported crime, ignoring the complex and important nuances of systemic biases, the statistical representations for which are largely unresolved.

In terms of diversity, ethnic diversity might relate to crime in different ways, including offending and victimization, including hate crime as a direct effect of ethnicity (Shepherd, 2006). Moreover, Vermeulen, Tillie, and van de Walle (2012) have argued that the negative effects of ethnic diversity on social networks would probably be stronger in terms of interpersonal trust, as well as differences in interests and needs between groups which weakens networks of social interaction. Though counter-arguments can be made in areas where everyone is essentially in a minority population, the nuances of tension and disadvantage in communities of multiple ethnicities and country of origin are likely to be complex. We therefore include diversity of country of origin to capture elements of isolation and integration, and diversity of ethnicity to capture the complex elements of offending and victimhood associated with ethnicity in mixed communities. We hypothesized that these sets of diversities would be positively associated with burglary rates.

2.2.4 | Employment and income

While there is a wide range of criminality across the socio-economic spectrum, for burglary, offenders in the vast majority of cases are drawn from the poor and unemployed (Bursik Jr & Grasmick, 1993; Sariaslan et al., 2013). Given this relationship, we include the level of unemployment in those age categories that could be working as a key variable.

Additionally, in Leeds the number of students is important because of the presence of large residential educational institutions (especially the two universities). Students are more likely to fall victims of crime, especially burglary, because multiple-occupancy homes are attractive to burglars, and because students are less likely to be at home (Kongmuang, 2006; Shepherd, 2006).

Furthermore, students' residences are attractive to burglars because students are more likely to possess valuable items, especially electronic gadgets (e.g. DVDs, laptops, iPads, and mobile phones), and be less careful about the security of their personal belongings (Barberet & Fisher, 2009). Additionally, some students reside in poor accommodation that lacks security surveillance devices such as closed circuit television (CCTV) and may not be adequately patrolled (Masike & Mofokeng, 2014). Wealth diversity within a community may act to increase crime. Given that most burglars only travel a short distance to commit crimes (Ashby, 2005), there is some evidence that disparities of wealth within short distances encourage burglary (Chiu & Madden, 1998; Rufrancos, Power, Pickett, & Wilkinson, 2013; Tseloni, Osborn, Trickett, & Pease, 2002). In addition, disparity of wealth within a community can influence crime by weakening social cohesion (Fajnzlber, Lederman, & Loayza, 2002; Rufrancos et al., 2013). Equally, low wealth diversity can enhance social cohesion (Cooper & Innes, 2009). Although the picture is complicated across other types of crime (Rufrancos et al., 2013), researchers have found support for the relationship between property crime and income inequality (Demombynes & Özler, 2005; Kelly, 2000; Reilly & Witt, 2008; Witt, Clarke, & Fielding, 1998). We therefore include diversity of employment type in our assessment, as a proxy for wealth, in the absence of a household income variable not captured in the UK census statistics (House of Commons, 2011). We hypothesized a positive relationship between employment diversity and burglary rates.

2.2.5 | Deprivation

Deprivation has been defined as a lack of resources to meet the basic necessities of life (DCLG, 2015). Literature on the relationship between deprivation and crime suggests that deprived communities tend to have more crime than affluent communities (Sampson & Wooldredge, 1987; Bursik Jr & Grasmick, 1993; Krivo & Peterson, 1996; Malczewski & Poetz, 2005). Furthermore, deprivation widens the gap between rich and poor, which can reduce social cohesion (Morenoff, Sampson, & Raudenbush, 2001; Takagi & Kawachi, 2014). The Index of Multiple Deprivation is a multidimensional metric that is measured, in England, through a combination of seven distinct domains: income, employment, education, health, crime, barriers to housing & services, and living environment (DCLG, 2015).

Although deprivation is often seen as a key indicator of social cohesion, as well as propensity to commit a crime such as burglary, UK deprivation statistics include crime and therefore it is inappropriate to use them in this analysis. Deprivation is covered by the other variables, as far as demographics are concerned.

2.2.6 | Educational attainment

Educational attainment has a great influence on individuals' social behavior, as well as on participation in community activities (Sabates, 2008). The theory of human capital suggests that skills and qualifications determine wages, and the wider the distribution of qualifications, the wider the distribution of wages (Green, Preston, & Janmaat, 2006). Reynolds, Temple, Robertson, and Mann (2001) found that the propensity of individuals to commit crime is associated with their level of educational attainment, and so we include lack of qualifications as a traditional variable.

However, there is also a likely indirect relationship between educational inequality and crime. Sabates, Feinstein, and Shingal (2008) found that educational inequality is associated with violent crime. While it is unclear whether such a relationship acts at the intra-area scale independent of any effect of wealth, we include a diversity statistic centered on education to test the potential relationship and hypothesized a positive association with burglary rates.

2.2.7 | Residential instability

Residential instability has been defined as two or more residential moves within the course of one year (Foulkes & Newbold, 2008). Residential stability in a neighborhood is an important factor for the generation of social capital and place-based attachment, so it is expected that residential duration will affect crime via this effect on social cohesion (Thomas, Stillwell, & Gould, 2016). Studies have demonstrated that the creation of social ties is associated with the length of residence in an area. For example, Yamamura (2011) argued that personal relationships are built over time, tend to be more solid when people reside in a particular neighborhood, and are influenced by length of residence and home ownership. Similarly, Keene, Bader, and Ailshire (2013) point out that it takes time to create supportive social ties, therefore length of neighborhood residency may be an important determinant of social integration. Additionally, Oh (2003) shows that length of residence has a positive effect on friendships, social cohesion, and trust, which also enhance the probability of working together to solve local problems. In contrast, residential instability in a neighborhood is associated with weak social ties and a low probability of residents connecting (Sampson et al., 1997).

Crime is also more likely to occur in transient neighborhoods. For example, in the UK, the tendency to commit crime is related to length of residence, in other words, crime reduces as length of residence in a neighborhood increases (Bell & Machin, 2011). Residential instability also influences crime from the social disorganization perspective (Shaw & McKay, 1942). Specifically, research has established the relationship between residential instability and violent crime (e.g. Boggess & Hipp, 2010). However, the relationship between residential instability and burglary is likely to be complex (Markowitz, Bellair, Liska, & Liu, 2001; Martin, 2002). Given the complexity of the relationship, we include length of residence less than two years as a standard variable and diversity of length of residence as a proxy for residential instability. The hypothesis is that a positive association would be expected with burglary rates.

3 | DATA, METHODS, AND ANALYSIS

This section describes first the study area and the data used for the study. Measures of diversity statistics and the analysis approach for the study are then provided. Figure 1 shows our methodology workflow diagram.

3.1 | Study area

The city of Leeds in the north of England (UK) is a medium-sized post-industrial city of ~750,000 people. It comprises 33 wards, which are divided into 482 lower super output areas (LSOAs). Figure 2 shows the location of Leeds. The LSOAs have a minimum population of 1,000 (with an average of 1,500) and a minimum resident household number of 400 (with an average of 630) (ONS, 2011). It contains some of the poorest wards in England (DCLG, 2015), but equally has wards containing the homes of some of the most affluent individuals in the country (BBC, 2003). Leeds is an area with an increasing number of BME groups. For example, in the 2001 population census the population of BME groups was 77,530 (about 10.8% of the resident population), and this increased to 141,771 (representing 18.9% of the resident population) by 2011 (ONS, 2011). The city also has a relatively large number of burglaries (12.83/1,000 population) compared to the national average (7.5/1,000 population) (ONS, 2017) and

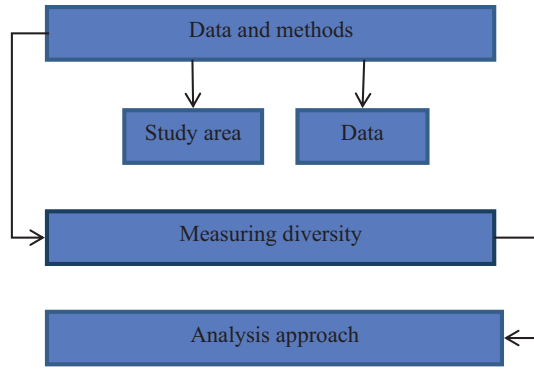


FIGURE 1 Methodology workflow diagram

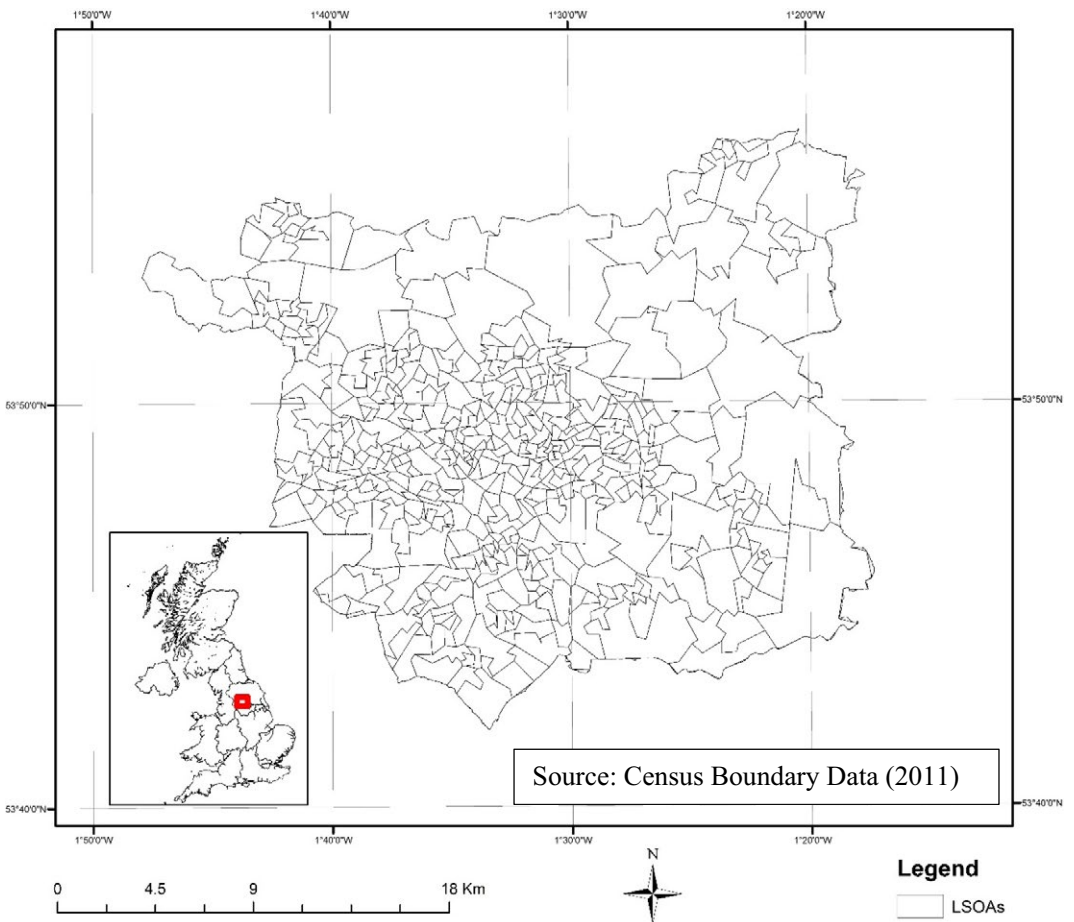


FIGURE 2 Location map of Leeds with LSOA boundaries

characteristically different types of neighborhood, which makes it suitable for examining relationships between socio-economic and demographic diversity and burglary (Hirschfield, Birkin, Brunson, Malleson, & Newton, 2013).

3.2 | Data

Burglaries reported between 2011 and 2015 in the city of Leeds were obtained from the “police open public monthly data of reported crimes” (<https://data.police.uk/data/>), a portal that provides customized crime data downloads for all police forces in England and Wales. In this case, West Yorkshire Police for the period 2011–2015 ($n = 51,800$). Rates per 1,000 population were then calculated over the whole data for each of the 482 LSOAs of Leeds. LSOA geography has been chosen because it is small enough to capture neighborhood effects but large enough to represent coherent community groups. The remaining data (age distribution, family structure, identity, employment, educational attainment, and length of residence) were derived from UK 2011 census data, supplied by the UK Data Service (downloaded from <http://infuse.ukdataservice.ac.uk/>). Figures 3a and b show the spatial distribution of independent variables (standard and diversity) in the study area.

3.3 | Measuring diversity statistics

Researchers have used a number of methods to measure diversity (Morris et al., 2014). Nevertheless, in this initial study we concentrate on diversity indices, which report the probability that two individuals taken at random are different. Such diversity indices therefore use equivalent classes weighted on the same scale, irrespective of the total community size, with each class within the community having members that share common attributes (Jost, 2006). The most widely used diversity index is Simpson’s (1949) diversity index (D) (Johnson & Lichter, 2010). The range of values of D is 0 to 1; values toward 0 indicate no diversity and values toward 1 indicate the presence of absolute diversity. Simpson’s diversity index for area i is written as:

$$D_i = 1 - \frac{\sum_i n_i(n_i - 1)}{N(N - 1)} \quad (1)$$

where n_i is the proportion of a population in an area falling into category i and N is the total population of that area.

3.4 | Analysis approach

In this analysis, the categories were determined by census data availability. Table 1 shows the core components of crime and community cohesion, and the variables used to represent them in the model. Table 2 shows the different categories included to measure diversity. Utilizing the variables in Table 1, we constructed a model of correlates with crime. Identifying the best model fit requires an iterative process that examines different combinations of explanatory variables. Exploratory regression analysis is important for selecting the best explanatory variables for a given model (Braun & Oswald, 2011). Exploratory regression builds ordinary least square (OLS) models using all possible combinations of explanatory variables and assesses which models pass the OLS checks (Rosenshein, Scott, & Pratt, 2011). This process is useful for ensuring that only variables with the highest significance are retained. Here, to test the strength of the relationship between the variables and crime, we utilize stepwise (a combination of forward and backward selection) linear regression. Stepwise methods are commonly used to select the best variables in a regression model, especially multiple regressions with many predictors such as in this study (Sinha, Malo, & Kuosmanen, 2015; Wooldridge, 2012). However, the process of adding and dropping variables associated with stepwise regression has been criticized in that it is possible to miss the optimal model, as removing less significant predictors increases the significance of the others, which may lead researchers to overstate the

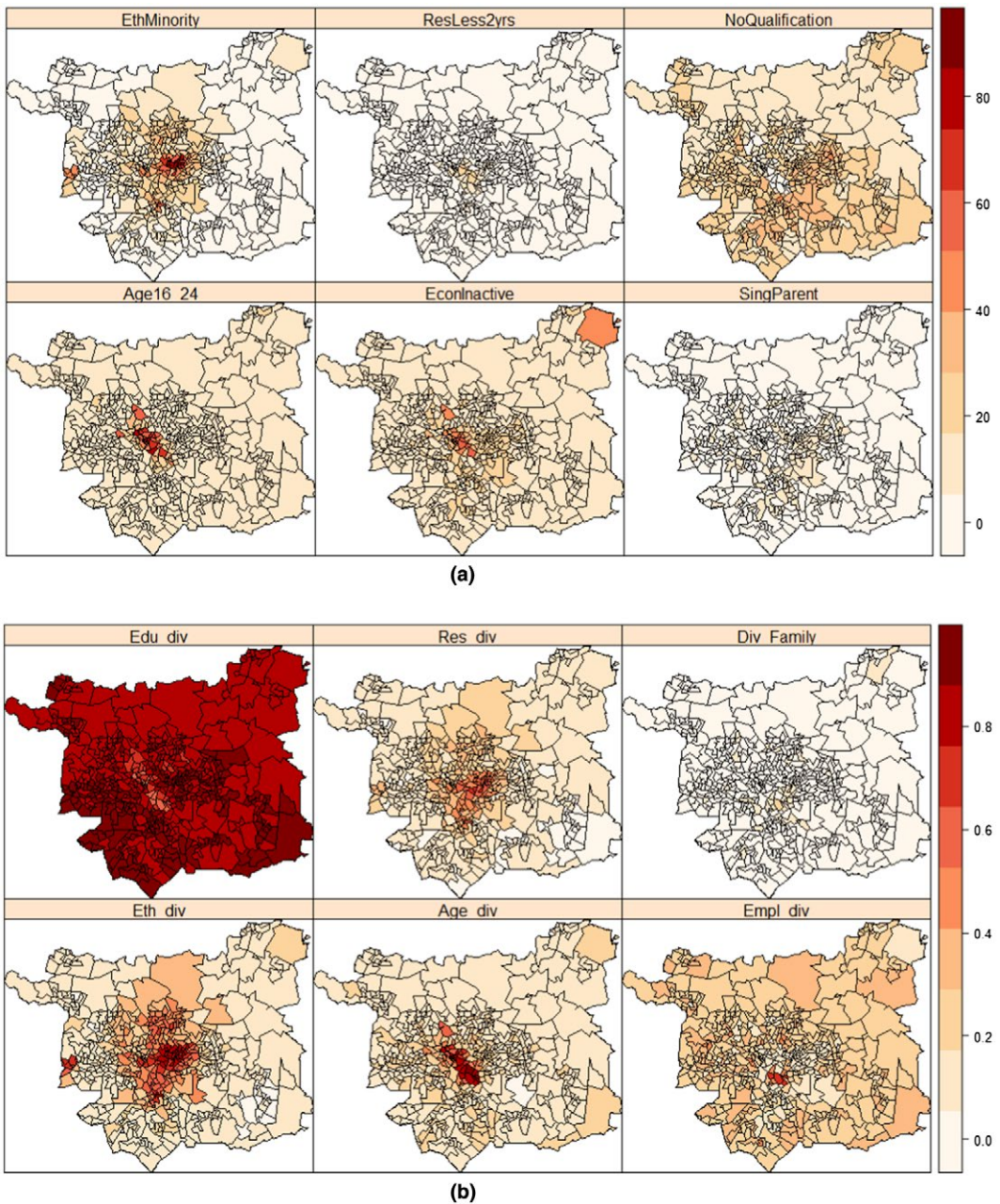


FIGURE 3 (a, b) Spatial distributions of standard and diversity metrics. Source: ONS (2011)

importance of the remaining variables (Rawlings, Pantula, & Dickey, 1998). Despite the limitations of the stepwise multiple regression method, it is widely used in different ecological studies (Caplan, Kennedy, Barnum, & Piza, 2015; Collins, Babyak, & Moloney, 2007; Meera & Jayakumar, 1995; Pitner, Yu, & Brown, 2012; Raftery, Madigan, & Hoeting, 1997).

In this study, the model was built by sequentially adding significant ($p \leq 0.05$) variables; the order of correlation between the dependent variables determines the order by which they are added to the model. The stopping

TABLE 1 The core components of crime and community cohesion, and the variables used to represent them in the model

Component	Standard variable	Diversity variable
Age distribution	Number of young persons (16–24)	Age diversity
Family structure	Lone parents	Diversity of family structure
Identity	Ethnic minority population	Ethnic diversity
Employment/income	Age 16–64, economically inactive	Diversity of employment type
Educational attainment	Age 16 over, no qualification	Diversity of educational attainment
Residential instability	Resident less than 2 years	Length of residence diversity

TABLE 2 Components used to measure different diversity metrics

Diversity	Components included
Age	10–14, 15, 16–17, 18–19, 20–24, 25–29, 30–44, 45–59, 60–64, 65–74
Family structure	Lone parent, no dependent; Lone parent, one dependent child; Lone parent, two or more dependent children; Married couple, no children; Married couple, one dependent child; Married couple, two or more children
Ethnicity	All 18 ethnic groups included
Employment	16–64 Managers/Directors, 16–64 Professionals, 16–64 Associate Professionals, 16–64 Administration and Secretariat, 16–64 Skilled Trade, 16–64 Caring, Leisure and Services, 16–64 Customer Services, 16–64 Process, Plants and Machines, 16–64 Elementary Occupation
Education	16-over qualification level 1, 16-over qualification level 2, 16-over qualification level 3, 16-over qualification level 4
Residence length	Length of residence: Less than 2 years; Less than 5 years; More than 5 years; 10 years and above; Born in the UK

criterion for the stepwise process is reached when none of the remaining variables are significant ($p \geq 0.1$), then the process terminates. We first included a model using only standard variables and subsequently compared that to one including all variables (standard and diversity).

Equation (2) for linear multiple regression is given based on Charlton, Fotheringham, and Brunson (2009). The stepwise method adds variables (standard and diversity in this context) to the model through a series of iterations and ensures that the variables are still significant contributors to the model, removing those which are not:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + \varepsilon \quad (2)$$

Here, Y is the value of the dependent variable, β_0 is the constant intercept, $\beta_1, \beta_2, \beta_3, \dots, \beta_n$ are the slope coefficients of $X_1, X_2, X_3, \dots, X_n$ and $X_1, X_2, X_3, \dots, X_n$ are the independent variables, while ε is the standard error of coefficients. The standard error is calculated by summing the squared values of the residuals and dividing by the difference between the number of parameters subtracted from the total number of observations.

Optimal models are a balance of correlation against parsimony. Although such balances are largely subjective and centered around use cases, traditionally scree graphs have been used to help in the decision making as there is often a natural kink in the graph of, for example, R^2 versus number of model components, which indicates considerably decreasing explanatory power being provided by additional components (Mehmood, Martens, Sæbø, Warringer, & Snipen, 2011; Preacher, 2006).

Prior to building the model, the independent variables were tested for multicollinearity. Multicollinearity is present when there is a high degree of correlation among independent variables. This can significantly affect

TABLE 3 Model summary of stepwise regression

(a)	Model	R	R ²	Adjusted R ²	Std. error of the estimate
	1	0.302	0.091	0.089	32.729
	2	0.337	0.114	0.110	32.357
	3	0.359	0.129	0.123	32.115
(b)	Model	R	R ²	Adjusted R ²	Std. error of the estimate
	1	0.358	0.128	0.126	32.061
	2	0.447	0.200	0.196	30.746
	3	0.469	0.220	0.215	30.384
	4	0.481	0.232	0.225	30.191
	5	0.492	0.242	0.235	30.010

model performance and reliability (Wang, 1996). There is no standard rule for filtering out variables based on the issue; here, correlations of $r > 0.70$ are regarded as very significant. The proportion of *economically inactive population* correlates with the *proportion of young persons aged 16–24* (0.876); the *proportion of single parents* correlates with the *proportion of persons with no qualification* (0.709); and *ethnic diversity* correlates with *length of residence diversity* (0.971). Therefore, the following variables were removed to avoid redundancy and because of their relatively lower correlation with the dependent variable: *young persons aged 16–24* (correlation with burglary rate: young persons (0.269) compared to economically inactive population (0.276)); *single parents* (constitutes a larger proportion of those with no/lower qualifications in the UK); and *length of residence diversity* (also a useful indicator of ethnicity and has weaker correlation with burglary (0.310) compared to ethnic diversity (0.313)). No significant correlation was found between the standard variables and their diversity equivalents.

4 | RESULTS

Tables 3a and b summarize the results of the standard and combined stepwise regression models used to assess the relative importance of each variable in the models. The statistics reported are Pearson's product moment correlation (r), which shows the correlation with the dependent variable for each model. R^2 reports the percentage of variation in rate of burglary crime explained by the variables used in the model. Adjusted R^2 is the fraction by which the square of the standard error of the regression is less than the variance of the dependent variable. It increases only if the variables improve the model. It is usually used to evaluate which model performs better, where a model with a smaller standard error of estimate is likely to produce a higher adjusted R^2 (Kongmuang, 2006). In the combined analysis, model 5 is the best-performing model, represented in the form of Equation (3), while in the standard variables analysis, model 3 performed best, explaining 14% of the variation in burglary rates. Model 5 will be the subject of discussion in Section 6.

Tables 4a and b present the coefficients of the standard and combined models. The elements reported are standardized and unstandardized coefficients, standard error, t statistics and significance tests. In regression analysis, standardized coefficients are estimates standardized so that the variance of the dependent variable produced by changes in the independent variables is between -1 and 1 ; unstandardized coefficients express values of the relationship in raw values (Landis, 2005). The standard error is a measure of the accuracy of predictions obtained from the difference between the observed and predicted values; smaller values indicate that observations are closer to the fitted regression line (Altman & Bland, 2005). The stepwise regression results indicate that the parameters are within acceptable standards for regression modeling. An important guide for understanding this are the t statistics (Dunn, 1989). A t statistic is the estimated coefficient divided by its own standard error.

TABLE 4 Coefficients and tests of model performance

(a)	Model	Unstandardized coefficients		Standardized coefficients		
		B	Std. error	Beta	t	Sig.
1	(Constant)	63.411	1.700		37.308	0.000
	Length of residence less than 2 years(%)	3.893	0.561	0.302	6.945	0.000
	(Constant)	50.232	4.146		12.116	0.000
2	Length of residence less than 2 years(%)	4.327	0.568	0.336	7.616	0.000
	Age 16 over, no qualification(%)	0.651	0.187	0.153	3.477	0.001
3	(Constant)	37.493	4.957		7.564	0.000
	Length of residence less than 2 years(%)	2.860	0.646	0.222	4.429	0.000
	Age 16 over, no qualification(%)	0.955	0.196	0.225	4.880	0.000
	Age 16–24(%)	0.630	0.140	0.240	4.491	0.000
(b)	Model	Unstandardized coefficients		Standardized coefficients		
		B	Std. error	Beta	t	Sig.
1	(Constant)	53.775	2.336		23.018	0.000
	Age diversity	74.289	8.851	0.358	8.393	0.000
2	(Constant)	23.531	5.130		4.587	0.000
	Age diversity	102.110	9.490	0.492	10.759	0.000
	Age 16 over, no qualification(%)	1.273	0.194	0.300	6.554	0.000
3	(Constant)	22.092	5.086		4.344	0.000
	Age diversity	86.719	10.341	0.418	8.386	0.000
	Age 16 over, no qualification(%)	1.145	0.195	0.270	5.867	0.000
	Ethnic diversity	24.609	6.964	0.157	3.534	0.000

(Continues)

Table 4 (Continued)

(b)	Model	Unstandardized coefficients			Standardized coefficients		
		B	Std. error	Beta	t	Sig.	
4	(Constant)	101.265	30.075		3.367	0.001	
	Age diversity	60.622	14.180	0.292	4.275	0.000	
	Age 16 over, no qualification(%)	1.344	0.208	0.316	6.469	0.000	
	Ethnic diversity	28.885	7.102	0.185	4.067	0.000	
	Educational diversity	-93.151	34.882	-0.181	-2.670	0.008	
5	(Constant)	106.442	29.962		3.553	0.000	
	Age diversity	85.607	17.060	0.412	5.018	0.000	
	Age 16 over, no qualification(%)	1.509	0.216	0.355	6.985	0.000	
	Ethnic diversity	32.582	7.202	0.208	4.524	0.000	
	Educational diversity	-97.983	34.723	-0.190	-2.822	0.005	
	Age 16-64, economically inactive(%)	-0.675	0.260	-0.163	-2.600	0.010	

Dependent variable: Burglary rate.

Significant *t* statistics should be approximately 1.96 in magnitude, corresponding to a *p* value less than 0.05% or 95% confidence level (Coe, 2002). The results obtained from the stepwise regression in this analysis indicate that the values of the *t* statistics for all variables in the models were greater than 1.96, meaning that all variables are statistically significant. At each iteration of the stepwise regression, variables that are not significant are dropped and model variables that are significant are retained.

It is common practice to assess the appropriateness of a model using the coefficient of determination, although this is not an absolute indicator of goodness of fit (Reisinger, 1997) and a low effect size does not mean that the model is inefficient (Martin, 2014; Weisburd & Piquero, 2008). Although the analysis explained approximately 24% of the variation in burglary crime, that is good compared to other studies: Zhao, Lawton, and Longmire (2015), Karyda (2015), Hino, Uesugi, and Asami (2016), and Boateng (2016) have models explaining 21%, 10%, 14%, and 12%, respectively. Crime, especially burglary, is difficult to understand, predict, and model (Malleeson & Birkin, 2012). The percentage of variation of the dependent variable explained in a model can sometimes be misleading, as small effect sizes can produce better and more meaningful outcomes than larger ones (Lieberson, 1985). However, this depends on the unit of analysis, the type of crime, and the underpinning theory (Weisburd & Piquero, 2008).

The final regression equation (model 5) is given here by computing the values of the unstandardized (B) coefficients:

$$\begin{aligned} \text{Burglary rate} = & 106.442 + 85.606 * \text{Age diversity} + 1.509 * \\ & \% \text{Age 16 over, no qualification} + 32.582 * \text{Ethnic diversity} + -97.983 * \\ & \text{Educational diversity} + -0.675 * \% \text{Age 16 to 64, economically inactive} \end{aligned} \quad (3)$$

5 | DISCUSSION

The most notable result of the above analysis is the almost complete exclusion of standard variables in preference for diversity statistics (Table 4a). As seen from the unstandardized (B) coefficients of the standard only (Table 4a) and combined variables (Table 4b) models, diversity variables have shown a higher relationship in explaining burglary rates than the standard variables. Additionally, the order of the variable correlation with the dependent variables (except for the proportion of those with no educational qualification, which is the second most important variable in both models) also indicated that the diversity variables are more important. The results highlight the importance of diversity in the crime system, with a concomitant suspicion that this acts through community cohesion, but also highlight that the standard statistics probably, in part, represent community cohesion, and are being excluded here simply because the new metrics are potentially stronger correlates of burglary rates. It could equally be that diversity indicates the proximity of "haves" and "have nots" and opportunity/targets within a community, as mechanisms by which diversity impacts on crime.

In this study, diversity of age was the most important variable when regressed against the dependent variables consistently throughout the models (see model coefficients in Table 4a). As hypothesized, age diversity was significant ($p < 0.01$) and positively associated with burglary rates. Age diversity has shown that offenders are commonly drawn from younger age groups rather than elderly people, and the findings in this study are consistent with previous literature which found a relationship between age and crime (e.g. Farrington, 1986; Gottfredson & Hirschi, 1990; Sampson & Laub, 2003; McVie, 2005; Blonigen, 2010; McCall et al., 2013; Sweeten et al., 2013). However, it is likely that a wide age range puts young offenders in close proximity with older victims with, potentially, more to steal. Equally, however, we know that the young are also targets for crime (Finkelhor, Ormrod, Turner, & Hamby, 2005), and it makes some sense that the broader the range of population characteristics in an area, the more likely that there will be suitable target criteria for burglars making decisions about risk (Bernasco & Nieuwebeerta, 2005).

In this study, unexpectedly, *diversity* of educational attainment was significant ($p < 0.05$) and negatively correlated with burglary rates, meaning that the smaller the diversity of educational attainment, the more burglary occurs in an area. This finding should be interpreted with caution, as there are sophisticated crimes (such as cybercrimes) that are perpetrated by educated individuals. Although previous studies have found that educational attainment increases returns through legitimate means (Green et al., 2006), it also raises the opportunity cost of illegal behavior (Machin, Marie, & Vujić, 2011). Consistent with previous studies, we also found a significant ($p < 0.01$) positive relationship between the proportion of those with no educational qualification and burglary (Machin et al., 2011).

We also found strong support for a positive relationship between ethnic diversity and rates of burglary crime. This finding contradicts Papadopoulos (2014), who found no significant relationship between an increase in the size of the immigrant population and property crime. The findings of this study, however, are consistent with the findings of previous studies, which found a positive relationship between the size of the immigrant population in an area and the incidence of property crime (e.g. Bell & Machin, 2011). Previous research has shown that ethnically heterogeneous communities are often characterized by distrust, low levels of social cohesion, and disputes (Sturgis et al., 2014), which negatively affect individual behaviors (Mellgren, 2011). Recent studies into the spatial distribution of neighborhood crime consistently show that areas which are characterized by ethnic diversity have high rates of crime (Gartner, 2013; Takagi & Kawachi, 2014). However, the significant positive relationship found in this study could also be because migrants often lack formal crime prevention connections to protect themselves against crime victimization (Sharp & Atherton, 2007).

The feeling of disparity between wealthy and poor people increases antagonism, with a resultant increase in crime (Fajnzlber et al., 2002; Ruffancos et al., 2013). Disparity within an area also, however, implies a potential mix of richer targets and poorer offenders within the area. Given that burglars tend to be poor, and have a fairly short travel distance (see above), more diverse communities may have more targets (Demombynes & Özler, 2005; Kelly, 2000; Reilly & Witt, 2008; Witt et al., 1998). Nevertheless, in this study we found no statistically significant relationship between diversity of employment and burglary crime rate. Further study is needed to explore this relationship.

In this study, we found a significant negative correlation between the proportion of economically inactive population and burglary crime, which might be seen as counter-intuitive. Previous studies have found support for relationships between income inequality and property crime (Demombynes & Özler, 2005; Kelly, 2000; Reilly & Witt, 2008; Witt et al., 1998). However, the difference between measuring offences committed by those residing in a community and measuring offences occurring in a community could be a reason for the following preposition; this relationship might only suggest that unemployment might contribute to offending elsewhere. Recent statistics in the UK show that economically inactive people are twice as likely to be victims of burglary crime as those who are economically active (ONS, 2014), considering that this category of population comprises students, those who are retired, and people with long-term health challenges, the relationship for Leeds needs further investigation.

6 | CONCLUSIONS

This study explored the impact of diversity on burglary crime in the Leeds district, UK. We used stepwise regression models to assess the relationships between both standard and diversity-based socio-demographic variables and burglary crime rate. We showed that diversity-based statistics are a better correlate with crime than most standard metrics, highlighting the importance of diversity in the crime system, and suggesting the potential importance of social cohesion in preventing crime. It seems likely that standard statistics go some way, normally, to explaining neighborhood variations in burglary, but that this is better captured through diversity statistics.

The variables used in this study have provided useful insights into the relationship between neighborhood social context (diversity) and the spatial variability of burglary rates in Leeds. The most important predictor for modeling burglary crime rates in this analysis was age diversity. However, other predictors—such as ethnic diversity, distribution of educational attainment, proportion of those with no educational qualifications, and proportion of economically inactive population—also made a valuable contribution to the models. Notably, economically inactive population had a slight negative relationship with crime, and this needs further investigation.

It seems likely that community cohesion is an important factor in establishing social control and collective efficacy in neighborhoods with regard to crime. Here we have used a simple set of diversity statistics to highlight the possibilities for investigating this. However, there is scope, having identified the importance of diversity statistics, to investigate alternative metrics in this area to reveal different aspects of community cohesion—for example, it may be that age distributions are better represented by statistics which utilize the frequency distribution of the population in a more nuanced fashion than the standard Simpson's diversity index. As this study considered burglary crime rates, we also recommend future research to consider applying the present approach against other types of crime, in order to uncover relationships between crime and diversity metrics.

The results obtained in this study are potentially useful in prioritizing areas of policy planning for crime prevention. The study suggests that in terms of crime prevention alone, there is a need for extra support in areas dedicated to encouraging community building, rather than poverty specifically being the key, at least in Leeds¹—although clearly poverty is at the root of additional social issues.

ACKNOWLEDGEMENTS

The authors wish to acknowledge and thank West Yorkshire Police and the UK data service whose data were used in this study.

CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

AUTHORSHIP

Usman Lawal Gulma: Conceived the idea of the study, made substantial contributions in data acquisition, analysis and interpretation of the results.

Andy Evans: Critically reviewed the paper to ensure quality. Alison Heppenstall. Sufficiently contributed in revising the paper and made valuable comments.

Nick Malleson: Contributed by ensuring that issues related to accuracy of the work are appropriately followed.

NOTE

¹It is worth noting, in this respect, that financial gain was the dominant factor identified across all burglars in recent interviews; however, over a fifth of offenders in Leeds talked about how they will also offend “for the buzz” it provides them (N. Addis, pers. comm., 2016). This may not be the case in other areas, where poverty may be more of a direct driver.

ORCID

Usman L. Gulma  <http://orcid.org/0000-0002-2986-6700>

REFERENCES

- Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *BMJ*, 331(7521), 903.
- Andresen, M. A. (2014). *Environmental criminology: Evolution, theory, and practice*. Abingdon, UK: Routledge.
- Ariely, G. (2014). Does diversity erode social cohesion? Conceptual and Methodological Issues. *Political Studies*, 62(3), 573–595.
- Ashby, D. I. (2005). Policing neighbourhoods: Exploring the geographies of crime, policing and performance assessment. *Policing & Society*, 15(4), 413–447.
- Baek, H., Roberts, A. M., & Higgins, G. E. (2018). The impact of family indifference on delinquency among American Indian youth: A test of general strain theory. *Journal of Ethnicity in Criminal Justice*, 16(1), 57–75.
- Bandyopadhyay, S., Bhattacharya, S., & Han, L. (2010). *Determinants of violent and property crimes in England and Wales: A panel data analysis*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1691801
- Barberet, R., & Fisher, B. S. (2009). Can security beget insecurity? Security and crime prevention awareness and fear of burglary among university students in the East Midlands. *Security Journal*, 22(1), 3–23.
- BBC. (2003). *Northern tops 'real' rich league (Carole Mitchell report)*. Retrieved from <http://news.bbc.co.uk/1/hi/business/3025321.stm#text>
- Bell, B., & Machin, S. (2011). *The impact of migration on crime and victimization: A report for the Migration Advisory Committee*. Retrieved from <https://www.gov.uk/government/publications/impact-of-migration-on-crime-and-victimisation>
- Bernasco, W., & Nieuwebeerta, P. (2005). How do residential burglars select target areas? A new approach to the analysis of criminal location choice. *British Journal of Criminology*, 45(3), 296–315.
- Berrington, A. (2014). *The changing demography of lone parenthood in the UK* (Working Paper No. 48). Southampton, UK: University of Southampton, ESRC Centre for Population Change.
- Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and Kmeans in WSN. *International Journal of Computer Applications*, 105(9), 17–24.
- Blonigen, D. M. (2010). Explaining the relationship between age and crime: Contributions from the developmental literature on personality. *Clinical Psychology Review*, 30(1), 89–100.
- Boateng, F. D. (2016). Fearfulness in the community empirical assessments of influential factors. *Journal of Interpersonal Violence*, 0886260516642295.
- Bobbio, A., Lorenzino, L., & Arbach, K. (2016). Family, neighborhood, and society: A comparative study carried out in Argentina among youth with and without criminal backgrounds. *Revista Criminalidad*, 58(1), 81–95.
- Boggess, L. N., & Hipp, J. R. (2010). Violent crime, residential instability and mobility: Does the relationship differ in minority neighborhoods? *Journal of Quantitative Criminology*, 26(3), 351–370.
- Bourdieu, P. (1989). Social space and symbolic power. *Sociological Theory*, 7(1), 14–25.
- Braun, M. T., & Oswald, F. L. (2011). Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behavior Research Methods*, 43(2), 331–339.
- Browning, C. R., Buntington, L. A., Leventhal, T., & Brooks-Gunn, J. (2008). Neighborhood structural inequality, collective efficacy, and sexual risk behavior among urban youth. *Journal of Health & Social Behavior*, 49(3), 269–285.
- Bruce, C., & Santos, R. B. (2011). *Crime pattern definitions for tactical analysis* (International Association of Crime Analysts White Paper). Retrieved from http://www.iaca.net/Publications/Whitepapers/iacawp_2011_01_crime_patterns.pdf
- Brunsdon, C., Corcoran, J., & Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment & Urban Systems*, 31(1), 52–75.
- Bursik, R. J. Jr, & Grasmick, H. G. (1993). Economic deprivation and neighborhood crime rates, 1960–1980. *Law & Society Review*, 27, 263.
- Caplan, J. M., Kennedy, L. W., Barnum, J. D., & Piza, E. L. (2015). Risk terrain modeling for spatial risk assessment. *Cityscape*, 17(1), 7.
- Chandra, K. (2006). What is ethnic identity and does it matter? *Annual Reviews in Political Science*, 9, 397–424.
- Charlton, M., Fotheringham, S., & Brunsdon, C. (2009). *Geographically weighted regression* (White Paper). Maynooth, UK: National Centre for Geocomputation, National University of Ireland Maynooth.
- Cheung, A.-K.-L., & Park, H. (2016). Single parenthood, parental involvement and students' educational outcomes in Hong Kong. *Marriage & Family Review*, 52(1&2), 15–40.
- Chiu, W. H., & Madden, P. (1998). Burglary and income inequality. *Journal of Public Economics*, 69(1), 123–141.
- Coe, R. (2002). *It's the effect size, stupid*. Paper presented at the British Educational Research Association Annual Conference, University of Exeter, England.
- Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588–608.
- Collins, K., Babyak, C., & Moloney, J. (2007). *Spatial modeling of geocoded crime data* (Upcoming Methodology Branch Working Paper). Ottawa, ON: Statistics Canada.

- Cooper, H., & Innes, M. (2009). *The causes and consequences of community cohesion in Wales: A secondary analysis*. Cardiff, UK: Universities' Police Science Institute, Cardiff University.
- DCLG. (2015). *English indices of deprivation*. Retrieved from <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>
- Demombynes, G., & Özler, B. (2005). Crime and local inequality in South Africa. *Journal of Development Economics*, 76(2), 265–292.
- Dunaway, R. G., Cullen, F. T., Burton, V. S., & Evans, T. D. (2000). The myth of social class and crime revisited: An examination of class and adult criminality. *Criminology*, 38(2), 589–632.
- Dunn, R. (1989). A dynamic approach to two-variable color mapping. *The American Statistician*, 43(4), 245–252.
- Eck, J. E., & Weisburd, D. (1995). Crime places in crime theory. *Crime & Place, Crime Prevention Studies*, 4, 1–33.
- Fagan, A. A., & Western, J. (2005). Escalation and deceleration of offending behaviours from adolescence to early adulthood. *Australian & New Zealand Journal of Criminology*, 38(1), 59–76.
- Fajnzlber, P., Lederman, D., & Loayza, N. (2002). Inequality and violent crime. *Journal of Law & Economy*, 45, 1.
- Farrington, D. P. (1986). Age and crime. *Crime & Justice*, 7, 189–250.
- Feldmeyer, B., & Steffensmeier, D. (2007). Elder crime: Patterns and current trends, 1980–2004. *Research on Aging*, 29(4), 297–322.
- Finkelhor, D., Ormrod, R., Turner, H., & Hamby, S. L. (2005). The victimization of children and youth: A comprehensive, national survey. *Child Maltreatment*, 10(1), 5–25.
- Foulkes, M., & Newbold, K. B. (2008). Poverty catchments: Migration, residential mobility, and population turnover in impoverished rural Illinois communities. *Rural Sociology*, 73(3), 440–462.
- Gartner, R. (2013). *Neighbourhood change and the spatial distribution of violent crime*. Retrieved from <http://neighbourhoodchange.ca/wpcontent/uploads/2014/04/2de90c57bf6b615f986f6bdd82d49c6c8.pdf>
- Gilchrist, A., & Kyprianou, P. (2011). *Social networks, poverty and ethnicity*. Retrieved from <https://www.jrf.org.uk/report/social-networks-poverty-and-ethnicity>
- Goodin, R. E. (2014). *Just the facts 101: The Oxford handbook of political science* (1st ed.). Moore Park, CA: Content Technologies.
- Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime*. Palo Alto, CA: Stanford University Press.
- Green, A., Preston, J., & Janmaat, G. (2006). *Education, equality and social cohesion: A comparative analysis*. Basingstoke, UK: Palgrave.
- Gruner, S. (2010). 'The others don't want...' small-scale segregation: Hegemonic public discourses and racial boundaries in German neighbourhoods. *Journal of Ethnic & Migration Studies*, 36(2), 275–292.
- Hino, K., Uesugi, M., & Asami, Y. (2016). Official crime rates and residents' sense of security across neighborhoods in Tokyo. *Japan. Urban Affairs Review*, 54(1), 165–189.
- Hirschfield, A., Birkin, M., Brunsdon, C., Malleon, N., & Newton, A. (2013). How places influence crime: The impact of surrounding areas on neighbourhood burglary rates in a British city. *Urban Studies*, 51(5), 1057–1072.
- Hirschi, T., & Gottfredson, M. (1983). Age and the explanation of crime. *American Journal of Sociology*, 89(3), 552–584.
- House of Commons. (2011). *Official statistics: Census questions*. Retrieved from <https://www.parliament.uk/documents/upload/memosweb3.pdf>
- Johnson, K. M., & Lichter, D. T. (2010). Growing diversity among America's children and youth: Spatial and temporal dimensions. *Population & Development Review*, 36(1), 151–176.
- Johnston, R., & Matthews, J. S. (2004). *Social capital, age, and participation*. Paper presented at the Youth Participation Workshop of the Annual Meeting of the Canadian Political Science Association. Winnipeg, MB.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2), 363–375.
- Kanazawa, S. (2003). Why productivity fades with age: The crime–genius connection. *Journal of Research in Personality*, 37(4), 257–272.
- Karyda, M. (2015). *The effect of crime in the community on becoming not in education, employment or training (NEET) at 18–19 years in England*. London, UK: UCL Institute of Education.
- Kathryn, H. (2016). *Associations between police recorded ethnic background and being sentenced to prison in England and Wales*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/479874/analysis-of-ethnicity-and-custodial-sentences.pdf
- Keene, D., Bader, M., & Ailshire, J. (2013). Length of residence and social integration: The contingent effects of neighborhood poverty. *Health & Place*, 21, 171–178.
- Kelly, M. (2000). Inequality and crime. *Review of Economics & Statistics*, 82(4), 530–539.
- Kindler, M., Ratcheva, V., & Piechowska, M. (2014). *Social networks, social capital and migrant integration at local level: European literature review*. Edgbaston, UK: Institute for Research into Superdiversity, University of Birmingham.
- Kongmuang, C. (2006). *Modelling crime: A spatial microsimulation approach* (Unpublished Ph.D. Dissertation). University of Leeds, Leeds, UK.

- Kristjánsson, Á. L. (2007). On social equality and perceptions of insecurity: A comparison study between two European countries. *European Journal of Criminology*, 4(1), 59–86.
- Krivo, L. J., & Peterson, R. D. (1996). Extremely disadvantaged neighborhoods and urban crime. *Social Forces*, 75(2), 619–648.
- Landis, R. S. (2005). Standardized regression coefficients. In B. Everitt & D. Howell (Eds.), *Encyclopedia of behavioral statistics*. Chichester, UK: Wiley.
- Laurence, J. (2011). The effect of ethnic diversity and community disadvantage on social cohesion: A multi-level analysis of social capital and interethnic relations in UK communities. *European Sociological Review*, 27(1), 70–89.
- Lee, D., & McLanahan, S. (2015). Family structure transitions and child development: Instability, selection, and population heterogeneity. *American Sociological Review*, 80(4), 738–763.
- Letki, N. (2008). Does diversity erode social cohesion? Social capital and race in British neighbourhoods. *Political Studies*, 56(1), 99–126.
- Lieberman, S. (1985). *Making it count: The improvement of social research and theory*. Berkeley, CA: University of California Press.
- Loeber, R., Menting, B., Lynam, D. R., Moffitt, T. E., Stouthamer-Loeber, M., Stallings, R., ... Pardini, D. (2012). Findings from the Pittsburgh youth study: Cognitive impulsivity and intelligence as predictors of the age–crime curve. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(11), 1136–1149.
- Louderback, E. R., & Sen Roy, S. (2018). Integrating social disorganization and routine activity theories and testing the effectiveness of neighbourhood crime watch programs: Case study of Miami-Dade County, 2007–15. *British Journal of Criminology*, 58(4), 968–992.
- Ludwig, A., & Marshall, M. (2015). Using crime data in academic research: Issues of comparability and integrity. *Records Management Journal*, 25(3), 228–247.
- Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education. *Economic Journal*, 121(552), 463–484.
- Macpherson, R. (1999). *The Inquiry into the matters arising from the death of Stephen Lawrence*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/277111/4262.pdf
- Maginnis, R. (1997). Single-parent families cause juvenile crime. In A. E. Sadler (Ed.), *Juvenile crime: Opposing viewpoints* (pp. 62–66). Farmington Hills, MI: Greenhaven Press.
- Malczewski, J., & Poetz, A. (2005). Residential burglaries and neighborhood socioeconomic context in London, Ontario: Global and local regression analysis. *Professional Geographer*, 57(4), 516–529.
- Malleson, N., & Birkin, M. (2012). Analysis of crime patterns through the integration of an agent-based model and a population microsimulation. *Computers, Environment & Urban Systems*, 36(6), 551–561.
- Markowitz, F. E., Bellair, P. E., Liska, A. E., & Liu, J. (2001). Extending social disorganization theory: Modeling the relationships between cohesion, disorder, and fear. *Criminology*, 39(2), 293–319.
- Martin, D. (2002). Spatial patterns in residential burglary assessing the effect of neighborhood social capital. *Journal of Contemporary Criminal Justice*, 18(2), 132–146.
- Martin, K. (2014). *Can regression model with small r-squared be useful?* Retrieved from <http://www.theanalysisfactor.com/small-r-squared/>
- Masike, L. G., & Mofokeng, J. (2014). Safety of students in residences of a university in the Tshwane Metropolitan. *Acta Criminologica: Southern African Journal of Criminology*, 2014(Special ed. 2), 64–80.
- McCall, P. L., Land, K. C., Dollar, C. B., & Parker, K. F. (2013). The age structure–crime rate relationship: Solving a long-standing puzzle. *Journal of Quantitative Criminology*, 29(2), 167–190.
- McVie, S. (2005). Patterns of deviance underlying the age–crime curve: The long term evidence. *British Society of Criminology e-Journal*, 7, 1–15.
- Meer, T. V. D., & Tolsma, J. (2014). Ethnic diversity and its effects on social cohesion. *Annual Review of Sociology*, 40, 459–478.
- Meera, A. K., & Jayakumar, M. D. (1995). Determinants of crime in a developing country: A regression model. *Applied Economics*, 27(5), 455–460.
- Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011). A partial least squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology*, 6(1), 27.
- Mellgren, C. (2011). *What's neighbourhood got to do with it? The influence of neighbourhood context on crime and reactions to crime*. Malmö, Sweden: Malmö University.
- Moore, M. H., & Trojanowicz, R. C. (1988). *Policing and the fear of crime*. Washington, DC: U.S. Department of Justice, National Institute of Justice.
- Morenoff, J. D., Sampson, R. J., & Raudenbush, S. W. (2001). Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. *Criminology*, 39(3), 517–558.
- Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., ... Prati, D. (2014). Choosing and using diversity indices: Insights for ecological applications from the German biodiversity exploratories. *Ecology & Evolution*, 4(18), 3514–3524.

- Murray, A. T., McGuffog, I., Western, J. S., & Mullins, P. (2001). Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment. *British Journal of Criminology*, 41(2), 309–329.
- Nam, C. B. (2004). The concept of the family: Demographic and genealogical perspectives. *Sociation Today*, 2(2), 1–9.
- O'Connor, A. (2004). The sociology of youth subcultures. *Peace Review*, 16(4), 409–414.
- Oh, J. H. (2003). Assessing the social bonds of elderly neighbors: The roles of length of residence, crime victimization, and perceived disorder. *Sociological Inquiry*, 73(4), 490–510.
- ONS. (2011). *Census aggregate data*. Retrieved from <https://census.ukdataservice.ac.uk/getdata/aggregate-data>
- ONS. (2014). *Crimestatistics: Focus on property crime*. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/focusonpropertycrime/yearendingmarch2016>
- ONS. (2015). *Statistics on race and the criminal justice system*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/480250/bulletin.pdf
- ONS. (2017). *Crime in England and Wales: Police force area data*. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/policeforceareadata>
- Papadopoulos, G. (2014). Immigration status and property crime: An application of estimators for underreported outcomes. *IZA Journal of Migration*, 3(1), 12.
- Parker, K. F., & Johns, T. (2002). Urban disadvantage and types of race-specific homicide: Assessing the diversity in family structures in the urban context. *Journal of Research in Crime & Delinquency*, 39(3), 277–303.
- Piquero, A. R., & Brame, R. W. (2008). Assessing the race-crime and ethnicity-crime relationship in a sample of serious adolescent delinquents. *Crime & Delinquency*, 54(3), 390–422.
- Piquero, A. R., Farrington, D. P., & Blumstein, A. (2003). The criminal career paradigm. *Crime & Justice*, 30, 359–506.
- Pitner, R. O., Yu, M., & Brown, E. (2012). Making neighborhoods safer: Examining predictors of residents' concerns about neighborhood safety. *Journal of Environmental Psychology*, 32(1), 43–49.
- Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, 41(3), 227–259.
- Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437), 179–191.
- Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis: A research tool*. New York, NY: Springer Science & Business Media.
- Rees, P., & Butt, F. (2004). Ethnic change and diversity in England, 1981–2001. *Area*, 36(2), 174–186.
- Reilly, B., & Witt, R. (2008). Domestic burglaries and the real price of audio-visual goods: Some time series evidence for Britain. *Economics Letters*, 100(1), 96–100.
- Reisinger, H. (1997). The impact of research designs on R2 in linear regression models: An exploratory meta-analysis. *Journal of Empirical Generalisations in Marketing Science*, 2(1), 1–12.
- Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2001). Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: A 15-year follow-up of low-income children in public schools. *Journal of the American Medical Association*, 285(18), 2339–2346.
- Rosenstein, L., Scott, L., & Pratt, M. (2011). *Exploratory regression: A tool for modeling complex phenomena*. Retrieved from <http://www.esri.com/news/arcuser/0111/files/exploratory.pdf>
- Rufrancos, H. G., Power, M., Pickett, K. E., & Wilkinson, R. (2013). Income inequality and crime: A review and explanation of the time-series evidence. *Social Criminology*, 1(1), 1000103.
- Sabates, R., Feinstein, L., & Shingal, A. (2008). *Educational inequality and juvenile crime: An area-based analysis* (World Benefits of Learning Research Report No. 26). Retrieved from <http://webarchive.nationalarchives.gov.uk/20130401151715/http://www.education.gov.uk/publications/eOrderingDownload/WBL26.pdf>
- Sabates, R. (2008). Educational attainment and juvenile crime: Area-level evidence using three cohorts of young people. *British Journal of Criminology*, 48(3), 395–409.
- Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology*, 94(4), 774–802.
- Sampson, R. J., & Laub, J. H. (2003). Life-course desisters? Trajectories of crime among delinquent boys followed to age 70. *Criminology*, 41(3), 555–592.
- Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multi-level study of collective efficacy. *Science*, 277(5328), 918–924.
- Sampson, R. J., & Wooldredge, J. D. (1987). Linking the micro- and macro-level dimensions of lifestyle routine activity and opportunity models of predatory victimization. *Journal of Quantitative Criminology*, 3(4), 371–393.
- Sariaslan, A., Långström, N., D'Onofrio, B., Hallqvist, J., Franck, J., & Lichtenstein, P. (2013). The impact of neighbourhood deprivation on adolescent violent criminality and substance misuse: A longitudinal, quasi-experimental study of the total Swedish population. *International Journal of Epidemiology*, 42(4), 1057–1066.
- Shammas, V. L., & Sandberg, S. (2015). Habitus, capital, and conflict: Bringing Bourdieusian field theory to criminology. *Criminology & Criminal Justice*, 16(2), 195–213.

- Sharp, D., & Atherton, S. (2007). To serve and protect? The experiences of policing in the community of young people from black and other ethnic minority groups. *British Journal of Criminology*, 47(5), 746–763.
- Shaw, C. R., & McKay, H. D. (1942). *Juvenile delinquency and urban areas*. Chicago, IL: University of Chicago Press.
- Shepherd, P. J. (2006). *Neighbourhood profiling and classification for community safety* (Unpublished Ph.D. Dissertation), University of Leeds, Leeds, UK.
- Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688.
- Sinha, A., Malo, P., & Kuosmanen, T. (2015). A multi-objective exploratory procedure for regression model selection. *Journal of Computational & Graphical Statistics*, 24(1), 154–182.
- Steffensmeier, D., Zhong, H., & Lu, Y. (2017). Age and its relation to crime in Taiwan and the United States: Invariant, or does cultural context matter? *Criminology*, 55(2), 377–404.
- Sturgis, P., Brunton-Smith, I., Kuha, J., & Jackson, J. (2014). Ethnic diversity, segregation and the social cohesion of neighbourhoods in London. *Ethnic & Racial Studies*, 37(8), 1286–1309.
- Sweeten, G., Piquero, A. R., & Steinberg, L. (2013). Age and the explanation of crime, revisited. *Journal of Youth & Adolescence*, 42(6), 921–938.
- Takagi, D., & Kawachi, I. (2014). Neighborhood social heterogeneity and crime victimization in Japan: Moderating effects of social networks. *Asian Journal of Criminology*, 9(4), 271–284.
- Tasgin, S., & Morash, M. (2016). Social context, family process, and Turkish boys' pathway to incarceration: An application of the age-graded theory of informal social control. *International Journal of Comparative & Applied Criminal Justice*, 40(4), 307–323.
- Thomas, M. J., Stillwell, J. C., & Gould, M. I. (2016). Modelling the duration of residence and plans for future residential relocation: A multilevel analysis. *Transactions of the Institute of British Geographers*, 41(3), 297–312.
- Tittle, C. R., Ward, D. A., & Grasmick, H. G. (2003). Gender, age, and crime/deviance: A challenge to self-control theory. *Journal of Research in Crime & Delinquency*, 40(4), 426–453.
- Tonry, M. (1997). Ethnicity, crime, and immigration. *Crime & Justice*, 21, 1–29.
- Tricia, J. (2016). Synthesizing structure and agency: A developmental framework of Bourdieu's constructivist structuralism theory. *Journal of Theoretical & Philosophical Criminology*, 8(1), 1.
- Tseloni, A., Osborn, D. R., Trickett, A., & Pease, K. (2002). Modelling property crime using the British crime survey. What have we learnt? *British Journal of Criminology*, 42(1), 109–128.
- Umberson, D., Pudrovska, T., & Reczek, C. (2010). Parenthood, childlessness, and well-being: A life course perspective. *Journal of Marriage & Family*, 72(3), 612–629.
- Unnever, J. D. (2018). Ethnicity and crime in the Netherlands. *International Criminal Justice Review*, 28, in press.
- Vermeulen, F., Tillie, J., & van de Walle, R. (2012). Different effects of ethnic diversity on social capital: Density of foundations and leisure associations in Amsterdam neighbourhoods. *Urban Studies*, 49(2), 337–352.
- Wang, G. C. (1996). How to handle multi-collinearity in regression modeling. *Journal of Business Forecasting*, 15(1), 23.
- Weisburd, D., & Piquero, A. R. (2008). How well do criminologists explain crime? Statistical modeling in published studies. *Crime & Justice*, 37(1), 453–502.
- Wikström, S., & Wikström, P.-O.-H. (2001). Why are single parents more often threatened with violence? A question of ecological vulnerability? *International Review of Victimology*, 8(2), 183–198.
- Witt, R., Clarke, A., & Fielding, N. (1998). Crime, earnings inequality and unemployment in England and Wales. *Applied Economics Letters*, 5(4), 265–267.
- Wooldridge, J. (2012). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western.
- Yamamura, E. (2011). How do neighbors influence investment in social capital? Homeownership and length of residence. *International Advances in Economic Research*, 17(4), 451–464.
- Zhao, J. S., Lawton, B., & Longmire, D. (2015). An examination of the micro-level crime–fear of crime link. *Crime & Delinquency*, 61(1), 19–44.

How to cite this article: Gulma UL, Evans A, Heppenstall A, Malleson N. Diversity and burglary: Do community differences matter? *Transactions in GIS*. 2018;00:1–22. <https://doi.org/10.1111/tgis.12511>