Causal inference methods and simulation approaches in observational health research within a geographical framework

Lauren Berrie

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

The University of Leeds

School of Medicine

August 2019

The candidate confirms that the work submitted is their own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter 6 of the thesis has appeared in publication as follows:

L Berrie, GTH Ellison, PD Norman, PD Baxter, RG Feltbower, PWG Tennant, and MS Gilthorpe. The association between childhood leukemia and population mixing: An artifact of focusing on clusters? Epidemiology, 30:75–82, 2019.

L Berrie, GTH Ellison, PD Norman, PD Baxter, RG Feltbower, PWG Tennant, and MS Gilthorpe. Authors' respond: Re: The association between childhood leukemia and population mixing: An artifact of focusing on clusters? Epidemiology, 30, 2019.

Lauren Berrie researched the literature, developed the project design, ran the simulations and analyses, compiled results into tables and figures and drafted the manuscript. Professor Gilthorpe was involved in the conceptualisation of the study and, along with Drs Baxter and Norman, was involved in revising the manuscript and supervising the study. Dr Feltbower aided in real–world data acquisition and all authors gave advice on and approved the final version of the manuscript.

Lauren Berrie wrote the follow–up letter and all authors approved the final version.

Acknowledgements

It is difficult to find the words to thank all of the people who have had a role in this work. Although a PhD is ultimately awarded to an individual, it could not have happened without the valuable input and support of a number of people.

First of all, my thanks must go to my supervisors Mark Gilthorpe, Paul Norman and Paul Baxter. To Mark, in many ways I would not and could not have done this without you; thank you for seeing my potential and encouraging me in this endeavour. Thank you for your reassurance when I faltered and for always being there to provide a healthy dose of perspective. To Paul N, thank you for your unconditional support and for always being only an e–mail away. To Paul B, thank you for your encouragement and for being there to dot the i's and cross the t's.

This work would not have been possible without the financial support of the Medical Research Council, in particular the generous training budget which has afforded me the opportunity to attend training and present at conferences in many different places; all of them unforgettable experiences.

I am grateful to have been able to conduct my research in such a supportive and stimulating environment. This has allowed me to pursue teaching and research collaborations outside of my PhD which has been invaluable to my development as an academic.

Thank you to Dr Johannes Textor of the Radboud University Medical Centre; being able to

spend time on a research visit to your group was a particular highlight of the last 3.5 years and my research has benefited from your input immensely. I cannot thank you enough.

Special thanks to Claire Owen, Kellyn Arnold, Peter Tennant and Wendy Harrison. Thank you for getting me out from behind my desk and for talking to me about things other than my research; I feel very lucky to have shared this time with you. In particular, Claire, thank you for your advice and friendship from the very beginning and for the times you have brought me back from the brink of panic. Kellyn, thank you for being a voice of reason and for being the best conference travel buddy.

Thank you to all of the above for the valuable discussions of their and my work; it has been an honour to work alongside you. Thank you also to my examiners Dr Myles Gould and Dr Rosie Green for what turned out to be a very enjoyable discussion of this thesis and the best examination experience I could have hoped for.

Thank you to my adopted family in Leeds. Janne and Andrew, thank you for welcoming me into your home and especially for the long chats over gallons of tea at the end of the day. Finella, Orson, Elfina and Yolanda, you may never know how much you have brightened my days and provided much needed distraction from office woes. I cannot (indeed I do not want to) imagine what the last 3.5 years would have been like without you. I look forward to seeing you grow and to celebrating your own achievements.

To mum, thank you for your patience, love and support and your unwavering belief in my ability to pull this off. You always told me that I could do anything I put my mind to; I hope I have made you proud.

Finally, and most importantly, thank you to Vanessa. I could take on anything with you by my side.

Abstract

Statistical methods are often used habitually, perhaps without sufficient reflection on their robustness in a range of novel circumstances. Increasingly, there is a desire to unravel the complexities of humans interacting with their environments, to improve our understanding and explanation of what influences population health in the wider context of our living environment. A framework is provided for using simulation and causal inference methods to evaluate analytical approaches in health geography, to introduce the reader to some of the considerations around complexity of context and data generation that may need to be reflected upon carefully when applying such methods in their own work. These methods have the potential to aid researchers in their explanation of what factors are important for population health and well–being in the context of our geographical environment while avoiding potential pitfalls in their work and allowing for greater critical evaluation of the methods employed by themselves and others.

This thesis considers the utility of simulation to investigate applied problems related to mathematical coupling and specific considerations that need to be made in relation to research on the relationship between limiting long–term illness and deprivation and the challenges encountered while investigating the relationship between population mixing and childhood leukaemia —with all such considerations examined through the lens of cause and effect. The datasets chosen are representative of many others in health geography and span the full range of outcome prevalence rates likely encountered.

Methods in causal inference and simulation are demonstrated to be powerful tools in

understanding potential bias in research analyses. With careful planning, forethought and reflection on the data generating processes of the context of interest, causal inference and simulation methodologies are accessible to all researchers to improve their understanding of the methods they employ to address the research questions they pose.

Contents

List of Figures

List of Tables

Table 1: Table of notation

Notation	**Meaning**
ANCOVA	Analysis of Covariance
BMI	Body Mass Index
CI	Confidence Interval
CVD	Cardiovascular Disease
DAG	Directed Acyclic Graph
EW	Electoral Ward
GDP	Gross Domestic Product
km	kilometre
LLTI	Limiting long–term illness
LSOA	Lower Layer Super Output Area
MAUP	Modifiable Areal Unit Problem
MSAS	Minimally Sufficient Adjustment Set
MSOA	Middle Layer Super Output Area
MTUP	Modifiable Temportal Unit Problem
OA	Output Area
RCT	Randomised Controlled Trial
RR	Risk Ratio
SEP	Socio–economic position
SIR	Standardised illness ratio
UK	United Kingdom

Chapter 1

Introduction

Statistical methods are used habitually; researchers tend to analyse data in the same ways and go on to teach the same methods.[1,2] This thesis calls for the mindful application of methods to statistical problems aided by causal inference theory and advocates for the practice of developing simulations alongside all applied research to fully understand and appreciate any methods used so that they are adopted conscientiously.

The work that appears in this thesis was heavily influenced by the topics covered in the MSc Epidemiology and Biostatistics which immediately preceded the undertaking of this PhD research. As part of this course, the author was introduced to methods in causal inference which the author wished to use in the course of this work which, as per the original funding proposal was to include health geography problems. The health geography problems approached in this thesis were chosen due to previous awareness of the research by the student (e.g. population mixing) and their supervisors (e.g. limiting long–term illness). However, the datasets used in this thesis are representative of a whole swathe of research in the field of health geography. The methods used throughout this thesis and their implementation on these specific datasets aim to be translatable to further research in the field of health geography.

1.1 Working definition of health geography

Throughout this thesis, the fields of epidemiology and health geography are referred to. These two fields have developed separately[3] and therefore require definition and a brief discussion at the beginning of this work.

Both epidemiology and health geography aim to understand disease processes and subsequently develop interventions.[4] Indeed, epidemiology has been defined as the "study of how often diseases occur in different groups of people and why",[5] whilst the subtopic of social epidemiology is concerned with "social phenomena that influence health inequalities in populations"[3] (p.3) which can include geographical concepts. Health geography has been defined as "the study of the relationship between health and place" and in relation to the work in this thesis, "in particular, health geography is concerned with... the socio–spatial relations of health,...".[6] Some researchers may not consider there to be any substantial difference between health geography and (social) epidemiology as defined above, however, as the two fields have developed separately,[3] so have the methods used in each. For example, the causal inference methods used in this thesis have been developed within epidemiology, however, they have not been used in health geography (though some causal inference methods are becoming more popular within health geography, such as approaches using instrumental variables[7]).

With the above definitions in mind, the working definition of health geography used in this thesis assumes that the topic of epidemiology is subsumed within health geography. That is, the geographical information inherent to health geography extends (and can enhance) epidemiological data. The causal inference and simulation methods which are becoming more common in observational health research within epidemiology are applied throughout to understand health geographical research questions.

1.2 Thesis roadmap

There is an overarching introduction to causal inference and simulation in Chapter 2, this is supplemented with short literature reviews for the topics covered in each chapter. This is because there is not a linear story to this work and each chapter does not necessarily build on the last, however, there is a common theme to all of them: the aim of understanding bias in health geography research using causal inference and simulation.

As broad objectives, each chapter will: represent the problem in terms of a causal diagram or diagrams; tease out the problem using simulation and; present the appropriate method of analysis as determined by causal inference theory and confirmed by simulation.

The Vancouver referencing style is used since research included in this thesis is published in a journal which uses this style and future submissions are anticipated to be in journals which use this style.

1.2.1 Chapter 2

Chapter 2 introduces background literature related to causal inference and simulation studies and links these through the data generating process.

There is a growing literature on the implementation of causal inference methods for observational data and the aim here is to include enough background information for the reader to follow the subsequent examples. Causal inference methods have been used to uncover several pitfalls in observational data analysis and some examples of these are included as an indication of the power of these methods and as an illustration of how they are to be used when looking at the novel situations addressed in this thesis.

Causal inference methods naturally combine with simulation through the data generation process and this is put explicitly to highlight the potential of the methods when used together.

1.2.2 Chapter 3

Chapter 3 expands on Chapter 2 by describing how one would conduct a simulation study informed by causal inference methods along with some of the challenges and issues that one may want to consider when developing one's own simulation; in particular, accounting for elements of health geography applicable to later chapters.

This chapter begins with a brief overview of the history of health geography data before discussing the purpose of simulation studies for assessing statistical methods, foreseeing criticism, how simulations are approached in this thesis and considering complexity. The more specific elements of simulation that are covered are: compositional data and composite variables, the modifiable areal and temporal unit problems, and the 'most dangerous equation'.

1.2.3 Chapter 4

Chapter 4 introduces the problem of mathematical coupling of proportions with common denominators, illustrates the problem using causal diagrams and further probes the extent of the problem via changing parameters in simulations.

This chapter first shows how causal inference methods, through the use of causal diagrams, can expand understanding of this long–standing problem and how simulation can be used to extend this knowledge to the field of health geography specifically. It is shown under which circumstances the historical solution to mathematical coupling breaks down; a situation that is more easily understood through the use of causal diagrams.

1.2.4 Chapter 5

Chapter 5 uses census data regarding limiting long–term illness and area–level deprivation to investigate the problems from Chapter 4 in an applied health geography setting and

using observed data to inform simulations.

The simulations undertaken in this chapter follow the framework outlined in Chapter 3, under the null hypothesis and consider compositional data and composite variables in their undertaking.

1.2.5 Chapter 6

Chapter 6 uses simulation and observed data analysis to investigate what has been termed the 'population mixing hypothesis' and childhood leukaemia.

These simulations again follow the framework outlined in Chapter 3 and consider selection on the outcome, the 'most dangerous equation' and the modifiable areal and temporal unit problems.

1.2.6 Chapter 7

Chapter 7 summarises the findings of all chapters, discusses the strengths and limitations of the research and makes suggestions for future research.

Causal inference methods and simulation approaches can provide a powerful tool for researchers to evaluate their analytical and inferential methods. They have the potential to aid researchers in avoiding potential pitfalls in their work. These approaches are not limited to use by only a subset of researchers and this thesis aims to illustrate how these methods can be used by any researcher in the field of health geography. As these methods have not previously been applied to health geography problems, this thesis aims to provide a framework for using simulation and causal inference methods to evaluate analytical methods in epidemiology and health geography and to introduce the reader to some of the considerations that may need to be taken into account when applying such methods in their own work.

Chapter 2

Background

This chapter provides definitions and background information relating to causal inference, simulation studies and health geography, which will be required for later chapters. First, causal inference methods and simulation study designs are introduced, then some helpful causal inference language and definitions are provided.

To enable readers from a range of relevant disciplines to access this work, definitions of some key terms relevant to the forthcoming chapters are given at the end of this Chapter in Section 2.10. The words that are defined at the end of the Chapter appear in bold throughout this Chapter.

2.1 Causal Inference

In epidemiology and related fields, the overall aim, which may or may not be made explicit, is to determine the cause(s) of a particular outcome, or to predict the effect of an intervention.[8,9] Causal inference aims to emulate randomised controlled experiments on observed data. In observational studies, such as those conducted in epidemiology and health geography research, it is not possible to randomise individuals or areas to the

exposure of interest as is done in **Randomised Controlled Trials (RCTs)**. This could be for many reasons, such as it being time consuming, expensive, or unethical. A rigorous approach to answering causal questions using observational data has been in development since the topic was introduced to the field of epidemiology in 1999,[10] however, uptake of these methods across the field has been slow.[11]

The reason that an **RCT** is considered to be the 'gold standard' in statistics is that all variables that could control the outcome are either held static or vary completely at random, except the variable of interest. This means that any change seen in the outcome must be a result of changes in a specific input variable. This is what researchers wish to emulate using causal inference methods on observational data.[12] Huge insight can be gained in the analysis of observational data when causal inference methods are embraced and causal aims are made explicit. This means that the estimates generated are more likely to be robust and meaningful. Unfortunately, it is uncommon in quantitative social science research for these aims to be made explicit and for causal inference methods to be adopted, even though this is often the unspoken aim. Hernán[13] argues that "being explicit about the causal objective of a study reduces ambiguity in the scientific question, errors in the data analysis, and excesses in the interpretation of the results" (p.616). Many health researchers have been encouraged not to discuss causation when interpreting the results of observational studies as this was thought to "overreach the evidence" (p.81).[14] Causal inference methods are embraced in this thesis as they allow clear and logical thinking about research questions and make the assumptions around variable relationships explicit. And, as Holland wrote:[15] "Correlation does not imply causation, and yet causal conclusions drawn from a carefully designed experiment are often valid" (p.945).

2.1.1 What are causal inference methods?

Causal inference methods unite the **counterfactual** and probabilistic theories of causation into an algebraic and graphical framework.[12]

Under the **counterfactual** or potential outcome frameworks, the explanation of a cause is described in the following way: if the cause did not happen then the chance of the outcome would be different to if the cause had occurred, i.e. an event A may be considered a cause of an event Y if, *contrary to fact*, had A not occurred, then the probability distribution of Y would be different.

As an example (adapted from[12]), consider a driver, Jess, who is driving home and comes to a fork in the road. They choose to go right and arrive late for a dinner engagement. Upset, Jess says, “I should have gone left instead!". This statement implies that Jess’ decision to go right at the fork in the road *caused* them to be late home because, had they chosen to turn left, they would not have arrived late. There is no way to prove whether this statement is correct; Jess cannot travel back in time to the same moment and observe what would have happened if they had turned left (attempting this journey at any other time would not be directly comparable, as this requires being in the same space and time to hold all other factors constant). This is the ‘fundamental problem of causal inference’, once one outcome has been observed (the fact) it is not possible to know what would have happened otherwise (the **counterfactual**).

This example demonstrates the philosophical aim of causal thinking; to compare how things would have been different in a counterfactual universe. It is akin to a randomised controlled study where everything is kept the same except for the cause being studied. This is called the *counterfactual contrast* between *exchangeable units of analysis*, that is, units that are the same *in every way except for the presumed causal factor of interest*.

2.1.2 Causal Conditions

It is not possible to identify individual–level causal effects within a causal framework from observational data. However, there are three assumptions that, if met, can be used to identify average causal effects from observational data. These are:

- **exchangeability**;
- **positivity**; and
- **consistency**.

2.1.3 Exchangeability

Essentially, **exchangeability** is the condition of no **confounding**. In the counterfactual framework, "causal inference can be drawn when the distribution of observed outcomes among those who did not receive the intervention equals in expectation the distribution that would have been observed had those who received the intervention not received it" (p.82).[14] This means that had the exposed actually been unexposed, they would have experienced the same distribution of outcomes as those who were unexposed.[16]

Unconditional Exchangeability

As an example (adapted from[17]), in the case of a **randomised controlled trial (RCT)**, suppose that a researcher wants to assess the effectiveness of aspirin (A) for treating headache (Y). In order to test this, the researcher gathers a large, representative sample of individuals with headaches and assigns each of them to receive aspirin ($a = 1$) or a placebo ($a = 0$). Two hours later, the individuals are observed to see whether they have a headache ($y = 1$) or not ($y = 0$).

There may be some measured and unmeasured attributes of the individuals that may affect how likely they are to have a headache when they are checked after two hours (e.g. some of them may have been suffering a more intense headache than others), however, the randomisation of the individuals into treatment groups so that these attributes are equivalent between groups, ensures that, *on average*, the individuals who are given the aspirin are **exchangeable** with those who are given the placebo. In an **RCT**,

randomisation of individuals into treatment groups ensures that the units of analysis are **unconditionally exchangeable**. This means that the researcher can say what *did happen* (the 'outcome') to those who received the treatment provides a good estimate of *what would have happened* to the placebo group (the 'potential outcome') if they had received the treatment. In the aspirin example, the causal effect of receiving the aspirin as treatment is found by comparing the difference between the observed outcome in the placebo group with their potential outcome, estimated from the treatment group.

Conditional exchangeability

As mentioned earlier, there are many reasons why an **RCT** may not be feasible especially in social science and in order to estimate average causal effects, two units of analysis must be created that are exchangeable so that their outcomes can be compared. This is a challenge in observational data.

As an example, if a researcher wanted to estimate the causal effect of the influenza vaccine on the diagnosis of influenza, it is likely that those individuals who received the vaccine are systematically different to those who did not receive the vaccine (e.g. they are older, richer, etc.).[17] It is not appropriate in that case to simply compare the outcomes in those who received the vaccine with those who did not to estimate the average causal effect of the vaccine because the differences in the outcomes between the two groups may be due to other differences between the groups. However, by comparing the outcomes between subgroups in which the distributions of the relevant attributes are equivalent, the causal effect could be estimated. These subgroups are exchangeable conditional on these factors, i.e. they are **conditionally exchangeable**.

Causal diagrams (introduced in Section 2.1.7) aid researchers in determining which attributes (variables) are required to be measured and conditioned on to achieve **conditional exchangeability**. If it was possible to condition on all of the appropriate

variables in a model (as determined by a causal diagram), the subgroups would be directly comparable and causal inferences could be made. In reality, when using observational data, there will be unmeasured variables which cannot be accounted for which means assumptions and approximations must be made.

2.1.4 Positivity

The **positivity** assumption states that any individual has a positive probability of receiving all values of the treatment (or exposure) variable. This means that in an analysis it is required that some individuals (or units of analysis) receive the treatment (or exposure) and some do not. When the **positivity** assumption is violated the researcher will not have any information about the distribution of the outcome for a certain subset of the population and consequently will not be able to make any inferences about it. An example of a positivity violation would be where every patient in a critical condition received a heart transplant,[18] as there is no information on patients in a critical condition who did not receive a heart transplant.

2.1.5 Consistency

The **consistency** assumption states that, for an individual who received treatment, their potential outcome is equal to their observed outcome if they received treatment, and is therefore known. Similarly, for an untreated individual, their potential outcome would equal their observed outcome but their outcome had they been treated remains unknown.[18] This may seem like an obvious statement, however, it is only an assumption and does not always hold. For example, it may not hold when an intervention or exposure is not well–defined. If a researcher wanted to understand the causal effect of smoking and did not differentiate between smoking regular cigarettes and e–cigarettes they would have multiple **counterfactuals** that they have not accounted for, i.e. one for

individuals who smoke regular cigarettes and one for those who smoke e–cigarettes. These **counterfactuals** cannot be combined into a composite **counterfactual**, therefore the results of any analyses cannot be drawn on to estimate the causal effect of smoking.

Another example is that of body mass index (BMI). If a researcher was interested in the effect of BMI on the risk of diabetes, there are ways in which two people could have the same BMI but have completely different body compositions (e.g. one could be 'obese' due to high quantities of fat whereas the other could be 'obese' due to high quantities of muscle). Although these 'exposures' are the same, they would not lead to the same risk of diabetes. This is because the measure, BMI, is not consistent.[19]

2.1.6 Interference

Another assumption that is often made when estimating causal effects is that of no **interference**, as the presence of interference makes causal inference more complex.[20] Under no **interference** it is assumed that the outcome of one individual or unit is not affected by the treatment of any other individual or unit.[21] In many settings this assumption does not hold. A common example of **interference** is that of infectious diseases where the vaccinated population affects whether the rest of the population becomes infected. However, **interference** can also be present when: 1) the intervention is defined and measured on one type of observational unit, but 2) outcomes of interest are defined and measured on a second, distinct type of unit" (p.1)[22] This could be the case in health geography research where policies are implemented at area–level but outcomes are measured at the individual–level. It could be argued that invoking the clustering effect of health geography addresses this in part where appropriate methods are used to respect the data hierarchies (e.g. multilevel modelling).

2.1.7 Directed Acyclic Graphs

Directed Acyclic Graphs (DAGs) allow researchers to represent the causal relationships that they believe to exist between an **exposure**, **outcome** and **confounding** variables visually. **DAGs** provide researchers with a method to represent expert knowledge on the topic along with their assumptions.[23] Mathematical rules can then be applied to decide which variables must be adjusted for to remove **confounding** and reduce **bias** in future analyses.[18] This allows researchers to identify, measure, and compensate for every potential source of non–causal association between the two variables of interest, thereby allowing for **conditional exchangeability**.[24,25] When **DAGs** and data are combined, the results of interventions can be predicted without actually performing those interventions.[12]

The mathematical origins of DAGs lie in causal graph theory. They were developed extensively in the field of computer science to provide robustness in their application[26] and were subsequently introduced into epidemiology at the end of the last millennium.[10] Variables are represented as **nodes**, and **nodes** are connected by **arcs** (or arrows) to indicate the existence and direction of hypothesised causal relationships; **DAGs** encode a researcher's *a priori* assumptions among **exposure**, **outcome** and covariates.[25] A group of arrows that flow in the same direction from one **node** to a subsequent **node** form a **causal path**. **DAGs** are 'acyclic' because no variable can cause itself at an instantaneous point in time. A **path** indicates that there is a statistical dependency between the **nodes** that is causal. **Endogenous nodes** are those that have at least one direct cause in the **DAG** and **exogenous nodes** are those that have no direct causes in the **DAG**.

Temporal information is included in a **DAG** implicitly, because a **node** that has an arrow leading into it from another **node** must proceed the first **node** in time, this means that a **node** can never be returned to.

It is a greater assumption to omit an **arc** in a graph than to include it and in some places

it has been advocated that **DAGs** should be drawn in a forward–saturated way (i.e. **nodes** are drawn in time order from left to right and all preceding **nodes** lead into all successive nodes), **arcs** can then be removed. It is argued that this approach is more likely to avoid omitting **arcs** accidentally.[17]

Confounding of an exposure–outcome relationship occurs when one or more variables cause both the **exposure** and the **outcome**. **Confounding** is a "concern in almost all observational studies in epidemiology that focus on causality"[27] (p.211), and it is often used as a criticism of published research; that some other factor is involved in the relationship between the **exposure** and the **outcome**. That is, groups are not exchangeable. When **confounders** are identified their effects can be eliminated by adjusting for them, stratifying on them or **conditioning** on them.[28]

Diagrams have often been used to express hypothesised relationships between variables, but by applying formal rules their utility can be greatly expanded.[25] **DAGs** assist with identifying a **minimally sufficient adjustment set (MSAS)** of **confounders**, which, when **conditioned** on (e.g. by including as covariates in a regression model), minimise the assumed **confounding**. This helps with robustly estimating the total causal effect of an **exposure** of interest on an **outcome** of interest. **DAGs** provide a relatively simple approach to identifying a **MSAS** of variables that should be controlled for to identify the causal effect of interest; this is achieved by using systematic graphical criteria that have mathematical underpinnings.[29]

DAGs have greater utility than aiding in the identification of **confounders** however and some examples of this will be shown in Section 2.2 and throughout later chapters.

2.1.8 d–Separation

From a **DAG** it is possible to learn the conditional independencies between variables, even though DAGs are non–parametric (i.e. it is known which variables are functions of others,

but not what the nature of those functions are). These independencies are true for all data sets that can be generated from that particular graphical causal model.

The rules of **d–separation** are used to determine whether any pair of **nodes** in a **DAG** are **d–connected**, i.e. there is a connecting **path** between them or **d–separated** i.e. there is no open **path** between them.

When **nodes** are **d–separated** they are definitely independent, however, when they are **d–connected** they are possibly, or likely, dependent; for this reason, it is a stronger assumption to omit an arrow between **nodes** in a **DAG** than it is to retain one.

Two **nodes** in a **DAG** are **d–separated** if every **path** between them is blocked; even if one **path** between them is unblocked they remain **d–connected**.

In order to formally introduce the definition of **d–separation**, the four ways in which 3 variables can be connected are considered:

1. B is a **mediator** in a chain: $A \rightarrow B \rightarrow C$
2. B is a **mediator** in a chain: $A \leftarrow B \leftarrow C$
3. B is a **confounder** of A and C in a fork: $A \leftarrow B \rightarrow C$
4. B is a **collider** on the **causal path** between A and C: $A \rightarrow B \leftarrow C$

In Cases 1–3, B is a **non–collider**, in Case 4, B is a **collider**. Each of these relationships are illustrated in the DAG in Figure 2.1.1.

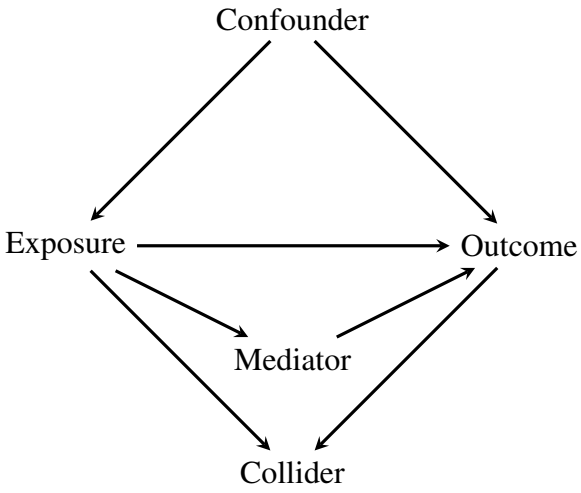

Figure 2.1.1: DAG illustrating each possible type of node.

Non–colliders are normally open and **colliders** are normally closed. **Colliders** and **non–colliders** are defined in relation to a specific **path**, i.e. B could be a **collider** on one **path** and a **non–collider** on another. These terms are clarified below.

A **path** which is blocked by a **collider** can be 'opened' by **conditioning** on the **collider** itself, or any descendant of that **collider** (Case 1 above). **Conditioning** on a variable between the **exposure** and **outcome** of interest that is not a **collider** will cause a **path**, which was otherwise open, to become blocked (Cases 1–3 above).

If a set, **{Z}**, blocks every **path** between two **nodes** X and Y, then X and Y are **d–separated**, conditional on **{Z}**, and are thus independent conditional on **{Z}**.

When there is no **conditioning**, only **colliders** can block a **path**; when there is a **collider** on a **path** between two **nodes** they are conditionally dependent or marginally independent. When a set of **nodes**, **{Z}**, are conditioned on then these are the **nodes** that can block a **path**:

- a **collider** that is not conditioned on (it is not in **{Z}**) nor does it have any descendants in **{Z}**;

- a chain or fork whose middle node is in **{Z}**.

More formally, a **path**, p, is blocked by a set of **nodes {Z}**, if and only if:[12]

- p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node, B, is in **{Z}** (i.e. B is conditioned on), or
- p contains a **collider** $A \rightarrow B \leftarrow C$ such that the collision node, B is not in **{Z}**, and no descendant of B is in **{Z}**.

Pearl[12] helpfully uses the analogy of water pipes to describe 'open' and 'closed' paths within a **DAG**. If a **path** between variables is thought of as a pipe, when variables are dependent, water will flow through these pipes. Only one unblocked **path** is required for water to flow through the pipes and the variables at either end of this **path** will be dependent. On the other hand, a **path** only needs to be blocked in one place for the water to be unable to pass through the pipe.

2.1.9 Minimally Sufficient Adjustment Sets

The systematic graphical criteria of **DAGs** along with **d–separation** provide an algorithm for identifying the **MSAS**:[29]

1. Delete all arrows that start at the **exposure**.
2. Check whether there are any unblocked **back–door paths** between the **exposure** and the **outcome**, where a **back–door path** is a non–causal set of arcs between the **exposure** and the **outcome** of interest.
3. Any covariates on the unblocked **back–door paths** between the **exposure** and the **outcome** will have to be controlled for to remove **confounding** bias.

It is possible to extend this to test whether this set of covariates is minimally sufficient to find the total causal effect of the **exposure** on the **outcome**.[29]

There are also programs available (such as 'Dagitty'[30]) which automate the identification of these adjustment sets which are particularly useful in studies with a large number of variables, as is often the case in epidemiological and health geography studies. When datasets are large and complex it is possible that **conditioning** on a set of variables could block some **paths** whilst opening others and this can become difficult to keep track of manually.[30] Essentially, testing whether the **exposure** and **outcome** are **d–separated** involves examining all **paths** in the **DAG** which have more than three variables which is made much easier using automated procedures.

Confounders, as identified by the **MSAS** and **DAGs**, are not the only factors that **DAGs** are useful for. Indeed, they can aid researchers in identifying **mediators** (Section 2.1.10) and **competing exposures** along with biases, including **collider bias** and the related **selection bias** .

2.1.10 Mediators

Mediators are variables on the **causal path** between the **exposure** and the **outcome**. **Conditioning** on **mediators** does not provide appropriate statistical adjustment for **confounding**. This is because controlling for the **mediator** 'blocks' the **causal path** between the **exposure** and the **outcome**, thereby controlling away some of the processes being investigated.[31] Instead, such adjustment can introduce an inferential **bias** into the estimated exposure–outcome relationship, known as 'reversal paradox';[32] so–called because it can reverse the apparent effect, although it is more likely to simply alter the estimate either towards or away from the **null** (i.e. where there is no relationship between the exposure and the outcome), depending upon the correlation structure amongst the variables being modelled.

2.1.11 Colliders

As noted above, B is a **collider** on the **path** $A \rightarrow C$ in the following **DAG**: $A \rightarrow B \leftarrow C$. The **path** between A and C is closed unless B is conditioned on, in which case it is open. Collider bias occurs when there is a change in association between two variables as a result of **conditioning** on their common effect.[33] To illustrate collider bias, consider the following classic example (adapted from[25]). Suppose there are two factors that determine success as a basketball player: height and speed (successful players must be either extremely tall or extremely fast). In the general population, these two attributes are statistically independent of each other, but if the population of professional basketball players was examined there is a high probability that the short ones are very fast. This is because the short players must compensate for their lack of height with speed to become great players. In the language of causal inference, restricting the population to only professional basketball players is **conditioning** on a common effect of height and speed, and within that specific population height and speed are inversely related; this is different to the relationship found in the general population.

2.2 An example of where DAGs have been used to understand bias

Causal inference methods, and more specifically **DAGs**, have been used to illustrate common problems in observational health research. One such example is Simpson's Paradox.[34,35] Simpson's Paradox is a form of "reversal paradox"[32] where an effect appears to be present when data are analysed in different groups but this effect disappears or is reversed when the group data are combined for analysis. By depicting the data generating structure in a **DAG** one is able to understand how this problem arises; it is not possible to solve this problem with statistical analysis alone[36] and the data must be supplemented

with causal knowledge.[35]

2.2.1 Simpson's Paradox

Simpson's Paradox has been known about by statisticians for a long time,[34] however, it is only through causal inference that it has truly been understood. It cannot be explicated using a statistical approach alone. Pearl[35] uses simulated data to illustrate this problem and shows how the apparent efficacy of a drug is reversed when the data are divided into male and female groups. The drug appears to be harmful to both males and females but beneficial to the population as a whole; intuitively, this is impossible. However, it is possible to understand which of these results is appropriate by thinking about the 'story' behind the data, i.e. the data generating process. This can be represented in a causal diagram where the rules of **d–separation** can be applied to determine whether the **conditioned** or non–conditioned model is correct.

Simpson's Paradox is commonly taught on courses in epidemiology and it is included here as it is an early example of how causal inference is required to fully understand some biases as a result of data analyses or inferences. The rest of the thesis uses causal inference in a similar way, to delve further into the data analyses and inferences of some select epidemiological and health geographical analyses.

2.3 DAGs and area–level problems

DAGs are not widely used to depict area–level variables, however, if one is interested in area–to–area variation in area–level outcomes it is plausible to construct a **DAG** to answer area–level questions.[29] Rubin[37] writes that "'summary' causal effects can also be defined at the level of collections of units, such as the mean unit–level causal effect for all units" (p.323). Morgan and Winship[38] suggest that "a variety of possible population–based (and

'collection'–based) definitions of potential outcomes, treatment assignment patterns, and observed outcomes can be used" (p.51).

As **DAGs** essentially depict the data generating process of a set of variables and have been developed to incorporate both individual–and area–level variables,[22] this suggests that it is possible to construct **DAGs** based purely at the area–level. For example **DAGs** that have incorporated area–level data have been used to investigate the following: "cooking and season as risk factors for acute lower respiratory infection";[39] "the association of cigarette price differentials with infant mortality";[40] "the association of community sanitation usage with soil–transmitted helminth infections among school–aged children";[41] and "does employee resistance during a robbery increase the risk of customer injury?"[42]

An important consideration when thinking about the level at which to conduct analyses is the ecological fallacy and its inverse, the individualistic fallacy.[43] The ecological fallacy is a **bias** that "may occur when an observed relationship between aggregated variables differs from the true, i.e. causal, association at an individual level" (p. 1).[43] That is, it is assumed that conclusions drawn from analyses at the aggregate level hold at the individual level. Simpson's Paradox is an example of the ecological fallacy where analyses on two or more populations does not generate the same conclusion as when analyses are conducted on the population as a whole.

There is an argument for conducting analyses at the level at which the research question is posed, i.e. the level at which any intervention would take place.[44] For example, if an intervention would involve the change of policy at the local area level, then one should conduct one's analyses on aggregated data at the local area level, but, if the intervention was to change a person's calorific intake then analyses should be implemented at the individual–level.

In a multilevel framework, however, analyses can respect the hierarchical structure of data and area–level attributes can be separated from individual–level attributes, irrespective of the level at which the research question is focused. This can help in overcoming the

ecological fallacy as cause and effect can be partitioned across levels, notwithstanding the potential for cross–level interactions (which are not within the scope of this thesis).

In analyses that are not conducted at the lowest level or when multilevel modelling is not employed, there is a risk of inferential **bias** if the research question cannot be clearly proposed as operating at an aggregate level. This remains an under–developed area of research and it is a challenge for causal inference; it will be returned to in Chapter 7.

2.4 Causal inference and statistical associations

Statistical analyses are data–driven and do not necessarily require any prior knowledge of the directions of the associations between variables. They often focus on maximising the proportion of explained variation in an outcome; the greater the coefficient of determination (R^2), the better. A reason for using a solely statistical approach to analysis is because in observational data there are often many measured covariates but the sample size may be small, resulting in poor convergence properties of the statistical models.[27] Another common consideration in solely statistical analyses (i.e. those not informed by *a priori* causal knowledge) is that of collinearity between covariates; the concern is that this will lead to numerical instability. However, the importance of considering the causal framework under study when specifying regression models in the presence of highly correlated data has been shown; when the correct causal structure is specified for a model the parameter estimates of the effect of interest are unbiased.[45]

Statistical dependence between an exposure and an outcome could be a result of one of the following (as outlined in[25]): random fluctuation, X caused Y, Y caused X, X and Y share a common cause, the statistical association was induced by **conditioning** on a common effect of X and Y, i.e. **selection bias**. As introduced above, causal diagrams can help researchers dismiss some of these statistical associations as being causal by assessing whether they are consistent with the data and the data generating process.

DAGs automatically include the temporal order of the variables, so if interest lies in the relationship $X \rightarrow Y$ the explanation that Y causes X can be dismissed. X and Y share a common cause can be dismissed as a causal explanation of the data because these common causes are **confounders** and their effects can be eliminated by **conditioning** on them in analyses. The last explanation, that a statistical association was induced by **conditioning** on a common effect of X and Y, is the result of **conditioning** on a common cause (i.e. a **collider**), which would be avoided if a causal diagram was considered providing the data provenance is well understood and any conditional selection of the data is completely known and accounted for within the **DAG**.

2.5 The difference between causal inference and prediction modelling

Both causal inference and prediction modelling are important for generating and testing hypotheses. However, the distinction between the two is not always made clear and prediction models are different to the models used for causal inference.[46] The aim of prediction modelling is to accurately predict the outcome of interest. When variables are included in a prediction model they are the ones that are likely to be associated with the outcome but not necessarily causally related to it. Methods for narrowing down a list of possible covariates to include in a prediction model can be automated (e.g. forwards/backwards step–wise regression) and the best group of covariates are selected for the final model. The group is chosen that is both parsimonious and maximises the amount of variation in the **outcome** explained. It is likely that adding covariates to the model will increase its predictive capability but will reduce the model's external validity, i.e. whether the same model would apply to another dataset or the general population.

In contrast, the aim of causal modelling is to estimate the causal association between an **exposure** and an **outcome** by removing all other hypothesised relationships which affect

that focal relationship. Methods for implementing this approach have been the focus of this Chapter so far. These methods rely on *a priori* assumptions and theory and cannot yet be automated. As these assumptions are not accounted for in prediction models they are not interchangeable with causal models; their goals differ.[47] Expert knowledge is an important aspect of causal inference methods which set it apart from prediction modelling. Experts are required for designing the research question, identifying/generating suitable datasets for analysis and in describing the causal structure of the data being studied.[48]

Related to the concepts of prediction modelling and causal modelling is the 'mutual adjustment fallacy'. In a prediction model it is often assumed that each covariate can be interpreted from a given model, whereas in a causal model this is not the case. The reasoning behind this is described in the next Section.

2.6 The mutual adjustment fallacy

Often, in epidemiological and health geographical studies, effect estimates are obtained from a single model and they are therefore presented in a single table (often the second table in a research article). Each individual effect estimate is then interpreted individually as if each estimate has the equivalent interpretation. This interpretation is flawed and is referred to as the 'Table 2 fallacy'[49] or the 'mutual adjustment fallacy'.[50]

By way of illustration, if the effect of X on Y is to be estimated and it is known from a **DAG** (Figure 2.6.1) that there exists only one **confounder** of this relationship, Z, then a regression model, $Y \sim X + Z$, can be run. If all the usual assumptions of linear regression hold, then the coefficient of X obtained from this model estimates the **total causal effect** of X on Y. The 'mutual adjustment fallacy' occurs when the coefficient of Z is interpreted as if it estimates the effect of Z on Y. In models with more variables, this fallacy occurs when it is assumed that all of the estimated coefficients have a similar interpretation with respect to the outcome, Y.

Looking closer at the **DAG** in Figure 2.6.1, it can be seen why this is problematic. When investigating the effect $X \rightarrow Y$, Z is a **confounder** and adjustment for Z removes all **confounding**, however, if the effect of interest was $Z \rightarrow Y$, X is a **mediator**. It was seen above (in Section 2.1.10) why adjustment for a **mediator** is not appropriate for estimating the **total causal effect**. Instead, the **direct causal effect** could be estimated in this way, i.e. the effect of Z on Y whilst X is held constant, but this can be quite different to the **total causal effect** of Z on Y.

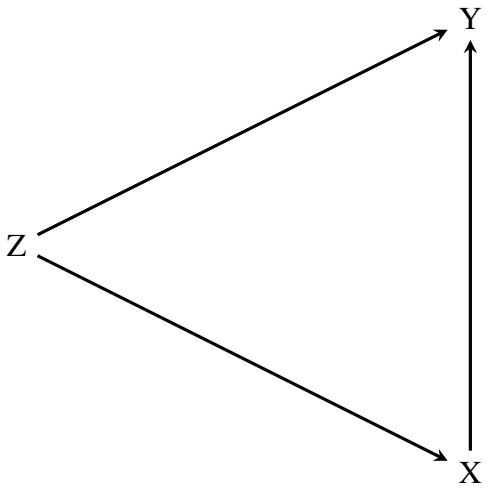

Figure 2.6.1: DAG illustrating the 'mutual adjustment fallacy' with three variables

This example can be complicated further if another variable, U, is considered which affects both Z and Y (Figure 2.6.2). In this case, if the total effect of X on Y was sought, it would still be appropriate to adjust for Z as it acts as a **confounder** of the exposure–outcome relationship. However, as U is a **confounder** of the $Z \rightarrow Y$ relationship, interpreting the coefficient of Z from the regression model $Y \sim X + Z$ would give the **direct effect** of Z on Y **confounded** by U which is not a very useful estimate.

Instead, a totally different model would be required to estimate the total effect of Z on Y to that of X on Y because X is the **exposure** in the latter and a **mediator** in the former.

This is true in general; different regression models are required if multiple causal effects are to be estimated.

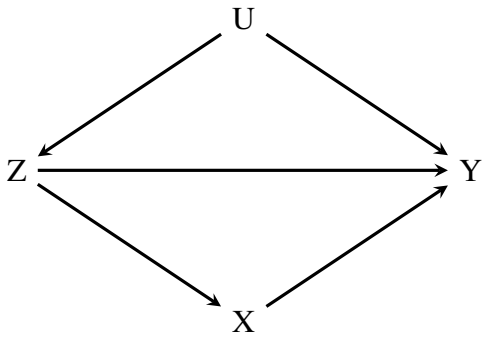

Figure 2.6.2: DAG illustrating the 'mutual adjustment fallacy' with four variables.

2.7 Simulation Studies

Simulation studies use computer experiments to assess uncertainty in observational data analyses. Simulated data allow one to know the 'truth' about the data that one is working with and the level of **bias** attributable to the statistical method used can then be assessed. This is often found by counting the number of deviations from what is expected, usually averaged over many iterations.[51] In social science research, the aim of simulation is to create a simpler model representing an observed mechanism from which conclusions can be drawn which are generalisable to the more complex real–world.[52]

There are several types of simulation that can be utilised and each has a different role and purpose; however, there is no specific simulation approach for the field of epidemiology or health geography. Some common approaches to simulation are: simulating from observed data distributions, microsimulation, and agent–based modelling.[53]

As mentioned in Chapter 1, many statistical methods are used habitually[1,2] and there are many flowcharts available that suggest which methods should be adopted depending on

the type of data available. However, this is not the only thing that should be considered when deciding which method is most appropriate; one important aim is to minimise **bias** in the methods one uses and "simulations can evaluate the robustness of a statistical procedure under ideal and non–ideal conditions" (p.34).[54] Simulation is considered in much more detail in Chapter 3.

2.8 Data Generation

Common to both simulation and causal inference, is the thinking behind the data generation process; in simulation studies one considers the relationships found in nature that one wishes to recreate in order to devise a model (an empirical abstraction of reality) that is simpler to study than the target data[52] and causal models "represent the mechanism by which data were generated" (p.35).[12] In this way, they are a sort of "blueprint" for the 'part of the universe' that researchers wish to simulate (p.35).[12]

The data–generating process is obtained from a **DAG** by supplementing the information contained within the **DAG** with parametric assumptions. The data–generating process is the way in which the endogenous variables within the system obtain their values. If all of the values of the **exogenous variables** in the system are known then the value of any **endogenous variable** can be found.[53]

Section 2.10 defines language that is important to the reading of this thesis. These words appeared in **bold** text throughout this chapter.

2.9 Conclusion

In this thesis, DAGs and causal graph theory will be used to illustrate causal relationships between variables and to shed light on biases that arise from the methods explored in the

upcoming chapters. These will be linked with theoretical simulations, based directly on simplified DAGs under the null hypothesis and, where appropriate, also on observed data.

Causal graph theory can provide important insights into research questions, not least by making researchers be explicit about their research question of interest. This chapter has introduced important background to the causal inference framework and how this is linked to simulation through the data generation process. The power of these methods was demonstrated by illustrating how they have been used to understand some paradoxes and biases in epidemiology.

These methods have not previously been used to look at the research questions addressed in this thesis (and are not a feature of health geography research more generally), namely: mathematical coupling (Chapter 4), Limiting Long–term Illness and deprivation (Chapter 5), and population mixing and childhood leukaemia (Chapter 6).

2.10 Definitions

Ancestor: A variable on a path which precedes another in time.

Arc: In a directed acyclic graph (DAG), an arc is a directed line (arrow) symbolising a hypothesised causal relationship between two variables.

Back–door path: A back–door path is an alternative path to the causal path of interest which connects the exposure and the outcome.

Bias: Bias occurs when the results or inferences of an analysis deviate from the truth.

Causal path: A causal path is a path between the hypothesised exposure and outcome.

Child: A direct descendant of another variable.

Collider: A collider between a pair of variables is any variable that is causally influenced by both variables in the pair. Conditioning on a collider can create a spurious (or

non–causal) association between its causes. In a DAG, a collider is in the middle of a fork, i.e. $X \rightarrow C \leftarrow Y$.

Composite Variable: Variables that have been algebraically constructed (e.g. by addition, subtraction, multiplication, or division) from two (or more) source variables.

Compositional Data: Data that are parts of a whole. Theoretically, all variables can be divided into 'parts'.

Conditioning (on a variable): Introducing a variable into an analysis, this could be through, for example, stratification or including a variable as a covariate in a regression model.

Confounder: Within a causal inference framework a confounder is defined to be:

- a cause of the outcome in unexposed people;
- a cause of the exposure; and
- unaffected by the exposure (not on the causal path between exposure and outcome.

In a DAG a confounder (C) is an ancestor of both the exposure (X) and outcome (Y), e.g. $X \leftarrow C \rightarrow Y$. If a confounder is conditioned on it will 'open' a back–door path.

Confounding: A mechanism that creates a non–causal path between exposure and outcome, i.e. the presence of at least one 'open' back–door path between the exposure and the outcome.

Consistency: One of the three assumptions required to identify a causal effect. Consistency requires that the exposure and outcome are sufficiently well–defined in order for the causal effect to be well–defined.

Counterfactual: An event, state, or situation, that did not happen but could potentially have happened.

Descendant: A variable on a path which follows another variable in time.

Directed Acyclic Graph (DAG): A graph that represents the causal assumptions of the data–generating process, i.e. the hypothesised causal relationships between variables. All defined relationships between variables have a direction (hence 'directed') and no variable can cause itself (hence 'acyclic').

Direct effect: The part of the effect between exposure and outcome that does not go through any intermediate variables.[55]

d–Separation: Two variables are d–separated if there is no open path between them in the DAG.

Endogenous Node: A node that has at least one direct cause in a DAG.

Endogenous Selection Bias: Bias that is a result of conditioning on a collider, or the descendant of a collider, on a non–causal path between the exposure and the outcome.[56]

Exchangeability: One of the three assumptions required to identify a causal effect. Exchangeability assumes that once all confounders are conditioned on, values of the exposure are randomly assigned.

Exogenous Node: A node that has no direct causes in a DAG.

Exposure: The variable whose causal effect is to be estimated. This can also be any conceivable concept that the population of interest 'experiences'.

Indirect effect: The parts of the total effect of an exposure on an outcome that is transmitted via intermediate variables.[55]

Interference: Under the assumption of no interference it is assumed that the outcome of one individual or unit is not affected by the treatment of any other individual or unit.

Mediator: A variable in the middle of a chain of causal arrows between the exposure and the outcome. A mediator 'mediates' part of the total causal effect between the exposure

and the outcome.

Minimally Sufficient Adjustment Set (MSAS): Sets of covariates, that when adjusted for, block all back–door paths between the exposure and outcome.

Node: A variable in a DAG.

Null hypothesis: The hypothesis that there is no relationship between the exposure and outcome of interest.

Outcome: A variable whose causal determinants are to be estimated.

Path: A path is a set of arrows connecting any two variables in a DAG, regardless of the direction of the arrows.

Parent: An immediate ancestor of another node.

Positivity: One of the three assumptions required to identify a causal effect. Positivity assumes that an individual or unit has a positive probability of receiving all values of the treatment (or exposure) variable.

Randomised Controlled Trial (RCT): A study design in which participants are assigned to groups to test a treatment. One group receives a new treatment whilst the other receives an alternative or no treatment. The groups are compared after a period of treatment to determine whether there are any systematic differences in outcome between them.

Selection Bias: Selection bias occurs when selection of observations into a study are not independent of the outcome.

Total Causal Effect: Combined direct and indirect effects between exposure and outcome.[55]

Chapter 3

Simulation in Health Geography

3.1 Introduction

This chapter introduces the history of health geography data and considerations regarding data provenance before proceeding to discuss simulation methods and related issues. Everyone can perform simulations; they do not need to be "restricted to researchers with advanced skills in statistics and computer programming" (p.43),[54] however, they do need careful planning and execution[51] as well as a healthy dose of careful forethought and reflection on the real world as it pertains to the data generating processes in any one context. This chapter shows how researchers can build up their simulations, informed by causal inference methods, by illustrating some of the issues mentioned in Chapter 2, discussing some important considerations resulting from the research reported later in the thesis and outlining how simulations are approached throughout the rest of the thesis.

3.2 A brief history of epidemiology/health geography data

The collection of epidemiological and health geography data has a long history. The original, pioneering collection of population health related data is often credited to the demographer John Graunt. Together with Sir William Petty, Graunt collated the initial life–tables: calculating survival probabilities for each age, he later went on to write the 'Bills of Mortality'.[57] Since ancient times, societies have tried to collect data on population attributes, and even 'Before the Common Era' censuses were being conducted by those in charge. Although these may have been initially instigated to maximise tax collection, census questions have developed and the utility of collecting data on the same questions periodically, aiming for 100% population coverage, should be recognised. Since 1991, the UK Census has collected data on health outcomes, for example, morbidity (in the form of a self–assessed question on whether a person considers themselves to have a limiting long–term illness) which complements the Vital Registration data on mortality events.[58]

Often, in health geography research, researchers rely on secondary data sources, sometimes called 'routine' data. These are data which are collected for administrative purposes and consequently they are collected without a research question or hypothesis in mind, but they are later made available for research purposes.[59,60] Some data sources which are readily available to researchers are outlined in Sections 3.2.1–3.2.3, below.

3.2.1 Census data

As mentioned above, the idea of collecting data about a population is to get information about 100% of the population at a single point in time. Census forms are sent out to each household to be filled in on the same day. This process is usually repeated every

5 or 10 years with the questions and general topic areas being repeated at each census to get an idea of changes in the population over time. The collation of census data is a long process, and a criticism of collecting this form of data is that it may be out of date before it is available to researchers for analysis. However, this form of data collection does allow analysis of the size and key characteristics of the population at several different area–levels; from national level to output areas.[61]

3.2.2 Survey data

Large scale surveys are often conducted by government departments motivated by a need to know more about a particular topic such as health, crime, labour force participation, etc. Unlike a census, these studies can be conducted over several months by professional interviewers and even though the survey may be repeated another year, different people will be interviewed. Efforts are often made to ensure a similar cross–section of the population are surveyed. The benefit of the data collected by this kind of survey rather than a census is that a larger variety of questions can be asked and the data can be released to researchers much more quickly. These data can often be linked up with census data because the cross–section of the population that is chosen for questioning is often informed by the census to ensure sufficient coverage of population types. As introduced in Section 2.4, there are serious implications when this kind of survey does not accurately include a representative sample of the population that the researchers are interested in; this can be exacerbated by non–response bias.[61]

3.2.3 Administrative data

Administrative data are data that are collected for "an organisation's activities"(p.22).[61] These organisations are often government departments collecting large amounts of administrative data regarding areas such as welfare, tax, health and education. They are

often used to inform official statistics which are then used to inform government policies. When they are released for use by others they are often aggregated into census or electoral geographies which means that they can be linked to census data straightforwardly. These datasets can be much bigger than datasets collected by surveys because the data are collected routinely which would be expensive and logistically difficult to conduct by survey or census. This type of data often includes parts of the population that are difficult to reach by survey (e.g. because they are in temporary accommodation), this means it can be particularly useful in identifying areas of society which need more investment or healthcare interventions, for example.

A difficulty with using administrative data for research purposes is that the researcher has little control over how and what data are collected. They can also miss out certain sub–populations (e.g. homeless people) because they do not use certain services or when changes are made to service provision.[61]

3.2.4 Future data sources

There has been a lot of discussion around so–called 'Big Data' and its possible utility for aiding researchers by overcoming some of the shortfalls in the data types mentioned in the previous sections. Big Data has been defined as data that is high in volume, velocity and variety and therefore requires new technology and analytical methods to make use of it.[62] Big Data can offer very detailed information in near real–time from a variety of different sources, but this can present challenges for researchers as existing methods of analysis may not be adequate because of the volume of data available.

With these evolving methods of data collection and collation, and as the research focus moves more towards 'Big Data', researchers should be mindful of what this means for their research; 'Big Data' can offer researchers exciting opportunities, but there is a danger that it could confuse rather than clarify research,[61] especially as data provenance has

implications on the robustness of causal inference sought, as will be shown in Chapter 6.

This chapter discusses simulation and its usefulness in assessing the statistical methods used for bias and the role that causal inference can play in this. This thesis focuses on more traditional data sources rather than entering the world of 'Big Data' as there are a lot of issues that still need to be addressed within the fields of health geography, causal inference and simulation before these issues can be researched in relation to 'Big Data', particularly regarding the merging of multiple datasets.[63]

This thesis primarily uses census data to inform simulations and upon which to conduct direct analyses. This reduces some of the difficulties due to selection bias resulting from other data sources and allows focus to be directed at some other potentially more analytical biases.

Although there are many advanced methods available for analysing health data within a geographical context (e.g. geographically weighted regression[64]), simpler methods are often called upon to get a 'quick' idea of the data. It will be seen in Chapter 5, how this approach can produce erroneous results when investigating the relationship between deprivation and limiting long–term illness. It appears as though more advanced methods are sometimes considered to have diminishing returns on effort and that regression models are 'good enough'.[50] Throughout this thesis, it will be shown why this approach should either be avoided, as it can lead to incorrect conclusions and prompt further research to follow the wrong direction, or how simple simulations and a graphical causal framework can be combined to avoid these analytical pitfalls. This will be achieved by re–visiting some go–to methodologies using simulation and approached from a causal inference perspective to illustrate how they can be accurately implemented, and to highlight more contemporary methods that can be used to overcome biases. If researchers often revert back to regression modelling[1] rather than more 'advanced' methods then it is important that regression modelling is integrated well with causal inference methods and simulation

to avoid biases.

3.3 Where to begin: the data generating process

It is often reported that there are four elements required for data analysis, these are: data collection, data collation, data cleaning followed by data analysis (and subsequently dissemination of results). However, in order to think about addressing research questions from a causal inference perspective it is important to also think about data provenance, i.e. the (actual or assumed) data generating process and therefore the causal relationships between the variables of interest. This can be greatly aided by the use of Directed Acyclic Graphs (DAGs - introduced in Section 2.1.7). It is not always possible to tell directly from the data what the correct data generating process is and in that case either external knowledge or underlying theory is then needed. Model development must be driven by *a priori* understanding of the data generation process and the same is true when simulating datasets - they should be informed by the assumed data generation process and where this is not known, several different data generation processes can be used and compared. The distinction between DAGs drawn to represent observed data may not be known *a priori* and examination of DAG–data consistency may then be insightful for confirming plausible and therefore likely correct underlying data generating processes in line with external theory or hypothesis.[30]

Increasingly, population health data scientists rely on data collected from one or more external sources which may have to be combined together in order for analyses to be performed. These data are often multifaceted and can be at the individual–or area–level.

The most contemporary methods of analysis may suggest that multilevel models are most appropriate to address a particular research question, especially in health geography, however, it may be the case that the data are not in a format that facilitates this. It is important, then, to think about the methods that are at the researchers' disposal and to find

the best way for these methods to be implemented in order to minimise bias. This is not as straightforward as acknowledging the format of the data available and implementing "mechanical calculations with little attention to scientific context, experimental design, assumptions and limitations of methods, or the interpretation of results" (p.42).[1] It is paramount to accurately attributing cause and to inform subsequent administering of interventions. For better or worse, researchers will always be required to answer research questions based on the data that are available to them.

3.4 What are simulation studies?

The simulation studies used in this thesis generate data by "pseudo-random sampling from known probability distributions" by computer experiment (p.2074).[51] One purpose in relation to the analysis of health data is to enable the researcher to evaluate their methods in order to assess the suitability of that method to answer the research question at hand. Simulations are used in the case where mathematical proofs are not suitable, i.e. where no closed form solution exists or the problem is intractable[54] but they can also be used to supplement closed form solutions by providing empirical evidence.

In terms of causal inference studies, mathematical solutions do not necessarily lend themselves well to respecting the causal data generation process; this is much more easily achieved by using simulation methods where parameters can be changed for code to be re–run in a matter of seconds.

3.5 Simulation studies informed by observed data

In a causal inference context, observed data informed simulation studies allow researchers to take the observed data on exposures and confounders and then simulate outcomes

under different causal scenarios. For the simulated outcomes, the true answer is known and investigations can then be made into how alternative methods compare to otherwise used methods and the extent to which the use of one approach over another alters the concluding inferences of any analyses. Simulation studies are a powerful technique that allow researchers to answer a broad set of methodological and theoretical questions and "provide a flexible framework to answer specific questions relevant to one's own research" (p.43).[54] When creating a simulated dataset, based on observed data, relevant information is extracted so that the simulated data is statistically equivalent to the observed data. This means that distributions of the simulated data are quasi–identical to the distributions of the variables in the observed population and the marginal distributions are accurately represented via the correlation structure.[65]

The purposes of simulation studies in this setting are not to replace formally collected datasets and they do not reduce the need to collect more and better data. The suggestion in this thesis is that simulation studies should be used to supplement applied analyses in order to identify and avoid inferential biases from the limitations of methods employed for complex real–world analytical challenges.

3.6 What is the purpose of simulation for assessing statistical methods?

The purpose of simulation is to remove as much superfluous uncertainty from data as possible. Messiness is removed from the data and the researcher can have a certain level of confidence that any bias that is found is due to the methods that have been implemented for analysis. The data generation process is known, therefore methods can be assessed by comparing analytical results from a simulation with those results that are known to be 'true' because they were simulated to be a particular way. This allows bias to be estimated by comparing the estimates from an analysis with the known, true values. It is important

to have dedicated research questions to investigate how appropriate the methods used are for the purpose they are being used for.[66]

It is also important to take time to step back and look at the methodology to understand whether the methods used are fit for purpose, i.e. how robust are the methods being used? Have they somehow been taken out of context: has the "ritualistic miming of statistics rather than conscientious practice" (p.40)[1] taken over? Indeed, Maxwell and Cole[67] describe this well when they say, "simulations can be extraordinarily valuable because they allow the author to describe properties of statistics under suboptimal conditions where underlying assumptions have not been met" (p.196). However, simulation is not always straightforward. Some important considerations relevant to the health geography situations in the rest of the thesis are now discussed, starting by discussing the criticisms often levelled at simulation studies.

3.7 Foreseeing criticisms of simulation

It gets more and more complicated to find closed form solutions when dealing with complex real–world scenarios[45] and one criticism of simulation is that it is abstract and does not relate to data found in the 'real–world'.[68] For the simulations here and throughout this thesis, observed datasets are used to inform the simulations. However, it can sometimes be useful to build simulations 'from the ground up' by using hypothetical data as examples.

Simplifying scenarios as much as possible can help one understand what is truly going on. Simplification, is in fact a necessary element of simulation, as one must understand one's data to know the truth for comparison with analytical results.[51] Building a simulation up from scratch, as much as it is feasible, can help to avoid unintentional marginal constraints which might accidentally be introduced when using a top down approach to simulation. The simulations used in Chapter 4 are an example of simulations being built up, whereas

those in Chapters 5 and 6 are based on existing data. Since a statistical model is an abstraction of reality and should be parsimonious, this leads on to questions around how complex the simulation should be.

3.8 How complex should the simulation be?

There are several things to consider when deciding how complex a simulation needs to be to get a true picture of how biased a particular statistical method is. It can be helpful to build–up simulations to fully understand the processes and where/how bias is introduced.

The datasets that are generated via simulation are often 'cleaner' than those that are encountered in the real–world as they are often generated under unrealistic conditions.[54] Whether these more complicated aspects should be incorporated in simulations is sometimes difficult to gauge and largely comes down to the knowledge of the researcher, however, it gets more and more complicated when research is dealing with complex real–world scenarios.[53]

When deciding how complex the researcher should make a simulation they may want to consider what particular aspect of a simulation they are interested in and avoid being distracted by other issues or nuances that are not directly pertinent to the research question at hand.

As a simplification in this thesis, latent variables are either one of the last issues to be considered in the simulations (Chapter 4) or are considered only abstractly or in the discussion (Chapters 5 and 6). Latent variables are unobserved variables –they could be unobserved for several reasons, such as: it is not possible to measure them, they are missing in the dataset that is available for analysis, or even that they are not known (but can be assumed) to exist. In observational data analysis, latent variables complicate analyses and it is important to be aware of them and account for them where possible,

or to be explicit about the assumptions that are made in relation to omitting or including them. Models can only ever be approximations of the 'truth' and it can be assumed that simulations including latent variables (when they are not directly relevant to the research method being investigated) would only serve to complicate results and distort the bias attributable to the use of an inappropriate method.

Later, it is realised that a latent variable is essential to capture cluster heterogeneity though it may not always be viewed as a 'variable', rather an intrinsic part of the data generating process. Of course, simulations are most valuable when they can capture the complexities of real–world data generating processes,[68] and the rest of this chapter introduces some approaches to simulation and discusses some of the complexities in simulating health geography data which are important for the rest of the thesis and for such research, generally.

3.9 How to simulate?

Using the data generation process that is common to both thinking behind simulation and causal inference, a graphical causal model (of which a DAG is an example) can be adapted to a path diagram using Sewell Wright's path tracing rules.[69–71] Data can then be simulated from this simple model where the truth about the underlying relationships in the data are known, i.e. it is known what regression coefficients and covariance structures should be generated from any models that are run. This is a straightforward way of simulating under the null and non–null scenarios. This method is used in Chapter 4 to illustrate the problem of mathematical coupling.

Path diagrams are now defined before some other complexities are introduced to achieve simulations which are more truly representative of reality and avoid criticism as mentioned in Section 3.7.

3.9.1 Path diagrams

DAGs are non–parametric, this means that in order to generate simulated data informed by a DAG some assumptions must be made which allow the DAG to be developed into a path model.[72] A simple way to do this is to assume that all of the variables in a DAG follow a multivariate normal distribution and that all relationships between the pairs of variables in the DAG are linear. The arcs on a DAG can then be assigned standardised path coefficients. These path coefficients are not to be confused with correlation coefficients as they do not represent bivariate correlations.

A way of deconstructing the correlations among a set of variables to estimate standardised path coefficients was developed by Sewell Wright.[69–71] To find the correlation between any two variables in a path diagram, all routes connecting the variables must be traced using the following 'path tracing' rules:

- trace backward along an arrow and then forward, or only forwards from one variable to the other, but never forward and then back,
- pass through each variable only once in each chain of paths, and
- trace through at most one two–way arrow (correlation) in each chain of paths.

The contribution of the correlation of each chain traced between two variables is the product of the standardised coefficients in the chain. Wright[73] reports that the primary purpose for the method of path coefficients was to combine the quantitative information from a system of correlation coefficients with knowledge of their causal relations. This method provides a fairly simple approach to simulating data according to a DAG and making certain assumptions about the nature of the quantitative relationships between variables when only a few variables are being considered. However, it can become extremely tricky to implement this method when there are many variables and many

relationships between those variables; once there are 5 covariates in a DAG there are up to $2^5 = 32$ bivariate relationships, but by doubling this to 10 covariates there are up to $2^{10} = 1024$. This is why programs and `R` packages, such as 'dagitty',[30,74] that are able to perform these tasks for the researcher, are so useful.

As an example, consider the DAG in Figure 3.9.1 where each arrow is labelled with a standardised path coefficient. The bivariate correlations between each pair of variables are calculated using the above path tracing rules in the following way.

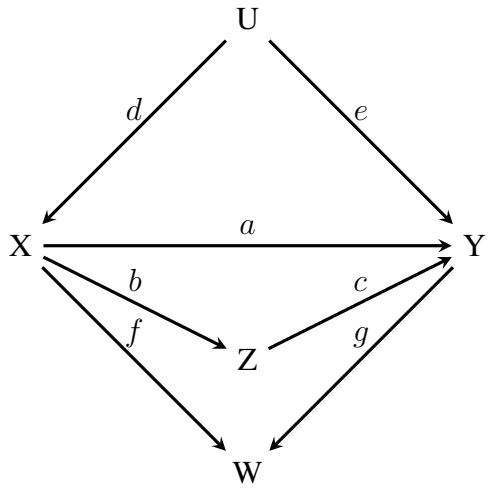

Figure 3.9.1: Example DAG with path coefficients displayed on edges.

Correlation between X and Y: $r_{XY} = a + bc + de$

Correlation between X and U: $r_{XU} = d$

Correlation between X and W: $r_{XW} = f + bcg + ag + deg$

Correlation between X and Z: $r_{XZ} = b$

Correlation between Y and U: $r_{YU} = e$

Correlation between Y and W: $r_{YW} = g + af + edf$

Correlation between Y and Z: $r_{YZ} = c + ab + edb$

Correlation between U and W: $r_{UW} = df + dag + dbcg$

Correlation between U and Z: $r_{UZ} = db$

Correlation between W and Z: $r_{WZ} = fb$

3.9.2 Simulating using 'dagitty'

The R package 'dagitty'[30,74] can simulate simple datasets once it is given certain information. These are: the variables that the researcher wishes to simulate, the causal structure between those variables (i.e. the DAG), and the covariance matrix of the variables. This assumes that the variables are linearly related and follow standard normal distributions but they can be transformed subsequently, for example, a nominal value can be added to avoid negative values.

The covariance structure of the variables is related to the path coefficients introduced in Section 3.9.1. It is important to consider the choices made when defining this covariance structure for simulation to make sure it represents what it is supposed to and does not produce fluke or confusing results, as will be discussed in Chapter 4.

3.9.3 Simulating directly

It is more complicated to simulate non–normal data with a specific covariance matrix because simply applying a non–linear transformation to variables generated in the way described above will usually change the target correlation matrix. Ruscio and Kaczetow[75] provide (freely available) full R code to implement an algorithm that allows simulation of non–normal correlated data. Their algorithm generates data from specified distributions and iterates through intermediate correlation matrices until the target matrix is reproduced. This particular algorithm takes the specified number of variables to generate, the population distribution for each variable, the sample size, and the target

correlation matrix as arguments from which it generates datasets. It can also take an observed dataset directly and calculate the sample size and number of variables to generate and resample with replacement from the supplied variables to generate a new dataset, however, upon experiment with this method it was found that the variation in datasets generated was not sufficient. The idea behind simulating in this way is that many different worlds are generated (i.e. one for each iteration of the simulation) which could have occurred under the specified variable distributions and covariance structure; this is not achieved by resampling from the observed dataset. This algorithm is used to simulate data in Chapters 5 and 6 using assigned distributional assumptions that are hypothesised to fit the underlying data generating process.

The following sections will go on to consider some specific simulation considerations which are pertinent to the work that follows in the rest of the thesis.

3.10 Specific health geography simulation considerations

3.10.1 Simulating compositional/composite data

Later in the thesis the use of ratios, composite variables and compositional data will be investigated. These three concepts are intrinsically linked to the data generation process, and some of the more philosophical insights will be introduced here before discussing them in a more health geographical context later.

These relationships are not usually considered in DAGs and this is likely because the relationships between these tautological variables are *deterministic* rather than *probabilistic*.[76] How these variables can be dealt with in a causal framework is different to how probabilistic variables are normally handled and this is briefly considered here.

Considering the DAG in Figure 3.10.1, in which the exposure, X, is a probabilistic

function of the variable, C, and the outcome, Y, is a probabilistic function of both X and C. C is a classical confounder of the $X \rightarrow Y$ relationship in this case. From the back–door criterion (Section 2.1.8) it is known that the total causal effect of X on Y can be estimated by conditioning on C.[12]

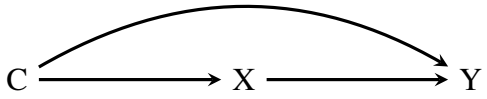

Figure 3.10.1: Causal Diagram in which C is a confounder of the exposure–outcome relationship between X and Y. X is a probabilistic function of C.

Now, considering the scenario where X is a deterministic function of C rather than a probabilistic one (indicated by the double circles around X;[77–80] Figure 3.10.2). Without further information, it may appear as though C is still a classical confounder in this case, however, the relationship between C and X is tautological which means that conditioning on C, as if it were a confounder, reduces the association between X and Y to zero. For any DAG containing a variable that is fully determined by its parents, conditioning on the parents makes X independent of all other variables in the DAG.

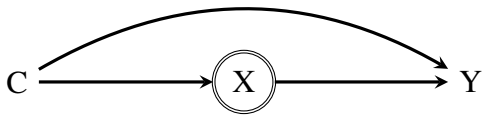

Figure 3.10.2: Causal Diagram in which C is a confounder of the exposure–outcome relationship between X and Y. X is a deterministic function of C; indicated by the double circle around X.

To accommodate deterministic variables in DAGs, Geiger[80] developed D–separation (capital 'D'–separation) and 'Deterministic Node Reduction' which ultimately results in eliminating the fully determined node from the DAG by passing the arcs starting from this deterministic node back to its parents. This algorithm provides a way of making a DAG

compatible with the usual rules of d–separation, however, there are many circumstances in which deterministic nodes are important/favoured in analyses, not least in the context of health geography; researchers have become very familiar with using these deterministic concepts in their analyses (e.g. socio–economic position or deprivation indices). In the course of the research for this thesis, it was discovered that these issues require closer consideration for both causal inference and simulation and this can be split into the case of composite variables or compositional data.

3.10.2 Composite Variables

When considering building up a simulation, composite variables may have to be considered. Composite variables (whether at individual–or area–level) are constructed from two or more parent variables and cannot be measured directly (and arguably do not exist before they are constructed by the researcher). The components used to construct composite variables may be on different scales, which means that the composite has its own unique scale. For example, body mass index (BMI) which is formed of weight (measured in kilograms) and height (measured in metres) squared, creating a composite variable which is on a scale measured in kilograms divided by height squared.

Composite variables are generally constructed for one of the following purposes: (1) to create a variable that aims to summarise multiple related concepts in a convenient or parsimonious way (e.g. a deprivation index), or (2) to standardise one variable by another (e.g. GDP per capita). The distinction between these two purposes is not trivial and indeed has important implications for determining the appropriate analytical strategy. On one hand, summarisation implies an interest in modelling and understanding the average effect of a series of related concepts on an outcome of interest, whilst, on the other hand, standardisation implies an interest in modelling the effect on the outcome of an individual variable, *conditional* on another variable thought to confound the focal relationship, i.e. to standardise a measure with respect to a perceived ‘norm’, such as average body height

or a typical cross–section of society.

To illustrate the implications of this, the 'effect' of body mass index (BMI) on risk of cardiovascular disease (CVD) is considered. Although use of BMI is ubiquitous in health and medical research, the fact that it is an algebraic construct determined by weight and height is often overlooked. Using deterministic notation, this scenario can be depicted by the DAG in Figure 3.10.3.

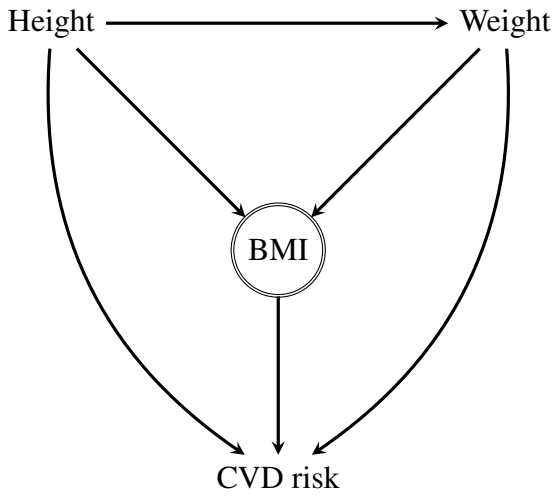

Figure 3.10.3: Causal Diagram indicating the relationships between height, weight, body mass index (BMI) and cardiovascular disease (CVD) risk. BMI is a deterministic function of height and weight indicated by the double circles around this variable.

The application of Deterministic Node Reduction produces Figure 3.10.4, which is mathematically equivalent to Figure 3.10.3. If there is less interest in the individual effects of height and weight, or the source information on height and weight is no longer separately available, the scenario might be depicted with the DAG in Figure 3.10.5. Which of the two DAGs, Figure 3.10.4 or Figure 3.10.5, is most appropriate to answer this causal question depends upon the value and meaning given to the average causal effect of the composite exposure, bearing in mind that a composite variable is not itself a measurable feature of nature.

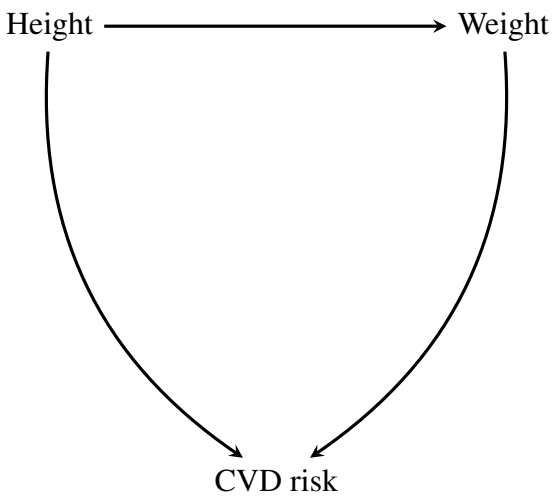

Figure 3.10.4: Causal Diagram produced by applying the Deterministic Node Reduction algorithm to Figure 3.10.3.

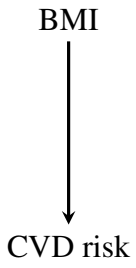

Figure 3.10.5: Causal Diagram produced by removing deterministic parents of BMI from Figure 3.10.3.

Fundamental to this inquiry is whether BMI represents a meaningful summary of height and weight which serves as a useful proxy for another more clearly–defined concept (e.g. adiposity), or whether it is simply a measure of weight standardised by height (to account for the fact that taller people are generally heavier). If BMI is considered to be a valid and useful proxy for adiposity, then analysing the composite as a distinct variable is arguably acceptable. If, however, BMI is considered to be a measure of weight standardised by height, then it must be carefully considered whether $\frac{\text{weight}}{\text{height}^2}$ represents the most effective

parameterisation of this relationship, and whether the causal effect of weight is really captured by the variable BMI. The total causal effect of BMI on CVD risk will likely differ from the total causal effect of weight on CVD risk, conditional on height; although both can be theoretically estimated without statistical bias, inferential bias may result if the effect estimate obtained does not accurately reflect the causal mechanism that the researcher seeks to understand and may eventually wish to target for intervention. This issue is returned to in the health geographical context of composite variables constructed to capture the latent concept of deprivation in Chapter 5.

3.10.3 Compositional Data

Compositional data differ from composite variables in that the individual components of compositional data can be measured directly and on the same scale as a larger whole, or subdivided into smaller parts. An example of this would be total number of calories consumed, divided into calories from fat, protein and carbohydrates. This does not pose a particular issue unless interest lies in the role of one or more components in relation to the whole.

As in the case of composite variables, there is a tautological relationship, this time between the component variables and the total variable. For example, if a researcher was interested in the causal effect of the economically active population on gross domestic product (GDP), they might consider analysing this by conditioning on the total population, or not (Figure 3.10.6). The utility of these two approaches depends on context.

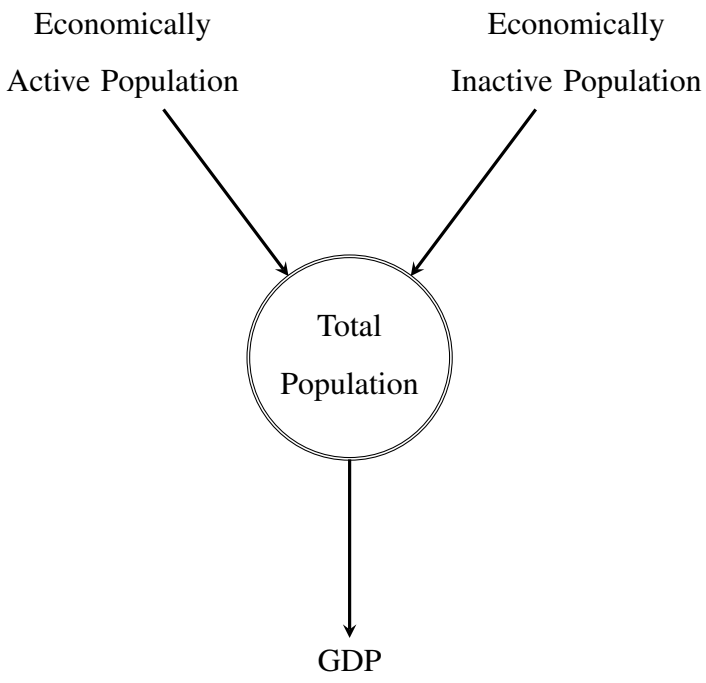

Figure 3.10.6: Causal Diagram indicating the relationships between the economically active and inactive populations, the total population and gross domestic product (GDP). Total population is a deterministic variable calculated by adding the economically active and inactive population together; this is indicated by the double circle around total population.

The effect of the economically active population without conditioning on the total population represents the average change in GDP that results from adding economically active individuals to the area, thereby increasing both the number of economically active individuals and the total number of individuals, whilst doing nothing to the population of economically inactive individuals. An estimate of this effect may be desirable if, for example, the government were considering a policy aimed at increasing immigration.

In contrast, the effect of the economically active population whilst simultaneously conditioning on the total population represents the average change in GDP achieved by swapping economically inactive individuals for economically active individuals –either by adding economically active individuals and removing an equal number of

economically inactive (different) individuals, or by effectively converting economically inactive individuals to economically active (same) individuals, or some combination of both.

This effect is therefore a combination of the effects of both subgroups on GDP –the positive effects of simultaneously increasing the economically active population and decreasing the economically inactive population by equal numbers, thereby retaining the same overall total population. An estimate of this effect may be desirable if, for example, the government were considering implementing a job–training programme for currently unemployed individuals.

In this scenario, both the effects reflect the population–level average effects of changing the relative numbers (i.e. the proportions) of economically active individuals to alter GDP, but by different mechanisms; they therefore reflect distinct causal quantities, the utility of which must be determined by context.

Ordinarily, conditioning on a collider may be considered as introducing 'collider bias' (Section 2.1.11) into an analysis, however, in this context, conditioning on a collider provides an interpretable causal quantity which has real utility in certain situations.[81]

When simulating composite variables and compositional data it is necessary to simulate the smallest components of interest and use these for construction of the variables of interest. In the case of composite variables, this is because the composite is not measurable directly in nature and its unit of measurement is combined from its components. In the case of compositional data, this is because the whole is a constraint on its components.

When causal inference is undertaken it has been recommended that the (often hypothetical) intervention is well–defined.[19] This is linked to the key condition of consistency introduced in Section 2.1.2. Returning to the example of BMI and CVD risk, there is a question over whether BMI is a useful proxy of a more clearly–defined concept (e.g. adiposity), or whether it is simply a measure of weight standardised by height (i.e.

to account for the fact that taller people are generally heavier). If BMI is considered to be a useful proxy for adiposity, then it is possible that analysing it as its own variables is acceptable. However, if BMI is considered to be a measure of weight standardised by height, then it must be considered whether $\frac{\text{weight}}{\text{height}^2}$ truly represents the most effective parameterisation of this concept. The total causal effect of BMI on CVD risk will likely differ from the total causal effect of weight on CVD risk, conditional on height; although both can be theoretically estimated without statistical bias, there is a risk of inferential bias if the effect estimate obtained does not accurately reflect the causal mechanism that is sought and may eventually be a target for intervention.

Whether 'obesity' can be interpreted as a definable exposure with an identifiable causal effect has previously been challenged; in particular, there are concerns that obesity fails to satisfy the consistency assumption required for causal inference (Section 2.1.5) because it can represent multiple states, including high adiposity and high muscle mass.[19,82] The same concern is relevant for BMI (and all composite variables) since any value of the composite may represent various combinations of the determining component parents.

Hypothesising that BMI 'causes' an increased risk of CVD implies that intervening to lower BMI would result in a decreased risk of CVD. Theoretically, this could be achieved by lowering BMI by either decreasing weight or increasing height. Realistically, however, weight is the more likely target for intervention. Regardless of the philosophical perspective on the utility and validity of BMI this suggests that it might actually be more useful to estimate the causal effect of weight adjusted for height.

These issues are considered for each simulation based on observed data later in the thesis, with a focus on health geography data and further complexities that may arise from composite variables are explored in the final chapter.

In the next section, an illustration of the 'modifiable areal unit problem' (MAUP) is given as this is important for the upcoming chapters, particularly Chapter 6.

3.10.4 Random events and the Poisson distribution

As a simple illustration (adapted from[83]), imagine a country that is 10000km^2, with 300 cases of a disease randomly distributed across it. If the area is divided into equal sized areas that are 100km^2 (Figure 3.10.7) the cases in each square of the grid can be counted (Table 3.1). In this example, there is one area with nine cases, six with seven cases and four with no cases. There is an implicit assumption in this example that the population of each square of the grid is the same and that every square has the same probability of having a case.

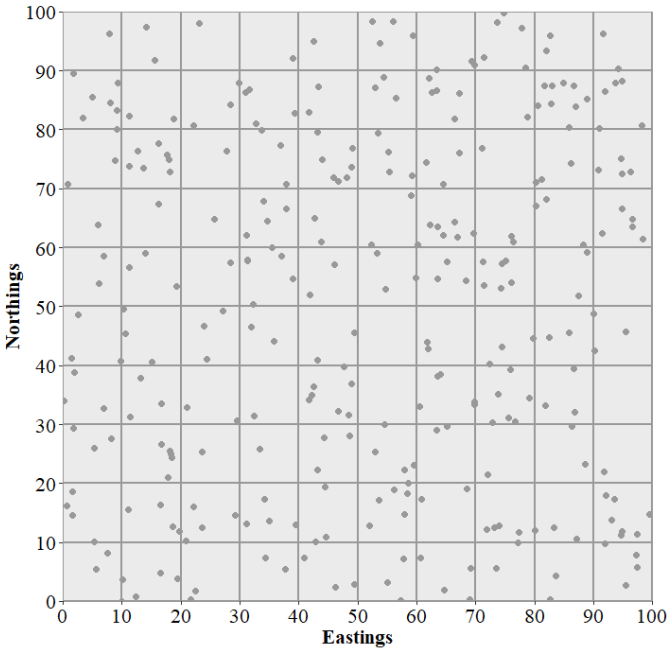

Figure 3.10.7: Plot showing randomly generated cases of a disease over a hypothetical 10000km^2 country divided into equal areas of 100km^2 each.

Table 3.1: Count of the number of areas with each number of cases.

Cases	0	1	2	3	4	5	6	7	8	9
Number of Areas	4	17	23	25	11	11	2	6	0	1

Highlighting the areas with large numbers of cases and those with none (Figure 3.10.8), it can be seen that there is no pattern to the numbers in each of the areas. The cases across the areas follow a Poisson distribution; most areas will have a number of cases that are at the lower end of the distribution, but as this is a skewed distribution (and it therefore has a tail) there will be a few areas that have a very high number of cases that thus stand out.

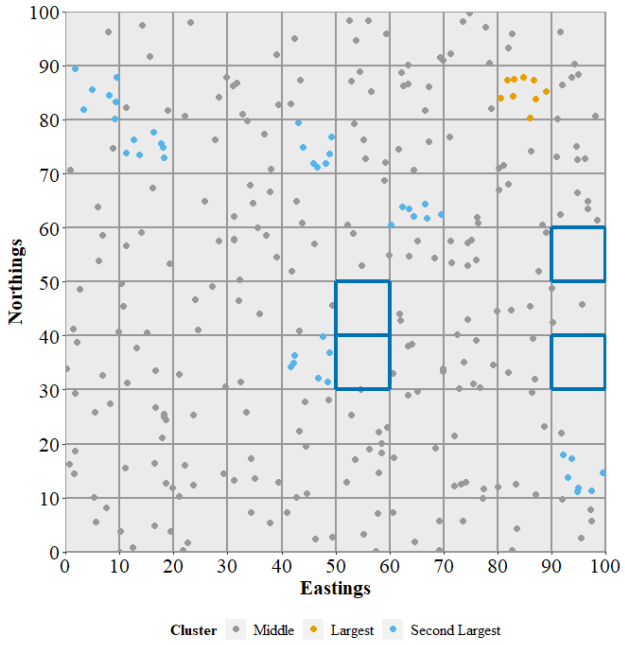

Figure 3.10.8: Plot showing randomly generated cases of a disease over a hypothetical 10000km^2 country divided into equal areas of 100km^2 each. Cases in areas with the largest and second largest number of cases are highlighted along with areas that have no cases at all.

Thinking about this problem, it can be seen that if the area boundaries were drawn differently, it would be possible for the boundaries to be selected in order to create areas with higher and lower numbers of cases. This problem has been reported widely, and is known as the Modifiable Areal Unit Problem.[84,85] Code for this example is available in Appendix A.

A similar effect can be generated when choosing the unit of time for analyses. It will be seen in Chapter 6 that, in the case of a rare outcome, a 5–year period is most often chosen

for analysis to avoid certain analytical problems, however, if this was changed, radically different results could ensue. This is known as the Modifiable Temporal Unit Problem.[86]

3.10.5 The 'most dangerous equation'

The problem introduced in Section 3.10.4 is linked to what has been called the 'most dangerous equation'.[87,88] The 'most dangerous equation' is based around De Moivre's equation which calculates the standard deviation of the sampling distribution of the mean (Equation 3.10.1).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \tag{3.10.1}$$

where $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of the mean, σ is the standard error of the mean and N is the sample size. The equation states that the standard error of the mean is inversely related to the sample size. This means that the larger the sample size, the less the sample mean will vary. If the example of Section 3.10.4 above is expanded to be closer to the real–world where area boundaries are drawn for various different reasons, e.g. for political reasons in the case of electoral wards, there will be a variety of population sizes across the areas. This means that areas with smaller populations are more likely to have high and low occurrence of disease by chance, whereas those areas with larger population sizes are more likely to have a stable disease occurrence through time.

This issue is something to be aware of when singling out areas for their high or low occurrence of disease without accounting for their relative size. Often areas are singled out due to extreme values of an outcome, however, little or no thought is given to the nature of the areas at the other extreme. Both of these areas may have similar characteristics.

3.10.6 Simulating under the null hypothesis

The simulations in this thesis predominantly assume the null hypothesis, i.e. that there is no relationship between the measured covariates and the outcome being investigated. This also assumes that any geographical/spatial factors do not influence the outcome which is highly unrealistic but the effects of geographical/spatial factors will be assumed to be superfluous to the individual factors being investigated, except where discussed explicitly. It is thus assumed that geographical/spatial factors have only a modest influence as 'noise', adding 'messiness' to the real–world data, but in some instances, as will be discussed explicitly, this assumption may need to be challenged and more complicated simulations may be warranted. This is explored briefly in Chapter 4. The purpose of the simulations is to determine whether truly biased results occur before further complexity is accounted for. Under the non–null scenario, these factors would need to be carefully considered and this is discussed further in the final chapter.

3.10.7 DAG–data consistency

It is not possible to ascertain causality from observational data unequivocally, however, it is useful to assume potential causality and use graphical model theory to evaluate data where possible. Causal models can then be updated according to what is found and this process of trying to falsify the causal claims can then continue to be repeated following the scientific method. This can be done by testing DAG–data consistency, that is, whether the assumed causal diagram could have produced the DAG hypothesised to have generated a specific dataset.[30]

3.11 Assessing simulation results

Recently, work has been published which proposes methods for assessing the results of simulations.[51] This work builds on that of others (e.g.[66]) and provides a comprehensive guide to follow when conducting simulation studies to assess statistical methods, it is therefore used to measure the performance of the simulations used in this thesis. The performance measures used are: coverage, the probability that the confidence intervals of the coefficient for each iteration of the simulation contain zero; measurement of the inferential bias present, whether the estimated coefficient averages the true value, zero in this case; and Monte Carlo standard error, the simulation uncertainty, i.e. an estimate of the standard error of the performance as a result of using a finite number of simulation iterations.

Given all of the considerations above, the final section of this chapter will outline a step–by–step protocol for setting up a simulation study with these topics incorporated.

3.12 A step–by–step walk–through of the simulation set–up

This thesis explicitly details each simulation undertaken using the following steps. These steps are drawn specifically from other step–by–step procedures reported for: conducting simulation studies in `R`,[54] integrating causal modelling and statistical estimations[9] and using simulation studies to evaluate statistical methods.[51]

1. Specify the effect estimates to be estimated and compared (informed by literature search of common methods used) and what parameter estimates will be retained.

2. Specify the assumptions made about the nature and parameters of the dataset. When combining this with causal inference methods this would include drawing

a causal diagram (such as a DAG, as is the case in this thesis) of the assumed causal relationships between variables.

3. Specify assumptions made about the variable distributions and covariance structure of these variables including how simple or complex the model will be and whether it should be based on real–world data.

4. Specify the factors that will be varied in the simulation.

5. List the performance measures to be estimated and choose the number of simulations necessary to achieve acceptable Monte Carlo standard errors for the key performance measures.

6. Set the seed for the random number generator so that equivalent results can be replicated by others.

7. Generate a dataset according to these assumptions.

8. Perform statistical analyses on this dataset and retain the parameter estimates obtained.

9. Steps 7 and 8 are repeated many times with newly generated datasets in order to obtain an empirical distribution of parameter estimates.

10. In some instances, Steps 1–4 are repeated according to a new causal diagram, new parameters and/or new assumptions.

11. The empirical distributions of the parameter estimates from the simulated datasets are analysed to evaluate the question of interest.

12. Compute the performance estimates.

The benefit of following a step–by–step guide for a simulation is that researchers cannot simply report the most or least favourable results (depending on what they wish to

show) from using different configurations of the simulation parameters or from judicious selection of random number seeds.[51] A step–by–step guide to the simulations acts as a protocol and holds the researcher accountable to their simulation choices and gives them opportunity to justify those choices.

This chapter has introduced methods that researchers can use to conduct simulation studies informed by causal knowledge. Causal inference and simulation are naturally linked via the data generation process and this chapter has added to existing step–by–step simulation guides[51] by suggesting steps to incorporate causal considerations. It has also introduced some specific long–acknowledged issues encountered within the field of health geography such as the modifiable areal[84,85] and temporal[86] unit problems (MAUP and MTUP, respectively) and what has been termed 'the most dangerous equation'.[87] A particularly novel topic considered in this chapter is the introduction of how one may think about composite and compositional data, which are particularly common in quantitative health geography, within a causal inference framework.

This thesis now goes on to use these methods of simulation and causal inference to investigate some long–standing issues in health geography. These problems have not been investigated using these methods previously. Chapter 4 illustrates the issue of mathematical coupling using causal inference methods building simulations based on these causal assumptions to investigate this problem further. Chapter 5 investigates studies into deprivation and limiting long–term illness and encounters simulation of composite variables and compositional data. Chapter 6 uses simulated data to investigate the historical methods used to study the 'population mixing hypothesis' and encounters selection on the outcome amongst issues relating to the 'most dangerous equation' and MAUP. Along with drawing conclusions about these problems specifically and how they have previously been investigated, the following work serves as a demonstration of how simulation and causal inference can be used to evaluate methods used in epidemiology and health geography.

Chapter 4

Mathematical Coupling and Causal Inference

4.1 Introduction

This chapter introduces mathematical coupling and uses the methods introduced in Chapter 3 combining causal inference via graphical causal models and simulation to investigate this problem. The historical solution to avoiding the form of bias introduced by mathematical coupling is critiqued from a causal inference perspective and expanded into the area of population health/health geography.

4.1.1 What is Mathematical Coupling?

Mathematical coupling is a form of composite variable bias that occurs when two or more variables are analysed by correlation or regression while sharing an algebraic dependency.[89] One instance of mathematical coupling occurs when analysing proportions (i.e. composite variables formed by dividing one variable by another) where two proportions share a common denominator (e.g. $\frac{X}{Z}$ and $\frac{Y}{Z}$).

4.1.2 The history of Mathematical Coupling

Although mathematical coupling was first recognised by Pearson in 1896[90] and has been repeatedly discussed in the literature,[91] it has remained largely overlooked in observational research. For instance, Pearson demonstrated that if three variables (e.g. X, Y, Z) are random, have similar coefficients of variation, and are otherwise unrelated (i.e. they are mutually uncorrelated), then the 'spurious' correlation between any two of these variables, when commonly divided by the third (e.g. $\frac{X}{Z}$ and $\frac{Y}{Z}$), would average $r = 0.5$.[90]

Equation 4.1.1 shows that if the correlation between the raw variables were all equal to zero, the correlation of the ratios with a common denominator would not be equal to zero, i.e. the 'spurious' correlation generated when dividing by the common denominator when all the variables are uncorrelated is 0.5.[92]

$$r_{(Y/Z)(X/Z)} = (1 - r_{YZ} - r_{XZ} + r_{XY})/[2(1 - r_{YZ})^{1/2}(1 - r_{XZ})^{1/2}] \tag{4.1.1}$$

Put simply, this is because any change in the common denominator (Z) affects both proportions simultaneously: increases in Z reduce both $\frac{X}{Z}$ and $\frac{Y}{Z}$ simultaneously, while decreases in Z increase both $\frac{X}{Z}$ and $\frac{Y}{Z}$ simultaneously. The null–hypothesis (of no relationship between X and Y) therefore suggests that $\frac{X}{Z}$ and $\frac{Y}{Z}$ should be positively correlated, the extent of which depends on the variance and covariance structure of the three variables.[93]

As a solution, Pearson[90] proposed calculating the partial correlation between numerators whilst conditioning on the common denominator by including it as a separate covariate in a linear regression model. Later, Neyman[94] repeated Pearson's warnings and agreed that the ratios should be separated before analysis. Neyman advocated the use of analysis of covariance (ANCOVA) to achieve this. The approach of separating the variables is often used when modelling rare outcomes using log–linear (Poisson) regression, with

the denominator included as a logged–covariate 'offset'.[95] This circumnavigates the tautological bias that would otherwise have been introduced, though it appears that this was not by intentional design. It is fortunate that Poisson modelling, as applied in health research, happens to follow this separation approach.

In geographical and population health research, proportions are often preferred to raw count variables because they offer a 'relative' measure of each characteristic within local populations of different sizes between areas. Dividing the raw count by the population size is intended to 'control' for the 'dominating influence' of the varying population denominator[96] and allows researchers to compare proportions (e.g. prevalence of an exposure or incidence of a disease) across geographically–defined populations. 'Standardising' for population size (N) in this way transforms the relationship into an algebraic dependency, but little attention has been paid to investigating this in the literature and whether it risks introducing biased or 'spurious' relationships when analysed by correlation or regression. A key question is therefore: does the common denominator distort the null hypothesis and any estimated correlation or regression coefficients obtained, making it difficult to draw robust inferences (i.e. inferences unbiased by mathematical coupling; defined as 'robust' throughout) in the same way as that discussed by Pearson, Neyman and Fisher?[90,94,97]

Contemporary causal inference methods are used to explore the original problem as described by Pearson[90] and demonstrate how these methods can be utilised to explicate when the historical solution is appropriate. It is shown that no obvious solution exists where the common denominator does not cause both numerators but is instead a consequence of one or more of them. This context was overlooked by Pearson, Neyman and Fisher, whose work preceded contemporary causal inference methods. The data structure of the three variables is investigated to determine its influence on the level of 'spurious' correlation present.

The difference between these examples and the common health geography scenario of

analysing area–level data where the common denominator (e.g. population size) causes the number of units exposed and/or the number of events that arise is discussed and the problem is simulated under the null hypothesis.

4.2 Methods

4.2.1 Exploring proportions using causal graphs

The issue of mathematical coupling is explored using a directed acyclic graph (DAG; introduced in Section 2.1.7), which allows researchers to depict causal relationships they believe to operate between two or more variables.

Firstly, the example introduced by Pearson[90] is illustrated in a DAG and this example is expanded upon using simulations. This example is then adapted to the case where there are associations between the three variables (two numerators and the common denominator), again, this is further investigated using simulations. This principle is then generalised in terms of an arbitrary population–level exposure measure and a population–level health outcome.

Simulated data allows the confusing influence of unobserved confounding to be excluded, which would be present in observed data. The primary assumption in these population health examples is that Z represents the population count in each area–level unit of analysis (e.g. ward, Clinical Commissioning Group, region, country). Initially, Z is considered as the 'driver' of both numerator variables, i.e. Z causes both X and Y, since this is the most ubiquitous generic illustration of the common denominator problem within bio–medicine and health geography. This three–variable configuration may be quite common in other circumstances: wherever two or more features of clustered observational measures are examined for putative causal relationships and these (numerator) features are 'standardised' relative to their area–level sizes (common denominator), which vary

(e.g. across companies in commerce, across markets in finance, across urban landscapes in planning, across networks in communications or transport, etc.).

The circumstances under which mathematical coupling is introduced to the analysis of the proportions are explored. In all instances, the objective is to estimate the total causal effect of the exposure on the outcome while accommodating the variation in area–level sizes.

These steps are outlined in Section 4.2.4 using the step–by–step guide to simulation developed in Chapter 3 (Section 3.12). All simulations are written in R[98] and the code used in this chapter is available in Appendix B.

4.2.2 Historical examples

The original example given by Pearson[90] of 'spurious' correlation between variables with a common denominator was a biological one. In this example, Pearson described a situation in which 1,000 skeletons were randomly rearranged to make up new skeletons. If a researcher was to try to ascertain whether they were in their originating skeletons by correlating the lengths of the bones, e.g. $\frac{\text{femur}}{\text{humerus}}$ and $\frac{\text{tibia}}{\text{humerus}}$, they would report a correlation of around 0.45 when it should have been zero (on average) due to the bones being randomly reassigned. Pearson attributes this correlation to the arithmetic used and suggests that any correlation above this value (0.45) is 'organic' (i.e. true) correlation between the variables and that the 0.45 correlation is 'spurious'. As this is a biological example, the use of the normal distribution for the variables would seem acceptable. The next historical example is a geographical one.

To illustrate the problem of mathematical coupling, Neyman[94] asked 'do storks bring babies?' and to answer this question he used hypothetical independent data and correlated $\frac{\text{number of storks}}{\text{number of women}}$ and $\frac{\text{number of babies}}{\text{number of women}}$ over 54 counties. Neyman showed that, although the original data on the population of storks and number of babies were independent, the

correlation between the ratios (these variables divided by the number of women) were statistically significant.

This example is simulated in this chapter as an intermediary example between the biological example of Pearson (where the normal distribution can be assumed) and other examples where the data are compositional, that is, the variables are constrained by a total, e.g. if only singleton births are considered, the number of babies is constrained by the number of women but the number of storks is not constrained in the same way. However, simulating the number of babies and the number of women as following a normal distribution with no such constraints may not be an issue when the number of births is low because the situation where births exceeds the number of women may not occur. This constraint would have to be taken into consideration when other research questions are investigated.

4.2.3 Pearson's historical example

Using a DAG, the generic example discussed by Pearson, in which there are three variables (X, Y and Z) which are completely independent of each other (Figure 4.2.1), is depicted.[90]

To simulate this situation, a DAG with no causal relationships (i.e. no causal arrows) is drawn. This means that the correlation matrix between the variables approximates the identity matrix:

$$\begin{bmatrix} & X & Y & Z \\ X & 1.00 & 0.00 & 0.00 \\ Y & 0.00 & 1.00 & 0.00 \\ Z & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

10,000 data points for X, Y and Z were simulated from a standard normal distribution and 5 was added to each value to ensure that there were no negative values but the covariance

matrix was unchanged.

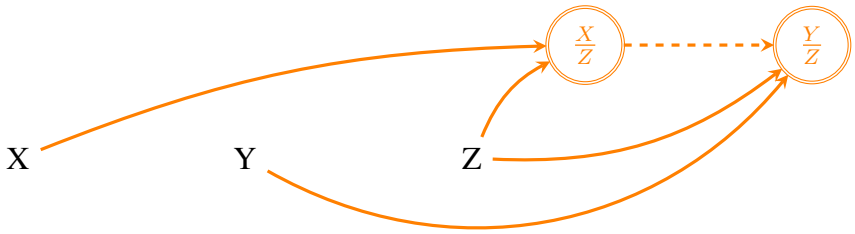

Figure 4.2.1: **Directed Acyclic Graph depicting the example of Pearson;[90] three completely independent variables.** Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

The simple correlation between X and Y was 0.00, the correlation between $\frac{X}{Z}$ and $\frac{Y}{Z}$ was 0.56 and the association between X and Y conditional on Z was 0.00. All of these results are consistent with the results of Pearson,[90] Neyman[94] and Fisher.[97]

Next, five possible causal scenarios are considered for this hypothetical context in separate DAGs. When depicting these hypothetical causal relationships in a DAG, for illustration only, constructed (wholly deterministic) proportions and their component raw variables (X, Y, Z) are included, simultaneously. Separately, the causal relationships between the raw variables alone are illustrated, and from this an appropriate minimally sufficient adjustment set (MSAS) is identified (which can be achieved using the rules of d–separation via DAGitty[30,74] at www.dagitty.net, or the package `dagitty` in the `R` statistical software;[98] see Section 2.1.9).

For each scenario, the following effects are calculated and their implications are considered: 1) analysing the simple effect of X on Y, 2) analysing the effect of $\frac{X}{Z}$ on $\frac{Y}{Z}$, and 3) analysing the effect of X on Y conditional on Z, as recommended by

Pearson, Neyman, and Fisher.[90,94,97] The results are presented as standardised regression coefficients, as these are easily comparable.

These examples differ from those outlined by Pearson, Neyman and Fisher because of the associations that are simulated between the three variables (via path coefficients), where their originals example had none. Like the historical examples, the examples introduced here begin with the same coefficients of variation ($\frac{\sigma}{\mu}$), however, this simplicity is expanded considerably by differing these along with the path coefficients defining each bivariate relationship in the DAG. The data simulation process is reported below along with the results. The details of the simulations are outlined in the following step–by–guide.

4.2.4 Step–by–step guide to the simulations

1. Correlation, partial correlation and linear regression modelling are used to analyse the simulated data, this means that the correlation coefficients, partial correlation coefficients and the regression coefficient are retained, respectively.

2. The simulations used in this chapter are built–up from the example of Pearson where three completely independent variables were analysed. DAGs are drawn for each of the possible scenarios (Section 2.1.7). The path coefficients between the simulated variables are then varied and the coefficient of variation of the three variables is changed. The path coefficients and coefficients of variation are then changed simultaneously.

3. The variables are first assumed to be from Normal distributions with estimated mean and standard deviations informed by researcher knowledge and internet searches. The covariance matrix for each example, calculated from the path coefficients using path tracing rules (introduced in Section 3.9.1), are recorded and represent the bivariate relationships that are approximated in each simulated dataset. All possible variations are then investigated, followed by the analysis

of geographical data where a population are simulated from a negative binomial distribution and from these two other variables are generated from this population using a binomial distribution; equivalent to flipping a coin for every member of the population and recording how many tails occurred.

4. The simulations in this chapter are used to illustrate the concept of mathematical coupling and whether the historical solutions to the problem are appropriate, for this reason, performance measures are not taken for the simulations.

5. Set the seed for the random number generator so that exact results can be replicated by other researchers.

6. Generate a dataset according to these assumptions.

7. Perform statistical analyses on this dataset and retain the parameter estimates obtained.

8. Compare these parameter estimates with the simulated to be true values.

4.2.5 Five causal scenarios

Scenario 1) Z causes X and Y, but X does not cause Y, nor does Y cause X (Figure 4.2.2); this is the null scenario for any $X - Y$ association;

> For example, the population size (Z) of a geographical area causes both the number of cats resident in that area (X) and the number of apples consumed per month (Y) in that area, but the number of cats (X) does not cause the number of apples consumed (Y) within any area, or vice versa. Note: It is difficult to find two variables that are not causally related in at least some way as, however unlikely, there is usually some physically plausible causal relationship.[24] In this case,

however, an example was chosen where any causal relationship between X and Y is deemed to be extremely unlikely.

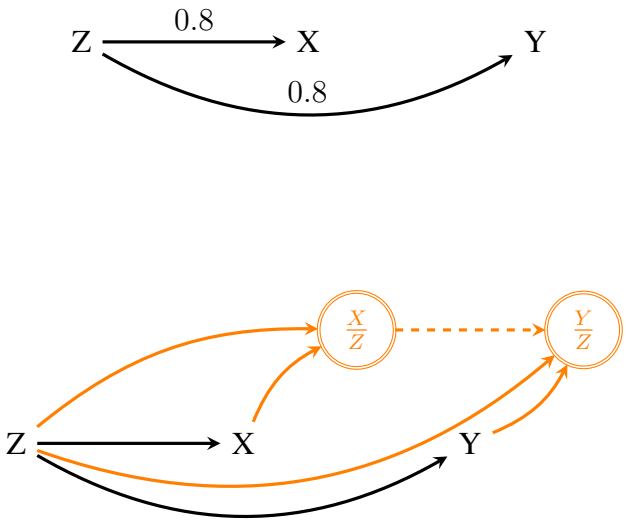

Figure 4.2.2: **Scenario 1: Null association: no causal relationship between X and Y.** Directed Acyclic Graph (DAG) above with path coefficients on edges and corresponding graph including functionally deterministic ratio variables below. Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

In Scenario 1, X (e.g. number of cats) has no causal effect on Y (e.g. number of apples consumed in a month), but both X and Y are caused by Z (e.g. population size). The correlation matrix is approximately:

$$\begin{bmatrix} & X & Y & Z \\ X & 1.00 & 0.64 & 0.80 \\ Y & 0.64 & 1.00 & 0.80 \\ Z & 0.80 & 0.80 & 1.00 \end{bmatrix}$$

The causal effect of X on Y is 0.00; there is no direct path between X and Y. The back–door path between X and Y is 0.64.

The variables are drawn from the following distributions:

Population: $Z \sim Normal(mean = 1000, sd = 1000/5)$

Population of cats: $X \sim Normal(mean = 200, sd = 200/5)$

Number of apples consumed in a month: $Y \sim Normal(mean = 12000, sd = 12000/5)$

According to causal graph theory, a simple estimate of the association between X on Y would therefore be biased by confounding from Z (beta $= 0.64$). Similarly, the association between $\frac{X}{Z}$ and $\frac{Y}{Z}$ would be biased by mathematical coupling from the shared denominator Z (beta $= -0.71$). The association between X and Y conditional on Z however would close the 'spurious' path $X \leftarrow Z \rightarrow Y$ to produce a robust estimate of the null effect of X on Y (beta $= -0.004$). Results from each Scenario are summarised in Table 4.1.

Scenario 2) where Z causes X and Y, and X is a cause of Y (Figure 4.2.3, consistent with what is hypothesised as the 'exposure' causing the 'outcome', 'confounded' by Z);

> For example, the population size (Z) of a geographical area causes both the minutes exercised weekly (X) and the number of anti–depressant prescriptions (Y), within an area. Both the population size (Z) and number of weekly minutes exercised (X) are causes of the number of anti–depressant prescriptions (Y).

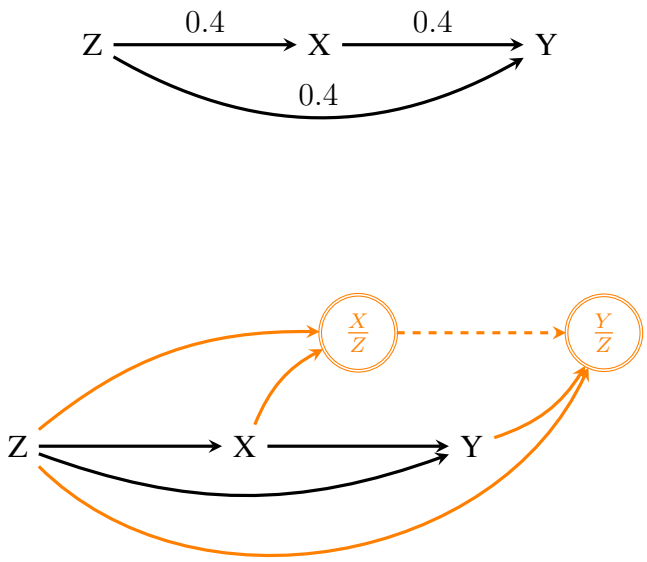

Figure 4.2.3: **Scenario 2: X causes Y, Z is a confounder.** Directed Acyclic Graph (DAG) with path coefficients on the edges above and corresponding graph including functionally deterministic ratio variables below. Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

In Scenario 2, X (e.g. weekly minutes of exercise) causes Y (e.g. number of anti-depressant prescriptions), and both X and Y are caused by Z (e.g. population size). The correlation matrix is approximately:

$$\begin{bmatrix} & X & Y & Z \\ X & 1.00 & 0.56 & 0.40 \\ Y & 0.56 & 1.00 & 0.56 \\ Z & 0.40 & 0.56 & 1.00 \end{bmatrix}$$

The *direct* causal effect of X on Y is 0.40; there is a direct path between X and Y

and the bivariate correlation between these is 0.40. The *total* causal effect of X on Y (according to the correlation matrix) is $0.4 + 0.4 * 0.4 = 0.56$.

The variables are drawn from the following distributions:

Population: $Z \sim Normal(mean = 1000, sd = 1000/5)$

Minutes of exercise: $X \sim Normal(mean = 150000, sd = 150000/5)$

Number of anti-depressant prescriptions: $Y \sim Normal(mean = 80, sd = 80/5)$

According to causal graph theory, a simple estimate of the association between X on Y would therefore be biased by confounding from Z (beta $= 0.57$). Similarly, the association between $\frac{X}{Z}$ and $\frac{Y}{Z}$ would be biased by mathematical coupling from the shared denominator Z (beta $= 0.997$). The association between X and Y conditional on Z however would close the 'spurious' path ($X \leftarrow Z \rightarrow Y$) to produce a robust estimate of the total causal effect of X on Y (beta $= 0.41$, sufficiently close to the true value 0.40).

Scenario 3) it is no longer assumed that the common denominator Z causes both X and Y, but instead it allows for the possibility that the exposure X causes Z, and both X and Z continue to cause Y (Figure 4.2.4);

> For example, within a geographical area, the number of job opportunities (X) causes the population size (Z). Both the number of job opportunities (X) and the population size (Z) cause healthcare expenditure (Y).

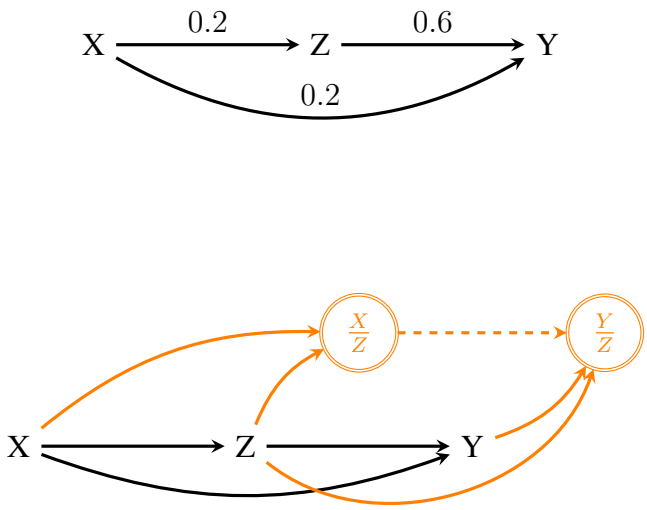

Figure 4.2.4: **Scenario 3: Z is a mediator on the causal path between X and Y.** Directed Acyclic Graph (DAG) with path coefficients above and corresponding graph including functionally deterministic ratio variables below. Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

In Scenario 3, X (e.g. job opportunities) causes Y (e.g. healthcare expenditure) partly through mediator Z (e.g. population size). The correlation matrix is approximately:

$$\begin{bmatrix} & X & Y & Z \\ X & 1.00 & 0.32 & 0.20 \\ Y & 0.32 & 1.00 & 0.64 \\ Z & 0.20 & 0.64 & 1.00 \end{bmatrix}$$

The causal effect of X on Y is 0.32; via the direct path between X and Y (0.20) and the path via Z $(0.2 * 0.6 = 0.12)$. The *direct* causal effect of X on Y is 0.20.

The variables are drawn from the following distributions:

Population: $Z \sim Normal(mean = 1000, sd = 1000/5)$

Job Opportunities: $X \sim Normal(mean = 25, sd = 25/5)$

Healthcare Expenditure: $Y \sim Normal(mean = 3000000, 3000000/5)$

According to causal graph theory, a simple estimate of the association between X on Y would therefore produce a robust estimate since there is no confounding between X and Y (beta $= 0.32$). The association between $\frac{X}{Z}$ and $\frac{Y}{Z}$ would again be biased by mathematical coupling from the shared denominator Z (beta $= 0.999$). The association between X and Y conditional on Z would also produce a biased estimate of the total causal effect of X on Y, as it would inappropriately close the causal path ($X \rightarrow Z \rightarrow Y$) (beta $= 0.20$) and risk introducing further problems from the reversal paradox.

Scenario 4) X and Y cause Z, Z is now a collider. Temporally, the exposure, X, occurs before the outcome of interest, Y, represented in the causal diagram with Y illustrated to the right of X, i.e. time runs from left to right (Figure 4.2.5);

> For example, within a geographical area, the number of inward-migrants (X) and the number of births (Y) cause the population size (Z). It should be noted here that the assumption is made that the dominant direction of effect is from migration and births towards population size, however, the reverse may also be true.

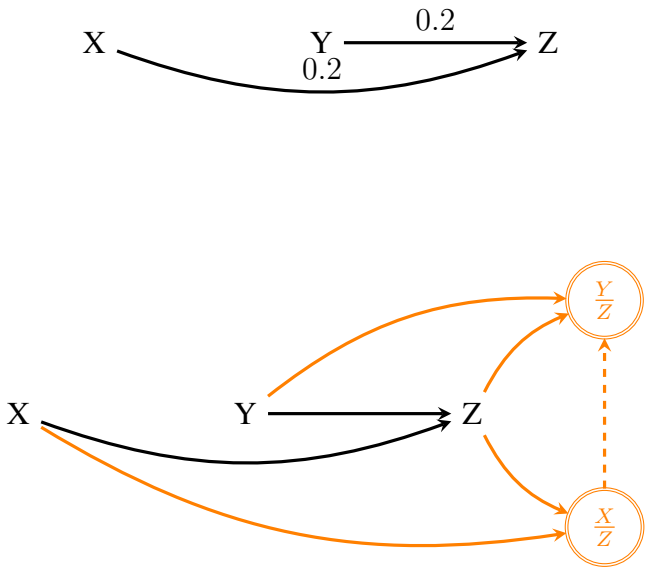

Figure 4.2.5: **Scenario 4: X and Y cause Z, a collider.** Directed Acyclic Graph (DAG) with path coefficients on edges above and corresponding graph including functionally deterministic ratio variables below. Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

In Scenario 4, X (e.g. migration) and Y (e.g. births) cause Z (e.g. population size); Z is a collider. The correlation matrix is approximately:

$$\begin{bmatrix} & X & Y & Z \\ X & 1.00 & 0.00 & 0.20 \\ Y & 0.00 & 1.00 & 0.20 \\ Z & 0.20 & 0.20 & 1.00 \end{bmatrix}$$

The causal effect of X on Y is 0.00; there is no direct path between X and Y.

The variables are drawn from the following distributions:

Population: $Z \sim Normal(mean = 1000, sd = 1000/5)$

Migration: $X \sim Normal(mean = 5, sd = 5/5)$

Births: $Y \sim Normal(mean = 12, sd = 12/5)$

According to causal graph theory, a simple estimate of the association between X on Y would therefore produce a robust estimate since there is no confounding between X and Y (beta $= -0.009$). The association between $\frac{X}{Z}$ and $\frac{Y}{Z}$ would again be biased by mathematical coupling from the shared denominator Z (beta $= -0.49$). The association between X and Y conditional on Z would also produce a biased estimate of the total causal effect of X on Y, as it would introduce an association between X and Y due to conditioning on a collider (beta $= -0.05$).

Scenario 5) an extension of Scenario 4 where X also causes Y (Figure 4.2.6);

> For example, within a geographical area, the number of new housing units (X) causes the number of immigrants (Y) and both of these cause the population size (Z).

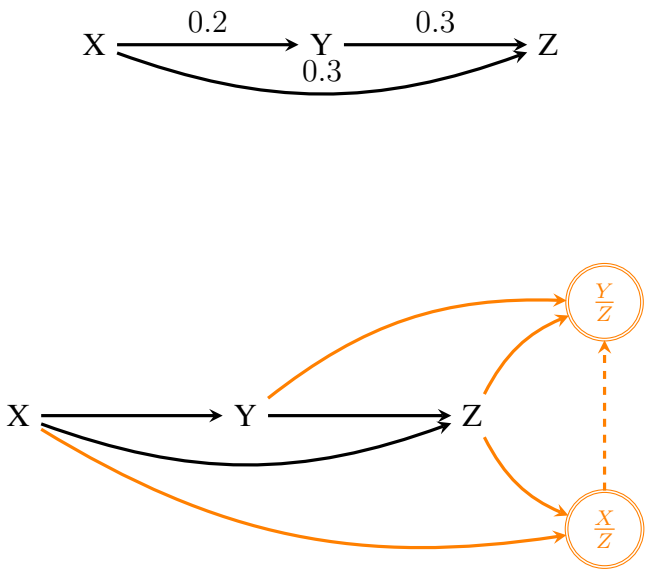

Figure 4.2.6: **Scenario 5: X causes Y and X and Y cause Z, a collider.** Directed Acyclic Graph (DAG) with path coefficients on edges above and corresponding graph including functionally deterministic ratio variables below. Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

In Scenario 5, X (e.g. count of new housing units) causes Y (e.g. number of immigrants) and both X and Y cause Z (e.g. population size); Z is a collider. The correlation matrix is approximately:

$$\begin{bmatrix} & X & Y & Z \\ X & 1.00 & 0.20 & 0.36 \\ Y & 0.20 & 1.00 & 0.36 \\ Z & 0.26 & 0.36 & 1.00 \end{bmatrix}$$

The causal effect of X on Y is 0.20; there is a direct path between X and Y and the bivariate correlation between these (according to the correlation matrix) is 0.20. Following the path tracing rules there are no other paths that contribute to the correlation between

these variables as one cannot travel forward along an arrow and then backward along another.

The variables are drawn from the following distributions:

Population: $Z \sim Normal(mean = 1000, sd = 1000/5)$

New housing: $X \sim Normal(mean = 5, sd = 5/5)$

Immigration: $Y \sim Normal(mean = 5, sd = 5/5)$

According to causal graph theory, a simple estimate of the association between X on Y would therefore produce a robust estimate since there is no confounding between X and Y (beta $= 0.20$). The association between $\frac{X}{Z}$ and $\frac{Y}{Z}$ is also unbiased by mathematical coupling from the shared denominator Z) (beta $= 0.20$) because the coefficients of the simulated variables have the same coefficients of variation, this is investigated later in this chapter. The association between X and Y conditional on Z would also produce a biased estimate of the total causal effect of X on Y, as it would introduce an association between X and Y due to conditioning on a collider (beta $= 0.08$).

Each of these simulations are generated in a way that means that the coefficient of variation for the three variables in each Scenario are the same. Simulations are now conducted to investigate the effect of changing these values on the amount of 'spurious' correlation present.

4.2.6 The importance of the simulated causal relationships between the three variables

Next, the $Z \rightarrow X$ and $Z \rightarrow Y$ causal relationships are now varied to understand the importance of these relationships. These investigations are informed by a causal diagram

Table 4.1: Summary results from analysing the simple effect of X on Y, analysing the effect of $\frac{X}{Z}$ on $\frac{Y}{Z}$, and analysing the effect of X on Y conditional on Z, as recommended by Pearson, Neyman, and Fisher.[90,94,97] The simulated to be true causal effect of X on Y is also shown.

	Estimated Coefficients				
	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
True causal effect	0.00	0.40	0.20	0.00	0.20
Correlation of X and Y	0.64	0.57	0.32	-0.01	0.20
Correlation of $\frac{X}{Z}$ and $\frac{Y}{Z}$	-0.71	1.00	1.00	-0.49	0.20
Partial Correlation of Y and X (controlling for Z)	0.00	0.41	0.20	-0.05	0.08

which has path coefficients (b_1 and b_2) assigned to the causal arcs (Figure 4.2.7). Firstly, the case where $b_1 = b_2$ for the relationships $Z \to X$ and $Z \to Y$ is considered. In this case, no causal relationship between X and Y is simulated and the 'spurious' correlation between $\frac{X}{Z}$ and $\frac{Y}{Z}$ against b is plotted.

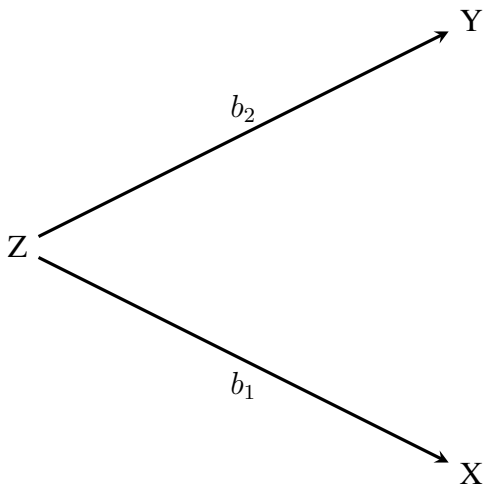

Figure 4.2.7: Causal Diagram in which Z is a confounder of the exposure–outcome relationship between X and Y; b_1 and b_2 represent the path coefficient assigned to the arcs for simulation.

Figure 4.2.8 shows the effect of increasing the true path coefficients ($b_1 = b_2$) between Z and X and Y. As b_1 and b_2 simultaneously increase to 1, the 'spurious' correlation

between $\frac{X}{Z}$ and $\frac{Y}{Z}$ reduces to 0. The 'spurious' correlation is highest (0.5) when the three variables are completely independent of each other. This means that if there is no $Z \rightarrow X$ and $Z \rightarrow Y$ causal relationship at all it would appear as if there was one. In contrast, if there is a strong $Z \rightarrow X$ and equally strong $Z \rightarrow Y$ causal relationship then the spurious effects of mathematical coupling are substantially diminished. In practice, there is no way of knowing to what extent mathematical coupling is affecting the correlation reported.

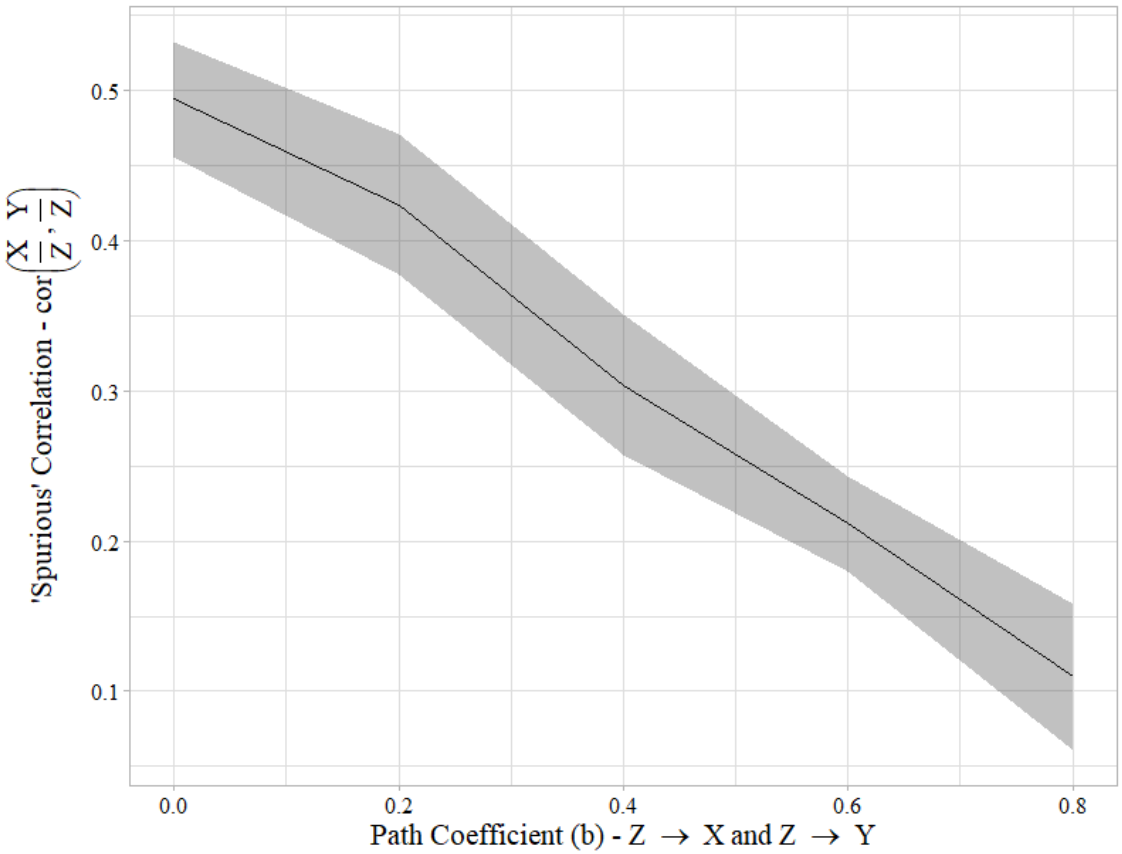

Figure 4.2.8: Plot showing the 'spurious' correlation from the analysis of $\frac{X}{Z}$ and $\frac{Y}{Z}$ against the true path coefficients simulated between $Z \rightarrow X$ and $Z \rightarrow Y$. The shaded area represents the 95% confidence interval of the 'spurious' correlation over the simulations.

4.2.7 The importance of the coefficient of variation

Now, the importance of the coefficients of variation of the three variables on the level of 'spurious' correlation generated from the common denominator is explored. Particular interest lies in the case where Z is a confounder (Scenarios 1 and 2, above) as it is proposed that this is the most likely situation in which a researcher would divide through by a common denominator (when researchers divide through by a common denominator they are aiming to remove confounding). The $Z \rightarrow X$ and $Z \rightarrow Y$ causal relationships are varied, then the coefficients of variation are varied and then both are varied simultaneously.

Figure 4.2.9 shows the effect of changing the ratio of the coefficients of variation of X, Y and Z when the true causal effect between the three variables is zero. When X, Y and Z vary by a similar amount, the amount of 'spurious' correlation is about 0.5. On the right hand side of the graph Z varies much less than X and Y, and dividing through by Z does not cause much mathematical coupling. On the left hand side of the graph, when Z varies much more than X and Y the amount of 'spurious' correlation introduced by dividing through by Z is very high (left hand side of the graph). This may occur when Z is not measured accurately or measures are taken to protect the identity of individuals once geographical location is included in a dataset. It could also occur when datasets are amalgamated, for example, if a dataset is combined with population data taken from the census there may be a time gap between the points at which the data were collected elevating the variation in Z.

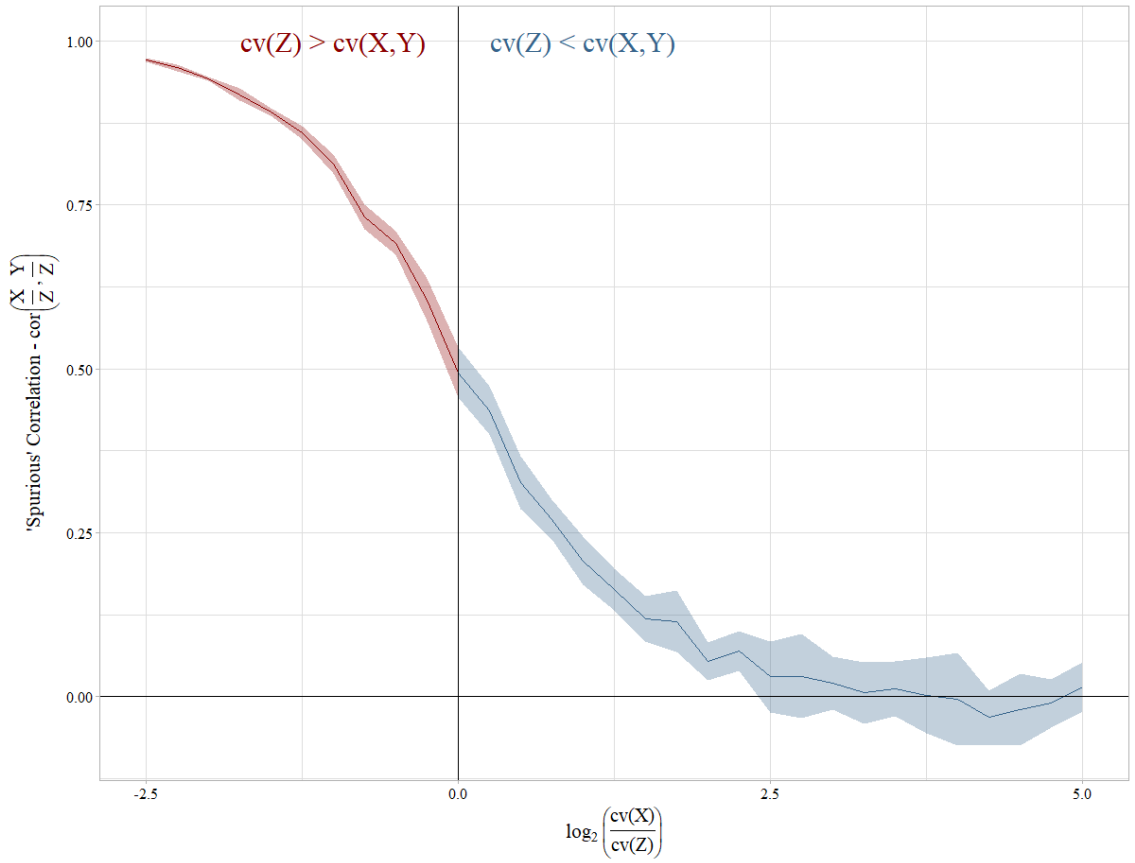

Figure 4.2.9: Plot showing the 'spurious' correlation from the analysis of $\frac{X}{Z}$ and $\frac{Y}{Z}$ against the log (base 2) ratio of the coefficients of variation (cv) of X and Z ($cv(X) = cv(Y)$), true associations simulated between $Z \rightarrow X$ and $Z \rightarrow Y$ are zero.

4.2.8 The effect of varying the simulated causal effect and the coefficients of variation simultaneously

Figure 4.2.10 shows the effect of changing both the true path coefficients ($b_1 = b_2$) along with the coefficient of variation. The darkest line is equivalent to that shown in Figure 4.2.9 when the true causal relationships between X, Y and Z are zero. As b_1 and b_2 simultaneously increase and the coefficient of variation of Z becomes less than the coefficient of variation of X and Y, the 'spurious' correlation decreases to zero before

increasing to b^2, i.e. the value of the backdoor path $X \rightarrow Z \rightarrow Y$.

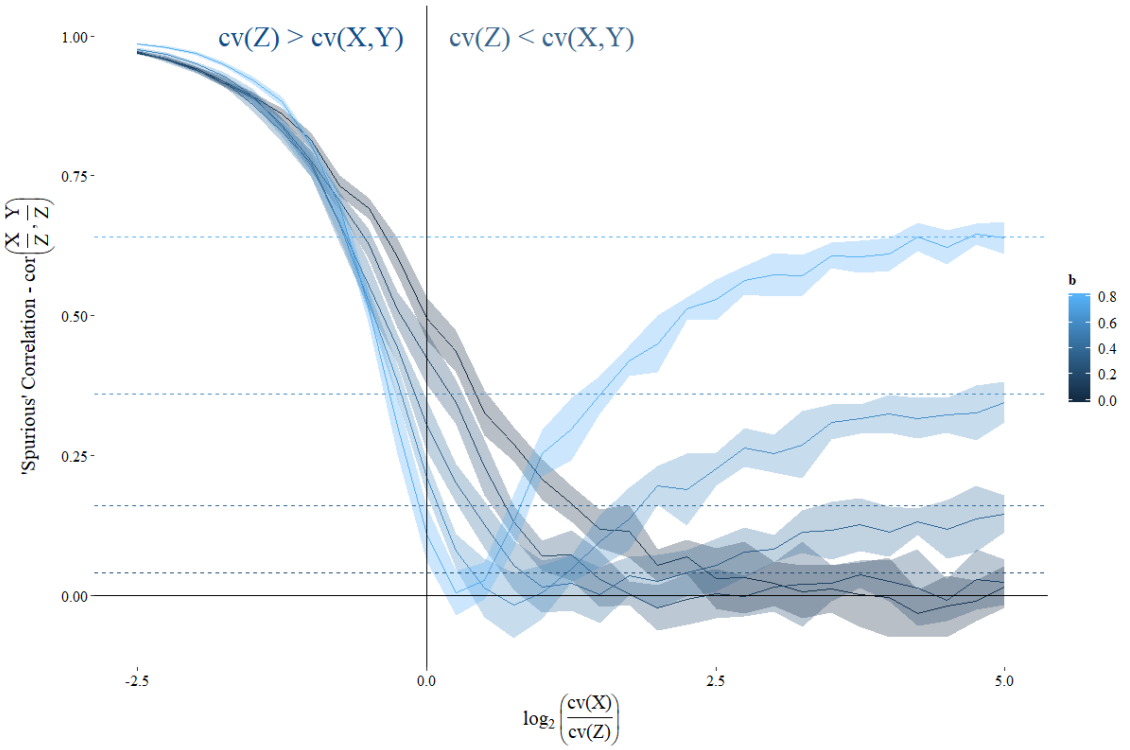

Figure 4.2.10: Plot showing the 'spurious' correlation from the analysis of $\frac{X}{Z}$ and $\frac{Y}{Z}$ against the log (base 2) ratio of the coefficients of variation (cv) of X and Z ($cv(X) = cv(Y)$ for different values of path coefficients (b; true associations simulated between $Z \rightarrow X$ and $Z \rightarrow Y$).

4.2.9 The effect of varying the simulated causal effect, the coefficients of variation and differing the path coefficients between the three variables

Figure 4.2.11 expands Figure 4.2.10 by varying the path coefficients (b_1 and b_2) between $Z \rightarrow X$ and $Z \rightarrow Y$ which, until now, have been equal. This shows that the difference between how much X and Y vary in comparison to Z are both important determinants of how much 'spurious' correlation is present when dividing through by Z.

When the coefficients of variation between X and Z and Y and Z are similar, the effect of mathematical coupling on correlating $\frac{X}{Z}$ with $\frac{Y}{Z}$ is small. This is apparent in Figure 4.2.11 where the 'spurious' correlation is zero when $log_2\frac{cv(Y)}{cv(N)} = log_2\frac{cv(X)}{cv(N)} = 0$, i.e. a cross appears in the plots as the path coefficient (b) between N and Y and N and X approaches 1. However, if the coefficient of variation of one, or both, of X and Z and Y and Z differ then the greater the difference in these coefficients of variation, the greater the effect of mathematical coupling from the common denominator. This is apparent in the plots where there is a stronger 'quadrant' effect as the path coefficient (b) between N and Y and N and X approaches 1.

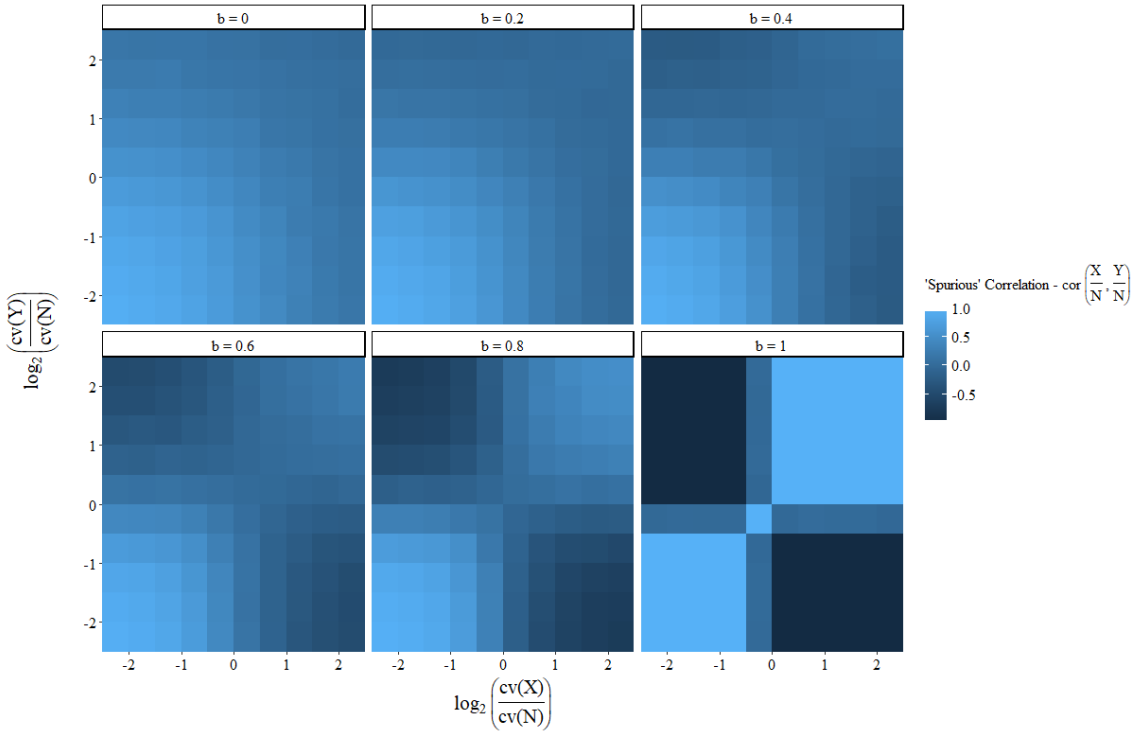

Figure 4.2.11: Plot showing the 'spurious' correlation from the analysis of $\frac{X}{Z}$ and $\frac{Y}{Z}$ when varying the differences between the coefficient of variation of both X and Y with the coefficient of variation of Z for different values of simulated true path coefficients, b.

4.2.10 Neyman's Historical Example

Neyman's historical example[94] asks whether storks bring babies. Neyman used hypothetical data from 54 counties to investigate this problem, these data are available in the literature[99] and in the `R` package '`TeachingDemos`'.[100] The variables in this dataset are completely independent and this is first repeated here by simulating three independent variables (X, Y, Z) from normal distributions as in the Pearson example[90] but with means and standard deviations similar to those used in the hypothetical data generated by Neyman.

In geographical examples such as this the number of women is a constraint on the number of babies that can be born and it perhaps makes more sense that the number of women is a cause of the number of babies (as in Figure 4.2.12). In order to simulate this scenario (under the null hypothesis that storks do not bring babies) the number of women (Z) is simulated from a negative binomial distribution for 1,000 areas and from this the number of births (X) is simulated using the binomial distribution with 'success' probability 0.1. The number of storks (Y) is simulated from a negative binomial distribution independent of the distributions used to simulate X and Z. The correlation between $\frac{X}{Z}$ and $\frac{Y}{Z}$ is then calculated and the linear regression model $Y \sim X + Z$ is run to check for 'spurious' relationships in the results.

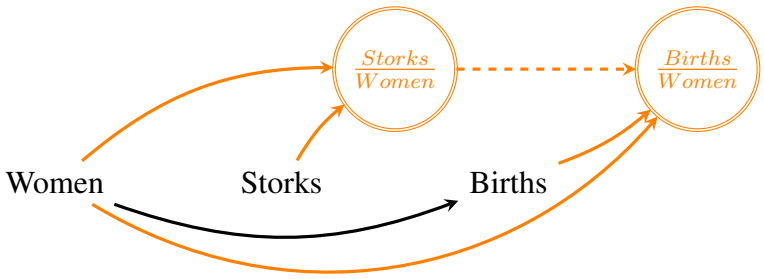

Figure 4.2.12: **Directed Acyclic Graph depicting the assumed relationships in the example of Neyman;[94] 'do storks bring babies?'** Black arrows indicate causal non–parametric relationships, while orange arrows indicate algebraic relationships (constructed by the researcher) which are both causal and parametric (since each proportion is algebraically determined by its components). Raw variables are in black and constructed variables are in double orange circles (by convention). Dashed orange arrows indicate the associations (and null hypothesis) that are commonly tested, despite introducing bias from mathematical coupling.

In the example considered by Neyman and simulated here to replicate his hypothesised dataset 10,000 areas are simulated.

The true causal effect between the number of storks and the number of babies is 0.00.

The areas are simulated from the following distributions:

$$Women \sim Normal(mean = 40000, sd = 10000)$$

$$Births \sim Normal(mean = 30, sd = 5)$$

$$Storks \sim Normal(mean = 5, sd = 1)$$

The simple correlation between the number of births and the number of storks is 0.00, the correlation between the number of births and the number of storks both divided by the number of women is 0.82 and the association between the number of births and the number of storks, conditional on the number of women, is 0.00.

These three variables are then simulated with the constraint that the number of births must

be fewer than the number of women, when there are no women there are no births, and that the number of women is a cause of the number of births instead.

The true causal effect between the number of storks and the number of babies is 0.00.

The variables are drawn from the following distributions:

$Women \sim Normal(mean = 40000, sd = 10000)$

$Births \sim Binomial(size = Women, successprobability = 0.1)$

$Storks \sim Normal(mean = 5, sd = 1)$

In this case, the simple correlation, the ratio correlation and the conditional relationship between the number of storks and the number of births are all 0.00. The reason that this is the case, is that the number of births is now a proportion of the number of women.[101] This means that dividing one by the other 'standardises' the numerator by the denominator because the relationship between the two is a straight line which goes through the origin. This is more realistic than the figures Neyman used because the number of births is zero when the number of women is zero, i.e. there cannot be any births without there being women.

4.2.11 Geographical Examples

The above examples are further expanded, developing the notion of 'constraint' introduced by Neyman's storks example, to look at what happens in a health geography context, where researchers are often interested in the exposures and outcomes that occur at the population–level with the potential that both X and Y are constrained by N. From this point on, N is referred to instead of Z and all variables are population counts. N is the total population and X is the number of people experiencing an exposure and Y is the number of people experiencing an outcome.

In this case, there are three variables, representing the number of individuals who experience an exposure (X), the number who experience an outcome (Y), and the total number of people in a geographically–defined population (N) which acts as a potential limit for both X and Y. These examples are an expansion of the Neyman example because X and Y are subsets of N. Note, however, that X and Y are not compositional —they cannot be combined to make another subset of N (Section 3.10.3).

In the previous examples the simulations were simplified so that the variables were generated from a multivariate normal distribution, however, that would not be appropriate under this situation as it does not allow for the constraint that $X \leq N$ and $Y \leq N$. In this case, X and Y are simulated from N by randomly generating the population size (N) and using the binomial distribution to generate X and Y from N; akin to flipping a coin for each member of the population and recording when it comes up tails, this can be done for various 'success' probabilities.

In this case, it is assumed that N causes both X and Y (Scenario 1 from earlier). Following the results from above, to analyse the relationship between X and Y, avoiding 'spurious' correlation, it was shown that N would have to be conditioned on as it is a confounder of the exposure–outcome relationship.

In these geographical examples, N represents the population size of each area and the DAG represents relationships at the area–level. Although this is not a common use of DAGs in the literature, DAGs can be used to describe any data–generation process.[37] In these examples, as each variable represents a property of a geographical area, to simplify it is assumed that the areas are independent of each other, i.e. changing the level of exposure in one area has no effect on the outcome in another area (consistent with the null hypothesis).

The examples are also simplified by assuming that there is no causal relationship between the exposure X and outcome Y and that the only true causal relationships are $N \rightarrow X$ and $N \rightarrow Y$. This means that if any correlations are found between X and Y or $\frac{X}{N}$ and

$\frac{Y}{N}$ they are 'spurious'. To simulate these situations, population sizes (N) are randomly generated from a negative binomial distribution for a large number of areas (1,000) and from these the binomial distribution (with various probabilities of success, ranging from 0 to 1) are used to simulate a number of people in the population experiencing the exposure (X) and the outcome (Y). This data generating process implies that X does not cause Y and that Y does not cause X, and both are simulated independently.

From the simulated data it is shown how varying the 'success probability' of the X and Y variables affects the 'spurious' element generated from correlating $\frac{X}{N}$ with $\frac{Y}{N}$.

Figure 4.2.13 shows the amount of 'spurious' correlation present when X and Y are generated for 1,000 areas with population sizes (N) simulated from a negative binomial distribution. X and Y are generated from a binomial distribution, the equivalent of flipping a coin for each member of the population and recording how many tails there were. The probability of getting a tail is varied between 0 and 1. This shows that the 'spurious' correlation varies between -0.01 and 0.01 for all success probabilities.

The simulations used in this case assume that the only driver of the exposure and the outcome is the population size and that the exposure and outcome are proportional to the population size.

In this situation, there is an assumption that there are no other variables involved in determining the outcome, and in particular, that area–level attributes do not affect the exposure–outcome relationship. In reality, the variables are likely to be more complex than that illustrated here and that is explored further in the discussion.

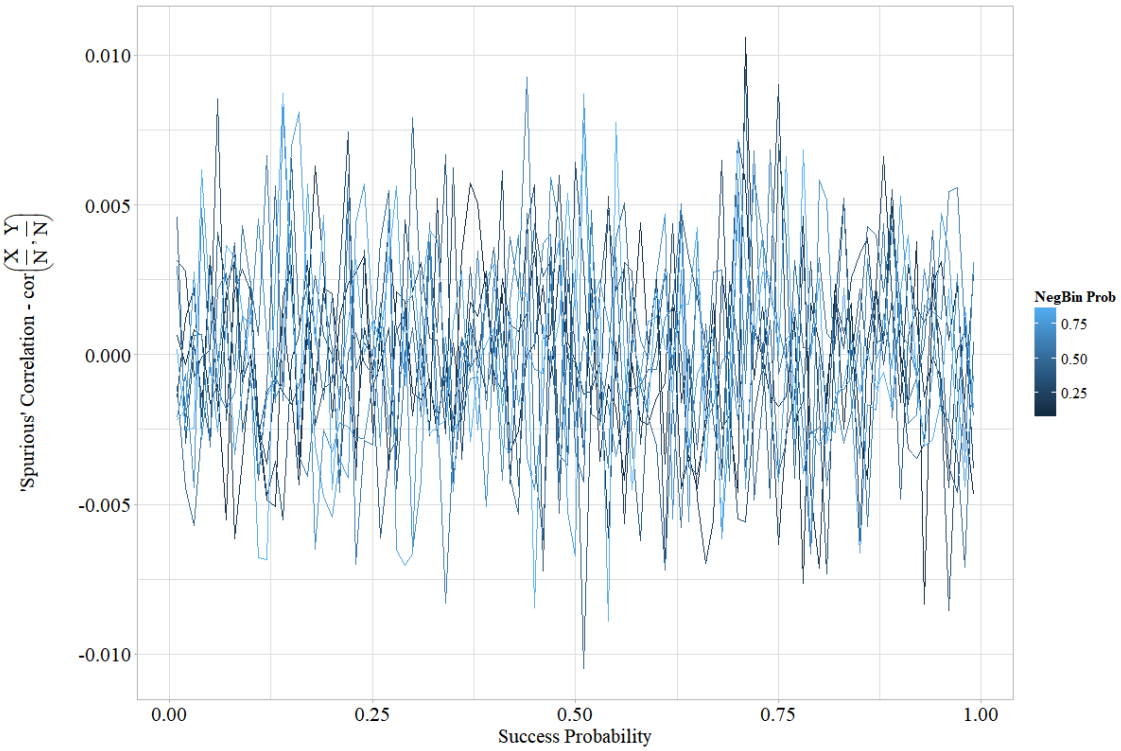

Figure 4.2.13: Plot showing the 'spurious' correlation from the analysis of $\frac{X}{N}$ and $\frac{Y}{N}$ when varying the success probability (between 0 and 1) of the exposure and outcome generated from a binomial distribution. A health geography example.

4.3 Discussion

This chapter revisits the long–acknowledged but enduring problem of mathematical coupling of proportions through the modern lens of graphical model theory. By embracing the utility of causal diagrams for exposing analytic errors, it is shown how large inferential bias can be introduced in a variety of contexts. Explorations have been made as to how and when Pearson, Neyman, and Fisher's suggested fix offers a robust solution, i.e. when the common denominator is a confounder of the exposure–outcome relationship.

The original example was extended, using simulation, to a health geography context to show that dividing through by a common denominator may be sufficient for

confounder adjustment in the setting when the numerator and denominator of the ratio are proportional. However, these are only preliminary findings under the null hypothesis of no causal effect of exposure on the outcome and with simplified conditions of cluster heterogeneity. The non–null scenario should be simulated to explore this case further. It is unlikely that, in complex data from which causal effects are estimated, that this proportionality is the only effect acting between these variables and these effects cannot be accounted for using ratios alone. In this case using the raw variables would always be preferred.

The examples here have been deliberately simplified by omitting many potential confounding and mediating variables and by retaining cluster heterogeneity as simple, i.e. by not envisaging putative causal agents differentiating cluster means, for example. The problems of mathematical coupling are not however lessened by extra complexity. On the contrary, contextual variables can have different causal relationships with the different elements of a composite, creating greater analytical problems. For these more complex scenarios, it will be important to take an in–depth look at these relationships and DAGs will be useful in elucidating the most appropriate analytical strategy, as demonstrated here. In most practical scenarios, evaluation of the methods deployed via simulation would be extremely beneficial.

The apparent suitability of dividing through by population size in the health geography example was discovered late in the development of the research undertaken for this thesis. It is further explored in the context of limiting long–term illness in Chapter 5 where simplification was maintained in the first instance to elucidate potential inferential biases that arise through constructed indices and the adverse influences of mathematical coupling. The implications of more complex data generation processes will be discussed and is a focus of the further work suggested in the Chapter 7.

Although DAGs are typically used to represent causal relationships between variables at the individual–level, in the health geography examples seen here they are also used to

represent relationships at an area–level, which is common in population health studies. Therefore, all simulations and subsequent analyses are conducted at the area–level (e.g. using aggregate data). In this situation, area–level heterogeneity is effectively equal to unknown confounding. Although this area–level mathematical coupling can be avoided by conducting analyses at the individual–level (e.g. using logistic regression models), this is often not possible in geographical or ecological studies. It could also be avoided within a multi–level analysis, as this can 'accommodate' any cluster heterogeneity present and if not explicitly accommodated, the results may be affected by area–level unmeasured confounding.

Mathematical guidance is available on the use of functionally deterministic variables in causal diagrams, which outlines how to define and determine conditional independence between the raw and composite variables.[24] Researchers wishing to conduct analyses using composite variables, must follow these rules to ensure appropriate analyses and interpretation. However, when one or more composite variables are wholly determined by one or more parent variables, as with proportion variables, it is important to decide whether any additional information is captured in these composites or not. If no additional information is provided by including both the raw and composite variables, and there is a clear risk of confusion, it may be more appropriate for DAGs to favour including raw component variables over derived composite variables.

Although robust causal analysis requires the use of raw variables within correlation or regression, it is recognised that proportions are still helpful and offer meaningful descriptive summaries. Analysts should not be afraid of transforming their final results into whatever format is easiest to interpret, but should be extra clear that the underlying analyses were performed robustly using raw variables or the risk of inferential bias through analyses of coupled variables must be quantifiable, perhaps through the implementation of simulations to evaluate the methods adopted.

The headline message is clear: analyses of proportions with shared common denominators

are at risk of substantial inferential bias from mathematical coupling. Pearson, Neyman, and Fisher's solution —of conditioning on the denominator —is robust for many situations, but only a judicious use of causal diagrams can provide the appropriate analytical strategy for all scenarios.

4.4 In the context of the thesis

'Spurious correlation' as a result of what has more recently been termed 'mathematical coupling' was first reported in 1896[90] and has repeatedly been reported in the literature, however, this is the first time that it has been approached from a causal inference perspective. This has uncovered when the historical solution is appropriate, i.e. where the denominator is a confounder of the two numerator variables. The extent of 'spurious correlation' due to mathematical coupling was also quantified in the geographical situation where an exposure (X) and outcome (Y) are both components of the common denominator (N), i.e. $X \leq N$ and $Y \leq N$ but $X + Y \neq N$. It is hoped that this more contemporary perspective on mathematical coupling bias (and the addition of a health geographical point of view) can bring it to the attention of a wider audience who will avoid it in their own work and recognise it in the work of others.

Simulations are an excellent place to start when thinking about data analysis as this allows one to address any avoidable analytical biases. However, there may be some biases present that cannot be addressed at this stage, for example, conditional data acquisition (Chapter 6) and cross–level interactions amongst the unmeasured/unmodelled confounders (discussed in Chapter 7).

The next chapter investigates how the issues of deprivation and morbidity are analysed in the literature in light of mathematical coupling bias using causal inference methods, including compositional and composite data as introduced in Section 3.10.3, and simulation.

Chapter 5

Limiting Long-Term Illness and Deprivation

5.1 Introduction

In Chapter 4, the 'spurious' correlation as a result of mathematical coupling was introduced. In health studies, proportions and percentages can seem more informative than raw counts and appear to be of greater interest to analysts.[102] However, it is unclear to what extent, if any, mathematical coupling biases results in this area.

5.2 Background

Proportions are ubiquitous in observational research, but their mathematical coupling has not been investigated extensively in relation to studies in health geography. 'Spurious' correlation has been investigated in the case of $X - Y$ and Y, for any random variables X and Y, related to geographical problems,[103] and although the case of $\frac{X}{Z}$ and $\frac{Y}{Z}$ has been reported in the field of demography,[104] studies using potentially mathematically coupled

data in related areas persist (for example[105]). The problems generated by mathematical coupling arise for proportions, rates, ratios, prevalence and incidence —these terms are used in this Chapter, along with 'ratio index variables', interchangeably.

Proportions and ratio index variables are new variables derived from the division of one variable by another. Some health geography–related research is concerned with prevalence and incidence (counts of total cases per population and counts of new cases per population per unit time, respectively), which are ratios that capture the relative frequency of a condition (e.g. prevalence of obesity, incidence of mortality) by accounting for differences in population sizes. Many variables are generated as ratios to capture human features (e.g. obesity), acknowledging that humans vary due to genetic predisposition (e.g. height). Such ratios seek to capture a relative construct (e.g. Body Mass Index as a measure of weight relative to height–squared), or to create a variable which aims to summarise related concepts into a composite variable that is considered more useful or parsimonious (including proxies for unmeasurable latent variables, e.g. social deprivation). Thus, researchers calculate ratio index variables from the available absolute values; they cannot be measured directly and must be constructed, and they can have their own unique scale if they are constructed from one or more variables that are measured on different scales.

If ratio index variables are analysed as if they were raw variables (i.e. as if they capture a single concept) this may introduce inferential bias. By forming ratios, information that is contained in separate components is compressed, but where denominators are the same in different ratio index variables that are being analysed using correlation or regression, coefficients will comprise an expected positive effect (due to the algebraic dependency introduced by the common denominator) along with the true effect,[96] which could be zero. The implications of bias due to mathematical coupling amongst ratio variables are numerous, yet almost no attention is given to the artefacts generated within epidemiology, health geography or observational research more generally.

It is common for area–level measures of health outcomes and mortality to be analysed in relation to indicators of social deprivation. This Chapter therefore seeks to illustrate the issue of mathematical coupling and resultant inferential bias within a health geography context using analyses similar to those present in the literature. It is reported how causal inference theory could be used to inform appropriate analyses. The example of limiting long–term illness (LLTI) and social deprivation as captured by area of residence and represented by the Townsend Deprivation Index[106] (and the individual components thereof) is used.

The Townsend Index aims to capture the concept of material deprivation, which cannot be measured directly. The Townsend Index is a composite variable that is itself made up of individual compositional variables. These individual compositional variables are ratios formed from numerators divided by denominators of the total population or the number of households. Each of the components of the Townsend Index may be affected by mathematical coupling when analysed along with another composite variable that has a component in common. This bias may be further complicated when the components are combined to form the Townsend Index.

To indicate how robust the analysis of a specific research question might be, the role of all data components are clarified using directed acyclic graphs (DAGs) and graphical model theory. Datasets are simulated at the same spatial scale and in which LLTI prevalence is determined solely by population size; i.e. the null hypothesis whereby the number of persons reporting an LLTI is simply a function of population size. Generating data under the null hypothesis ensures there is no 'true' causal relationship, thereby allowing evaluations of the magnitude of any artefact due to mathematical coupling. These simulated datasets are analysed using methods present in the literature and outputs are compared to results from analyses informed by a DAG. Finally, the observed data are analysed and comparisons are made between models which are specified to avoid mathematical coupling and those previously used in the literature.

5.3 Methods

5.3.1 Overview of Methods

Data were simulated based on those from the 1991 UK Census; these data were linked with a simulated number of the population reporting an LLTI for each electoral ward (see Section 5.3.5 for why electoral wards were chosen) in England and Wales. The number of people with an LLTI was simulated under the null hypothesis where only the population size determined the prevalence of LLTI and no other variables (i.e. those involved in the calculation of the deprivation index) caused LLTI.

5.3.2 Variables of Interest

Limiting Long–Term Illness

A question regarding LLTI was introduced to the 1991 British Census; before this, mortality was likely to be used as a proxy for morbidity,[107] however, LLTI is a broader health status measure[108] which records non–life–threatening illnesses that burden health services.[109] LLTI is a self–reported measure counting those who answered affirmatively to the question: "does the person have any long–term illness, health problem or handicap which limits his/her daily activities or the work he/she can do? Include problems due to old age". Data from the 1991 Census were used due to this being the first year that a question regarding LLTI was asked which prompted studies in this area.

Townsend Deprivation Index

Deprivation is a latent variable (i.e. it is not possible to measure it directly) and is generally recognised as a composite concept, with measurable 'proxy' variables

being combined to represent it.[110] The Townsend Index comprises four area–level ratio index variables which are standardised and summed: percentage of economically active individuals who are unemployed, percentage of households that are not owner–occupied, percentage of households without access to a car, and percentage of households that are overcrowded (have more than one person per room). These percentages are standardised using z–scores and the Index scores are the sum of these equally weighted components. Before summation, the unemployment and overcrowding percentages are logarithmically transformed as they tend to be highly skewed. A greater Townsend Index indicates a greater level of deprivation.[106] A variety of deprivation indexes have been constructed but the use of input variables expressed as proportions is ubiquitous (e.g. the Carstairs Index[111] and Jarman Index[112]). The Townsend Index was used here as it was readily available in free to access datasets.

All variables forming a composite variable that is subsequently used as an independent variable (exposure) in a model are implicitly assumed to precede the dependent variable (outcome) in time; the exposure must occur before the outcome to be a cause. There may however be some question surrounding whether components of the deprivation index always precede the dependent variable of interest, as components may be causally affected by the dependent variable. It has been suggested, for instance, that unemployment is a more straightforward measure of deprivation than a composite measure, as it generates a 'stronger' association with LLTI than the composite measure,[113] yet it is unclear whether unemployment is a cause or consequence of LLTI. Although a strong relationship is often found between unemployment and LLTI, this cannot provide any information regarding the direction of causation.[114] During periods of recession, the strength of association between unemployment and poor health has been known to decrease as more people become unemployed due to economic conditions.[115] Notwithstanding these important issues, to illustrate the methodological concepts this thesis addresses, it is assumed that unemployment precedes LLTI in time, as this is what is taken for granted in most studies

using composite variables.

5.3.3 Literature Search

A keyword search of 'Limiting Long*Term Illness' and 'Deprivation' using Web of Science returned 37 research articles since 1991 when a question regarding LLTI was added to the census. This list was reduced to 16 articles[107,108,113,114,116–127] by manual review of the methods; papers using linear or Poisson regression, as well as those calculating the correlation only, were included. Any papers using logistic or any form of multilevel regression were omitted because they focus on individuals rather than aggregated data (i.e. data based on geographical areas) and therefore avoid the form of mathematical coupling investigated here. However, methods considering how individual data linked to areas could be analysed in a causal framework will be discussed in Chapter 7. Papers conducting both individual and area–level analyses were retained, but only the area–level analyses were considered. The methods used in each of the retained articles were studied to form the basis of the comparisons between the unbiased analyses this thesis seeks and the analyses presented in the literature, the article summaries are presented in Table 5.1. Of the 16 articles retained, 7 considered correlations, 6 used linear regression, 2 used Poisson regression and 1 used both linear and Poisson regression. All 'statistically significant' results reported indicated a positive relationship between LLTI and the exposure of interest.

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[107]	Linear regression	LLTI or all cause mortality (SMRs and SIRs separate for males and females and under 65s, 65-74, and 75+)	Social deprivation	Indicators of social deprivation: % unemployment (males and total), Townsend index, Jarman index, Carstairs index, and DoE index	Carstairs (then Townsend, Jarman and DoE indices) best predictor of SMRs and SIRs. Unemployment rates simpler alternative measure for deprivation
[108]	Correlations (main method: multilevel logistic regression)	LLTI	Material deprivation	Age, age squared, social class 4/5 (binary), non-white ethnicity (binary), married (binary), and deprivation indicator (more than 1 person per room, non-owner occupied household, household without car, no access to separate bathroom, unemployed; scored 0-5)	Area factors have significant association with individual health outcome though effect smaller than properties of individuals

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[114]	Linear and Poisson regression (log of expected counts is an offset)	Age–sex standardised LLTI	See covariates	% males working in coal industry 1981, % males working in coal industry 1971, % residents working in energy and water industries 1991, % unemployed residents with most recent job (last 10 years) in energy and water 1991, % EA with unskilled/semi-skilled manual occupations 1991, % residents 16+ unemployed 1991, % households without car 1991, % households not owner-occupiers 1991, % households in terraced dwellings 1991, % households more than 1.5 persons per room 1991, % households no central heating 1991, % households lacking bath/shower or inside toilet 1991, % non-white residents 1991	Positive associations between LLTI and all 8 variables

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[116]	Linear regression	Age–sex standardised LLTI	20 socio-economic variables	Unemployment, long–term unemployment, children in non–earner households, educational attainment and income support, social class, religious affiliation, rented households, households without a car, households without central heating	Variation in morbidity ratios explained by socio–economic variables (77.9%). “Income support a particularly strong predictor”
[113]	Correlations	LLTI	Socio–economic variables: employment and economic activity, ethnicity, household amenities, household characteristics, household tenure, Jarman index, Townsend index	None –correlations	“Some specific areas of morbidity did indeed show strong associations with socio–economic disadvantage.”

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[117]	Correlations	Standardised mortality (all cause, and specific cause; 1993–95) and illness ratios (1991 Census)	Townsend index and urban–rural dichotomous variable	None –correlations	Large correlation between Townsend and LLTI (0.82); similar for urban–rural split. Large correlation between Townsend index and all–cause mortality
[118]	Linear regression	Standardised illness ratios based on LLTI	Townsend, Carstairs and Jarman indices	Region and labour market conditions	The four health measures were related to social deprivation indicators and region
[119]	Linear regression	Age–standardised LLTI	Townsend score,measure of variation in Townsend scores, locality measure of variation in Townsend scores, log migration	Combinations of variables used in single variable models	“Significant, positive relationship between age–standardised limiting, long–term illness and deprivation”. Townsend index “most significant”

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[120]	Correlations (main method: multilevel Poisson regression)	Indirectly standardised (by age and sex) premature (0–64 years) LLTI	Carstairs, Jarman, Townsend, and Department of Environment indices and customised deprivation profiles	None –correlations	"Premature LLTI is positively correlated with all of the deprivation indices"
[121]	Poisson regression (one at individual and one at ward–level)	Indirectly standardised (by age and sex) LLTI	Carstairs index (individual level), McLoone and Boddy index (ward level), and models with individual components of these indices	Age, sex, district	Ward level analysis was not sufficiently good at explaining variation in illness across region which the authors contribute to the ecological fallacy
[122]	Poisson regression (main method: multilevel Poisson regression)	Age and sex standardised illness ratios using LLTI	Rurality indicator	"Poisson regressions were carried out to find the socio–economic and demographic variables associated with illness for each area type"	Not reported for single level models

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[123]	Correlations	Premature LLTI (0–64) and premature mortality (0–64), indirectly standardised for sex and age (5 year groups)	Index of Multiple Deprivation (made up of 33 indicators from: income; employment; health deprivation and disability; education, skills and training; housing; and geographical access to services) compared to Townsend score	Rurality (14 categories, interest in "rural", "rural fringe" and remaining 12 categories were combined)	Correlation between LLTI and Townsend Score was 0.76, and with IMD was 0.79
[124]	Correlations	Age–sex standardised LLTI	Townsend score and its individual components	None –correlations	"the individual unemployment component is more strongly associated with LLTI... than the composite Townsend score"

Table 5.1: Summaries of the articles retained from the literature search

Ref.	Method	Outcome	Exposure	Covariates	Results
[125]	Linear regression	Logged standardised LLTI rate	Townsend score quintile		All significant regression coefficients
[126]	Correlations (main method: multiple correspondence analysis)	Standardised morbidity ratios	Area level 'health resilience' and ethnic composition, residential mobility, employment type, housing tenure, and an indicator of social cohesion	None –correlations	All morbidity indicators significantly correlated with each other
[127]	Linear regression	Logged standardised proportion of 'not good health' and LLTI rate	Area deprivation and a deprivation differential	Univariable regression	Positive relationship between deprivation and morbidity 'indicators'

5.3.4 Simulation of Datasets

All analyses were performed on simulated data. 1,900 datasets were generated to have the approximate correlation structure and variable distributions[75] as the electoral wards for England and Wales in the 1991 Census (see Section 5.3.5 for why electoral wards were chosen as the unit of analysis). Wards with resident populations fewer than 200 persons were removed from the dataset as these wards may not be representative of the population in general as they tend to be in very remote areas or financial districts. Cases of LLTI were generated under the null hypothesis; population size was the only variable determining the number of individuals reporting LLTI and the prevalence of cases was taken to be the mean national average in 1991 over all age groups (13.5%). The raw components of the Townsend Index and their opposite counterparts (number of economically active residents, number of economically active **unemployed**, number of economically active **employed**, number of private households, number of private households with **more than** one person per room, number of private households with **one or fewer** persons per room, number of private households **without** a car, number of private households **with** a car, number of private households that **are not** owner occupied, and number of private households that **are** owner occupied) were simulated and the Townsend Index was calculated from these. The opposites are needed because these are compositional data made up of two components each and by generating both, each pair can be scaled so that the totals (e.g. number of households) are the same.

Distributions from which variables were simulated were chosen by plotting the observed variable distributions and fitting general distributions to them, Figures 5.3.1–5.3.5. Summary statistics were calculated for all of the variables that were to be simulated (Table 5.3). These figures show that the log normal distribution is a good fit to the observed data, indicated by the simulated data from the log normal distribution having a line close to that of the observed data. This is also confirmed by the summary statistics of the observed data and log normal simulated data having similar values (Table 5.3). From

Section 3.10.3, simulations were built up from the raw variables, therefore **employment** and unemployment figures were generated, etc. Although a Poisson distribution is often the go–to distribution when looking at count data, it can be seen from the summary statistics (Table 5.3) that each of the variables is over–dispersed (the variance is larger than the mean) and that the Poisson distribution would not be suitable in this case. Table 5.2 shows correlation structure used to generate the datasets. The random number generator seed was set 9,499 values apart for each iteration (i.e. equal to the number of electoral wards) to avoid dependence between datasets.[51,66]

Table 5.2: Correlation matrix of the observed data to be emulated in the simulated datasets. Non Own = Households not owner–occupied; Own = Households that are owner–occupied; No Car = Households without a car; Car = Households with a car; Overcrowded = Households that are overcrowded; Not Overcrowded = Households that are not overcrowded; Unemployed = Population that is unemployed; Employed = Population that is employed; Population = Population in each ward.

	Employed	Unemployed	Non Own	Own	No Car	Car	Overcrowded	Not Overcrowded	Population
Employed	1.00	0.72	0.66	0.95	0.74	0.97	0.56	0.97	0.97
Unemployed	0.72	1.00	0.92	0.58	0.95	0.61	0.85	0.82	0.83
Non Own	0.66	0.92	1.00	0.46	0.95	0.54	0.88	0.78	0.75
Own	0.95	0.58	0.46	1.00	0.62	0.98	0.42	0.91	0.92
No Car	0.74	0.95	0.95	0.62	1.00	0.63	0.78	0.87	0.84
Car	0.97	0.61	0.54	0.98	0.63	1.00	0.44	0.93	0.94
Overcrowded	0.56	0.85	0.77	0.42	0.78	0.44	1.00	0.63	0.65
Not Overcrowded	0.97	0.82	0.78	0.91	0.87	0.93	0.63	1.00	0.99
Population	0.97	0.83	0.75	0.92	0.84	0.94	0.65	0.99	1.00

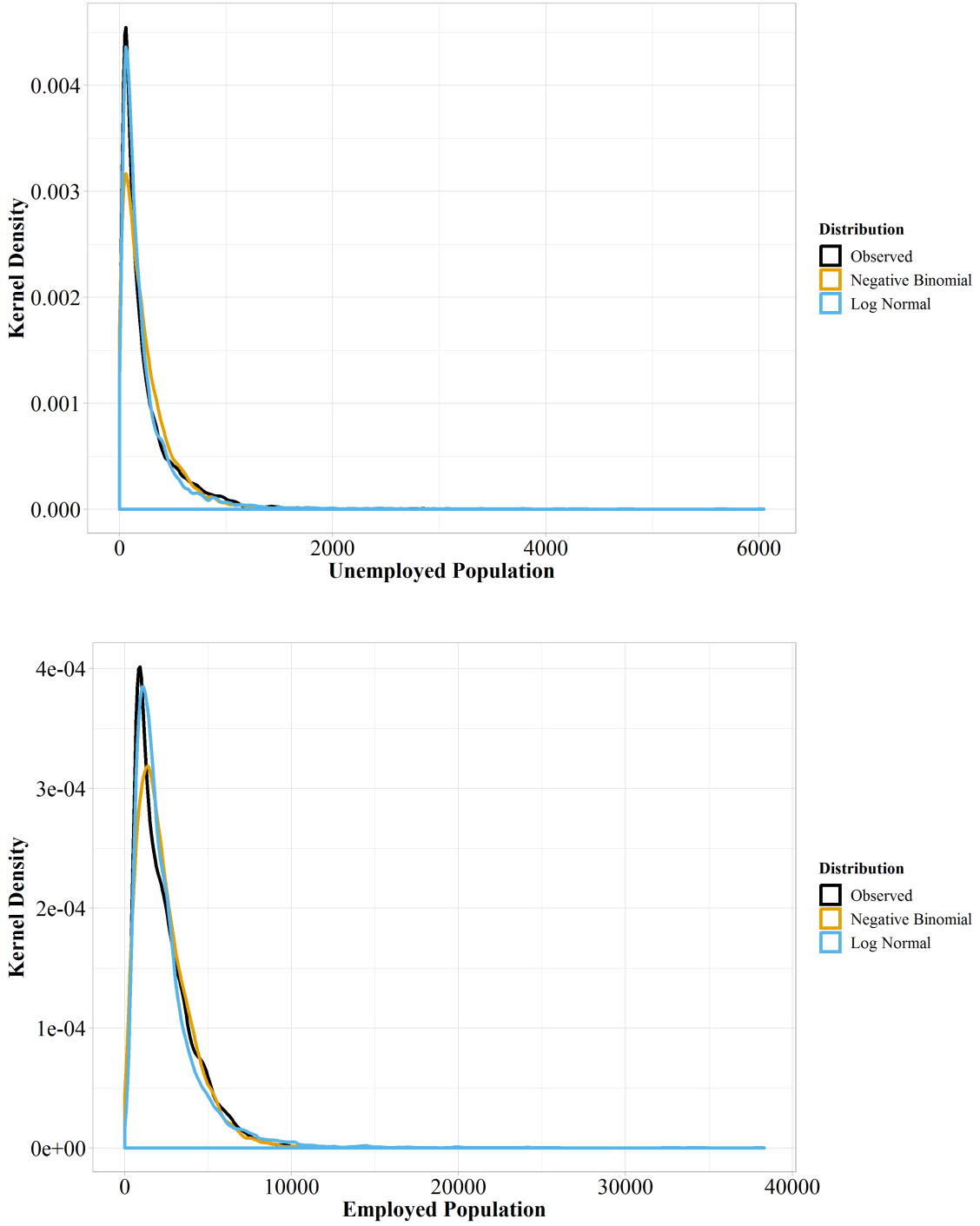

Figure 5.3.1: Distribution of observed employed and unemployed population variables with fitted negative binomial and log normal distributions.

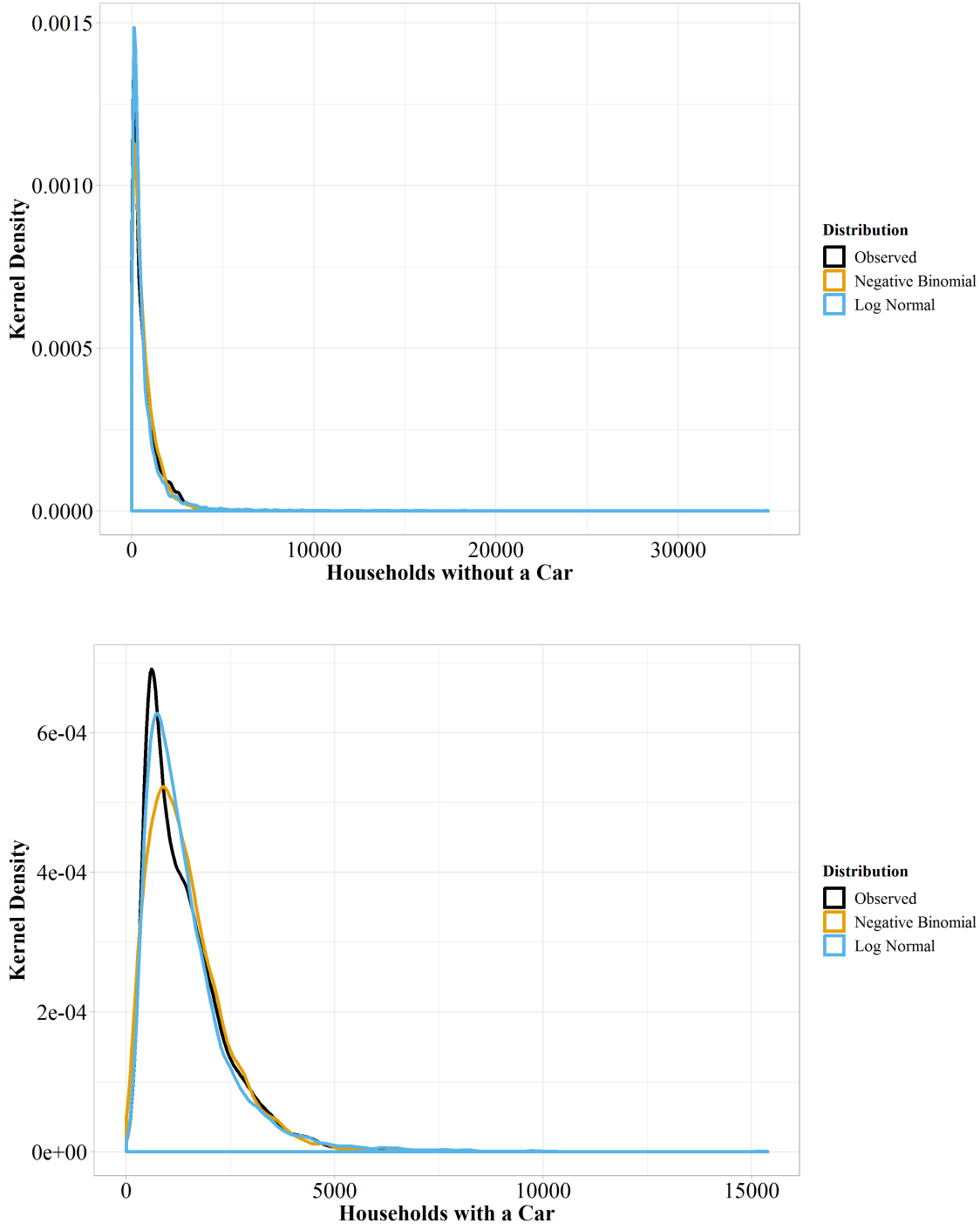

Figure 5.3.2: Distribution of observed variables for households with and without a car with fitted negative binomial and log normal distributions.

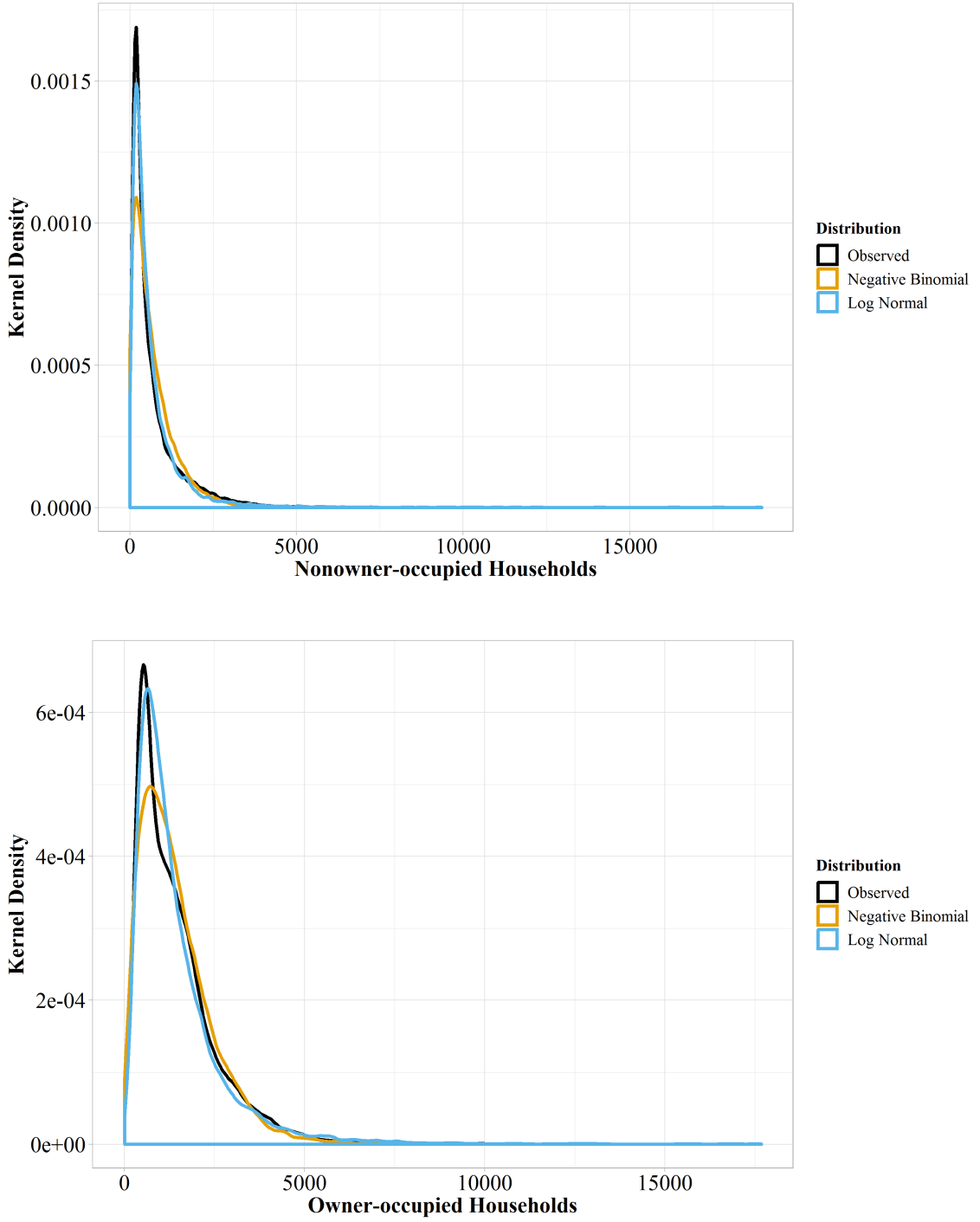

Figure 5.3.3: Distribution of observed variables for households that are non-owner occupied and owner occupied with fitted negative binomial and log normal distributions.

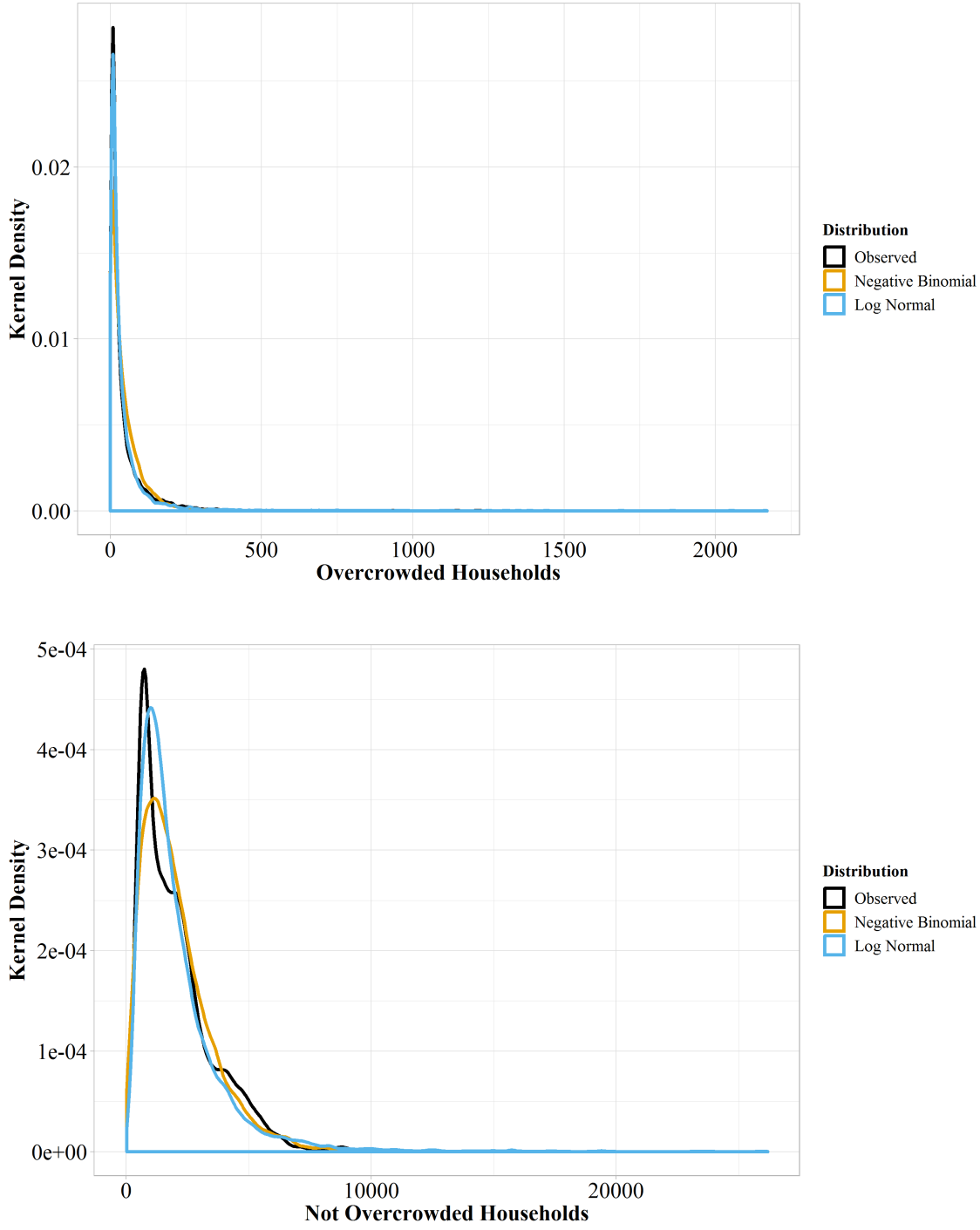

Figure 5.3.4: Distribution of observed variables for households that are overcrowded and not overcrowded with fitted negative binomial and log normal distributions.

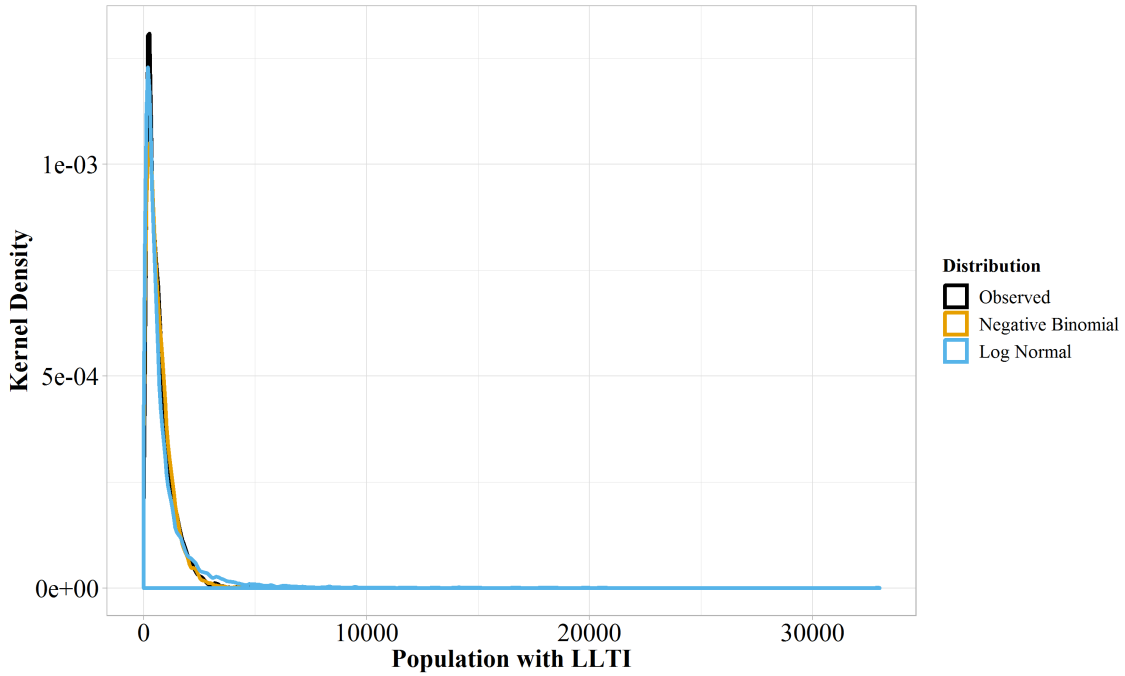

Figure 5.3.5: Distribution of observed variables for the population with a limiting long-term illness with fitted negative binomial and log normal distributions.

Table 5.3: Summary information for each variable to be simulated. Components of the Townsend Index and Limiting Long–term Illness. Non Own = Households not owner–occupied; Own = Households that are owner–occupied; No Car = Households without a car; Car = Households with a car; Overcrowded = Households that are overcrowded; Not Overcrowded = Households that are not overcrowded; Unemployed = Population that is unemployed; Employed = Population that is employed; LLTI = Population with a limiting long–term illness.

	Non Own	Own	No Car	Car	Overcrowded	Not Overcrowded	Unemployed	Employed	LLTI
Minimum	14	28	7	92	1	108	3	145	1
1st Quartile	186	604	143	681	8	844	60	1004	254
Median	379	1138	399	1175	19	1665	137	1888	511
Mean	671.2	1424	678.1	1426	43.65	2060	235.3	2352	683
3rd Quartile	841.5	1890	900	1884	49	2730	302	3210	900
Maximum	7272	10276	7132	9860	1215	12297	3206	15396	4576

5.3.5 Why electoral wards?

Before electoral wards were selected as the unit of analysis the effects of the choice of area size on analyses of the area–level variable Townsend Index were considered. The Modifiable Areal Unit Problem (MAUP; introduced in Section 3.10.4) is a well documented problem which is looked at here in relation to its potential relationship with mathematical coupling. In this context, it may be expected that a smaller granularity with less variation in the population size would be preferred given the effect seen on the extent of mathematical coupling for different sizes of the coefficient of variation when all variables are normally distributed (Chapter 4). In the previous Chapter, area population sizes were simulated from which exposures and outcomes were drawn using the binomial distribution (Section 4.2.11). Here, population sizes were obtained from the 2011 Census for electoral wards (EWs), Lower Layer Super Output Areas (LSOAs), Middle Layer Super Output Areas (MSOAs) and Output Areas (OAs) and exposure and outcome variables were generated from these using the binomial distribution with success probabilities varying between 0 and 1. This was conducted in order to determine whether any area granularity will inherently bias the analysis more than any other with regards mathematical coupling.

Figure 5.3.6 shows that, under the null hypothesis, there is no discernible difference or clear pattern in bias associated with the different area sizes under the null hypothesis (the 'spurious' correlation averages zero). As a result, electoral wards were used for analysis of this problem as they are most often used to answer such questions in the literature.

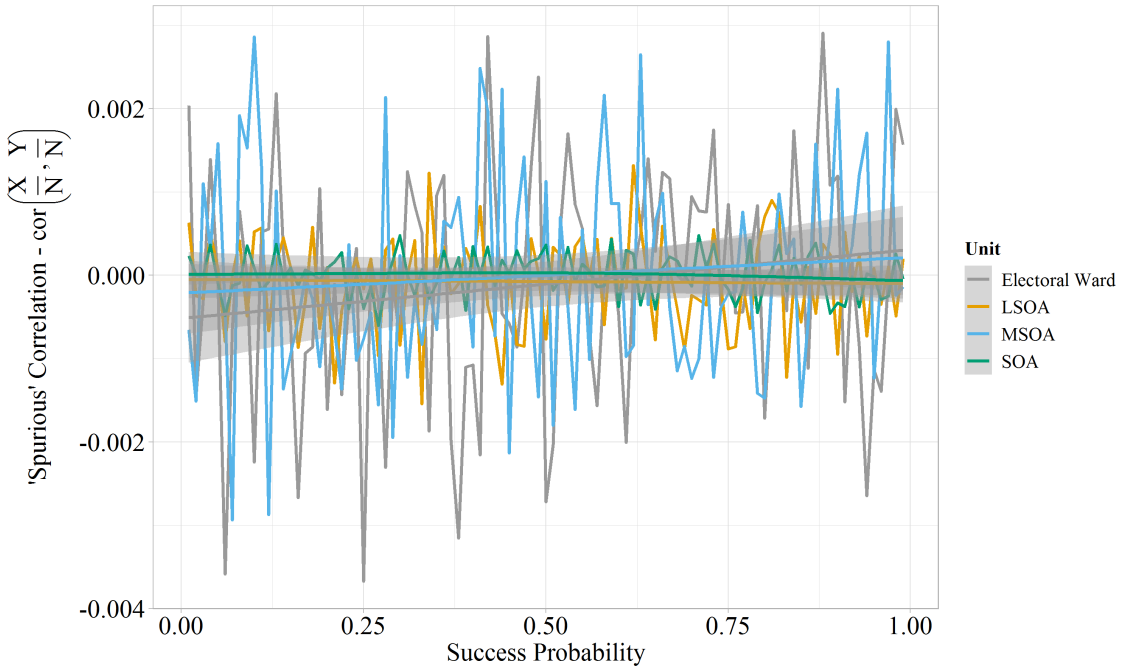

Figure 5.3.6: Plot showing the 'spurious' correlation from the analysis of $\frac{X}{N}$ and $\frac{Y}{N}$ when varying the success probability (between 0 and 1) of the exposure and outcome generated from a binomial distribution for each areal unit: Electoral Ward, Lower Layer Super Output Areas (LSOAs), Middle Layer Super Output Areas (MSOAs) and Output Areas (OAs).

5.3.6 Contrast between how Townsend Index components are simulated and analysed

To simulate this problem, the component variables of the Townsend Index were generated along with their opposite components. These are examples of compositional data, i.e. data that comprise parts of some whole, for which all parts sum to that whole.[128] They differ from composite variables as both the components and whole can be measured directly and are on the same scale (Section 3.10.3).

Figure 5.3.7 shows the compositional make up of the components of the Townsend Index. This diagram highlights the variables required to simulate the problem, and how each

variable is part of a pair which adds up to a preceding total. Although pairs are shown in this diagram, this is not how the data would be analysed and researchers should be particularly mindful of the question that they are trying to answer and how they can accurately achieve this when dealing with compositional data. For example, if one was interested in the effect of moving one person from the unemployed population to the employed population on limiting long–term illness they could regress the unemployed population on LLTI whilst adjusting for the total population. This is equivalent to keeping the total population fixed, whilst changing the size of the unemployed population with an equal and opposite change in the size of the employed population. This is the analysis that is performed most often in the literature, although it is not often made explicit.

It is more difficult when considering the Townsend Index because there are several compositional variables that make up this composite variable. It is not possible to consider the effect of moving one person from the unemployed group to the employed group in this case (or moving a household from being overcrowded to not overcrowded, for example). However, when the individual components of the Townsend Index are considered, previous work has mostly looked at the effect of moving from one category to another, rather than increasing the number in one category *and* the total.

5.3.7 Causality and Directed Acyclic Graphs

Directed acyclic graphs (DAGs) were used to illustrate all assumed associations between variables when analysing LLTI in relation to deprivation as measured by the Townsend Index. DAGs are used to inform robust model choices (i.e. unaffected by confounding) using minimally sufficient adjustment sets (MSAS).[74] The results from these DAG–informed models were analysed and compared to results from models replicating analyses conducted in the literature.

The focus here is only on the Townsend Index and its components and other variables

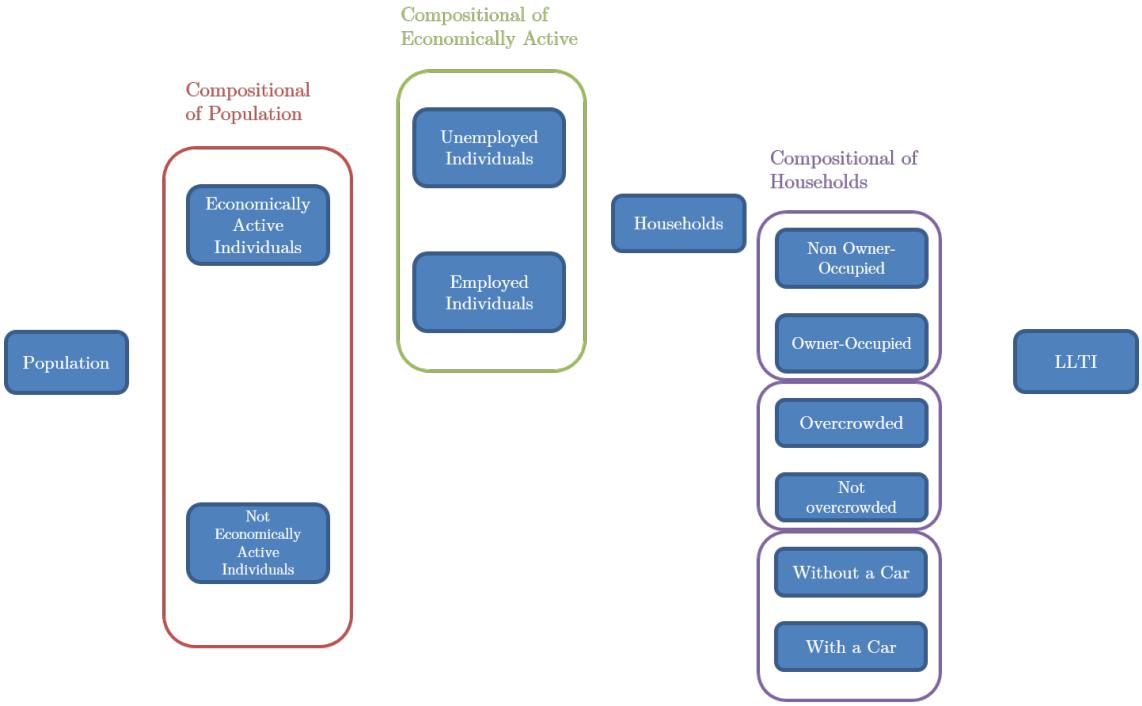

Figure 5.3.7: Diagram showing the compositional components of the Townsend Index components for simulation.

that were found in the literature search were not included in the models. When including many covariates in a model, and not consulting causal graph theory, researchers should be aware of the 'Table 2 Fallacy'[49] and the 'reversal paradox'[32] (introduced in Sections 2.6 and 2.1.10, respectively).

5.3.8 Analytical Methods

The literature search informed the choice of models used in the analysis. Each of the chosen analytical methods was applied to the 1,900 simulated datasets and the relevant regression or correlation coefficient (i.e. for deprivation as characterised by the Townsend Index or components thereof) was recorded along with its confidence interval. The `R` statistical software[98]) was used throughout and the code is available in Appendix C.

Three analytical techniques were applied to the data: Poisson regression, because this

often appears in the literature due to the outcome being a count; linear regression because it is often used to analyse the individual components of the deprivation composite with the outcome (sometimes as a crude rate, often standardised directly or indirectly); and correlation which is used for descriptive purposes. The standardised rate of illness (of LLTI in this case) is known as the Standardised Incidence Rate (SIR), which is similar to the Standardised Mortality Rate (SMR), and in this case represents the ratio of the population in an electoral ward with an LLTI and the number of people in the general population with an LLTI.

Different specifications of variables were modelled; they were included as either proportions or counts to determine which modelling techniques and variable definitions lead to more or less bias, if any.

Linear regression models and simple correlations are the analytical methods most often used in the analysis of such data. Correlation may be appropriate for hypothesis generation, but it can only tell the researcher about statistical associations and not causal associations as it cannot account for the relationships between other variables related to those being correlated, e.g. confounding. Linear regression can be used to account for some of these complexities and this will be returned to in the discussion of this chapter (Section 5.5).

The Poisson distribution is used when data are counts. This distribution is often used as a starting point for count data as it assumes that the model variation is the same as the model expectation (mean), i.e. in the context of this thesis, that would be the assumption that the probability of observing the next individual or event is constant in time or space for each geographical unit.[129] This may not be a realistic assumption to make in population–level data and the negative binomial or log–normal distributions would be preferred. The log normal distribution was used here because it provided the best fit to the observed LLTI data.

The negative binomial distribution could also be used for modelling count data as it is less

restrictive in that it allows for the data to be 'over–dispersed' this means that the variance is larger than the mean. This is often more appropriate in population–level data because populations are heterogeneous across geographical units. However, negative binomial regression was not present in the literature so only Poisson regression models were evaluated here. When Poisson regression is used to model data that follows a log–normal distribution the standard errors are biased which has repercussions for the confidence intervals which will be inaccurate.

5.3.9 Performance Measures

Performance measures were used here as suggested by Morris *et al.*[51] to assess the amount of inferential bias present (which calculates whether the estimated coefficient averages the true value, zero in this case), the empirical standard error (which measures the precision or efficiency of the coefficient), and the coverage (the probability that the confidence interval of the coefficient contains zero) of each model, illustrated using 'zip plots'.

5.3.10 Analysis of Observed Data

The same regression models were applied to the observed, original 1991 Census data and the results were compared to those simulated under the null hypothesis to add context and highlight any biases present.

5.3.11 Step–by–step guide to simulation of LLTI data

1. Three analytical techniques will be investigated: correlation, linear regression and Poisson regression as these are all found in the literature. The correlation

or regression coefficients and the 95% confidence intervals will be retained for analysis.

2. Observed data are taken from the 1991 Census and all relevant variable distributions are plotted against general distributions and summary statistics are used to aid the choice of distribution parameters. Raw components of the Townsend Index are used along with their opposite counterparts, e.g. unemployed and employed population. The assumed data generation process is illustrated in a DAG.

3. Assumptions are made that the components of the Townsend Index and their opposite counterparts add up to the total population count and total household count and the simulated variables are re–scaled to account for this.

4. Simulations assume that the null hypothesis is true and that neither the Townsend Index nor its components cause LLTI.

5. The performance measures to be estimated are: bias, coverage and standard errors of the estimates. 1,900 iterations of the simulation are performed, calculated based on equations presented by Morris et al.[51] to achieve acceptable Monte Carlo standard errors for the key performance measures.

6. The seed for the random number generator was set so that exact results can be replicated by others. The random number generator seed is set 9,499 values apart (equal to the number of electoral wards) to avoid dependence between datasets.[51,66]

7. A dataset is generated according to the assumptions covered above.

8. Statistical analyses are performed on this dataset and the parameter estimates obtained are retained (i.e. correlation and regression coefficients and related 95% confidence intervals).

9. The steps above are repeated 1,899 times with newly generated datasets in order to obtain an empirical distribution of parameter estimates.

10. The empirical distributions of the parameter estimates from the simulated datasets are analysed to estimate the bias from each analytical method.

11. The performance estimates are calculated and reported.

5.4 Results

5.4.1 Causal Diagram

The causal diagram (Figure 5.4.1) suggests that 'Population' should be adjusted for in the regression models as it is a confounding variable that acts upon both the exposure and the outcome. This is true for each of the components of the Townsend Index if used individually in regression models within the literature.

Composite Variables

The introduction of deterministic variables (e.g. composites such as the Townsend Index) in causal diagrams introduces additional conditional independencies that need to be accounted for in subsequent analyses.[24] Including composite variables which are wholly determined by raw parent variables needs to be carefully considered; it is straightforward to only include the composite variable (and exclude its components) as the usual rules of causal diagrams then apply. However, there may be philosophical issues and parametric constraints that arise by considering deterministic nodes in causal diagrams where it is truly believed they capture more information than the individual components themselves (this was considered in Section 3.10.3).

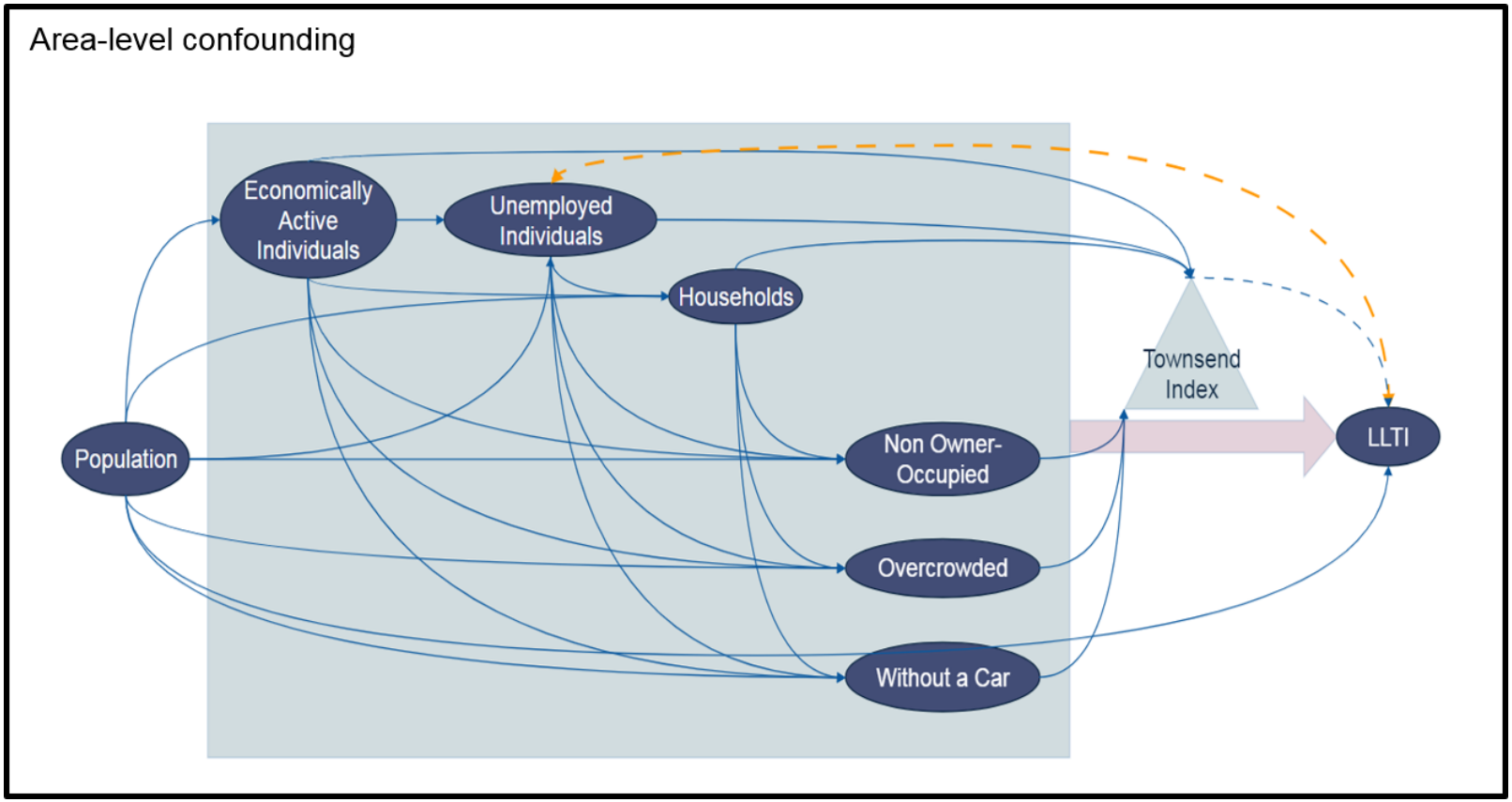

Figure 5.4.1: Causal associations between deprivation (as components of the Townsend Index) and Limiting Long–Term Illness (LLTI). The shaded grey box represents variables that generate the Townsend Index which is explored for its realtion to LLTI. Ordinarily, a composite variable (the Townsend Index, in this case) would not be included in a DAG as well as its components; for this reason, the Townsend Index is included in a triangle. The blue dotted arrow between the Townsend Index and LLTI indicates the relationship that is commonly investigated in the literature. The orange dotted double–headed arrow between unemployment and LLTI highlights that there is a complex, time–varying relationship between these two variables and is another relationship commonly investigated in the literature. The 'Area–level confounding' box surrounding the diagram indicates that there is an exogenous latent 'confounder' that causes area–level heterogeneity, however, this is not directly simulated but is added here for completeness in relation to future work.

Time of variable crystallisation

Further complexity is realised if it is acknowledged that the four components of the Townsend Index may not crystallise at the same time (although they are measured at the same time); for instance, unemployment may arise before, and thus help determine, living in non–owner–occupied accommodation, living in overcrowded accommodation, or living in a household without a car. When looking at unemployment as the exposure of interest, the minimally sufficient adjustment set (MSAS) thus comprises the economically active population and total population only. For this reason an investigation of unemployment is conducted separately for comparison to analyses encountered in the literature.

Time of variable crystallisation is also an additional complexity when there are other variables that would need to be adjusted to get more accurate model estimates. This needs to be considered more carefully when data are compressed into a latent variable such as the Townsend Index as this could mean that these variables go unnoticed, affecting model estimates or they cannot be adjusted for appropriately.

5.4.2 Mutual adjustment fallacies

Green and Popham[50] expand on previous literature[49] regarding the mutual adjustment fallacy (introduced in Section 2.6) using the example of research into the effects of Socioeconomic Position (SEP) on health. This can be applied here to the related concept of the Townsend Index. As a brief reminder, 'Mutual adjustment fallacies' refer to when all coefficients in a model are assumed to have an equivalent interpretation.

Using the DAG drawn in Figure 5.4.1, when unemployment is the exposure of interest the economically active population was adjusted for as it is a common cause of both the unemployed population and LLTI. When households without a car is the exposure of interest, the number of households, the unemployed population and the economically active population were adjusted for. Adjustments were also made for the number of

households, the unemployed population and the economically active population when the number of overcrowded households and the number of households that were non–owner occupied were the exposures of interest.

The information in the DAG informed a series of regression models: Poisson and linear to analyse the outcome, LLTI (as a count and as a standardised illness ratio) with respect to the Townsend Index and its components, as both proportions and as counts adjusting for population size along with models adjusting for the MSAS. As the outcome was simulated under the null hypothesis, a record of the number of times the coefficient of interest from each model deviates from zero (i.e. the type 1 error rate) was recorded, if the method of analysis is unbiased it is expected that this will occur on only 5% of occasions.

Along with the regression models, the outcome, LLTI (again, as a count and as a standardised illness ratio), was correlated with the Townsend Index and its components, as both proportions and counts but no adjustments (e.g. for the confounder, population, or other variables) could be made.

5.4.3 Correlation

Analyses of the 1,900 synthetic populations generated under the null hypothesis correlating the Townsend Index with LLTI produced a median correlation that was biased towards a positive relationship between LLTI and the Townsend Index (95% CI: 0.10, 0.18; Figure 5.4.3). When LLTI was standardised by the number of the population expected to experience an LLTI (SIR) the median correlation coefficient was zero (95% CI: -0.2, 0.02).

When the components of the Townsend Index were correlated with the count of the population experiencing a LLTI, under the null, a positive relationship was suggested. However, when the SIR was correlated with the components of the Townsend Index there was no bias present (Figure 5.4.3).

The ‘zip plots’ (Figures 5.4.4 and 5.4.5) show these biases clearly. In particular, Figure 5.4.5 shows an expected Type 1 error rate for the correlations undertaken using the SIR (approximately 5%). Figure 5.4.4 shows a very high Type 1 error rate when the correlations were undertaken using a count of the population experiencing a LLTI (in some cases the Type 1 error rate was 100%) with all confidence intervals biased towards a positive relationship between the Townsend Index, and its components, with LLTI.

For readers unfamiliar with ‘zip plots’ a larger plot is included here (Figure 5.4.2) in order to explain the concept. Each horizontal bar on the plot shows the 95% confidence interval of the estimate from correlating the number of the population with an LLTI with the number of unemployed people in the population. Bars shown in blue are confidence intervals that cover the true value (i.e. the null or zero in this case), whereas the red bars do not cover the true value. The black horizontal line shows at which point there would be a split in coverers (blue bars) and non–converers (red bars) if the results were not biased (i.e. at the 1,805 iteration of the simulation; the 95^{th} centile of the 1,900 iterations).

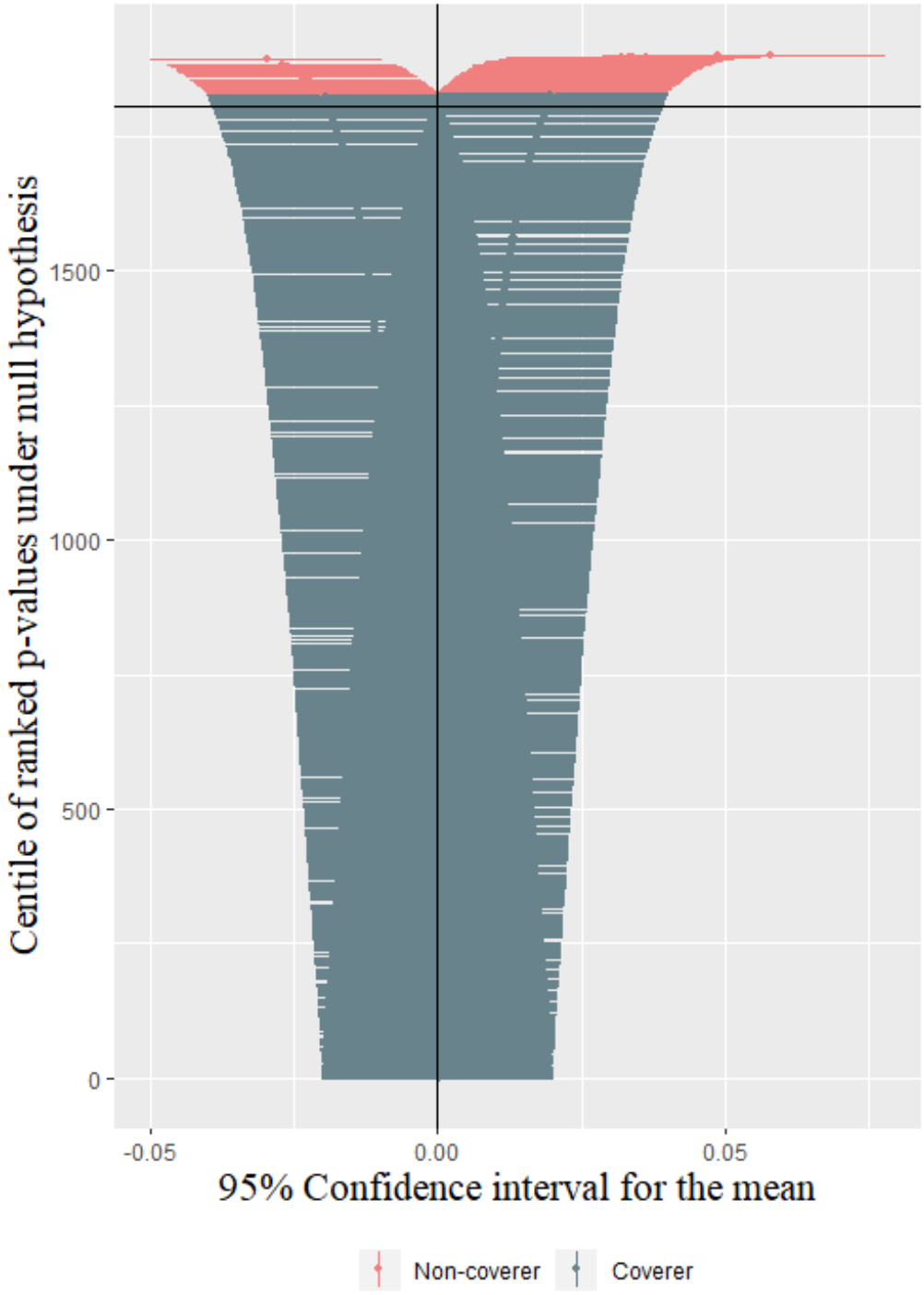

Figure 5.4.2: Enlarged 'Zip plot' for explaining the concept of these plots. Taken from the result of correlating the number of the population with an LLTI with the unemployed population. The plot shows the direction of bias of each correlation for the 1,900 simulations. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

5.4.4 Linear regression

Analyses of the 1,900 synthetic populations generated under the null hypothesis using linear regression, with the Townsend Index as the exposure variable, produced a median coefficient which was biased towards a positive relationship between the Townsend Index and LLTI (95% CI: 275, 452). When this regression model was adjusted for population size the relationship was no longer biased towards a positive relationship (95% CI: -90.8, 73.4). When the standardised illness ratio (SIR) was used as the outcome the 95% confidence intervals were small and contain zero.

When the components of the Townsend Index were used as the exposure variable in the models, the coefficients were biased when they were included as proportions (e.g. the proportion of the population that is unemployed), and there was no such bias when the numerator and denominator were included separately in the model with the number of people in the population treated as a confounding variable (e.g. the absolute number in the population that is unemployed and the number of people in the population), Figure 5.4.6.

The Type 1 error rates of the coefficients were substantially greater than the expected 5% using linear regression when the outcome was not standardised (Figure 5.4.7). When the outcome was standardised the Type 1 error rate was approximately the expected 5% for both count and proportion exposures (Figure 5.4.8).

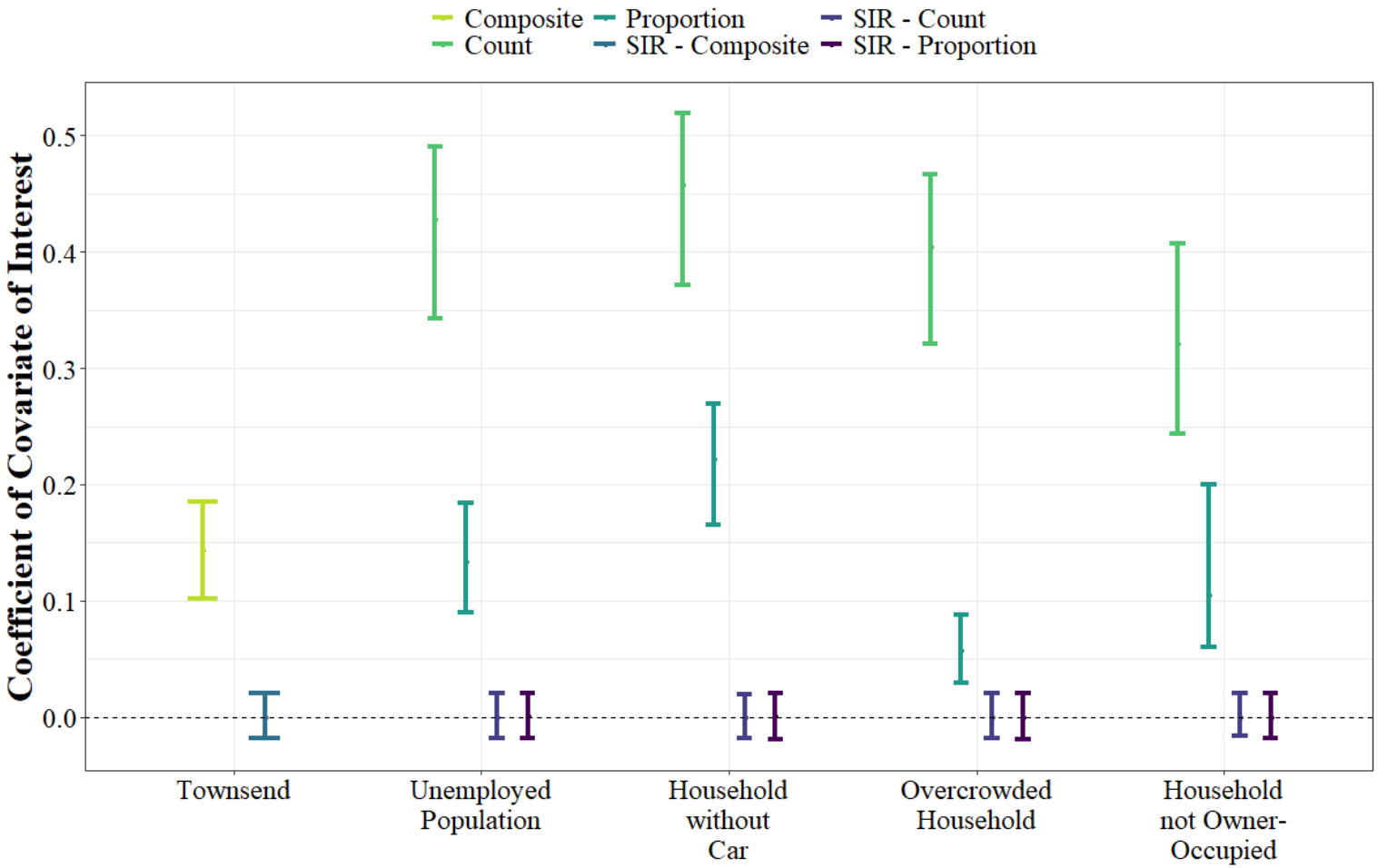

Figure 5.4.3: 95% confidence intervals calculated over 1,900 simulations for the median coefficient approximated using correlation

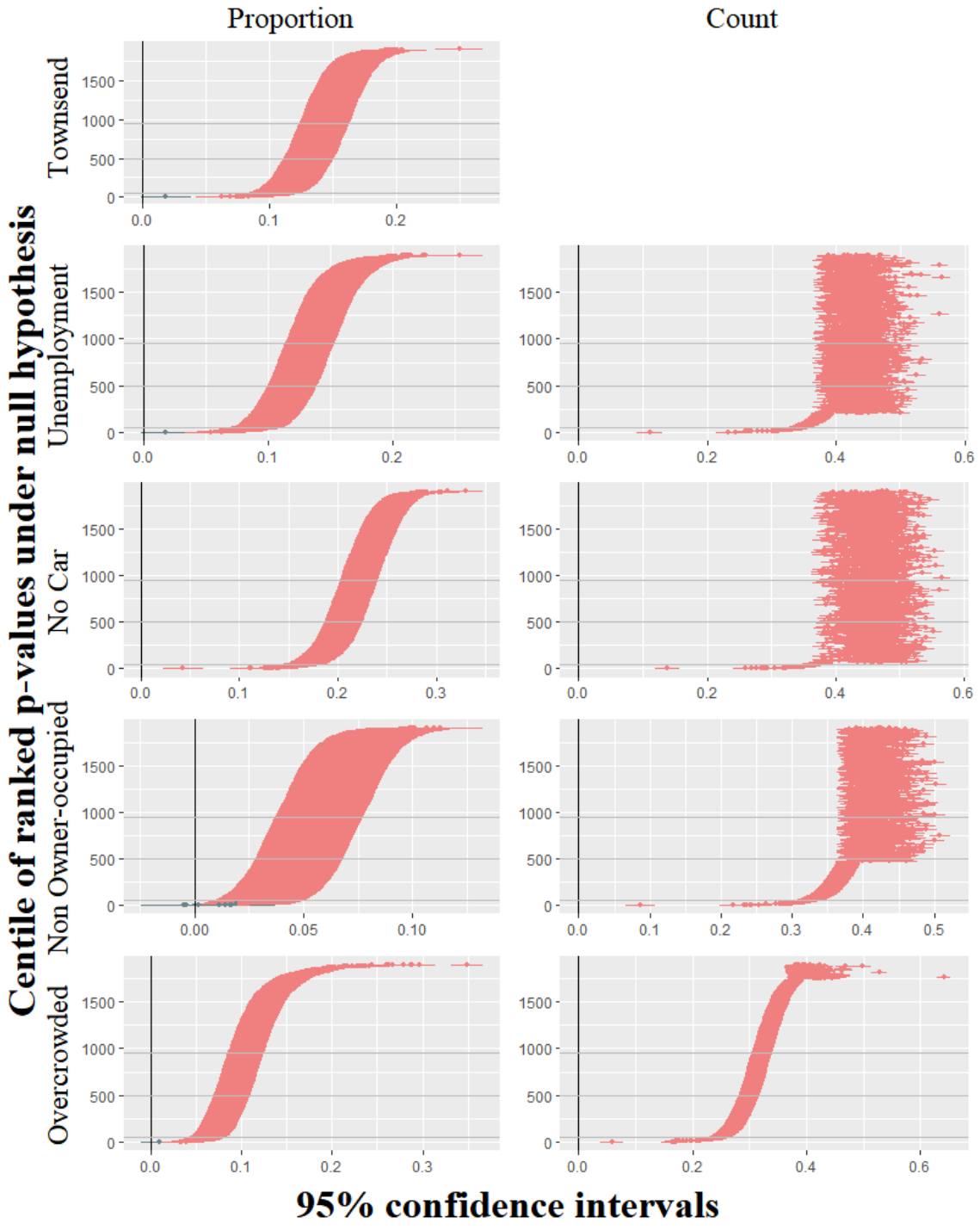

Figure 5.4.4: 'Zip plot' showing the direction of bias of each correlation for the 1,900 simulations when the outcome is not standardised. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

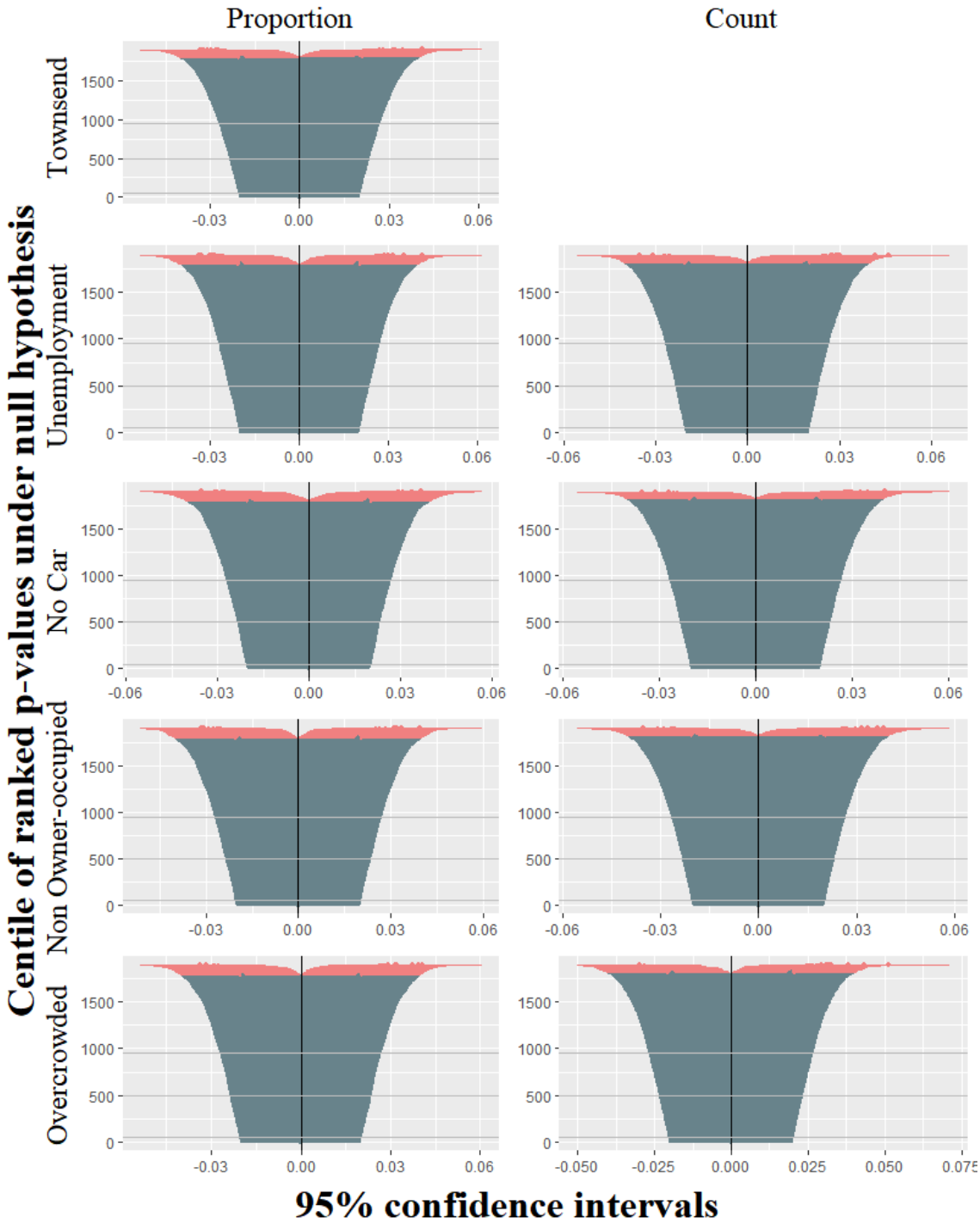

Figure 5.4.5: 'Zip plot' showing the direction of bias of each correlation for the 1,900 simulations when the outcome is standardised. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

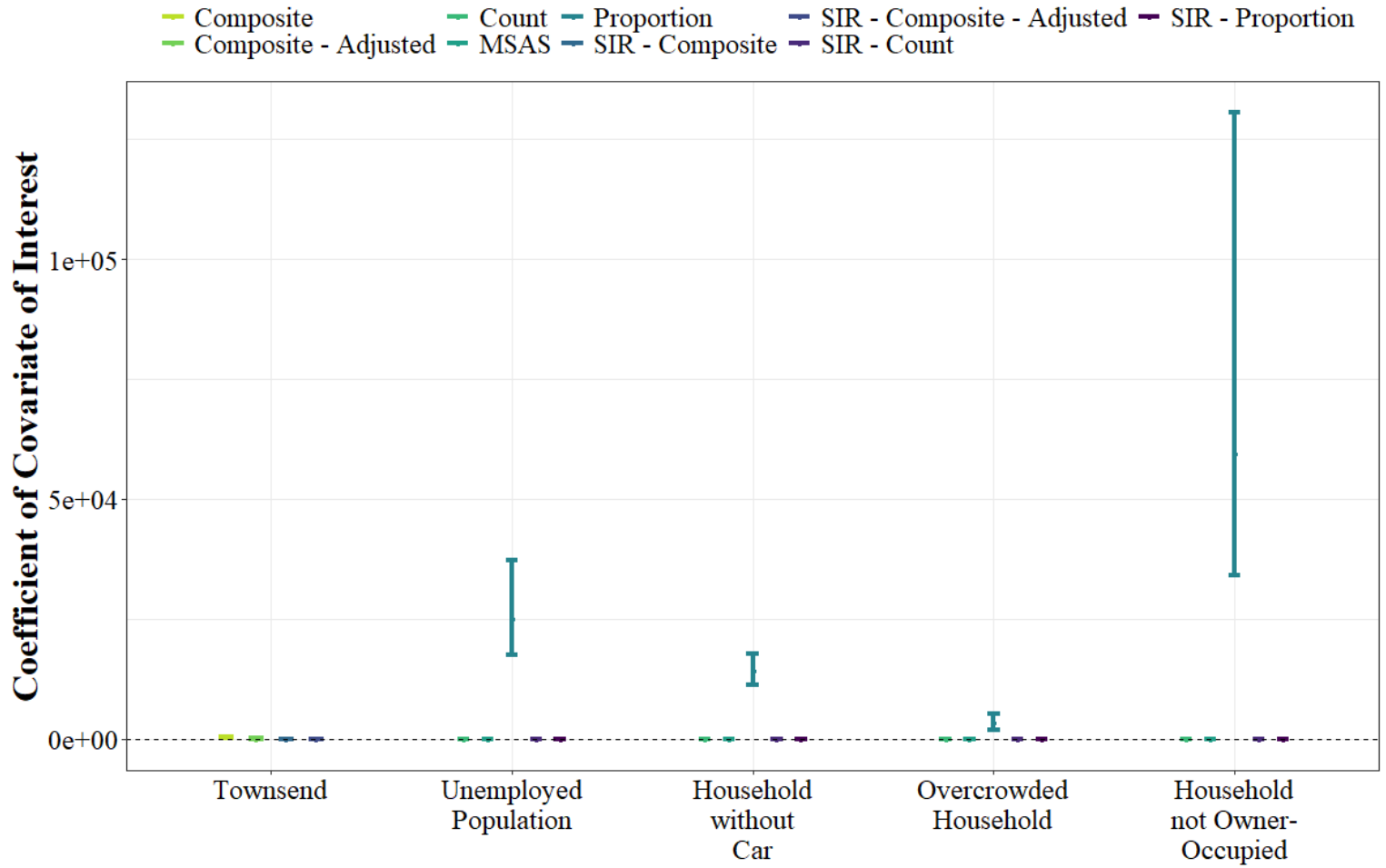

Figure 5.4.6: 95% confidence intervals calculated over 1,900 simulations for the median coefficient approximated using linear regression

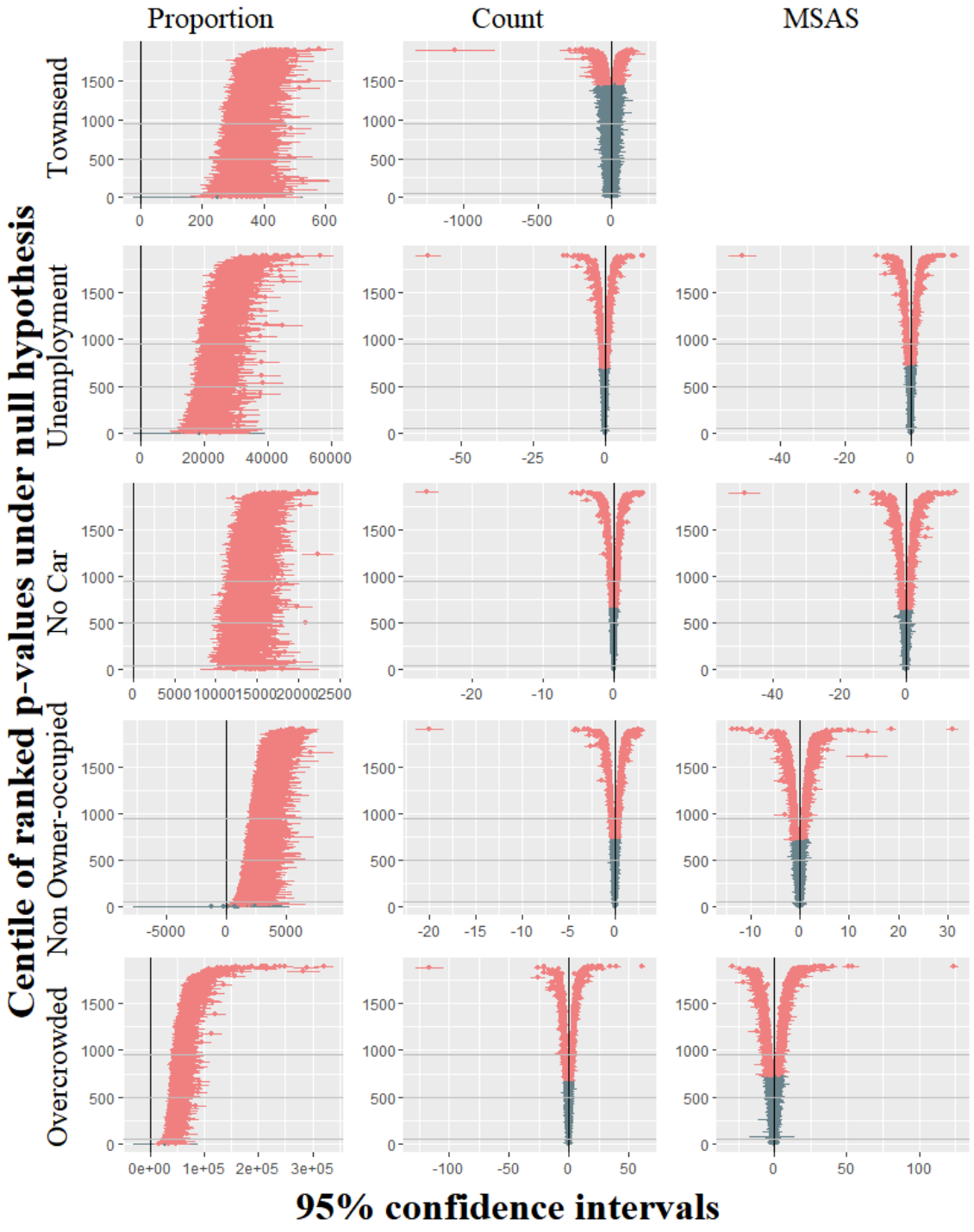

Figure 5.4.7: 'Zip plot' showing the direction of bias of each linear regression model for the 1,900 simulations when the outcome is not standardised. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

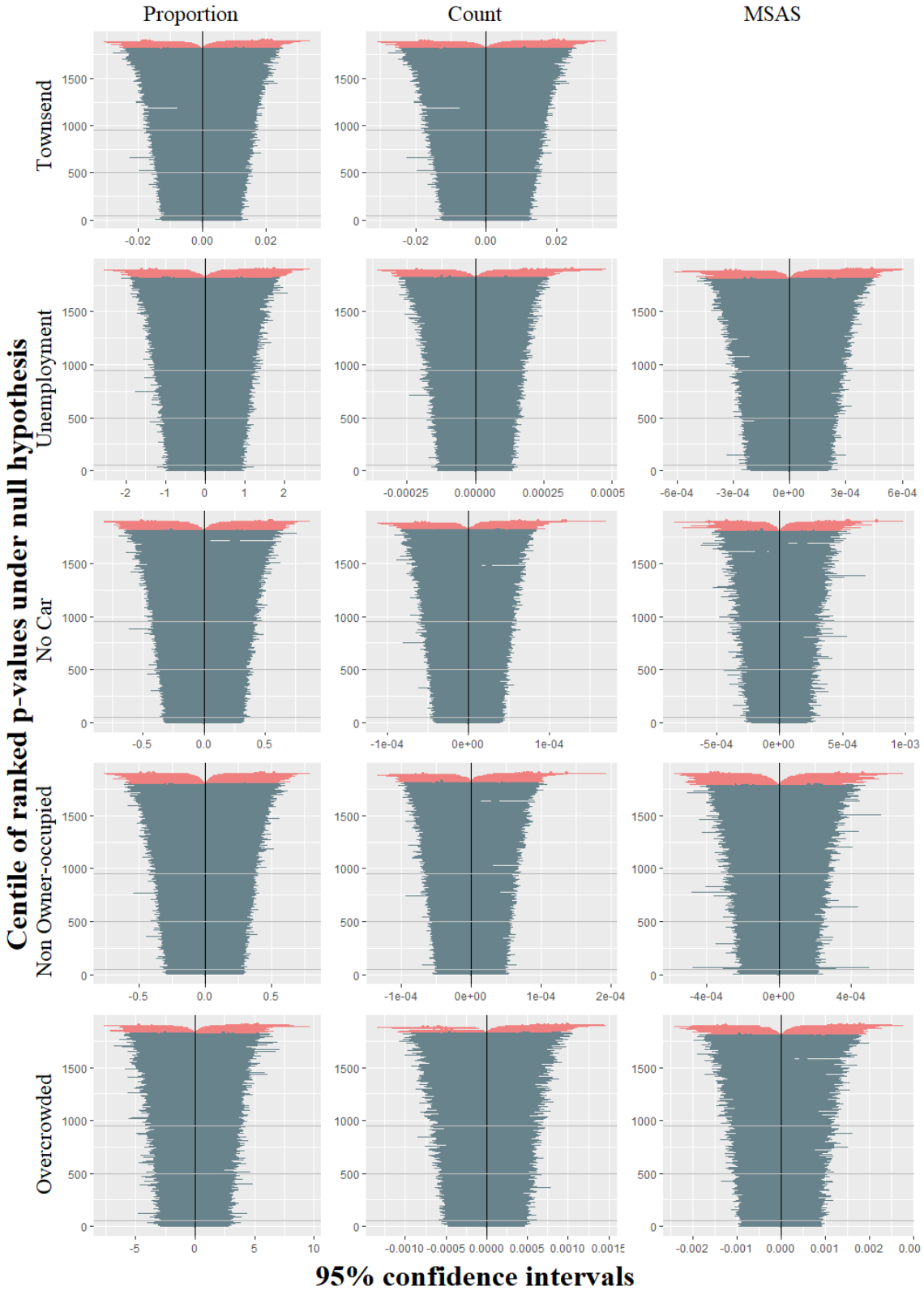

Figure 5.4.8: 'Zip plot' showing the direction of bias of each linear regression model for the 1,900 simulations when the outcome is standardised. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

5.4.5 Poisson Regression

Analyses of the 1,900 synthetic populations generated under the null hypothesis using Poisson regression with the Townsend Index as the exposure variable produced a median coefficient which was biased towards a positive relationship between the Townsend Index and LLTI (95% CI: 1.07 to 1.10). When this regression model was adjusted for population size the relationship was no longer biased towards a positive relationship (95% CI: 0.99 to 1.01).

When the components of the Townsend Index were used as the exposure variables in the models, the coefficients were biased when they were included as proportions (e.g. the proportion of the population that is unemployed), and there was no such bias when the numerator and denominator were included separately in the model with the number of people in the population treated as a confounding variable (e.g. the absolute number of unemployed people in the population and population size), Figure 5.4.9.

The Type 1 error rate of the coefficients were almost 100% for all the models using Poisson regression (Figure 5.4.10). These correspond with 95% confidence intervals that contained zero and the high Type 1 error rates were a result of using Poisson regression on a log normal distributed outcome; as mentioned above, the standard errors are larger when a Poisson regression model is used on an outcome that follows the log normal distribution. Simulations with a Poisson distributed outcome were conducted for empirical verification.

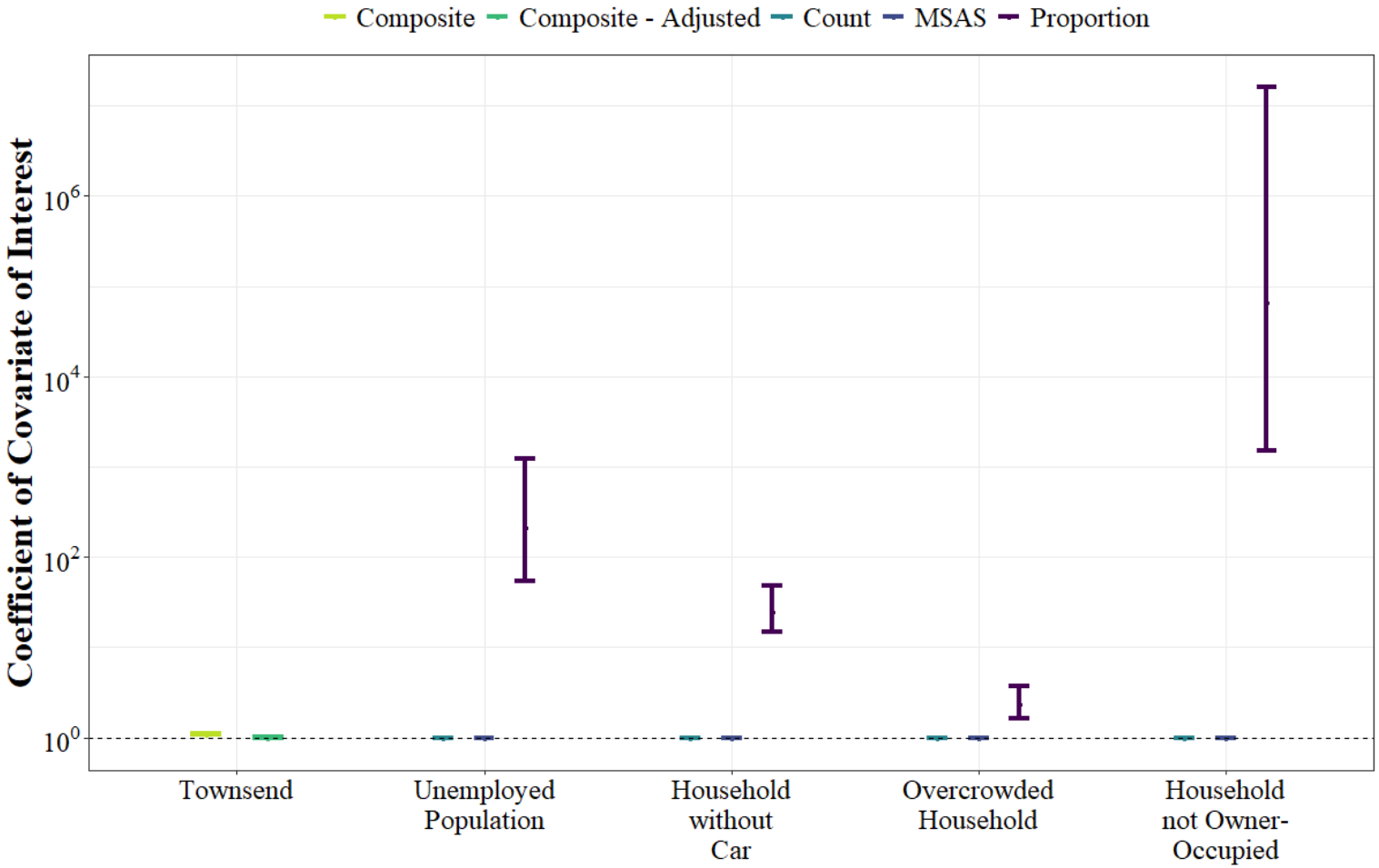

Figure 5.4.9: 95% confidence intervals calculated over 1,900 simulations for the median coefficient approximated using Poisson regression

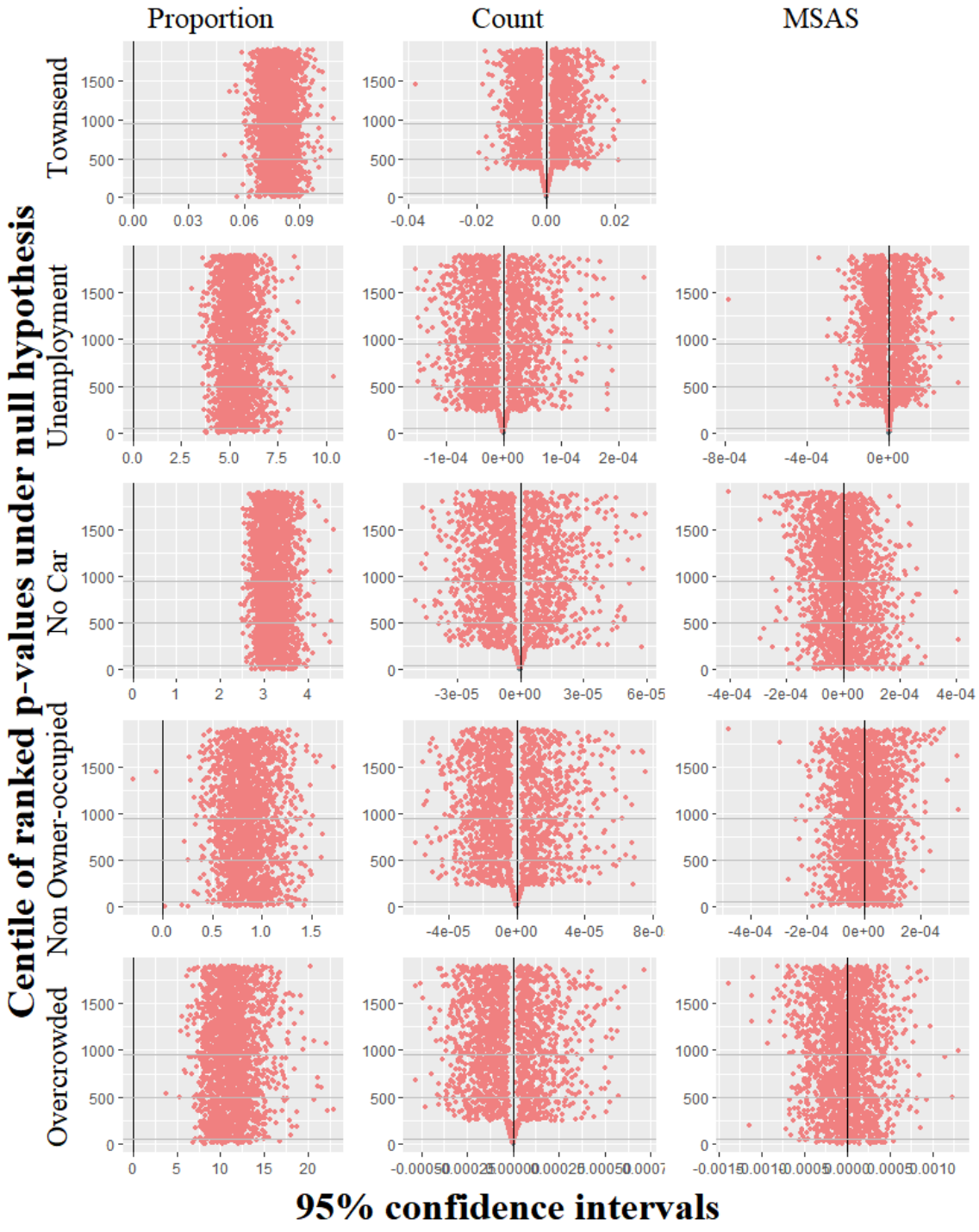

Figure 5.4.10: 'Zip plot' showing the direction of bias of each Poisson regression model for the 1,900 simulations. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

Performance measures

The 'zip plots' for each method (Figures 5.4.4–5.4.5 & 5.4.7–5.4.8 & 5.4.10) show bias and large standard errors when using proportions rather than count variables in the regression models. These Figures also show that the highest level of coverage (i.e. when the confidence intervals 'cover' the true value, zero) is achieved when count variables were used in the case of the regression models and that when proportions were used in the regression models a positive relationship between the exposure and outcome always results. When the outcome was standardised (i.e. when the SIR is used) no bias was present for either the exposure expressed as a count or as a percentage.

The reason that it is acceptable to divide through by a common denominator to 'control' for population size was explained in Section 4.2.10. This is because the exposures and outcome are proportional to each other and if they were to be plotted against each other they would form a straight line through the origin. There are two main issues regarding dividing through by a common denominator for unbiased analyses which are both related to this only being the case when the null hypothesis is true. Firstly, as soon as any other variable is involved in determining the outcome, the outcome would not be proportional to the confounder and therefore it would not be sufficient to divide through by the confounder to adjust for it. This would especially be an issue in the case of unobserved confounding due to other variables as it would not be possible to ascertain the level of 'spurious' correlation present. Secondly, there is likely to be unobserved confounding at the area–level present which represents area–level heterogeneity. These issues will be discussed further in Section 5.5 and in Chapter 7.

Results of observed data analysis

Analysis of the observed data using correlation suggested positive relationships between the Townsend Index and each of its components (as counts and as proportions) with

limiting long–term illness (as a count or SIR; Tables 5.4 & 5.5, respectively). However, it is recognised from the simulations that when the count of the population with an LLTI is used (rather than the SIR) the correlations will be biased in a positive direction because the confounding variable (population size) was not adjusted for. This likely accounts for some of the difference in correlation between using the counts of LLTI and the SIR. Incidentally, the 95% confidence intervals of the correlations of the count of LLTI and the SIR with the percentage of non–owner occupied households are the same.

Table 5.4: 95% Confidence intervals from correlating the count of LLTI in the population with the Townsend Index and its components.

Count of LLTI correlated with:	95% Confidence Interval
Townsend	(0.51, 0.54)
Unemployed population	(0.86, 0.87)
Percentage unemployed population	(0.52, 0.55)
Households without a car	(0.91, 0.92)
Percentage households without a car	(0.61, 0.64)
Overcrowded households	(0.64, 0.66)
Percentage overcrowded households	(0.35, 0.39)
Non–owner occupied households	(0.81, 0.83)
Percentage non–owner occupied households	(0.28, 0.31)

Table 5.5: 95% Confidence intervals from correlating the standardised rate of LLTI in the population with the Townsend Index and its components.

SIR correlated with:	95% Confidence Interval
Townsend	(0.42, 0.45)
Unemployed population	(0.24, 0.27)
Percentage unemployed population	(0.47, 0.50)
Households without a car	(0.31, 0.35)
Percentage households without a car	(0.56, 0.58)
Overcrowded households	(0.13, 0.17)
Percentage overcrowded households	(0.15, 0.19)
Non–owner occupied households	(0.26, 0.29)
Percentage non–owner occupied households	(0.28, 0.31)

Using linear regression, large regression coefficients were found when using the count of the population with LLTI as the outcome and the Townsend Index (both adjusted and unadjusted for the population confounder) as was suggested would be the case from the simulations. These were attenuated (though still positive) when the counts of the Townsend components were used as the exposure and the population size was adjusted for in the model. The simulations suggest that the least biased method of analysis for these data is using the SIR as the outcome and counts of the Townsend Index components as exposures adjusting for population size (Tables 5.6 & 5.7, respectively). Results using this approach on the observed data suggest very small effect sizes in the positive direction, but for households without a car, overcrowded households and non–owner occupied households these became negative when the MSAS was considered.

Table 5.6: 95% Confidence intervals of linear regression coefficients on observed LLTI data using the count of LLTI in the population as the outcome and the Townsend Index and its components in count and percentage format as covariates

Model Exposure:	95% Confidence Interval
Townsend	(87.5, 90.4)
Townsend (adjusted for population size)	(33.4, 34.6)
Unemployed population (adjusted for population size)	(0.71, 0.74)
Percentage unemployed population	(8830, 8580)
Households without a car (adjusted for population size)	(0.36, 0.36)
Percentage households without a car	(4280, 4390)
Overcrowded households (adjusted for population size)	(0.72, 0.81)
Percentage overcrowded households	$(1.5 \times 10^4, 1.6 \times 10^4)$
Non–owner occupied households (adjusted for population size)	(0.22, 0.23)
Percentage non–owner occupied households	$(2.0 \times 10^3, 2.1 \times 10^3)$
Unemployed population (adjusted for MSAS)	(0.81, 0.83)
Households without a car (adjusted for MSAS)	(0.10, 0.14)
Overcrowded households (adjusted for MSAS)	(-0.83, -0.77)
Non–owner occupied households (adjusted for MSAS)	(-0.07, -0.06)

Table 5.7: 95% Confidence intervals of linear regression coefficients on observed LLTI data using the standardised rate of LLTI as the outcome and the Townsend Index and its components in count and percentage format as covariates

Model Exposure:	95% Confidence Interval
Townsend	(0.04, 0.04)
Townsend (adjusted for population size)	(0.04, 0.04)
Unemployed population (adjusted for population size)	$(6.1 \times 10^{-4}, 6.5 \times 10^{-4})$
Percentage unemployed population	(4.25, 4.40)
Households without a car (adjusted for population size)	$(3.3 \times 10^{-4}, 3.5 \times 10^{-4})$
Percentage households without a car	(2.20, 2.26)
Overcrowded households (adjusted for population size)	$(5.9 \times 10^{-4}, 7.2 \times 10^{-4})$
Percentage overcrowded households	(3.83, 4.29)
Non–owner occupied households (adjusted for population size)	$(1.8 \times 10^{-4}, 2.0 \times 10^{-4})$
Percentage non–owner occupied households	(1.12, 1.18)
Unemployed population (adjusted for MSAS)	$(7.1 \times 10^{-4}, 7.5 \times 10^{-4})$
Households without a car (adjusted for MSAS)	$(-1.1 \times 10^{-4}, -4.2 \times 10^{-5})$
Overcrowded households (adjusted for MSAS)	$(-5.5 \times 10^{-4}, -4.0 \times 10^{-4})$
Non–owner occupied households (adjusted for MSAS)	$(-1.5 \times 10^{-4}, -1.3 \times 10^{-4})$

Using Poisson regression, a small positive relationship between the Townsend Index and LLTI was suggested for models both adjusted and not adjusted for population size (Table 5.8). In all models using the Townsend Index components as counts adjusted for population size no relationship was evident. In models using proportions as the exposure the risk ratios were very high as was the case for the simulations which suggests that this is a result of inappropriate model and variable specifications.

Table 5.8: 95% Confidence intervals of Poisson regression coefficients on observed LLTI data using the count of LLTI in the population as the outcome and the Townsend Index and its components in count and percentage format as covariates

Model Exposure:	95% Confidence Interval
Townsend	(1.10, 1.10)
Townsend (adjusted for population size)	(1.03, 1.03)
Unemployed population (adjusted for population size)	(1.00, 1.00)
Percentage unemployed population	(9060, 9200)
Households without a car (adjusted for population size)	(1.00, 1.00)
Percentage households without a car	(337, 340)
Overcrowded households (adjusted for population size)	(1.00, 1.00)
Percentage overcrowded households	$(1.4 \times 10^6, 1.5 \times 10^6)$
Non–owner occupied households (adjusted for population size)	(1.00, 1.00)
Percentage non–owner occupied households	(15.7, 15.8)
Unemployed population (adjusted for MSAS)	(1.00, 1.00)
Households without a car (adjusted for MSAS)	(1.00, 1.00)
Overcrowded households (adjusted for MSAS)	(1.00, 1.00)
Non–owner occupied households (adjusted for MSAS)	(1.00, 1.00)

Further investigation of the relationship between the denominator (population size) and the Townsend Index and each of its components suggest that these variables are not exactly proportional to each other in the observed data (which is to be expected; Figure 5.4.11). This means that dividing through by population size is inappropriate for controlling for the population confounder and that linear regression with SIR as the outcome and controlling for population size as a covariate is the most appropriate method to avoid bias due to mathematical coupling in this case.

It is unclear why models using the components of the Townsend Index as exposures were more biased than those with the Townsend Index as exposure and this warrants further investigation.

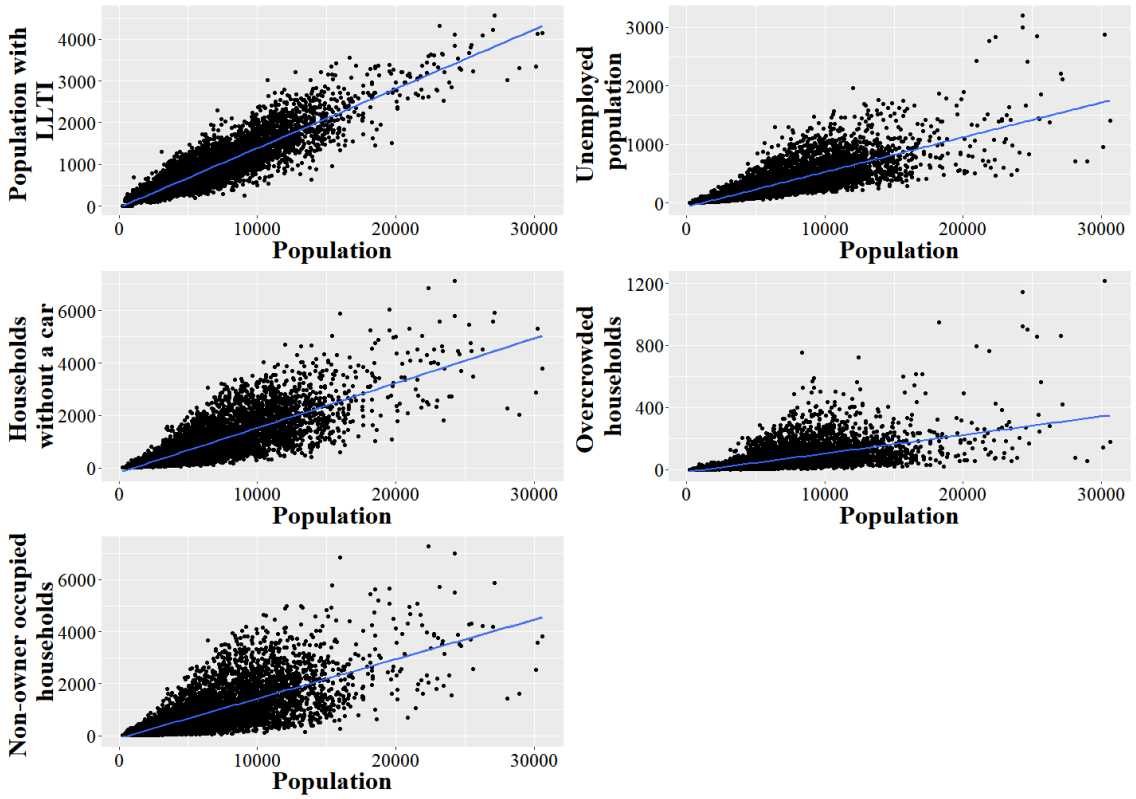

Figure 5.4.11: Scatter plots showing the Townsend Index and its components plotted against the population size.

5.5 Discussion

This research has used analyses of simulated and observed data to investigate the potential ramifications of mathematical coupling on observational health geography data. The specific example used explores the proposed relationship between area–level deprivation and limiting long–term illness.

The methods investigated here were used to build on the foundations of the theory that

deprivation is a cause of limiting long–term illness (LLTI). The study presented here does not disprove a link between the two, but suggests that revisiting investigations into these two concepts ought to consider the risks of mathematical coupling. Indeed, the literature review discovered journal articles that covered multi–level models and this may prove to be a more appropriate avenue for future research which could account for area–level heterogeneity presuming that the individual–level data are available.

It has previously been shown mathematically that the bias due to mathematical coupling is a result of incorrect model specification; equations commonly used in the literature are not equivalent to the models using components.[101] The accurate regression equation for using ratios is achieved by dividing through the equation for the component model by the variable that should be controlled for (population size in this case). This means that both $\frac{X}{N}$ and $\frac{1}{N}$ must be regressed on $\frac{Y}{N}$ to get the correct result. Although this means that there is a way of dealing with ratio variables that avoids mathematical coupling, this method has been shown to perform poorly when the measure of population size is unreliable.[104] The interpretation of the ratio model becomes much more complex when there are other variables involved which will always be the case when the research question involves observational health geography data. This issue, combined with the fact that the correct specification of the ratio model has not been adopted since the publication of this work in 1986,[104] provides a good argument for the acceptance of component variables for observational health geography research, unless it is thought that composite variables truly represent more information than their components. In terms of causal inference, Firebaugh and Gibbs[101] also pose the question: "how does one know that a ratio truly has causal effects and should be used, or whether the ratio is no more than a statistical invention whose use reifies a nonentity?" (p.715); further reason for the use of only the component model.

DAGs are not widely used to depict area–level variables, but if researchers are interested in area–to–area variation in area–level outcomes it should be plausible to construct a

DAG to answer area–level questions.[29] DAGs have been developed which incorporate both individual–and area–level variables, showing it is possible to construct DAGs based purely at the area level.[22] It may be necessary, however, to carefully consider the impacts of exogenous confounding that leads to heterogeneity that is a hallmark of health geography, with greatest effects where the denominator, N, varies considerably (e.g. electoral wards as opposed to SOAs) which links to the MAUP.

When thinking of the research question addressed here (what is the causal effect of deprivation on limiting long term illness?), the consideration is what would happen if an intervention was made to change the deprivation level of an area, which is not a well–defined problem. The causal inference principle of consistency requires that however the level of deprivation is changed, whether by reducing unemployment, reducing overcrowding, increasing owner–occupation or increasing household car ownership, this would always have the same effect on the outcome. With this in mind, it may not be possible to estimate the causal effect of deprivation explicitly in a single concept; the question would have to be reframed around potentially multiple aspects of deprivation that could be intervened upon, such as unemployment.[82] Further research is needed to elicit how composite concepts can be used in a causal framework and how they can be represented in DAGs along with the natural hierarchy or whether they must be broken down into their component parts for meaningful analysis, as has been done here.

Where the models suggest bias arises in the coefficients of proxies for deprivation, the bias always tends towards a positive relationship with LLTI due to the mathematical dependency that using common denominators creates.[90] If studies consistently report a stronger association between exposure and outcome (there may be a true non–null association in an observed dataset), there will always be a ‘spurious’ element to this when ratios are used. Interventions to reduce the outcome (LLTI in this case) may be focussed on a domain that will not achieve the reduction in incidence that would be expected given the results of the algebraically coupled analyses. The burden on health services

could be effectively reduced and priority given to more appropriate areas in health policy, according to unbiased analyses of the observed data that suggest no relationship between the individual components of the Townsend Index and LLTI.

Often analyses will break down the data into age–sex categories, however, this is not done here as the results would be similarly distorted in the simulated examples and this only serves to complicate the exposure–outcome relationship making it more difficult to illustrate the problem of mathematical coupling. However, the results of the observed data analysis may be more robust if the data were age–standardised or additionally adjusted for age, sex and their interaction.

Composite measures are often used in analyses of geographical health data, particularly when a variable of interest (e.g. deprivation) can only be represented by proxies. However, the algebraic dependencies that are introduced in the construction of ratio index variables for analysis are not usually considered, resulting in inadvisable recommendations for latent variable proxies, e.g. the proportion of unemployment as a proxy for deprivation. This may explain why some studies have noted higher correlations between the individual components of the composite variables with the outcome than the composite variables themselves.[113] Additionally, the complex relationship between unemployment and LLTI suggests that a time–varying unemployment variable may be more appropriate in a causal diagram representing this problem; however, this is not attempted here as the goal is to explore the methods present in the existing literature.

5.6 Conclusion

A move away from the use of proportions or percentages in area–level health geography research may be difficult to implement as they are often the core element of quantitative analyses in this discipline; it can be difficult to comprehend the meaning of a variable except in ratio index form.[96] Analysts must however endeavour to do this for their analyses

that incorporate correlation or regression to avoid the inferential biases that result from mathematical coupling. This problem is wide ranging; it not only affects the area of health geography, but many others too. It has been demonstrated how the adoption of a DAG framework to consider such composite relationships can aid understanding of these issues along with understanding of the historical solutions.

5.7 In the context of the thesis

This chapter has illustrated how simulation and causal inference influenced thinking can help researchers understand biases that may be present in the historical approaches to data analyses. This is only one such example, but it highlights how causal thinking and subsequent analysis could be brought into the field of health geography to avoid inferential biases so that research can more accurately be used to inform health policy. Integrating causal inference methods with health geography will be discussed more thoroughly in Chaper 7.

The next chapter uses the framework outlined in Chapter 3 to consider selection on the outcome, the ‘most dangerous equation’ and the modifiable areal and temporal unit problems.

Chapter 6

Population Mixing and Childhood Leukaemia

6.1 Introduction

Following on from the previous two chapters which have focused on mathematical coupling and composite variables, this chapter looks at a long–standing research hypothesis in health geography –the so–called 'population mixing hypothesis' (see Section 6.2.1). Initially, it may appear that the bias introduced when trying to answer this research question in a certain way is a result of mathematical coupling, however, this is not the case and using simulation and causal inference knowledge built on from Chapters 2 and 3 the true cause of this bias is shown. This work has been published in *Epidemiology*[130] and a follow–up letter[131] has been responded to.[132] A recent citation of this work[133] recognises the conclusion that the methods used to analyse such data can influence the results and that in the case of clustering, region–wide analytical strategies should be used.

The code for the simulations is available in Appendix D and on GitHub, however, code

related to the observed data analyses are not publicly available due to the data being confidential. A video abstract is also available at `https://journals.lww.com/epidem/pages/videogallery.aspx?videoId=87`.

The following chapter combines the two concepts the 'most dangerous equation' and the modifiable areal and temporal unit problems (introduced in Sections 3.10.5 and 3.10.4, respectively) along with selection on the outcome (Section 2.1.11) to show the pitfalls of common analyses looking at the hypothesised relationship between population mixing and childhood leukaemia.

6.2 Background to the 'population mixing hypothesis'

If the example introduced in Section 3.10.4 is extended to the scenario where disease incidence between areas is being compared, those areas with small populations are more likely to appear as spatial clusters of high incidence by chance alone. Focusing on these supposed clusters is therefore a poor basis on which to generate or test causal hypotheses.[87] Nonetheless, such clusters are hard to ignore,[134] and can generate substantial pressure for plausible explanations. This may explain the considerable public and political interest given to the high incidence of childhood leukaemia in Seascale (Cumbria, UK) during 1963–1983, and the relative lack of attention to the absence of such cases during 1991–2006.[135,136]

The challenges of examining clusters between areas with different population sizes are likely to have influenced the development and testing of the 'population mixing hypothesis'. The idea emerged from analyses purporting to show an association between 'population mixing' and childhood leukaemia, interpreted as evidence for the involvement of infectious agents.

6.2.1 What is the 'population mixing hypothesis'?

The hypothesis proposes that: the immune systems of children resident in more isolated and/or less densely populated communities are more likely to have been exposed to a less diverse range of infectious agents than residents in less isolated and/or more densely populated communities. These children are therefore believed to be more likely to develop leukaemia if they are exposed to novel infections from inward–migrants.[137]

6.2.2 The assumptions of the 'population mixing hypothesis'

This hypothesis is both persuasive and enduring.[138,139] Indeed, one recommendation of the Seventeenth Report of the Committee of the Medical Aspects of Radiation in the Environment (COMARE), published in 2016, was that "prospective studies be made of the incidence of childhood leukaemia in rural areas in which any large–scale construction projects (both non–nuclear and nuclear) are to be carried out"[140] (p.151). This recommendation was made on the basis of evidence of the "influence of rural population mixing upon the risk of childhood leukaemia"[140] (p.151). However, the hypothesis relies on several untested assumptions, and involves a lack of clarity around how many of its key concepts should be defined, measured and analysed.[141] One assumption is that isolated communities, and those with lower population densities, are less likely to experience the frequency and or intensity of contact required to sustain infections. Another is that communities with lower rates of 'inward-migration' are less frequently exposed to exogenous infections. While these assumptions reflect established tenets of infectious disease epidemiology they require levels of isolation, population dispersion, and (im)mobility that remain unspecified, and may be neither plausible nor applicable where the hypothesis has been examined. There also remains extensive disagreement regarding the roles that the immune system and early exposures to infection play in the aetiology of childhood leukaemia.[142–144]

6.2.3 Measures used to capture population mixing

These mechanistic uncertainties are compounded by a lack of consensus concerning: what constitutes an isolated or less dense population; criteria used to distinguish between migrants and residents; and how these concepts are operationalised as measures of population mixing. Researchers exploring the association between population mixing and childhood leukaemia have therefore used a range of different measures as proxies for population mixing including: differences and/or changes in population size/density; the proportion and/or diversity of inward–migrants; and versions of the Shannon Diversity index.[141]

The variety of measures confirms a lack of conceptual precision/consensus, and reflects the practical constraints imposed by: the distribution and migration patterns of populations within regions where suitable data exist; the collation/organisation of data on these parameters; and challenges differentiating leukaemia cases amongst residents and inward–migrants. Good quality, area–level data on population size/density, migration, and childhood leukaemia incidence are only available for high–/middle–income countries where large regions are usually subdivided into small areas along political/administrative rather than demographic lines. These small areas display substantial variation in geo–spatial features (size, shape, and distance apart), and in the size/distribution of their constituent populations. Consequently, along with the socio–demographic detail of data available from sources such as a decennial census, the geographical specification of these areas constrains what measures of isolation, density, migration, and mixing can be generated. Such sub–division also creates larger–than–expected chance variations in incidence amongst smaller populations simply due to chance[84,87] as explained in Sections 3.10.4 and 3.10.5.

6.2.4 Analytical strategies for investigating the 'population mixing hypothesis'

Different researchers have used different analytical strategies, generating contradictory results.[137,145] Some of the earliest studies followed the identification of an apparent cluster of leukaemia cases in a single area, and sought to verify whether this constituted a *bona fide* cluster (i.e. a higher number of cases than expected given the national/regional incidence proportion –the number of new cases per population at risk during a particular period of time).[146] Unfortunately, such studies provide little evidence of whether the elevated incidence is associated with any characteristics of the area concerned. In these studies, it is often unclear how/when the specific measures for population mixing were selected (i.e. before or after the areas of study were selected for their apparent excess of cases). Substantial methodological variations make it challenging to identify commonalities in analytical approach for closer examination. However, many such studies focused specifically on areas displaying childhood leukaemia clusters/higher incidence of childhood leukaemia. Indeed, where other studies adopted a non–selective region–wide analytical strategy –examining associations between area–based measures of population mixing and leukaemia incidence across the whole region, or in a random sample of areas –these tend to generate contradictory findings to those adopting non–random, selective, or focused analytical strategies.[137,146–152]

Much work is needed to strengthen the concepts, measures, and datasets used to test the population mixing hypothesis. There is a pressing need to establish why different analytical strategies generate such contradictory findings. This chapter uses simulation and analysis of observed data to examine the two principal analytical strategies used by previous ecological studies and explores the relationship between commonly used measures of population mixing and childhood leukaemia. Such measures typically draw on the concept of population mixing as proposed by the first study to use this term (Section 6.2.1),[137] which was subsequently defined as an "increase in population density

produced by a marked influx into a rural area" (p.1163; where 'rural' was considered a less densely populated area).[153] On this basis, the two most common measures of population mixing used by previous studies were chosen: population density and inward–migration. Population density provides a measure of the number of individuals capable of spreading a putative leukaemia–promoting infectious agent,[141,154] expressed as the population per unit area. Inward–migration provides a measure of the relative number of new arrivals capable of bringing such agents with them, expressed as the proportion of migrants within the population. Both measures were calculated using existing data dis–aggregated by administrative areas, and were used to undertake each of the analytic strategies as follows:

(i) Selective sub–region analysis. Areas with contrasting values of population density, inward–migration, and/or childhood leukaemia incidence (i.e. representing areas of specific interest as potentially 'highly exposed' vs. reference areas) were non–randomly selected for direct comparison; and (ii) Region–wide analysis. The relationship between population density, inward–migration, and childhood leukaemia incidence is examined using standard regression techniques across all small areas within a larger region, or a random sample of areas.

6.3 Methods

The selective sub–region analysis and region–wide analysis were applied to observed data from the Yorkshire and Humber region of the UK using data from a previous study of the population mixing hypothesis.[145] Data were also simulated in which the number of childhood leukaemia cases was determined solely by population size and not by population density or inward–migration (i.e. the null hypothesis).

6.3.1 Observed Data

Population density and inward–migration were calculated for each of the 532 census wards in the Yorkshire and Humber region using 1991 Census data on: total population; ward area (km^2); number of inward–migrants (those with a different address one year prior to the Census); and number of 0–14 year olds (the population deemed to be 'at–risk'). Population density prior to inward–migration was calculated. Inward–migration in relation to each ward's pre–migration population was calculated, such that the proportion of inward–migration could exceed one (i.e. for wards where inward–migration resulted in a doubling, or more, of the population).

Leukaemia cases (for 0–14 year olds) were identified from the Yorkshire Specialist Register of Cancer in Children and Young People, diagnosed within the Yorkshire Regional Health Authority between 1988 and 1993 (the closest date to the 1991 Census for which data were available).[145] These were mapped to census wards, to permit estimation of childhood leukaemia incidence rates (Figure 6.3.1). Situating these analyses around the 1991 Census facilitated comparison with previously published studies, most using data before subsequent declines in incidence reported elsewhere.[135,136]

6.3.2 Childhood leukaemia data

Childhood leukaemia data were available for five–year periods over the 25 years, 1978-2003. Data from the period 1988–1993 was deemed most appropriate because the inward–migration figures related to those who had moved into each area during the year prior to the census date in 1991. All available childhood leukaemia data are used in this section for illustrative purposes.

Section 3.10.4 illustrates how the random nature of the Poisson distribution creates areas that appear to have a significantly high number of cases along with areas that have no cases

at all. Here, this illustration is developed using the childhood leukaemia data (in contrast to the previous more synthetic example where areas were of equal, regular size and shape; Section 3.10.4) by showing the ratio of observed to expected cases of childhood leukaemia on maps of the electoral wards of Yorkshire and the Humber, UK (Figures 6.3.1–6.3.6).

From these maps, it can be seen that for each 5–year period, different electoral wards stand out as having more cases than would be expected given the size of the population of 0–14 year olds. These tend to be larger, rural areas with smaller populations. It can be seen, therefore, that choosing areas for analysis on such a basis would result in an incorrect idea of the true relationship between these attributes and the occurrence of childhood leukaemia.

Focussing on ratios in geographical health research can be misleading when visualising information using a choropleth map where the denominator in the ratio of interest is not geographic area (nor a variable that is not correlated with geographic area) because the map gives undue attention to larger areas in this case[155] and it is human nature to be drawn to these areas.[134] Unfortunately, a demand can be placed on researchers to look into apparent clusters by the general public or government organisations as this problem is not well understood. A better way in which to represent these data visually would be to use a map that shows the electoral wards as uniform shapes whose sizes are standardised by the population of the electoral ward that they represent thereby removing visual bias.

An analogous concept to the Modifiable Areal *Unit* Problem is that of the Modifiable *Temporal* Unit Problem which refers to the aggregation of temporal scales and the effects of this on subsequent statistical analyses.[86] In this context, the occurrence of childhood leukaemia happens over continuous time, but time can be discretised into temporal units in many different ways. Figures 6.3.1–6.3.6 highlight this as they show cases of childhood leukaemia over 5 separate 5–year periods and the aggregation of these into a 25 year period. From these, it can be seen that if selection on the time period for analysis was made *post–hoc* this could lead to analysis being conducted on data that is most likely to produce

'positive' results. As the data is aggregated over the total period, no areas show 'extreme' incidence of childhood leukaemia and most areas have a ratio of observed versus expected cases in–around 1 and that these occur mostly in the most populated electoral wards in the larger area.

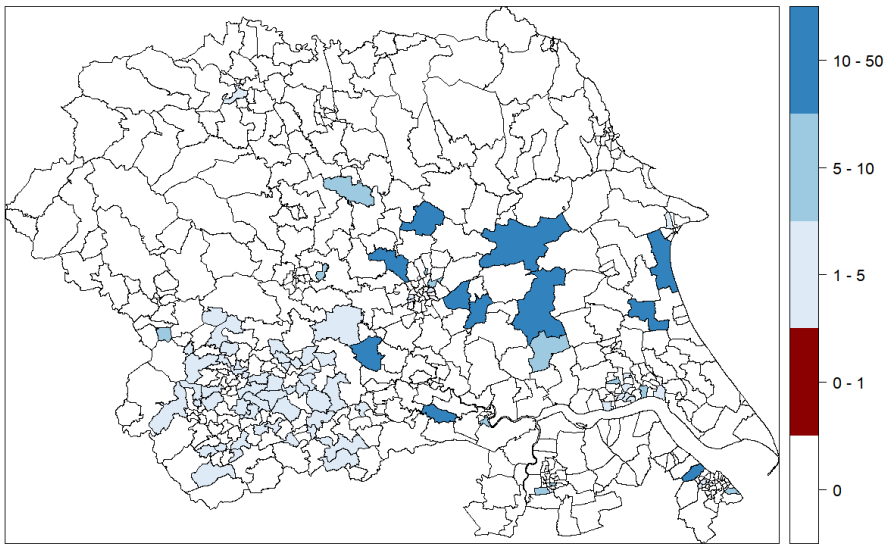

Figure 6.3.1: Ratio of observed to expected (based on average national incidence) cases of childhood leukaemia in Yorkshire and Humber (UK), 1978–1982, by ward

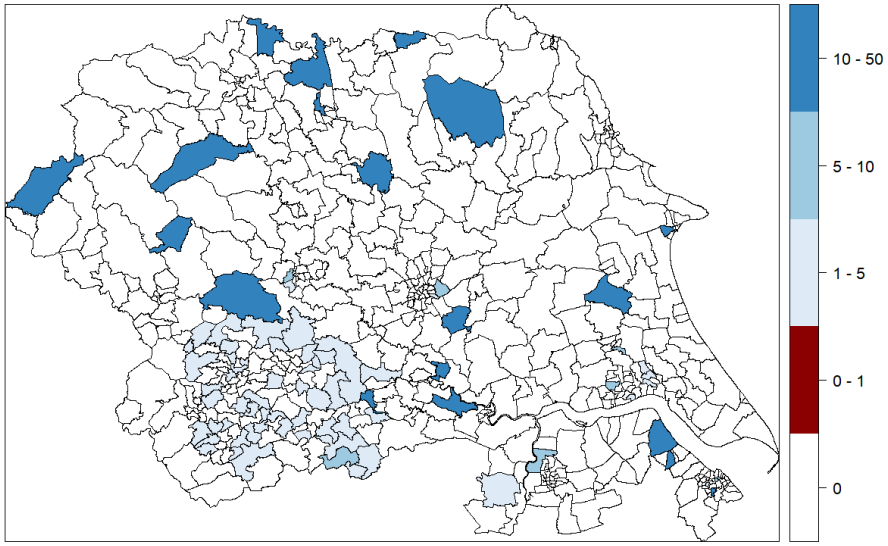

Figure 6.3.2: Ratio of observed to expected (based on average national incidence) cases of childhood leukaemia in Yorkshire and Humber (UK), 1983–1987, by ward

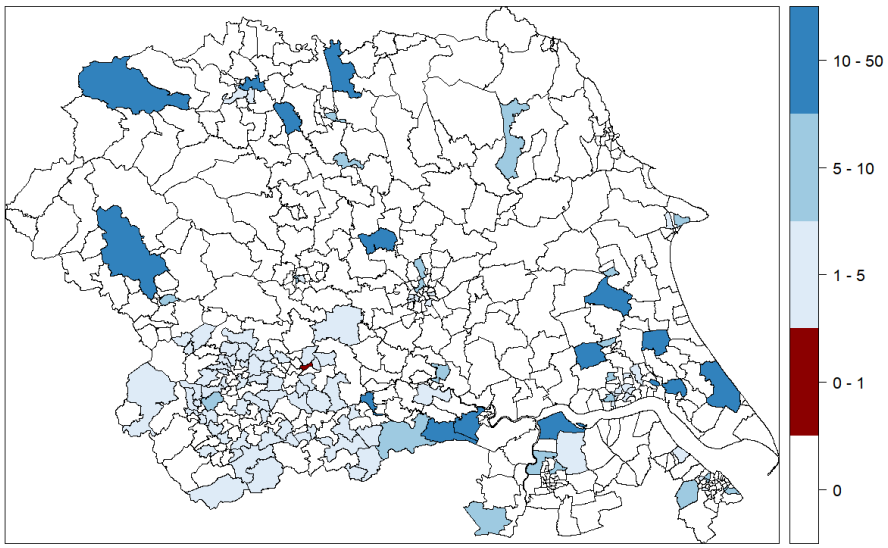

Figure 6.3.3: Ratio of observed to expected (based on average national incidence) cases of childhood leukaemia in Yorkshire and Humber (UK), 1988–1993, by ward

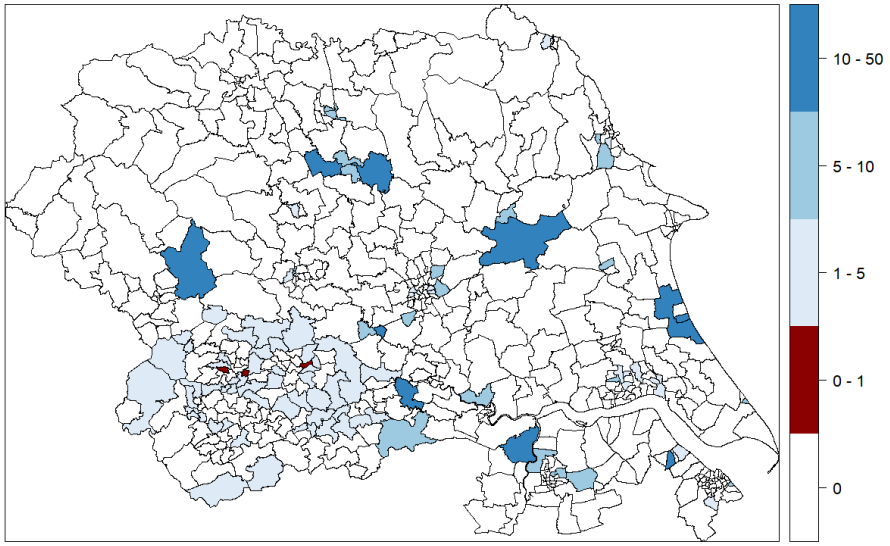

Figure 6.3.4: Ratio of observed to expected (based on average national incidence) cases of childhood leukaemia in Yorkshire and Humber (UK), 1994–1998, by ward

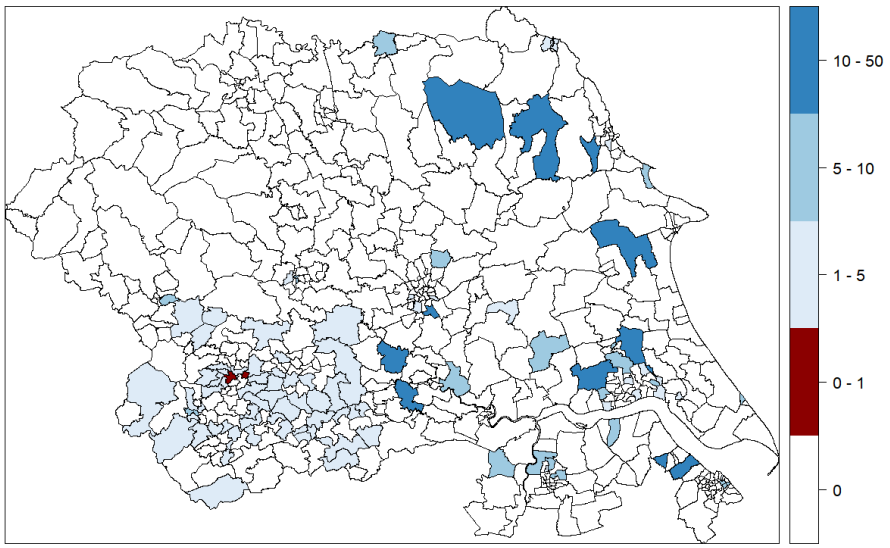

Figure 6.3.5: Ratio of observed to expected (based on average national incidence) cases of childhood leukaemia in Yorkshire and Humber (UK), 1999–2003, by ward

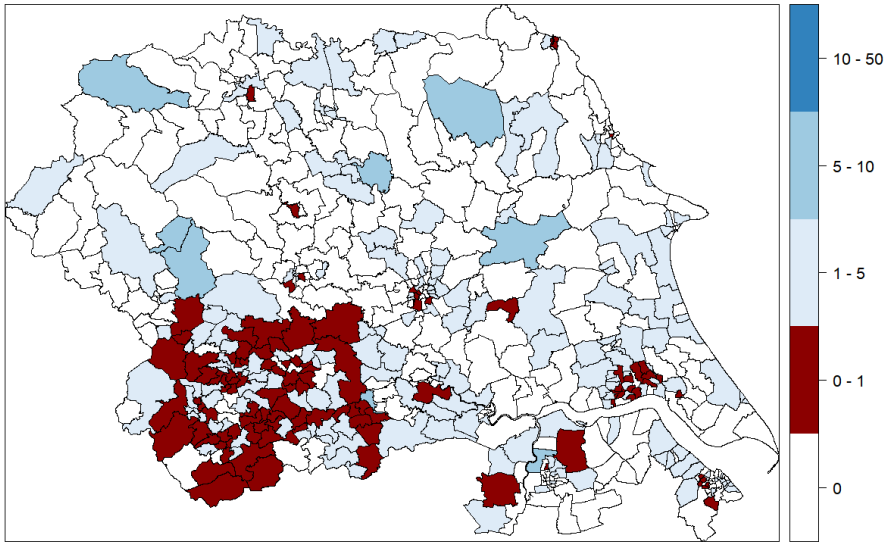

Figure 6.3.6: Ratio of observed to expected (based on average national incidence) cases of childhood leukaemia in Yorkshire and Humber (UK), 1978–2003, by ward

6.3.3 Selection on the outcome

Selecting on the outcome in this way can be explained from a causal inference perspective as introduced in Section 2.1.11. Selection on the outcome is essentially selecting on a collider which then introduces relationships between the exposure and the outcome which are artefacts of the data analysis approach.

6.3.4 Simulated data

Multivariate ward–level data on population density and inward–migration were simulated such that their distributions (Table 6.1) and correlation structure (Table 6.2) approximated those in the observed data using an algorithm to simulate multivariate non–normal data using an iterative algorithm,[75] as was the procedure for simulating data in Chapter 5. First, summary statistics were calculated for all of the variables that were to be simulated (Table

6.3). The distributions from which variables were simulated were chosen by plotting the observed variable distributions and fitting general distributions to them (Figures 6.3.7–6.3.10) before determining the particular distribution parameters (Figures 6.3.11).

Table 6.1: Distributions from which simulated variables were drawn.

Variable	**Distribution**
Total population	Negative Binomial (mean = 6500, theta = 2.0)
0–14 population	Negative Binomial (mean = 1300, theta = 1.6)
Area	Negative Binomial (mean = 26, theta = 0.7)
Inward–migration	Negative Binomial (mean = 500, theta = 1.8)

Table 6.2: Correlation matrix of the observed data to be emulated in the simulated datasets.

	Total Pop	0–14 Pop	Area	Inward–mig
Total Pop	1.00	0.97	−0.29	0.92
0–14 Pop	0.97	1.00	−0.30	0.89
Area	−0.29	−0.30	1.00	−0.32
Inward–mig	0.92	0.89	−0.32	1.00

Table 6.3: Summary information for each variable to be simulated. Total Pop = Total population of each electoral ward; 0–14 Pop = 0–14 year old population of each electoral ward; Area = Area (km^2) for each electoral ward; Inward–migrants = Total number of inward–migrants to each electoral ward.

	Total Pop	0–14 Pop	Area	Inward–migrants
Minimum	476	70	0.17	24
1st Quartile	2142	404	3.7	175
Median	4051	746	11.3	332
Mean	6455	1301	25.9	564
3rd Quartile	10 368	2025	34.6	936
Maximum	24 578	5883	216.8	3445

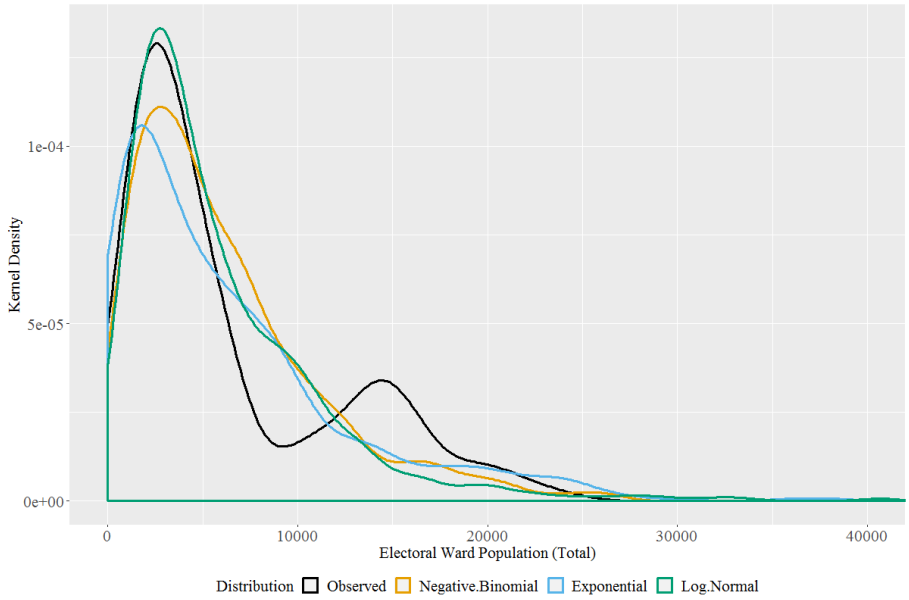

Figure 6.3.7: Distribution of observed variable for total population with fitted estimated distributions.

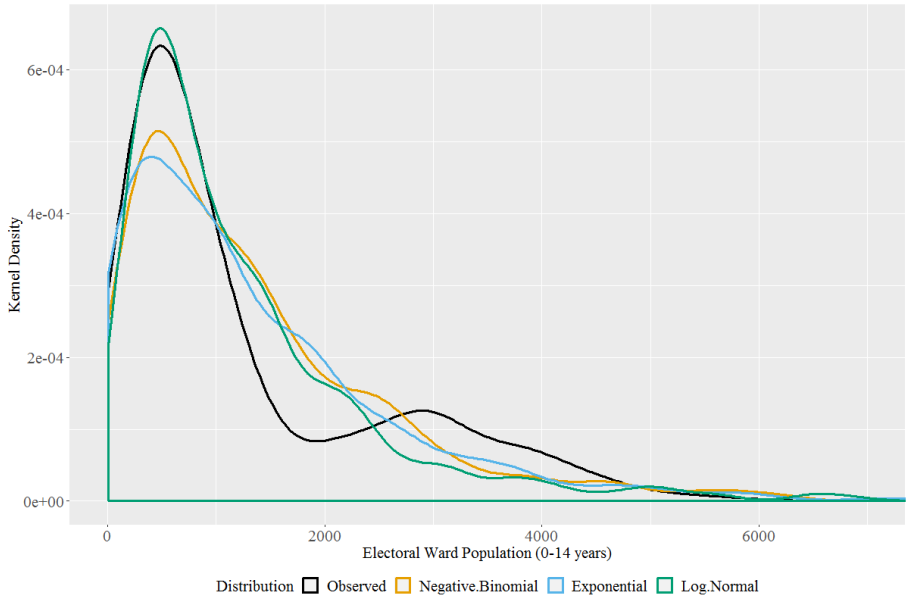

Figure 6.3.8: Distribution of observed variable for 0–14 year old population with fitted estimated distributions.

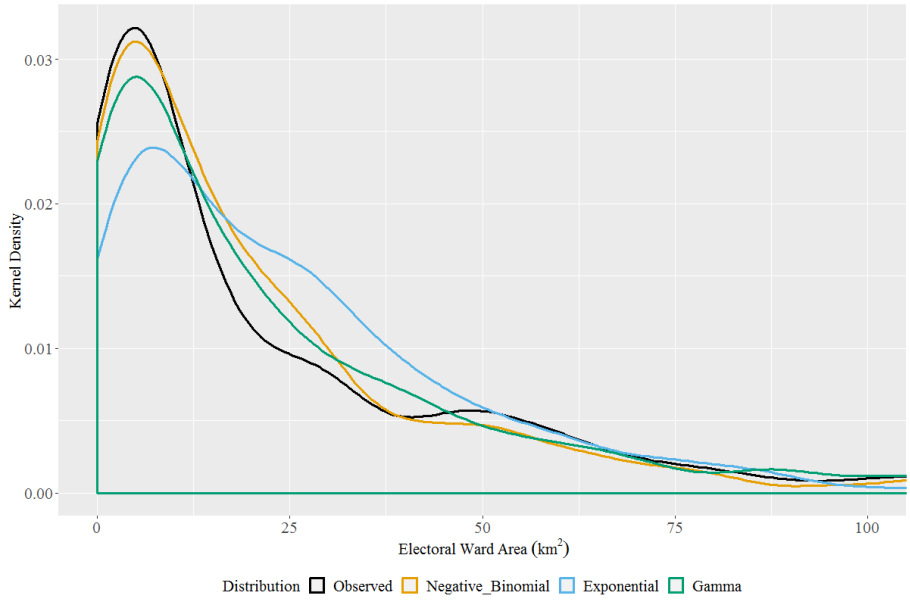

Figure 6.3.9: Distribution of observed variable for electoral ward area (km^2) with fitted negative binomial distributions varying the size and mean parameters of the distribution.

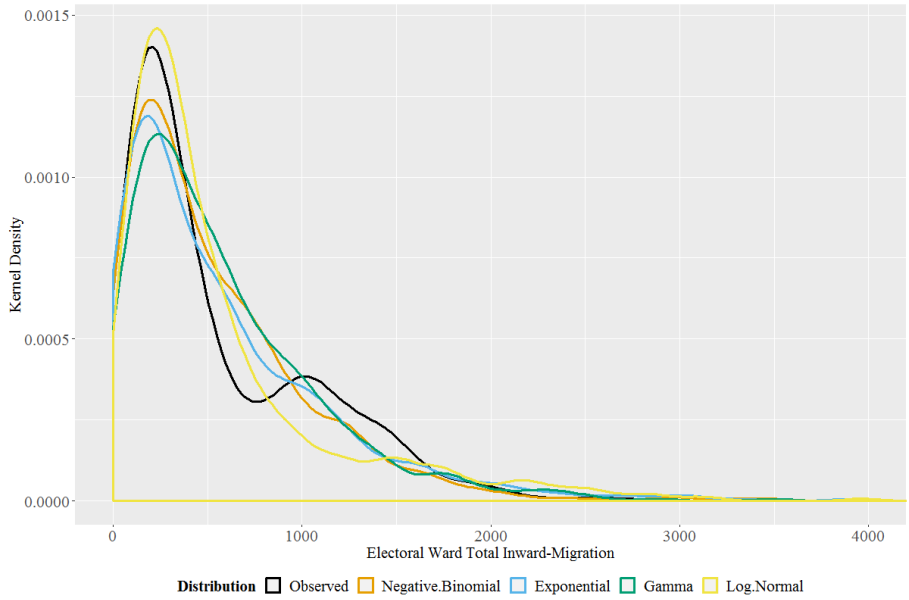

Figure 6.3.10: Distribution of observed variable for total count of inward–migrants with fitted estimated distributions.

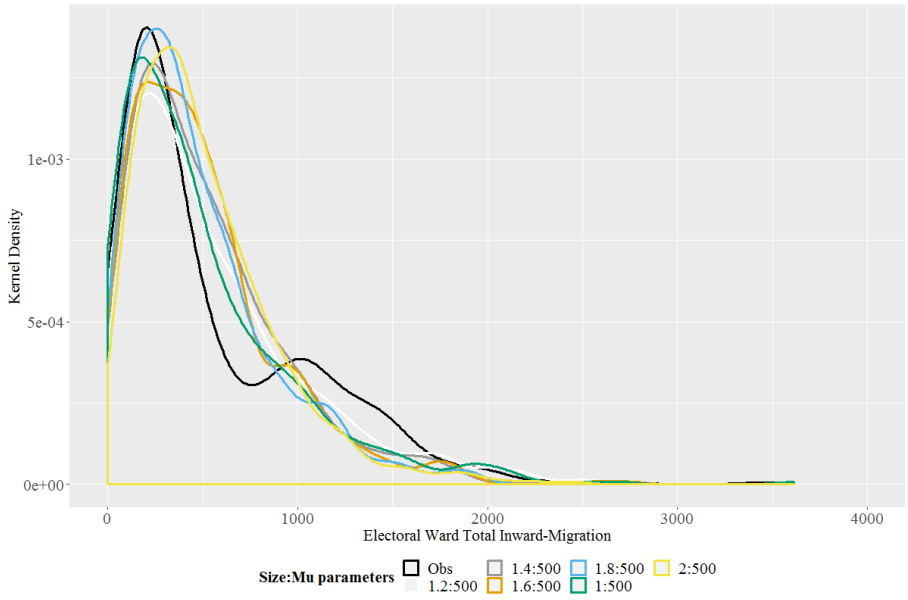

Figure 6.3.11: Distribution of observed variable for total count of inward–migrants with fitted negative binomial distributions varying the size and mean parameters of the distribution.

Simulated cases were based on the national childhood leukaemia incidence proportion over a comparable 5–year period;[156] a time interval chosen to emulate previous studies, and overcome key challenges with modelling rare events. Simulations of the outcome used the Poisson distribution (i.e. as evident in cases of childhood leukaemia in the observed data) under the null hypothesis that the number of cases of childhood leukaemia in each area is determined only by the number of 0–14 year olds. By approximating the observed population structure under the null assumption that the only driver of the number of cases of childhood leukaemia is population size, deviations from a null result in the analyses of simulated data must be due to selection or analytic errors. To ensure sufficient data were available to reduce the standard error of the simulation process,[51,66] and to more precisely learn the operating characteristics of the different estimation procedures, 10,000 simulated datasets were generated.

This chapter was the first in the thesis to be completed and it was submitted and

underwent peer review before the publication of the paper by Morris *et al.*[51] which provides accessible guidance on how to measure the performance of simulations and how to decide how many simulations are appropriate. 10,000 simulations were undertaken at the time as that was thought an appropriately large number and did not take too much computational time. However, subsequently, the performance of the simulations in this chapter have been assessed using the criteria of Morris *et al.*[51] and a step–by–step guide to the simulation has been written in the style of that developed in Section 3.12.

6.3.5 Step–by–step guide to the simulation

1. The two main methods for investigating a hypothesised relationship between population mixing and childhood leukaemia are undertaken and compared. These are:

   - selective sub–region analysis, and
   - region–wide analysis

   The parameter estimates obtained from the selective sub–region analysis, i.e. estimates obtained from the binomial exact test and their corresponding p–values are retained.
   The region–wide analysis uses Poisson regression models and the risk ratios and corresponding p–values are retained.

2. Observed data are taken from the 1991 Census and the Yorkshire Specialist Register of Cancer in Children and Young People. All relevant variable distributions are plotted against general distributions and summary statistics are used to aid the choice of distribution parameters for simulation. Each part of population density (i.e. population size and area size) and the proportion of inward–migration (i.e. the number of inward–migrants and population size at the beginning of the study

period) were generated separately as they are composite variables. The population at the beginning of the study period is calculated by subtracting the number of migrants over the study period from the population size at the time of the census. The assumed data generation process is illustrated in a DAG.

3. There are complex relationships between the variables of interest which are better suited to real–world informed simulation.

4. Simulations assume that the null hypothesis is true and that neither of the population mixing proxies (population density and the proportion of inward–migrants) cause childhood leukaemia and that the only determinant of childhood leukaemia incidence is the size of the 0–14 year old population.

5. List the performance measures to be estimated: bias, coverage and standard errors of the estimated. 10,000 iterations of the simulation are performed (chosen as this is a large number of iterations but does not require an excessive amount of computing power). This was later deemed an acceptable number of simulations using the equations presented by Morris *et al.*[51]

6. Set the seed for the random number generator so that the exact results can be replicated by others. The random number generator seed is set 532 values apart (equal to the number of electoral wards) to avoid dependence between datasets.[51,66]

7. Generate a dataset according to the assumptions covered above.

8. Perform statistical analyses on this dataset and retain the parameter estimates obtained (i.e. risk ratios from the Poisson regression and the statistic obtained from the binomial exact test, along with associated *p*–values).

9. Repeat the previous two steps 9,999 times with newly generated datasets in order to obtain an empirical distribution of the parameter estimates.

10. The empirical distributions of the parameter estimates from analysing the simulated datasets are analysed to estimates the bias from each analytical method.

11. The performance measures are calculated and reported.

6.3.6 'Selective sub–region' analytical strategy

To emulate the selective sub–region strategy, 16 wards were selected, the mean number of areas in those studies that used this approach in the literature,[137,146–152,157] based on extreme values of: low population density; high inward–migration; high childhood leukaemia incidence; or combinations of all three. 15 selection scenarios were examined (based on all combinations of these three selection variables) in order to account for the disparate methods found in the literature.

Scenarios 1–3 involved ranking wards according to low population density, high inward–migration or high incidence alone, then randomly selecting 16 of the highest ranked 50% of wards for analysis. Scenarios 4–9 involved ranking wards according to each possible pair of variables: ranking first on the initial variable and selecting the highest 50%, next ranking these on the second variable and selecting the highest 50%, then randomly selecting 16 wards for analysis. Finally, Scenarios 10–15 involved (1) ranking the wards according to every possible ordering of all three variables –ranking on the initial variable and selecting the highest 50%; (2) on the second variable, selecting the highest 50%; (3) on the third variable, again selecting the highest 50%, before (4) randomly selecting 16 wards for analysis. To match the number of random selections available from the 10,000 simulated datasets, random selection of the 16 wards were taken 10,000 times on the observed data.

For each of these 15 scenarios, median values of the estimated childhood leukaemia incidence were reported with their empirically derived 95% ranges (95% range: 2.5% and 97.5% estimates from the 10,000 datasets). Figures were aggregated from the 16 selected

wards and compared the total number of cases observed with the number expected from the national incidence in people aged 0–14 years using the binomial exact test.[158] The proportion of significant *p*–values (5% level) for each test, together with the direction of the corresponding estimates (above/below the national incidence rate) was recorded. For simulated data, the proportion of significant *p*–values (5% level) is equivalent to the estimated type I error rate. *P*–values have been included along with confidence intervals as the original studies reported these.

6.3.7 'Region–wide' analytical strategy

To replicate the 'region–wide' strategy of previous studies,[145,159–166] Poisson regression models were used to match the distribution evident in the observed data and that used in the generation of the simulated datasets. Three separate regression models were conducted on a random selection of 50% of wards using 'population density' or 'inward–migration', or both, as covariates (corresponding to Scenarios 1, 2, 4 and 5 of the 'selective sub–region' analytical strategy, above). The arbitrary choice of selecting a random sample of 50% of the data for analysis was to ensure that the impact of random sampling variation across the simulations was present in both region–wide and selective sub–region approaches. Each model was generated 10,000 times for the observed data to facilitate comparisons with analysis of the 10,000 simulated datasets. Median risk ratios and their empirically derived 95% Ranges (95% Range: 2.5 and 97.5 centile estimates from the 10,000 datasets) are described for a 25% increase in population due to inward-migration and for a population density increase of 500 persons per km^2. Since population density is a continuous variable, a contrast between two states cannot be easily described; instead the effect of an absolute increase in population density is reported. The *p*–values corresponding to each risk ratio were recorded, combined with whether the risk ratio was above (harmful effect) or below (protective effect) one.

6.3.8 A note on *p*–values

Null hypothesis significance tests are inappropriate for observational data analyses and should not typically be used.[2] Unfortunately, they remain extremely common in the wider literature and all the historical studies that are emulated here used null hypothesis significance tests based on *p*–value thresholds. For comparison to these previous studies, the results here are explored in terms of the likelihood of obtaining $p < 0.05$ and (the currently preferred) absolute effect size.

6.4 Results

6.4.1 Results for the 'selective sub–region' analytical strategy

Analyses of 10,000 random samples drawn from the observed dataset using each 'selective sub–region' scenario (Table 6.4) indicate that, where selection was based on low population density or high inward–migration alone, or both (Scenarios 1, 2, 4 and 5), the proportions of significant *p*–values were low (ranging from $1.3\% - 3.6\%$). Where selection was based on either a high incidence of leukaemia, either alone or together with one or both exposures (Scenarios 3 and 6 –15), the proportions of significant *p*–values were substantially greater than the expected 5% (ranging from $18.4\% - 97.2\%$).

For analyses of data simulated under the null hypothesis, type I error rates of $2.8\% - 3.7\%$ were observed under Scenarios 1, 2, 4 and 5 (Table 6.4), consistent with random sub–region selection (i.e. 3.5% type I error rate). Where selections were based on a high incidence of leukaemia either alone or together with one or both exposures (Scenarios 3 and 6 –15), type I error rates were far higher (ranging from $18.3\% - 99.3\%$; Figure 6.4.1).

Table 6.4: Type 1 error rates of the 'selective sub–region' analytical strategy under each of the Scenarios examined.

Scenario	Observed Data Percentage statistically significant (5%)	Simulated Data Type 1 error rate (5%)
1. Low population density	2.93	2.90
2. High inward–migration	1.96	3.66
3. High incidence	34.30	67.02
4. Low population density–high inward–migration	1.32	3.17
5. High inward–migration–low population density	3.60	2.79
6. Low population density–high incidence	45.56	43.96
7. High incidence–low population density	44.88	18.34
8. High inward–migration–high incidence	27.01	41.63
9. High incidence–high inward–migration	97.22	67.28
10. Low population density–high inward–migration–high incidence	18.39	33.02
11. Low population density–high incidence–high inward–migration	22.77	45.16
12. High inward–migration–low population density–high incidence	60.06	44.18
13. High inward–migration–high incidence–low population density	30.88	99.28
14. High incidence–low population density–high inward–migration	41.21	19.22
15. High incidence–high inward–migration–low population density	47.67	20.90

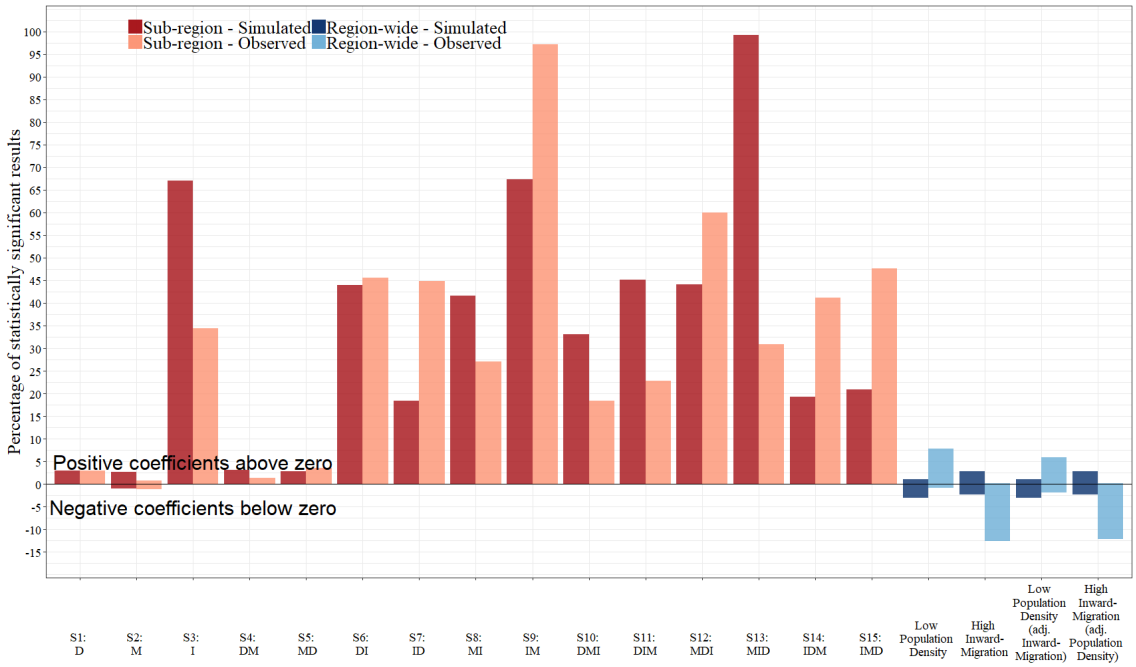

Figure 6.4.1: Percentage of statistically significant results at the 5% level by analytical strategy for both simulated and observed data. Selective subregion analytical strategy results were analyzed using the binomial exact test; direction of the bars indicates whether the estimated probabilities of the significant test results were greater (above zero) or less than (below zero) the national average. Region–wide analytical strategy results were analysed using Poisson regression; direction of the bars indicates whether statistically significant coefficients were greater (above zero) or less than (below zero) zero. D, population density; M, inward migration; and I, childhood leukaemia incidence; order of letters indicates the order used to select data for analysis.

The estimated 5–year incidence of childhood leukaemia ranged between 0 per 10,000 and 6 per 10,000 children across the 10,000 simulated datasets, indicating that up to 6 cases per 10,000 children might occur by chance in any five–year period. This is in contrast to what was simulated, i.e. 2 cases per 10,000 population. The range of estimates were similar in the observed datasets (Figure 6.4.2).

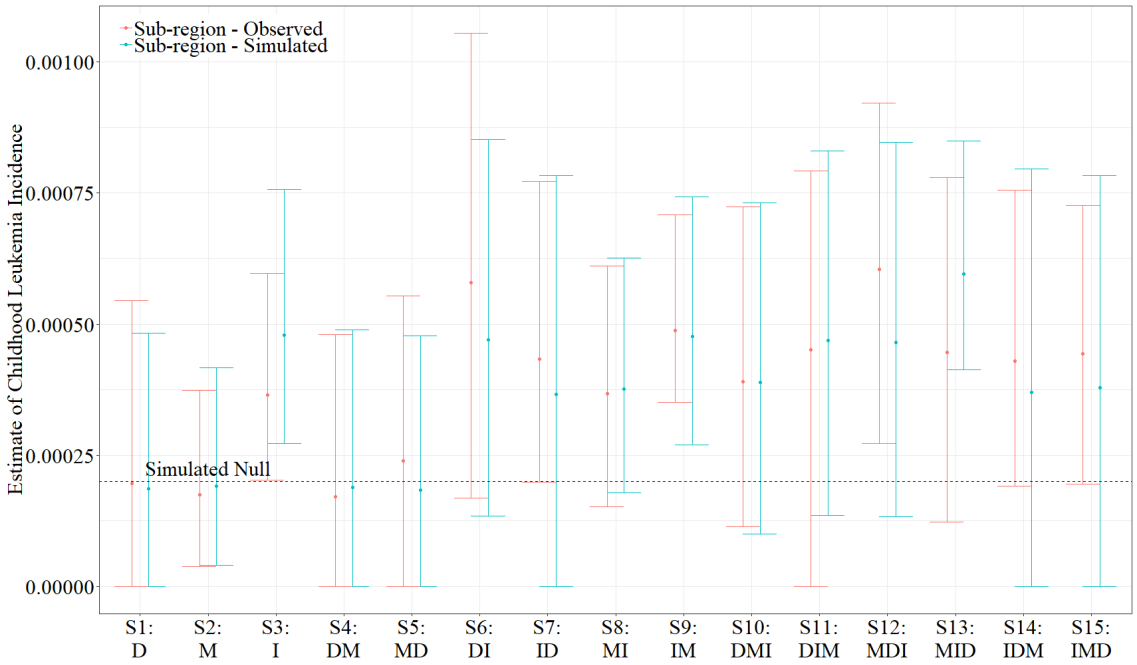

Figure 6.4.2: 95% empirically derived ranges (95% range: 2.5 and 97.5% centile estimates from the 10,000 datasets, points indicate the median) of the distribution of childhood leukaemia incidence from binomial exact test of the selective subregion analytical strategy. The dashed line indicates the incidence rate used in the simulated datasets, that is, two cases per 10,000 0–14 year olds in a 5–year period. This is the incidence expected under the null hypothesis; any deviation from this indicates bias. D, population density; M, inward–migration; and I, childhood leukaemia incidence; order of letters indicates the order used to select data for analysis.

6.4.2 Results for the 'region–wide' analytical strategy

The proportions of significant p–values in 'region–wide' analyses of observed data all exceeded 5% ($7.9\% - 13.0\%$; Table 6.5); suggesting that high inward–migration and low population density were associated with a lower and higher childhood leukaemia incidence, respectively. 'Region–wide' analyses of simulated data returned type I error rates between $4.2\% - 5.1\%$ for all model coefficients (Table 6.5). The distribution of the coefficient values is not centred on zero (i.e. $-2.5\% - 2.5\%$) due to small, but non–zero, correlations between cases of childhood leukaemia, population density and

inward–migration, which arise from a mathematical dependency between these variables.

Table 6.5: Type 1 error rates of the 'region–wide' analytical strategy according to the covariate examined in the model.

Covariate	Observed Data: Percentage statistically significant (5%)	Simulated Data: Type 1 error rate (5%)
Population density	8.71	4.15
Inward–migration	12.99	5.07
Population density (adjusted for inward–migration)	7.91	4.16
Inward–migration (adjusted for population density)	12.43	5.14

In the simulated data, the median risk ratios for the effects of inward–migration were consistently 1.0, indicating agreement with the null hypotheses (e.g. RR vs 0% migration: $25\% = 1.0$ [95% CI $= 0.08 - 8.81$]). In the observed data, however, increasing levels of inward–migration were associated with lower incidence of leukaemia (e.g. RR vs 0%: $25\% = 0.33$ [95% CI $= 0.02 - 2.05$]); the 95% confidence interval is not symmetric.

All risk ratios for the effect of population density in both the simulated data and observed data were close to 1.0, indicating consistent agreement with the null hypotheses (RRs per unit increase in person/km^2 in simulated data: 500–people/km^2 = 1.0 [95% CI $= 0.95 - 1.03$]; in observed data: 500–people/km^2 $= 0.98$ [95% CI $= 0.90 - 1.05$]. Coefficients of adjusted regression models (including both inward–migration and population density as covariates) did not materially differ from those in unadjusted models (Figure 6.4.3).

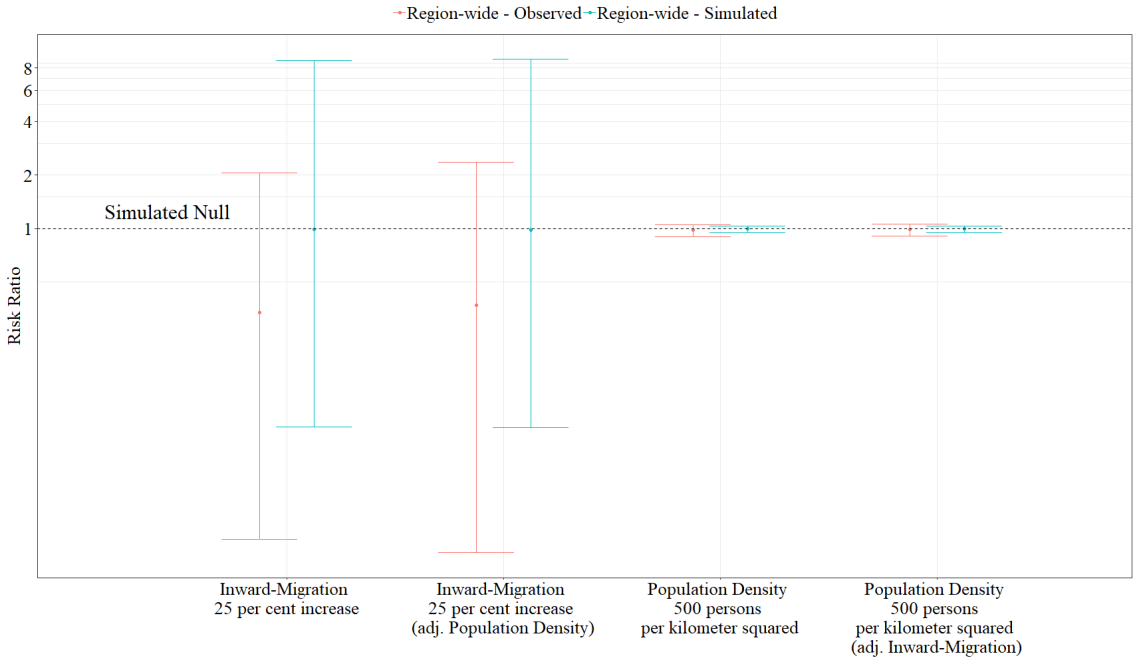

Figure 6.4.3: 95% empirically derived ranges (95% range: 2.5 and 97.5% centile estimates from the 10,000 datasets, points indicate the median) of the distribution of childhood leukaemia incidence of the percentage increase or decrease in childhood leukaemia incidence from the regression models of the region–wide analytical strategy with an increase of inward–migration of 25% and an increase in population density of 500 persons/km^2. The dashed line indicates no change in childhood leukaemia incidence as expected under the null hypothesis. Results shown with log scaling.

The ‘zip plot’ (Figure 6.4.4 clearly shows that the 95% confidence intervals of the region–wide analyses of the 10,000 datasets are unbiased (approximately 5%).

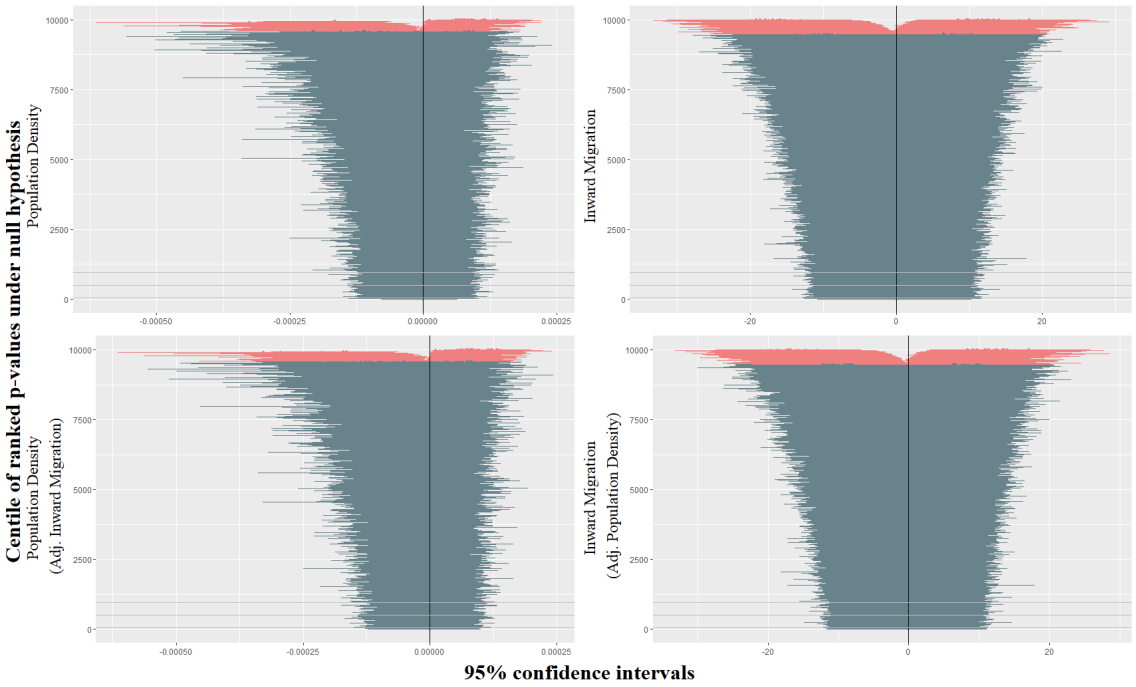

Figure 6.4.4: 'Zip plot' showing the 95% confidence intervals for analysis of each of the 10,000 datasets using the region–wide approach. Blue confidence intervals contain zero (the true value) whereas red confidence intervals do not.

The demographic data were simulated such that the correlation structure equals that of the observed data for the Yorkshire and Humber region as per the causal structure depicted in Figure 6.4.5: i.e. under the null hypothesis, only the population size causally influences the number of childhood leukaemia cases and there is no causal arrow between inward–migration and the size of the area. There will be a non–zero correlation between the number of 'Cases' and all four area measures ('Area size', 'Population density', number of 'Inward–migrants', and the 'Proportion of Inward–migrants') because 'Population' is causally related to them all. Since 'Population' is an offset term in the Poisson regression model, conditional independence between 'Cases' and both 'Area size' and the number of 'Inward–migrants' is assured due to 'controlling' for 'Population'. Conditional independence is not achieved between 'Cases' and either derived ratio variable ('Population density' and 'Proportion of inward–migrants') by

'controlling' for the 'Population' offset because both derived ratio variables contain an element of 'Population' explicitly; this explains a lack of symmetry in some of the *p*–values in Table 6.5.

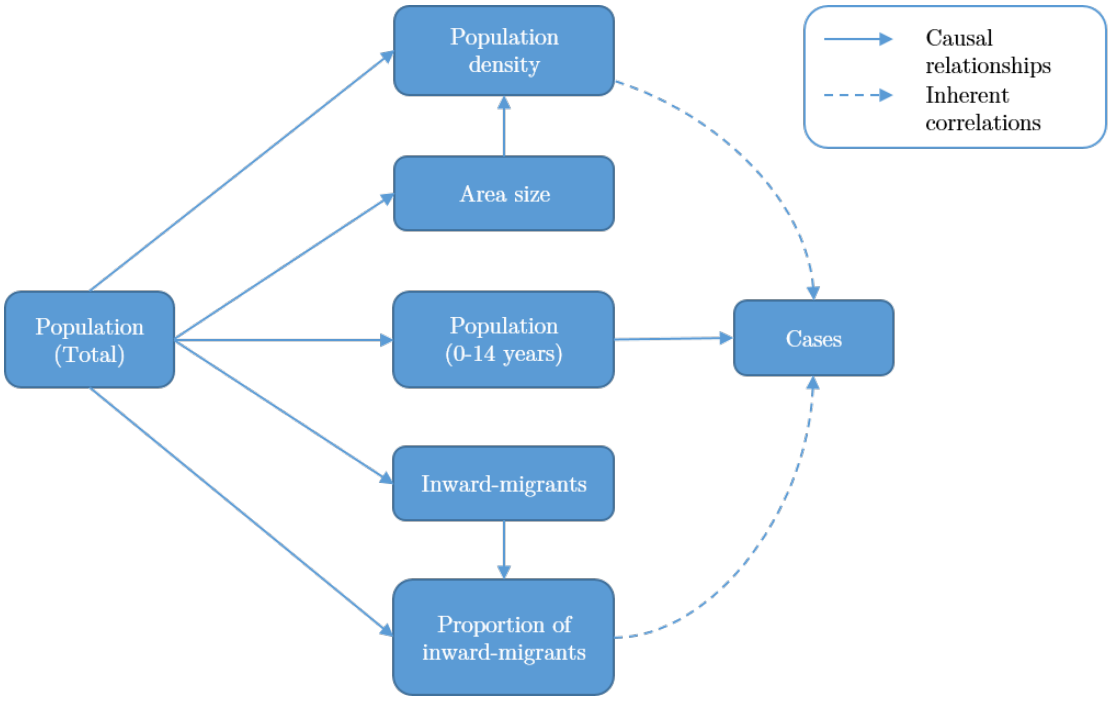

Figure 6.4.5: Graph representing the simulated relationships of variables within the dataset: assumed causal relationships are represented by solid arrows and implicit correlations are represented by dashed arrows.

6.5 Discussion

In this chapter it has been demonstrated how the different analytical strategies used to examine the relationship between 'population mixing' and childhood leukaemia incidence can generate radically divergent results. Considerable bias occurs if geographical areas are selected prior to analysis and selection is influenced by elevated childhood leukaemia incidence (i.e. by 'clusters'). Bias is also evident where selection involves measures of 'population mixing', as population mixing appears adversely associated with elevated

childhood leukaemia incidence.

All of the measures of 'population mixing' used by previous studies were not examined, and no attempt was made to generate alternative proxies for 'population mixing'. The aim, however, was to examine the impact of the two most common analytical strategies, using two typical measures of 'population mixing'. Furthermore, since the observed and simulated data did not differentiate between cases of childhood leukaemia amongst 'residents' and 'inward–migrants', it was not possible to assess whether 'population mixing' might be associated with a differential risk of childhood leukaemia in each. However, the small numbers of 'inward–migrants', and of 0–14 year–olds therein, would make such analyses challenging, and may explain why few previous studies have sought to do this. A further limitation of the present study is that possible temporal effects related to the timing of population mixing events and/or age at exposure were not examined. Despite substantial variation in these criteria amongst previously published studies, few sought to examine their impact on the direction or strength of the associations found. This, however, is less relevant to this chapter's focus on the comparison of analytical strategies.

Analyses of rare diseases such as childhood leukaemia are challenging because disease registries often only collect summary information on the denominator (or population 'at–risk') within areas for aggregated blocks of time. By far the most common type of analysis is therefore to conduct aggregated analyses of incidence proportions, i.e. comparing cases per population within fixed units of time. Because of the size of the areas typically examined, and the rarity of childhood leukaemia, such analyses are prone to an abundance of zero cell counts. The most common solution to this is to preserve the area level granularity and collapse the time frame into longer periods, with 5–year periods being the most common approach in the literature; the same approach has been adopted here for indicative analyses.

Alongside the limitations imposed from using data from disease registries mentioned above, there is also the limitation that averaged area–level migration patterns from census

data had to be used, these measure migration based on change of address from a year prior to the census date. Therefore, the likely time lag between exposure and event cannot be accounted for, however, this is also a limitation of many of the studies which are being emulated here.

Notwithstanding these limitations, the work in this chapter confirms that analyses based on non–random area/sub–region selections that are influenced by or associated with elevated childhood leukaemia incidence can generate entirely erroneous findings. In all Scenarios with such selection, associations were profoundly biased; falsely suggesting that low population density and/or high inward–migration were associated with elevated childhood leukaemia incidence.

Unfortunately, the lack of methodological clarity in research adopting a 'selective sub–region' analytical strategy means it is not possible to establish which studies might be prone to biases associated with this strategy. Even if studies sought to select areas using only variables chosen as measures for 'population mixing', it is feasible that selection was affected instrumentally (by co–dependence on demographic characteristics), or implicitly (by knowledge of, interest in, or attention to the outcome). The latter is likely to be central to the importance afforded to clusters of similarly rare events. It seems likely that focusing on clusters of childhood leukaemia, together with the confirmatory 'results' produced by 'selective sub–region' analyses, researchers are encouraged to use this analytical strategy, unaware of the bias it generates. This would explain the publication bias among studies examining the population mixing hypothesis.[153] Studies using the unbiased region–wide approach are more challenging to publish because they fail to identify the large artefact found in 'selective sub–region' analyses. Nevertheless, region–wide analytical strategies: avoid the risk of explicit or implicit attention to clusters; ensure that selection biases cannot occur; and can be extended to cover any available geographical characteristics. For this reason, ecological studies of the 'population mixing' hypothesis that have used a non–random 'selective sub–region' approach should be viewed with extreme caution.

6.6 Conclusion

Future studies investigating the association between population mixing and childhood leukaemia (or other 'clustered' events) should adopt a 'region–wide' analytical strategy to avoid the potential biases inherent in a non–random 'selective sub–region' approach. Where an entire dataset is not available for analysis, sampling should be random to avoid potential sub–region selection biases. Syntheses of previous studies examining this association should place greater emphasis on findings from studies adopting 'region–wide' analyses, and only consider findings from those studies using 'selective sub–region' analyses where the authors have explicitly used random selection methods to avoid the potential risk of focussing on areas exhibiting apparent clusters (i.e. a high incidence) of leukaemia.

6.7 In the context of the thesis

This chapter has illustrated the bias inherent to analyses focussing on clusters. It has used causal inference methods to consider how selection on the outcome influences analyses and links this with the 'most dangerous equation'[87,88] and the modifiable areal and temporal unit problems.[84–86]

Simulation and causal inference influenced thinking can help researchers understand biases that may be present in the historical approaches to data analysis. This chapter is an example of how these two domains can be brought together to uncover unbiased analytical analyses in health geography.

The next chapter summarises the findings of all chapters, discusses the strengths and limitations of the research and makes suggestions for future research.

Chapter 7

Conclusion

7.1 Overview

This chapter summarises the findings of the entire thesis and critically evaluates these findings. The three major elements of the thesis are discussed: causal inference methods, simulation, and health geography, and particularly these topics combined. Future research directions are suggested, particularly in relation to similar work in the literature and lastly, a final overview of the thesis is provided.

Statistical methods are often used habitually.[1,2] This thesis aimed to provide a framework for using simulation and causal inference methods in health geography so that other researchers can critically evaluate the methods used in their own work and that of others. In the course of doing this, this thesis has considered these methods to investigate applied problems related to mathematical coupling (Chapter 4) and specific considerations that need to be made in relation to research on the relationship between limiting long–term illness and deprivation (Chapter 5) and the challenges encountered while investigating the relationship between population mixing and childhood leukaemia (Chapter 6). The datasets chosen for this thesis are representative of many others in health geography; LLTI

is a relatively common outcome and childhood leukaemia is a relatively rare outcome. The methods used here aim to be translatable to further health geography research questions.

Methods in causal inference and simulation are powerful tools in understanding bias and with careful planning, forethought and reflection on the data generating processes of the context of interest they can be made accessible to all researchers.

7.1.1 Causal Inference

Chapter 2 showed the intersection between causal inference and simulation as background to the thesis as a whole. Although this knowledge is taken for granted in some domains due to the link between the two topics via the data generation process, the applicability of this was made explicit. This was achieved, in part, by showing examples of how causal inference methods have been used in the past to elicit deeper understanding of paradoxes and biases, such as Simpson's Paradox[34] and the mutual adjustment fallacy.[49,50]

By introducing a causal perspective to the research challenges covered in this thesis some important aspects have been considered that would need to be fully explored if causal inference methods are to be usefully integrated within the field of health geography. Section 7.5 revisits the three causal assumptions, exchangeability, positivity and consistency, and additionally discusses the issue of interference, introduced in Chapter 2, and discusses how these have started to be addressed in the literature in relation to population health. Ideas are discussed around possible ways that interference might be explored in future health geography research, with wider implications for addressing interference more generally within a DAG framework.

7.1.2 Simulation

Chapter 3 detailed how simulation and causal inference are naturally linked via the data generating mechanism and how these can be applied in the case of health geography. This chapter showed methods on how to undertake simulation informed by causal inference techniques. This is an expansion from using mathematical closed–form solutions because these cannot account for all aspects of causality, for example causal direction.

An important contribution of Chapter 3 is the consideration of composite variables and compositional data in causal diagrams and therefore their importance in simulation because of their relation to the data generation process. Composite variables and compositional data are often a feature of health geography research and, in order to fully integrate causal thinking into the field of health geography, these concepts must be taken into account. Future work could look deeper into incorporating composite variables and compositional data into causal thinking although this may have implications for the consistency assumption.

Chapter 3 also introduced a step–by–step guide to simulation which combined literature on conducting simulation studies in R,[54] integrating causal modelling and statistical estimations[9] and yielding a systematic approach to simulation studies to evaluate statistical methods.[51] This guide can be used in any future studies evaluating and comparing methods that explicitly consider causality. It is often taken for granted that it is known that DAGs represent the data generation process and that it is obvious how they can be combined with simulation. It is hoped that, by making this explicit and providing a framework which researchers can use in their own work, others will more readily be able to critically evaluate the methods that they use by developing simulations alongside applied research where possible.

7.1.3 Health Geography

Geography is important in relation to how particular locations and types of places relate to variation in health but also in how geography influences the planning of data collection and as the choice of variables and their utility in the dissemination of (causal) information for research purposes. When geographical structure and causal inference come together in the dissemination and analysis of health data, it can affect the results in important ways that are not always intuitively understood. Causal inference methods and simulation, when combined, can be used to ensure that the influence of geography on the ensuing analyses does not introduce bias.

Observational data are not always collected to answer specific research questions[59] and this can have consequences for the analyses that are conducted. Chapter 3 introduced some considerations due to data provenance. These are important considerations in observational research and, in particular, health geography research. When using secondary data, the purposes for which the data were collected should be considered as seen in the population mixing and childhood leukaemia example (6) where selection directly or indirectly related to the outcome introduced bias to the results of some analyses.

In some cases, researchers must work with the data that are available to them and they have no control over how the data were collected. In these circumstances it is important to consider the best methods to minimise bias and what would have been different if the optimal dataset could have been acquired and what, if any, are the biases that result from the predetermined data provenance; quantifying the extent of any bias is then important.

7.2 Findings

7.2.1 Mathematical Coupling

Chapter 4 introduced important insights into the mathematical coupling of ratios with common denominators that could only be understood using causal inference methods. This showed under what circumstances the historical solution of including the common denominator as a covariate in a regression model is appropriate, i.e. when it acts as a 'confounder' of the exposure–outcome relationship. Chapter 4 also proceeded to quantify the extent of this problem and developed this into the geographical situation where an exposure (X) and outcome (Y) are both components of the common denominator (N), i.e. $X \leq N$ and $Y \leq N$ but $X + Y \neq N$. This showed that if X and Y are proportional to N it is acceptable to divide through by N in order to condition on it. However, this is unlikely to be the case in complex observational data, especially when secondary data sources have to be combined.

Only when the numerator and denominator are proportional to each other is it appropriate to correlate two ratios with a common denominator. However, the real–world is complex and understanding the data generating process is difficult (as has been illustrated here in simulating several scenarios). It is unlikely that two variables would be exactly proportional and that no variables other than the exposure of interest caused the outcome. It has been said that any variable that precedes an outcome in time could be considered a cause,[24] although this would be impossible to analyse so simplifications are made because the effect of some variables would be incredibly small.

Often, if there are restricted circumstances under which a method generates appropriate results, over time, these restricted circumstances are forgotten. It is prudent to adhere to methods that will produce appropriate inferences that do not assume these restrictions. Avoiding ratio variables in correlation and regression analyses is therefore the preferred

approach to avoid mathematical coupling bias.

Mathematical coupling was first mentioned in the literature by Pearson in 1896.[90] However, bias associated with it still appears in the literature today. The chapter on this topic contributes to the literature by providing a more contemporary perspective which will hopefully reach a wider audience of researchers who will avoid mathematical coupling bias in the future and recognise it in the literature that they consume.

7.2.2 Limiting Long–Term Illness and Deprivation

Chapter 5 introduced the relatively common condition, limiting long–term illness (LLTI) and showed the application of compositional data, composite variables and mathematical coupling in an applied setting. The chapter used simulation to demonstrate the complex relationships between the variables included in the analyses investigating the 'causal effects' of deprivation on LLTI.

This chapter gives an example of how a researcher would go about answering a causal question on such a dataset and some aspects that would need to be thought through. The aspects covered are by no means all those that could be raised in the analysis of such a research question, however, they illustrate the kinds of things a researcher should look out for should they wish to consider causal relationships. Research has often focused on the theoretical side whereas, particularly Chapter 5, has used an existing dataset to explore aspects of health geography that need to be considered when introducing and applying causal inference methods to this field.

7.2.3 Population Mixing and Childhood Leukaemia

Chapter 6 introduced the example of 'population mixing' and, in contrast to Chapter 5, the relatively rare condition of childhood leukaemia. This built upon the author's MSc

research which had already ascertained that analyses in the literature were not subject to mathematical coupling bias but required further exploration to understand the disparity in results obtained from the two analytical methods that predominate in the literature.

The simulations undertaken in this chapter followed the simulation framework introduced in Chapter 3 but this was formalised later in the undertaking of this research. By formulating the problem in a causal inference framework, upon which subsequent simulations were based, it was made clear that the disparity in results in the literature was a result of possible conditioning on the outcome by focusing on clusters of childhood leukaemia. This problem was then exacerbated by the outcome being rare and bias as a result of what has been termed 'the most dangerous equation'[87,88] and the modifiable and temporal unit problems.[84–86] It was shown how using a region–wide approach to this research question avoids these problems and can be extended to cover any available geographical characteristics.

7.3 Contributions to the Literature

The work featured in Chapter 6 has already been published in *Epidemiology*[130] and has been the subject of a letter to the editor from the originator of the 'population mixing hypothesis'[131] which has been responded to.[132] It has also already made an impact in the cancer epidemiology literature[133] where it was noted that focusing on clusters in related research can produce biased results.

Large amounts of coding was required for the undertaking of the research in this thesis and, as "no simulation study is definitive and new methods or refinements of methods are inevitable"(p.25),[51] the code for the simulations is freely available in the Appendices. This means that others can replicate and extend the work presented in this thesis along with using it as a basis for, and to inform, their own similar studies.

A lot of the causal inference literature focusses on methods research, however, it is important to show its utility (along with simulation) in an applied setting. Hopefully, this approach can introduce more researchers to the advantages of using these important tools. Chapter 4 on mathematical coupling is in preparation for journal submission along with Chapter 5 on limiting long–term illness and deprivation.

Chapter 3 is contributing to another paper around compositional data analysis in collaboration with others, though not led by the author of this thesis. Two 'How to...' guides are also planned for inclusion in SAGE Methods around simulation and regression modelling similar to another paper the author was involved in.[167]

7.4 Limitations

The simulations throughout the thesis have been conducted under the null hypothesis. This has been sufficient, generally, to show the presence of bias under certain analytical strategies. However, these simulations could be expanded to generate more in–depth understanding of the methods used. On the other hand, where bias has been shown under the null condition it could be assumed that this bias will be present when the null hypothesis is not satisfied and research efforts could be directed towards more promising (i.e. less biased) methods instead.

In much observational research aiming to incorporate causal inference, there is a focus on the exchangeability assumption. This means that units of analysis are "identical on average for characteristics that may affect the outcome except the outcome itself" (p.3).[168] This has also been the case in this thesis, where reducing confounder bias via conditioning has been a feature of each chapter. However, in order to fully integrate causal inference methodology into health geography, the other assumptions required to identify a causal effect (positivity and consistency) must be satisfied or at least considered. Along with these assumptions, traditional regression techniques "assume that there is no interference

between units, but that is often not a realistic assumption"(p.101).[169] These assumptions are considered further in Section 7.5.

Developments late in the research for this thesis regarding mathematical coupling in a health geography setting suggest that the simulations are slightly limited in scope in ways that may have repercussions for the investigations into LLTI and deprivation. This stems from the complication generated by geographical contexts of inherent data hierarchy, where individual influences become conflated with geographical area influences, and thought must be given to what is meant by 'geographical structure'. The role of 'area' may therefore be considered distinct from the role of individuals residing within an area, which introduces potentially more complex heterogeneity through either simple random variation or variation brought about through unmeasured confounding, that operates at the area–level and not the individual–level. Unpicking these concepts is not trivial. For instance, the exposure and outcome from the examples in Chapter 4 and the simulations in Chapter 5 used the binomial distribution to generate variables from the total population. When these exposures and outcomes are divided through by the population size (in an attempt to acknowledge the geographical area structure and thus to control for area population size) the numerator and denominator are proportional to each other and no 'spurious' correlation is generated. However, this ignores the possibility (indeed likelihood) that heterogeneity amongst geographical areas is independent to the heterogeneity within each area. This has implications on the data–generating simulation considerations to be adopted under various circumstances.

Although many issues have been accounted for in the simulation of this problem (e.g. compositional data and composite variables), this has not considered all possible scenarios involving different underlying causes to heterogeneity both within and between geographical structures. Consequently, it is unlikely that this has created a true picture of the health geography context in which these methods would be used. It is likely that there are many other variables involved in the relationship between the exposure and the

outcome that cannot be accounted for by dividing through by the population size alone. A suggested issue to consider in work following on from this thesis would be to investigate potential influences upon geographical area heterogeneity independently to investigating the potential influences of within–area heterogeneity. In effect, consideration might be given to a hierarchical or multilevel causal structure, where a DAG may be used to describe individual causal relationships distinct from area–level causal relationships. This is an extensive area of future research but its implications are worth briefly considering, as it also has relevance to 'interference', as is discussed in the Section on future work (Section 7.5).

7.5 Future Work

There are several directions that work following on from this thesis could take. Some of these directions are now discussed.

7.5.1 Acknowledging causal hierarchy within health geography

It is difficult to reach causal conclusions in the social sciences. By investigating causation from many angles, however, it may be possible to have a greater understanding of causal processes. It is important to understand aggregate i.e. population–level phenomena, particularly if a researcher believes that there is a causal mechanism which acts at that aggregate level.[170] Smith suggests that "causation should be thought of as operating at the lowest level at which a policy could conceivably by implemented" (p.464).[44] There has been a focus in the causal inference literature on individual–level causal analysis and on the interventionist approach (which suggests that there is "no causation without manipulation"[15]) but this is not often appropriate in health geography research.[31] Going forward, it will be important to understand how to engage causal inference

thinking outside of the individual–level approach and embrace the aggregate–level that health geography research necessitates and integrate causal thinking within an explicit hierarchical structure.

As mentioned in Section 7.4, the causal assumptions of positivity and consistency are often overlooked in observational research, however, positivity has been considered in a commentary by Westreich and Cole[171] who discuss how positivity violations in one's dataset can be uncovered and how they can be dealt with in practice and commend other authors for considering this often overlooked issue.[172] VanderWeele[173] specifically discusses the positivity assumption in relation to 'neighbourhood effects' and it would be useful to expand this to looking at health geography scenarios where there are individual and area–level effects.

The consistency assumption has been mentioned in regard the composite variables and compositional data (Section 3.10.3) discussed in this thesis. This is an important assumption because if an exposure is not well–defined then the related effects on the outcome are not well–defined and suggested interventions may not be effective.[168] It may, thus, be inappropriate to use composite variables in the causal analysis of health geography datasets and their constituent parts should be used instead. However, further research could examine the effect of different violations of this assumption. Rehkopf *et al.*,[168] suggest that "particularly, early in the arc of a research question, it may be important to cast a wide net, examine unclearly defined constructs, and try to integrate evidence across studies with measures that do not clearly correspond to a specific intervention" (p.72).

Most work in causal inference has assumed that there is no interference between units, however, there are many contexts in which this assumption is not appropriate.[20] Health geography is one of those contexts. Tchetgen Tchetgen and VanderWeele[20] considered a method for dealing with interference in observational studies and these have been extended by Papadogeorgou *et al.*[174] to consider population–level interventions over a

collection of clusters and Zigler and Papadogeorgou[22] have developed bipartite causal inference with interference. These methods could be valuable in health geography and should be investigated further.

As one of the limitations of the simulations used in this thesis was that they did not consider area–level heterogeneity, future work should endeavour to include this. These methods could utilise multi–level models and the principle of maximising the proportion of explained variation in an outcome taken from statistical (as opposed to causal) inference discussed in Section 2.4. Preliminary investigations could be made, for example, into controlling variables at the area–level to achieve the maximum proportion of explained variation, thereby controlling away all area–level confounding. This could allow the researcher to elucidate the causal relationships at the lower or individual levels if such data were available. This may also provide a way of dealing with interference between units of analysis by partitioning effects at different geographical levels.

A limitation of the mathematical coupling and LLTI simulations (Chapters 4 and 5, respectively) was that they were conducted under the null hypothesis. Future work could consider the non–null scenario and the circumstance of area–level heterogeneity in a way similar to that suggested above. It could also investigate the circumstances in which the common denominator is a collider or mediator variable. However, it is apparent that dividing through by the common denominator is only a valid approach when the numerator and denominator are proportional which is unlikely in observational datasets. However, another avenue that further research could take would be to investigate the accuracy of the measurement of population denominators in health geography research. This would lead on to questions around confidentiality of individuals' data and the utility of data that has deliberately been blurred to preserve confidentiality.

7.5.2 Causal inference in applied health geography research

Causal inference would benefit from applied research demonstrating how relevant methods can be implemented. Studies similar to those in Chapters 5 and 6 could be undertaken to gain further clarity on the methods used but also to bring causal understanding to research questions in health geography. By introducing a framework from which researchers can build simulations that incorporate causal inference methods this could prove to be a fruitful avenue for further research. In particular, it would allow others to evaluate their methods so that they can make conscientious choices around how they conduct their research. All of the findings of the applied research should be taken into account when evaluating the literature and can be used to critically evaluate one's own and others' research.

"The determination that an association is causal indicates the possibility for intervention" (p.61).[175] However, even if it is not possible to fully meet the three assumptions required to identify a causal effect within health geography, there are many useful tools from causal inference thinking that can be applied to health geography problems which will strengthen research in this domain.

7.6 Summary

This thesis has illustrated how simulation and causal inference influenced thinking can help researchers understand biases that may be present in the historical approaches to health geography data analyses. The thesis provides a framework for undertaking causal inference informed simulations and analyses. It demonstrates how this can be used theoretically in the case of understanding mathematical coupling (which is ubiquitous in health geography) and the appropriateness of the historical solutions to this problem and in the case of two applied scenarios. These two applied scenarios were chosen to

represent a large range of research questions in health geography as one considers a relatively common outcome and the other a relatively rare outcome.

Great insights can be made when considering research within a causal framework which can be further expanded using simulation to quickly consider many alternative scenarios. This thesis provides a framework for considering both of these aspects, by aiding researchers to critically evaluate the methods they use with the hope of moving away from the habitual use of statistical methods.

It is hoped that causal inference informed research will lead to more robust results and reliable interventions.

Appendices

Appendix A

Simulations illustrating the Modifiable Areal Unit Problem

```
1  ##################################
2  ## Introductory Example - MAUP ##
3  ##################################
4
5  # load packages
6  require(ggplot2)
7  cbPal   <- c("#999999","#E69F00","#56B4E9","#009E73","#F0E442","#0072B2","#D55E00","#CC79A7")
8
9  # A hypothetical nation
10 # 100km^2; 500 cases of childhood leukaemia spread randomly
11 # across the nation
```

```
12
13 set.seed(1123)
14 cl <- 300
15 x  <- runif(cl)*100
16 y  <- runif(cl)*100
17
18 dat  <- data.frame(x, y)
19
20 # Map of points
21 ggplot( dat, aes( x, y ) ) +
22   geom_point( size = 2, colour = cbPal[1] ) +
23   theme( axis.text = element_text( colour = "black", size = 16, family = "Times New Roman" ),
24        axis.title = element_text( size = 16, family = "Times New Roman", face = "bold" ),
25        legend.title = element_text( size = 12, family = "Times New Roman", face = "bold" ),
26        legend.text = element_text( size = 12, family = "Times New Roman" ),
27        plot.title = element_text( size = 12, family = "Times New Roman", face = "bold" ) )
28
29
30 ggplot( dat, aes( x, y ) ) +
31   geom_point( size = 2, colour = cbPal[1] ) +
32   theme( panel.grid.major = element_line( colour = cbPal[6], size = 1, linetype = "solid" ),
33          axis.text = element_text( colour = "black", size = 16, family = "Times New Roman" ),
34          axis.title = element_text( size = 16, family = "Times New Roman", face = "bold" ),
35          legend.title = element_text( size = 12, family = "Times New Roman", face = "bold" ),
36          legend.text = element_text( size = 12, family = "Times New Roman" ),
37          plot.title = element_text( size = 12, family = "Times New Roman", face = "bold" ) ) +
38   scale_y_continuous( breaks = seq( 0, 100, 10 ) ) +
39   scale_x_continuous( breaks = seq( 0, 100, 10 ) )
40
41 # Assign points to the square they are in
```

```
dat$row  <- 0
dat$col  <- 0
dat$pos  <- 0
for ( j in 1:nrow( dat ) ) {
  dat$row  <- ceiling( dat$y/10 )
  dat$col  <- ceiling( dat$x/10 )
  dat$pos  <- 10*( dat$row - 1 ) + dat$col
}

# How many points are in each square?
dat$count  <- 1
gridpoint  <- aggregate( count ~ pos, dat, sum )
dat        <- dat[, -6]
dat$clus   <- "0"

# What about squares with no cases?
gridpointZero  <- nrow(gridpoint)
if( gridpointZero < 100 ) {
  zero  <- which( !( 1:100 %in% gridpoint$pos ) )
  addrows  <- ( gridpointZero + 1 ):( gridpointZero + length( zero ) )
  gridpoint[addrows, 1]  <- zero
  gridpoint[addrows, 2]  <- 0
  gridpoint  <- gridpoint[order( gridpoint$pos), ]
}

# What is the distribution of cases across the squares?
table  <- table(gridpoint$count)
print(table)

# What is the biggest cluster?
```

```
big  <- max( gridpoint$count )
loc  <- which( gridpoint$count == max( gridpoint$count ) )
dat[dat$pos %in% loc, 6]  <- "1"

# What is the second biggest cluster?
loc  <- which( gridpoint$count == 7 )
dat[dat$pos %in% loc, 6]  <- "2"

# Where are the grid squares with no cases?
loc  <- which( gridpoint$count == 0 )
dat[dat$pos %in% loc, 6]  <- NA

d  <- data.frame(x = c( 50, 90, 50, 90 ), y = c(30, 30, 40, 50) )

ggplot( dat, aes( x, y, colour = clus ) ) +
  geom_point( size = 2 ) +
  theme( panel.grid.major = element_line( colour = cbPal[1], size = 1, linetype = "solid" ),
         axis.text = element_text(colour = "black", size = 16, family = "Times New Roman"),
         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
         legend.title = element_text(size = 12, family = "Times New Roman", face = "bold"),
         legend.text = element_text(size = 12, family = "Times New Roman"),
         legend.position = "bottom") +
  scale_y_continuous( breaks = seq( 0, 100, 10 ) ) +
  scale_x_continuous( breaks = seq( 0, 100, 10 ) ) +
  scale_colour_manual( values = cbPal, breaks = c("0", "1", "2" ),
                       name = "Cluster", labels = c("Middle", "Largest", "Second Largest") ) +
  geom_segment( data = d, mapping = aes(x = x, y = y, xend = x + 10, yend = y), size = 1.5, color = cbPal[6]) +
  geom_segment( data = d, mapping = aes(x = x, y = y, xend = x, yend = y + 10), size = 1.5, color = cbPal[6]) +
  geom_segment( data = d, mapping = aes(x = x, y = y + 10, xend = x + 10, yend = y + 10), size = 1.5, color = cbPal[6]) +
  geom_segment( data = d, mapping = aes(x = x + 10, y = y, xend = x + 10, yend = y + 10), size = 1.5, color = cbPal[6])
```

```
102
103
104 library(plotly)
105 library(reshape2)
106
107 p <- dat %>%
108   melt() %>%
109   ggplot(aes(x, y, fill = clus)) + geom_tile()
110
111 p <- ggplotly(p)
112
113 loc  <- which( gridpoint$count == 0 )
114 zeroes  <- data.frame( loc )
115 zeroes$row  <- ceiling( zeroes$loc/10 )
116 zeroes$col  <- zeroes$loc %% 10
117 w  <- which( zeroes$col == 0 )
118 if( length( w ) > 0 ) {
119   zeroes$col[w] = 10
120 }
121 print( zeroes )
```

Appendix B

Simulations illustrating mathematical coupling due to a common denominator

```
#########################################################################
## Simulations of the examples in: Mathematical coupling of propotions: ##
## revisiting Pearson, Neyman and Fisher with causal graphs             ##
#########################################################################

####################
## Load packages ##
####################

require(Matrix); require(matrixcalc)
require(base64enc); require(devtools); require(RVAideMemoire); require(MASS); require(ppcor)
```

```
#devtools::install_github("jtextor/dagitty/r")
require(dagitty)
set.seed(1123)

#####################
## DATA SIMULATION ##
#####################

#########################################################
## Pearson Example with totally independent variables ##
#########################################################

d1 <- "dag{X Y N}"

d <- simulateSEM(d1,N=10000)

d$N <- d$N + 5
d$X <- d$X + 5
d$Y <- d$Y + 5

d$XN <- d$X/d$N
d$YN <- d$Y/d$N

par(mfrow=c(2,2))
scatter.smooth(d$X,d$N, main=signif(cor(d$X,d$N)))
scatter.smooth(d$Y,d$N, main=signif(cor(d$Y,d$N)))
scatter.smooth(d$X,d$Y, main=signif(cor(d$X,d$Y)))
scatter.smooth(d$XN, d$YN, main=signif(cor(d$XN,d$YN)))

cor(d$X, d$Y)
```

```
cor(d$XN, d$YN)

summary(lm(Y ~ X + N, data = d))

################
## Example 1 ##
################

# Specify DAG1 from which data are simulated
DAG1 <- dagitty('dag{
                Pop [pos="0,0.5"]
                Cats [pos="1,0"]
                Apples  [pos="1,1"]
                Pop -> Cats [beta= 0.4]
                Pop -> Apples  [beta= 0.4]
                }')
plot(DAG1)



d2 <- simulateSEM(DAG1,N=10000)

d2$Pop    <- d2$Pop + 5
d2$Cats   <- d2$Cats + 5
d2$Apples <- d2$Apples + 5

d2$XN <- d2$Cats/d2$Pop
d2$YN <- d2$Apples/d2$Pop

par(mfrow=c(2,2))
```

```
72
73 scatter.smooth(d2$Cats,d2$Pop, main=signif(cor(d2$Cats,d2$Pop)))
74 scatter.smooth(d2$Apples,d2$Pop, main=signif(cor(d2$Apples,d2$Pop)))
75 scatter.smooth(d2$Cats,d2$Apples, main=signif(cor(d2$Cats,d2$Apples)))
76 scatter.smooth(d2$XN, d2$YN, main=signif(cor(d2$XN,d2$YN)))
77
78 summary(lm(Apples ~ Cats, data = d2))
79 summary(lm(YN ~ XN, data = d2))
80 summary(lm(Apples ~ Cats + Pop, data = d2))
81
82 cor(d2$Cats, d2$Apples)
83 cor(d2$XN, d2$YN)
84
85 summary(lm(Apples ~ Cats + Pop, data = d2))
86
87
88 ################
89 ## Example 2 ##
90 ################
91
92 # Specify DAG2 from which data are simulated
93 DAG2 <- dagitty('dag{
94                 Pop [pos="0,0.5"]
95                 Exercise [pos= "1,0"]
96                 Antidepressants  [pos= "1,1"]
97                 Pop -> Exercise [beta = 0.4]
98                 Pop -> Antidepressants [beta = 0.4]
99                 Exercise -> Antidepressants [beta = 0.4]
100                }')
101
```

```
plot(DAG2)


d3 <- simulateSEM(DAG2,N=10000)

## Simulate data based on DAG2
MyCov2 <- impliedCovarianceMatrix(DAG2)
N2 <- 10000
Mu2 <- c(150000, 1000, 75) # Minutes exercised per week, Pop, Antidepressant prescriptions
SD2 <- Mu2/5
MyData2 <- data.frame(mvrnorm(N2, Mu2, MyCov2, empirical = FALSE))
MyData2$pcExercise <- MyData2$Exercise/MyData2$Pop
MyData2$pcAntidepressants <- MyData2$Antidepressants/MyData2$Pop

## Standardise MyData2
MyData2 <- data.frame(scale(MyData2))

## Get confidence intervals of correlation and partial correlation coefficients
summary(lm(Antidepressants ~ Exercise, data = MyData2))
confint(lm(Antidepressants ~ Exercise, data = MyData2))
summary(lm(pcAntidepressants ~ pcExercise, data = MyData2))
confint(lm(pcAntidepressants ~ pcExercise, data = MyData2))
summary(lm(Antidepressants ~ Exercise + Pop, data = MyData2))
confint(lm(Antidepressants ~ Exercise + Pop, data = MyData2))

################
## Example 3 ##
################

# Specify DAG3 from which data are simulated
```

```
DAG3 <- dagitty('dag{
                JobOpps [pos="0,0.5"]
                Pop [pos= "0.5,0.3"]
                HealthCare  [pos= "1,0.5"]
                JobOpps -> Pop [beta = 0.4]
                JobOpps -> HealthCare [beta = 0.4]
                Pop -> HealthCare [beta = 0.4]
                }')

plot(DAG3)
d4 <- simulateSEM(DAG3,N=10000)

## Simulate data based on DAG3
MyCov3 <- impliedCovarianceMatrix(DAG3)
N3 <- 10000
Mu3 <- c(25, 1000, 3) # JobOpps, Pop, HealthCare (millions ?GBP)
SD3 <- Mu3/5
MyData3 <- data.frame(mvrnorm(N3, Mu3, MyCov3, empirical = FALSE))
MyData3$JobOppsPC <- MyData3$JobOpps/MyData3$Pop
MyData3$HealthPC <- MyData3$HealthCare/MyData3$Pop

## Standardise MyData3
MyData3 <- data.frame(scale(MyData3))

## Get confidence intervals of correlation and partial correlation coefficients
summary(lm(HealthCare ~ JobOpps, data = MyData3))
confint(lm(HealthCare ~ JobOpps, data = MyData3))
summary(lm(HealthPC ~ JobOppsPC, data = MyData3))
confint(lm(HealthPC ~ JobOppsPC, data = MyData3))
summary(lm(HealthCare ~ JobOpps + Pop, data = MyData3))
```

```
confint(lm(HealthCare ~ JobOpps + Pop, data = MyData3))

################
## Example 4 ##
################

# Specify DAG4 from which data are simulated
DAG4 <- dagitty('dag{
                Migration [pos="0,0"]
                Pop [pos= "0.25,0.25"]
                Births  [pos= "0,0.5"]
                Migration -> Pop [beta = 0.4]
                Births -> Pop [beta = 0.4]
                }')

plot(DAG4)
d5 <- simulateSEM(DAG4,N=10000)


## Simulate data based on DAG4
MyCov4 <- impliedCovarianceMatrix(DAG4)
N4 <- 10000
Mu4 <- c(5, 12, 1000) # Migration, Births, Pop
SD4 <- Mu4/5
MyData4 <- data.frame(mvrnorm(N4, Mu4, MyCov4, empirical = FALSE))
MyData4$MigrationPC <- MyData4$Migration/MyData4$Pop
MyData4$BirthsPC <- MyData4$Births/MyData4$Pop

## Standardise MyData4
MyData4 <- data.frame(scale(MyData4))
```

```
192
193 ## Get confidence intervals of correlation and partial correlation coefficients
194 summary(lm(Births ~ Migration, data = MyData4))
195 confint(lm(Births ~ Migration, data = MyData4))
196 summary(lm(BirthsPC ~ MigrationPC, data = MyData4))
197 confint(lm(BirthsPC ~ MigrationPC, data = MyData4))
198 summary(lm(Births ~ Migration + Pop, data = MyData4))
199 confint(lm(Births ~ Migration + Pop, data = MyData4))
200
201 ################
202 ## Example 5 ##
203 ################
204
205 # Specify DAG5 from which data are simulated
206 DAG5 <- dagitty('dag{
207                 NewHousing [pos="0,0"]
208                 Pop [pos= "0.25,0.25"]
209                 Immigration  [pos= "0.125,0.5"]
210                 NewHousing -> Immigration [beta = 0.4]
211                 NewHousing -> Pop [beta = 0.4]
212                 Immigration -> Pop [beta = 0.4]
213                 }')
214
215 plot(DAG5)
216 d6 <- simulateSEM(DAG5,N=10000)
217
218
219 ## Simulate data based on DAG5
220 MyCov5 <- impliedCovarianceMatrix(DAG5)
221 N5 <- 10000
```

```
Mu5 <- c(5, 5, 1000) # NewHousing, Immigration, Pop
SD5 <- Mu5/5
MyData5 <- data.frame(mvrnorm(N5, Mu5, MyCov5, empirical = FALSE))
MyData5$NewHousingPC <- MyData5$NewHousing/MyData5$Pop
MyData5$ImmigrationPC <- MyData5$Immigration/MyData5$Pop

## Standardise MyData5
MyData5 <- data.frame(scale(MyData5))

## Get confidence intervals of correlation and partial correlation coefficients
summary(lm(Immigration ~ NewHousing, data = MyData5))
confint(lm(Immigration ~ NewHousing, data = MyData5))
summary(lm(ImmigrationPC ~ NewHousingPC, data = MyData5))
confint(lm(ImmigrationPC ~ NewHousingPC, data = MyData5))
summary(lm(Immigration ~ NewHousing + Pop, data = MyData5))
confint(lm(Immigration ~ NewHousing + Pop, data = MyData5))


###################################
## Neyman Stork and Baby example ##
###################################

W  <- round(rnorm(10000, 40000, 10000), 0)
B  <- round(rnorm(10000, 30, 5), 0)
#B  <- rbinom(length(W), W, 0.1)
S  <- round(rnorm(10000, 5, 1))

SW <- S/W
BW <- B/W

```

```
252 cor.test(SW, BW)
253 cor.test(S, B)
254 summary(lm(B ~ S + W))
255
256 ############################################
257 ## Simulated Neyman example with constraint ##
258 ############################################
259
260 W  <- abs(round(rnorm(10000, 40000, 10000), 0))
261 B  <- rbinom(length(W), W, 0.1)
262 S  <- round(rnorm(10000, 5, 1))
263
264 SW <- S/W
265 BW <- B/W
266
267 cor.test(SW, BW)
268 cor.test(S, B)
269
270 summary(lm(B ~ S + W))
271
272 ##############################################################################
273 ## Investigating coefficient of variation and affect of changing path coefficents ##
274 ##############################################################################
275
276 library( dagitty )
277 library( ggplot2 )
278 library( dplyr )
279 library( ppcor )
280 library( tikzDevice )
281 library( boot )
```

```
282
283
284 sim.cv <- function( N, cv, distr="norm" ){
285
286   x <- rnorm( N, mean = 1/cv  )
287   return(x)
288
289 }
290
291
292 sim.data <- function( cvn, cvx, cvy, N, distr = "norm" ){
293
294   n <- sim.cv( N, cvn, distr=distr )
295   x <- sim.cv( N, cvx, distr=distr )
296   y <- sim.cv( N, cvy, distr=distr )
297
298   cor(x/n,y/n)
299
300 }
301
302 sim.cor.data <- function( cvn, cvx, cvy, N, cor = .1 ){
303
304   d <- simulateSEM( 'dag{ N -> {X Y} }', b.default = cor )
305   d$N <- d$N + 1/cvn
306   d$X <- d$X + 1/cvx
307   d$Y <- d$Y + 1/cvy
308
309   cor( d$X/d$N, d$Y/d$N )
310
311 }
```

```
312
313 # add mean afterwards.
314
315
316 cvn <- c( 0.1 )
317 b <- seq(0,0.8,by=0.2)
318 #cv.FC <- c( 1, 2, 5 )
319 cv.FC2 <- 2^seq(-2.5, 5, 0.25 )
320
321 cvdata <- expand.grid( b, cv.FC2 )
322 colnames(cvdata) <- c( "b", "cvFC" )
323 cvdata$cvn <- cvn
324 cvdata$cvx <- cvdata$cvn*cvdata$cvFC
325 nsim <- 10
326 N <- 1000
327
328 meancor <- numeric()
329 sdcor <- numeric()
330
331 for( i in 1:nrow(cvdata) ){
332
333   c <- replicate( nsim, sim.cor.data( cvdata$cvn[i], cvdata$cvx[i], cvdata$cvx[i], N, cvdata$b[i] ) )
334   # c <- abs(c)
335   meancor <- c( meancor, mean(c) )
336   sdcor <- c( sdcor, sd(c) )
337
338 }
339
340 cvdata$meancor <- meancor
341 cvdata$sdcor <- sdcor
```

```
cvdata$true <- cvdata$b^2


# CV equal
d2 <- cvdata %>% filter( log2(cvFC) == 0, b != 1 )

par(family = "LM Roman 10")
ggplot( d2, aes( x = b, y = meancor ) ) +
  geom_line() +
  geom_ribbon( aes( ymin = meancor-sdcor, ymax=meancor+sdcor), alpha = 0.3 ) +
  theme_light() +
  theme(axis.text = element_text(colour = "black", size = 12, family = "Times New Roman"), axis.title = element_text(size =
      16, family = "Times New Roman", face = "bold")) +
  labs( y = expression(paste("'Spurious' Correlation - cor"~bgroup("(",over("X","N")~","~over("Y","N"),")"))),
        x = expression("Path Coefficient (b) - N " %->% " X and N " %->% " Y"))

# b = 0
d3 <- cvdata %>% filter( b == 0 )
d4 <- d3[ d3$cvn <= d3$cvx , ]
d3 <- d3[ d3$cvn >= d3$cvx, ]

ggplot( d3, aes( x = log2(cvx/cvn), y = meancor ) ) +
  geom_line( color="red4") +
  geom_ribbon( fill="red4", aes( ymin = meancor-sdcor, ymax = meancor+sdcor ),alpha = 0.3, color=NA, show.legend=FALSE) +
  geom_line( data=d4, color = "steelblue4" ) +
  geom_ribbon(data=d4, aes( ymin = meancor-sdcor, ymax = meancor+sdcor ),alpha = 0.3, color=NA, fill = "steelblue4", show.
      legend=FALSE) +
  geom_hline( yintercept= 0 ) +
  geom_vline( xintercept = 0 ) +
  labs(  y = expression(paste("'Spurious' Correlation - cor"~bgroup("(",over("X","N")~","~over("Y","N"),")"))),
```

```
370         x = expression(paste("log"[2]~bgroup("(",over("cv(X)","cv(N)"),")")))) +
371   theme_light() +
372   theme( axis.line = element_blank(), axis.text = element_text(colour = "black", size = 12, family = "Times New Roman"),
373         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold") ) +
374   annotate( "text", label = "cv(N) > cv(X,Y)", color = "red4", x = -1, y = 1, family = "Times New Roman", size = 8) +
375   annotate( "text", label = "cv(N) < cv(X,Y)", color = "steelblue4", x = 1, y = 1, family = "Times New Roman", size = 8)
376
377 # 2D
378 ggplot( cvdata, aes( x = log2(cvx/cvn), y = meancor, group = b, color = b, fill = b ) ) +
379   geom_line() +
380   geom_ribbon( aes( ymin = meancor-sdcor, ymax = meancor+sdcor ),alpha = 0.3, color=NA) +
381   geom_hline( aes(yintercept=true, color = b ), lty = 2 ) +
382   geom_hline( yintercept= 0 ) +
383   geom_vline( xintercept = 0 ) +
384   labs(  y = expression(paste("'Spurious' Correlation - cor"~bgroup("(",over("X","N")~","~over("Y","N"),")"))),
385         x = expression(paste("log"[2]~bgroup("(",over("cv(X)","cv(N)"),")")))) +
386   theme_classic() + theme( axis.line = element_blank(), axis.text = element_text(colour = "black", size = 12, family = "Times
        New Roman"),
387                            axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
388                            legend.text = element_text(colour = "black", size = 12, family = "Times New Roman"),
389                            legend.title = element_text(colour = "black", size = 12, family = "Times New Roman", face = "bold"
                               )) +
390   annotate( "text", label = "cv(N) > cv(X,Y)", color = "dodgerblue4", x = -1, y = 1, family = "Times New Roman", size = 8) +
391   annotate( "text", label = "cv(N) < cv(X,Y)", color = "steelblue4", x = 1, y = 1, family = "Times New Roman", size = 8)
392
393 ggplot( test2, aes( x = log2(cvx/cvn), y = atanh(meancor), group = interaction(b,cvy), color = b, fill = b ) ) +
394   geom_line() +
395   #geom_ribbon( aes( ymin = meancor-sdcor, ymax = meancor+sdcor ),alpha = 0.3, color=NA) +
396   #geom_hline( aes(yintercept=true, color = b ), lty = 2 ) +
397   geom_hline( yintercept= 0 ) +
```

```
  geom_vline( xintercept = 0 ) +
  labs(  y = "atanh cor( x/N, y/N)" ) +
  theme_classic() + theme( axis.line = element_blank())


# Now vary both cvx and cvy
cvn <- c( 0.1 )
b <- seq(0,1.0,by=0.2)
#cv.FC <- c( 1, 2, 5 )
cv.FC2 <- 2^seq(-2.5, 2.5, 0.5 )
cvx <- cvn*cv.FC2
cvy <- cvn*cv.FC2


cvdata <- expand.grid( b, cvx, cvy )
colnames(cvdata) <- c( "b", "cvx","cvy" )
cvdata$cvn <- cvn
nsim <- 10
N <- 500

meancor <- numeric()
sdcor <- numeric()

for( i in 1:nrow(cvdata) ){

  c <- replicate( nsim, sim.cor.data( cvdata$cvn[i],
                                      cvdata$cvx[i],
                                      cvdata$cvy[i],
                                      N, cvdata$b[i] ) )
  meancor <- c( meancor, mean(c) )
```

```
  sdcor <- c( sdcor, sd(c) )

}

cvdata$meancor <- meancor
cvdata$sdcor <- sdcor
cvdata$true <- cvdata$b^2

cvdata$b2 <- paste0("b = ", cvdata$b)
ggplot( cvdata, aes( x = log2(cvx/cvn), y = log2(cvy/cvn), z = meancor ) ) +
  stat_summary_2d(bins = 10, fun = "mean" ) +
  facet_wrap( ~b2 ) +
  scale_x_continuous( expand=c(0,0) )+
  scale_y_continuous(expand=c(0,0) ) +
  theme_classic() +
  theme( axis.line = element_blank(),  axis.text = element_text(colour = "black", size = 12, family = "Times New Roman"),
         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
         legend.text = element_text(colour = "black", size = 12, family = "Times New Roman"),
         legend.title = element_text(colour = "black", size = 12, family = "Times New Roman", face = "bold"),
         strip.text.x = element_text(colour = "black", size = 12, family = "Times New Roman")) +
  labs(  y = expression(paste("log"[2]~bgroup("(",over("cv(Y)","cv(N)"),")"))),
         x = expression(paste("log"[2]~bgroup("(",over("cv(X)","cv(N)"),")"))),
         fill = expression(paste("'Spurious' Correlation - cor"~bgroup("(",over("X","N")~","~over("Y","N"),")"))))

################################################################################
################################################################################
## What happens when we use a binomial distribution instead of a normal distribution? ##
## i.e. Simulate areas with certain population sizes and for each person in the       ##
## population flip a coin and count all the tails - i.e. those who experience the     ##
## conditions X and Y.                                                                ##
```

```
458 ##############################################################################################
459 ##############################################################################################
460
461
462
463 meancor <- numeric()
464 sdcor <- numeric()
465
466 cv  <- function(x) {
467   sd(x)/mean(x)
468 }
469
470 nsim <- 100
471 dat  <- NULL
472
473 prseq <- rep(seq(0.01, 0.99, by = 0.01), nsim)
474 Prob  <- seq(0.1, 0.9, by = 0.1)
475
476 dat <- expand.grid( Prob, prseq)
477 colnames(dat) <- c("Prob", "prseq")
478
479 for (i in 1:length(dat[,1])){
480   N <- rnbinom( 1000, 1000, prob = dat$Prob[i] )
481   X <- sapply( N, function(x) rbinom( 1, x, prob = dat$prseq[i] ) )
482   Y <- sapply( N, function(x) rbinom( 1, x, prob = dat$prseq[i] ) )
483
484   df <- data.frame(X, Y, N)
485
486   #dat$NProb[i] <- Prob[j]
487   #dat$pr[i]  <- prseq[i]
```

```
  dat$cvx[i] <- cv(X)
  dat$cvy[i] <- cv(Y)
  dat$cvn[i] <- cv(N)
  dat$corxy[i] <- cor(X, Y)
  dat$corxn[i] <- cor(X, N)
  dat$corxyn[i] <- cor(X/N, Y/N)
  dat$pcorxy[i] <- pcor(df)$estimate[1, 2]
}

dat$cvxn <- dat$cvx/dat$cvn

meandat  <- aggregate(corxyn ~ Prob + prseq, data = dat, function(x) c(mean = mean(x)))
meancvxn <- aggregate(cvxn ~ Prob + prseq, data = dat, function(x) c(mean = mean(x)))
meandat2 <- aggregate(corxyn ~ cvxn + Prob, data = dat, function(x) c(mean = mean(x)))

summary(dat$pcorxy)

ggplot( meandat, aes( x = prseq, y = corxyn, group = Prob, color = Prob, fill = Prob ) ) +
  geom_line() +
  theme_light() +
  theme( axis.line = element_blank(), axis.text = element_text(colour = "black", size = 16, family = "Times New Roman"),
         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
         legend.title = element_text(size = 12, family = "Times New Roman", face = "bold"),
         legend.text = element_text(size = 12, family = "Times New Roman")) +
  labs(  y = expression(paste("'Spurious' Correlation - cor"~bgroup("(",over("X","N")~","~over("Y","N"),")"))),
         x = expression(paste("Success Probability")),
         color = "NegBin Prob")

ggplot( meancvxn, aes( x = prseq, y = log2(cvxn), group = Prob, color = Prob, fill = Prob ) ) +
  geom_line() +
```

```
  theme_light() +
  theme( axis.line = element_blank(), axis.text = element_text(colour = "black", size = 16, family = "Times New Roman"),
         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
         legend.title = element_text(size = 12, family = "Times New Roman", face = "bold"),
         legend.text = element_text(size = 12, family = "Times New Roman")) +
  labs(  y = expression(paste(bgroup("(",over("cv(X)","cv(N)"),")"))),
         x = expression(paste("Success Probability")),
         color = "NegBin Prob")

ggplot( meandat2, aes( x = cvxn, y = corxyn , group = Prob, color = Prob, fill = Prob ) ) +
  geom_line() +
  theme_light() +
  theme( axis.line = element_blank(), axis.text = element_text(colour = "black", size = 16, family = "Times New Roman"),
         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
         legend.title = element_text(size = 12, family = "Times New Roman", face = "bold"),
         legend.text = element_text(size = 12, family = "Times New Roman")) +
  labs(  y = expression(paste("'Spurious' Correlation - cor"~bgroup("(",over("X","N")~","~over("Y","N"),")"))),
         x = expression(paste(bgroup("(",over("cv(X)","cv(N)"),")"))),
         color = "NegBin Prob")
```

MC.R

Appendix C

Simulations of area–level data to investigate analyses of limiting long–term illness and deprivation

```
############################################################
## Simulations investigating LLTI and deprivation using   ##
## correlations, linear regression and Poisson regression ##
############################################################

ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}
```

```
12
13
14 packages <- c("MASS", "Matrix", "matrixcalc", "ggplot2", "tidyr", "reshape", "extrafont", "gridExtra",
15              "grid", "VGAM", "boot", "pscl", "SuppDists", "distr", "distrEx", "stringr", "epitools",
16              "scales","extrafontdb")
17 ipak(packages)
18
19 cbPal   <- c("#999999","#E69F00","#56B4E9","#009E73","#F0E442","#0072B2","#D55E00","#CC79A7")
20 Dark    <- c("#3333FF","#CC0000"); Lite <- c("#99CCFF","#FF9999")
21
22
23 Models  <- function(dataset, outcome, covariate, off, model, ...) {
24   # outcome  <- substitute(dataset$outcome)
25   # covariate  <- substitute(dataset$covariate)
26   # off  <- substitute(dataset$off)
27   if (model == "negativebinomial") {
28     if (is.null(off) == FALSE) {
29       if (length(covariate) == 1) {
30         Tmp  <- glm.nb(dataset[[outcome]] ~ dataset[[covariate]] + offset(log(dataset[[off]])), data = dataset, maxit =
               10000)
31         Err  <- try(confint(Tmp)[c(2, 4)])
32         if(isTRUE(class(Err)=="try-error")) {
33           c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8])
34         }
35         else {c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8])}
36       }
37       else if (length(covariate) == 2) {
38         Tmp  <- glm.nb(dataset[[outcome]] ~ dataset[[covariate[1]]] + dataset[[covariate[2]]]
39                        + offset(log(dataset[[off]])), data = dataset, maxit = 10000)
40         Err  <- try(confint(Tmp)[c(2, 5)])
```

```
41        if(isTRUE(class(Err)=="try-error")) {
42          c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 5)], summary(Tmp)$coefficients[11])
43          }
44        else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 5)], summary(Tmp)$coefficients[11])}
45      }
46      else if (length(covariate) == 4) {
47        Tmp  <- (glm.nb(dataset[[outcome]] ~ dataset[[covariate[1]]] + dataset[[covariate[2]]] + dataset[[covariate[3]]]
48                        + dataset[[covariate[4]]] + offset(log(dataset[[off]])), data = dataset, maxit = 10000))
49        Err  <- try(confint(Tmp)[c(2, 7)])
50        if(isTRUE(class(Err)=="try-error")) {
51          c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 7)],
52                                              summary(Tmp)$coefficients[17])
53          }
54        else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 7)], summary(Tmp)$coefficients[17]) }
55      }
56    }
57    else if (is.null(off) == TRUE) {
58      Tmp  <- glm.nb(dataset[[outcome]] ~ dataset[[covariate]], data = dataset, maxit = 10000)
59      Err  <- try(confint(Tmp)[c(2, 4)])
60      if(isTRUE(class(Err)=="try-error")) {
61        c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8])
62        }
63      else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8]) }
64    }
65  }
66  else if (model == "poisson") {
67    if (is.null(off) == FALSE) {
68      if (length(covariate) == 1) {
69        Tmp  <- glm(dataset[[outcome]] ~ dataset[[covariate]], offset = log(dataset[[off]]), family = "poisson", data =
              dataset, maxit = 10000)
```

```
70        Err  <- try(confint(Tmp)[c(2, 4)])
71        if(isTRUE(class(Err)=="try-error")) {
72          c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8])
73          }
74        else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8]) }
75      }
76      else if (length(covariate) == 2) {
77        Tmp  <- glm(dataset[[outcome]] ~ dataset[[covariate[1]]] + dataset[[covariate[2]]] + offset(log(dataset[[off]])),
78                    family = "poisson", data = dataset, maxit = 10000)
79        Err  <- try(confint(Tmp)[c(2, 5)])
80        if(isTRUE(class(Err)=="try-error")) {
81          c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 5)], summary(Tmp)$coefficients[11])
82          }
83        else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 5)], summary(Tmp)$coefficients[11]) }
84
85      }
86      else if (length(covariate) == 4) {
87        Tmp  <- glm(dataset[[outcome]] ~ dataset[[covariate[1]]] + dataset[[covariate[2]]] + dataset[[covariate[3]]]
88                    + dataset[[covariate[4]]] + offset(log(dataset[[off]])), family = "poisson", data = dataset, maxit =
                        10000)
89        Err  <- try(confint(Tmp)[c(2, 7)])
90        if(isTRUE(class(Err)=="try-error")) {
91          c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 7)], summary(Tmp)$coefficients[11])
92        }
93        else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 7)], summary(Tmp)$coefficients[11]) }
94      }
95
96    }
97    else if (is.null(off) == TRUE) {
98      Tmp  <- glm(dataset[[outcome]] ~ dataset[[covariate]], family = "poisson", data = dataset, maxit = 10000)
```

```
    Err  <- try(confint(Tmp)[c(2, 4)])
    if(isTRUE(class(Err)=="try-error")) {
      c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8])
      }
    else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8]) }
  }
}
else if (model == "linear") {
  if (is.null(off) == FALSE) {
    if (length(covariate) == 1) {
      Tmp  <- lm(dataset[[outcome]] ~ dataset[[covariate]] + dataset[[off]], data = dataset)
      Err  <- try(confint(Tmp)[c(2, 5)])
      if(isTRUE(class(Err)=="try-error")) {
        c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 5)], summary(Tmp)$coefficients[11])
        }
      else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 5)], summary(Tmp)$coefficients[11]) }
      # confint(Tmp[c(2,5)]) because confidence intervals for Pop are also generated, in other models it is an offset
    }
    else if (length(covariate) == 2) {
      Tmp  <- lm(dataset[[outcome]] ~ dataset[[covariate[1]]] + dataset[[covariate[2]]] + dataset[[off]], data = dataset)
      Err  <- try(confint(Tmp)[c(2, 6)])
      if(isTRUE(class(Err)=="try-error")) {
        c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 6)], summary(Tmp)$coefficients[14])
        }
      else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 6)], summary(Tmp)$coefficients[14]) }
    }
    else if (length(covariate) == 4) {
      Tmp <- lm(dataset[[outcome]] ~ dataset[[covariate[1]]] + dataset[[covariate[2]]] + dataset[[covariate[3]]] +
                  dataset[[covariate[4]]] + dataset[[off]], data = dataset)
      Err <- try(confint(Tmp)[c(2, 8)])
```

```
129         if(isTRUE(class(Err)=="try-error")) {
130           c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 8)], summary(Tmp)$coefficients[20])
131           }
132         else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 8)], summary(Tmp)$coefficients[20]) }
133       }
134     }
135     else if (is.null(off) == TRUE) {
136       Tmp  <- lm(dataset[[outcome]] ~ dataset[[covariate]], data = dataset)
137       Err  <- try(confint(Tmp)[c(2, 4)])
138       if(isTRUE(class(Err)=="try-error")) {
139         c(summary(Tmp)$coefficients[2], confint.default(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8])
140       }
141       else { c(summary(Tmp)$coefficients[2], confint(Tmp)[c(2, 4)], summary(Tmp)$coefficients[8]) }
142     }
143   }
144 }
145
146 ## Read in data
147 LLTI_Covariates <- read.csv("LLTIMortTownsend.csv", sep = ",")
148 LLTI_Covariates[,4:21] <- sapply(LLTI_Covariates[,4:21], as.integer)
149
150 ## Fit lognormal distributions to the data
151
152 LLTI_Covariates_0 <- LLTI_Covariates[,4:21]
153 LLTI_Covariates_0[LLTI_Covariates_0 == 0] <- 1
154
155
156 head(LLTI_Covariates_0)
157
```

```
158 LLTI_Covariates_0  <- data.frame(cbind(Nonown = LLTI_Covariates_0$Nonown_n, Own = LLTI_Covariates_0$Nonown_d - LLTI_
        Covariates_0$Nonown_n,
159                                    NoCar = LLTI_Covariates_0$Nocar_n, Car = LLTI_Covariates_0$Nocar_d - LLTI_Covariates_0
                                           $Nocar_n,
160                                    Overcr = LLTI_Covariates_0$Ovcr_n, NotOvercr = LLTI_Covariates_0$Ovcr_d - LLTI_
                                           Covariates_0$Ovcr_n,
161                                    Unemp = LLTI_Covariates_0$Unemp_n, Emp = LLTI_Covariates_0$Unemp_d - LLTI_Covariates_0
                                           $Unemp_n,
162                                    LLTI = LLTI_Covariates_0$llti.t))
163
164 DFn <- length(LLTI_Covariates_0[,1])
165
166
167 Parameters  <- matrix(nrow = length(LLTI_Covariates_0[1,]), ncol = 3)
168 for (i in 1:length(LLTI_Covariates_0[1, ])) {
169
170   tmp     <- fitdistr(LLTI_Covariates_0[, i], "negative binomial"); Pars <- tmp$estimate
171   Dat     <- data.frame(NB = round(rnbinom(DFn, size = Pars[1], mu = Pars[2])))
172   tmp     <- fitdistr(LLTI_Covariates_0[, i], "lognormal"); Pars <- tmp$estimate
173   Dat     <- data.frame(cbind(Dat, LN = round(rlnorm(DFn, meanlog = Pars[1], sdlog = Pars[2]))))
174
175
176   #Parameters[i,1]  <- Pars[1]
177   #Parameters[i,2]  <- Pars[2]
178   #Parameters[i,3]  <- colnames(LLTI_Covariates_0[i])
179
180   tmpData <- data.frame(cbind(ID = 1:DFn, Observed = LLTI_Covariates_0[,i], "Negative Binomial" = Dat$NB, "Log Normal" = Dat$
          LN))
181
```

```
182   Variable  <- cbind("Nonowner-occupied Households", "Owner-occupied Households", "Households without a Car", "Households
          with a Car",
183                    "Overcrowded Households", "Not Overcrowded Households", "Unemployed Population", "Employed Population",
184                    "Population with LLTI")
185
186   #apply(tmpData[,-1],2,summary)
187   #apply(tmpData[,-1],2,sum)
188   dd      <- melt(tmpData, id = c("ID")); names(dd) <- c("ID", "Dist", "Variable")
189   Xlim    <- c(0, max(dd[3])); Xlab <- Variable[i]; Ylab <- "Kernel Density"
190   Mlab    <- Variable
191   print(ggplot(dd) + geom_density(aes(x = Variable, group = Dist, colour = Dist), size = 1.2) + labs(x = Xlab, y = Ylab,
          colour = NULL) +
192           coord_cartesian(xlim = Xlim, ylim = NULL) +
193           theme_light() +
194           theme( axis.line = element_blank(),
195                  axis.text = element_text(colour = "black", size = 16, family = "Times New Roman"),
196                  axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
197                  legend.title = element_text(size = 12, family = "Times New Roman", face = "bold"),
198                  legend.text = element_text(size = 12, family = "Times New Roman"),
199                  legend.position = "right") + scale_colour_manual(values = c("black", cbPal[-1]), name = "Distribution",
200                                                                   labels = c("Observed", "Negative Binomial", "Log Normal")))
201   ggsave(filename = paste(Variable[i]," - 190318.png"), width = 10, height = 6, units = "in", dpi = 300)
202   ##
203
204 }
205
206
207 ############################################################
208 ## GenData Function (adapted from Ruscio & Kaczetow, 2008) ##
209 ############################################################
```

```
GenData <- function(Supplied.Data = NULL, rho, N = 1000, N.Factors = 0, Max.Trials = 5, Initial.Multiplier = 1, seed = NA, k = NULL)
{
  # Initialize variables and (if applicable) set random number seed (step 1) ------------------------------------

  # k <- length(Pop)
  Data <- matrix(0, nrow = N, ncol = k)            # Matrix to store the simulated data
  Iteration <- 0                                   # Iteration counter
  Best.RMSR <- 1                                   # Lowest RMSR correlation
  Trials.Without.Improvement <- 0                  # Trial counter
  if (!is.na(seed)) set.seed(seed)                 # If user specified a nonzero seed, set it
  Distributions <- matrix(NA, nrow = N, ncol = k)

  # Target.Corr <- matrix(c(1, rho, rho, 1), nrow=2)

  # Generate distribution for each variable (step 2) ------------------------------------------------------------

  Distributions[, 1] <- sort(rlnorm(N, 7.51, 0.73))  # Employed Economically Active Population
  Distributions[, 2] <- sort(rlnorm(N, 4.93, 1.05))  # Unemployed Population
  Distributions[, 3] <- sort(rlnorm(N, 6.00, 1.01))  # Households non-owner occupied
  Distributions[, 4] <- sort(rlnorm(N, 7.00, 0.75))  # Households owner occupied
  Distributions[, 5] <- sort(rlnorm(N, 5.91, 1.16))  # Households without a car
  Distributions[, 6] <- sort(rlnorm(N, 7.05, 0.67))  # Households with a car
  Distributions[, 7] <- sort(rlnorm(N, 3.00, 1.27))  # Households overcrowded
  Distributions[, 8] <- sort(rlnorm(N, 7.37, 0.75))  # Households not overcrowded
  Distributions[, 9] <- sort(rlnorm(N, 9.22, 1.00))  # Total Population - changed from mean of 8.22 to make sure Pop is bigger than EA, spread originally 1.22


```

```
238 # for (i in 1:k) {
239 #   Distributions[, i] <- sort(sample(Pop[[i]], N, replace = TRUE))
240 # }
241
242 #     This implementation of GenData bootstraps each variable's score distribution from a supplied data set.
243 #     Users should modify this block of the program, as needed, to generate the desired distribution(s).
244 #
245 #     For example, to sample from chi-square distributions with 2 df, replace the 2nd line in this block with:
246 #        Distributions[,i] <- sort(rchisq(N, df = 2))
247 #
248 #     Or, one can drop the loop and use a series of commands that samples variables from specified populations:
249 #        Distributions[,1] <- sort(rnorm(N, 0, 1))          # Standard normal distribution
250 #        Distributions[,2] <- sort(runif(N, 0, 1))          # Uniform distribution ranging from 0 - 1
251 #        Distributions[,3] <- sort(rlnorm(N, 0, 1))         # Log-normal distribution, log scale M = 0, SD = 1
252 #        Distributions[,4] <- sort(rexp(N, rate = 1))       # Exponential distribution with rate = 1
253 #        Distributions[,5] <- sort(rpois(N, lambda = 4))    # Poisson distribution with lambda = 4
254 #        Distributions[,6] <- sort(rbinom(N, 10, .25)       # Binominal distribution, size = 10 and p = .25
255 #        Distributions[,7] <- sort(rbinom(N, 2, .25)        # Binary distribution with p = .25
256 #
257 #     All of the commands shown above draw random samples from specified population distributions.  As an
258 #     alternative, one can reproduce distributions without sampling error.  For example, working with a
259 #     supplied data set, one can replace the 2nd line in this block with:
260 #        Disrributions[,i] <- Supplied.Data[,i]
261 #     Alternatively, idealized distributions can be reproduced.  For example, uniform quantiles can be
262 #     created and used to generate data from common distributions:
263 #        Uniform.Quantiles <- seq(from = 0, to = 1, length = (N + 2))[2:(N + 1)]   # quantiles 0, 1 dropped
264 #        Distributions[,1] <- qnorm(Uniform.Quantiles, 0, 1)      # Standard normal distribution
265 #        Distributions[,2] <- qunif(Uniform.Quantiles, 0, 1)      # Uniform distribution ranging from 0 to 1
266 #        Distributions[,3] <- qchisq(Uniform.Quantiles, df = 2)   # Chi-square distribution with 2 df
267 #
```

```
#     Note that when score distributions are generated from specified populations rather than bootstrapped from
#     a supplied data set, the user must provide the target correlation matrix (see the next block).  This is
#     true regardless of whether the distributions incorporate sampling error.

# Calculate and store a copy of the target correlation matrix (step 3) ----------------------------------------

Target.Corr <- cor(Supplied.Data)
Intermediate.Corr <- Target.Corr

#     This implementation of GenData calculates the target correlation matrix from a supplied data set.
#     Alternatively, the user can modify the program to generate data with user-defined sample size, number of
#     variables, and target correlation matrix by redefining the function as follows:
#     GenData <- function(N, k, Target.Corr, N.Factors = 0, Max.Trials = 5, Initial.Multiplier = 1, seed = 0)
#     In this case, one would also remove the program lines that calculate N, k, and Target.Corr.
#     To generate data in which variables are uncorrelated, one would remove the sort function from step 2
#     and terminate the program before step 3 begins by returning the Distributions object as the data set.

# If number of latent factors was not specified, determine it through parallel analysis (step 4) --------------

if (N.Factors == 0)
{
  Eigenvalues.Observed <- eigen(Intermediate.Corr)$values
  Eigenvalues.Random <- matrix(0, nrow = 100, ncol = k)
  Random.Data <- matrix(0, nrow = N, ncol = k)
  for (i in 1:100)
  {
    for (j in 1:k)
      Random.Data[,j] <- sample(Distributions[,j], size = N, replace = TRUE)
    Eigenvalues.Random[i,] <- eigen(cor(Random.Data))$values
  }
```

```
298   Eigenvalues.Random <- apply(Eigenvalues.Random, 2, mean) # calculate mean eigenvalue for each factor
299   N.Factors <- max(1, sum(Eigenvalues.Observed > Eigenvalues.Random))
300 }
301
302 # Generate random normal data for shared and unique components, initialize factor loadings (steps 5, 6) --------
303
304 Shared.Comp <- matrix(rnorm(N * N.Factors, 0, 1), nrow = N, ncol = N.Factors)
305 Unique.Comp <- matrix(rnorm(N * k, 0, 1), nrow = N, ncol = k)
306 Shared.Load <- matrix(0, nrow = k, ncol = N.Factors)
307 Unique.Load <- matrix(0, nrow = k, ncol = 1)
308
309 # Begin loop that ends when specified number of iterations pass without improvement in RMSR correlation --------
310
311 while (Trials.Without.Improvement < Max.Trials)
312 {
313   Iteration <- Iteration + 1
314
315   # Calculate factor loadings and apply to reproduce desired correlations (steps 7, 8) --------------------------
316
317   Fact.Anal <- Factor.Analysis(Intermediate.Corr, Corr.Matrix = TRUE, N.Factors = N.Factors)
318   if (N.Factors == 1) {
319     Shared.Load[,1] <- Fact.Anal$loadings
320   } else {
321     Shared.Load <- Fact.Anal$loadings
322   }
323   Shared.Load[Shared.Load > 1] <- 1
324   Shared.Load[Shared.Load < -1] <- -1
325   if (Shared.Load[1,1] < 0) Shared.Load <- Shared.Load * -1
326   Shared.Load.sq <- Shared.Load * Shared.Load
327   for (i in 1:k)
```

```
    if (sum(Shared.Load.sq[i,]) < 1) {
      Unique.Load[i,1] <- (1 - sum(Shared.Load.sq[i,]))
    } else {
      Unique.Load[i,1] <- 0
    }
  Unique.Load <- sqrt(Unique.Load)

  for (i in 1:k) {
    Data[,i] <- (Shared.Comp %*% t(Shared.Load))[,i] + Unique.Comp[,i] * Unique.Load[i,1]
  }
  # the %*% operator = matrix multiplication, and the t() function = transpose (both used again in step 13)

  # Replace normal with nonnormal distributions (step 9) ------------------------------------------------------

  for (i in 1:k)
  {
    Data <- Data[sort.list(Data[,i]),]
    Data[,i] <- Distributions[,i]
  }

  # Calculate RMSR correlation, compare to lowest value, take appropriate action (steps 10, 11, 12) -------------

  Reproduced.Corr <- cor(Data)
  Residual.Corr <- Target.Corr - Reproduced.Corr
  RMSR <- sqrt(sum(Residual.Corr[lower.tri(Residual.Corr)] * Residual.Corr[lower.tri(Residual.Corr)]) /
                 (.5 * (k * k - k)))
  if (RMSR < Best.RMSR) {
    Best.RMSR <- RMSR
    Best.Corr <- Intermediate.Corr
    Best.Res <- Residual.Corr
```

```
358       Intermediate.Corr <- Intermediate.Corr + Initial.Multiplier * Residual.Corr
359       Trials.Without.Improvement <- 0
360     } else {
361       Trials.Without.Improvement <- Trials.Without.Improvement + 1
362       Current.Multiplier <- Initial.Multiplier * .5 ^ Trials.Without.Improvement
363       Intermediate.Corr <- Best.Corr + Current.Multiplier * Best.Res
364     }
365   }  # end of the while loop
366
367   # Construct the data set with the lowest RMSR correlation (step 13) ------------------------------------------
368
369   Fact.Anal <- Factor.Analysis(Best.Corr, Corr.Matrix = TRUE, N.Factors = N.Factors)
370   if (N.Factors == 1) {
371     Shared.Load[,1] <- Fact.Anal$loadings
372   } else {
373     Shared.Load <- Fact.Anal$loadings
374   }
375   Shared.Load[Shared.Load > 1] <- 1
376   Shared.Load[Shared.Load < -1] <- -1
377   if (Shared.Load[1,1] < 0) {Shared.Load <- Shared.Load * -1}
378   Shared.Load.sq <- Shared.Load * Shared.Load
379   for (i in 1:k) {
380     if (sum(Shared.Load.sq[i,]) < 1) {
381       Unique.Load[i,1] <- (1 - sum(Shared.Load.sq[i,]))
382     } else {
383       Unique.Load[i,1] <- 0
384     }
385   }
386   Unique.Load <- sqrt(Unique.Load)
387   for (i in 1:k) {
```

```
    Data[,i] <- (Shared.Comp %*% t(Shared.Load))[,i] + Unique.Comp[,i] * Unique.Load[i,1]
  }
  Data <- apply(Data, 2, scale) # standardizes each variable in the matrix
  for (i in 1:k)
  {
    Data <- Data[sort.list(Data[,i]),]
    Data[,i] <- Distributions[,i]
  }
  Data <- Data[sample(1:N, N, replace = FALSE), ] # randomize order of cases

  # Report the results and return the simulated data set (step 14) ---------------------------------------------

  Iteration <- Iteration - Max.Trials
  cat("\nN =",N,", k =",k,",",Iteration,"Iterations,",N.Factors,"Factors, RMSR r =",round(Best.RMSR,3),"\n")
  return(Data)
}

#########################################################################################################################
Factor.Analysis <- function(Data, Corr.Matrix = FALSE, Max.Iter = 50, N.Factors = 0)
{
  Data <- as.matrix(Data)
  k <- dim(Data)[2]
  if (N.Factors == 0) N.Factors <- k
  if (!Corr.Matrix) Cor.Matrix <- cor(Data)
  else Cor.Matrix <- Data
  Criterion <- .001
  Old.H2 <- rep(99, k)
  H2 <- rep(0, k)
  Change <- 1
  Iter <- 0
```

```
418   Factor.Loadings <- matrix(nrow = k, ncol = N.Factors)
419   while ((Change >= Criterion) & (Iter < Max.Iter))
420   {
421     Iter <- Iter + 1
422     Eig <- eigen(Cor.Matrix)
423     L <- sqrt(Eig$values[1:N.Factors])
424     for (i in 1:N.Factors)
425       Factor.Loadings[,i] <- Eig$vectors[,i] * L[i]
426     for (i in 1:k)
427       H2[i] <- sum(Factor.Loadings[i,] * Factor.Loadings[i,])
428     Change <- max(abs(Old.H2 - H2))
429     Old.H2 <- H2
430     diag(Cor.Matrix) <- H2
431   }
432   if (N.Factors == k) N.Factors <- sum(Eig$values > 1)
433   return(list(loadings = Factor.Loadings[,1:N.Factors], factors = N.Factors))
434 }
435
436
437 ##############################################################################################################################
438 ##############################################################################################################################
439
440 ## Read in data
441 LLTI_Covariates <- read.csv("LLTIMortTownsend.csv", sep = ",")
442 LLTI_Covariates[,4:21] <- sapply(LLTI_Covariates[,4:21], as.integer)
443
444 ## Replace 0 with 1
445 LLTI_Covariates_0 <- LLTI_Covariates[,4:21]
446 LLTI_Covariates_0[LLTI_Covariates_0 == 0] <- 1
447
```

```
## Create data frame of covariates
LLTI_Dataset  <- cbind(Emp = LLTI_Covariates_0$Unemp_d - LLTI_Covariates_0$Unemp_n, Unemp = LLTI_Covariates_0$Unemp_n,
                           Nonowner = LLTI_Covariates_0$Nonown_n, Owner = LLTI_Covariates_0$Nonown_d - LLTI_Covariates_0$
                               Nonown_n,
                           NoCar = LLTI_Covariates_0$Nocar_n, Car = LLTI_Covariates_0$Nocar_d - LLTI_Covariates_0$Nocar_n,
                           Overcrowd = LLTI_Covariates_0$Ovcr_n, NotOvercrowd = LLTI_Covariates_0$Ovcr_d - LLTI_Covariates_0
                               $Ovcr_n,
                           Pop = LLTI_Covariates_0$Persons, LLTI = LLTI_Covariates_0$llti.t)


ObsCor <- cor(LLTI_Dataset[,1:9])

#########################################################
## Function to generate lognormal distributed cases  ##
#########################################################

NullSim   <- function(N, id, x, sizex, setPRTot){
  #y      <- rpois(N, x*setPRTot)
  #y       <- rbinom(N, x, setPRTot)
  y       <- round(rnbinom(N, mu = setPRTot*x, size = sizex), 0)
  Dat     <- data.frame(Id = id, Pop = x, Obs = y)
  PR      <- sum(Dat$Obs)/sum(Dat$Pop)
  Dat     <- cbind(Dat, Exp = round(Dat$Pop*PR, 0))
  return(Dat)
}


Beg   <- Sys.time()
Nsim  <- 1000

```

```
476 repmin <- function(simcol, obscol){
477   simcol <- replace(simcol, simcol[which(simcol < min(obscol))], sample(min(obscol):median(obscol), 1, replace = TRUE))
478 }
479
480 repmax  <- function(simcol, obscol){
481   simcol  <- replace(simcol, simcol[which(simcol > max(obscol))], sample(median(obscol):max(obscol), 1, replace = TRUE))
482 }
483
484 # Dataframe of 9 variables to generate
485 LLTI_Dat <- data.matrix(LLTI_Dataset[,1:9])
486
487 # Create empty vectors
488 NullObj <- function(obnames){
489     for (q in 1:length(obnames)) {
490       nam1  <- paste("c", obnames[q], sep = "")
491       nam2  <- paste("p", obnames[q], sep = "")
492       nam3  <- paste("pLin", obnames[q], sep = "")
493       nam4  <- paste("cLin", obnames[q], sep = "")
494       nam5  <- paste("pLin", obnames[q], "Pop", sep = "")
495       nam6  <- paste("cLin", obnames[q], "Pop", sep = "")
496       assign(nam1, NULL, envir = .GlobalEnv); assign(nam2, NULL, envir = .GlobalEnv)
497       assign(nam3, NULL, envir = .GlobalEnv); assign(nam4, NULL, envir = .GlobalEnv)
498       assign(nam5, NULL, envir = .GlobalEnv); assign(nam6, NULL, envir = .GlobalEnv)
499     }
500
501
502 }
503
504 NullObj(c("Townsend", "UnempPC", "UnempCount", "EmpPC", "EmpCount", "NoCarPC", "NoCarCount",
505         "CarPC", "CarCount","NonOwnPC", "NonOwnCount", "OwnPC", "OwnCount", "OvercrowdPC", "OvercrowdCount",
```

```
        "NotOvercrowdPC", "NotOvercrowdCount", "", "Adj", "Unemp", "Emp", "Car", "NoCar", "NonOwn", "Own",
        "Overcrowd", "NotOvercrowd"))

cTownsendCov <- NULL; pTownsendCov  <- NULL
cUnempCov  <- NULL; pUnempCov  <- NULL

PoisTownsendUnadj   <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisTownsend        <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisUnempCount      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisUnempPC         <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisNoCarCount      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisNoCarPC         <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisOvercrowdCount  <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisOvercrowdPC     <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisNonOwnCount     <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisNonOwnPC        <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))

PoisUnempMSAS       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisNoCarMSAS       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisOvercrowdMSAS   <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
PoisNonOwnMSAS      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))

NegBinTownsendUnadj   <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinTownsend        <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinUnempCount      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinUnempPC         <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinNoCarCount      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinNoCarPC         <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinOvercrowdCount  <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
NegBinOvercrowdPC     <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
```

```
536 NegBinNonOwnCount     <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
537 NegBinNonOwnPC        <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
538
539 NegBinUnempMSAS       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
540 NegBinNoCarMSAS       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
541 NegBinOvercrowdMSAS   <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
542 NegBinNonOwnMSAS      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
543
544 LinTownsendUnadj    <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
545 LinTownsend         <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
546 LinUnempCount       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
547 LinUnempPC          <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
548 LinNoCarCount       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
549 LinNoCarPC          <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
550 LinOvercrowdCount   <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
551 LinOvercrowdPC      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
552 LinNonOwnCount      <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
553 LinNonOwnPC         <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
554
555 LinUnempMSAS        <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
556 LinNoCarMSAS        <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
557 LinOvercrowdMSAS    <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
558 LinNonOwnMSAS       <- data.matrix(data.matrix(matrix(ncol = 4, nrow = Nsim)))
559
560 for (itn in 1:Nsim) {
561   GenSim <- data.matrix(GenData(Supplied.Data = LLTI_Dat, rho, N = length(LLTI_Dat[, 1]), k = length(LLTI_Dat[1, ]),seed = (
        itn*length(LLTI_Dat[, 1]))))
562   GenSim <- cbind(Id = 1:length(LLTI_Dat[, 1]), round(GenSim,0));
563   colnames(GenSim) <- paste(c("ID", "Emp", "Unemp", "Nonowner", "Owner", "NoCar", "Car", "Overcrowd", "NotOvercrowd", "Pop"))
564
```

```
565  # Replace zeros and maximum values
566
567  repmin(GenSim[,"Car"], LLTI_Dataset[,"Car"])
568  repmin(GenSim[,"NotOvercrowd"], LLTI_Dataset[,"NotOvercrowd"])
569  repmin(GenSim[,"Owner"], LLTI_Dataset[,"Owner"])
570  repmin(GenSim[,"NoCar"], LLTI_Dataset[,"NoCar"])
571  repmin(GenSim[,"Overcrowd"], LLTI_Dataset[,"Overcrowd"])
572  repmin(GenSim[,"Nonowner"], LLTI_Dataset[,"Nonowner"])
573  repmin(GenSim[,"Emp"], LLTI_Dataset[,"Emp"])
574  repmin(GenSim[,"Unemp"], LLTI_Dataset[,"Unemp"])
575  repmin(GenSim[,"Pop"], LLTI_Dataset[,"Pop"])
576
577  repmax(GenSim[,"Car"], LLTI_Dataset[,"Car"])
578  repmax(GenSim[,"NotOvercrowd"], LLTI_Dataset[,"NotOvercrowd"])
579  repmax(GenSim[,"Owner"], LLTI_Dataset[,"Owner"])
580  repmax(GenSim[,"NoCar"], LLTI_Dataset[,"NoCar"])
581  repmax(GenSim[,"Overcrowd"], LLTI_Dataset[,"Overcrowd"])
582  repmax(GenSim[,"Nonowner"], LLTI_Dataset[,"Nonowner"])
583  repmax(GenSim[,"Emp"], LLTI_Dataset[,"Emp"])
584  repmax(GenSim[,"Unemp"], LLTI_Dataset[,"Unemp"])
585  repmax(GenSim[,"Pop"], LLTI_Dataset[,"Pop"])
586
587  ## Force number of households to be the same across car, overcrowd and owner-occupier variables
588  MeanHH <- NULL
589  VecOwn <- data.matrix(matrix(ncol = 2, nrow = 9499))
590  VecCar <- data.matrix(matrix(ncol = 2, nrow = 9499))
591  VecOvr <- data.matrix(matrix(ncol = 2, nrow = 9499))
592  for (i in 1:9499) {
593    MeanHH[i] <- (sum(GenSim[,"Nonowner"][i], GenSim[,"Owner"][i]) +
594                      sum(GenSim[,"NoCar"][i], GenSim[,"Car"][i]) +
```

```
595                     sum(GenSim[,"Overcrowd"][i], GenSim[,"NotOvercrowd"][i]))/3
596
597
598   #VecOwn[i,]  <- round((MeanHH[i]/sum(GenSim[,"Nonowner"][i], GenSim[,"Owner"][i]))*data.matrix(GenSim[,"Nonowner"][i],
          GenSim[,"Owner"][i]), 0)
599   VecOwn[i, 1]  <- (sum(MeanHH[i])/sum(GenSim[,"Nonowner"][i], GenSim[,"Owner"][i]))*(GenSim[,"Nonowner"][i])
600   VecOwn[i, 2]  <- (sum(MeanHH[i])/sum(GenSim[,"Nonowner"][i], GenSim[,"Owner"][i]))*(GenSim[,"Owner"][i])
601   #
602   #VecCar[i,]  <- round((MeanHH[i]/sum(GenSim[,"NoCar"][i], GenSim[,"Car"][i]))*data.matrix(GenSim[,"NoCar"][i], GenSim[,"
          Car"][i]), 0)
603   VecCar[i, 1]  <- (sum(MeanHH[i])/sum(GenSim[,"NoCar"][i], GenSim[,"Car"][i]))*(GenSim[,"NoCar"][i])
604   VecCar[i, 2]  <- (sum(MeanHH[i])/sum(GenSim[,"NoCar"][i], GenSim[,"Car"][i]))*(GenSim[,"Car"][i])
605
606   #VecOvr[i,]  <- round((MeanHH[i]/sum(GenSim[,"Overcrowd"][i], GenSim[,"NotOvercrowd"][i]))*data.matrix(GenSim[,"Overcrowd
          "][i], GenSim[,"NotOvercrowd"][i]), 0)
607   VecOvr[i, 1]  <- (sum(MeanHH[i])/sum(GenSim[,"Overcrowd"][i], GenSim[,"NotOvercrowd"][i]))*(GenSim[,"Overcrowd"][i])
608   VecOvr[i, 2]  <- (sum(MeanHH[i])/sum(GenSim[,"Overcrowd"][i], GenSim[,"NotOvercrowd"][i]))*(GenSim[,"NotOvercrowd"][i])
609
610   #
611 }
612
613 VecOwn <- round(VecOwn, 0)
614 VecCar <- round(VecCar, 0)
615 VecOvr <- round(VecOvr, 0)
616
617 # Make data frame of relevant variables
618 GenSim2  <- data.frame(HH = VecOwn[, 1] + VecOwn[, 2], NoCar = VecCar[, 1], Car = VecCar[, 2], Overcrowd = VecOvr[, 1],
        NotOvercrowd = VecOvr[, 2],
619                        Nonowner = VecOwn[, 1], Owner = VecOwn[, 2], Unemp = GenSim[,"Unemp"], Emp = GenSim[,"Emp"], EA =
                           GenSim[,"Emp"] + GenSim[,"Unemp"], Pop = GenSim[,"Pop"])
```

```
620
621 GenSim3  <- cbind(GenSim2, NoCarPC = GenSim2[,"NoCar"]/GenSim2[,"HH"], CarPC = GenSim2[,"Car"]/GenSim2[,"HH"], OvercrowdPC
        = GenSim2[,"Overcrowd"]/GenSim2[,"HH"],
622                   NotOvercrowdPC = GenSim2[,"NotOvercrowd"]/GenSim2[,"HH"], NonOwnerPC = GenSim2[,"Nonowner"]/GenSim2[,"HH"
                          ], OwnerPC = GenSim2[,"Owner"]/GenSim2[,"HH"],
623                   UnempPC = GenSim2[,"Unemp"]/GenSim2[,"EA"], EmpPC = GenSim2[,"Emp"]/GenSim2[,"EA"])
624
625
626 # Calculate Townsend score
627
628 NoCarPC     <- (GenSim3[,"NoCar"]/GenSim3[,"HH"])*100
629 OvercrowdPC <- (GenSim3[,"Overcrowd"]/GenSim3[,"HH"])*100
630 NonOwnPC    <- (GenSim3[,"Nonowner"]/GenSim3[,"HH"])*100
631 UnemplPC    <- (GenSim3[,"Unemp"]/GenSim3[,"EA"])*100
632
633 NoCarZ     <- scale(NoCarPC, center = TRUE, scale = TRUE)
634 OvercrowdZ <- scale(OvercrowdPC, center = TRUE, scale = TRUE)
635 NonOwnZ    <- scale(NonOwnPC, center = TRUE, scale = TRUE)
636 UnemplZ    <- scale(UnemplPC, center = TRUE, scale = TRUE)
637
638 TownsendZ   <- NoCarZ + OvercrowdZ + NonOwnZ + UnemplZ
639
640 GenSim3     <- cbind(GenSim3, Townsend = TownsendZ)
641 GenSim3  <- as.matrix(GenSim3)
642
643 ## Simulate cases under null hypothesis
644 LLTI  <-  NullSim(N = 9499, id = 1:9499, x = (GenSim3[,"Pop"]), sizex = 1.5, setPRTot = 0.135)
645 # Alter sd.x in NullSim
646
647 #LLTI  <-  NullSim(N = 9499, id = 1:9499, x = (GenSim3[,"Pop"]), setPRTot = 0.135)
```

```
648
649 GenSim3 <- cbind(GenSim3, LLTIObs = LLTI[,"Obs"], LLTIExp = LLTI[,"Exp"])
650 GenSim3  <- as.matrix(GenSim3)
651
652 repmin(GenSim3[,"LLTIObs"], LLTI_Dataset[,"LLTI"])
653 repmax(GenSim3[,"LLTIObs"], LLTI_Dataset[,"LLTI"])
654
655 LLTIPC      <- (GenSim3[,"LLTIObs"]/GenSim3[,"Pop"])*100
656 # if (min(GenSim3$LLTIObs) <= min(LLTI_Dataset$LLTI)) GenSim3$LLTIObs[which(GenSim3$LLTIObs <= min(LLTI_Dataset$LLTI))] <-
        sample(min(LLTI_Dataset$LLTI):median(LLTI_Dataset$LLTI), 1, replace = TRUE)
657 # if (max(GenSim3$LLTIObs) >= max(LLTI_Dataset$LLTI)) GenSim3$LLTIObs[which(GenSim3$LLTIObs >= max(LLTI_Dataset$LLTI))] <-
        sample(median(LLTI_Dataset$LLTI):max(LLTI_Dataset$LLTI), 1, replace = TRUE)
658 #
659 GenSim3  <- as.data.frame(GenSim3)
660 # Investigating Observed counts of LLTI with a population offset
661
662 PoisTownsendUnadj[itn, ]  <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "Townsend", off = NULL, model = "
        poisson")
663 PoisTownsend[itn, ]       <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "Townsend", off = "Pop", model = "
        poisson")
664 PoisUnempCount[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "Unemp", off = "Pop", model = "
        poisson")
665 PoisUnempPC[itn, ]        <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "UnempPC", off = NULL, model = "
        poisson")
666 PoisNoCarCount[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "NoCar", off = "Pop", model = "
        poisson")
667 PoisNoCarPC[itn, ]        <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "NoCarPC", off = NULL, model = "
        poisson")
668 PoisOvercrowdCount[itn, ] <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "Overcrowd", off = "Pop", model = "
        poisson")
```

```
PoisOvercrowdPC[itn, ]    <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "OvercrowdPC", off = NULL, model =
    "poisson")
PoisNonOwnCount[itn, ]    <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "Nonowner", off = "Pop", model = "
    poisson")
PoisNonOwnPC[itn, ]       <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = "NonOwnerPC", off = NULL, model = "
    poisson")

PoisUnempMSAS[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Unemp", "EA"),
                                   off = "Pop", model = "poisson")
PoisNoCarMSAS[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("NoCar", "HH", "Unemp", "EA"),
                                   off = "Pop", model = "poisson")
PoisOvercrowdMSAS[itn, ] <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Overcrowd", "HH", "Unemp", "EA"),
                                   off = "Pop", model = "poisson")
PoisNonOwnMSAS[itn, ]    <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Nonowner", "HH", "Unemp", "EA"),
                                   off = "Pop", model = "poisson")

NegBinTownsendUnadj[itn, ]  <- Models(GenSim3, "LLTIObs", "Townsend", NULL, "negativebinomial")
NegBinTownsend[itn, ]       <- Models(GenSim3, "LLTIObs", "Townsend", "Pop", "negativebinomial")
NegBinUnempCount[itn, ]     <- Models(GenSim3, "LLTIObs", "Unemp", "Pop", "negativebinomial")
NegBinUnempPC[itn, ]        <- Models(GenSim3, "LLTIObs", "UnempPC", NULL, "negativebinomial")
NegBinNoCarCount[itn, ]     <- Models(GenSim3, "LLTIObs", "NoCar", "Pop", "negativebinomial")
NegBinNoCarPC[itn, ]        <- Models(GenSim3, "LLTIObs", "NoCarPC", NULL, "negativebinomial")
NegBinOvercrowdCount[itn, ] <- Models(GenSim3, "LLTIObs", "Overcrowd", "Pop", "negativebinomial")
NegBinOvercrowdPC[itn, ]    <- Models(GenSim3, "LLTIObs", "OvercrowdPC", NULL, "negativebinomial")
NegBinNonOwnCount[itn, ]    <- Models(GenSim3, "LLTIObs", "Nonowner", "Pop", "negativebinomial")
NegBinNonOwnPC[itn, ]       <- Models(GenSim3, "LLTIObs", "NonOwnerPC", NULL, "negativebinomial")

NegBinUnempMSAS[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Unemp", "EA"),
                                     off = "Pop", model = "negativebinomial")
NegBinNoCarMSAS[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("NoCar", "HH", "Unemp", "EA"),
```

```
                                    off = "Pop", model = "negativebinomial")
NegBinOvercrowdMSAS[itn, ] <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Overcrowd", "HH", "Unemp", "EA"),
                                    off = "Pop", model = "negativebinomial")
NegBinNonOwnMSAS[itn, ]    <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Nonowner","HH", "Unemp", "EA"),
                                    off = "Pop", model = "negativebinomial")


LinTownsendUnadj[itn, ]   <- Models(GenSim3, "LLTIObs", "Townsend", NULL, "linear")
LinTownsend[itn, ]        <- Models(GenSim3, "LLTIObs", "Townsend", "Pop", "linear")
LinUnempCount[itn, ]      <- Models(GenSim3, "LLTIObs", "Unemp", "Pop", "linear")
LinUnempPC[itn, ]         <- Models(GenSim3, "LLTIObs", "UnempPC", NULL, "linear")
LinNoCarCount[itn, ]      <- Models(GenSim3, "LLTIObs", "NoCar", "Pop", "linear")
LinNoCarPC[itn, ]         <- Models(GenSim3, "LLTIObs", "NoCarPC", NULL, "linear")
LinOvercrowdCount[itn, ]  <- Models(GenSim3, "LLTIObs", "Overcrowd", "Pop", "linear")
LinOvercrowdPC[itn, ]     <- Models(GenSim3, "LLTIObs", "OvercrowdPC", NULL, "linear")
LinNonOwnCount[itn, ]     <- Models(GenSim3, "LLTIObs", "Nonowner", "Pop", "linear")
LinNonOwnPC[itn, ]        <- Models(GenSim3, "LLTIObs", "NonOwnerPC", NULL, "linear")

LinUnempMSAS[itn, ]      <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Unemp", "EA"),
                                    off = "Pop", model = "linear")
LinNoCarMSAS[itn, ]      <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("NoCar", "HH", "Unemp", "EA"),
                                    off = "Pop", model = "linear")
LinOvercrowdMSAS[itn, ]  <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Overcrowd", "HH", "Unemp", "EA"),
                                    off = "Pop", model = "linear")
LinNonOwnMSAS[itn, ]     <- Models(dataset = GenSim3, outcome = "LLTIObs", covariate = c("Nonowner", "HH", "Unemp", "EA"),
                                    off = "Pop", model = "linear")

}
```

```
End  <- Sys.time()

End - Beg

# Calculate lengths of 'significant' results

SigRes <- function(model) {
  length(model[which(model[,4] < 0.05)])
}

SigRes(PoisTownsendUnadj); SigRes(PoisTownsend); SigRes(PoisUnempCount); SigRes(PoisUnempPC); SigRes(PoisNoCarCount); SigRes(
    PoisNoCarPC);
SigRes(PoisNonOwnCount); SigRes(PoisNonOwnPC); SigRes(PoisOvercrowdCount); SigRes(PoisOvercrowdPC)
SigRes(PoisUnempMSAS); SigRes(PoisNonOwnMSAS); SigRes(PoisNoCarMSAS); SigRes(PoisOvercrowdMSAS)

SigRes(NegBinTownsendUnadj); SigRes(NegBinTownsend); SigRes(NegBinUnempCount); SigRes(NegBinUnempPC); SigRes(NegBinNoCarCount
    ); SigRes(NegBinNoCarPC);
SigRes(NegBinNonOwnCount); SigRes(NegBinNonOwnPC); SigRes(NegBinOvercrowdCount); SigRes(NegBinOvercrowdPC)
SigRes(NegBinUnempMSAS); SigRes(NegBinNonOwnMSAS); SigRes(NegBinNoCarMSAS); SigRes(NegBinOvercrowdMSAS)

SigRes(LinTownsendUnadj); SigRes(LinTownsend); SigRes(LinUnempCount); SigRes(LinUnempPC); SigRes(LinNoCarCount); SigRes(
    LinNoCarPC);
SigRes(LinNonOwnCount); SigRes(LinNonOwnPC); SigRes(LinOvercrowdCount); SigRes(LinOvercrowdPC)
SigRes(LinUnempMSAS); SigRes(LinNonOwnMSAS); SigRes(LinNoCarMSAS); SigRes(LinOvercrowdMSAS)

# Summarise esimated coefficients

SumCoef  <- function(model) {
  #summary(model[,3])
  quantile(sort(model[,1]),c(0.025,0.5,0.975))
```

```
752 }
753
754 PTUCoef <- exp(SumCoef(PoisTownsendUnadj));  PTCoef    <- exp(SumCoef(PoisTownsend));
755 PUCoef  <- exp(SumCoef(PoisUnempCount));     PUPCCoef  <- exp(SumCoef(PoisUnempPC));
756 PCCoef  <- exp(SumCoef(PoisNoCarCount));     PCPCCoef  <- exp(SumCoef(PoisNoCarPC));
757 POCoef  <- exp(SumCoef(PoisNonOwnCount));    POPCCoef  <- exp(SumCoef(PoisNonOwnPC));
758 POCCoef <- exp(SumCoef(PoisOvercrowdCount)); POCPCCoef <- exp(SumCoef(PoisOvercrowdPC))
759 PUMSAS  <- exp(SumCoef(PoisUnempMSAS));      PCMSAS    <- exp(SumCoef(PoisNoCarMSAS));
760 POMSAS  <- exp(SumCoef(PoisNonOwnMSAS));     POCMSAS   <- exp(SumCoef(PoisOvercrowdMSAS))
761
762 NTUCoef <- exp(SumCoef(NegBinTownsendUnadj));  NTCoef    <- exp(SumCoef(NegBinTownsend));
763 NUCoef  <- exp(SumCoef(NegBinUnempCount));     NUPCCoef  <- exp(SumCoef(NegBinUnempPC));
764 NCCoef  <- exp(SumCoef(NegBinNoCarCount));     NCPCCoef  <- exp(SumCoef(NegBinNoCarPC));
765 NOCoef  <- exp(SumCoef(NegBinNonOwnCount));    NOPCCoef  <- exp(SumCoef(NegBinNonOwnPC));
766 NOCCoef <- exp(SumCoef(NegBinOvercrowdCount)); NOCPCCoef <- exp(SumCoef(NegBinOvercrowdPC))
767 NUMSAS  <- exp(SumCoef(NegBinUnempMSAS));      NCMSAS    <- exp(SumCoef(NegBinNoCarMSAS));
768 NOMSAS  <- exp(SumCoef(NegBinNonOwnMSAS));     NOCMSAS   <- exp(SumCoef(NegBinOvercrowdMSAS));
769
770 LTUCoef <- SumCoef(LinTownsendUnadj); LTCoef <- SumCoef(LinTownsend); LUCoef <- SumCoef(LinUnempCount); LUPCCoef <- SumCoef(
        LinUnempPC); LCCoef <- SumCoef(LinNoCarCount); LCPCCoef <- SumCoef(LinNoCarPC);
771 LOCoef <- SumCoef(LinNonOwnCount); LOPCCoef <- SumCoef(LinNonOwnPC); LOCCoef <- SumCoef(LinOvercrowdCount); LOCPCCoef <-
        SumCoef(LinOvercrowdPC)
772 LUMSAS <- SumCoef(LinUnempMSAS); LCMSAS <- SumCoef(LinNoCarMSAS); LOMSAS <- SumCoef(LinNonOwnMSAS); LOCMSAS <- SumCoef(
        LinOvercrowdMSAS);
773
774
775 PoisEsts <- data.frame(Estimates = rbind(PTUCoef, PTCoef, PUCoef, PUPCCoef, PCCoef, PCPCCoef,
776                                          POCoef, POPCCoef, POCCoef, POCPCCoef,
777                                          PUMSAS, PCMSAS, POMSAS, POCMSAS),
778                        Label=c("Townsend","Townsend", "Unemployed Population", "Unemployed Population",
```

```
                                "Household without Car", "Household without Car",
                                "Overcrowded Household", "Overcrowded Household",
                                "Household not Owner-Occupied", "Household not Owner-Occupied",
                                "Unemployed Population", "Household without Car",
                                "Household not Owner-Occupied", "Overcrowded Household"),
                        Mod=c("Composite", "Composite - Adjusted", "Count", "Proportion", "Count", "Proportion", "Count", "
                            Proportion", "Count", "Proportion", "MSAS",
                              "MSAS", "MSAS", "MSAS"))

PoisEsts$Label <- factor(PoisEsts$Label, levels = unique(PoisEsts$Label))

x11()
ggplot(PoisEsts,aes(x = Label, y = Estimates.50., group = Mod, colour = Mod)) +
  geom_point(position = position_dodge(width = .5)) +
  geom_errorbar(data = PoisEsts, aes(ymin = Estimates.2.5., ymax = Estimates.97.5., colour = Mod), width = .5, size = 1,
      position = position_dodge(width = .5)) +
  geom_hline(yintercept = 1.00, size = 0.2, colour = "black", linetype = "dashed") + theme_bw() +
  scale_x_discrete(name = "") +
  theme(legend.title = element_blank()) +
  theme(axis.text=element_text(colour="black",size=12,family="Arial"),axis.title=element_text(size=16, family = "Arial", face
       = "bold"),plot.title=element_text(size=16,face="bold", family = "Arial"),legend.position="top",
        legend.text = element_text(size = 12, family = "Arial")) +
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)), name = "Coefficient of Covariate of Interest")+
  scale_colour_viridis(discrete = TRUE, direction = -1, begin = 0, end = 0.9)


LinEsts <- data.frame(Estimates = rbind(LTUCoef, LTCoef, LUCoef, LUPCCoef, LCCoef, LCPCCoef,
                                        LOCoef, LOPCCoef, LOCCoef, LOCPCCoef,
                                        LUMSAS, LCMSAS, LOMSAS, LOCMSAS),
```

```
806                       Label=c("Townsend", "Townsend", "Unemployed Population", "Unemployed Population",
807                               "Household without Car", "Household without Car",
808                               "Overcrowded Household", "Overcrowded Household",
809                               "Household not Owner-Occupied", "Household not Owner-Occupied",
810                               "Unemployed Population", "Household without Car",
811                               "Household not Owner-Occupied", "Overcrowded Household"),
812                       Mod=c("Composite", "Composite - Adjusted", "Count", "Proportion", "Count", "Proportion", "Count", "
                              Proportion", "Count", "Proportion", "MSAS",
813                                 "MSAS", "MSAS", "MSAS"))
814
815 LinEsts$Label <- factor(LinEsts$Label, levels = unique(LinEsts$Label))
816
817 x11()
818 ggplot(LinEsts,aes(x = Label, y = Estimates.50., group = Mod, colour = Mod)) +
819   geom_point(position = position_dodge(width = .5)) +
820   geom_errorbar(data = LinEsts, aes(ymin = Estimates.2.5., ymax = Estimates.97.5., colour = Mod), width = .5, size = 1,
          position = position_dodge(width = .5)) +
821   geom_hline(yintercept = 0.00, size = 0.2, colour = "black", linetype = "dashed") + theme_bw() +
822   scale_y_continuous(name = "Coefficient of Covariate of Interest") + scale_x_discrete(name = "") +
823   theme(legend.title = element_blank()) +
824   theme(axis.text=element_text(colour="black",size=12,family="Arial"),axis.title=element_text(size=16, family = "Arial", face
           = "bold"),plot.title=element_text(size=16,face="bold", family = "Arial"),legend.position="top",
825          legend.text = element_text(size = 12, family = "Arial"))+
826   scale_colour_viridis(discrete = TRUE, direction = -1, begin = 0, end = 0.9)
827
828
829 ##################################################
830 ## Save all of the regression details into tables ##
831 ##################################################
832
```

```
write.table(PoisTownsendUnadj, "PoisTownsendUnadj - 181012.csv", sep =",", col.names = TRUE)
write.table(PoisTownsend, "PoisTownsend - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisUnempCount, "PoisUnempCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisUnempPC, "PoisUnempPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisNoCarCount, "PoisNoCarCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisNoCarPC, "PoisNoCarPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisOvercrowdCount, "PoisOvercrowdCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisOvercrowdPC, "PoisOvercrowdPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisNonOwnCount, "PoisNonOwnCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisNonOwnPC, "PoisNonOwnPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisUnempMSAS, "PoisUnempMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisNoCarMSAS, "PoisNoCarMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisOvercrowdMSAS, "PoisOvercrowdMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(PoisNonOwnMSAS, "PoisNonOwnMSAS - 181012.csv", sep = ",", col.names = TRUE)

write.table(NegBinTownsendUnadj, "NegBinTownsendUnadj - 181012.csv", sep =",", col.names = TRUE)
write.table(NegBinTownsend, "NegBinTownsend - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinUnempCount, "NegBinUnempCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinUnempPC, "NegBinUnempPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinNoCarCount, "NegBinNoCarCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinNoCarPC, "NegBinNoCarPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinOvercrowdCount, "NegBinOvercrowdCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinOvercrowdPC, "NegBinOvercrowdPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinNonOwnCount, "NegBinNonOwnCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinNonOwnPC, "NegBinNonOwnPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinUnempMSAS, "NegBinUnempMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinNoCarMSAS, "NegBinNoCarMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinOvercrowdMSAS, "NegBinOvercrowdMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(NegBinNonOwnMSAS, "NegBinNonOwnMSAS - 181012.csv", sep = ",", col.names = TRUE)

```

```
#
write.table(LinTownsendUnadj, "LinTownsendUnadj - 181012.csv", sep =",", col.names = TRUE)
write.table(LinTownsend, "LinTownsend - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinUnempCount, "LinUnempCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinUnempPC, "LinUnempPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinNoCarCount, "LinNoCarCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinNoCarPC, "LinNoCarPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinOvercrowdCount, "LinOvercrowdCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinOvercrowdPC, "LinOvercrowdPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinNonOwnCount, "LinNonOwnCount - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinNonOwnPC, "LinNonOwnPC - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinUnempMSAS, "LinUnempMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinNoCarMSAS, "LinNoCarMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinOvercrowdMSAS, "LinOvercrowdMSAS - 181012.csv", sep = ",", col.names = TRUE)
write.table(LinNonOwnMSAS, "LinNonOwnMSAS - 181012.csv", sep = ",", col.names = TRUE)

####################
## Read in tables ##
####################

PoisTownsendUnadj  <- read.table("PoisTownsendUnadj - 181012.csv", sep = ",")
PoisTownsend       <- read.table("PoisTownsend - 181012.csv", sep = ",")
PoisUnempCount     <- read.table("PoisUnempCount - 181012.csv", sep = ",")
PoisUnempPC        <- read.table("PoisUnempPC - 181012.csv", sep = ",")
PoisNoCarCount     <- read.table("PoisNoCarCount - 181012.csv", sep = ",")
PoisNoCarPC        <- read.table("PoisNoCarPC - 181012.csv", sep = ",")
PoisOvercrowdCount <- read.table("PoisOvercrowdCount - 181012.csv", sep = ",")
PoisOvercrowdPC    <- read.table("PoisOvercrowdPC - 181012.csv", sep = ",")
PoisNonOwnCount    <- read.table("PoisNonOwnCount - 181012.csv", sep = ",")
PoisNonOwnPC       <- read.table("PoisNonOwnPC - 181012.csv", sep = ",")
```

```
PoisUnempMSAS      <- read.table("PoisUnempMSAS - 181012.csv", sep = ",")
PoisNoCarMSAS      <- read.table("PoisNoCarMSAS - 181012.csv", sep = ",")
PoisOvercrowdMSAS  <- read.table("PoisOvercrowdMSAS - 181012.csv", sep = ",")
PoisNonOwnMSAS     <- read.table("PoisNonOwnMSAS - 181012.csv", sep = ",")

NegBinTownsend        <- read.table("NegBinTownsend - 181012.csv", sep = ",")
NegBinTownsendUnadj   <- read.table("NegBinTownsendUnadj - 181012.csv", sep =",")
NegBinUnempCount      <- read.table("NegBinUnempCount - 181012.csv", sep = ",")
NegBinUnempPC         <- read.table("NegBinUnempPC - 181012.csv", sep = ",")
NegBinNoCarCount      <- read.table("NegBinNoCarCount - 181012.csv", sep = ",")
NegBinNoCarPC         <- read.table("NegBinNoCarPC - 181012.csv", sep = ",")
NegBinOvercrowdCount  <- read.table("NegBinOvercrowdCount - 181012.csv", sep = ",")
NegBinOvercrowdPC     <- read.table("NegBinOvercrowdPC - 181012.csv", sep = ",")
NegBinNonOwnCount     <- read.table("NegBinNonOwnCount - 181012.csv", sep = ",")
NegBinNonOwnPC        <- read.table("NegBinNonOwnPC - 181012.csv", sep = ",")
NegBinUnempMSAS       <- read.table("NegBinUnempMSAS - 181012.csv", sep = ",")
NegBinNoCarMSAS       <- read.table("NegBinNoCarMSAS - 181012.csv", sep = ",")
NegBinOvercrowdMSAS   <- read.table("NegBinOvercrowdMSAS - 181012.csv", sep = ",")
NegBinNonOwnMSAS      <- read.table("NegBinNonOwnMSAS - 181012.csv", sep = ",")
#
LinTownsend       <- read.table("LinTownsend - 181012.csv", sep = ",")
LinTownsendUnadj  <- read.table("LinTownsendUnadj - 181012.csv", sep = ",")
LinUnempCount     <- read.table("LinUnempCount - 181012.csv", sep = ",")
LinUnempPC        <- read.table("LinUnempPC - 181012.csv", sep = ",")
LinNoCarCount     <- read.table("LinNoCarCount - 181012.csv", sep = ",")
LinNoCarPC        <- read.table("LinNoCarPC - 181012.csv", sep = ",")
LinOvercrowdCount <- read.table("LinOvercrowdCount - 181012.csv", sep = ",")
LinOvercrowdPC    <- read.table("LinOvercrowdPC - 181012.csv", sep = ",")
LinNonOwnCount    <- read.table("LinNonOwnCount - 181012.csv", sep = ",")
LinNonOwnPC       <- read.table("LinNonOwnPC - 181012.csv", sep = ",")
```

```
923 LinUnempMSAS      <- read.table("LinUnempMSAS - 181012.csv", sep = ",")
924 LinNoCarMSAS      <- read.table("LinNoCarMSAS - 181012.csv", sep = ",")
925 LinOvercrowdMSAS  <- read.table("LinOvercrowdMSAS - 181012.csv", sep = ",")
926 LinNonOwnMSAS     <- read.table("LinNonOwnMSAS - 181012.csv", sep = ",")
927
928
929 ####################################################
930 ## Run Fitdistr on the adjusted household variables ##
931 ####################################################
932
933 ## Create data frame from the six household variables
934
935 HouseVars  <- data.matrix(Nonowner = VecOwn[,1], Owner = VecOwn[,2], Nocar = VecCar[,1], Car = VecCar[,2], Overcrowd = VecOvr
        [,1], NotOvercrowd = VecOvr[,2])
936
937
938 HousePars  <- matrix(nrow = length(HouseVars[1,]), ncol = 3)
939 for (i in 1:length(HouseVars[1,])) {
940
941   tmp     <- fitdistr(HouseVars[,i],"lognormal"); Pars <- tmp$estimate
942   Dat     <- data.matrix(LN = round(rlnorm(DFn, meanlog = Pars[1], sdlog = Pars[2])))
943
944   HousePars[i,1]  <- Pars[1]
945   HousePars[i,2]  <- Pars[2]
946   HousePars[i,3]  <- colnames(HouseVars[i])
947
948   tmpData <- data.matrix(ID = 1:DFn, Observed = HouseVars[, i], LogNormal = Dat$LN)
949
950   Variable  <- colnames(HouseVars[i])
951
```

```
952   #apply(tmpData[,-1],2,summary)
953   #apply(tmpData[,-1],2,sum)
954   dd      <- melt(tmpData, id = c("ID")); names(dd) <- c("ID", "Dist", "Variable")
955   Xlim    <- c(0,max(dd[3])); Xlab <- Variable; Ylab <- "Kernel Density"
956   Mlab    <- Variable
957   print(ggplot(dd) + geom_density(aes(x = Variable, group = Dist, colour = Dist), size = 1.2) + labs(x = Xlab, y = Ylab,
          colour = NULL) +
958           coord_cartesian(xlim = Xlim, ylim = NULL) + theme(axis.title = element_text(size = 16), axis.text.x = element_text(
                  size = 16),
959                                                             axis.text.y = element_text(size = 16), plot.title = element_text(size
                                                                 = 16),
960                                                             legend.position = "right") + scale_colour_manual(values = c("black",
                                                                 cbPal[-1]), name = "Distribution"))
961   ggsave(filename = paste("LogNormal Household Variables - ", Variable," - 170822.png"), width = 16, height = 8.17)
962   ##
963
964 }
965
966 ###################################
967 ## 'Zip plots' to illustrate bias ##
968 ###################################
969
970 zipplot <- function(modelsum) {
971   modelsum  <- as.data.frame(modelsum)
972   modelsumz <- modelsum[order(-modelsum[ ,4]), ]
973   modelsumz$Sig <- as.factor((modelsumz[ ,4] > 0.05)*1)
974   df <- data.frame(x = c(1:length(modelsumz[ ,1])), y = modelsumz[,1], ylo = modelsumz[,2], yhi = modelsumz[ ,3], sig =
          modelsumz$Sig)
975   p <- ggplot(df, aes(x = x, y = y , ymin = ylo, ymax = yhi, color = sig)) +
976     geom_pointrange(size = 0.2) +
```

```
977     geom_point(size = 0.1) +
978     geom_hline(yintercept = 0, linetype = 1) +
979     scale_color_manual(name = "", values = c("lightcoral", "lightblue4"), labels = c("Non-coverer", "Coverer")) +
980     coord_flip() +
981     geom_vline(xintercept = c(50, 500, 950), linetype = 1, color = "grey") +
982     theme(legend.position = "none", axis.title.x = element_blank(),
983           axis.title.y = element_blank(), axis.text.x = element_text(vjust = 1))
984   return(p)
985
986 }
987
988 #aspect.ratio = 0.75,
989
990 p1  <- zipplot(PoisTownsendUnadj)
991 p2  <- zipplot(PoisTownsend)
992
993 p3  <- zipplot(PoisUnempPC)
994 p4  <- zipplot(PoisUnempCount)
995 p5  <- zipplot(PoisUnempMSAS)
996
997 p6  <- zipplot(PoisNoCarPC)
998 p7  <- zipplot(PoisNoCarCount)
999 p8  <- zipplot(PoisNoCarMSAS)
1000
1001 p9  <- zipplot(PoisNonOwnPC)
1002 p10  <- zipplot(PoisNonOwnCount)
1003 p11  <- zipplot(PoisNonOwnMSAS)
1004
1005 p12  <- zipplot(PoisOvercrowdPC)
1006 p13  <- zipplot(PoisOvercrowdCount)
```

```
p14  <- zipplot(PoisOvercrowdMSAS)

label1  <- textGrob("Proportion", vjust = 0.5)
label2  <- textGrob("Count", vjust = 0.5)
label3  <- textGrob("MSAS", vjust = 0.5)
label4  <- textGrob("Townsend", vjust = 0.5)
label5  <- textGrob("Unemployment", vjust = 0.5)
label6  <- textGrob("No car", vjust = 0.5)
label7  <- textGrob("Nonowner-occupied", vjust = 0.5)
label8  <- textGrob("Overcrowded", vjust = 0.5)

windows()
grid.arrange(arrangeGrob(ggplotGrob(p1),
                         top = textGrob("Proportion", gp = gpar(fontsize = 18, fontfamily = "Times New Roman")),
                         left = textGrob("Townsend", vjust = 0.5, hjust = 0.5, rot = 90,
                                         gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
             arrangeGrob(ggplotGrob(p2), top = textGrob("Count", gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
             arrangeGrob(rectGrob(gp=gpar(col=NA)), top = textGrob("MSAS", gp = gpar(fontsize = 18, fontfamily = "Times New
                 Roman"))),
             arrangeGrob(ggplotGrob(p3), left = textGrob("Unemployment", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                 fontsize = 18, fontfamily = "Times New Roman"))),
             arrangeGrob(ggplotGrob(p4)),
             ggplotGrob(p5),
             arrangeGrob(ggplotGrob(p6), left = textGrob("No Car", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(fontsize =
                 18, fontfamily = "Times New Roman"))),
             ggplotGrob(p7),
             ggplotGrob(p8),
             arrangeGrob(ggplotGrob(p9), left = textGrob("Non Owner-occupied", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                 fontsize = 18, fontfamily = "Times New Roman"))),
             ggplotGrob(p10),
```

```
1033              ggplotGrob(p11),
1034              arrangeGrob(ggplotGrob(p12), left = textGrob("Overcrowded", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                      fontsize = 18, fontfamily = "Times New Roman"))),
1035              ggplotGrob(p13),
1036              ggplotGrob(p14),ncol = 3,
1037              left = textGrob("Centile of ranked p-values under null hypothesis", gp = gpar(fontsize = 24, fontface = "bold",
                      fontfamily = "Times New Roman"), rot = 90),
1038              bottom = textGrob("95% confidence intervals", gp = gpar(fontsize = 24, fontface = "bold", fontfamily = "Times
                      New Roman")))
1039
1040
1041 p1   <- zipplot(LinTownsendUnadj)
1042 p2   <- zipplot(LinTownsend)
1043 p3   <- zipplot(LinUnempPC)
1044 p4   <- zipplot(LinUnempCount)
1045 p5   <- zipplot(LinUnempMSAS)
1046 p6   <- zipplot(LinNoCarPC)
1047 p7   <- zipplot(LinNoCarCount)
1048 p8   <- zipplot(LinNoCarMSAS)
1049 p9   <- zipplot(LinNonOwnPC)
1050 p10  <- zipplot(LinNonOwnCount)
1051 p11  <- zipplot(LinNonOwnMSAS)
1052 p12  <- zipplot(LinOvercrowdPC)
1053 p13  <- zipplot(LinOvercrowdCount)
1054 p14  <- zipplot(LinOvercrowdMSAS)
1055
1056 label1  <- textGrob("Proportion", vjust = 0.5)
1057 label2  <- textGrob("Count", vjust = 0.5)
1058 label3  <- textGrob("MSAS", vjust = 0.5)
1059 label4  <- textGrob("Townsend", vjust = 0.5)
```

```
label5  <- textGrob("Unemployment", vjust = 0.5)
label6  <- textGrob("No car", vjust = 0.5)
label7  <- textGrob("Nonowner-occupied", vjust = 0.5)
label8  <- textGrob("Overcrowded", vjust = 0.5)

windows()
grid.arrange(arrangeGrob(ggplotGrob(p1),
                         top = textGrob("Proportion", gp = gpar(fontsize = 18, fontfamily = "Times New Roman")),
                         left = textGrob("Townsend", vjust = 0.5, hjust = 0.5, rot = 90,
                                         gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
             arrangeGrob(ggplotGrob(p2), top = textGrob("Count", gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
             arrangeGrob(rectGrob(gp=gpar(col=NA)), top = textGrob("MSAS", gp = gpar(fontsize = 18, fontfamily = "Times New
                 Roman"))),
             arrangeGrob(ggplotGrob(p3), left = textGrob("Unemployment", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                 fontsize = 18, fontfamily = "Times New Roman"))),
             arrangeGrob(ggplotGrob(p4)),
             ggplotGrob(p5),
             arrangeGrob(ggplotGrob(p6), left = textGrob("No Car", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(fontsize =
                 18, fontfamily = "Times New Roman"))),
             ggplotGrob(p7),
             ggplotGrob(p8),
             arrangeGrob(ggplotGrob(p9), left = textGrob("Non Owner-occupied", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                 fontsize = 18, fontfamily = "Times New Roman"))),
             ggplotGrob(p10),
             ggplotGrob(p11),
             arrangeGrob(ggplotGrob(p12), left = textGrob("Overcrowded", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                 fontsize = 18, fontfamily = "Times New Roman"))),
             ggplotGrob(p13),
             ggplotGrob(p14),ncol = 3,
```

```
1084              left = textGrob("Centile of ranked p-values under null hypothesis", gp = gpar(fontsize = 24, fontface = "bold",
                      fontfamily = "Times New Roman"), rot = 90),
1085              bottom = textGrob("95% confidence intervals", gp = gpar(fontsize = 24, fontface = "bold", fontfamily = "Times
                      New Roman")))
1086
1087
1088 p1  <- zipplot(BLinTownsend)
1089 p2  <- zipplot(BLinUnempPC)
1090 p3  <- zipplot(BLinUnempCount)
1091 p4  <- zipplot(BLinNoCarPC)
1092 p5  <- zipplot(BLinNoCarCount)
1093 p6  <- zipplot(BLinNonOwnPC)
1094 p7  <- zipplot(BLinNonOwnCount)
1095 p8  <- zipplot(BLinOvercrowdPC)
1096 p9  <- zipplot(BLinOvercrowdCount)
1097
1098 grid.arrange(arrangeGrob(ggplotGrob(p1),
1099                          top = textGrob("Proportion", gp = gpar(fontsize = 18, fontfamily = "Times New Roman")),
1100                          left = textGrob("Townsend", vjust = 0.5, hjust = 0.5, rot = 90,
1101                                          gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
1102              arrangeGrob(rectGrob(gp=gpar(col=NA)), top = textGrob("Count", gp = gpar(fontsize = 18, fontfamily = "Times New
                      Roman"))),
1103              arrangeGrob(ggplotGrob(p2), left = textGrob("Unemployment", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                      fontsize = 18, fontfamily = "Times New Roman"))),
1104              ggplotGrob(p3),
1105              arrangeGrob(ggplotGrob(p4), left = textGrob("No Car", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(fontsize =
                      18, fontfamily = "Times New Roman"))),
1106              ggplotGrob(p5),
1107              arrangeGrob(ggplotGrob(p6), left = textGrob("Non Owner-occupied", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                      fontsize = 18, fontfamily = "Times New Roman"))),
```

```
1108            ggplotGrob(p7),
1109            arrangeGrob(ggplotGrob(p8), left = textGrob("Overcrowded", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                    fontsize = 18, fontfamily = "Times New Roman"))),
1110            ggplotGrob(p9),ncol = 2,
1111            left = textGrob("Centile of ranked p-values under null hypothesis", gp = gpar(fontsize = 24, fontface = "bold",
                    fontfamily = "Times New Roman"), rot = 90),
1112            bottom = textGrob("95% confidence intervals", gp = gpar(fontsize = 24, fontface = "bold", fontfamily = "Times
                    New Roman")))
1113
1114 ############################
1115 ## Correlation Zip Plots ##
1116 ############################
1117
1118 p1  <- zipplot(CorTownsend)
1119 p2  <- zipplot(CorUnempPC)
1120 p3  <- zipplot(CorUnempCount)
1121 p4  <- zipplot(CorNoCarPC)
1122 p5  <- zipplot(CorNoCarCount)
1123 p6  <- zipplot(CorNonOwnPC)
1124 p7  <- zipplot(CorNonOwnCount)
1125 p8  <- zipplot(CorOvercrowdPC)
1126 p9  <- zipplot(CorOvercrowdCount)
1127
1128 grid.arrange(arrangeGrob(ggplotGrob(p1),
1129                          top = textGrob("Proportion", gp = gpar(fontsize = 18, fontfamily = "Times New Roman")),
1130                          left = textGrob("Townsend", vjust = 0.5, hjust = 0.5, rot = 90,
1131                                          gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
1132             arrangeGrob(rectGrob(gp=gpar(col=NA)), top = textGrob("Count", gp = gpar(fontsize = 18, fontfamily = "Times New
                    Roman"))),
```

```
1133              arrangeGrob(ggplotGrob(p2), left = textGrob("Unemployment", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                      fontsize = 18, fontfamily = "Times New Roman"))),
1134              ggplotGrob(p3),
1135              arrangeGrob(ggplotGrob(p4), left = textGrob("No Car", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(fontsize =
                      18, fontfamily = "Times New Roman"))),
1136              ggplotGrob(p5),
1137              arrangeGrob(ggplotGrob(p6), left = textGrob("Non Owner-occupied", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                      fontsize = 18, fontfamily = "Times New Roman"))),
1138              ggplotGrob(p7),
1139              arrangeGrob(ggplotGrob(p8), left = textGrob("Overcrowded", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                      fontsize = 18, fontfamily = "Times New Roman"))),
1140              ggplotGrob(p9),ncol = 2,
1141              left = textGrob("Centile of ranked p-values under null hypothesis", gp = gpar(fontsize = 24, fontface = "bold",
                      fontfamily = "Times New Roman"), rot = 90),
1142              bottom = textGrob("95% confidence intervals", gp = gpar(fontsize = 24, fontface = "bold", fontfamily = "Times
                      New Roman")))
1143
1144 p1  <- zipplot(BCorTownsend)
1145 p2  <- zipplot(BCorUnempPC)
1146 p3  <- zipplot(BCorUnempCount)
1147 p4  <- zipplot(BCorNoCarPC)
1148 p5  <- zipplot(BCorNoCarCount)
1149 p6  <- zipplot(BCorNonOwnPC)
1150 p7  <- zipplot(BCorNonOwnCount)
1151 p8  <- zipplot(BCorOvercrowdPC)
1152 p9  <- zipplot(BCorOvercrowdCount)
1153
1154 grid.arrange(arrangeGrob(ggplotGrob(p1),
1155                          top = textGrob("Proportion", gp = gpar(fontsize = 18, fontfamily = "Times New Roman")),
1156                          left = textGrob("Townsend", vjust = 0.5, hjust = 0.5, rot = 90,
```

```
1157                                    gp = gpar(fontsize = 18, fontfamily = "Times New Roman"))),
1158            arrangeGrob(rectGrob(gp=gpar(col=NA)), top = textGrob("Count", gp = gpar(fontsize = 18, fontfamily = "Times New
                    Roman"))),
1159            arrangeGrob(ggplotGrob(p2), left = textGrob("Unemployment", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                    fontsize = 18, fontfamily = "Times New Roman"))),
1160            ggplotGrob(p3),
1161            arrangeGrob(ggplotGrob(p4), left = textGrob("No Car", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(fontsize =
                    18, fontfamily = "Times New Roman"))),
1162            ggplotGrob(p5),
1163            arrangeGrob(ggplotGrob(p6), left = textGrob("Non Owner-occupied", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                    fontsize = 18, fontfamily = "Times New Roman"))),
1164            ggplotGrob(p7),
1165            arrangeGrob(ggplotGrob(p8), left = textGrob("Overcrowded", vjust = 0.5, hjust = 0.5, rot = 90, gp = gpar(
                    fontsize = 18, fontfamily = "Times New Roman"))),
1166            ggplotGrob(p9),ncol = 2,
1167            left = textGrob("Centile of ranked p-values under null hypothesis", gp = gpar(fontsize = 24, fontface = "bold",
                    fontfamily = "Times New Roman"), rot = 90),
1168            bottom = textGrob("95% confidence intervals", gp = gpar(fontsize = 24, fontface = "bold", fontfamily = "Times
                    New Roman")))
```

LLTI.R

Appendix D

Simulations of area–level data to investigate analyses of ‘population mixing’ and childhood leukaemia

```
1  ##############################################################
2  ## Accompanying code for paper: 'Is the association between  ##
3  ## childhood leukaemia and population mixing an artefact of  ##
4  ## focusing on 'clusters' of cases? (Berrie et al., 2018)    ##
5  ##############################################################
6
7  ###################
8  ## Load packages ##
9  ###################
10 rm(list=ls());
11 library(MASS); library(ggplot2); library(reshape);
```

```
library(extrafont); library(gridExtra); library(grid); library(VGAM);
library(Matrix); library(boot); library(pscl);library(SuppDists);
library(distr);library(distrEx);library(stringr)

##########################################
## Read in the Yorkshire & Humber dataset ##
##########################################

## Omitted as dataset not publicly available

############################################################
## GenData Function (adapted from Ruscio & Kaczetow, 2008) ##
############################################################

GenData <- function(Supp.Data=NULL,n.Fact=0,Max.Trials=5,Init.Mult=1,
                    seed=0,Emp=TRUE,Target.Corr=NULL,N=NULL,k=NULL) {
  ##############################################################################################
  # Initialize variables and (if applicable) set random number seed (step 1)
  if (Emp) {
    N   <- dim(Supp.Data)[1]                  # Number of cases
    k   <- dim(Supp.Data)[2] }                # Number of variables
  Data    <- matrix(0,nrow=N,ncol=k)          # Matrix to store the simulated data
  Distributions <- matrix(0,nrow=N,ncol=k)    # Matrix to store each variable?s score distribution
  Iteration <- 0                              # Iteration counter
  Best.RMSR <- 1                              # Lowest RMSR correlation
  Trials.Without.Improvement <- 0             # Trial counter
  if (seed != 0) set.seed(seed)               # If user specified a nonzero seed, set it
  ##############################################################################################
  # Generate distribution for each variable (step 2) ------------------------
  if (Emp) for (i in 1:k) Distributions[,i] <- sort(sample(Supp.Data[,i],N,replace=TRUE)) else {
```

```
  Distributions[,1] <- sort(rnegbin(N,1300,1.6))              # 0 - 14 population
  Distributions[,2] <- sort(rnegbin(N,26,0.7))                # Area
  Distributions[,3] <- sort(rnegbin(N,500,1.8))              # 'Post' in-migration
  Distributions[,4] <- sort(rnegbin(N,6500,2.0))  }           # Total population
# This implementation of GenData bootstraps each variable?s score distribution from a supplied data set.
# Users should modify this block of the program, as needed, to generate the desired distribution(s).
# For example, to sample from chi-square distributions with 2 df, replace the 2nd line in this block with:
#   Distributions[,i] <- sort(rchisq(N,df=2))
# Or, one can drop the loop and use a series of commands that samples variables from specified populations:
#   Distributions[,1] <- sort(rnorm(N,0,1))         # Standard normal distribution
#   Distributions[,2] <- sort(runif(N,0,1))         # Uniform distribution ranging from 0 - 1
#   Distributions[,3] <- sort(rlnorm(N,0,1))        # Log-normal distribution, log scale M = 0, SD = 1
#   Distributions[,4] <- sort(rexp(N,rate=1))       # Exponential distribution with rate = 1
#   Distributions[,5] <- sort(rpois(N,lambda=4))    # Poisson distribution with lambda = 4
#   Distributions[,6] <- sort(rbinom(N,10,0.25)     # Binominal distribution, size = 10 and p = 0.25
#   Distributions[,7] <- sort(rbinom(N,2,0.25)      # Binary distribution with p = 0.25
#########################################################################################################
# All of the commands shown above draw random samples from specified population distributions. Alternatively,
# one can reproduce distributions without sampling error. For example, working with a supplied data set, one can
# replace the 2nd line in this block with:
#   Distributions[,i] <- Supp.Data[,i]
#########################################################################################################
# Alternatively, idealized distributions can be reproduced. For example, uniform quantiles can be created and
# used to generate data from common distributions:
#   Uniform.Quantiles <- seq(from = 0, to = 1, length = (N + 2))[2:(N + 1)] # quantiles 0, 1 dropped
#   Distributions[,1] <- qnorm(Uniform.Quantiles,0,1)   # Standard normal distribution
#   Distributions[,2] <- qunif(Uniform.Quantiles,0,1)   # Uniform distribution ranging from 0 to 1
#   Distributions[,3] <- qchisq(Uniform.Quantiles,df=2) # Chi-square distribution with 2 df
#########################################################################################################
# Note that when score distributions are generated from specified populations rather than bootstrapped from a
```

```
# supplied dataset, the user must provide the target correlation matrix (see the next block). This is true
# regardless of whether the distributions incorporate sampling error.
##############################################################################################################
# Calculate and store a copy of the target correlation matrix (step 3) ---------------------------------------
if (Emp) Target.Corr <- cor(Supp.Data)
Intermediate.Corr    <- Target.Corr
# This implementation of GenData calculates the target correlation matrix from a supplied dataset.
# Alternatively, the user can modify the program to generate data with user-defined sample size, number of
# variables and target correlation matrix by redefining the function as follows:
#   GenData <- function(N,k,Target.Corr,n.Fact=0,Max.Trials=5,Init.Mult=1,seed=0)
# In this case, one would also remove the program lines that calculate N, k, and Target.Corr.
# To generate data in which variables are uncorrelated, one would remove the SsortT function from step 2
# and terminate the program before step 3 begins by returning the Distributions object as the dataset.
##############################################################################################################
# If number of latent factors was not specified, determine it through parallel analysis (step 4) --------------
if (n.Fact == 0) {
  Eigenvalues.Observed <- eigen(Intermediate.Corr)$values
  Eigenvalues.Random <- matrix(0, nrow = 100, ncol = k)
  Random.Data <- matrix(0, nrow = N, ncol = k)
  for (i in 1:100) {
    for (j in 1:k) Random.Data[,j] <- sample(Distributions[,j], size = N, replace = TRUE)
    Eigenvalues.Random[i,] <- eigen(cor(Random.Data))$values }
  Eigenvalues.Random <- apply(Eigenvalues.Random, 2, mean)        # calculate mean eigenvalue for each factor
  n.Fact <- max(1, sum(Eigenvalues.Observed > Eigenvalues.Random)) }
##############################################################################################################
# Generate random normal data for shared and unique components, initialize factor loadings (steps 5, 6) --------
Shared.Comp <- matrix(rnorm(N*n.Fact,0,1),nrow=N,ncol=n.Fact)
Unique.Comp <- matrix(rnorm(N*k,0,1),nrow=N,ncol=k)
Shared.Load <- matrix(0,nrow=k,ncol=n.Fact)
Unique.Load <- matrix(0,nrow=k,ncol=1)
```

```
################################################################################################################
# Begin loop that ends when specified number of iterations pass without improvement in RMSR correlation --------
while (Trials.Without.Improvement < Max.Trials) {
  Iteration <- Iteration + 1
  ##############################################################################################################
  # Calculate factor loadings and apply to reproduce desired correlations (steps 7, 8) -----------------------
  Fact.Anal <- Factor.Analysis(Intermediate.Corr, Corr.Matrix = TRUE, n.Fact = n.Fact)
  if (n.Fact == 1) Shared.Load[,1] <- Fact.Anal$loadings else
    Shared.Load <- Fact.Anal$loadings
  Shared.Load[Shared.Load > 1] <- 1
  Shared.Load[Shared.Load < -1] <- -1
  if (Shared.Load[1,1] < 0) Shared.Load <- Shared.Load * -1
  Shared.Load.sq <- Shared.Load * Shared.Load
  for (i in 1:k)
    if (sum(Shared.Load.sq[i,]) < 1) Unique.Load[i,1] <- (1 - sum(Shared.Load.sq[i,])) else
      Unique.Load[i,1] <- 0
  Unique.Load <- sqrt(Unique.Load)
  for (i in 1:k)
    Data[,i] <- (Shared.Comp %*% t(Shared.Load))[,i] + Unique.Comp[,i] * Unique.Load[i,1]
  # the %*% operator = matrix multiplication, and the t() function = transpose (both used again in step 13)
  ##############################################################################################################
  # Replace normal with nonnormal distributions (step 9) ------------------------------------------------------
  for (i in 1:k) {
    Data <- Data[sort.list(Data[,i]),]
    Data[,i] <- Distributions[,i] }
  ##############################################################################################################
  # Calculate RMSR correlation, compare to lowest value, take appropriate action (steps 10, 11, 12) ----------
  Reproduced.Corr <- cor(Data)
  Resid.Corr <- Target.Corr - Reproduced.Corr
  RMSR <- sqrt(sum(Resid.Corr[lower.tri(Resid.Corr)]*Resid.Corr[lower.tri(Resid.Corr)])/(0.5*(k*k-k)))
```

```
132   if (RMSR < Best.RMSR) {
133     Best.RMSR <- RMSR
134     Best.Corr <- Intermediate.Corr
135     Best.Res <- Resid.Corr
136     Intermediate.Corr <- Intermediate.Corr + Init.Mult * Resid.Corr
137     Trials.Without.Improvement <- 0 } else {
138       Trials.Without.Improvement <- Trials.Without.Improvement + 1
139       Current.Multiplier <- Init.Mult * .5 ^ Trials.Without.Improvement
140       Intermediate.Corr <- Best.Corr + Current.Multiplier * Best.Res }
141 } # end of the while loop
142 ##############################################################################################################
143 # Construct the data set with the lowest RMSR correlation (step 13) ------------------------------------------
144 Fact.Anal <- Factor.Analysis(Best.Corr, Corr.Matrix = TRUE, n.Fact = n.Fact)
145 if (n.Fact == 1) Shared.Load[,1] <- Fact.Anal$loadings else
146   Shared.Load <- Fact.Anal$loadings
147 Shared.Load[Shared.Load > 1] <- 1
148 Shared.Load[Shared.Load < -1] <- -1
149 if (Shared.Load[1,1] < 0) Shared.Load <- Shared.Load * -1
150 Shared.Load.sq <- Shared.Load * Shared.Load
151 for (i in 1:k)
152   if (sum(Shared.Load.sq[i,]) < 1) Unique.Load[i,1] <- (1 - sum(Shared.Load.sq[i,])) else
153     Unique.Load[i,1] <- 0
154 Unique.Load <- sqrt(Unique.Load)
155 for (i in 1:k)
156   Data[,i] <- (Shared.Comp %*% t(Shared.Load))[,i] + Unique.Comp[,i] * Unique.Load[i,1]
157 Data <- apply(Data, 2, scale)       # standardizes each variable in the matrix
158 for (i in 1:k) {
159   Data <- Data[sort.list(Data[,i]),]
160   Data[,i] <- Distributions[,i] }
161 ##############################################################################################################
```

```
  # Report the results and return the simulated data set (step 14) ----------------------------------------------
  Iteration <- Iteration - Max.Trials
  #cat("\nN =",N,", k =",k,",",Iteration,"Iterations,",n.Fact,"Factors, RMSR r =",round(Best.RMSR,3),"\n")
  return(Data) }

Factor.Analysis <- function(Data,Corr.Matrix=FALSE,Max.Iter=50,n.Fact=0) {
  Data <- as.matrix(Data)
  k <- dim(Data)[2]
  if (n.Fact == 0) n.Fact <- k
  if (!Corr.Matrix) Cor.Matrix <- cor(Data) else
    Cor.Matrix <- Data
  Criterion <- .001
  Old.H2 <- rep(99, k)
  H2 <- rep(0, k)
  Change <- 1
  Iter <- 0
  Factor.Loadings <- matrix(nrow = k, ncol = n.Fact)
  while ((Change >= Criterion) & (Iter < Max.Iter)) {
    Iter <- Iter + 1
    Eig <- eigen(Cor.Matrix)
    L <- sqrt(Eig$values[1:n.Fact])
    for (i in 1:n.Fact)
      Factor.Loadings[,i] <- Eig$vectors[,i] * L[i]
    for (i in 1:k)
      H2[i] <- sum(Factor.Loadings[i,] * Factor.Loadings[i,])
    Change <- max(abs(Old.H2 - H2))
    Old.H2 <- H2
    diag(Cor.Matrix) <- H2 }
  if (n.Fact == k) n.Fact <- sum(Eig$values > 1)
  return(list(loadings = Factor.Loadings[,1:n.Fact], factors = n.Fact)) }
```

```
192
193 ##################################################
194 ## Function to generate Poisson distributed cases  ##
195 ##################################################
196 NullSim   <- function(N, id, x){
197   y      <- rpois(N, x*setPR)
198   #y       <- rbinom(N, x, setPR)
199   Dat     <- data.frame(Id = id, Pop = x, Obs = y)
200   PR      <- sum(Dat$Obs)/sum(Dat$Pop)
201   Dat     <- cbind(Dat, Exp = Dat$Pop*PR)
202   return(Dat)
203 }
204
205 ####################################
206 ## Set-up target correlation matrix ##
207 ####################################
208 ObsCor  <- matrix(c(1, -0.2961142, 0.8946106, 0.9707670,
209                     -0.2961142, 1, -0.3227724, -0.2930421,
210                     0.8946106, -0.3227724, 1, 0.9178227,
211                     0.9707670, -0.2930421, 0.9178227, 1), nrow = 4, ncol = 4)
212
213 ##########################################################################
214 ## Simulate 10000 datasets using GenData function to create correlation  ##
215 ## structure and approximate distributions of observed dataset           ##
216 ## Use NullSim function to generate cases of childhood leukaemia based   ##
217 ## solely on the population of 0-14 year olds. Perform sub-region and    ##
218 ## region-wide methods on simulated datasets and store results.          ##
219 ##########################################################################
220
221 ###############################################################################################
```

```
## Set-up simulation with:                                                          ##
## (1). Seed (2.) Record start time (3). Set 5-year incidence rate of leukaemia     ##
## (4). Set N = number of electoral wards (5). Set k = number of variables to generate: ##
## 0-14 population, area, 'post' in-migration, total population                     ##
## (6). Set Nsim = number of simulations (10000)                                    ##
#####################################################################################
set.seed(1123)
Beg     <- Sys.time()
setPR   <- 0.0002         # Set incidence rate for 5 year period - test against this
N       <- 532            # Number of EWs to generate
k       <- 4              # Number of variables: 0-14 population, area, 'post' in-migration, total population
Nsim    <- 10000

#############################################################################################################
## Create empty vectors to store estimates and p-values from each method performed on the 10000 datasets ##
#############################################################################################################
## Sub-region approach empty vectors
p.binom1 <- p.binom2 <- p.binom3 <- p.binom4 <- p.binom5 <- p.binom6 <- p.binom7 <- NULL
p.binom8 <- p.binom9 <- p.binom10 <- p.binom11 <- p.binom12 <- p.binom13 <- p.binom14 <- p.binom15 <- NULL
p.binom16 <- p.binom17 <- NULL

c.binom1 <- c.binom2 <- c.binom3 <- c.binom4 <- c.binom5 <- c.binom6 <- c.binom7 <- c.binom8 <- NULL
c.binom9 <- c.binom10 <- c.binom11 <- c.binom12 <- c.binom13 <- c.binom14 <- c.binom15 <- NULL
c.binom16 <- c.binom17  <- NULL

## Region-wide approach empty vectors
p.PmD    <- NULL
p.PmM    <- NULL
p.PmB    <- matrix(NA, nrow = Nsim, ncol = 2)
c.PmD    <- NULL
```

```
252 c.PmM    <- NULL
253 c.PmB    <- matrix(NA, nrow = Nsim, ncol = 2)
254
255 ##############################################################################################################
256 ## Initiate simulation using for loop. Each loop: Generates a dataset using GenData and the set up defined above,   ##
257 ## certain restraints are placed on the data as explained in the paper, each method is performed on the dataset and ##
258 ## results are stored, this is repeated until 10000 datasets have been generated and analysed.                      ##
259 ##############################################################################################################
260 for (itn in 1:Nsim){
261   # Generate 4 variables using GenData, set seed N values apart, target correlation matrix = correlation matric from observed
        data
262   # N = number of electoral wards to generate data for, k = number of variables to generate
263   YHsim     <- data.frame(GenData(seed = (itn*N), Emp = FALSE, Target.Corr = ObsCor, N = N, k = k))
264   # Assign column names and IDs to generated variables
265   YHsim     <- cbind(Id = 1:N, YHsim); names(YHsim) <- c("Id", "Pop", "Area", "InMig", "Tot_Pop")
266
267   # Replace lower than observed generated values with samples from values between minimum and median observed values
268   if (min(YHsim$Tot_Pop) < 450) YHsim$Tot_Pop[which(YHsim$Tot_Pop < 450)] <- sample(450:6000, 1, replace = TRUE)
269   if (min(YHsim$Pop) < 70) YHsim$Pop[which(YHsim$Pop < 70)] <- sample(70:1300, 1, replace = TRUE)
270   if (min(YHsim$Area) < 0.17)  YHsim$Area[which(YHsim$Area < 0.17)] <- sample(0.17:16, 1, replace = TRUE)
271
272   PrePop   <- YHsim[, 5] - YHsim[, 4]
273   if (min(PrePop) < 450)  PrePop[which(PrePop < 450)] <- sample(450:3600, 1, replace = TRUE)
274
275   # Calculate 'pre' in-migration proportions
276   PreInMig <-  (YHsim[, 4]/PrePop)
277   YHsim    <- cbind(YHsim, PreInMig = PreInMig, PreDen = PrePop/YHsim[, 3])
278
279   # Simulate Poisson distributed cases
280   OutPois  <- NullSim(length(YHsim$Id), YHsim$Id, YHsim$Pop)
```

```
SimPois  <- cbind(OutPois, Den = YHsim$PreDen, Mig = YHsim$PreInMig)

#########################
## Sub-region strategy ##
#########################
# Selection 1 - Low population density
SimTwnDO   <- SimPois[order(SimPois$Den), ]
SimTwnD2   <- SimTwnDO[0:ceiling((1/2)*length(SimTwnDO[, 1])), ]
SimTwnD    <- SimPois[sample(SimTwnD2[, 1], 16, replace = FALSE), ]

df1        <- data.frame(Cases = SimTwnD[, 3], NonCases = SimTwnD[, 2] - SimTwnD[, 3])
p.binom1[itn] <- binom.test(sum(df1[, 1]), sum(df1[, 1]) + sum(df1[, 2]), p = setPR)$p.value
c.binom1[itn] <- binom.test(sum(df1[, 1]), sum(df1[, 1]) + sum(df1[, 2]), p = setPR)$estimate

# Selection 2 - High inward-migration
SimTwnMO   <- SimPois[order(-SimPois$Mig), ]
SimTwnM2   <- SimTwnMO[0:ceiling((1/2)*length(SimTwnMO[, 1])), ]
SimTwnM    <- SimPois[sample(SimTwnM2[, 1], 16, replace = FALSE), ]

df2        <- data.frame(Cases = SimTwnM[, 3], NonCases = SimTwnM[, 2] - SimTwnM[, 3])
p.binom2[itn] <- binom.test(sum(df2[, 1]), sum(df2[, 1]) + sum(df2[, 2]), p = setPR)$p.value
c.binom2[itn] <- binom.test(sum(df2[, 1]), sum(df2[, 1]) + sum(df2[, 2]), p= setPR)$estimate

# Selection 3 - High incidence
SimObs     <- cbind(SimPois, Inc = SimPois$Obs/SimPois$Exp)
SimTwnI1   <- SimObs[order(-SimObs$Inc), ]
SimTwnI2   <- SimTwnI1[0:ceiling((1/2)*length(SimTwnI1[, 1])), ]
SimTwnI    <- SimObs[sample(SimTwnI2[, 1], 16, replace = FALSE), ]

df3        <- data.frame(Cases = SimTwnI[, 3], NonCases = SimTwnI[, 2] - SimTwnI[, 3])
```

```
311 p.binom3[itn] <- binom.test(sum(df3[, 1]), sum(df3[, 1]) + sum(df3[, 2]), p = setPR)$p.value
312 c.binom3[itn] <- binom.test(sum(df3[, 1]), sum(df3[, 1]) + sum(df3[, 2]), p = setPR)$estimate
313
314 # Selection 4 - Low population density and high inward-migration
315 OrdD4      <- SimObs[order(SimObs$Den), ]
316 SmpD4      <- OrdD4[1:ceiling(0.5*length(OrdD4[, 1])), ]
317 OrdM4      <- SmpD4[order(-SmpD4$Mig), ]
318 SmpM4      <- OrdM4[1:ceiling(0.5*length(OrdM4[, 1])), ]
319 Smp4       <- SimObs[sample(SmpM4[, 1], 16, replace = FALSE), ]
320
321 df4      <- data.frame(Cases = Smp4[, 3], NonCases = Smp4[, 2] - Smp4[, 3])
322 p.binom4[itn] <- binom.test(sum(df4[, 1]), sum(df4[, 1]) + sum(df4[, 2]), p = setPR)$p.value
323 c.binom4[itn] <- binom.test(sum(df4[, 1]), sum(df4[, 1]) + sum(df4[, 2]), p = setPR)$estimate
324
325 # Selection 5 - High inward-migration and low population density
326 OrdM5      <- SimObs[order(-SimObs$Mig), ]
327 SmpM5      <- OrdM5[1:ceiling(0.5*length(OrdM5[, 1])), ]
328 OrdD5      <- SmpM5[order(SmpM5$Den), ]
329 SmpD5      <- OrdD5[1:ceiling(0.5*length(OrdD5[, 1])), ]
330 Smp5       <- SimObs[sample(SmpD5[, 1], 16, replace = FALSE), ]
331
332 df5      <- data.frame(Cases = Smp5[, 3], NonCases = Smp5[, 2] - Smp5[, 3])
333 p.binom5[itn] <- binom.test(sum(df5[, 1]), sum(df5[, 1]) + sum(df5[, 2]), p = setPR)$p.value
334 c.binom5[itn] <- binom.test(sum(df5[, 1]), sum(df5[, 1]) + sum(df5[, 2]), p = setPR)$estimate
335
336 # Selection 6 - Low population density and high incidence
337 OrdD6      <- SimObs[order(SimObs$Den), ]
338 SmpD6      <- OrdD6[1:ceiling(0.5*length(OrdD6[, 1])), ]
339 OrdI6      <- SmpD6[order(-SmpD6$Inc), ]
340 SmpI6      <- OrdI6[1:ceiling(0.5*length(OrdI6[, 1])), ]
```

```
Smp6       <- SimObs[sample(SmpI6[, 1], 16, replace = FALSE), ]

df6      <- data.frame(Cases = Smp6[, 3], NonCases = Smp6[, 2] - Smp6[, 3])
p.binom6[itn] <- binom.test(sum(df6[, 1]), sum(df6[, 1]) + sum(df6[, 2]), p = setPR)$p.value
c.binom6[itn] <- binom.test(sum(df6[, 1]), sum(df6[, 1]) + sum(df6[, 2]), p = setPR)$estimate

# Selection 7 - High incidence and low population density
OrdI7      <- SimObs[order(-SimObs$Inc), ]
SmpI7      <- OrdI7[1:ceiling(0.5*length(OrdI7[, 1])), ]
OrdD7      <- SmpI7[order(SmpI7$Den), ]
SmpD7      <- OrdD7[1:ceiling(0.5*length(OrdD7[, 1])), ]
Smp7       <- SimObs[sample(SmpD7[, 1], 16, replace = FALSE), ]

df7      <- data.frame(Cases = Smp7[, 3], NonCases = Smp7[, 2] - Smp7[, 3])
p.binom7[itn] <- binom.test(sum(df7[, 1]), sum(df7[, 1]) + sum(df7[, 2]), p = setPR)$p.value
c.binom7[itn] <- binom.test(sum(df7[, 1]), sum(df7[, 1]) + sum(df7[, 2]), p = setPR)$estimate

# Selection 8 - High inward-migration and high incidence
OrdM8      <- SimObs[order(-SimObs$Mig), ]
SmpM8      <- OrdM8[1:ceiling(0.5*length(OrdM8[, 1])), ]
OrdI8      <- SmpM8[order(-SmpM8$Inc), ]
SmpI8      <- OrdI8[1:ceiling(0.5*length(OrdI8[, 1])), ]
Smp8       <- SimObs[sample(SmpI8[, 1], 16, replace = FALSE), ]

df8      <- data.frame(Cases = Smp8[, 3], NonCases = Smp8[, 2] - Smp8[, 3])
p.binom8[itn] <- binom.test(sum(df8[, 1]), sum(df8[, 1]) + sum(df8[, 2]), p = setPR)$p.value
c.binom8[itn] <- binom.test(sum(df8[, 1]), sum(df8[, 1]) + sum(df8[, 2]), p = setPR)$estimate

# Selection 9 - High incidence and high inward-migration
OrdI9      <- SimObs[order(-SimObs$Inc), ]
```

```
371 SmpI9      <- OrdI9[1:ceiling(0.5*length(OrdI9[, 1])), ]
372 OrdM9      <- SmpI9[order(-SmpI9$Mig), ]
373 SmpM9      <- OrdM9[1:ceiling(0.5*length(OrdM9[, 1])), ]
374 Smp9       <- SimObs[sample(SmpM9[, 1], 16, replace = FALSE), ]
375
376 df9      <- data.frame(Cases = Smp9[, 3], NonCases = Smp9[, 2] - Smp9[, 3])
377 p.binom9[itn] <- binom.test(sum(df9[, 1]), sum(df9[, 1]) + sum(df9[, 2]), p = setPR)$p.value
378 c.binom9[itn] <- binom.test(sum(df9[, 1]), sum(df9[, 1]) + sum(df9[, 2]), p= setPR)$estimate
379
380 # Selection 10 - Low population density, high inward-migration and high incidence
381 OrdD10    <- SimObs[order(SimObs$Den), ]
382 SmpD10    <- OrdD10[1:ceiling(0.5*length(OrdD10[, 1])), ]
383 OrdM10    <- SmpD10[order(-SmpD10$Mig), ]
384 SmpM10    <- OrdM10[1:ceiling(0.5*length(OrdM10[, 1])), ]
385 OrdI10    <- SmpM10[order(-SmpM10$Inc), ]
386 Smp10     <- OrdI10[1:ceiling(0.5*length(OrdI10[, 1])), ]
387 Smp10     <- SimObs[sample(Smp10[, 1], 16, replace = FALSE), ]
388
389 df10      <- data.frame(Cases = Smp10[, 3], NonCases = Smp10[, 2] - Smp10[, 3])
390 p.binom10[itn] <- binom.test(sum(df10[, 1]), sum(df10[, 1]) + sum(df10[, 2]), p = setPR)$p.value
391 c.binom10[itn] <- binom.test(sum(df10[, 1]), sum(df10[, 1]) + sum(df10[, 2]), p = setPR)$estimate
392
393 # Selection 11 - Low population density, high incidence and high inward-migration
394 OrdD11    <- SimObs[order(SimObs$Den), ]
395 SmpD11    <- OrdD11[1:ceiling(0.5*length(OrdD11[, 1])), ]
396 OrdI11    <- SmpD11[order(-SmpD11$Inc), ]
397 SmpI11    <- OrdI11[1:ceiling(0.5*length(OrdI11[, 1])), ]
398 OrdM11    <- SmpI11[order(-SmpI11$Mig), ]
399 Smp11     <- OrdM11[1:ceiling(0.5*length(OrdM11[, 1])), ]
400 Smp11    <- SimObs[sample(Smp11[, 1], 16, replace = FALSE), ]
```

```
401
402 df11      <- data.frame(Cases = Smp11[, 3], NonCases = Smp11[, 2] - Smp11[, 3])
403 p.binom11[itn] <- binom.test(sum(df11[, 1]), sum(df11[, 1]) + sum(df11[, 2]), p = setPR)$p.value
404 c.binom11[itn] <- binom.test(sum(df11[, 1]), sum(df11[, 1]) + sum(df11[, 2]), p = setPR)$estimate
405
406 # Selection 12 - High inward-migration, low population density and high incidence
407 OrdM12    <- SimObs[order(-SimObs$Mig), ]
408 SmpM12    <- OrdM12[1:ceiling(0.5*length(OrdM12[, 1])), ]
409 OrdD12    <- SmpM12[order(SmpM12$Den), ]
410 SmpD12    <- OrdD12[1:ceiling(0.5*length(OrdD12[, 1])), ]
411 OrdI12    <- SmpD12[order(-SmpD12$Inc), ]
412 Smp12     <- OrdI12[1:ceiling(0.5*length(OrdI12[, 1])), ]
413 Smp12    <- SimObs[sample(Smp12[, 1], 16, replace = FALSE), ]
414
415 df12      <- data.frame(Cases = Smp12[, 3], NonCases = Smp12[, 2] - Smp12[, 3])
416 p.binom12[itn] <- binom.test(sum(df12[, 1]), sum(df12[, 1]) + sum(df12[, 2]), p = setPR)$p.value
417 c.binom12[itn] <- binom.test(sum(df12[, 1]), sum(df12[, 1]) + sum(df12[, 2]), p = setPR)$estimate
418
419 # Selection 13 - High inward-migration, high incidence and low population density
420 OrdM13    <- SimObs[order(-SimObs$Mig), ]
421 SmpM13    <- OrdM13[1:ceiling(0.5*length(OrdM13[, 1])), ]
422 OrdI13    <- SmpM13[order(-SmpM13$Inc), ]
423 SmpI13    <- OrdI13[1:ceiling(0.5*length(OrdI13[, 1])), ]
424 OrdD13    <- SmpI13[order(SmpI13$Den), ]
425 Smp13     <- OrdI13[1:ceiling(0.5*length(OrdD13[, 1])), ]
426 Smp13    <- SimObs[sample(Smp13[, 1], 16, replace = FALSE), ]
427
428 df13      <- data.frame(Cases = Smp13[, 3], NonCases = Smp13[, 2] - Smp13[, 3])
429 p.binom13[itn] <- binom.test(sum(df13[, 1]), sum(df13[, 1]) + sum(df13[, 2]), p = setPR)$p.value
430 c.binom13[itn] <- binom.test(sum(df13[, 1]), sum(df13[, 1]) + sum(df13[, 2]), p = setPR)$estimate
```

```
431
432 # Selection 14 - High incidence, low population density and high inward-migration
433 OrdI14    <- SimObs[order(-SimObs$Inc), ]
434 SmpI14    <- OrdI14[1:ceiling(0.5*length(OrdI14[, 1])), ]
435 OrdD14    <- SmpI14[order(SmpI14$Den), ]
436 SmpD14    <- OrdD14[1:ceiling(0.5*length(OrdD14[, 1])), ]
437 OrdM14    <- SmpD14[order(-SmpD14$Mig), ]
438 Smp14     <- OrdM14[1:ceiling(0.5*length(OrdM14[, 1])), ]
439 Smp14    <- SimObs[sample(Smp14[, 1], 16, replace = FALSE), ]
440
441 df14      <- data.frame(Cases = Smp14[, 3], NonCases = Smp14[, 2] - Smp14[, 3])
442 p.binom14[itn] <- binom.test(sum(df14[, 1]), sum(df14[, 1]) + sum(df14[, 2]), p = setPR)$p.value
443 c.binom14[itn] <- binom.test(sum(df14[, 1]), sum(df14[, 1]) + sum(df14[, 2]), p = setPR)$estimate
444
445 # Selection 15 - High incidence, high inward-migration and low population density
446 OrdI15    <- SimObs[order(-SimObs$Inc), ]
447 SmpI15    <- OrdI15[1:ceiling(0.5*length(OrdI15[, 1])), ]
448 OrdM15    <- SmpI15[order(-SmpI15$Mig), ]
449 SmpM15    <- OrdM15[1:ceiling(0.5*length(OrdM15[, 1])), ]
450 OrdD15    <- SmpM15[order(SmpM15$Den), ]
451 Smp15    <- OrdD15[1:ceiling(0.5*length(OrdD15[, 1])), ]
452 Smp15    <- SimObs[sample(Smp15[, 1], 16, replace = FALSE), ]
453
454 df15     <- data.frame(Cases = Smp15[, 3], NonCases = Smp15[, 2] - Smp15[, 3])
455 p.binom15[itn] <- binom.test(sum(df15[, 1]), sum(df15[, 1]) + sum(df15[, 2]), p = setPR)$p.value
456 c.binom15[itn] <- binom.test(sum(df15[, 1]), sum(df15[, 1]) + sum(df15[, 2]), p = setPR)$estimate
457
458 # Selection 16 - Random selection of 16 wards
459 Smp16     <- SimObs[sample(SimObs[, 1], 16, replace = FALSE), ]
460
```

```
df16     <- data.frame(Cases = Smp16[, 3], NonCases = Smp16[, 2] - Smp16[, 3])
p.binom16[itn] <- binom.test(sum(df16[, 1]), sum(df16[, 1]) + sum(df16[, 2]), p = setPR)$p.value
c.binom16[itn] <- binom.test(sum(df16[, 1]), sum(df16[, 1]) + sum(df16[, 2]), p = setPR)$estimate

# Selection 17 - Incidence less than average
SimObs     <- cbind(SimPois, Inc = SimPois$Obs/SimPois$Exp)
SimTwnI17   <- SimObs[order(-SimObs$Inc), ]
SimTwnI17   <- SimTwnI17[ceiling((1/2)*length(SimTwnI17[, 1])):length(SimTwnI17[, 1]), ]
SimTwnLI    <- SimObs[sample(SimTwnI17[, 1], 16, replace = FALSE), ]

df17        <- data.frame(Cases = SimTwnLI[, 3], NonCases = SimTwnLI[, 2] - SimTwnLI[, 3])
p.binom17[itn] <- binom.test(sum(df17[, 1]), sum(df17[, 1]) + sum(df17[, 2]), p = setPR)$p.value
c.binom17[itn] <- binom.test(sum(df17[, 1]), sum(df17[, 1]) + sum(df17[, 2]), p = setPR)$estimate

##########################
## Region-wide strategy ##
##########################

# Create Poisson models
SimHalf    <- SimPois[sample(SimPois[, 1], 266, replace = FALSE), ]
PmDtmp     <- glm(Obs ~ offset(log(Pop)) + Den, data = SimHalf, family = poisson(link = log))
PmMtmp     <- glm(Obs ~ offset(log(Pop)) + Mig, data = SimHalf, family = poisson(link = log))
PmBtmp     <- glm(Obs ~ offset(log(Pop)) + Den + Mig, data = SimHalf, family = poisson(link = log))

# Store point estimates
c.PmD[itn]   <- PmDtmp$coefficients[2]
c.PmM[itn]   <- PmMtmp$coefficients[2]
c.PmB[itn, 1] <- PmBtmp$coefficients[2]
c.PmB[itn, 2] <- PmBtmp$coefficients[3]

```

```
491   # Store p-values
492   p.PmD[itn]   <- summary(PmDtmp)$coefficients[8]
493   p.PmM[itn]   <- summary(PmMtmp)$coefficients[8]
494   p.PmB[itn, 1] <- summary(PmBtmp)$coefficients[11]
495   p.PmB[itn, 2] <- summary(PmBtmp)$coefficients[12]
496 }
497 End   <- Sys.time()
498
499 ## Find length of time code took to run
500 End-Beg
501
502 ######################################################
503 ## Calculate Type I error rate of sub-region method ##
504 ######################################################
505
506 Bpdists <- data.frame(BE1 = p.binom1,  BE2 = p.binom2,  BE3 = p.binom3, BE4 = p.binom4, BE5 = p.binom5, BE6 = p.binom6, BE7 =
        p.binom7, BE8 = p.binom8,
507                       BE9 = p.binom9, BE10 = p.binom10, BE11 = p.binom11, BE12 = p.binom12, BE13 = p.binom13, BE14 = p.
                          binom14, BE15 = p.binom15, BE16 = p.binom16, BE17 = p.binom17)
508
509 ## Calculate percentages of critical p-values
510 B.1pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE1 <= x)}))/itn)
511 B.2pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE2 <= x)}))/itn)
512 B.3pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE3 <= x)}))/itn)
513 B.4pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE4 <= x)}))/itn)
514 B.5pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE5 <= x)}))/itn)
515 B.6pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE6 <= x)}))/itn)
516 B.7pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE7 <= x)}))/itn)
517 B.8pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE8 <= x)}))/itn)
518 B.9pPct    <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE9 <= x)}))/itn)
```

```
B.10pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE10 <= x)}))/itn)
B.11pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE11 <= x)}))/itn)
B.12pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE12 <= x)}))/itn)
B.13pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE13 <= x)}))/itn)
B.14pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE14 <= x)}))/itn)
B.15pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE15 <= x)}))/itn)
B.16pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE16 <= x)}))/itn)
B.17pPct   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(Bpdists$BE17 <= x)}))/itn)

B.1pPct; B.2pPct; B.3pPct; B.4pPct; B.5pPct; B.6pPct; B.7pPct; B.8pPct; B.9pPct; B.10pPct; B.11pPct; B.12pPct; B.13pPct; B.14
    pPct; B.15pPct; B.16pPct; B.17pPct

## Store p-values of sub-region method
write.table(Bpdists, "p-distributions binomial test - all combinations.csv", sep = ",", col.names = TRUE)

#Bpdists <- read.csv("p-distributions binomial test - all combinations .csv", sep = ",", header = TRUE)

#########################################################
## Calculate Type I error rates for region-wide method ##
#########################################################

Pct.p.PmM     <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(p.PmM <= x)}))/itn)
Pct.p.PmD     <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(p.PmD <= x)}))/itn)
Pct.p.PmB.M   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(p.PmB[, 2] <= x)}))/itn)
Pct.p.PmB.D   <- c(100*do.call(c, lapply(c(0.05, 0.01, 0.001), function(x){sum(p.PmB[, 1] <= x)}))/itn)

Pct.p.PmD; Pct.p.PmM; Pct.p.PmB.D; Pct.p.PmB.M;

# Store p-values of region-wide method
pdists <- data.frame(PmM = p.PmM, PmD = p.PmD, PmBM = p.PmB[, 2], PmBD = p.PmB[, 1])
```

```
548 write.table(pdists, "p-distributions regression - all combinations.csv", sep = ",", col.names = TRUE)
549
550 ## Plot the p-values for the region-wide method
551 pdists <- data.frame("Inward-migration" = pdists[, 1], "Population density" = pdists[, 2], "Inward-migration adj. Population
        density" = pdists[, 3],
552                      "Population density - adj. Inward-migration" = pdists[, 4])
553 pdists <- stack(pdists)
554 Xlab <- "p-values"; Ylab <- "Kernel Density"
555 Mlab    <- "Distribution of p-values for the 'whole region' selection strategy"; windows()
556 ggplot(pdists, aes(x = values)) + geom_density(aes(group = ind, colour = ind), size = 1.2) +
557   labs(title = Mlab, x = Xlab, y = Ylab, colour = NULL) +
558   scale_colour_manual(values = cbPal, name = "Covariate", label = c("Inward-Migration", "Population Density", "Inward-
          Migration (adj. Population Density)",
559                                                                     "Population Density (adj. Inward-Migration)")) +
560   theme_bw() + theme(axis.title = element_text(family = "Times New Roman", size = 16),
561                      plot.title = element_text(size = 16, family = "Times New Roman", face = "bold"), legend.position = "top"
                           ,
562                      legend.text = element_text(family = "Times New Roman", size = 12),
563                      legend.title = element_text(family = "Times New Roman", size = 12, face = "bold"))
564
565
566 ## Plot the p-values for the sub-region method
567 Bpdists <- data.frame("High Inward-Migration" = Bpdists[, 1], "Low Population Density" = Bpdists[, 2], "High In-ward
        Migration and Low Population Density" = Bpdists[, 3],
568                       "High Incidence" = Bpdists[, 4])
569 Bpdists <- stack(Bpdists)
570 Xlab <- "p-values"; Ylab <- "Kernel Density"
571 Mlab    <- "Distribution of p-values for the 'sub-sample' selection strategy"; windows()
572 ggplot(Bpdists, aes(x = values)) + geom_density(aes(group = ind, colour = ind), size = 1.2) +
573   labs(title = Mlab, x = Xlab, y = Ylab, colour = NULL) +
```

```
574   scale_colour_manual(values = cbPal, name = "Selection", label = c("High Inward-Migration", "Low Population Density", "High
          Inward-Migration and Low Population Density",
575                                                                    "High Incidence")) +
576   theme_bw() + theme(axis.title = element_text(family = "Times New Roman", size = 16), plot.title = element_text(size = 16,
          family = "Times New Roman", face = "bold"), legend.position = "top",
577                      legend.text = element_text(family = "Times New Roman", size = 12), legend.title = element_text(family =
                          "Times New Roman", size = 12,face = "bold"))
578
579 ####################
580 ## Visualisation  ##
581 ####################
582
583 ##############################################################################
584 ## Region-wide and sub-sample selection analyses performed on observed data ##
585 ##############################################################################
586
587 ## Omitted as observed dataset not publically available
588
589 Estimatesdf  <- data.frame(Estimates =  rbind(S1CI, S2CI, S3CI, S4CI, S5CI, S6CI, S7CI, S8CI,
590                                               S9CI, S10CI, S11CI, S12CI, S13CI, S14CI, S15CI),
591                            Label = c("S1: D", "S2: M", "S3: I", "S4: DM", "S5: MD", "S6: DI", "S7: ID", "S8: MI",
592                                      "S9: IM", "S10: DMI", "S11: DIM", "S12: MDI", "S13: MID", "S14: IDM", "S15: IMD"),
593                            Mod = c(rep("Sub-region - Simulated", 15)))
594
595 EstimatesdfSim <- Estimatesdf[1:15, ]
596 EstimatesdfSim$Label <- factor(EstimatesdfSim$Label, as.character(EstimatesdfSim$Label))
597
598 x11()
599 ggplot(EstimatesdfSim, aes(x = Label, y = Estimates.50., group = Mod)) +
600   geom_point() +
```

```
601   geom_errorbar(data = EstimatesdfSim, aes(ymin = Estimates.2.5., ymax = Estimates.97.5.), width = .2)
602
603 Estimatesdf$Label  <- factor(Estimatesdf$Label, unique(as.character(Estimatesdf$Label)))
604
605 ###################################################
606 ## Figure 3 of paper - showing simulated data only ##
607 ###################################################
608
609 x11()
610 ggplot(Estimatesdf, aes(x = Label, y = Estimates.50., group = Mod, colour = Mod)) +
611   geom_point(position = position_dodge(width = .5)) +
612   geom_errorbar(data = Estimatesdf, aes(ymin = Estimates.2.5., ymax = Estimates.97.5., colour = Mod), width = .5, position =
          position_dodge(width = .5)) +
613   geom_hline(yintercept = 0.0002, size = 0.2, colour = "black", linetype = "dashed") + theme_bw() +
614   scale_y_continuous(name = "Estimate of Childhood Leukemia Incidence") + scale_x_discrete(name = "") +
615   annotate("text", label = "Simulated Null", x = 1.0, y = 0.000225, size = 5, colour = "black")  +
616   theme(axis.text = element_text(colour = "black", size = 12, family = "Times New Roman"),
617         axis.title = element_text(colour = "black", size = 20, face = "bold", family = "Times New Roman"),
618         plot.title = element_text(size = 16, face = "bold", family = "Times New Roman"), legend.position = c(0.1, 0.95),
619         legend.text = element_text(size = 16, family = "Times New Roman"), axis.text.x = element_text(vjust = 0.5),
620         legend.key = element_blank(), legend.title = element_blank())
621
622 ###########################################################################
623 ## Create dataframe of coefficients according to specific covariate values ##
624 ###########################################################################
625
626 # Log scale
627 Coef10pc <- exp(0.1*MCI); Coef25pc <- exp(0.25*MCI); Coef50pc <- exp(0.5*MCI)
628 Coef10pcadj <- exp(0.1*BMCI); Coef25pcadj <- exp(0.25*BMCI); Coef50pcadj <- exp(0.5*BMCI)
629 Coef100d <- exp(100*DCI); Coef500d <- exp(500*DCI); Coef1000d <- exp(1000*DCI)
```

```
630 Coef100dadj <- exp(100*BDCI); Coef500dadj <- exp(500*BDCI); Coef1000dadj <- exp(1000*BDCI)
631
632 ##################################################################################################################
633 ## Plot confidence intervals for 500 persons/km^2 increase in population density and 25% increase in population ##
634 ##################################################################################################################
635
636 Coeffsdf  <- data.frame(Coeffs = rbind(Coef25pc, Coef25pcadj, Coef500d, Coef500dadj),
637                         Label= c("Inward-Migration \n25%",
638                                  "Inward-Migration \n(adj. Population Density) - 25%",
639                                  "Population Density \n500 persons per km^2",
640                                  "Population Density \n(adj. Inward-Migration) \n500 persons per km^2"),
641                         Mod=c(rep("Region-wide - Simulated", 4)))
642
643 Coeffsdf$Label  <- factor(Coeffsdf$Label, levels = unique(Coeffsdf$Label))
644
645 ###########################################
646 ## Figure 4 of paper - simulated data only ##
647 ###########################################
648
649 x11()
650 ggplot(Coeffsdf,aes(x = Label, y = Coeffs.50., group = Mod, colour = Mod)) +
651   scale_y_continuous(name = "Risk Ratio", breaks = c(0, 1, 2, 4, 6, 8, 10)) +
652   geom_point(position = position_dodge(width = .5)) +
653   geom_errorbar(data = Coeffsdf, aes(ymin = Coeffs.2.5., ymax = Coeffs.97.5., colour = Mod), width = .5, position = position_
        dodge(width = .5)) +
654   geom_hline(yintercept = 1.00, size = 0.2, colour = "black", linetype = "dashed") + theme_bw() +
655   scale_x_discrete(name = "") +
656   theme(legend.title = element_blank()) +
657   theme(axis.text = element_text(colour = "black", size=12, family = "Times New Roman"),
658         axis.title = element_text(size = 16, family = "Times New Roman", face = "bold"),
```

```
659         plot.title = element_text(size = 16, face = "bold", family = "Times New Roman"), legend.position = "top",
660         legend.text = element_text(size = 12, family = "Times New Roman")) + theme(axis.text.x = element_text(angle = 90,
                hjust = 1)) +
661   annotate("text", label = "Simulated Null", x = 0.45, y = 1.25, size = 5, colour = "black", family = "Times New Roman") +
662   coord_trans(y = "log10")
663
664 ########################################
665 ## Summary of significant coefficients ##
666 ########################################
667
668 ## Summary of proportion of inward-migration coefficient
669 summary(Sp.PmMSubHi)
670 summary(Sp.PmMSubLo)
671 summary(Sp.PmMtmpSub)
672
673 ## Summary of population density coefficient
674 summary(Sp.PmDSubHi)
675 summary(Sp.PmDSubLo)
676 summary(Sp.PmDtmpSub)
677
678 ## Summary of proportion of inward-migration coefficient (adjusted for population density)
679 summary(Sp.PmBMSubHi)
680 summary(Sp.PmBMSubLo)
681 summary(Sp.PmBMtmpSub)
682
683 ## Summary of proportion of inward-migration coefficient (adjusted for population density)
684 summary(Sp.PmDSubHi)
685 summary(Sp.PmDSubLo)
686 summary(Sp.PmBDtmpSub)
687
```

```
###
S.binomS1tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS2tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS3tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS4tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS5tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS6tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS7tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS8tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS9tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS10tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS11tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS12tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS13tmp    <- matrix(NA, nrow = 10000, ncol = 2); S.binomS14tmp    <- matrix(NA, nrow = 10000, ncol = 2)
S.binomS15tmp    <- matrix(NA, nrow = 10000, ncol = 2);

S.binomS1tmp[, 1] <- c.binom1; S.binomS1tmp[, 2] <- p.binom1
S.binomS2tmp[, 1] <- c.binom2; S.binomS2tmp[, 2] <- p.binom2
S.binomS3tmp[, 1] <- c.binom3; S.binomS3tmp[, 2] <- p.binom3
S.binomS4tmp[, 1] <- c.binom4; S.binomS4tmp[, 2] <- p.binom4
S.binomS5tmp[, 1] <- c.binom5; S.binomS5tmp[, 2] <- p.binom5
S.binomS6tmp[, 1] <- c.binom6; S.binomS6tmp[, 2] <- p.binom6
S.binomS7tmp[, 1] <- c.binom7; S.binomS7tmp[, 2] <- p.binom7
S.binomS8tmp[, 1] <- c.binom8; S.binomS8tmp[, 2] <- p.binom8
S.binomS9tmp[, 1] <- c.binom9; S.binomS9tmp[, 2] <- p.binom9
S.binomS10tmp[, 1] <- c.binom10; S.binomS10tmp[, 2] <- p.binom10
S.binomS11tmp[, 1] <- c.binom11; S.binomS11tmp[, 2] <- p.binom11
S.binomS12tmp[, 1] <- c.binom12; S.binomS12tmp[, 2] <- p.binom12
S.binomS13tmp[, 1] <- c.binom13; S.binomS13tmp[, 2] <- p.binom13
S.binomS14tmp[, 1] <- c.binom14; S.binomS14tmp[, 2] <- p.binom14
S.binomS15tmp[, 1] <- c.binom15; S.binomS15tmp[, 2] <- p.binom15

S.binomS1tmpSub <- S.binomS1tmp[which(S.binomS1tmp[, 2] < 0.05), ]
S.binomS1SubHi  <- S.binomS1tmpSub[which(S.binomS1tmpSub[, 1] > 0.0002), ]
S.binomS1SubLo  <- S.binomS1tmpSub[which(S.binomS1tmpSub[, 1] < 0.0002), ]
S.binomS1Sub <- data.frame(cbind(-length(S.binomS1SubLo[, 1])/100, length(S.binomS1SubHi[, 1])/100))
```

```
718
719 S.binomS2tmpSub <- S.binomS2tmp[which(S.binomS2tmp[, 2] < 0.05), ]
720 S.binomS2SubHi  <- S.binomS2tmpSub[which(S.binomS2tmpSub[, 1] > 0.0002), ]
721 S.binomS2SubLo  <- S.binomS2tmpSub[which(S.binomS2tmpSub[, 1] < 0.0002), ]
722 S.binomS2Sub <- data.frame(cbind(-length(S.binomS2SubLo[, 1])/100, length(S.binomS2SubHi[, 1])/100))
723
724 S.binomS3tmpSub <- S.binomS3tmp[which(S.binomS3tmp[, 2] < 0.05), ]
725 S.binomS3SubHi  <- S.binomS3tmpSub[which(S.binomS3tmpSub[, 1] > 0.0002), ]
726 S.binomS3SubLo  <- S.binomS3tmpSub[which(S.binomS3tmpSub[, 1] < 0.0002), ]
727 S.binomS3Sub <- data.frame(cbind(-length(S.binomS3SubLo[, 1])/100, length(S.binomS3SubHi[, 1])/100))
728
729 S.binomS4tmpSub <- S.binomS4tmp[which(S.binomS4tmp[, 2] < 0.05), ]
730 S.binomS4SubHi  <- S.binomS4tmpSub[which(S.binomS4tmpSub[, 1] > 0.0002), ]
731 S.binomS4SubLo  <- S.binomS4tmpSub[which(S.binomS4tmpSub[, 1] < 0.0002), ]
732 S.binomS4Sub <- data.frame(cbind(-length(S.binomS4SubLo[1])/100, length(S.binomS4SubHi[, 1])/100))
733
734 S.binomS5tmpSub <- S.binomS5tmp[which(S.binomS5tmp[, 2] < 0.05), ]
735 S.binomS5SubHi  <- S.binomS5tmpSub[which(S.binomS5tmpSub[, 1] > 0.0002), ]
736 S.binomS5SubLo  <- S.binomS5tmpSub[which(S.binomS5tmpSub[, 1] < 0.0002), ]
737 S.binomS5Sub <- data.frame(cbind(-length(S.binomS5SubLo[1])/100, length(S.binomS5SubHi[, 1])/100))
738
739 S.binomS6tmpSub <- S.binomS6tmp[which(S.binomS6tmp[, 2] < 0.05), ]
740 S.binomS6SubHi  <- S.binomS6tmpSub[which(S.binomS6tmpSub[, 1] > 0.0002), ]
741 S.binomS6SubLo  <- S.binomS6tmpSub[which(S.binomS6tmpSub[, 1] < 0.0002), ]
742 S.binomS6Sub <- data.frame(cbind(-length(S.binomS6SubLo[, 1])/100, length(S.binomS6SubHi[, 1])/100))
743
744 S.binomS7tmpSub <- S.binomS7tmp[which(S.binomS7tmp[, 2] < 0.05), ]
745 S.binomS7SubHi  <- S.binomS7tmpSub[which(S.binomS7tmpSub[, 1] > 0.0002), ]
746 S.binomS7SubLo  <- S.binomS7tmpSub[which(S.binomS7tmpSub[, 1] < 0.0002), ]
747 S.binomS7Sub <- data.frame(cbind(-length(S.binomS7SubLo[, 1])/100, length(S.binomS7SubHi[, 1])/100))
```

```
748
749 S.binomS8tmpSub <- S.binomS8tmp[which(S.binomS8tmp[, 2] < 0.05), ]
750 S.binomS8SubHi  <- S.binomS8tmpSub[which(S.binomS8tmpSub[, 1] > 0.0002), ]
751 S.binomS8SubLo  <- S.binomS8tmpSub[which(S.binomS8tmpSub[, 1] < 0.0002), ]
752 S.binomS8Sub <- data.frame(cbind(-length(S.binomS8SubLo[1])/100, length(S.binomS8SubHi[, 1])/100))
753
754 S.binomS9tmpSub <- S.binomS9tmp[which(S.binomS9tmp[, 2] < 0.05), ]
755 S.binomS9SubHi  <- S.binomS9tmpSub[which(S.binomS9tmpSub[, 1] > 0.0002), ]
756 S.binomS9SubLo  <- S.binomS9tmpSub[which(S.binomS9tmpSub[, 1] < 0.0002), ]
757 S.binomS9Sub <- data.frame(cbind(-length(S.binomS9SubLo[, 1])/100, length(S.binomS9SubHi[, 1])/100))
758
759 S.binomS10tmpSub <- S.binomS10tmp[which(S.binomS10tmp[, 2] < 0.05), ]
760 S.binomS10SubHi  <- S.binomS10tmpSub[which(S.binomS10tmpSub[, 1] > 0.0002), ]
761 S.binomS10SubLo  <- S.binomS10tmpSub[which(S.binomS10tmpSub[, 1] < 0.0002), ]
762 S.binomS10Sub <- data.frame(cbind(-length(S.binomS10SubLo[, 1])/100, length(S.binomS10SubHi[, 1])/100))
763
764 S.binomS11tmpSub <- S.binomS11tmp[which(S.binomS11tmp[, 2] < 0.05), ]
765 S.binomS11SubHi  <- S.binomS11tmpSub[which(S.binomS11tmpSub[, 1] > 0.0002), ]
766 S.binomS11SubLo  <- S.binomS11tmpSub[which(S.binomS11tmpSub[, 1] < 0.0002), ]
767 S.binomS11Sub <- data.frame(cbind(-length(S.binomS11SubLo[, 1])/100, length(S.binomS11SubHi[, 1])/100))
768
769 S.binomS12tmpSub <- S.binomS12tmp[which(S.binomS12tmp[, 2] < 0.05), ]
770 S.binomS12SubHi  <- S.binomS12tmpSub[which(S.binomS12tmpSub[, 1] > 0.0002), ]
771 S.binomS12SubLo  <- S.binomS12tmpSub[which(S.binomS12tmpSub[, 1] < 0.0002), ]
772 S.binomS12Sub <- data.frame(cbind(-length(S.binomS12SubLo[, 1])/100, length(S.binomS12SubHi[, 1])/100))
773
774 S.binomS13tmpSub <- S.binomS13tmp[which(S.binomS13tmp[, 2] < 0.05), ]
775 S.binomS13SubHi  <- S.binomS13tmpSub[which(S.binomS13tmpSub[, 1] > 0.0002), ]
776 S.binomS13SubLo  <- S.binomS13tmpSub[which(S.binomS13tmpSub[, 1] < 0.0002), ]
777 S.binomS13Sub <- data.frame(cbind(-length(S.binomS13SubLo[, 1])/100, length(S.binomS13SubHi[, 1])/100))
```

```
778
779 S.binomS14tmpSub <- S.binomS14tmp[which(S.binomS14tmp[, 2] < 0.05), ]
780 S.binomS14SubHi  <- S.binomS14tmpSub[which(S.binomS14tmpSub[, 1] > 0.0002), ]
781 S.binomS14SubLo  <- S.binomS14tmpSub[which(S.binomS14tmpSub[, 1] < 0.0002), ]
782 S.binomS14Sub <- data.frame(cbind(-length(S.binomS14SubLo[, 1])/100, length(S.binomS14SubHi[, 1])/100))
783
784 S.binomS15tmpSub <- S.binomS15tmp[which(S.binomS15tmp[, 2] < 0.05), ]
785 S.binomS15SubHi  <- S.binomS15tmpSub[which(S.binomS15tmpSub[, 1] > 0.0002), ]
786 S.binomS15SubLo  <- S.binomS15tmpSub[which(S.binomS15tmpSub[, 1] < 0.0002), ]
787 S.binomS15Sub <- data.frame(cbind(-length(S.binomS15SubLo[, 1])/100, length(S.binomS15SubHi[, 1])/100))
788
789 Sdfbinom    <- data.frame(Significant = rbind(S.binomS1Sub, S.binomS2Sub, S.binomS3Sub, S.binomS4Sub, S.binomS5Sub, S.
        binomS6Sub, S.binomS7Sub, S.binomS8Sub,
790                                           S.binomS9Sub, S.binomS10Sub, S.binomS11Sub, S.binomS12Sub, S.binomS13Sub, S.
                                                  binomS14Sub, S.binomS15Sub),
791                           Label = c("S1: D", "S2: M", "S3: I", "S4: DM", "S5: MD", "S6: DI", "S7: ID", "S8: MI",
792                                     "S9: IM","S10: DMI","S11: DIM","S12: MDI","S13: MID","S14: IDM","S15: IMD"),
793                           Mod = rep("Sub-region - Simulated", 15))
794
795 Simdf       <- data.frame(rbind(Sdfbinom, Sdf))
796 Simdf$Label <- factor(Simdf$Label, as.character(Simdf$Label))
797
798 Sim <- Simdf
799
800 #write.table(Simdf, "Simulated data - significant results.csv", sep = ",", col.names = TRUE)
801 #Sim <- read.table("Simulated data - significant results.csv", sep = ",", header = TRUE)
802 Sim$Label <- factor(Sim$Label, as.character(Sim$Label))
803
804 Int <- rbind(Sim)
805 DenU <- -Int[16, 1]
```

```
DenL <- -Int[16, 2]

Int[16, 1] <- DenL
Int[16, 2] <- DenU

DenMU <- -Int[18, 1]
DenML <- -Int[18, 2]

Int[18, 1] <- DenML
Int[18, 2] <- DenMU

Int$Label  <- c("S1: D", "S2: M", "S3: I", "S4: DM", "S5: MD", "S6: DI", "S7: ID", "S8: MI",
                "S9: IM", "S10: DMI", "S11: DIM", "S12: MDI", "S13: MID", "S14: IDM", "S15: IMD",
                "Low Population Density", "High Inward-Migration", "Low Population Density (adj. Inward-Migration)",
                "High Inward-Migration (adj. Population Density)")

Int$Label <- factor(Int$Label, unique(as.character(Int$Label)))

##########################################
## Figure 2 of Paper - observed data only ##
##########################################

x11()
ggplot(Int)+
  geom_bar(aes(Label, Significant.X1, fill = Mod, alpha = Mod, order = Mod), position = "dodge", stat = "identity") +
  geom_bar(aes(Label, Significant.X2, fill = Mod, alpha = Mod, order = Mod), position = "dodge", stat = "identity") +
  geom_hline(yintercept = 0, size = 0.5) + scale_fill_manual(values = c("#08306b", "#a50f15", "#fc9272", "#6baed6")) +
  theme_bw() + scale_alpha_manual(values = c(0.8, 0.8, 0.8, 0.8)) +
  theme(axis.text = element_text(colour = "black", size = 12, family = "Times New Roman"),
        axis.title = element_text(colour = "black", size = 20, face = "bold", family = "Times New Roman"),
```

```
836       plot.title = element_text(size = 16, face = "bold", family = "Times New Roman"), legend.position = c(0.2, 0.95),
837       legend.text = element_text(size = 16, family = "Times New Roman"), axis.text.x = element_text(vjust = 0.5),
838       legend.key = element_blank()) +
839  coord_cartesian(ylim = c(-15, 100)) + scale_x_discrete(name = "", labels = function(Method)
840    str_wrap(c("S1: D", "S2: M", "S3: I", "S4: DM", "S5: MD", "S6: DI", "S7: ID", "S8: MI",
841               "S9: IM", "S10: DMI", "S11: DIM", "S12: MDI", "S13: MID", "S14: IDM", "S15: IMD",
842               "Low Population Density", "High Inward-Migration", "Low Population Density (adj. Inward-Migration)",
843               "High Inward-Migration (adj. Population Density)"), width=10))+
844  scale_y_continuous(name = "Percentage of statistically significant results",
845                     breaks = c(-15, -10, -5, 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95,
                           100)) +
846  theme(legend.title = element_blank(),legend.background = element_rect(fill = NA, colour = NA)) + guides(fill = guide_legend
       (nrow = 2, byrow = TRUE)) +
847  annotate("text", label = "Positive coefficients\nabove zero", x = 1.5, y = 10, size = 5, colour = "black") +
848  annotate("text", label = "Negative coefficients\nbelow zero", x = 1.5, y = -10, size = 5, colour = "black")
```

PopMix.R

Bibliography

[1] Stark P, Saltelli A. Cargo-cult statistics and scientific crisis. Significance. 2018;August.

[2] Wasserstein R, Lazar N. The ASA's Statement on p–Values: Context, Process, and Purpose. The American Statistician. 2016;70(2):129–133.

[3] Cutchin M. The need for the "new health geography" in epidemiologic studies of environment and health. Health Place. 2007;13(3):725–742.

[4] Glass G. Update: spatial aspects of epidemiology: the interface with medical geography. Epidemiologic Reviews. 2000;22(1):725–742.

[5] Coggon D, Rose G, Barker D. Epidemiology for the uninitiated. London: BMJ Books; 2003.

[6] Rogers A, Castree N, Kitchin R. health geography. In: A Dictionary of Human Geography. Oxford University Press; 2013. .

[7] Vertosick E, Assel M, Vickers A. A systematic review of instrumental variable analyses using geographic region as an instrument. Cancer Epidemiology. 2017;51:49–55.

[8] Joffe M, Gambhir M, Chadeau– Hyam M, Vineis P. Causal diagrams in systems epidemiology. Emerging Themes in Epidemiology. 2012;9(1).

[9] Petersen M, van der Laan M. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014;25(3):418–426.

[10] Greenland S, Pearl J, Robins J. Causal Diagrams for Epidemiologic Research. Epidemiology. 1999;10(1):37–48.

[11] Tennant P, Textor J, Gilthorpe M, Ellison G. OP87 Dagitty and directed acyclic graphs in observational research: a critical review. Journal of Epidemiology and Community Health. 2017;71(Suppl 1):A43–A43.

[12] Pearl J, Glymour M, Jewell N. Causal inference in statistics: a primer. Chichester: John Wiley and Sons; 2016.

[13] Hernán M. The C-word: scientific euphemisms do not improve causal inference from observational data. American Journal of Public Health. 2018;108(5):616–619.

[14] Glymour M, Spiegelman D. Evaluating Public Health Interventions: 5. Causal inference in public health research - do sex, race, and biological factors cause health outcomes? American Journal of Public Health. 2017;107(1):81–85.

[15] Holland P. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945–960.

[16] Robins J, Hernán M. Estimation of the causal effects of time–varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, editors. Longitudinal Data Analysis. Boca Raton: Chapman and Hall/CRC; 2009. p. 553–599.

[17] Tennant P, Arnold K, Berrie L, Ellison G, Gilthorpe M. Advanced modelling strategies: challenges and pitfalls in robust causal inference with observational data. In: Advanced Modelling Strategies: Challenges and pitfalls in robust causal inference with observational data. Leeds Institute for Data Analytics; 2017. .

[18] Hernán M, Robins J. Causal Inference. Boca Raton: Chapman and Hall/CRC; 2019; forthcoming.

[19] Herán M, Taubman S. Does obesity shorten life? The importance of well–defined interventions to answer causal questions. International Journal of Obesity. 2008;32(S3):S8–S14.

[20] Tchetgen Tchetgen E, VanderWeele T. On causal inference in the presence of interference. Statistical Methods in Medical Research. 2012;21(1):55–75.

[21] Hudgens M, Halloran M. Toward causal inference with interference. Journal of the American statistical Association. 2008;103(482):832–842.

[22] CM Z, Papadogeorgou G. Bipartite Causal Inference with Interference. arXiv e-prints. 2018 Jul;p. arXiv:1807.08660.

[23] Robins J. Data, design and background knowledge in etiologic inference. Epidemiology. 2001;11(3):313–320.

[24] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. 2nd Edition. Cambridge, Massachussets: MIT Press; 2000.

[25] Glymour M. Using causal diagrams to understand common problems in social epidemiology. In: Oakes M, Kaufman J, editors. Methods in Social Epidemiology. San Francisco, CA: Jossey-Bass; 2006. p. 387–422.

[26] Pearl J. Causal inference from direct experiments. Artificial Intelligence in Medicine. 1995;7(6):567–582.

[27] VanderWeele T. Principles of confounder selection. European Journal of Epidemiology. 2019;p. 1–9. Available from: https://doi.org/10.1007/s10654-019-00494-6.

[28] Hernán M, Hernández– Díaz S, Werler M, Mitchell A. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology. 2002;155(2):176–184.

[29] Fleischer N, Diez Roux A. Using directed acyclic graphs to guide analyses of neighbourhood health effects: An introduction. Journal of Epidemiology and Community Health. 2008;62(9):842–846.

[30] Textor J, Zander Bvd, Gilthorpe M, Liśkiewicz M, Ellison G. Robust causal inference using directed acyclic graphs: The R package 'dagitty'. International Journal of Epidemiology. 2016;45(6):1887–1894.

[31] Rohrer J. Thinking clearly about correlations and causation: graphical causal models for observational data. Advances in Methods and Practices in Psychological Science. 2018;1(1):27–42.

[32] Tu YK, Gunnell D, Gilthorpe M. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon–the reversal paradox. Emerging Themes in Epidemiology. 2008;5(2).

[33] Hernán M, Hernández– Díaz S, Robins J. A structural approach to selection bias. Epidemiology. 2004;15(5):615–625.

[34] Simpson E. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B. 1951;13:238–241.

[35] Pearl J. Comment: Understanding Simpson's Paradox. The American Statistician. 2014;68(1):8–13.

[36] Lindley D, Novick M. The Role of Exchangeability in Inference. The Annals of Statistics. 1981;9:45–58.

[37] Rubin D. Causal inference using potential outcomes. Journal of the American Statistical Association. 2005;100(469):322–331.

[38] Morgan S, Winship C. Counterfactuals and Causal Inference: Methods and Principles for Social Research. 2nd Edition. Cambridge, UK: Cambridge University Press; 2015.

[39] Buchner H, Rehfuess E. Cooking and season as risk factors for acute lower respiratory infections in African children: A cross-sectional multi-country analysis. PLoS One. 2015;10(6).

[40] Filippidis F, Laverty A, Hone T, Been J, Millett C. Association of Cigarette Price Differentials With Infant Mortality in 23 European Union Countries. JAMA Pediatrics. 2017;171(11):1100–1106.

[41] Oswald W, Stewart A, Kramer M, Endeshaw T, Zerihun M, Melak B, et al. Association of community sanitation usage with soil-transmitted helminth infections among school-aged children in Amhara Region, Ethiopia. Parasites and Vectors. 2017;10(1):91.

[42] Yau R, Casteel C, Nocera M, Bishop S, Peek– Asa C. Does employee resistance during a robbery increase the risk of customer injury? Journal of Occupational and Environmental Medicine. 2015;57(4):417–420.

[43] Loney T, Nagelkerke N. The individualistic fallacy, ecological studies and instrumental variables: a causal interpretation. Emerging Themes in Epidemiology. 2014;11(18):1–6.

[44] Smith H. Some thoughts on causation as it relates to demography and population studies. Population and Development Review. 2003;29(3):459–469.

[45] Schisterman E, Perkins N, Mumford S, Ahrens K, Mitchell E. Collinearity and causal diagrams –a lesson on the importance of model specification. Epidemiology. 2017;28(1):47–53.

[46] Shmueli G. To explain or to predict? Statistical Science. 2010;25:289–310.

[47] Arnold K, Davies V, de Kamps M, Tennant P, Mbotwa J, Gilthorpe M. Generalised linear models for prognosis and intervention: Theory, practice, and implications for machine learning. arXiv e-prints. 2019;p. arXiv:1906.01461.

[48] Hernán M, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. CHANCE. 2019;32(1):42–49.

[49] Westreich D, Greenland S. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. American Journal of Epidemiology. 2013;177:292–298.

[50] Green M, Popham F. Interpreting mutual adjustment for multiple indicators of socioeconomic position without committing mutual adjustment fallacies. BMC Public Health. 2019;19(10).

[51] Morris T, White I, Crowther M. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38:2074–2102.

[52] Gilbert N, Troitzsch K. Simulation for the Social Scientist. Maidenhead, Berkshire: Open University Press; 2005.

[53] Arnold K, Harrison W, Heppenstall A, Gilthorpe M. DAG-informed regression modelling, agent-based modelling and microsimulation modelling: a critical comparison of methods for causal inference. International Journal of Epidemiology. 2019;1(48):243–253.

[54] Hallgren K. Conducting Simulation Studies in the R Programming Environment. Tutorials in quantitative methods for psychology. 2013;9(2):43–60.

[55] Steyer R, Mayer A, Fiege C. Causal inference on total, direct, and indirect effects. In: Michalos A, editor. Encyclopedia of Quality of Life Research. Dordrecht: Springer Science and Business Media; 2013. p. 123–134.

[56] Elwert F, Winship C. Endogenous selection bias: the problem of conditioning on a collider variable. Annual Review of Sociology. 2014;40(1):31–53.

[57] Susser E, Bresnahan M. Origins of Epidemiology. Annals of the New York Academy of the Sciences. 2001;954(1):6–18.

[58] Charlton J. ONS data: other health sources. In: Leadbeter D, editor. Harnessing Official Statistics. Oxford: Radcliffe Medical Press; 2000. p. 35–50.

[59] Boslaugh S. An introduction to secondary data analysis. New York: Cambridge University Press; 2007.

[60] Fraser L, Norman P. The use of routine data in health research: An example from palliative care. SAGE Research Methods Cases. 2017;.

[61] Norman P, Marshall A, Lomax N. Data analytics: on the cusp of using new sources? Radical Statistics. 2017;115:19–30.

[62] De Mauro A, Greco M, Grimaldi M. A formal definition of big data based on its essential features. Library Review. 2016;65(3):122–135.

[63] Barenboim E, Pearl J. Causal inference and the data–fusion problem. Proceedings of the National Academy of the Sciences of the United States of America. 2016;113(27):7345–7352.

[64] Brunsdon C, Fotheringham A, Charlton M. Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis. 1996;28(4):281–298.

[65] Templ M, Meindl B, Kowarik A, Dupriez O. Simulation of Synthetic Complex Data: The R Package simPop. Journal of Statistical Software, Articles. 2017;79(10):1–38.

[66] Burton A, Altman D, Royston P, Holder R. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;9(4):4279–4292.

[67] Maxwell S, Cole D. Tips for writing (and reading) methodological articles. Psychological Bulletin. 1995;118(2):193–198.

[68] Sofrygin O, van der Laan M. simcausal R package: Conducting transparent and reproducible simulation studies of causal effect estimation with complex longitudinal data. Journal of Statistical Software. 2017;81(2):1–47.

[69] Wright S. On the nature of size factors. Genetics. 1918;3:367–374.

[70] Wright S. The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. Proceedings of the National Academy of Sciences of the United States of America. 1920;6(6):320–332.

[71] Wright S. Correlation and causation. Journal of Agricultural Research. 1921;20:557–585.

[72] Chen B, Pearl J, Kline R. Graphical tools for linear path models. Psychometrika. 2018;.

[73] Wright S. The method of path coefficients. Annals of Mathematical Statistics. 1934;5:161–215.

[74] Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. Epidemiology. 2011;22(5):745–751.

[75] Ruscio J, Kaczetow W. Simulating Multivariate Nonnormal Data Using an Iterative Algorithm. Multivariate Behavioural Research. 2008;43(3):355–381.

[76] Pearl J. Causality: models, reasoning and inference Vol. 1. Cambridge, UK: Cambridge University Press; 2000.

[77] Shacter R. A graph–based inference method for conditional independence. In: D'Ambrosio B, Smets P, Bonissone P, editors. Uncertainty Proceedings 1991. San Francisco (CA): Morgan Kaufmann; 1991. p. 353–360.

[78] Shacter R. An ordered examination of influence diagrams. Networks. 1990;20(0):535–563.

[79] Shacter R. Probabilistic inference and influence diagrams. Operations Research. 1988;36(4):589–604.

[80] Geiger D, Verma T, Pearl J. Identifying independence in Bayesian Networks. Networks. 1990;20(0):507–534.

[81] Berrie L, Arnold K, Textor J, Gilthorpe M, Tennant P. Depicting deterministic relationships in directed acyclic graphs (DAGs): An aid for analysing and interpreting compositional data. In: Book of Abstracts of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa, 3–8 June, 2019. Universitat Politècnica de Catalunya–BarcelonaTECH; 2019. .

[82] Hernán M. Does water kill? A call for less casual causal inferences. Annals of Epidemiology. 2016;26:674–680.

[83] Levy Y. Cancer clusters and the Poisson distributions; 2019. [Online; accessed 23-July-2019]. https://www.r-bloggers.com/cancer-clusters-and-the-poisson-distributions/.

[84] Flowerdew R, Manley D, Sabel C. Neighbourhood effects on health: Does it matter where you draw the boundaries? Social Science and Medicine. 2008;66(6):1241–1255.

[85] Openshaw S. The modifiable areal unit problem. Norwich: Geo Books: Farrar, Strous and Giroux; 1983.

[86] Cheng T, Adepeju M. Modifiable temporal unit problem (MTUP) and its effect on space–time cluster detection. PLOS ONE. 2014;9(6):1–10.

[87] Wainer H. The most dangerous equation. American Scientist. 2007;65:249–256.

[88] Tu YK, Gilthorpe M. The most dangerous hospital or the most dangerous equation? BMC health services research. 2007;7(1):185–189.

[89] Archie J. Mathematic coupling of data: a common source of error. Annals of Surgery. 1981;193(3):296–303.

[90] Pearson K. Mathematical Contributions to the Theory of Evolution.–On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. Proceedings of the Royal Society of London. 1896;60:489–498. Available from: http://www.jstor.org/stable/115879.

[91] Tu YK, Law G, Ellison G, Gilthorpe M. Ratio index variables or ANCOVA? Fisher's cats revisited. Pharmaceutical Statistics. 2010;9(1):77–83.

[92] Dunlap W, Dietz J, Cortina J. The spurious correlation of ratios that have common variables: a Monte Carlo examination of Pearson's formula. The Journal of General Psychology. 1997;124(2):182–193.

[93] Andersen B. Methodological errors in medical research. London: Blackwell; 1990.

[94] Neyman J. Lectures and Conferences on Mathematical Statistics and Probability. 2nd Edition. Washington: US Department of Agriculture; 1952.

[95] McCullagh P, JA N. Chapter 6: Log-linear models. London and New York: Chapman and Hall/CRC Monographs on Statistics & Applied Probability; 1989.

[96] Evans I, Jones K. Ratios and closed number systems. In: Wrigley N, Bennett R, editors. Quantitative Geography: A British View. London, Boston and Henley: Routledge and Kegan Paul; 1981. p. 123–134.

[97] Fisher R. The Analysis of Covariance Method for the Relation between a Part and the Whole. Biometrika. 1947;3(2):65–68.

[98] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018. Available from: https://www.R-project.org/.

[99] Kronmal R. Spurious correlation and the fallacy of the ratio standard revisited. Journal of the Royal Statistical Society Series A. 1993;156(3):379–392.

[100] Snow G. TeachingDemos: Demonstrations for Teaching and Learning; 2016. R package version 2.10. https://CRAN.R-project.org/package=TeachingDemos.

[101] Firebaugh G, Gibbs J. User's guide to ratio variables. American Sociological Review. 1985;50(5):713–722.

[102] Lloyd C, Pawlowsky-Glahn V, Egozcue J. Compositional Data Analysis in Population Studies. Annals of the Association of American Geographers. 2012;102(6):1251–1266.

[103] Kanaroglou P. On spurious correlation in geographical problems. Canadian Geographer / Le Géographe canadien. 1996;40:194–202.

[104] Firebaugh G, Gibbs J. Using ratio variables to control for population size. Sociological Methods and Research. 1986;15(1–2):101–117.

[105] Bambra C, Norman P. What is the association between sickness absence, mortality and morbidity? Health and Place. 2006;12(4):728–733.

[106] Townsend P, Phillimore P, Beattie A. Health and Deprivation: Inequality and the North. London, UK: Croom Helm; 1988.

[107] Bentham G, Eimermann J, Haynes R, Lovett R, Brainard J. Limiting long-term illness and its associations with mortality and indicators of social deprivation. Journal of Epidemiology and Community Health. 1995;49:S57–S64.

[108] Shouls S, Congdon P, Curtis S. Modelling inequality in reported long term illness in the UK: combining individual and area characteristics. Journal of Epidemiology and Community Health. 1996;50:366–376.

[109] Haynes R, Bentham G, Lovett A, Eimermann J. Effect of labour market conditions on reporting of limiting long term illness and permanent sickness in England and Wales. Journal of Epidemiology and Community Health. 1997;51(3):283–288.

[110] Folwell K. Single measures of deprivation. Journal of Epidemiology and Community Health. 1995;49:S51–S56.

[111] Carstairs V, Morris R. Deprivation and health in Scotland. Aberdeen, UK: Aberdeen University Press; 1991.

[112] Jarman B. Identification of underprivileged areas. British Medical Journal. 1983;286:1705–1708.

[113] Saul C, Payne N. How does the prevalence of specific morbidities compare with measures of socio-economic status at small area level? Journal of Public Health Medicine. 1999;21(3):340–347.

[114] Senior M. Area Variations in Self-perceived Limiting Long Term Illness in Britain, 1991: Is the Welsh Experience Exceptional? Regional Studies. 1998;32(3):265–280.

[115] Möller H, Haigh F, Harwood C, Kinsella T, Pope D. Rising unemployment and increasing spatial inequalities in England: further extension of the North-South divide. Journal of Public Health. 2013;35(2):313–321.

[116] O'Reilly D, Stevenson M. The two communities in Northern Ireland: deprivation and ill health. Journal of Public Health Medicine. 1998;20(2):161–168.

[117] Huff N, Macleod C, Ebdon D, Phillips D, Davies L, Nicholson A. Inequalities in mortality and illness in Trent NHS region. Journal of Public Health Medicine. 1999;21(1):81–87.

[118] Haynes R, Gale S. Mortality, long–term illness and deprivation in rural and metropolitan wards of England and Wales. Health and Place. 1999;5:301–312.

[119] Boyle P, Gatrell A, Duke– Williams O. The effect on morbidity of variability in deprivation and population stability in England and Wales: an investigation at small–area level. Social Science and Medicine. 1999;49:791–799.

[120] Barnett S, Roderick P, Martin D, Diamond I. A multilevel analysis of the effects of rurality and social deprivation on premature limiting long term illness. Journal of Epidemiology and Community Health. 2001;55:44–51.

[121] Lancaster G, Green M. Deprivation, ill–health and ecological fallacy. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2002;164(2):263–278.

[122] Levin K. Urban–rural differences in self–reported limiting long–term illness in Scotland. Journal of Public Health. 2003;25(4):295–302.

[123] Jordan H, Roderick P, Martin D. The Index of Multiple Deprivation 2000 and accessibility effects on health. Journal of Epidemiology and Community Health. 2004;58(3):250–257.

[124] Cockings S, Martin D. Zone design for environment and health studies using pre–aggregated data. Social Science and Medicine. 2005;60(12):2729–2742.

[125] Adams J, Holland L, White M. Changes in socioeconomic inequalities in census measures in England and Wales, 1991–2001. Journal of Epidemiology and Community Health. 2006;60(3):218–220.

[126] Cairns J, Curtis S, Bambra C. Defying deprivation: a cross–sectional analysis of area level health resilience in England. Health and Place. 2012;18(4):928–933.

[127] Zhang X, Cook P, Lisboa P, Jarman I, Bellis M. The effects of deprivation and relative deprivation on self–reported morbidity in England: an area–level ecological study. International Journal of Health Geographics. 2012;18(4):928–933.

[128] Aitchison J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B. 1982;44(2):139–177.

[129] Lindén A, Mäntyniemi S. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology. 2011;92(7):1414–1421.

[130] Berrie L, Ellison G, Norman P, Baxter P, Feltbower R, Tennant P, et al. The association between childhood leukemia and population mixing: An artifact of focusing on clusters? Epidemiology. 2019;30:75–82.

[131] Kinlen L, Peto J. Re: The association between childhood leukemia and population mixing: An artifact of focusing on clusters? Epidemiology. 2019;30.

[132] Berrie L, Ellison G, Norman P, Baxter P, Feltbower R, Tennant P, et al. Authors' Respond: Re: The association between childhood leukemia and population mixing: An artifact of focusing on clusters? Epidemiology. 2019;30.

[133] Omidakhsh N, Hansen J, Ritz B, Olsen J, Heck J. High parental occupational social contact and risk of childhood hematopoietic, brain and bone cancers. Cancer Epidemiology. 2019;62(1):1–8.

[134] Kahneman D. Thinking, fast and slow. New York: Farrar, Strous and Giroux; 2011.

[135] Bunch K, Vincent T, Black R, Pearce M, McNally R, McKinney P, et al. Updated investigations of cancer excesses in individuals born or resident in the vicinity of Sellafield and Dounreay. British Journal of Cancer. 2014;111(9):1814–1823.

[136] McNally R, James P, Blakey K, Basta N, Norman P, Pearce M. Can changes in population mixing and socio–economic deprivation in Cumbria, England explain changes in cancer incidence around Sellafield? Spatial and Spatio-temporal Epidemiology. 2017;21:23–36.

[137] Kinlen L. Evidence for an infective cause of childhood leukaemia: comparison of a Scottish New Town with nuclear reprocessing sites in Britain. Lancet. 1988;332(8624):1323–1327.

[138] Lupatsch J, Kuehni C, Niggli F, Ammann R, Egger M, Spycher B. Population mixing and the risk of childhood leukaemia in Switzerland: a census–based cohort study. European Journal of Epidemiology. 2015;30(12):1287–1298.

[139] Imam A, Fairley L, Parslow R, Feltbower R. Population mixing and incidence of cancers in adolescents and young adults between 1990 and 2013 in Yorkshire, UK. Cancer Causes and Control. 2016;27(10):1287–1292.

[140] COMARE. Seventeenth Report. Further consideration of the incidence of cancers around the nuclear installations at Sellafield and Dounreay. Committee on Medical Aspects of Radiation in the Environment (COMARE); 2016.

[141] Law G, Feltbower R, Taylor J, Parslow R, Gilthorpe M, Boyle P, et al. What do epidemiologists mean by 'population mixing'? Pediatric Blood Cancer. 2008;51:155–160.

[142] Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. Nature Reviews, Cancer. 2006;6(3):193–203.

[143] Wiemels J. Persepectives on the causes of childhood leukaemia. Chemico-Biological Interactions. 2012;196(3):59–67.

[144] Rudant J, Lightfoot T, Urayama K, Petridou E, Dockerty J, Magnani C, et al. Childhood acute lymphoblastic leukemia and indicators of early immune stimulation: a Childhood Leukemia International Consortium study. American Journal of Epidemiology. 2015;181(8):549–562.

[145] Parslow R, Law G, Feltbower R, Kinsey S, McKinney P. Population mixing, childhood leukaemia, CNS tumours and other childhood cancers in Yorkshire. European Journal of Cancer. 2002;38(15):2033–2050.

[146] Clark B, Ferketich A, Fisher J, Ruymann F, Harris R, Wilkins J. Evidence of population mixing based on the geographical distribution of childhood leukemia in Ohio. Pediatric Blood Cancer. 2007;49:797–802.

[147] Kinlen L, Clarke K, Hudson C. Evidence from population mixing in British New Towns 1946–85 of an infective basis for childhood leukaemia. Lancet. 1990;336:577–582.

[148] Kinlen L, Hudson C, Stiller C. Contacts between adults as evidence for an infective origin of childhood leukaemia: an explanation for the excess near nuclear establishments in West Berkshire? British Journal of Cancer. 1991;64(3):549–554.

[149] Langford I. Childhood leukaemia mortality and population change in England and Wales 1969–73. Social Science and Medicine. 1991;33(4):435–440.

[150] Kinlen L, Dickson M, Stiller C. Childhood leukaemia and non-Hodgkin's lymphoma near larage rural construction sites, with a comparison with Sellafield nuclear site. British Medical Journal. 1995;310:763–768.

[151] Labar B, Rudan I, Ivankovic D, Biloglav Z, Mrsic M, Strnad M, et al. Haematological malignancies in childhood in Croatia: Investigating the theories of depleted uranium, chemical plant damage and 'population mixing'. European Journal of Epidemiology. 2004;19(1):55–60.

[152] Kinlen L. Childhood leukaemia and ordnance factories in west Cumbria during the Second World War. British Journal of Cancer. 2006;95(1):102–106.

[153] Kinlen L. An examination, with a meta–analysis, of studies of childhood leukaemia in relation to population mixing. British Journal of Cancer. 2012;107(7):1163–1168.

[154] Taylor J, Law G, Boyle P, Feng Z, Gilthorpe M, Parslow R, et al. Does population mixing measure infectious exposure in children at the community level? European Journal of Epidemiology. 2008;23:593–600.

[155] Cromley R, Cromley E. Choropleth map legend design for visualizing community health disparities. Internatinal Journal of Health Geographics. 2009;8(52):1–11.

[156] Stiller C, Ardanaz E, Pannelli F, EA M, Can A. Geographical patterns of childhood cancer incidence in Europe, 1988–1997. Report from the Automated Childhood Cancer Information System project; 2006.

[157] Laplanche A, de Vathaire F. Leukaemia mortality in French communes (administrative units) with a large and rapid population increase. British Journal of Cancer. 1994;69(1):110–113.

[158] Conover W. Practical Nonparametric Statistics. New York: John Wiley and Sons; 1971.

[159] Stiller C, Boyle P. Effect of population mixing and socioeconomic status in England and Wales, 1979–85, on lymphoblastic leukaemia in children. British Medical Journal. 1996;313:1297–1300.

[160] Dickinson H, Parker L. Quantifying the effect of population mixing on childhood leukaemia risk: the Seascale cluster. British Journal of Cancer. 1999;81(1):144–151.

[161] Koushik A, King W, McLaughlin J. An ecologic study of childhood leukemia and population mixing in Ontario, Canada. Cancer Causes and Control. 2001;12(6):483–490.

[162] Boutou O, Guizard AV, Slama R, Pottier D, Spira A. Population mixing and leukaemia in young people around the La Hague nuclear waste reprocessing plant. British Journal of Cancer. 2002;87(7):740–745.

[163] Dickinson H, Hammal D, Bithell J, Parker L. Population mixing and childhood leukaemia and non–Hodgkin's lymphoma in census wards in England and Wales, 1966–87. British Journal of Cancer. 2002;86:1411–1413.

[164] Nyari T, Kajtar P, Bartyik K, Thurzo L, Parker L. Childhood Acute Lymphoblastic Leukaemia in Relation to Population Mixing Around the Time of Birth in South Hungary. Pediatric Blood Cancer. 2006;47:944–948.

[165] Adelman A, Groves F, O'Rourke K, Sinha D, Hulsey T, Lawson A, et al. Residential mobility and risk of childhood acute lymphoblastic leukaemia: an ecological study. British Journal of Cancer. 2007;97(1):140–144.

[166] Stiller C, Kroll M, Boyle P, Feng Z. Population mixing, socioeconomic status and incidence of childhood acute lymphoblastic leukaemia in England and Wales: analysis by census ward. British Journal of Cancer. 2008;98(5):1006–1011.

[167] Norman P, Berrie L, Exeter D. Introductory Guide: Calculating a deprivation index using census data. Australian Population Studies. 2019;3(1):30–39.

[168] Rehkopf D, Glymour M, Osypuk T. The consistency assumption for causal inference in social epidemiology: when is a rose not a rose? Current Epidemiology Report. 2016;3(1):63–71.

[169] Fink D, Keyes K, Cerdá M. Social determinants of population health: a systems sciences approach. Current Epidemiology Report. 2016;3(1):98–105.

[170] Ní Bhrolcháin M, Dyson T. On causation in demography: issues and illustrations. Population and Development Review. 2007;33(1):1–36.

[171] Westreich D, Cole S. Invited Commentary: Positivity in practice. American Journal of Epidemiology. 2010;171(6):674–677.

[172] Messer L, Oakes J, Mason S. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. American Journal of Epidemiology. 2010;171(6):664–673.

[173] VanderWeele T. Ignorability and stability assumptions in neighborhood effects research. Statistics in Medicine. 2008;27(11):1934–1943.

[174] Papadogeorgou G, Mealli F, Zigler C. Causal inference for interfering units with cluster and population level treatment allocation programs. arXiv e-prints. 2017;p. arXiv:1711.01280.

[175] Glass T, Goodman S, Hernán M, Samet J. Causal inference in public health. Annual Review of Public Health. 2013;34:61–75.