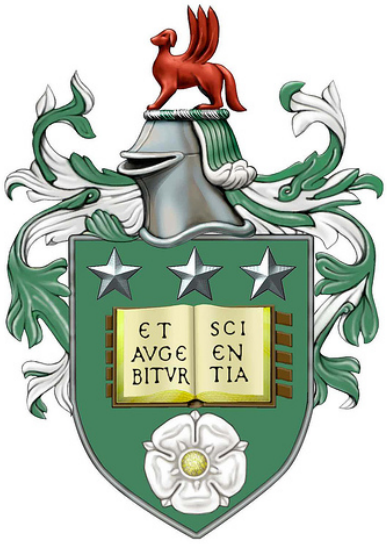
Classification of Biomedical Data Using Spatial Features



Wafa Ali J Almohri

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

Department of Statistics

September 2019

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

To Almighty Allah

Acknowledgements

In the name of Allah, the most gracious and merciful. Alhamdulillah I thank you Allah for supporting me to complete this work, helping me to the end with health and knowledge, giving me the best when I ask. The word thank you is not enough, what I can only say is please Allah help me to thank you always and to worship you in the best way.

This thesis is completed due to the assistance and guidance of many people. First and foremost, to my husband Hamdi Al-Mezgagi who is my best friend and my soul mate. The words can not be enough to say thank you very much for supporting, helping and encouraging. I would like to express my deepest gratitude to my mum and dad who are always supporting me with their prayers and doing their best to help me.

I also gratefully acknowledge the funding support provided by the Taibah University, Madinah. Also the Saudi Arabia cultural bureau in London for supporting Saudi students and allowing me to pursue my PhD at the University of Leeds.

I would like to express my sincere gratitude to my supervisors Prof. Charles C. Taylor, Dr. Robert G Aykroyd and Dr. Darren Treanor for keeping supporting and encouraging me through my Ph.D study, for their patience, motivation, and wide knowledge. In fact, words cannot express my heartfelt gratitude, appreciation and thanks for all their support, guidance and time they have provided during my PhD life. Another big thanks to my children Tareq, Mazen, Hazem and Mohammed who are always supporting and helping me at home and for their prayers. A more sweet thanks to my little daughter Sarah who makes me forget my depression and makes me smile.

I express my sincere gratitude to Dr. Nick West, Dr. Heike Grabsch, Dr. Takashi Oshima, Dr. Takaki Yoshikawa and Dr. Yohei Miyagi for their data support in this project. I would like to give my thanks to all staff in the School of Mathematics for their helping and supporting. Special thanks to Jeanne Shuttleworth who supported me all the time and helped me to create the PhD women in Mathematics group.

Abstract

Histopathologists typically collect biopsies which leads to image data. They examine the images to obtain various diagnostic summaries, e.g. proportion of tumor. They do this by overlaying a regular grid of points which are then classified. This classification allows them to estimate the proportion of tumor and other statistics. In this thesis, we focus on investigating heterogeneity. We do this by considering measures of clustering in the classified points spatially. We consider the use of cluster statistics in the diagnosis of patient cancer (stomach and rectum cancers). We further consider tests of anisotropy/direction of heterogeneity/clustering. Binary Markov random field parameter estimation is also investigated as an alternative approach for detecting heterogeneity of the image both overall and in a specific direction. Furthermore, we consider spatial prediction and consistency of spot classifications for overlapping regions sampled at different resolutions.

In the first part of this thesis, we aim to identify an appropriate spatial autocorrelation statistic measure, under a normal approximation of the statistical test. We investigate the power of Moran's I statistic which has power in the large sample setting. More importantly, the I statistic is then modified to measure the heterogeneity/clustering in different directions. In particular in the cancer studies, associating the cluster direction with that of the lumen surface, which is an important pathological feature, is investigated.

Following this, a new simulation-based iterative method for estimating binary Markov random field parameters is explained. Estimated parameters give similar information to the spatial measurements, and this method leads to a statistical test which does not depend on normal approximations. Based on simulation, the accuracy of the iterative method is checked and compared favourably with an existing parameter estimation method.

We address the sampling issue by investigating the spatial consistency for pairs of images sampled from the same area but with different resolutions. Finally, we address several clinical questions. For instance, explaining the differences in survival of patients is investigated and it was found that heterogeneity is related to expected survival times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Introduction

This project is the result of co-operative work between statisticians and pathologists and is based on the digital slides of human pathology cancer tissue. Pathology is the study and understanding of disease, knowledge that is essential in evaluating human tissue samples and identifying diseases and treatments. The pathologist is a bridge between medical doctors and scientists and is an expert in illness and disease.

Data for this project are derived from digital photographs of human tissue slides using 2D microscopy at 20X magnification. From these photographs, a web-based software tool, called RandomSpot (Treanor et al., 2008; Wright et al., 2015), generates a systematic grid of spot/cell locations within a target area. This software, which is based on systematic random sampling (SRS), provides a framework upon which to quickly build an accurate estimate of the distribution of classes within a tissue sample. The first spot of the grid created by RandomSpot is placed randomly, with the subsequent spots placed systematically following a hexagonal regular grid of spots. Each spot can then be quantitatively evaluated by an expert pathologist to determine the feature present at that spot; e.g. a tumor, or a stroma cell (the cells that surround tumor cells) to generate a sample of tissue-type classifications.

In this thesis, the spatial arrangement of spots on a hexagonal grid with their classes will be referred to as a "biomedical image". These images contain a tissue-type classification at 50-300 spots from the region of interest. The image data contains positions of

the sites given by the coordinates (u_i, v_i), where $i = 1, \ldots, n$ are the indexes of n spots, together with the classification of the spots given by $x_i, i = 1, \ldots, n$. The ordering of spots in each row is from left to right, and the rows are ordered from bottom to top.

Pathological assessments of the tissue on the slides is a key part of the diagnosis of cancer. An example of analysis is the use of overall summary images which can be derived from the classification of the spots, such as the ratio of tumor to stroma cells. Stereological methods, which include spot-counting tasks, provide quantification of tumor characteristics that can then be used to compare tumor structure and composition objectively to diagnose and understand diseases like cancer. The pathological analysis of biomedical images follows a standard approach to characterise cancer. However, traditional pathological diagnoses are subjective and descriptive, making comparison of quantitative features difficult.

This project is the first numerical characterisation of stereologically derived biomedical images from pathological samples collected from multiple patients. Examining patterns or spatial features of appearance by numerical quantification is one of the primary tasks in this project. Examination of tissue appearance helps in assessing tumor heterogeneity. Pathologists use a standard subjective process to assess the heterogeneity of spatial features and to classify images. This is a very time-consuming process and its success depends heavily on the expertise and experience of the pathologist.

Various medical questions arose during the project, such as how the new spatial feature measurements of assessing heterogeneous tumors is related to patient survival. Spatial measurements are also adjusted to examine the heterogeneity of tumors in various directions. We also sought to compare different resolution levels to see if the images are spatially consistent by prediction. More questions are addressed within each chapter. Ethical approval for the study was obtained by Dr. D. Treanor from the NHS ethical approval committee to use and analyse the biomedical image data (Leeds West LREC reference 05/Q1205/220).

In this chapter, our motivation and a review of related previous work is described. Background information into biomedical images is presented in Section 1.3, and the spot classification is determined by pathologists in Section 1.4. More detail about the contribution of each chapter is given within their introductory sections. An overview of the research covered in this thesis is given in Section 1.5.

1.2 Why spatial analysis?

Spatial analysis can be applied in various fields, such as epidemiology (Graham et al., 2004), geography (Ebdon, 1985; Lee and Wong, 2001), sociology (Logan, 2012) and geostatistics (Cressie, 1993). However, in biomedical images, work on investigating the spatial heterogeneity is very limited and subjectively evaluated.

Dworak et al. (1997) initially described a standardised 5-point grading system for what is called "tumor regression" (TRG), which is based on the spatial presence or absence of macroscopic disease. The TRG is the amount of macroscopic disease after chemotherapy but recorded before removing the tumor surgically. The TRG describes the varying degree of replacing tumor with fibrosis, and it ranges from 4, when there is no viable tumor cells detected, to 0 when there is no tumor regression. TRG$=3$ is defined as more than 50% with fibrosis outgrowing the tumor mass, TRG$=2$ is with less than 50%, and TRG$=1$ is defined as a morphologically unaltered tumor mass. Figure 1.1 shows the varying stages of TRG due to radiation treatment from A to F. The chemotherapy type before surgery called, "neoadjuvant treatment", is where one or more chemotherapy medicine is involved in helping to reduce the risk of the cancer coming back after operation.

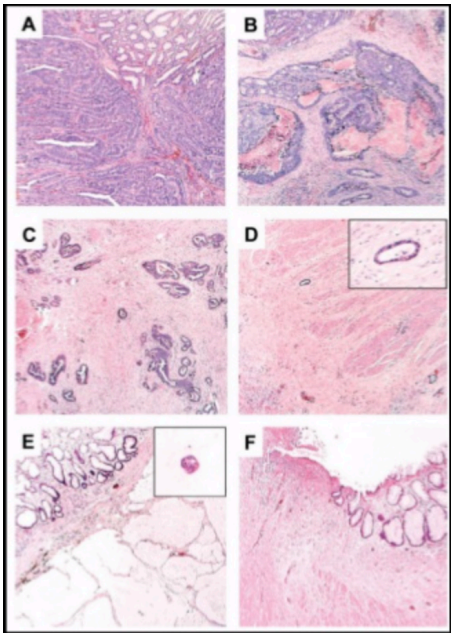


Figure 1.1: An example of TRG, (A) shows TRG equal to zero when there is no tumor regression, (B) dominant tumor with fibrosis, (C) significant fibrosis with clustering of tumor cells (D-E) fibrosis with very few tumor cells and (F) no viable tumor cells, taken from Cserni, Gabor (2011).

Wheeler et al. (2002) and Rdel et al. (2005) have also emphasised that the modified 3-point rectal cancer regression grade staging system, which has considered the combination of TRG 4 and 2 as well as TRG 3 and 0 but kept TRG 1, has led to stronger results than would have been obtained through other grouping. All these methods are, however, subjective assessments of tumor spatial behaviour. Rdel et al. (2005) suggested that the accuracy, reliability, and validity of modified staging systems still needs to be investigated further.

High spatial heterogeneity is a feature of especially gastric and rectal cancer which are the most leading causes of cancer mortality worldwide. This feature may influence the characterisation of tumor biology. Understanding cancer heterogeneity is substantial for a more accurate diagnosis, for selecting appropriate therapy regimens, and for monitoring remaining disease. Recent studies based on a histopathological subjective evaluation to understand heterogeneous tumor, for example, Gullo et al. (2017) and Aoyama et al. (2018), but they suggested that further investigations are still needed.

Pathologists have also analysed biomedical images of colorectal cancer in very simple numerical ways by comparing the overall proportion of tumor (POT) in the image (West et al., 2010a). They classified POT as either POT-high or POT-low which were defined using the mode. They found that POT-low in colorectal cancer is related to poor survival, but there was no significant correlation between POT and any of the clinical variables. Likewise, the gastric cancer dataset showed that a low proportion of tumor is related to poor survival (Aurello et al., 2017; Lee et al., 2017; Peng et al., 2018), but the sampling areas for measuring PoT were not at the luminal surface. A lumen is the inside space of a tubular structure, such as in the bowel and the interior of the gastrointestinal tract. This surface is important because it is the location where cancers first develop, and they spread into the deeper stomach from lumen. Furthermore, West et al. (2010a) hypothesised that patients with low proportion of tumor might be more likely to have more responded than POT-high and they recommend that this area warrants further investigations. Aoyama et al. (2018) showed a similar study using gastric cancer, where the POT was measured at the luminal surface, but the result was opposite. Patients with low POT survived significantly longer than whose with high POT. The POT measurement is commonly used in analysing biomedical images, for instance in the following papers: Huijbers et al. (2013), Mouliere et al. (2013) and Hale et al. (2016).

Pathologists also use an objective quantitative tumor cell density (TCD) analysis which has been found to be a useful prognostic indicator of the response to preoperative therapy (West et al., 2010b). Mesker et al. (2007) and West et al. (2010a) also observed significant heterogeneity subjectively in the POT within individual tumors, but reported that it is difficult to measure objectively.

Biomedical images have also been analysed in Almohri (2012) as compositional data, where the proportion of spots of many different types was given; not just tumor and stroma. The aim was to classify the images into two or more groups, but there was no definite answer regarding the number of groups.

Nevertheless, pathologists still need better ways to understand biomedical images objectively and new quantitative summaries which can be used in further analysis. Pathologists see spatial pattern in tumors, which they believe can be quantified statistically, in order to aid patient diagnosis. However, they currently have no objective measurement tools and an obvious idea is to use spatial statistical techniques. Hence, spatial analysis could be used to describe images instead of only, for example, considering the overall POT in relation to other factors (e.g. to patient survival).

1.3 Background to biomedical images

Two datasets of biomedical images were provided for two different cancer studies: 1) gastric cancer images (one image per patient) and 2) rectal cancer images (multiple images per patient).

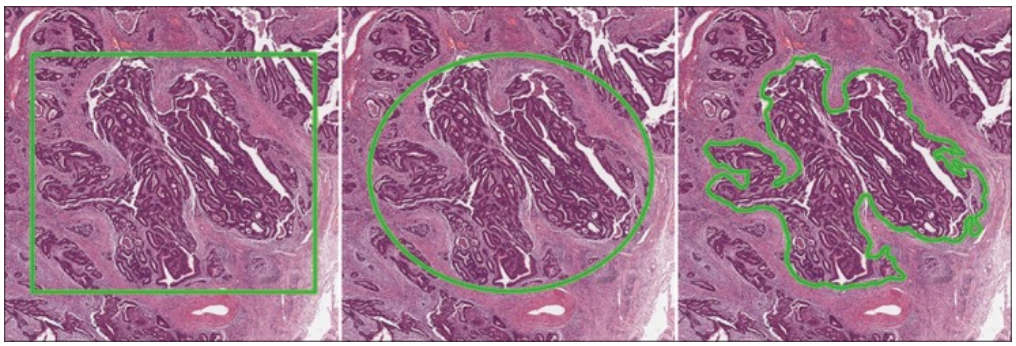


Figure 1.2: Examples of delineation types, drawn on the same digital tissue slide (viewing magnification 20X): rectangular, elliptical and polygonal, taken from Wright et al. (2015).

We will start by explaining the process of how the biomedical images are generated

and recorded, and then represented using statistical software. From digital tissue slides, the area of interest is highlighted manually by a pathologist in black ink. One of three types of image delineation can be used: square, elliptical/circle or polygonal, as shown in Figure 1.2. The delineated area is then scanned with capture resolutions of 0.5 microns/pixel (20X magnification). The target number of spots is then determined and they are spaced equally using a hexagonal mesh by the RandomSpot algorithm (Treanor et al., 2008; Wright et al., 2015). The ratio of distances between spots on vertical lines divided by that of horizontal lines was approximately 0.79 in all images, and hence the distance of vertical lines is actually shorter than the distance between horizontal lines. Therefore the edges of the hexagons are not the same length. Figure 1.3 shows how a grid of spots is added to a virtual slide of a whole tumor. Virtual slide viewing software was then used to view the spots at different resolutions. The tissue-type classes were then recorded manually.

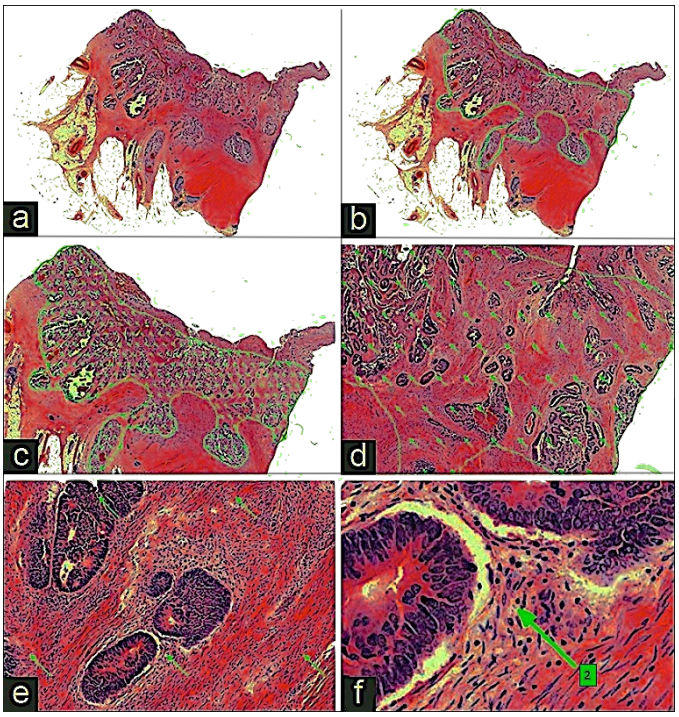


Figure 1.3: Example of preparation of a virtual slide where (a) shows delineating of a tumor boundary by hand. (b)-(e) show the spots added by RandomSpot, at different zoom levels (from 2.5X to 40X), and (f) shows the manual classification of a spot, taken from Wright et al. (2015).

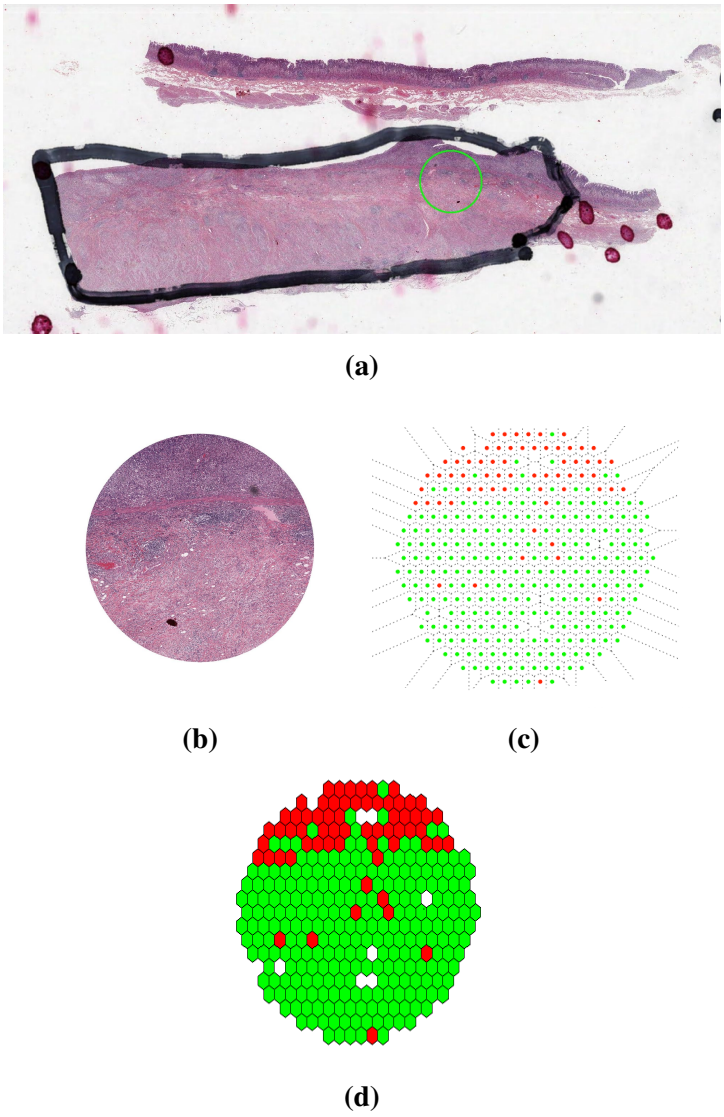


Figure 1.4: An example of highlighting the region of interest: (a) by the pathologist in black ink with the sampling target area shown as a green circle, (b) displays a zoomed in version of the sampling area and (c) Dirichlet tessellation of a set of spots after sampling the area based on Lee and Schachter (1980), (d) illustrates the same sampling area plotted using R, where the red spots show tumor spots, green indicates stroma spots, and missing data are shown in white.

The digitised images are exported in one or many files in XLS format containing two dimensional co-ordinates of the spots, as well as the classes of the spots. The classification of spots is subjective, as it is done by manual inspection, and takes a fully trained pathologist about 25 minutes to score 300 spots (Wright et al., 2015). The spot classification images can then be plotted in a new way using R software, where the classification of a spot is shown as a hexagon. Figure 1.4(a) displays a digital tissue slide example from a gastric cancer where the tumor is delineated in black and an area of interest shown as a green circle. This area of interest is shown at higher magnification

(20X) in Figure 1.4(b). Once the spots are labelled, they can be plotted as dots in Figure 1.4(c) where a gap indicates a missing spot. Then, in 1.4(d), the spots are illustrated as hexagonal shapes for each label, where the actual spot location is in the centre of the hexagon. Background definitions and related clinical data for the two sets of biomedical images, which will be considered later, are given in Sections 1.3.1 and 1.3.2.

1.3.1 Gastric cancer images

Gastric cancer is the third most common cause of cancer death in the world (Ferlay et al., 2013). This dataset was taken from a clinical trial where 50% of patients are randomised to receive chemotherapy after surgery but they were not treated with neoadjuvant treatments which means no TRG clinical variable is provided. Some digital tissue slides of gastric cancer are available online (Grabsch, 2013) without the spot information files. The 246 gastric cancer images to be studied belong to patients from the Kanagawa Cancer Center Hospital (KCCH), Yokohama, Japan who had surgery between January 2000 and February 2004 (Yamada et al., 2016a). The area of interest of these gastric cancer images is the luminal site of the tumor, which is the inner open space in the bowel. In this set of images, each patient has one image, but the image can have either single or multi-regions; though only 19 of the images have multi-regions. An example of a single-region image which contains 300 spots, is shown in Figure 1.4(c).

Table 1.1: Discrete covariates for the gastric cancer study (223 patients).

Covariate			# of patients
Pathological tumor stage	$pT = 1$	Tumor invades lamina propria, mucosae, or submucosa	4
	$pT = 2$	Tumor invades muscularis propria	35
	$pT = 3$	Tumor penetrates subserosal connective tissue without adjacent structures	30
	$pT = 4$	Tumor invades serosa with adjacent structures	154
Japanese Classification of tumor	$JS = 1$	Benign epithelial tumor	5
	$JS = 2$	Malignant epithelial tumor	17
	$JS = 3$	Non-epithelial tumor	78
	$JS = 4$	Lymphoma	46
	$JS = 5$	Metastatic tumor	56
	$JS = 6$	Tumor-like lesion	12
	$JS = 7$	Gastrointestinal polyposis	9
Lauren Classification of tumor	$LS = 1$	Intestinal type	107
	$LS = 2$	Diffuse type	116
Chemotherapy	$chemo = 1$	No chemotherapy	92
	$chemo = 2$	Chemotherapy	131
Survival Status	$Status = 0$	Alive	116
	$Status = 1$	Deceased	107

Clinical data was provided for most patients, see Table 1.1, however, some images

had incomplete clinical data and thus were removed. Hence, the total number of biomedical images matched with the clinical data was 223 out of 246.

The response variable of interest is the survival time in years (range, 0.27-9.53 years and median, 3.23 years) This is the length of survival post treatment (either time to death or time alive since treatment up to the end of the study), with survival status recorded as either deceased or alive. The other covariates are: having chemotherapy, where $chemo = 1$ indicates the patients who had only an operation and $chemo = 2$ indicating those who had chemotherapy after the surgery; slightly more patients had chemotherapy treatment (59%).

Three types of classification of tumor are considered as covariates. The American Joint Committee on Cancer (AJCC) defines the pathological classification (pT) using a staging system for gastric cancer, where pT refers to the term used in the pathological laboratory system (Shamudheen Rafiyath, 2018). This classification has four stages and depends on the diagnosis of tumor depth and how far it is progressed, where stage $pT = 1$ for the smallest and $pT = 4$ for the largest size. The meaning of each stage is shown in Table 1.1 for the layers of stomach illustrated in Figure 1.5.

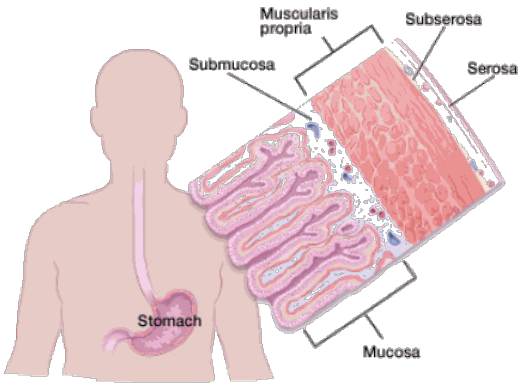


Figure 1.5: The structure of the stomach which helps to determine the stage of cancer, taken from American Cancer Society (2017).

Another type of tumor classification has seven stages. This is based on the Japanese Pathology System (JS) and it depends on the appearance of the tumor (Japanese Gastric Cancer Association, 2011). Lauren's Classification is another approach which is based on histologic features, genotypes and molecular phenotypes (Hu et al., 2012).

An essential step in understanding the clinical data is applying exploratory analysis

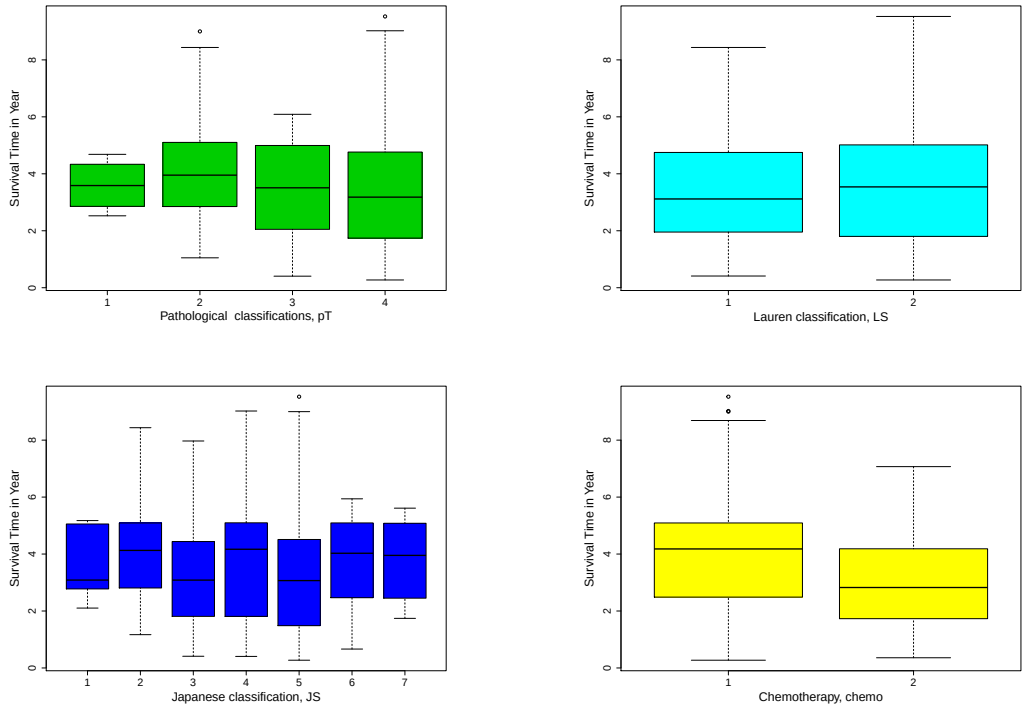


Figure 1.6: Box plots of the clinical variables for gastric cancer data occurring over the years.

to investigate whether there is an association between potential explanatory variables and the response variable. In this analysis, the information about patient survival status was ignored, and therefore, the outputs are indicative but not conclusive. We start with a test of independence for all possible pairs of categorical variables. The Pearson χ^2 test of independence is used with the null hypothesis that the two variables are independent. The assumptions of this test are: a large sample (e.g. > 100) so that the test statistic has an approximate χ^2 distribution, independence of variables and that all cells in the table have expected frequency higher than 1 and approximately 80% higher than 5 (Cochran, 1952). To achieve the required frequencies, some categories were combined, for instance, JS & pT, where $pT = 1$, $pT = 2$ and $pT = 3$ have been combined; the degrees of freedom (df) now equals 6 instead of 18. The p-value can then be found as the χ^2 right-tail probability above the observed test statistic.

Table 1.2 shows the statistical test result for each pair of categorical variables. At $\alpha = 0.05/10 = 0.005$ level of significance, using a Bonferroni correction for multiple testing (Bland and Altman, 1995), there is evidence of association between JS & LS and pT & $chemo$. As the analysis in this section is exploratory to investigate the dataset, the

Table 1.2: The tests of independence using a χ^2 test where Bonferroni correction is used for significant p-values.

Pairs of variables	χ^2	df	p-value
JS & LS	216.77	6	0.000
JS & $chemo$	9.92	6	0.118
JS & pT	7.25	6	0.299
JS & $Status$	8.23	6	0.221
LS & $chemo$	1.40	1	0.236
LS & pT	4.62	3	0.202
LS & $Status$	0.26	1	0.608
$chemo$ & pT	13.26	3	0.004
$chemo$ & $Status$	2.26	1	0.133
pT & $Status$	11.56	3	0.009

survival time response variable was only compared to each explanatory variable using one-way ANOVA test.

In general, the normality test is a method to tell if a random sample comes from a normal distribution, where a statistic is calculated to test the null hypothesis that a random sample comes from a normal distribution. The null hypothesis is rejected when the statistical values are under a certain threshold. The larger the sample size, the more likely we will get a statistically significant result. Regarding to survival time, there is a strong evidence against it follows a normal distribution using the Shapiro-Wilks test (McDonald, 2009) with p-value$= 1.1 \times 10^{-05}$, the Kruskal-Wallis test (Shapiro and Wilk, 1965) will be performed which is a non-parametric one-way ANOVA test. This test is used to identify significant differences in survival times between groups defined by a categorical independent variable. Figure 1.6 displays the distribution of the categorical variables. There are only significant differences between chemotherapy groups (p-value$= 0.0003$); where the patients who had no chemotherapy before surgery tend to live longer than those who have chemotherapy beforehand. This result is unexpected but may be done to the data being a pilot study. The same gastric cancer dataset has been analysed by Yamada et al. (2016b), but considering only the patients who had no treatment. Thus we have no evidence in gastric cancer about how life expectancy depends on whether or not chemotherapy was administered.

1.3.2 Rectal cancer images

Bowel cancer is the second most common cause of cancer death in Europe, with around $215,000$ deaths in 2012 (Cancer Research UK, 2018). The dataset provided is called the Eindhoven dataset, named after the city where the images were gathered in 2014 (Stone, 2017). This set of images were from an observational study and not a designed clinical trial. Neither the images, nor the derived spot score data, have been published.

There are multiple images per patient, but the number of images vary. In all cases, the number of spots is 300 with a fixed magnification of 20X. There are five image types, some of them are pre-treatment and some are of post-treatment resection specimens. In the Eindhoven dataset, pathologists use an automatic method to measure tumor spot density (TCD) in high resolution pathology images. This measurement is commonly used to determine the target area for sampling post-treatment images. The types of images provided were:

1. A pre-treatment image from a biopsy tissue sample. When a tumor is detected, the first stage is usually to take a small sample cut from the tumor, using a thin flexible tube (endoscope). The biopsy has no fixed area and may contain single or multi-region images. The shorthand name for this type of image is Bx.

2. A post-treatment image of the whole tumor at low resolution. The sampling area has no fixed shape and it may contain either single or multi-region. The shorthand name for whole image is W.

3. A post-treatment image at high resolution of a 3x3 mm square sampled from W. The sample is only from a region at the lumenal surface with high TCD. The location is deliberately chosen close to the surface where the biopsy has been taken before treatment. This is called L for lumen.

4. A post-treatment image at high resolution of a 3x3 mm square sampled from W, but this time the area of sampling is in the region of highest TCD anywhere within the tumor. The shorthand name for this image type is G for greatest.

5. A post-treatment image at high resolution of a 3x3 mm square sampled from W. This image has both the highest TCD in the whole tumor and is allocated in

the lumenal site, so we use the shorthand name LG meaning both lumenal and greatest.

Table 1.3: The summary count of all the biomedical rectal cancer images.

Matched images	# of images
Before matching Bx and clinical data	
W & L & G	66
W & LG	202
W & L	1
W & G	12
Incompleted	12
Total	293
After matching Bx and clinical data	
Bx & W & L & G	29
Bx & W & LG	84
Total	113

The delineation of the study area in G, L and LG is square, but for Bx and W it is polygonal. The Bx is sampled from the lumenal region, thus the image is superficial, and so the lumen sample is likely to be more correlated with the L. The W, L, G and LG images have the same coordinate system, but the Bx is different. The spot region in the whole tumor image (W) is partly overlapping with the high resolution images (L, G and LG). Each patient has between one and four different images with approximately 300 spots each. For instance, the patient can have either Bx, W, L and G images, or Bx, W and LG images, or Bx, W and L images, or Bx, W and G images. The summary of all images is given in Table 1.3 before and after a preliminary processing of matching the images with clinical data. The preliminary processing of data is vital as the image data files, virtual slides information and clinical data from the pathologists were not organized well and were very complicated to match with related images so it was time-consuming to be ready for analysis. The final set of images with clinical information includes a set of 113 images, where 29 patients have Bx, W, L and G images, and 84 patients have Bx, W and LG images.

The clinical variables are shown in Table 1.4 and will be explained in more detail before an exploratory analysis is performed. The patients analysed included 76 males and 37 females, with a median age of 62 years (range, 32-84). The clinical data have two discrete survival response times: (i) follow-up time recorded to nearest month, which is the length of survival post treatment, and (ii) disease-free survival time, which helps

Table 1.4: Discrete covariates of the rectal cancer images (113 patients).

Covariate			# of patients
Pre-operative tumor assessment	$Pr.Tstage = 1$	Early stage	1
	$Pr.Tstage = 2$	Intermediate stage	45
	$Pr.Tstage = 3$	Advanced stage	67
Tumor stage	$pT = 0$	No evidence of primary tumor	15
	$pT = 1$	Tumor invades into submucosa	2
	$pT = 2$	Tumor invades muscularis propria	23
	$pT = 3$	Tumor invades through muscularis propria into subserosa	61
	$pT = 4$	Tumor directly invades other organs or structures	12
Lymph nodes stage	$pN = 0$	No regional lymph nodes metastasis	75
	$pN = 1$	Metastatic disease in 1-3 regional nodes	28
	$pN = 2$	Metastatic disease in > 3 regionals	10
Distant metastasis Stage	$pM = 0$	No distant metastasis	112
	$pM = 1$	Distant metastasis	1
Follow-up survival status	$FU.status = 0$	Disease-free	78
	$FU.status = 1$	Alive with disease	10
	$FU.status = 2$	Dead of disease	17
	$FU.status = 3$	Dead of other cause	7
Disease-free survival status	$DF.status = 0$	Disease-free	107
	$DF.status = 1$	Not disease-free	116
Therapy type	$therapy = 1$	RTx + 5FU*	37
	$therapy = 2$	RTx + Cap*	21
	$therapy = 3$	RTx + 5FU + Ox*	55
Gender	$Gender = 1$	Male	76
	$Gender = 2$	Female	37

*Therapy types explained in text.

to see how well a treatment works, recorded to nearest month. The follow-up survival time has a range of 0-98 months, median 29 months, and its status ($FU.status$) has four categories, but pathologists usually combine groups 0 and 1 as disease-free and groups 2 and 3 as not disease-free. The range of disease-free survival time was 0-87 months, median 22 months, and its status ($DF.status$) is either disease-free or not disease-free as a binary variable.

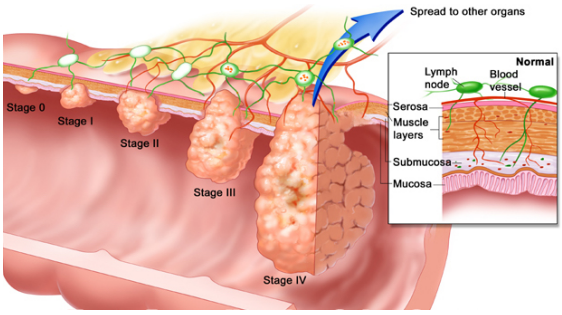
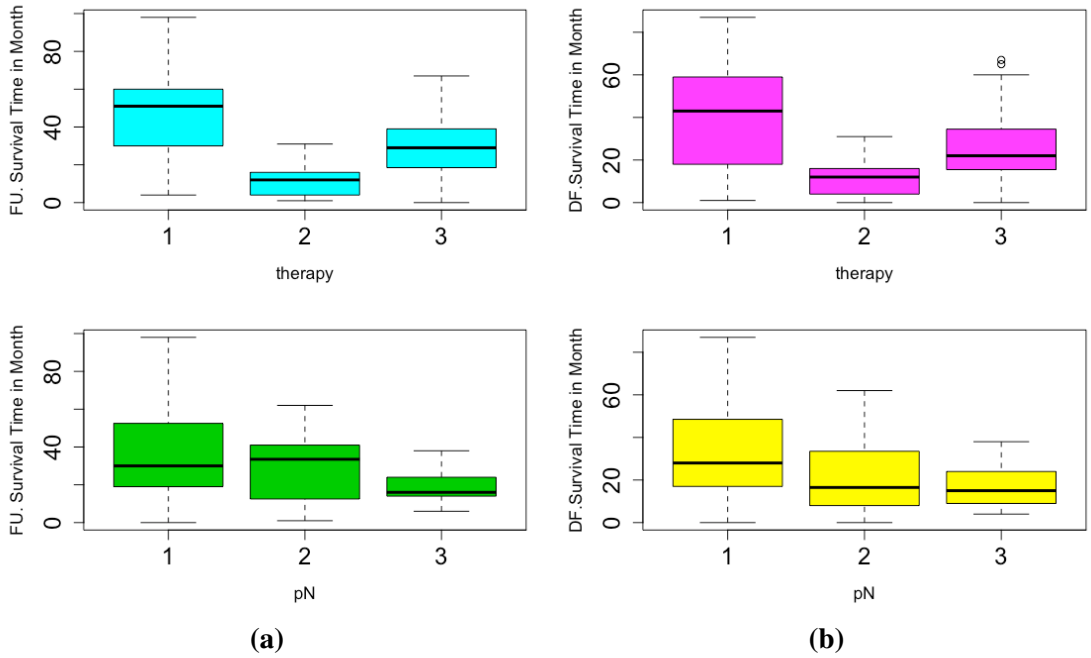


Figure 1.7: The structure of rectum including various pathologist stages of the rectal cancer (plantmedicine, 2018).

The other covariates are: preoperative tumor stage ($Pr.Tstage$) which was assessed from a biopsy sample using high resolution magnetic resonance imaging (MRI) (Greene et al., 2002; Guidelines for the Management of Colorectal Cancer, 2007). The MRI pro-

duced a series of detailed pictures of the affected areas inside the body. The $Pr.Tstage$ contains three stages, where 59% of patients have advanced $Pr.Tstage$. The American Joint Committee on Cancer (AJCC) Tumor-Node-Metastasis (TNM) staging model was also used to assess how much the cancer has spread (American Joint Committee on Cancer, 2009). The TNM classification is pathological tumor stage (pT) which has five stages where the lower stage ($pT = 0$) shows no tumor and the highest stage ($pT = 4$) shows that the tumor has invaded several organs or structures and all stages are illustrated in Figure 1.7. Guidelines for the Management of Colorectal Cancer (2007) also defines a classification of the lymph nodes (pN) which has three stages according to the number of metastatic sites. The $pN = 0$ corresponds to when there is no regional lymph node metastasis, $pN = 1$ for metastasis in 1 to 3 perirectal lymph nodes and $pN = 2$ for metastasis in 4 or more pericolic lymph nodes. Moreover, distant metastasis (pM) has two stages, when distant metastasis are present $pM = 1$, and $pM = 0$ otherwise. All patients had radiotherapy (RTx) after surgery using various chemotherapy regimens: $therapy = 1$ for Fluorouracil (5FU), $therapy = 2$ for Capecitabine (Cap) and $therapy = 3$ is a combination of 5FU and Oxaliplatin (Ox). The pathologists also provided the tumor spot density of W, L, G and LG as continuous variables, called $TCD(W)$, $TCD(L)$, $TCD(G)$ and $TCD(LG)$ respectively.

Exploratory analysis was applied to the data set of 67% male and 33% female patients to summarise its main characteristics. The aim from this step was to find out, for example, which variables suggest interesting relationships, or if there are any either categorical or continuous variables correlated with the response variable (survival time). Hence we will start by comparing pairs of categorical variables and then the response variable will be compared with all variables (either categorical or continuous variable). We start with each pair of categorical variables and use the χ^2 test of independence to determine if there is a significant relationship between the variables whilst recalling the assumptions of the test. The pT was associated with many covariantes: $therapy$ (p-value$= 0.045$), $Pr.Tstage$ (p-value$= 0.002$), pN (p-value$= 0.028$), $FU.status$ (p-value$= 0.026$) and $DF.status$ (p-value$= 0.019$). Similarly, $DF.status$ was associated with pN (p-value$= 0.028$) and pT.



(a) **(b)**

Figure 1.8: Box plots of survival time for categorical variables of rectal cancer.

Next, we consider the two survival time response variables ($FU.status$ and $DF.status$) against both categorical and continuous variables. Neither survival variable follows a normal distribution using the Shapiro-Wilks test (Royston, 1993), with p-value < 0.05.

By considering the categorical variables, the Kruskal-Wallis non-parametric test is performed. We found that various treatments are changed over FU (p-value$= 5.06 \times 10^{-10}$) and DF survival times (p-value$= 0.04$). Similarly, lymph node stage changes significantly over the FU survival time (p-value$= 1.95 \times 10^{-06}$) and the DF survival time (p-value$= 0.02$). Figure 1.8 shows how the survival times change according to $therapy$ types and pN stages for both FU and DF survival times. For both survival times, patients who had the first theory type tend to have better survival rate than other therapy types. Also, the survival time of the first stage of pN has better survival time than the third stage of pN. However, there is no correlation between the response survival time-variables and continuous variables, $TCD(W)$, $TCD(L)$, $TCD(G)$ and $TCD(LG)$.

1.4 The classification of spots

Biomedical image contains different classifications of spots. All spot types are identified, and then a combined version of spot is explained as defined by pathologists. Only the

gastric cancer images are considered as an example for spots comparison.

Table 1.5: Spots types and pathological classification.

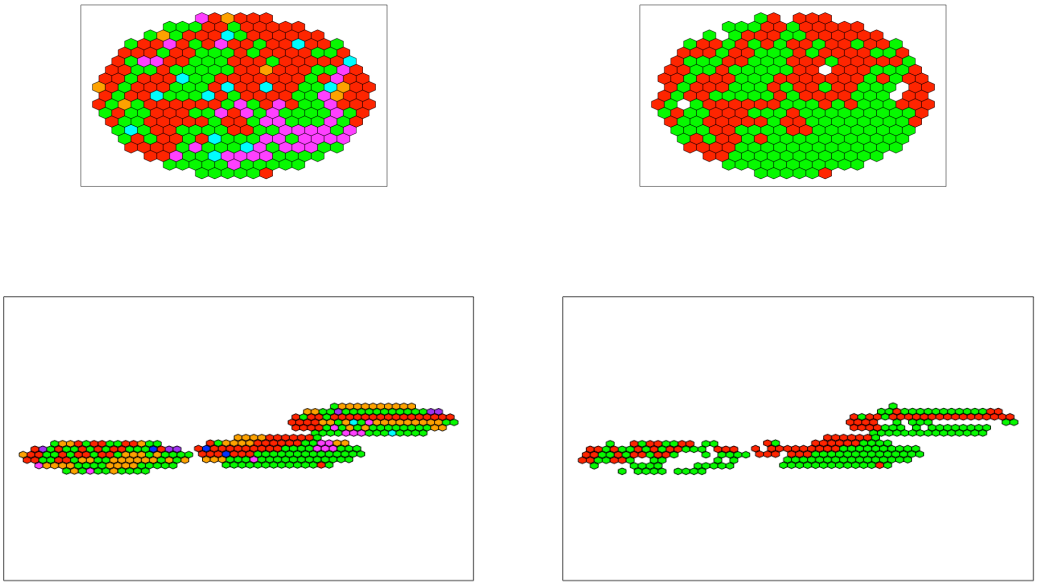
Spot type	Spot color	Description of spot	Pathological classification
0	Orange	Non informative	Exclude
1	Red	Tumor	Tumor
2	Green	Stroma	Stroma
3	Blue	Necrosis	Exclude
4	Cyan	Vessel	Stroma
5	Magenta	Inflammation	Stroma
6	Purple	Tumor lumen	Tumor
7	Yellow	Extracellular Mucus	Exclude
8	Brown	Muscle	Stroma

Table 1.6: The percentage of each spot type and combination of spot types for pathological classifications using 246 images.

Spot type	%	% of joint group	Group type
1	34.6%	36.3%	Tumor
6	1.7%		
2	52.4%	57.4%	Stroma
4	1.7%		
5	3.1%		
8	0.2%		
0	4.3%	6.3%	Excluded spots
7	0.5%		
3	1.5%		

There are nine types of spots, which are listed in Table 1.5 along with the colours used in later figures. In the same table, we define a combined grouping which has three types: tumor, stroma and excluded as recommended by an expert pathologist. The excluded spots should be removed before the analysis and are plotted with in white. This combined classification was applied before the images were analysed. The percentage of each spot type is shown in Table 1.6. The stroma has the highest percentage (57.4%) followed by tumor spot (36.3%). Figure 1.9 shows two examples of single and multi-region images of the original biomedical image and combined classification.

A box-plot and histogram of the proportion of each spot type are plotted for the 246 images and are shown in Figures 1.10a and 1.10b, respectively. It can be seen from both figures that some images have 40% of their spots of type 5. Also, most of the images have no spots of type 8, whereas some have between 20% and 30% of type 8.



(a) **(b)**

Figure 1.9: Example of gastric cancer image, plotted using R, where (a) shows the original biomedical image, and (b) shows the pathological classification where the red spots indicate tumor spots, the green spots indicate stroma spots, and the white spot shows the excluded spots.

Figure 1.10a shows that all types have outlier points except types 1 and 2. Also, some images have lots of non-informative spots (type 0), for example there is an image that has 62% of type 0 spot. Figure 1.10b shows the distribution of type frequency, using the probability density. For instance, the distribution of type 1 is positively skewed whereas the distribution of type 2 is more symmetric. The rest of types rarely appear in the images and hence their distributions are not important.

1.5 Thesis overview

In this thesis two dimensional biomedical images are considered, which are derived from pathology digital tissue slides. In Chapter 2, we cover the definition of the neighbourhood system in the hexagon grid, which has been generalised to be applicable in single and multi-region images. Some spatial statistics are defined and then compared numerically. A simulation is also used to investigate the suitability of distributional assumptions. The optimal spatial statistic to measure the degree of clustering is then determined to be Moran's I statistic (Moran, 1950) with an approximate distribution considered for

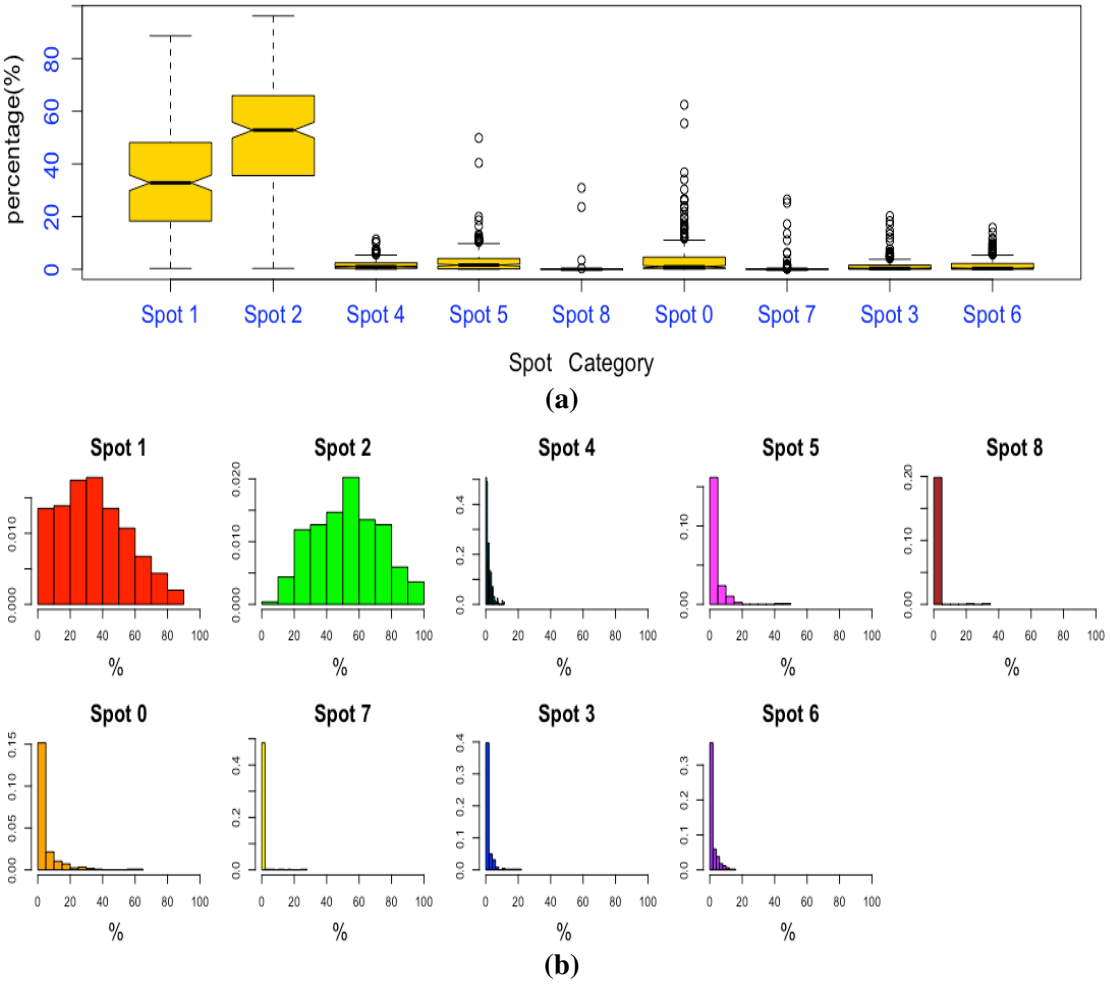


Figure 1.10: (a) Box-plot for each spot type, and (b) the density histogram of individual spot type using 246 images, where the spots are ordered similar to Table 1.6.

large images. This statistic is then computed for all the gastric and rectal images.

The I statistic is then extended in Chapter 3 to investigate anisotropy. A base for a neighbourhood system is adjusted to be defined for three directions, the directional I statistics are then computed. Several hypotheses are tested to determine if there is specified spatial structure in the different directions. From a clinical point of view, the dependency of spatial structure toward the direction of the lumen is particularly important. To calculate the I statistic and test the hypothesis that the anisotropy in the lumen site is different than the other two sites, the biomedical images are first rotated. This rotation is generalised for single and multi-region images. A new generalised statistical test is then defined under the null hypothesis that dependency in the direction of the

lumen is the same as in the other directions. This test is applied to the gastric cancer images as the lumenal direction is provided for this dataset only.

Another method of detecting the dependency of spots is explained in Chapter 4 using a binary Markov random field (*BMRF*), which explicitly includes clustering parameters. The test in this model is non-parametric with the significance computed without assuming forms from the data distribution. A new method for estimating parameters, which will be called the Iterative Method (*IM*), is introduced and explained in detail. Throughout the description of this method we include the steps of the estimation method, statistical inference and hypothesis testing. This estimation approach depends on a simulation method based on the given image, avoiding the need for the likelihood function. A generalisation of *IM* for detecting directions is explained theoretically and many applications of the *IM* for estimating parameters are included. Finally, the *IM* is compared to the I statistic in addition to comparing it with existing parameter estimation methods.

A method for prediction of spot classes using the spatial features is explained in Chapter 5. This process can help in predicting overlapped low-resolution from high-resolution images. The prediction may save clinical time and effort. The appropriate dataset to test the prediction method is the rectal cancer images as it contains several images per patient at different resolutions. In the prediction method, the spot classes are estimated by, for instance, weighted voting according to the distance between the predicted and observed spot type. Many spatial cases of prediction are covered, and the optimal method of prediction is determined.

In Chapter 6, several applications in pathology are considered statistically. Some of the questions are: can the spatial analysis help to predict chemotherapy, can the spatial statistics assist in predicting survival of patients, and can the I statistic help in predicting the survival time of patient.

Finally, in Chapter 7, the main results are summarised in addition to explaining possible future work and recommendations for pathologists.

Chapter 2

Spatial Statistics for Biomedical Images

2.1 Introduction

Mesker et al. (2007) and West et al. (2010a) subjectively observed remarkable heterogeneity in the proportion of tumour POT within individual tumours. As we introduced in Chapter 1, however, the pathologists' diagnoses are subjective which makes comparison of patterns or spatial features difficult. Thus a more objective technique is needed even in an exploratory analysis.

This chapter considers a study of many spatial statistics with a review to then recommend which statistics can be approximated by a normal distribution under the null hypothesis with the required sample size. This can then help to distinguish between images using spatial dependent features with correct statistical test. Spatial measure may help in future analysis of biomedical images to save pathologist's time as well as effort.

One of the main challenges in this chapter is determining the neighbourhood structure in a nearly regular hexagonal lattice. The resulting adjacency matrix, δ, is essential, as Cliff and Ord (1981) explained, to be able to calculate the spatial measurements. The proposed method of determining the δ matrix works well for single regions and it has been generalised in the case of multi-region images.

The spots on biomedical images are binary variables as described in Section 1.4, and hence in this chapter, the most common spatial statistics for binary data are covered: the black-white join-count, Moran's I and Geary's C statistics (Cliff and Ord, 1981; Geary, 1954; Moran, 1950). Cliff and Ord (1981, pp. 12) and Lee and Wong (2001, pp. 81)

derived the moments of each of these spatial statistics under two different assumptions. The first assumption, which has no restriction on the sampling process, is called free (*F*) sampling where the spot values $\{x_i\}$ are independently coded 0 or 1 with probability p and $p-1$ respectively. Alternatively, nonfree (*NF*) sampling (Cliff and Ord, 1981; Schabenberger and Gotway, 2005) fixes the number of spots of each type and hence only the spatial arrangement is random. Cliff and Ord (1981) and Sen (1976) proved the asymptotic normality of continuous spots. However, the output of this study is assuming normality is not good for sparsely connecting spots.

This chapter starts by defining the neighbourhood structure on the hexagonal grid of the biomedical images described in Section 2.2. The three spatial autocorrelation statistics are defined and explained in Section 2.3, and then we mathematically determine the relationship between them in Section 2.3.3. Extensive simulation studies under the null hypothesis are computed to investigate the normality of all defined spatial statistics in Section 2.4. Then, the power of I statistical tests is evaluated in Section 2.5. Some applications of spatial statistics on real biomedical images are shown in Section 2.6 with pathologists review about I. Finally, some discussion appears in Section 2.7.

2.2 Neighbourhood structure on a hexagonal grid

The nearly regular hexagon grid is not straightforward and even regular hexagonal grids are not as commonly used as square grids. This is because the distances of the six adjacent spots are not identical, and further determining neighbours becomes more tricky with missing spots. The spatial structure of the neighbourhood can, however, be summarised in elegant mathematical terms in order to calculate spatial measurements. The sharing a common border method is the most common approach to create neighbourhood structures. For example, Delaunay Triangulation (Bivand et al., 2008; Diggle, 1981) involves subdividing the hexagonal grid into triangles (mesh generation) where each triangle contains exactly one spot.

We started by applying this method to a nearly regular hexagon grid to see how well it works if there are no missing spots. Some spots, however, can be blanked off as there is no information allocated on the boundary of grid. This can be solved by adding a set of four dummy vertices as a square around the images (Turner, 2018); where each pair

of dummy spots are joined by an edge to have a rectangular window around the image and all spots lie inside the window. In this case, the method of sharing a common border can still work effectively. When there are missing spots in the grid, however, some spots which share a common boundary, should not be considered as neighbours.

The objective is to create a meaningful neighbouring structure even if we have missing spots anywhere in hexagon grid for a single region. In addition to determine the neighbourhood for multi-region image when some regions sometimes overlap. The first step is to define which cells are to be neighbours by making a Delaunay mesh of the spots based on Euclidean distance, that is to identify hexagons which share a boundary and choose a neighbour criterion to use. The second step is to assign a specific limited distance as a threshold to be used to avoid spots that are relatively far apart but share a boundary.

Now we consider the study area which has been partitioned into n nonoverlapping sub-areas. Suppose that a random variable X has been measured in each sub-area, and that the value in the typical sub-area i, is x_i, for $i = 1, \ldots, n$, where x_i is the classification of the ith spot, and a vector of classes for each dataset will be denoted $\boldsymbol{x}$. The cells have been classified into two types, which has been explained in Section 1.4, as follows:

$$x_i = \begin{cases} 1 & \text{if the class of spot } i \text{ is tumor,} \\ 0 & \text{otherwise.} \end{cases}$$

An effective combination of two steps for determining the neighbourhood structure has been introduced in this section which is a distance-based neighbour and boundary sharing approach. These steps define matrices $M^{(1)}$ and $M^{(2)}$ respectively with sizes $n \times n$. The element-wise multiplication of these matrices gives a "connection matrix", δ, determining the neighbouring structure where the contents of this matrix is explained by Moran (1948). It contains values zero and one, where one is an indicator of neighbouring spots. Thus $\delta_{ij} = 1$, if the ith and jth spots are joined, and $\delta_{ij} = 0$ otherwise. The definition of a "join" implies that $\delta_{ij} = \delta_{ji}$ for all i and j, so δ is symmetric (Cliff and Ord, 1981). Because the biomedical images have a nearly regular grid of locations, the maximum number of neighbours for each spots is six.

To define the first matrix $M^{(1)}_{n\times n} = \{M^{(1)}_{ij}\}$, the spots that share a boundary are deter-

mined by the dirichlet tessellation using the `deldir` function of tessellations (Lee and Schachter, 1980). Figure 2.1 (right) shows the shared boundaries of each spot. We let $M_{ij}^{(1)} = 1$ if the ith and jth spots share a boundary, and $M_{ij}^{(1)} = 0$ otherwise. Delaunay triangulation neighbours is a symmetric property by design, if i is a neighbour of j, then j is a neighbour of i.

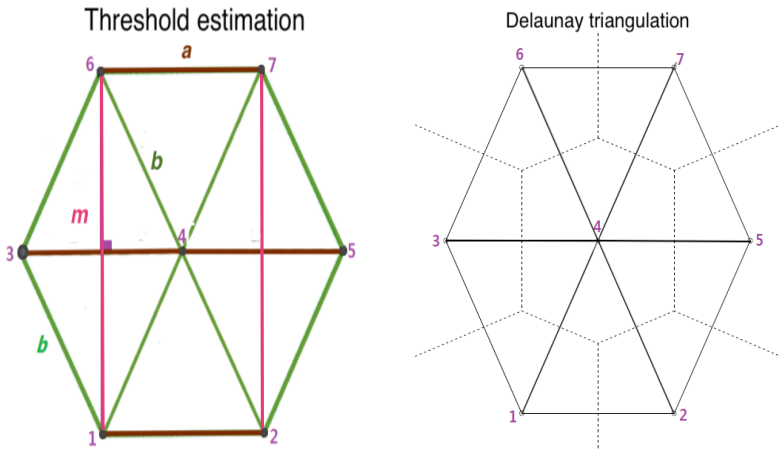


Figure 2.1: Distance-based neighbours (left) and boundary sharing approach (right) of a single hexagon, where the dotted lines represent the Delaunay Triangulation.

To define the second matrix $M_{n\times n}^{(2)} = \{M_{ij}^{(2)}\}$, a threshold must to be selected in order to exclude far away neighbours. First of all, the distances between all pairs of spots, with coordinates $(\boldsymbol{u}, \boldsymbol{v})$, are measured using an $n \times n$ Euclidean distance matrix with elements

$$D_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}, \quad i, j = 1, \ldots, n, \tag{2.1}$$

where (u_i, v_i) and (u_j, v_j) are the coordinate of the i^{th} and j^{th} spots respectively. Each element D_{ij} in Equation 2.1 is then divided by the smallest positive non-zero element which defines as $D_{\text{min}} = \min_{i,j} \{D_{ij}, i \neq j\}$, to give

$$D_{ij}^{*} = \frac{D_{ij}}{D_{\text{min}}}, \quad i, j = 1, \ldots, n. \tag{2.2}$$

The $D_{ij}^{*}, i \neq j$ is rounded off at the fifth decimal place. As the side lengths of the nearly regular hexagons are not exactly equal, we need now to define the six nearest neighbours for each spot. If $D_{ij}^{*} = 1$, this then defines the smallest distance, say a, which defines only two neighbours. The other four neighbours, which are a bit larger than a, have the

same length, say b (see Figure 2.1). Here, b occurs more than a, hence b has the highest frequency in image which can then be easily determined.

However, if we have a blanked spot, say spot 3 in Figure 2.1(left), the neighbourhood of spot 4 is reduced to five spots. Here, spots 1 and 6 are not neighbours even though they are sharing the same boundary since they are not close enough. The distance between these two spots, say m, which is used to define as a threshold to avoid spots 1 and 6 to be neighbours if spot 3 was missing. The b cannot be used as a threshold because there is some variation, by 0.00003, between b sides.

To compute the distance m for the nearly regular hexagon in Figure 2.1(left), we observe that in either the horizontal or vertical direction the spots lay exactly on a line which satisfies right angles at the junctions. Thus m is computed as follows

$$m = 2\left(\sqrt{b^2 - (a/2)^2}\right). \tag{2.3}$$

The value of a in D^*_{ij} is always 1 for any image, but b can vary a little. From Equation (2.3), the m can be approximately calculated if we assume that $a = b$. This corresponds to a regular hexagon with equal sides, the scaled distance m equals $\sqrt{3}b$. This means, in Figure 2.1, the angle between spots 3 and 6 at point 4 is 60^0 ($\angle 346 = 60^0$). Indeed, we need to select a threshold, which is smaller than $\sqrt{3}b$. We choose a threshold of maximum distance is $\sqrt{2}b$ (where $\angle 346 > 60^0$). Now the matrix $M^{(1)}$ has elements

$$M^{(2)}_{ij} = \begin{cases} 1 & \text{if } 1 < D^*_{ij} \leq \sqrt{2}b \quad \text{and} \quad i \neq j, \\ 0 & \text{otherwise,} \end{cases}$$

which defines when the spots are sufficiently close to each other. The thresholding method only works well if $a \leq b$ and $0.79 \leq |\frac{a}{b}| \leq 1$.

The matrix δ is now defined as the element-wise multiplication of two indicator matrices: $M^{(1)}_{n\times n}$, which specified the spots that share boundaries, and $M^{(2)}_{n\times n}$, which defines which spots are close to each other. Thus δ is the element-wise product of $M^{(1)}$ and $M^{(2)}$ as follows

$$\delta_{ij} = M^{(1)}_{ij} \times M^{(2)}_{ij}, \;\; i, j = 1, \ldots, n. \tag{2.4}$$

Example:

To clarify how the matrix δ can be calculated, we use a toy example of 7 spots, which has a similar structure to Figure 2.1, but is from a real image. The matrix $M^{(1)}$ is defined first determining which spots share the same boundaries as follows

$$M^{(1)}_{7\times 7} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

We need to define $M^{(2)}$ by finding the distance matrix ($D_{7\times 7}$), then all elements in this matrix are divided by the minimum non-zero positive element, 557.3666, producing the following scaled distance matrix,

$$D^{*}_{7\times 7} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & 1.00000 & 1.25830 & 1.25830 & 1.89297 & 2.30939 & 2.51660 \\ & 0 & 1.89297 & 1.25830 & 1.25830 & 2.51660 & 2.30939 \\ & & 0 & 1.00000 & 2.00000 & 1.25830 & 1.89297 \\ & & & 0 & 1.00000 & 1.25830 & 1.25830 \\ & & & & 0 & 1.89297 & 1.25830 \\ & & & & & 0 & 1.00000 \\ & & & & & & 0 \end{pmatrix} \end{array}.$$

From this matrix, the non-zero value which has the highest frequency was found, after rounding all elements to five decimal places, hence $b = 1.25830$. Then, the threshold $\sqrt{2}b = 1.7795$ is calculated, giving the matrix

$$M^{(2)}_{7\times 7} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

Now the matrix δ is calculated as the element-wise product multiplication of $M^{(1)}$ and $M^{(2)}$,

$$\delta_{ij} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

In this example, it is clear that $M^{(1)} = M^{(2)}$ as there are no missing spots (see also the first case of Figure 2.2). However, if the two matrices are not equal, neither of $M^{(1)}$ or $M^{(2)}$ is appropriate matrix and it is essential to use δ. When $M^{(1)} \neq M^{(2)}$, there are two cases: spots share a boundary but they are not close enough (Case 2 in Figure 2.2), and spots can be close but not sharing a boundary, this occurs in multi-region images when some rejoins are overlap (Case 3 in Figure 2.2).

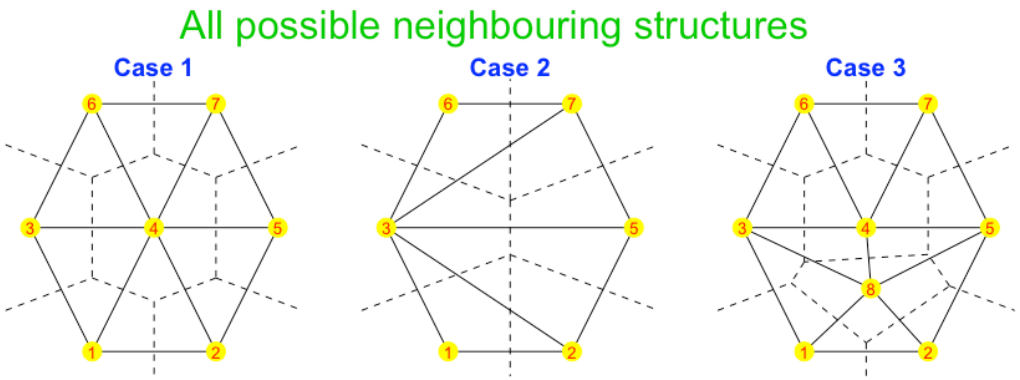


Figure 2.2: Sharing approach for all possible neighbouring structures on a single hexagon, where the dotted lines represent the Delaunay Triangulation.

The $M^{(1)}$ and $M^{(2)}$ matrices of Case 2 and Case 3 as follows:

Case 2:

$$M^{(1)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array} \text{ and } M^{(2)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

Case 3:

$$M^{(1)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \text{ and } M^{(2)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

where the different indexes are shown in red.

The method of calculating the matrix δ works well for single and multi-region images even if they have missing spots anywhere in the image. In the case of exactly regular hexagons, our approach to defining neighbours still works effectively but the threshold will be a constant $m = \sqrt{2}$. Moreover, from Equation (2.2), it never occurred in our dataset, particularly with multi-region images, that D_{min} equaled zero. If this happened in the general setting, we would need to define the next minimum number.

2.3 Spatial autocorrelation

Spatial autocorrelation is an important concept in spatial statistics, which measures the similarity between nearby observations. The similarity can also be described as clustering. This section defines in detail the various measures of autocorrelation that we considered in our project, and how they are mathematically described and computed.

After the definitions of statistics, the first two moments are given using free and non-free sampling assumptions.

All spatial statistics in this section are assumed to be asymptotically normally distributed under the null hypothesis of no spatial autocorrelation (Cliff and Ord, 1981). The alternative hypothesis, H_1, is that spatial autocorrelation exists where images can be either clustered or regular. Under H_0, the z-test is an appropriate two-tailed hypothesis test which follows a normal distribution under the Central Limit Theorem (CLT). This test seems appropriate to use as we have large samples of more than 30 spots. From a rule of thumb, this sample size choice is a boundary, however, between small and large samples. The images provided contain approximately 50 or 300 spots depending on the dataset. As soon as the theoretical expectation and standard deviation of spatial statistics are obtained, the corresponding p-value can be found in order to test for significance with an α value of 0.05, say.

To calculate the z-value (z_o) and p-value for z-tests, the critical value of z is found by subtracting the theoretical mean, and dividing by the theoretical standard deviation calculated under either *F* or *NF* sampling. Once the z value is calculated, a two-sided p-value can be found. For instance, suppose L is a spatial statistic, then z_o is $\frac{L-E(L)}{\sqrt{V(L)}}$, where $E(L)$ and $V(L)$ are the theoretical expectation and variance respectively, and then the p-value equals

$$\text{p-value } = 2P(Z < -|z_o|), \quad \text{with} \quad Z \sim N(0,1). \tag{2.5}$$

Now, the next sections are organised as follows. The join-count index for binary data, as the first set of clustering measures is described in Section 2.3.1. The second group, which are more commonly used, are Moran's I and Geary' C statistics which are explained in Section 2.3.2. At the end of this section, the relationship between spatial measurements is investigated in Section 2.3.3.

2.3.1 The join-count statistics

The join-count statistics are measures of autocorrelation within binary spatial datasets with values labelled as black and white which can be defined under both the *F* and *NF* sampling assumptions. The mathematical definition of these statistics, their first two

moments and the statistical tests are illustrated.

The join-count statistics include three coefficients: black-black (BB), black-white (BW) and white-white (WW), which count the number of joins between black and white areas, where here the black represents the tumor cell. The join may link two B spots, two W spots, or a B and a W spot. These joins are labeled BB, WW and BW respectively. Here, $x_i = 1$ if the ith spot is B (tumor spot), otherwise $x_i = 0$, for $i = 1, \ldots, n$. These spatial arrangements have been defined by Cliff and Ord (1981) and Bailey and Gatrell (1995) for testing the random scatter of, for example, black sites in black/white images. However, these statistics can only be applied to binary classes.

The observed numbers of BB, BW and WW joins in the spot structure are given by

$$BB = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}x_i x_j, \tag{2.6}$$

$$BW = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}(x_i - x_j)^2, \tag{2.7}$$

and WW is a linear function of BB and BW, where

$$WW = A - (BB + BW), \tag{2.8}$$

where A is the total number of joins in the image and $2(BB + BW + WW) = \sum_{i,j}\delta_{ij}$ *(Cliff and Ord, 1981)*, where δ is defined in Equation (2.4).

The moments of the BB, WW and BW coefficients can be evaluated under either *F*, or *NF* sampling (Bailey and Gatrell, 1995; Cliff and Ord, 1981). Under the *F* sampling assumption, the moments are equal to

$E_F(BW) = S_0 p(1-p)$,
$V_F(BW) = S_1 p(1-p) + \frac{1}{4}[S_2 p(1-p)(1-4p(1-p))]$,
$E_F(BB) = \frac{1}{2}S_0 p^2$,
$V_F(BB) = \frac{1}{4}[S_1(p^2 - p^4) + (S_2 - 2S_1)(p^3 - p^4)]$,

where

$$S_0 = \sum_{i,j} \delta_{ij}, \tag{2.9a}$$

$$S_1 = \frac{1}{2} \sum_{i,j} (\delta_{ij} + \delta_{ji})^2, \text{ and} \tag{2.9b}$$

$$S_2 = \sum_{i=1}^{n} (\sum_{j=1}^{n} \delta_{ij} + \sum_{j=1}^{n} \delta_{ji})^2. \tag{2.9c}$$

Likewise, $E_F(WW)$ and $V_F(WW)$ satisfy the same formula as BB but replacing p with $1 - p$, where E_F and V_F are the mean and variance under the *F* assumption respectively. Under the *NF* sampling assumption, the moments are equal to

$E_{NF}(BW) = \frac{S_0 n_1 n_2}{n^{(2)}}$,

$V_{NF}(BW) = \frac{1}{4}[\frac{2S_1 n_1 n_2}{n^{(2)}} + \frac{(S_2 - 2S_1) n_1 n_2 (n_1 + n_2 - 2)}{n^{(3)}} + \frac{4(S_0^2 + S_1 - S_2) n_1^{(2)} n_2^{(2)}}{n^{(4)}}] - (E_R(BW))^2$,

$E_{NF}(BB) = \frac{S_0}{2} \frac{n_1^{(2)}}{n^{(2)}}$,

$V_{NF}(BB) = \frac{1}{4} \left[S_1 [\frac{n_1^{(2)}}{n^{(2)}} - \frac{2n_1^{(3)}}{n^{(3)}} + \frac{n_1^{(4)}}{n^{(4)}}] + S_2 [\frac{n_1^{(3)}}{n^{(3)}} - \frac{n_1^{(4)}}{n^{(4)}}] + \frac{S_0^2 n_1^{(4)}}{n^{(4)}} - [\frac{S_0 n_1^{(2)}}{n^{(2)}}]^2 \right]$,

where $n^{(b)} = n(n-1)\dots(n-b+1)$, n_1 equals the number of black spots and n_2 equals the number of white spots.

Similarly, $E_{NF}(WW)$ and $V_{NF}(WW)$ satisfy the same formula as BB, but replacing n_1 by n_2 and n_2 by n_1.

The interpretation of the BB, WW and BW coefficients as follows: when the value of BW joins is small and the proportion of BB and WW joins are large, the image tends to be clustered. Whereas, if BW has a large value and the number of BB and WW joins is low, the image tends to be regular. However, if BB, WW and BW have different numbers, we will have a random image. These coefficients can also compared with the expected numbers of BB, WW and BW joins under the null hypothesis, H_0, of no spatial autocorrelation among the spots and H_1 of a spatial autocorrelation exist with either cluster or regular image. As explained in the introductory part of this section, the inference, typically based on BB, WW and BW, proceeds by assuming a normal distribution of the test statistic. For BB, for example, $z_o = \frac{BB - E(BB)}{\sqrt{V(BB)}}$, where the mean and variance come from a particular sampling assumption, is compared to the normal distribution to calculate the significance level.

2.3.2 The I and C spatial statistics

Now we will define the second group of statistics for assessing the degree of spatial autocorrelation: Moran's I and Geary's C statistics. Moran's I statistic is defined in terms of the difference between each value and the mean of all spot values (Lee and Wong, 2001) as

$$I = \frac{n}{2A} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{2.10}$$

where $z_i = x_i - \overline{x}$. The Geary's C statistic (Geary, 1954) is defined as

$$C = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} (x_i - x_j)^2}{4A \sum_{i=1}^{n} z_i^2}, \tag{2.11}$$

where δ is defined in Equation (2.4). The I and C spatial statistics can be extended to more than two spot values (Moran, 1950), but only a binary case has been considered in this work.

Inference for I and C statistics can proceed via approximate tests. These are based on the asymptotic z-test with *F* and *NF* sampling. Schabenberger and Gotway (2005) and Cliff and Ord (1981) presented the moments of I and C under the two sampling assumptions: The moments of I for both *F* and *NF* assumptions are

$E_F(I) = E_{NF}(I) = -(n-1)^{-1}$,

$E_F(I^2) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{S_0^2 (n^2 - 1)}$, and

$E_{NF}(I^2) = \frac{n[(n^2 - n3n + 3) S_1 - n S_2 + 3 S_0^2] - K[(n^2 - n) S_1 - 2n S_2 + 6 S_0^2]}{(n-1)^{(3)} S_0^2}$.

The moments of C for both *F* and *NF* assumptions are

$E_F(c) = E_{NF}(c) = 1$,

$V_F(c) = \frac{(2S_1 + S_2)(n-1) - 4S_0^2}{2(n+1) S_0^2}$, and

$V_{NF}(c) = \frac{(n-1) S_1 [n^2 - 3n + 3 - (n-1)K] + \frac{1}{4}(n-1) S_2 [n^2 + 3n - 6 - (n^2 - n + 2)K] + S_0^2 [n^2 - 3 - (n-1)^2 K]}{n(n-2)^{(2)} S_0^2)}$,

where $K = \frac{k_1}{(k_2)^2}$, $k_1 = \sum_{i \neq j} (x_{ij} - \bar{x})^2 / n$, $k_2 = \sum_{i \neq j} (x_{ij} - \bar{x})^4 / n$. and all other symbols have been introduced in Section 2.3.1.

As I is a coefficient of spatial autocorrelation, the interpretation of its value is similar to a correlation coefficient. It is restricted to the range $[-1, +1]$ with values near -1 or $+1$ indicating the image is highly dispersed or clustered respectively. However, Bailey and Gatrell (1995) explained that the C statistic, although still similar to a correlation coefficient, it is not restricted to $[-1, +1]$, and instead the p-value for the z-test is used

to interpret the value of the C statistic as has been done in Section 2.3.1.

Schabenberger and Gotway (2005) and Lee and Wong (2001, pp. 80) also interpreted the result of the I and C statistics as follows: if $I > E(I)$ and $0 < C < 1$, then spots tend to be connected to spots that have similar attribute values, so the spots are clustered. Alternatively, if $I < E(I)$ and $1 < C < 2$, attribute values of connected spots tend to be different and hence we see a dispersed pattern. If $I \simeq E(I)$ and $C \backsim 1$, spots do not show particular clustering or dispersity.

As the provided images have a large number of spots n, the test statistic formulated as $z_o = \frac{I - E(I)}{\sqrt{V(I)}}$, where the mean and variance come from either *F* or *NF* sampling, follows approximately a standard normal under the null hypotheses, where there is no spatial autocorrlation and alternative hypotheses of either cluster or regular image.

All possible values of I and C coefficients are investigated in Section 2.3.2.1 for a small n. The purpose here is to check how the arrangement of spots could affect the values of those statistics and their possible ranges.

2.3.2.1 Checking possible I and C values for small n

A toy example is considered with two sample sizes $n = 2$ and 3, with a vector $\boldsymbol{x}$, which contains the spot labels. The two samples are explained next in more detail with various arrangements of the spots.

1) $n = 2$**:**

In this example there is one join, $A = 1$, so the connection matrix δ is

$$\delta = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In this situation, $\boldsymbol{x}$ has two different cases. Firstly, if the two spots are from the same class (either $\boldsymbol{x} = (1, 1)$ or $\boldsymbol{x} = (0, 0)$), the I and C statistics are undefined. Secondly, if the two spots are from different classes (either $\boldsymbol{x} = (1, 0)$ or $\boldsymbol{x} = (0, 1)$), the I and C statistics are -1 and 1 respectively.

2) $n = 3$**:**

There are two main cases for the spot joins: all spots being joined (Case_1) and only some spots being joined (Case_2, Case_3, Case_4, Case_5 and Case_6). In the case of all spots being

joined (Case_1) the δ matrix is equal to

$$\delta_1 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \quad \text{with} \quad A = 3.$$

In this case, there are several different arrangements and values for the $\boldsymbol{x}$ class vector. All options are presented in Table 2.1 with their resulting values of the I and C statistics.

Similarly, in the case of some joined spots, there are two main subcases: when $A = 1$ (Case_2, Case_3 and Case_4) and when $A = 2$ (Case_5 and Case_6). For $A = 1$, there are three subcases: Case_2 with δ_2, where spots 2 and 3 are joined, Case_3 with δ_3, where spots 1 and 3 are joined, and Case_4 with δ_4, where spots 1 and 2 are joined. The δ matrix for these cases are shown below:

$$\delta_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \delta_3 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \text{ and } \delta_4 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

All these cases are illustrated in Figure 2.3, and their I and C statistics are presented in Table 2.1.

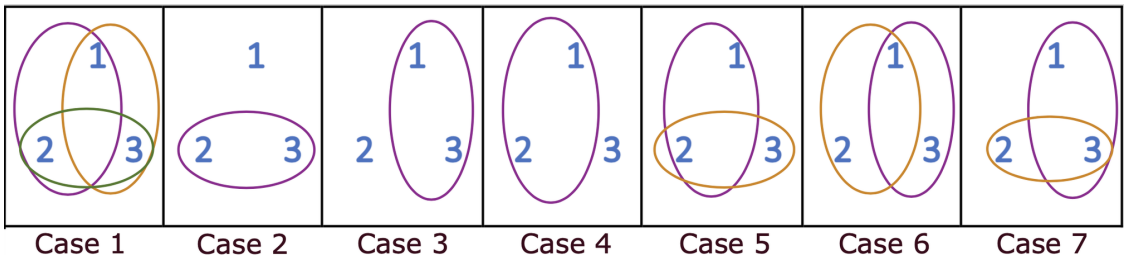


Figure 2.3: Seven different cases of joining spots with $n = 3$.

For $A = 2$, there are three possible connection matrices (Figure 2.3): Case_5 with δ_5, where spots 1 and 2 as well as 2 and 3 are joined, Case_6 with δ_6, where spots 1 and 2 as well as 1 and 3 are joined, and Case_7 with δ_7, where spots 1 and 3 as well as 2 and 3 are

Table 2.1: The results of I and C for $n = 3$

# of joined spots (A)	Case	δ	Statistic	$\boldsymbol{x}$ (1, 1, 0) (0, 0, 1)	$\boldsymbol{x}$ (1, 0, 1) (0, 1, 0)	$\boldsymbol{x}$ (0, 1, 1) (1, 0, 0)
3	Case_1	δ_1	I	-0.5	-0.5	-0.5
			C	0	0.5	0.5
1	Case_2	δ_2	I	-1	-1	0.5
			C	0	0	0
	Case_3	δ_3	I	-1	0.5	-1
			C	0	0	0
	Case_4	δ_4	I	0.5	-1	-1
			C	0	1.5	1.5
2	Case_5	δ_5	I	-0.25	-1	-0.25
			C	0.75	1.5	0.75
	Case_6	δ_6	I	-0.25	-0.25	-1
			C	0.75	0.75	1.5
	Case_7	δ_7	I	-1	-1	-0.25
			C	1.5	1.5	0.75

joined, where

$$\delta_5 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \delta_6 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \text{ and } \delta_7 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

The I and C statistics have been calculated for these three cases: (Case_5, Case_6 and Case_7) for all possible $\boldsymbol{x}$ (see Table 2.1).

As a result of using a very small example with various subcases, the range of the I statistic is between -1 and 1. However, the C value is sometimes not restricted to $[-1, 1]$, which has been confirmed by Bailey and Gatrell (1995).

2.3.3 The relationship between the spatial statistics

In this section, the I and C statistics will be written as functions of BB, BW and WW, and hence the C statistic can also be written as a function of I. The purpose here is to give another way of calculating one spatial statistic by knowing the others. For any statistic, which is calculated from basic equation, we can then calculate another statistic using a new formula as a function of known measurements.

The C statistic from Equation (2.11) can be written as a function of BW as follows,

$$C = \frac{(n-1)}{2A} \frac{BW}{\sum_{i=1}^{n} z_i^2}. \tag{2.12}$$

Then, from Equation (2.10), the I statistic can be written as a function of BB as follows:

$$I = \frac{n}{2A \sum_{i=1}^{n} z_i^2} \left[BB - \bar{x} \sum_{i,j} \delta_{ij}(x_i + x_j) + \bar{x}^2 \sum_{i,j} \delta_{ij} \right],$$

where

$$\bar{x} \sum_{i,j} \delta_{ij}(x_i + x_j) = \bar{x} \left[\sum_{i,j} \delta_{ij} x_i + \sum_{i,j} \delta_{ij} x_j \right],$$

here $\bar{x} \sum_{i,j} \delta_{ij} x_i = \bar{x} \sum_{i,j} \delta_{ij} x_j$ because of the symmetry of the δ matrix. Therefore, the I statistic as a function of BB is

$$I = \frac{n}{2A \sum_{i=1}^{n} z_i^2} \left[BB - 2\bar{x} \sum_{i,j} \delta_{ij} x_i + \bar{x}^2 \sum_{i,j} \delta_{ij} \right].$$

Finally, the C statistic can be written as a function of I, from Equation (2.11), as

$$C = \frac{(n-1)}{4A \sum_{i=1}^{n} z_i^2} \left[\sum_{i,j} \delta_{ij} (x_i - \bar{x} + \bar{x} - x_j)^2 \right].$$

Then, substituting $x_k - \bar{x}$ by z_k gives

$$C = \frac{(n-1)}{4A \sum_{i=1}^{n} z_i^2} \left[\sum_{i,j} \delta_{ij} (z_i - z_j)^2 \right].$$

Then, expanding $\sum_{i,j} \delta_{ij} (z_i - z_j)^2$ gives

$$C = \frac{(n-1)}{4A \sum_{i=1}^{n} z_i^2} \left[\sum_{i,j} \delta_{ij} (z_i^2 + z_j^2) - 2 \sum_{i,j} \delta_{ij} z_i z_j \right].$$

Next, Equation (2.10) can be rewritten as

$$\sum_{i,j} \delta_{ij} z_i z_j = I \ \frac{2A \sum_{i=1}^{n} z_i^2}{n},$$

giving

$$C = \frac{(n-1)}{4A\sum_{i=1}^{n} z_i^2} \sum_{i,j} \delta_{ij}(z_i^2 + z_j^2) - \frac{(n-1)}{n} \; I.$$

Since $\sum_{i,j} \delta_{ij} z_i^2 = \sum_{i,j} \delta_{ij} z_j^2$, the C statistic can therefore be written as

$$C = \frac{(n-1)}{2A\sum_{i=1}^{n} z_i^2} \sum_{i,j} \delta_{ij} z_i^2 - \frac{(n-1)}{n} \; I. \tag{2.13}$$

2.4 Simulation studies to investigate the distribution of the spatial statistics

A simulation study can help to assess the normality assumption of the spatial statistics, which were defined in Section 2.3, and determine which one is more informative and under which assumption (either *F* or *NF* sampling). This investigation reflects the complex situations seen in practice, such as the sample size (n) and the proportion of tumor p. The procedure of generating the datasets is explained in detail, in particular, how each study is performed, tested and reported.

As the classification of spots in the images is taken as binary, a useful motivation is to start using an example of a simple case when we have a binomial distribution as we want to introduce a guideline when normal approximation can be used. In this task, we examine in Section 2.4.1 how well can the binomial distribution approximates by the normal distribution as n increases. The approximation of the binomial distribution to the normal distribution helps to demonstrate the normality of the spatial statistics. Some normality tests are used, including the Shapiro-Wilk, to check the normality of the simulated statistics. Then in Section 2.4.2, the simulation studies, with various n and p, are implemented to asses the normality of spatial statistics and consistency of their outputs with different assumptions. Finally, the I and C statistics are then selected in Section 2.4.3 to be further investigated.

2.4.1 The approximation of the Binomial distribution

In this section, we investigate for which values of n and p how well can the binomial distribution approximated by the normal distrubution. The binomial distribution repre-

sents the probability of exactly x successes in n independent Bernoulli trials, where a given trial has two possible outcomes: a tumor with probability p and a not tumor with probability $1-p$. Here the probability of success is the same for each trial. In our experiment, the `rbinom` function is used to sample N random samples, which has been fixed to 100 replications, from a binomial distribution of spots over n trials with probability of success p. The general rule of thumb says if $n \times \min(p, 1-p) > 5$, the sample size n is sufficiently large. This principle is investigated for various sets of n and p.

To do the simulation study, samples sizes $n = 5, 10, 15, 20, 25, \ldots, 300$, and proportions $p = 0.1, 0.15, 0.20, \ldots, 0.95$ are used. For each combination of n and p, 100 datasets are simulated. The generated data is then used to see if it could be described by a normal distribution using the following normality tests: Kolmogorov-Smirnov and Shapiro-Wilk tests (Birnbaum and Tingey, 1951; Royston, 1993).

For each normality test, the p-value is saved and the results can be displayed visually using a "checkerboard"-type plot which is shown in Figure 2.4. Here, each square represents the p-value from a normality test using simulated data from the binomial distribution for specific pairs of n and p.

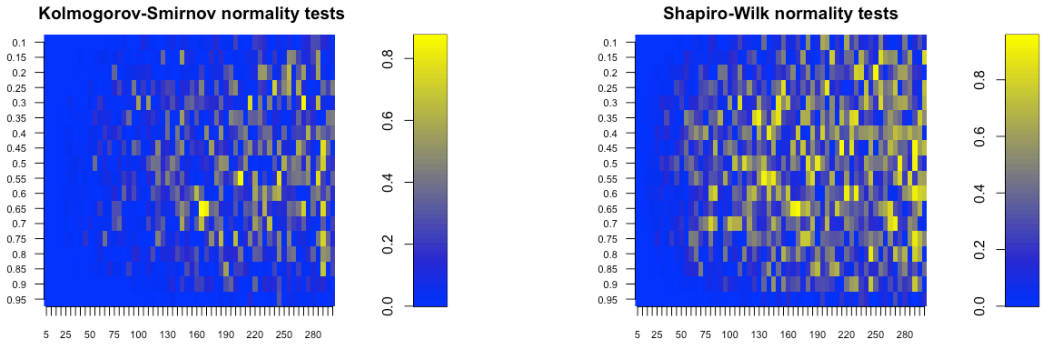


Figure 2.4: The p-value matrices from two normal tests of replicates for binomial data with various combinations of n (x-axis) and p (y-axis), where more than 0.05 refers to normal.

From Figure 2.4, it is clear that when the sample size becomes larger (>50), the binomial distribution is approximately normal. However, when the sample size is small (especially $n = 5$ and 10), the distribution tends away from normality. When the p-values are bigger than 0.05, the null hypothesis is retained at the 95% level of significance with normal approximation. In conclusion, the simulated data from the binomial distribution is approximately normal when the sample size becomes sufficiently large, say more than

$n = 100$ with various p and $20 < n < 100$ with p close to 0.5.

2.4.2 Accuracy and normality of the spatial statistic

The permutation test is a standard tool to assess the statistical significance in cases for which no distribution in known. The significance of a permutation test is shown by its p-value. Before using a normality test for the spatial statistics, we should perform some preliminary tests to make sure that the test assumptions are met.

To demonstrate the reliability of the spatial statistic p-values from the normal approximation with p-values from simulation based methods, we simulate data under the null hypothesis of no spatial autocorrelation. Both types of p-values can be computed under *F* and *NF* sampling. The idea behind these comparisons is to find which spatial statistic follows a normal distribution and under which assumption so that we can use the normal approximation in the statistical test. We also aim to test if each spatial statistic follows a normal distribution by applying the normality test.

Algorithm 1: Sampling Algorithm for generating $\boldsymbol{x}$ binary spots of length n under either *F* or *NF* assumptions.

function Binary Sampling (p, n, M);
Input : The proportion of tumor spot p, sample size n and the method of simulation M
Output: A binary classification of spots denoted by $\boldsymbol{x}$
if $M =$ F **then**
  $\boldsymbol{x} \sim bin(n, p)$;
else
  $n_1 =: np$;
  $n_2 =: n(1-p)$;
  $\boldsymbol{x} =: \langle rep(1, n_1), rep(0, n_2)\rangle$;
  $\boldsymbol{x} =: sample(\boldsymbol{x})$;
end
return $\boldsymbol{x}$;

To make the assessment of normality, the comparison between p-values requires generating sufficiently many spots under the null hypothesis to follow the Central Limit Theorem (CLT). To generate an image under H_0, the neighbouring structure of the hexagon grid defined by δ is fixed with a particular n. For given n, p and the assumption of simulation (either *F* or *NF*), the spots are obtained from Algorithm 1. In the simulated image, two sample sizes are considered $n = 50$ and 300 with two proportions of tumor spot

$p = 0.1$ and 0.5. Two main scenarios are considered in assessing normality of spatial measurements: a single sample simulation and a 1000 sample simulation. The generation methods are explained in detail, and then the result of the simulation studies and optimal spatial statistics are highlighted.

Algorithm 2: A single sample simulation of a spatial statistic, say, L with its theoretical and empirical p-values, where the empirical p-value used $k = 100$ replications.

function Single simulation $(M, n, p, Image)$;
Input : The probability of tumor p, sample size n, the method of simulation M (*F* or *NF*) and $Image_{n\times 2}$ which contains the coordinates (u_i, v_i)
output: Means, variances and p-values of normal and simulation bases
$Image[, 3] \leftarrow$ Sampling(M, n, p);
Calculate L_o;
Calculate $E(L_o)$ and $V(L_o)$ for assumption M (from Sections 2.3.1 and 2.3.2);
Calculate Th.p-value from Equation (2.5);
for $j = 1$ *to* k **do**
    $Image[, 3] \leftarrow$ Binary Sampling(p, n, M);
    Calculate spatial statistic $L[j]$;
end
$\boldsymbol{L} =: L_1, L_2, \ldots, L_k$;
Calculate $\bar{L} =: \frac{\sum_{i=1}^{k} L_i}{k}$;
Calculate $V_L =: \frac{\sum_{i=1}^{k} (L_i - \bar{L})^2}{k-1}$;
Calculate Em.p-values from Equation (2.14);
return *($E(L)$, $V(L)$,Th.p-value,$\bar{L}$,V_L,Em.p-values)*;

The steps of a single sample simulation experiment are shown in Algorithm 2, which returns a single theoretical mean, variance and p-value (Th. p-value) as well as the empirical mean, variance and p-value (Em. p-value) of each spatial statistic. By using this algorithm, we consider each possible pair of n and p to generate a random binary image under the null hypothesis for both sampling assumptions. The spatial statistics: I, C, BB, WW and BW, which are defined in Sections 2.3.1 and 2.3.2, are then calculated. From each spatial statistic, the first calculated spatial statistic is selected to be an observed value. Imagine L is one of the spatial statistics and its observed value is L_o, the theoretical mean ($E(L_o)$) and variance ($V(L_o)$) are then computed under both sampling assumptions. The test statistic is then calculated from the simulated images and we determined if the null hypothesis is accepted or rejected by computing the theoretical p-value (Th. p-value).

Now, to calculate the empirical p-value, from Algorithm 2, we sample $k = 100$ independently random images for a fixed pair of n and p to calculate 100 replicates of the spatial statistic ($\boldsymbol{L} = L_1, L_2, \ldots, L_k$). From our samples we can then calculate a sample mean $\bar{L}$, a sample variance V_L, and the empirical p-value. The empirical p-value is the probability, under the null hypothesis, of observing the observed value L_o more extreme than $\boldsymbol{L}$. For this we simply take twice the minimum proportion of either each statistic occurred less than or bigger than or equal the observed value. This p-value can be written mathematically as

$$\text{Em.p-values} = 2 \times \min\left(\frac{\sum_{i=1}^{m} I[L_i < L_o]}{m}, \frac{\sum_{i=1}^{m} I[L_i \geq L_o]}{m}\right), \tag{2.14}$$

where $I[.]$ is the indicator function.

From the single sample simulation, we will have one theoretical p-value and one empirical p-value for each spatial statistic and for each assumption. For 100 replicates of each spatial statistic, which have been used to calculate the empirical p-value, the normality test is also performed using the Shapiro-Wilk test. The results of each combination of n and p, in the case of single sample simulation, are shown in Table 2.2. From this table, the conclusions of using the theoretical and empirical p-values of all spatial statistics, with 0.05 level of significance, are almost the same with the same level of significance except the BB statistic under both sampling assumptions when $n = 50$ and $p = 0.1$. We can say that all spatial statistics adequately follow the normal distribution when $n \geq 300$ with any p for both sampling assumptions. From these results, the normal distribution may be a good approximation for the statistical test of all spatial statistics when $n = 300$, except BB statistic.

There is some evidence that some spatial statistics follow a normal distribution, in Table 2.2, when $n = 50$, but not all. However, most of the spatial statistics, when $n = 300$, are normally distributed (highlighted with red color) except I under *NF* assumption and BB. However, to decide about the normality of spatial statistics, more than a single sample simulation is needed.

Hence instead of a single sample simulation, an experiment of 1000 generated samples under the null hypothesis (that the spots are independent) is now considered to further investigate the results for the single simulation study. In this study, Algorithm 2

is used but with 1000 iterations, hence we have for each statistic and sampling method, 1000 Th. p-values and 1000 Em. p-values.

Figure 2.5 shows a plot of the differences between the theoretical and the empirical p-values against the empirical p-value. There is lots of variation between the two p-values when $n = 50$ for both values of p. When n equals 300, the variability reduces for I and for C when p is 0.5. However, there is still not enough evidence to choose among spatial statistics. The significance level can also be viewed as the percentage of times the p-value is less than α, the type I error. The level of significance is computed for all combinations of n and p for both theoretical and empirical p-values in Table 2.3. As soon as we have the same level of significance for both p-values under certain spatial statistics, assumptions, and values of n and p, a statistical test can be based on a normal approximation. An exact $\alpha = 0.05$ level of significance is considered in Table 2.3, where we expect 50 out of 1000 p-values to be less than 0.05. In addition to 0.05, the 95% confidence interval for $p = 0.05$ is $(0.04, 0.06)$ based on the binomial distribution. These lower and upper confidence limits are used as a threshold of acceptance to cover the true value $\alpha = 0.05$. From Table 2.3, the approximate agreement between the theoretical and the empirical p-values tends to be the same when $n = 300$ and $p = 0.5$, except C and BW statistics under *F* sampling.

As a result, the I statistic is normally distributed when $n = 300$ with 0.05 level of significance. Also, there is no evidence that the BB, WW and BW follow a normal distribution, and hence they will be excluded from the next experiment in the following section. Despite the fact that there is no evidence about the normality of the C statistic, this statistic will be still used to compare with I using different levels of significance. Another reason behind choosing C is because both of I and C can be generalised and applied to continuous spot values.

Table 2.2: One sample simulation study for combinations of $n = 50$ and 300 with $p = 0.1$ and 0.5 to calculate I, C, BB, WW and BW statistics with their theoretical expectations, variances and p-values under *F* and *NF* assumptions. Also under both assumptions, the empirical expectations, variances and p-values (based on a 100 simulations) are found for all statistics in addition to their normality tests.

$n = 50$ & $p = 0.1$

Statistic	Assumption	Obs. value	Empirical method			Theoretical method			Normality test
		L_o	$\bar{L}$	V_L	Em. p-value	$E(L_o)$	$V(L_o)$	Th. p-value	p-value
I	*F*	-0.027	-0.020	0.006	0.912	-0.02	0.007	0.942	0.000
	NF	-0.126	-0.023	0.006	0.082	-0.02	0.006	0.190	0.000
C	*F*	0.837	0.997	0.024	0.222	1.000	0.010	0.099	0.028
	NF	1.160	1.005	0.017	0.130	1.000	0.017	0.215	0.021
BB	*F*	0.000	1.233	2.069	0.000	1.220	2.135	0.404	0.000
	NF	0	0.977	0.785	0.000	0.996	0.828	0.274	0.000
WW	*F*	114	98.610	94.136	0.104	98.820	93.863	0.117	0.040
	NF	96	98.505	7.315	0.252	98.596	7.201	0.333	0.008
BW	*F*	8	22.157	77.822	0.108	21.960	117.302	0.197	0.097
	NF	26	22.518	8.366	0.130	22.408	8.410	0.215	0.021

$n = 50$ & $p = 0.5$

Statistic	Assumption	Obs. value	Empirical method			Theoretical method			Normality test
		L_o	$\bar{L}$	V_L	Em. p-value	$E(L_o)$	$V(L_o)$	Th. p-value	p-value
I	*F*	0.018	-0.016	0.008	0.664	-0.02	0.007	0.656	0.878
	NF	0.016	-0.018	0.007	0.596	-0.02	0.008	0.672	0.094
C	*F*	0.956	0.996	0.007	0.606	1.000	0.010	0.658	0.904
	NF	0.964	0.997	0.007	0.596	1.000	0.007	0.672	0.094
BB	*F*	43	30.617	91.830	0.184	30.500	87.250	0.181	0.422
	NF	31	30.001	12.856	0.624	29.878	12.353	0.749	0.379
WW	*F*	21	30.708	90.153	0.270	30.500	87.250	0.309	0.000
	NF	31	29.913	12.009	0.604	29.878	12.353	0.749	0.003
BW	*F*	58	60.675	30.924	0.534	61.000	122.000	0.786	0.278
	NF	60	62.086	27.258	0.596	62.245	28.167	0.672	0.094

$n = 300$ & $p = 0.1$

Statistic	Assumption	Obs. value	Empirical method			Theoretical method			Normality test
		L_o	$\bar{L}$	V_L	Em. p-value	$E(L_o)$	$V(L_o)$	Th. p-value	p-value
I	*F*	0.011	-0.006	0.001	0.564	-0.003	0.001	0.674	0.585
	NF	-0.019	-0.004	0.001	0.698	-0.003	0.001	0.645	0.042
C	*F*	0.973	1.002	0.002	0.482	1.000	0.001	0.451	0.369
	NF	1.019	1.001	0.002	0.662	1.000	0.002	0.654	0.105
BB	*F*	5	8.230	14.568	0.316	8.370	15.400	0.390	0.000
	NF	7	8.068	6.466	0.560	8.118	6.734	0.667	0.029
WW	*F*	721	677.236	691.013	0.08	677.970	705.016	0.105	0.984
	NF	676	677.601	23.940	0.74	677.718	23.976	0.726	0.186
BW	*F*	111	151.534	572.055	0.078	150.660	853.013	0.174	0.928
	NF	154	151.331	37.861	0.662	151.164	40.011	0.654	0.105

$n = 300$ & $p = 0.5$

Statistic	Assumption	Obs. value	Empirical method			Theoretical method			Normality test
		L_o	$\bar{L}$	V_L	Em. p-value	$E(L_o)$	$V(L_o)$	Th. p-value	p-value
I	*F*	-0.048	-0.005	0.001	0.192	-0.003	0.001	0.197	0.209
	NF	0.037	-0.004	0.001	0.222	-0.003	0.001	0.240	0.319
C	*F*	0.044	1.002	0.001	0.198	1.000	0.001	0.233	0.244
	NF	0.960	1.000	0.001	0.222	1.000	0.001	0.240	0.319
BB	*F*	213	208.602	640.204	0.830	209.25	650.938	0.883	0.382
	NF	211	208.382	63.105	0.694	208.55	66.498	0.764	0.706
WW	*F*	186	209.224	647.289	0.334	209.25	650.938	0.362	0.614
	NF	223	208.589	70.541	0.084	208.55	66.498	0.076	0.632
BW	*F*	438	419.174	205.832	0.162	418.5	837.000	0.500	0.109
	NF	403	420.029	208.202	0.222	419.9	206.525	0.240	0.319

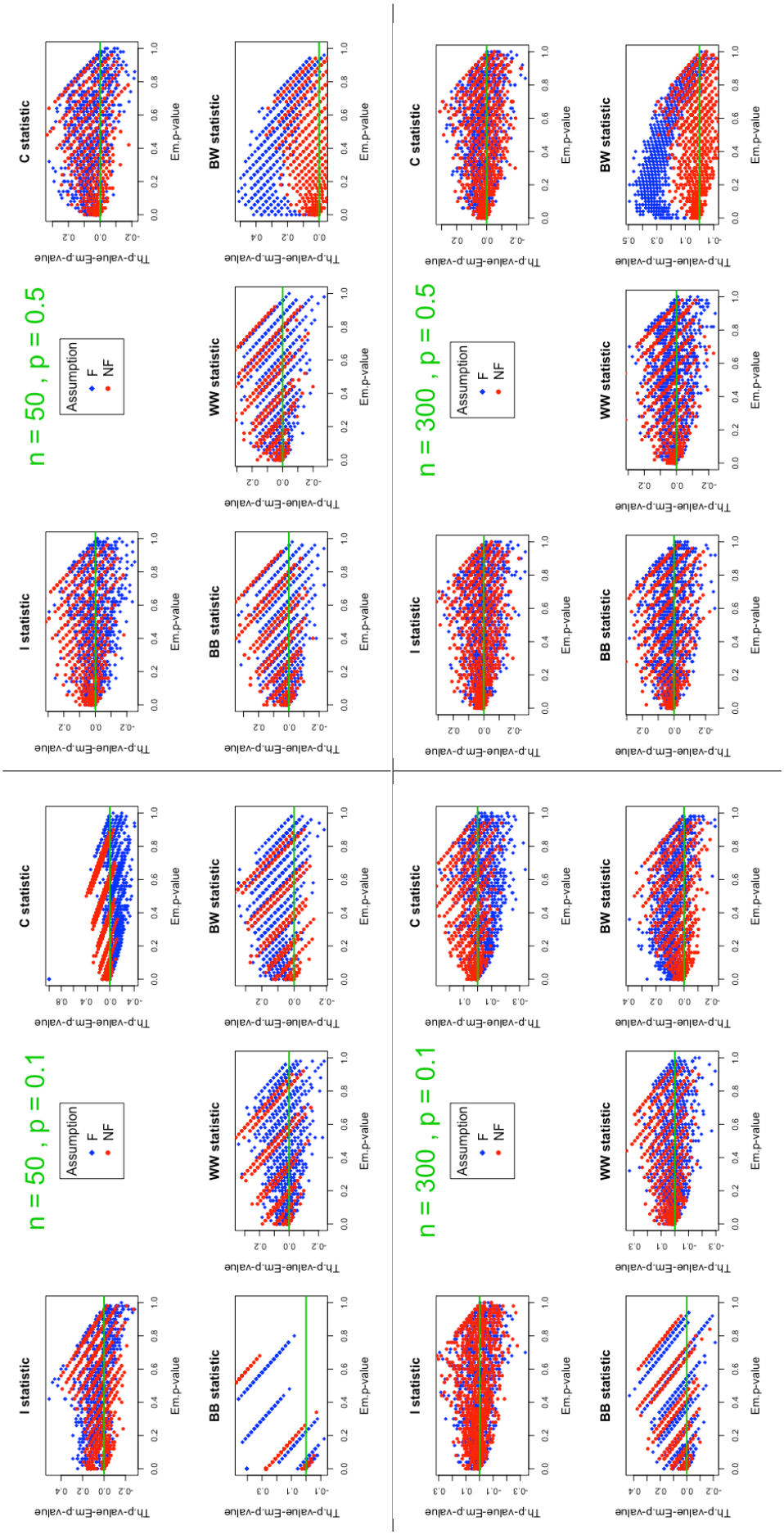


Figure 2.5: The differences of theoretical and empirical p-values against empirical p-value for 1000 simulated samples.

Table 2.3: The level of significance for 1000 sample simulations of theoretical and empirical p-values using different combinations of $n = 50$ and 300 with $p = 0.1$ and 0.5. All I, C, BB, WW and BW statistics are considered under *F* and *NF* assumptions where the shaded rows show approximate agreement between p-values.

		$n = 50$ & $p = 0.1$		$n = 50$ & $p = 0.5$	
Statistic	Assumption	Level of significance		Level of significance	
		Th.p-value	Em.p-value	Th.p-value	Em.p-value
I	*F*	0.05	0.07	0.06	0.05
	NF	0.05	0.08	0.06	0.08
C	*F*	0.15	0.05	0.02	0.06
	NF	0.04	0.09	0.06	0.08
BB	*F*	0.05	0.44	0.05	0.06
	NF	0.06	0.37	0.05	0.08
WW	*F*	0.04	0.06	0.05	0.06
	NF	0.05	0.11	0.05	0.09
BW	*F*	0.01	0.06	0.00	0.06
	NF	0.04	0.09	0.06	0.08
		$n = 300$ & $p = 0.1$		$n = 300$ & $p = 0.5$	
Statistic	Assumption	Level of significance		Level of significance	
		Th.p-value	Em.p-value	Th.p-value	Em.p-value
I	*F*	0.05	0.06	0.05	0.05
	NF	0.05	0.06	0.04	0.06
C	*F*	0.10	0.07	0.03	0.05
	NF	0.05	0.07	0.04	0.06
BB	*F*	0.04	0.08	0.04	0.05
	NF	0.05	0.08	0.05	0.06
WW	*F*	0.05	0.07	0.04	0.05
	NF	0.05	0.08	0.04	0.06
BW	*F*	0.02	0.07	0.00	0.06
	NF	0.05	0.07	0.04	0.06

2.4.3 Examining *I* and *C* statistics

In this section, the I and C statistics are considered in more detail. Even though Cliff and Ord (1981) strictly used the free sampling assumption to calculate the statistical moments of I and C when p is known, this section considers both assumptions in all simulation studies. In Section 2.4.2, we considered $n = 50$ and 300, and here the same number of spots are used in the plots of the distributions of both statistics to allow comparison between them. In Section 2.4.2, we used only one level of significance. In this section, however, more than one level of significances are considered with the same combination of n and p as in the pervious chapter. After that, different numbers of spots (n), which are between 50 and 300, are used to check the levels of significance. The aim is determining the minimum number of spots that lead to acceptable normality of the I and C spatial statistics.

Historically, the Moran (Moran, 1950) and Geary (Geary, 1954) statistics were for-

mally proved for the first time by Sen (1976), under fairly weak conditions, to be asymptotically normal for $n > 50$. He also showed that the Cliff-Ord theorem (Cliff and Ord, 1973) on asymptotic normality was incorrect. Cliff and Ord (1981) then confirmed that both the I and C statistics were asymptotically normally distributed as n increased. However, they explained that the I statistic was more robust than C as the variance of I was less sensitive to the distribution of the sample data than the differences-squared form used in Geary's C statistic. Now the experiments in this chapter will determine which spatial statistic follows the normal distribution and what is the appropriate n and sampling assumption method.

In Table 2.2, the I and C statistics sometimes adequately follow a normal distribution when $n = 300$. We still need to confirm which one is more appropriate to use. Two ways are considered to investigate and confirm the distribution of the I and C statistics: plotting the distribution of I and C statistics with large sample size, and doing a level check of significance.

Testing the assumption of distributional normality for I and C statistics can be checked by plotting their distribution using 100 replications for a fixed $n = 300$ with different proportions of tumor under both sampling methods. Previously in Table 2.2, the I statistic did not follow normality under the *NF* assumption when $p = 0.1$. From Figure 2.6, however, it is clear that the distribution of the I and C statistics (under both sampling methods) reasonably meet the normal assumption as the shape looks approximately symmetric and bell-shaped.

Secondly, a level check of significance is applied to investigate whether the p-value from the normal approximations of I and C statistics are reliable and compare them with the empirical p-value. The true significance level was estimated by simulation using the proportion of times that the null hypothesis was rejected, given that it is true. What we would like to argue here is if we use, for example, a 5% cut off, the normal approximation gives the right answer for a particular sample size. This experiment was performed using two sample sizes ($n = 50$ and 300) and two proportions of tumor ($p = 0.1$ and 0.5). For given n and p, the image was simulated under the null hypothesis of no spatial autocorrelation based on *F* and *NF* sampling methods.

Now the empirical and theoretical p-values are defined and the processes of calculating the significance level is then explained. In terms of the empirical p-value (Em.

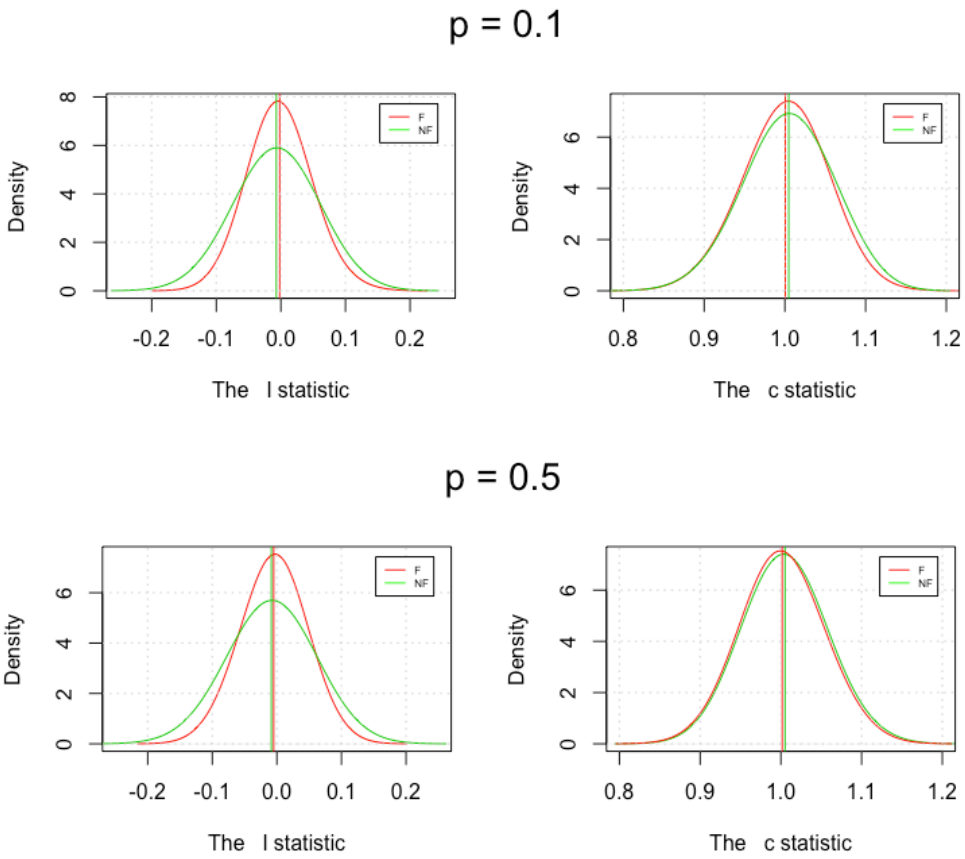


Figure 2.6: The distribution of 100 simulated I and C statistics for given $p = 0.1$ and 0.5 with $n = 300$ under free and nonfree sampling. The vertical lines show the mean of replications.

p-value), 1000 images were simulated under *F* and *NF* sampling and for each, the I and C statistics were calculated. A randomly chosen one was used in place of the observed value for each of the *F* and *NF* methods. Thus we will have $I_{o(F)}$ and $C_{o(F)}$ which are the observed values from an image simulated under free assumption, and $I_{o(NF)}$ and $C_{o(NF)}$ with the non-free assumption. The empirical p-value is calculated using Equation (2.14). This test is a two-tailed test for given α under the null hypothesis that there is no spatial autocorrelation among the spots. This procedure was repeated 1000 times, and then the empirical significance level was estimated as the proportion of p-values which are less than the nominal significance levels. In our experimental situation, we specify many values for the probability of a type I error, $\alpha = (0.01, 0.02, 0.3, 0.04, 0.05)$ which occur when a true null hypothesis is rejected.

For the theoretical p-value, the expectation and variance of the I and C statistics were calculated for each of 1000 simulated images under *F* and *NF* methods and the

usual normal distribution test was performed for both *F* and *NF* sampling. Here the theoretical p-value can be calculated under free (Th. p-value(*F*)) and non-free (Th. p-value(*NF*)) sampling. To have a wider view from the simulation study, the theoretical p-value under *F* sampling is calculated for both *F* and *NF* stimulated images. The true significance level was again estimated using $\alpha = (0.01, \ldots, 0.05)$ nominal significance levels.

All results are shown in Table 2.4. Even though the agreement of the level of significance of the C statistical test is sometimes reliable, for example, in the empirical p-value with any sampling size and $p = 0.5$ using free and nonfree sampling methods, lots of agreement for the level of significance are not good. For instance, the agreement of significance for theoretical p-values when $n = 300$ and and $p = 0.1$ under free sampling. However, it is clear that the level of significance of I statistical test is reliable for large n with any p and any sampling method. Sampling with replacement rather than without does not make any difference in demonstrating the normality of spatial statistics, but the free sampling is more appropriate as p is known. This result confirms the argument of Cliff and Ord (1981), that the statistical test of I is more accurate and better approximated by normality than C.

For the I statistic, the level check of significance was also applied for further sample sizes between 50 and 300 observations, $n = (79, 98, 111, 160, 173, 199, 271)$. However these sample sizes were not enough to have approximate normality. As a result, it is better to use a large sample, for example 300 or more, to calculate the I statistic in order to have a reliable p-value under normal approximation for any proportion of tumor.

It is important to note that when the p-value of I indicates statistical significance, a positive I value indicates a tendency toward clustering while a negative I value indicates a tendency toward regularity. To consider this case, an example of square grid for 380 spots was simulated to check several values of I. The square grid is used as it is easer in simulating different cases than hexagon. Figure 2.7 displays dispersed, random and clustered images with the relevant I statistic -1, close to zero and close to 1 respectively. However in the case of the image having only one colour, I is undefined and so the p-value can not be computed. Since I becomes closer to zero when the image becomes to one colour, we define the I statistic to be zero (with its p-value equal to 1) for images which have only one colour.

Table 2.4: The level check for I and C statistics in various cases for 5 levels of nominal significance using empirical and theoretical p-values with the assumption of free *F* and nonfree *NF* sampling.

Statistic	Simulated Image n	Simulated Image p	Level (α)	Em. p-value		Th. p-value(F)		Th. p-value (*NF*)	
				Simulate from					
				NF	*F*	*NF*	*F*	*NF*	*F*
I	50	0.1	0.05	0.06	0.06	0.04	0.04	0.05	0.03
			0.04	0.04	0.05	0.04	0.03	0.04	0.03
			0.03	0.03	0.04	0.03	0.02	0.03	0.02
			0.02	0.02	0.02	0.03	0.02	0.02	0.02
			0.01	0.01	0.01	0.02	0.01	0.02	0.01
I	50	0.5	0.05	0.05	0.05	0.05	0.05	0.05	0.06
			0.04	0.04	0.04	0.04	0.04	0.04	0.04
			0.03	0.03	0.02	0.03	0.03	0.03	0.03
			0.02	0.02	0.02	0.02	0.02	0.02	0.02
			0.01	0.01	0.01	0.01	0.01	0.01	0.01
I	300	0.1	0.05	0.05	0.05	0.05	0.05	0.05	0.05
			0.04	0.03	0.04	0.04	0.04	0.04	0.04
			0.03	0.02	0.03	0.03	0.03	0.03	0.03
			0.02	0.01	0.02	0.02	0.02	0.02	0.02
			0.01	0.01	0.01	0.01	0.01	0.01	0.01
I	300	0.5	0.05	0.05	0.05	0.05	0.05	0.05	0.05
			0.04	0.04	0.04	0.04	0.04	0.04	0.04
			0.03	0.03	0.03	0.03	0.03	0.03	0.03
			0.02	0.02	0.02	0.02	0.02	0.02	0.02
			0.01	0.01	0.01	0.01	0.01	0.01	0.01
C	50	0.1	0.05	0.06	0.07	0.05	0.12	0.06	0.12
			0.04	0.06	0.05	0.04	0.12	0.06	0.12
			0.03	0.05	0.04	0.03	0.12	0.04	0.12
			0.02	0.03	0.03	0.02	0.08	0.02	0.08
			0.01	0.01	0.02	0.01	0.05	0.01	0.05
C	50	0.5	0.05	0.05	0.05	0.05	0.03	0.05	0.03
			0.04	0.04	0.04	0.04	0.02	0.04	0.02
			0.03	0.03	0.03	0.03	0.01	0.03	0.02
			0.02	0.02	0.02	0.02	0.01	0.02	0.01
			0.01	0.01	0.01	0.01	0.01	0.01	0.004
C	300	0.1	0.05	0.06	0.05	0.05	0.09	0.06	0.09
			0.04	0.05	0.04	0.04	0.07	0.04	0.07
			0.03	0.03	0.02	0.03	0.06	0.03	0.06
			0.02	0.03	0.02	0.02	0.04	0.02	0.05
			0.01	0.02	0.002	0.01	0.03	0.01	0.02
C	300	0.5	0.05	0.05	0.05	0.05	0.04	0.05	0.03
			0.04	0.04	0.04	0.04	0.03	0.04	0.03
			0.03	0.03	0.03	0.03	0.02	0.03	0.02
			0.02	0.02	0.02	0.02	0.01	0.02	0.01
			0.01	0.01	0.01	0.01	0.01	0.01	0.01

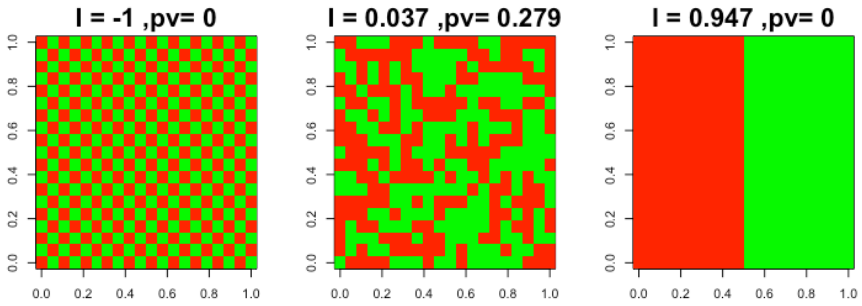


Figure 2.7: The range of images with relative I statistic, left to right, dispersed, random and clustered images with their values of I and p-values.

2.5 The power of the *I* statistical test

The purpose of studying the power of the I statistical test is to make sure that the test has the ability to correctly reject the null hypothesis. This can be done by estimating the probability of correctly rejecting a false null hypothesis of no spatial autocorrelation when H_1 is true at fixed level significance test ($\alpha = 0.05$). Let β represent the probability of a type II error when the power equals $1-\beta$. When the spots of an image become more autocorrelated, the expectation of rejecting H_0 is increased and if the power is close to 1 (or 100%), the hypothesis test is very good at detecting H_1. In this section two main points are considered: how can we generate spatially autocorrelated images and then a simulation study check is carried out to compute the power of the I statistical test.

To generate correlated spots $x_1, \ldots, x_n$, from the distribution specified by the alternative hypothesis, we can sample from a multivariate normal with zero mean vector ($\boldsymbol{\mu}$) and covariance matrix ($\boldsymbol{\Sigma}_{n\times n}$),where $\sigma_{ij} \neq 0, i \neq j$. Generating such data, the MASS em R package has a function *mvrnorm* which produces normally distributed samples with specified mean vector and covariance matrix (Venables and Ripley, 2010). This sample is then converted into a binary sample $\boldsymbol{x}$ by setting the mean as the cut-off point, where the negative values are replaced by zero, and 1 otherwise.

It is necessary, however, to appropriately define the covariance matrix. Spatial autocorrelation means that the spot at a given location depends on the spots at surrounding locations. To specify the close locations, the distance matrix in Equation (2.1) is used. To give spots that are further away, less weight and a positive-definite matrix, the covariance

matrix is defined as

$$\sigma_{ij} = e^{-\kappa D_{ij}}, \tag{2.15}$$

where κ is a parameter to control the amount of spots dependency. Now, as soon as we generated data under the alternative hypothesis, the proportion of rejections of H_0, when it is false, is then calculated. When κ is close to zero, we are expecting a greater occurrence of low p-values for these dependent spots.

Table 2.5 presents the simulation results. It shows that under the null hypothesis, the rejection rate is close to the nominal level of $\alpha = 0.05$ and that power to detect dependence increases with κ. Lastly, Figure 2.8 displays some examples of dependent spots shown as images; where the I statistic and p-value are also shown. Here as κ increases the spots becomes less correlated.

Table 2.5: Normal based tests for a fixed image of 300 spots with various κ. Dependence increases as κ decreases, and power ($= 1 - \beta$) is the proportion of 500 images in which the test rejected H_o

κ	p-value < 0.05	Power(%)	β (%)
0.001	500	100%	0%
0.003	364	73%	27%
0.005	69	14%	86%
0.007	31	6%	94%
0.009	29	6%	94%
0.010	36	7%	93%
0.030	33	7%	93%
0.050	25	5%	95%
0.070	31	6%	94%
0.090	29	6%	94%
0.100	24	5%	95%

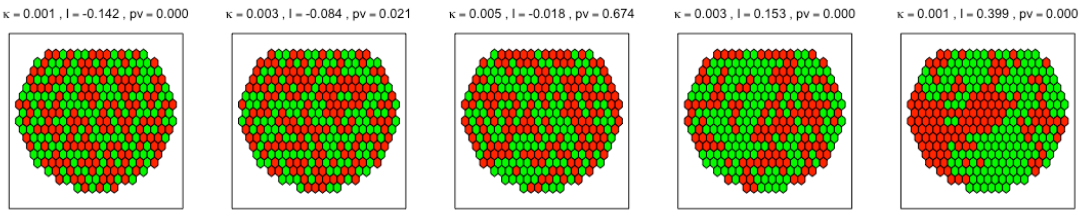


Figure 2.8: Simulating correlated images with various κ, the I statistic and its p-value are stated.

2.6 The *I* statistic for biomedical images

Moran's I statistic and proportion of tumor POT are calculated for both the gastric cancer dataset (which contains only one set) and the rectal cancer dataset (which has a set of three images). The classification method in Section 1.4 is used before calculating the I statistic. In the rectal images, the I statistic before and after treatment is compared using a paired t-test. The correlation between POT and the I statistic is also found for different cancer images, in addition to relating the I statistic with a pathologists review of the images in Section 2.6.1.

Table 2.6: The p-values of the I statistics for 231 images

p-value range	Frequency	Percentage
$0.00 - 0.04$	172	74%
$0.05 - 0.95$	59	26%

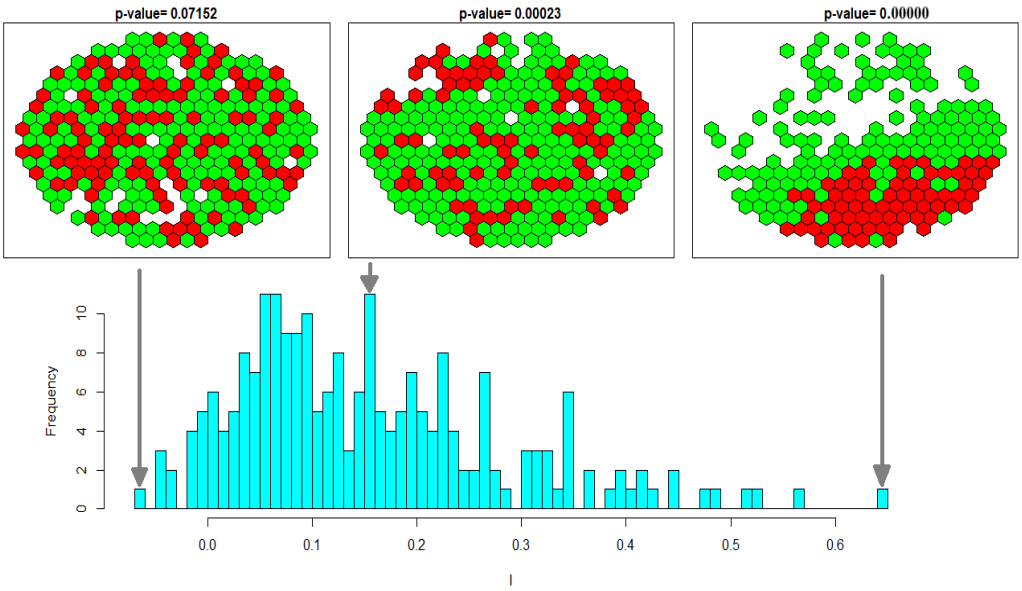


Figure 2.9: The distribution of the I statistic for all gastric cancer images, including image examples of maximum, mean and minimum of I with their p-values at the top.

In the gastric cancer dataset, there are 223 images containing single- and multi-regions. The ranges of p-values are shown in Table 2.6. Here 74% of the images are significant and 26% of the images have non-significant p-values, which means that they are independently distributed. It is important to note that all significant p-values correspond to positive I meaning that all significant images are clustered rather than dis-

persed. The distribution of I, for all gastric cancer images, is illustrated in Figure 2.9, where we picked the images of minimum, mean and maximum values of I.

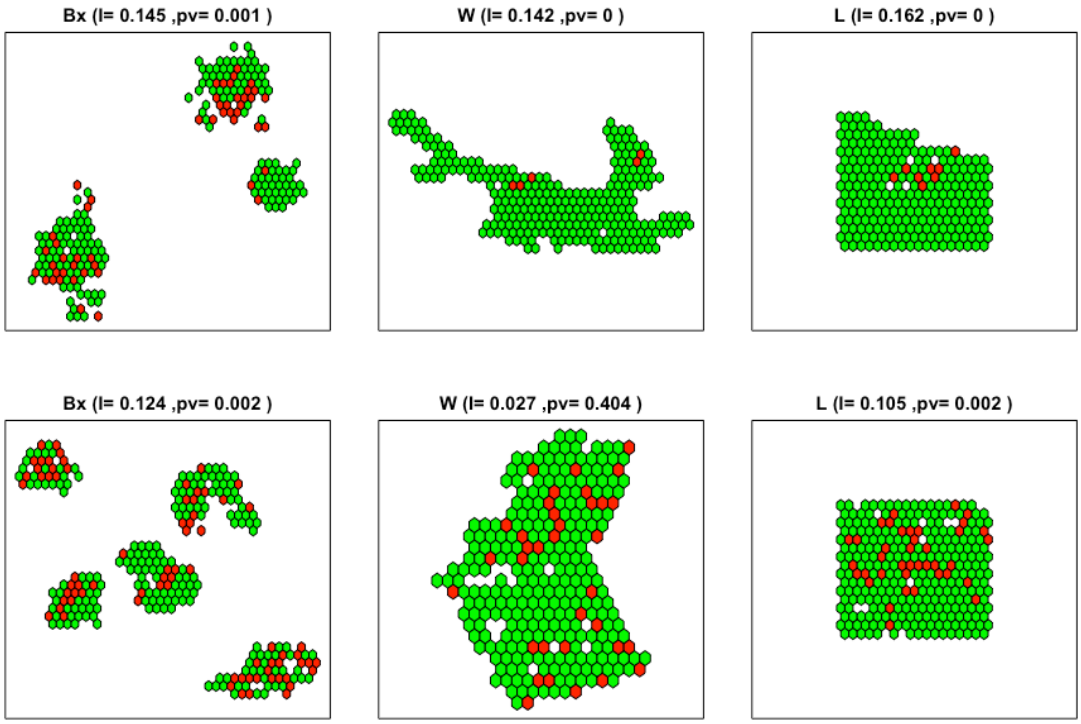


Figure 2.10: An example of matching whole (W), biopsy (BX), and L of two patients, where the I statistic and its p-value are shown at the top of each image.

The rectal cancer dataset contains 113 individuals. For each individual, there are Bx, W as well as L, G or LG images as appropriate, which have been described in Section 1.3.2. To simplify the presentation of the I statistic for all images and compare between them, L in this section and onwards will refer to either L or LG and thus G is removed from the analysis. We have chosen L as it had been sampled close the luminal site and this area is more related to Bx.

The I statistic is computed for all 113 individuals. Two examples of I for pre- (Bx) and post-treatment (W and L) are shown in Figure 2.10. The distribution of I for each individual is shown in Figure 2.11. A paired t-test is used to compare two population means for each combination of $I(Bx)$ vs $I(W)$ and $I(L)$ in Table 2.7. Here, there is a significant difference between the I statistic mean of the Bx and W images, in addition to similarly the I statistic mean of the W and L. However, there is no significant difference between the the mean of I of Bx and L, this may because the Bx sample is particularly taken from lumen surface before surgery.

Now, we will find the correlation between the POT and the I statistic. In the gastric cancer dataset, the correlation is -0.04, which is close to zero. In the rectal cancer images, the correlation between POT of L ($POT(L)$) and $I(L)$ as well as POT of W ($POT(W)$) and $I(W)$ are 0.43 and 0.53 respectively. however, there is low correlation between POT of Bx ($POT(Bx)$) and $I(Bx)$, which equals 0.01. Therefore, we can say that the I statistic gives different information than POT.

Table 2.7: The summary of the I statistic and its p-value for 113 rectal cancer images, and a paired t-test of $I(Bx)$ vs $I(W)$ and $I(L)$.

I	p-value (I)$\leq$ 0.05	$\bar{I}$	V_I	Paired t-test		
				t	Df	p-value
$I(Bx)$ vs $I(W)$						
$I(Bx)$	102 (90%)	0.23	0.014	4.57	112	1.3×10^{-5}
$I(W)$	58 (51%)	0.144	0.025			
$I(Bx)$ vs $I(L)$						
$I(Bx)$	102 (90%)	0.232	0.014	1.34	112	1.8×10^{-1}
$I(L)$	80 (71%)	0.205	0.033			
$I(W)$ vs $I(L)$						
$I(W)$	58 (51%)	0.144	0.025	-3.95	112	1.4×10^{-4}
$I(L)$	80 (71%)	0.205	0.033			

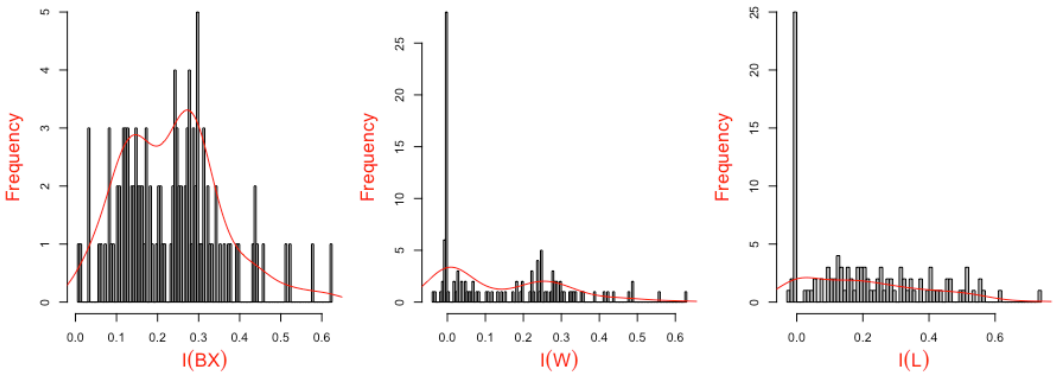


Figure 2.11: The distribution of the I statistic for Bx, W and L of 133 images.

2.6.1 Pathologist review and the I statistic

The pathologists believe that the tissue of cancers before treatment tends to be more homogeneous with high POT and TRG$=0$, which had been explained in Section 1.2. Dworak et al. (1997) showed that preoperative chemotherapy was able to reduce tumor

mass and the proportion of tumor was decreased. That means we expect after treatment the images change from a clustering pattern to be more dispersed or random. From Table 2.7, it is clear that we have more significant I statistic on the Bx images than on W and L. This differentiation is because the Bx images are before treatment, and therefore the distribution of spots is possibly affected by therapy. Also, some W and L images have only one colour, which may be the effect of treatment on the tissue.

2.7 Discussion

This chapter starts by defining the neighbourhood structure of nearly regular hexagonal grids in terms of an indicator matrix $\delta_{n\times n}$, which also works well for regular grids. This approach was developed to overcome the problem of missing spots inside the image as well as detecting neighbours in the multi-region images. The δ matrix is then used as a component of the most commonly used spatial statistics: the join-count statistics: BB, BW and WW as well as the I and C statistics. We assessed the normality of these statistics by different simulation based approaches. The assessment was under the null hypothesis and used free and non-free sampling methods with various image sizes and proportions of tumor. Then, the simulation studies are extended to check different levels of significances for I and C statistics and various sample sizes.

The test statistic based on Moran's I was found to follow a normal distribution with a large number of spots, 300 or more, hence it will be used as an accurate spatial auto-correlation statistic for large grids. Both *F* and *NF* sampling approaches give the same p-value output, but it is better to use the free sampling as p is known a priori. The study of the power of the I statistical test under the alternative hypothesis also confirms that it is appropriate to use I with a normal approximation for $n \geq 300$. The same number of spots was also pathologically determined by Wright et al. (2015) as the optimal target number of spot sampling to minimise image variation. Even though a binary setting of spots was used to calculate I, the I statistic can be generalised effectively if the classification of spots has been changed to a discrete variable.

Pathologists used hexagonal grids as they are more beneficial than a square grid. Each spot on a hexagon grid has six nearest neighbours where the distance to all these neighbours is the same. By using a hexagonal grid, we also minimise the edge effects in

a given region by considering more sides around the spot, whereas in the square shape the four sides are only considered. In practice, a hexagonal lattice of spots can be easily fitted with any given area of interest even on the curved regions, whereas a square lattice is difficult to fit to the curved regions.

Chapter 3

Detecting Anisotropy

3.1 Introduction and motivation

The way tumors spread varies depending on the structure of the surrounding tissue. Recent histopathological methods of detecting tumor directional spreading are objective and differentiable depending on the pathologists experience.

The behaviour of tumor growth is obvious in some organs, for instance, the tumor growth in a brain has the same rate in all directions, whereas in skin it starts by growing radially on the skin, then later grows vertically downwards (Cancer Research UK, 2018). However, how a cancer actually grows into the surrounding tissues in gastric and rectal cancer is not fully understood; it may grow out in a random direction from the place where it started. As the growth is anisotropic in stomach and rectum, it may have a spatial directional in which it grows faster. In another words, if there is a preferred direction, it may indicate a more aggressive or active tumor, which is subjectively evaluated by pathologists.

Underwood and Cross (2009) also explained how the tumor is spread in organs. Histological examination is of little or no value in patient management, however, and the role of the pathologist after cancer surgery is to determine the completeness of tumor removal and the extent of any spread. Only about 70% of patients with colorectal cancer undergo a potentially curative operation; in about 15-25% of patients only a palliative operation is possible because of widespread liver secondary tumors and the remainder are totally inoperable. However with pre-operative radiotherapy the proportion of poten-

tially curative operations is set to increase, and patients formerly considered inoperable because of liver metastases are now undergoing partial liver resections.

The motivation for quantitatively subjectively detecting directionality on images is: 1) to increase the chance of curing the cancer by extra treatments, 2) to help pathologists understand how cancer cells change shape as they move and spread to organs close by, 3) to avoid the spread of a tumor to another part of the body and start growing there (Cancer Research UK, 2018), 4) to evaluate the aggressiveness of the tumor, 5) to determine if the tumor grows through the layer vertically or horizontally and 6) to predict the next target area of cancer growth. However, no aggressive covariate (TRG) is provided which is tumor regression gird. Clinically, the alternative covariances for recognising aggressive of tumor are the patients who survived less and the second category of both Japanese and Lauren classifications ($JS = LS = 2$). Although the heterogeneity of overall biomedical images is important, pathologists intuitively acknowledge that the direction is also important and it may help as a diagnosis tool. Histologically, the investigation of directional pattern is a hypothesis rather than a guideline which is based on clinical practice and knowledge that the tumor in stomach and rectum can spread either linearly or radially.

In terms of directionality, pathologists are more interested in detecting pattern which is parallel to the lumen direction of the organ. The lumen, in general, refers to the inside space of a tubular structure, such as inside the stomach or rectum. Hence, it is possible statistically to investigate, in particular, if the homogeneity of spots toward lumen differs from other directions which has not been properly investigated. For directional application, the direction of luminal site is only provided for the gastric cancer dataset.

The aim of this chapter is to investigate if there is a difference between directions and, if there is a preferred direction is as the pathologists expect. New statistical tests for detecting dependency amongst spots in a specific direction are found. To do this, directional I statistics, labelled I_1, I_2 and I_3, are defined which consider three separate directions in the hexagonal grid with their corresponding neighbouring system, labelled δ_1, δ_2 and δ_3. The statistical tests in this chapter, which are specific to detecting if there is autocorrelation in particular direction, only strictly and hold under the assumptions of independent image since otherwise the distribution of statistical test is unknown. Hence only 26% of images from Chapter 2 in Table 2.6 are considered.

This chapter structured as follows. Section 3.2 defines a hexagonal neighbouring system of three directions. A statistical test for detecting anisotropy of pairs of directions is explained in Section 3.3.1. More generally, a multivariate statistical test is applied to each image in Section 3.3.2 under the null hypothesis that there is no preferred direction. The final set of work for this chapter is detecting anisotropy in a specified direction (toward the lumen site of the organ) in Section 3.3.3 using a new statistical test. The power of this test is then investigated in Section 3.4. Some discussions of key ideas of this chapter are highlighted in Section 3.5.

3.2 Connection matrices for the three directions

The general structure of the neighbourhood system of a hexagon grid was explained and defined, by the connection matrix δ, in Section 2.2. The diagonals of a hexagon, which connect diametrically opposite vertices, partition the hexagon into six triangles. These triangles help then in creating the three directional neighbouring system by picking spots in a relevant triangle creating $\delta^{(1)}$, $\delta^{(2)}$ and $\delta^{(3)}$, where each one is a subset from δ. The directional connection matrices are then used to calculate the directional I statistics, I_1, I_2 and I_3, which are defined as

$$I_r = \frac{n}{2A_r} \frac{\sum_{i,j} \delta_{ij}^{(r)} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \quad r = 1, 2, 3, \tag{3.1}$$

where $z_i = x_i - \overline{x}$, $A_r = \sum_{i,j} \delta_{ij}^{(r)}$ and $\delta_{ij}^{(r)}$ denotes the connection matrix for the neighbourhood structure in direction r. Note that here the summation of $\delta^{(1)}, \delta^{(2)}$ and $\delta^{(3)}$ gives δ.

The moments of I have been defined in Section 2.3.2 under free sampling. These formulas can be generalised for expectation and variance of the directional I statistics in given direction r as,

$$\begin{aligned} E(I_r) &= -(n-1)^{-1} \\ V(I_r) &= \frac{n^2 S_1^{(r)} - n S_2^{(r)} + 3\left(S_0^{(r)}\right)^2}{\left(S_0^{(r)}\right)^2 (n^2 - 1)} + (n-1)^{-2}, \end{aligned} \tag{3.2}$$

where

$$\left.\begin{aligned} S_0^{(r)} &= \sum_{i,j} \delta_{ij}^{(r)}, \\ S_1^{(r)} &= \frac{1}{2} \sum_{i,j} \left(\delta_{ij}^{(r)} + \delta_{ji}^{(r)}\right)^2, \text{ and} \\ S_2^{(r)} &= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \delta_{ij}^{(r)} + \sum_{j=1}^{n} \delta_{ji}^{(r)}\right)^2. \end{aligned}\right\} \tag{3.3}$$

In practice, the neighbourhood system for a direction is started by defining angles between the positive x-axis and the i^{th} spot, with coordinates (u_i, v_i), as the anti-clockwise direction to be able to identify the direction of each spot from another. These angles can be classified into three principle directions. Creating the neighbouring system of three directions is described and then a small image of six spots is used as an example. A real image is then used to illustrate the directional I statistic. The main target for using the directional I is to identify whether spots are autocorrelated in a specific direction.

We now explain how to define the $\delta^{(1)}, \delta^{(2)}, \delta^{(3)}$ matrices. A matrix of angles, say ζ, with dimensions $n \times n$ is needed to divide the angles into three groups of directions. Suppose we have n spots and the neighbourhood system defined by the connection matrix δ, then the matrix ζ is defined by the following steps:

1. To make the work more accessible to pathologists, we need to obtain the matrix of angles in degrees. To do this, an matrix called B with dimensions $n \times n$ is defined. This matrix contains the angles in radians between the positive x-axis and the n spots in the anti-clockwise direction. The arctangent function (`atan2` function in R), which returns angles in radians $(-\pi, \pi]$, is applied to a vector $(v_i - v_j, u_i - u_j)$. Mathematically, the B matrix is as follows:

$$B_{ij} = \text{atan}_2(v_i - v_j, u_i - u_j), \;\; i, j = 1, \ldots, n, \;\; i \neq j, \tag{3.4}$$

   where B_{ij} is an angle of complex number $x + iy$. As we would like to use the angle in degrees, angles in B are converted to degrees $B_{ij}^* = B_{ij} \times (180^\circ/\pi)$.

   This matrix can have a negative angle which is pointing in the opposite direction to that of a positive angle. In fact, the B^* matrix has the same properties as the distance matrix since it has a symmetric pattern.

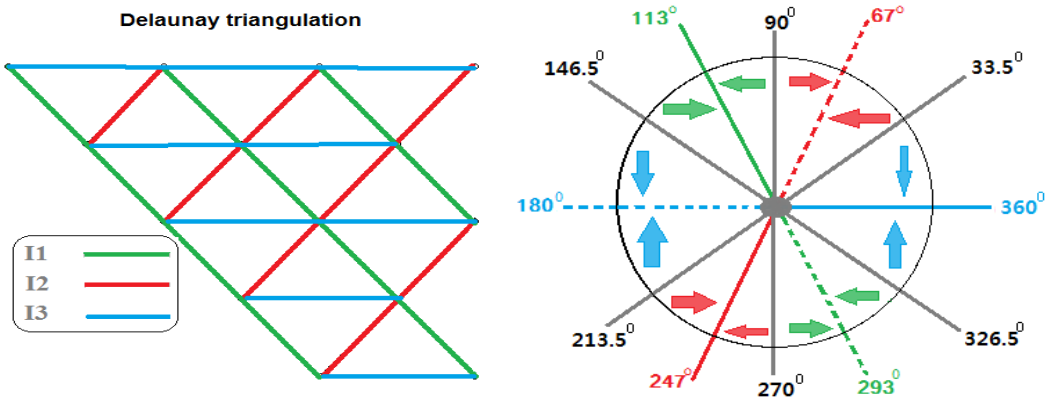


Figure 3.1: The left panel represents the location of the directional I statistic on a hexagonal grid, and the right shows the process of selecting spots allocated to the same direction and classifying them into three symmetric directions.

2. The B^*_{ij} is transformed into the interval $(0°,\ 360°]$ to facilitate the division of angles into three symmetric groups. The new matrix, say Q, is as follows:

$$Q_{ij} = \begin{cases} B^*_{ij} \mod 360° & B^*_{ij} \neq 0, \\ 360° & B^*_{ij} = 0. \end{cases}$$

   The $\mod$ keeps the positive elements and converts the negative angles to the range $180°$ and $359°$ degrees. All zero angles are replaced by $360°$ to distinguish between direction zero and an indicator of not being a neighbour.

3. Determining angles of neighbouring spots only. To do this, we define a matrix ζ which is the element-wise multiplication of Q by the matrix δ,

$$\zeta_{ij} = Q_{ij} \times \delta_{ij}, \quad i, j = 1, \ldots, n. \tag{3.5}$$

   If $\zeta_{ij} = 0$, spots i^{th} and j^{th} are not neighbours, otherwise $\zeta_{ij} \in (0°,\ 360°]$.

Now the angles in ζ are divided into three symmetric groups. To do this, the angles in this matrix are summarised as a frequency table to determine the groups of angles, and then sorted from the smallest angle, excluding zeros. Here, we will have six angles in order as shown in Figure 3.1. We then draw a straight line through the center of the circle between a pair of angles (grey lines) as the distance between angles are not equally spaced. The set of angles between a pair of grey lines represents one group, where each group refers to a hexagonal axis.

For example, let us suppose we consider Figure 3.1 as a hexagon setting, here there are three pairs of opposing angles: 360° vs 180°, 67° vs 247° and 113° vs 293°, where each pair represents a specific direction. To classify the angles into various directions and define the range of angles, six thresholds are set as a midpoint between two angles next to each other as shown by the gray lines in Figure 3.1 (right panel). In the first direction (I_1), for example, the range of angles for this direction is $90^\circ < \zeta_{ij} \leq 146.5^\circ$ or $270^\circ < \zeta_{ij} \leq 326.5^\circ$. So if ζ_{ij} is allocated in these ranges $\delta_{ij}^{(1)} = 1$, and zero otherwise. The matrices $\delta^{(2)}$, for the second direction, and $\delta^{(3)}$, for the third direction are defined in the same way. The method of defining the neighbouring structure for three directions on the hexagonal grid works efficiently for both single and multi-region region images, even after image rotation.

A small example of dividing angles into different directions:

The neighbouring system of three directions is explained on a small example of 4 spots (part of a real image in Figure 3.2), where the red spot refers to tumor and the green non-tumor and the numbers show the spots order. The connection matrix for this example is

$$\delta = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

The B matrix is firstly defined from Equation (3.4) in degrees as:

$$B = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.00000^\circ & 180.0000^\circ & -66.58666^\circ & -113.41334^\circ \\ 0.00000^\circ & 0.0000^\circ & -37.58894^\circ & -66.58666^\circ \\ 113.41334^\circ & 142.4111^\circ & 0.00000^\circ & 180.00000^\circ \\ 66.58666^\circ & 113.4133^\circ & 0.00000^\circ & 0.00000^\circ \end{pmatrix} \end{array}.$$

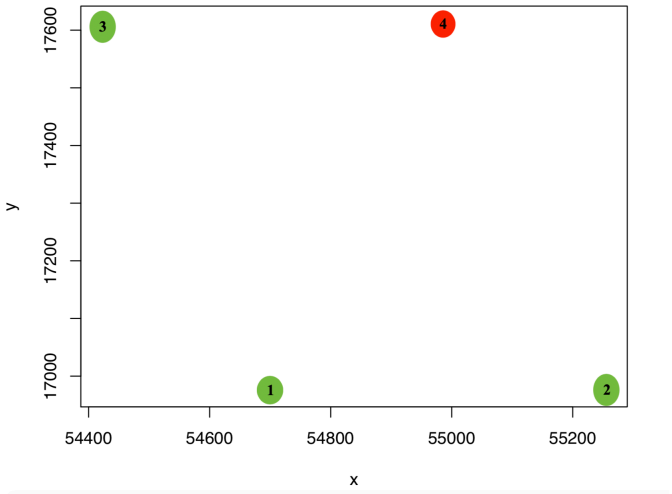


Figure 3.2: A small example of 4 spots.

Next, we convert all angles to positive values and replace all zeros by 360° as follows:

$$Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 360.00000° & 180.0000° & 293.4133° & 246.5867° \\ 360.00000° & 360.0000° & 322.4111° & 293.4133° \\ 113.41334° & 142.4111° & 360.0000° & 180.0000° \\ 66.58666° & 113.4133° & 360.0000° & 360.0000° \end{pmatrix} \end{array}.$$

The ζ matrix is then calculated using Equation (3.5) as:

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 180° & 293° & 247° \\ 360° & 0 & 0 & 293° \\ 113° & 0 & 0 & 180° \\ 67° & 113° & 360° & 0 \end{pmatrix} \end{array}.$$

Lastly, the angles in ζ matrix are classified into three groups after determining six midpoints between these angles. The matrices of the three directions are as follows:

$$\delta^{(1)} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \delta^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \text{ and } \quad \delta^{(3)} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

where the summation of these matrices equals δ.

3.3 Statistical tests for detecting anisotropy

The anisotropy can be effectively investigated in biomedical images using the directional I statistic. A couple of statistical tests can be carried out under the assumption of independent image. That means that a biomedical image can have no clustering but it may have direction. In this section, statistical tests are performed on the hypothesis about the direction. Each test statistic is constructed using the theoretical mean and variance. To assess the null hypothesis of each test, the p-value is compared to the significance level, we reject H_0 when the p-value is less than $\alpha = 0.05$. Three z-tests for an image are firstly defined in Section 3.3.1 and a generalisation, using a bivariate normal distribution, is explained in Section 3.3.2. More importantly, pathologists are interested in determining if the autocorrelation of spots in the direction of the lumen differs from the other directions. A null hypothesis of no significant difference in autocorrelation between the direction of lumen verses the other directions is stated and performed in Section 3.3.3. Applications of all tests are shown in Section 3.3.4 on real images.

3.3.1 Directional z-tests

As we have three directional I statistics, I_1, I_2 and I_3, the differences between pairs of directions for an image can be considered in detecting directionality. This can be achieved by using three statistical tests, say a pair I_r and I_s, where $r, s \in \{1, 2, 3\}$ and $r \neq s$. The assumption of this is that the distribution of I under the null hypothesis of no overall autocorrelation between spots. The directional pattern can be assessed under the

null hypotheses of directional I is that there is no direction versus directionality as an alternative hypothesis. The three statistical tests are defined in detail including direction of the first two moments which are used to define the probability distribution under the null hypothesis.

As the distribution of any I statistic is normal and the variance of difference between pairs of I is already known, a one-sample z-test can be used. Suppose that $E(I_r - I_s)$ and $Var(I_r - I_s)$ represent the theoretical mean and variance of the difference between two I statistics, and the null hypothesis for comparing the two statistics is $H_0 : I_s - I_r = 0$, meaning that there is no significant difference between pairs of directions I_r and I_s. The alternative hypothesis is two sided. The z-statistics are formed using the formulas:

$$z_1 = \frac{(I_1 - I_2) - E(I_1 - I_2)}{sd(I_1 - I_2)}, \tag{3.6a}$$

$$z_2 = \frac{(I_1 - I_3) - E(I_1 - I_3)}{sd(I_1 - I_3)}, \text{ and} \tag{3.6b}$$

$$z_3 = \frac{(I_2 - I_3) - E(I_2 - I_3)}{sd(I_2 - I_3)}. \tag{3.6c}$$

Now, we will need to calculate the expectations and variances of $I_r - I_s$. Some basic explanations are firstly defined. Under the null hypothesis, let $x_1, \ldots, x_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Then, if x_i is N(μ,σ^2) for each i, the first four moments are $E(x_i) = \mu$, $E[(x_i - \mu)^2] = \sigma^2$, $E[(x_i - \mu)^3] = 0$ and $E[(x_i - \mu)^4] = 3\sigma^4$ (Cliff and Ord, 1981). These moments are essential in deriving $Var(I_r - I_s)$, which are used later. For a given random sample with observed values $x_1, \ldots, x_n$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, Cliff and Ord (1981) found the variates z_i, corresponding to the observed values $z_i = x_i - \bar{x}$, to have expectations:

$$\left.\begin{aligned} E(z_i) &= 0, \\ E(z_i^2) &= (1 - \frac{1}{n})\sigma^2, \\ E(z_i z_j) &= -\frac{\sigma^2}{n}, \\ E(z_i^2 z_j^2) &= \frac{(n^2 - 2n + 3)\sigma^4}{n^2}, \\ E(z_i^2 z_j z_k) &= -\frac{(n-3)\sigma^4}{n^2}, \\ E(z_i z_j z_k z_l) &= \frac{3\sigma^4}{n^2}. \end{aligned}\right\} \tag{3.7}$$

All equations (3.7) had been proved by Cliff and Ord (1981), and we have checked them mathematically to verify the results. The moments of $I_r - I_s$ are explained generally, and these moments will then be calculated for each pair of directional I statistics in order to compute the z-tests under $H_0 : I_r = I_s = 0$ in Equation (3.6).

The expectation and variance of $I_r - I_s$, where $r \neq s$

The expectation of the non-directional I depends only on n and has no spatial information. That means the expectation of any directional I has the same value. Thus, the first moment of $I_r - I_s$ equals zero. The $E(I_r - I_s)$ is expressed algebraically as follows:

$$E(I_r - I_s) = E\left[\frac{n}{2}\left(\frac{\sum_{i\neq j}\left(\frac{\delta_{ij}^{(r)}}{A_r} - \frac{\delta_{ij}^{(s)}}{A_s}\right) z_i z_j}{\sum_{i=1}^{n} z_i^2}\right)\right].$$

As we introduced at the beginning of this chapter, the statistical test for detecting direction only holds under the assumption of no autocorrelation between spots which is the same assumption of I for independent spots. Therefore, the expected value of the ratio is equal to the ratio of the expected values shown in following equation as

$$E(I_r - I_s) = \frac{n}{2A_rA_s}\frac{\left(\sum_{i\neq j} A_s\delta_{ij}^{(r)} - A_r\delta_{ij}^{(s)}\right) E(z_i z_j)}{\sum_{i=1}^{n} E\left(z_i^2\right)}.$$

Using Equations (3.7), for $E(z_iz_j)$ and $E(z_i^2)$, in the above equation, we have

$$\begin{aligned} E(I_r - I_s) &= \frac{1}{2A_rA_s\left(n-1\right)}\left(\sum_{i\neq j} A_s\delta_{ij}^{(r)} - A_r\delta_{ij}^{(s)}\right) \\ &= \frac{1}{2A_rA_s\left(n-1\right)}\left(A_sS_0^{(r)} - A_rS_0^{(s)}\right) \qquad (3.8) \\ &= \frac{1}{2\left(n-1\right)}\left(\frac{S_0^{(r)}}{A_r} - \frac{S_0^{(s)}}{A_s}\right), \qquad (3.9) \end{aligned}$$

where $S_0^{(r)}$ and $S_0^{(s)}$ are given by Equation (3.3) as $2A_r$ and $2A_s$ respectively, the $E(I_r - I_s)$ is equal to zero.

Now moving to the variance of $I_r - I_s$ with dependent directions, which is as follows

$$Var(I_r - I_s) = Var(I_r) + Var(I_s) - 2Cov(I_r, I_s), \tag{3.10}$$

where

$$Cov(I_r, I_s) = E(I_r I_s) - E(I_r)E(I_s).$$

As it has been defined in Equation (3.2), the expectations of various directional I statistics are not dependent on the neighbouring system, therefore $E(I_r) = E(I_s)$. However the tricky term is $E(I_r I_s)$ which can be expressed as follows

$$E(I_r I_s) = \frac{n^2}{2A_r A_s} \left(\frac{E\left[\left(\sum_{i \neq j} \delta_{ij}^{(r)} z_i z_j \right) \left(\sum_{k \neq l} \delta_{kl}^{(s)} z_k z_l \right) \right]}{(n-1)(n+1)\sigma^4} \right).$$

Here $E_{rs} = E\left[\left(\sum_{i \neq j} \delta_{ij}^{(r)} z_i z_j \right) \left(\sum_{k \neq l} \delta_{kl}^{(s)} z_k z_l \right) \right]$ can be extended to include eight possible scenarios:

$$\begin{pmatrix} i = k \\ j = l \end{pmatrix}, \begin{pmatrix} i \neq k \\ j = l \end{pmatrix}, \begin{pmatrix} i = k \\ j \neq l \end{pmatrix}, \begin{pmatrix} i \neq k \\ j \neq l \end{pmatrix}, \begin{pmatrix} i = l \\ j \neq k \end{pmatrix},$$

$$\begin{pmatrix} i \neq l \\ j = k \end{pmatrix}, \begin{pmatrix} i = l \\ j = k \end{pmatrix}, \text{and} \begin{pmatrix} i \neq l \\ j \neq k \end{pmatrix}.$$

In the first scenario, for instance $\begin{pmatrix} i = k \\ j = l \end{pmatrix}$, this leads to

$$E_{rs} = E\left(\sum_{i=k \neq j=l} \delta_{ij}^{(r)} \delta_{ij}^{(s)} z_i^2 z_j^2 \right),$$

here for given i^{th} and j^{th}, we sum over the multiplication of $\delta_{ij}^{(r)} \delta_{ij}^{(s)}$.

Now E_{rs} is defined using all scenarios, which will then be extended using the same method with has been explained in Cliff and Ord (1981) which is used for the expectation of the I statistic. The expectation of various scenarios are as follows:

$$E_{rs} = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8, \tag{3.11}$$

$$
\begin{aligned}
E_{rs} = & E\left(\sum_{i=k\neq j=l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) + E\left(\sum_{l=j\neq i\neq k} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) \\
& + E\left(\sum_{k=i\neq j\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) + E\left(\sum_{i\neq j\neq k\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) \\
& + E\left(\sum_{l=i\neq j\neq k} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) + E\left(\sum_{k=j\neq i\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) \\
& + E\left(\sum_{j=k\neq l=i} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) + E\left(\sum_{l\neq i\neq j\neq k\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right), \qquad (3.12)
\end{aligned}
$$

which can be simplified as

$$
\begin{aligned}
E_{rs} = & E\left(\sum_{i\neq j} \delta_{ij}^{(r)}\delta_{ij}^{(s)} z_i^2 z_j^2\right) + E\left(\sum_{j\neq i\neq k} \delta_{ij}^{(r)}\delta_{kj}^{(s)} z_i z_j^2 z_k\right) \\
& + E\left(\sum_{i\neq j\neq l} \delta_{ij}^{(r)}\delta_{il}^{(s)} z_i^2 z_j z_l\right) + E\left(\sum_{i\neq j\neq k\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right) \\
& + E\left(\sum_{i\neq j\neq k} \delta_{ij}^{(r)}\delta_{ki}^{(s)} z_i^2 z_j z_k\right) + E\left(\sum_{j\neq i\neq l} \delta_{ij}^{(r)}\delta_{jl}^{(s)} z_i z_j^2 z_k\right) \\
& + E\left(\sum_{j\neq i} \delta_{ij}^{(r)}\delta_{ij}^{(s)} z_i^2 z_j^2\right) + E\left(\sum_{l\neq i\neq j\neq k\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} z_i z_j z_k z_l\right). \qquad (3.13)
\end{aligned}
$$

Equation (3.13) can be simplified by substituting some terms using Equation (3.7).

$$
\begin{aligned}
E_{rs} = & \sum_{i\neq j} \delta_{ij}^{(r)}\delta_{ij}^{(s)} \left(\frac{(n^2-2n+3)\sigma^4}{n^2}\right) + \sum_{j\neq i\neq k} \delta_{ij}^{(r)}\delta_{kj}^{(s)} \left(\frac{-(n-3)\sigma^4}{n^2}\right) \\
& + \sum_{i\neq j\neq l} \delta_{ij}^{(r)}\delta_{il}^{(s)} \left(\frac{-(n-3)\sigma^4}{n^2}\right) + \sum_{i\neq j\neq k\neq l} \delta_{ij}^{(r)}\delta_{kl}^{(s)} \left(\frac{3\sigma^4}{n^2}\right) \\
& + \sum_{i\neq j\neq k} \delta_{ij}^{(r)}\delta_{ki}^{(s)} \left(\frac{-(n-3)\sigma^4}{n^2}\right) + \sum_{j\neq i\neq l} \delta_{ij}^{(r)}\delta_{jk}^{(s)} \left(\frac{-(n-3)\sigma^4}{n^2}\right). \\
& + \sum_{j\neq i} \delta_{ij}^{(r)}\delta_{ij}^{(s)} \left(\frac{(n^2-2n+3)\sigma^4}{n^2}\right) + \sum_{l\neq i\neq j\neq k\neq l} \delta_{ij}^{(r)}\delta_{lk}^{(s)} \left(\frac{3\sigma^4}{n^2}\right).
\end{aligned}
$$

Here the S_1 and S_7 scenarios equal zero, as i's and j's have been repeated in both δ's and we suppose $\delta^{(r)}$ and $\delta^{(s)}$, in different directions, means the product of neighbouring systems for given i and j can not be 1. For instance, when $\delta_{ij}^{(r)} = 1$, the $\delta_{ij}^{(s)}$ is by definition equal to zero. The S_2, S_3, S_5 and S_6 cases are also identical, and similarly S_4

and S_8. All these results were checked numerically using examples.

Hence $E(I_r I_s)$ can be calculated as

$$E(I_r I_s) = \frac{1}{2A_r A_s (n-1)(n+1)} \left[\left(4 \times (3-n) \sum_{i \neq j \neq k} \delta_{ij}^{(r)} \delta_{kj}^{(s)} \right) + 6 \sum_{i \neq j \neq k \neq l} \delta_{ij}^{(r)} \delta_{kl}^{(s)} \right]. \tag{3.14}$$

The formula of covariance between two dependent directional I statistics is

$$Cov(I_r, I_s) = \frac{1}{h} \left[\left(4 \times (3-n) \sum_{i \neq j \neq k} \delta_{ij}^{(r)} \delta_{kj}^{(s)} \right) + 6 \sum_{i \neq j \neq k \neq l} \delta_{ij}^{(r)} \delta_{kl}^{(s)} \right] - (n-1)^{-2}, \tag{3.15}$$

where $r \neq s \in \{1, 2, 3\}$, $h = 2A_r A_s (n-1)(n+1)$ and n is the number of spots in the image. We can now substitute $Cov(I_r, I_s)$ in Equation (3.10) to obtain the variance of $I_r - I_s$, considering the dependency between directions, as follows:

$$\begin{aligned} V(I_r - I_s) = V(I_r) + V(I_s) - \frac{2}{h} & \left[\left(4 \times (3-n) \sum_{i \neq j \neq k} \delta_{ij}^{(r)} \delta_{kj}^{(s)} \right) + 6 \sum_{i \neq j \neq k \neq l} \delta_{ij}^{(r)} \delta_{kl}^{(s)} \right] \\ & + 2(n-1)^{-2}, \end{aligned} \tag{3.16}$$

where the variance of a particular direction is defined in Equation (3.2). As a result, $I_r - I_s$ for pairs of directions has distribution $N(0, Var(I_r - I_s))$. This test approximately follows a standard normal under $H_0 : I_r - I_s = 0$. A two-sided p-value can be found for each z-test. For instance, suppose $I_1 - I_2$ is the observed value, then z_1 is $\frac{I_1 - I_2}{\sqrt{V(I_1 - I_2)}}$ and p-value$= 2p(Z < -|z_1|)$, with $Z \sim N(0, 1)$ which is to be compared to $\alpha = 0.05/3 = 0.016$, using a Bonferroni correction for multiple testing (McDonald, 2009).

The z-tests in this section can be applied easily, but each image will have three p-values for the three pairs of directions. However, It is more appropriate to have a general statistical test with one p-value to distinguish images which have a preferred direction. This test is explained and obtained in the next section.

3.3.2 Bivariate normal test

In section 3.3.1, the three z-tests, which used pairs of directional I statistics (I_r and I_s), and the normal approximation under the null hypothesis $I_r - I_s = 0$, where $r \neq s$, can be used to detect if the image has a preferred direction. However, each image has three tests

with three corresponding p-values, and thus it is better to have a single test investigate the preferred direction. A generalisation of the one-dimensional normal distribution, to higher dimensions, is explained.

The bivariate normal distribution is often used to model pairs of dependent normal variables, where each is a linear combination of the other. The new bivariate statistic of directional I is defined with its hypothesis and mean and variance below and how to obtain the required p-value is also explained.

The bivariate case, involving 2 random dependent variables can be one of

$$\begin{aligned} \boldsymbol{I} &= (I_1 - I_2, I_1 - I_3), \\ \boldsymbol{I} &= (I_1 - I_2, I_2 - I_3), \\ \text{or } \boldsymbol{I} &= (I_1 - I_3, I_2 - I_3), \end{aligned} \tag{3.17}$$

and follows a bivariate normal distribution with 2-dimensional mean, $\boldsymbol{\mu}$, and 2×2 variance-covariance matrix, $\boldsymbol{\Sigma}$. All bivariate cases in Equation (3.17) contain the same information which aims to investigate if any direction is different. Now let us represent this in matrix algebra notation, for example, suppose we have $\boldsymbol{I} = (I_1 - I_2, I_1 - I_3)$ as a random vector, the shorthand notation we use is

$$\boldsymbol{I} \sim MVN\left(\boldsymbol{\mu} = \begin{pmatrix} E(I_1 - I_2) \\ E(I_1 - I_3) \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} V(I_1 - I_2) & Cov(I_1 - I_2, I_1 - I_3) \\ Cov(I_1 - I_3, I_1 - I_2) & Var(I_1 - I_3) \end{pmatrix}\right),$$

where

$$\begin{aligned} Cov(I_1 - I_2, I_1 - I_3) &= E(I_1^2 - I_1 I_3 - I_2 I_1 + I_2 I_3) - E(I_1 - I_2)E(I_1 - I_3) \\ &= E(I_1^2) - E(I_1 I_3) - E(I_2 I_1) + E(I_2 I_3), \end{aligned}$$

$V(I_r - I_s), r \neq s$, can be calculated from Equation (3.16), the expectations of I^2 can be defined from Equation (3.2) and $E(I_r I_s), r \neq s$ is defined in Equation (3.14). Furthermore, we have already shown in Section 3.3.1 under H_0 that $E(I_1 - I_2) = E(I_1 - I_3) = 0$. So we have $\boldsymbol{I} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma})$ under the null hypothesis $H_0 : \boldsymbol{I} = \mathbf{0}$, which means there is no preferred direction in the image.

To test the null hypothesis, Chatfield and Collins (1980) explained that if we have a one-sample multivariate test with known $\boldsymbol{\Sigma}$, the appropriate test is the likelihood ratio

test. The test statistic (T) is written as

$$T = \boldsymbol{I}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{I}, \tag{3.18}$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse matrix of $\boldsymbol{\Sigma}$. This test statistic converges in distribution to a central chi-squared distribution, as it is a quadratic form, with two degrees of freedom as there are two independent elements in $\boldsymbol{I}$. Based on the chi-square approximation, a p-value is computed. If this p-value is significant (< 0.05), then the significant directions can be specified by the three p-values of the z-test from the normal approximation for each direction as described in Section 3.3.1. The statistical test in Equation (3.18) can be computed using any of the two-dimensional random vectors in (3.17) because the results of the test for all vectors were identical. Therefore, any of these vectors can be used to investigate if the image has a preferred direction.

3.3.3 A z-test for anisotropy in the direction of the lumen

Pathologists are interested in detecting the clustering of spots in a particular direction. More importantly, they would like to find if the direction parallel to the lumen surface differs from the other directions. The direction of the lumen is only available for the gastric cancer dataset described in Section 1.3.1. In this section a new statistical test is established to find if there is a significant difference between the autocorrelation of spots in the lumen direction verses the other two hexagonal axes. To do this test, all images are first rotated so that the lumen surface is at the top of the image. The rotation procedure is explained and then the directional I statistics are calculated. The new statistical hypothesis testing problem is then defined with the null hypothesis that the autocorrelation of spots towards the lumen surface is equal to the other two directions. The hypothesised sample mean and covariance matrix are defined, because a pair of I_r and I_s, $r \neq s$, are not independent, in addition to defining the p-value.

The location of the lumen surface on images is defined as a "clock system" indicator. Suppose c is the subjective indicator variable of the lumen direction which takes the value $1, 2, 3, \ldots, 12$ where $c = 12$ indicates the direction of the lumen with $c = 12$ do not need rotation. To rotate the image toward the lumen, the image is firstly moved to the centre of the coordinate system, and then adjusted by rotation. Suppose we have an

image and the direction of lumen, c, then the image will be anticlockwise rotated by an angle, say φ_c, in radians about the origin to be directed towards the lumen. The angle of rotation is defined as a function of the clock system

$$\varphi_c = \frac{c\,\pi}{6}, \quad c = 1, \ldots, 12.$$

Here the direction of the lumen is $c = 12$, all possible values of lumen direction are illustrated in Figure 3.3 with angles in degrees. For given φ, a rotation matrix is used to perform a rotation in Euclidean space which is given by

$$R(\varphi) = \begin{bmatrix} cos(\varphi) & sin(\varphi) \\ -sin(\varphi) & cos(\varphi) \end{bmatrix}.$$

This matrix is used to rotate spots in the two-dimensional coordinate system anti-clockwise through an angle φ about the origin to give new coordinates for the rotated image. Suppose we have the coordinates of the spots in a $n \times 2$ matrix Y. A rotated matrix, Y', is obtained by using the matrix multiplication $YR(\varphi)$. This type of rotation, called rigid transformation, does not alter the size or the shape of any object.

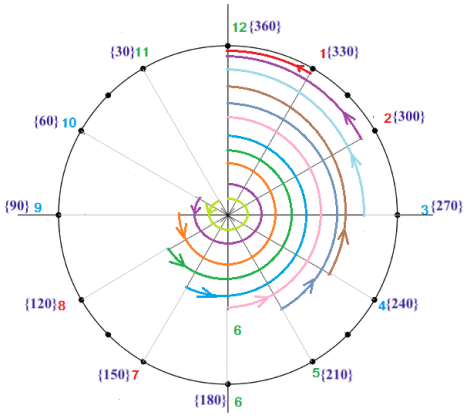


Figure 3.3: Counter-clockwise rotation for each possible clock values, c, to be toward the lumen surface (12 o'clock).

After rotation, the three directional I statistics can be computed as described in Section 3.2 where I_1 presents the direction of the lumen. The direction of the lumen is sometimes not lined up exactly on one of three main hexagon axes. Therefore, the divi-

sion of angles need to be generalised. The new set of directional I statistic is shown in Table 3.1 and Figure 3.4.

Table 3.1: The classification of angles after rotation with I_1 determining the direction of lumen.

Directional I	The range of directional I from the angle matrix ζ
I_1	$[330°, 30°)$ OR$[150°, 210°)$
I_2	$[30°, 90°)$ OR $[210°, 270°)$
I_3	$[90°, 150°)$ OR $[270°, 330°)$

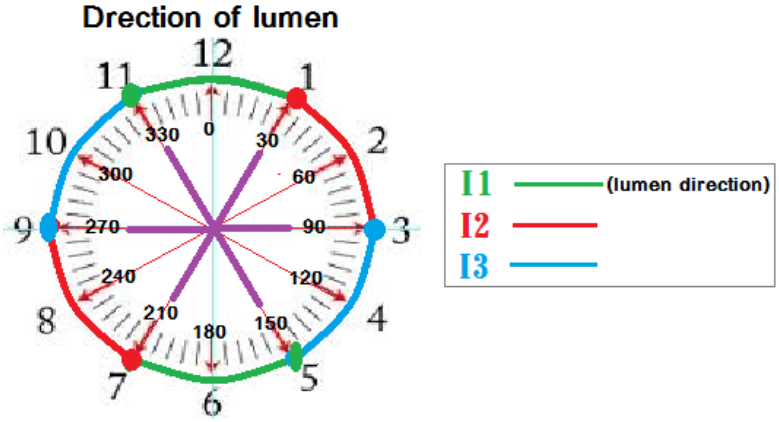


Figure 3.4: The three different directions of I after rotation and the classification of angles toward I_1 where I_1 indicates the direction of the lumen.

For example, Figure 3.5 shows all 12 possible rotations of 30 spots in the hexagonal grid. The green line represents the direction of the lumen (approximately 12 o'clock). Sometimes the direction of the lumen does not line up exactly toward 12 and it can be approximately between 11 and 1 o'clock.

The appropriate statistical test here is the z-test with null hypothesis that there is no significant difference in the autocorrelation of spots in the lumen direction verses the other directions, that is I_1 verses I_2 and I_3. The z-test is computed using the following formula

$$z = \frac{(2I_1 - I_2 - I_3) - 0}{sd(2I_1 - I_2 - I_3)} \sim \mathrm{N}(0, 1), \tag{3.19}$$

here $E(2I_1 - I_2 - I_3) = 0$ as the expectations of all directional I statistics are identical. The variance term can be obtained by considering the correlation between directional I

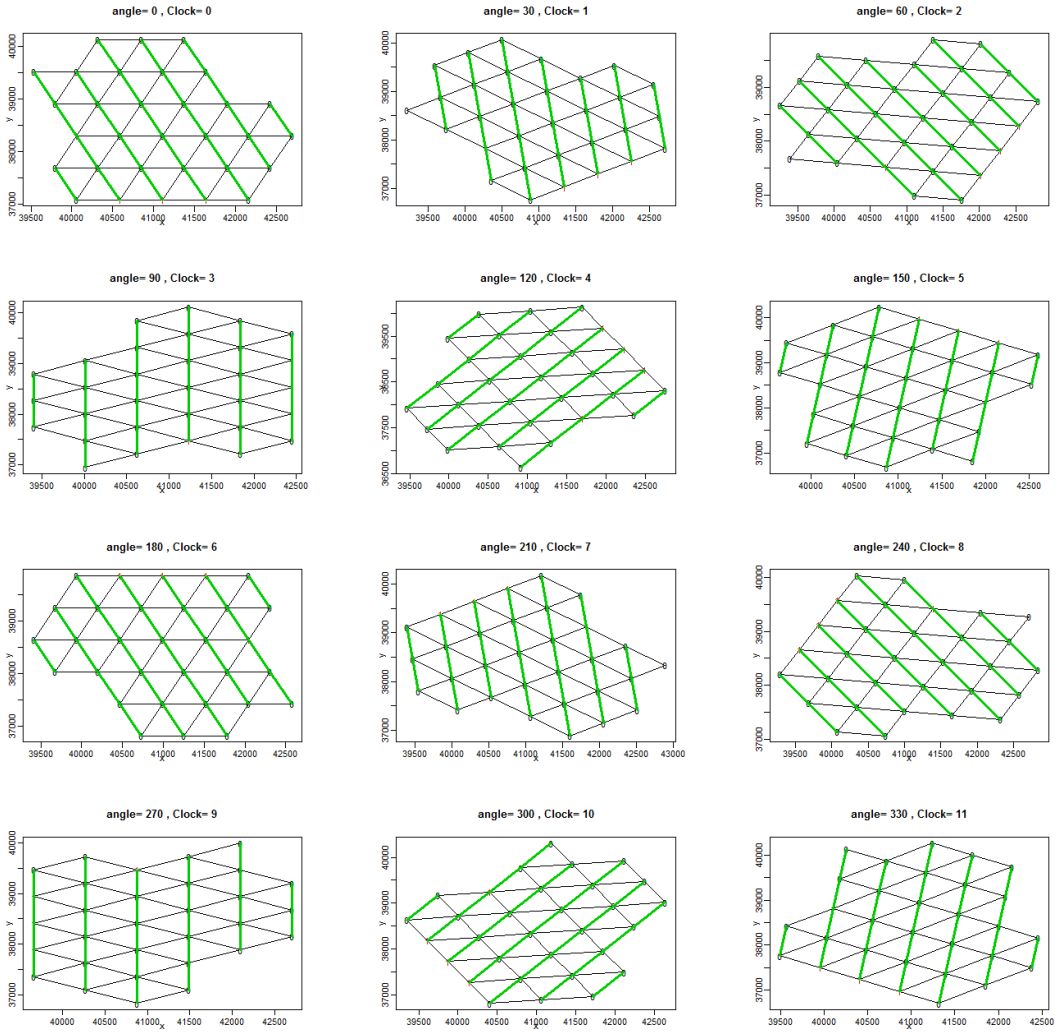


Figure 3.5: The 12 possible clock rotations for 30 spots where the green lines display the direction of the lumen (I_1).

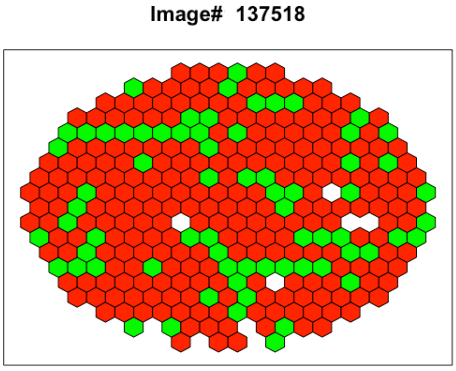
statistics as before

$$V(2I_1 - I_2 - I_3) = 4V(I_1) + V(I_2) + V(I_3) - 4Cov(I_1, I_2) \\ - 4Cov(I_1, I_3) + 2Cov(I_2, I_3), \tag{3.20}$$

where $Cov(I_r, I_s)$, $r \neq s$, was derived in Equation (3.15). Under the normal approximation, the p-value can then be calculated.

3.3.4 Applications

A random selection of images, with direction of lumen equal to 12, are used to illustrate all statistical tests defined in previous sections, and then the main statistical test described in Section 3.3.3 is preformed for all random images.

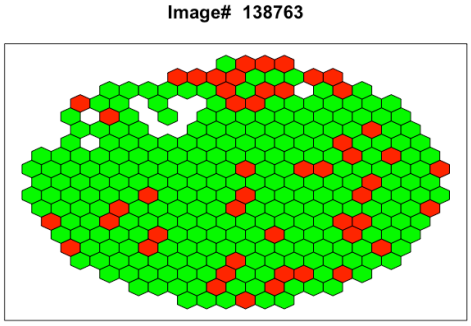


	Observed value(s)	$E(.)$	$V(.)$	Test statistic	p-value
The z-test for non- directional I					
I	0.0452	-0.0033	0.0012	1.4170	0.1565
The z-test for directional I					
I_1	-0.0812	-0.0033	0.0035	-1.3079	0.1909
I_2	0.0000	-0.0033	0.0036	0.0543	0.9567
I_3	0.2159	-0.0033	0.0035	3.6862	0.0002
The z-test for differences between directional I					
$I_1 - I_2$	-0.0811	0.0000	0.0071	-0.9611	0.3365
$I_1 - I_3$	-0.2971	0.0000	0.0071	-3.5250	0.0004
$I_2 - I_3$	-0.2159	0.0000	0.0071	-2.5598	0.0105
Bivariate normal test for detecting anisotropy					
$\begin{pmatrix} I_1 - I_2 \\ I_1 - I_3 \end{pmatrix}$	$\begin{pmatrix} -0.0811 \\ -0.2971 \end{pmatrix}$	$\begin{pmatrix} 0.0000 \\ 0.0000 \end{pmatrix}$	$\begin{pmatrix} 0.0071 & 0.0036 \\ 0.0036 & 0.0071 \end{pmatrix}$	13.2818	0.0013
The z-test for detecting if anisotropy in lumen direction is differ					
$2I_1 - I_2 - I_3$	-0.3782	0.0000	0.0213	-2.5888	0.0096

Figure 3.6 & Table 3.2: Different statistical tests for a single image (# 137518), where $E(.)$ and $V(.)$ are the mean and variance used in the test.

Consider two chosen images as shown in Figures 3.6 and 3.7. We start by obtaining the directional δ's matrices, to be able to calculate the directional I statistics. All statistical tests are then performed with results in Tables 3.2 and 3.3 respectively. These tables also include the non-directional I statistics with their p-values.

In Table 3.2, only the spots in the third direction are autocorrelated (p-value = 0.0002). For the same image the z-tests were applied for the difference between pairs of directions



	Observed value(s)	$E(.)$	$V(.)$	Test statistic	p-value
The z-test for non-directional I					
I	0.0525	-0.0033	0.0012	1.6004	0.1095
The z-test for directional I					
I_1	-0.0491	-0.0033	0.0037	-0.7541	0.4508
I_2	0.1142	-0.0033	0.0037	1.9200	0.0549
I_3	0.0936	-0.0033	0.0037	1.5979	0.1101
The z-test for differences between Directional I					
$I_1 - I_2$	-0.1633	0.0000	0.0075	-1.8918	0.0585
$I_1 - I_3$	-0.1427	0.0000	0.0074	-1.6607	0.0968
$I_2 - I_3$	0.0206	0.0000	0.0075	0.2390	0.8111
Bivariate normal test for detecting anisotropy					
$\begin{pmatrix} I_1 - I_2 \\ I_1 - I_3 \end{pmatrix}$	$\begin{pmatrix} -0.1633 \\ -0.1427 \end{pmatrix}$	$\begin{pmatrix} 0.0000 \\ 0.0000 \end{pmatrix}$	$\begin{pmatrix} 0.0075 & 0.0037 \\ 0.0037 & 0.0074 \end{pmatrix}$	4.2665	0.1184
The z-test for detecting if anisotropy in lumen direction is differ					
$2I_1 - I_2 - I_3$	-0.3060	0.0000	0.0222	-2.0530	0.0401

Figure 3.7 & Table 3.3: Different statistical tests for a single image (# 138763), where $E(.)$ and $V(.)$ are the mean and variance used in the test.

$I_r - I_s,\ r, s = 1, 2, 3;\ r \neq s$ and the p-values calculated. The p-value$(I_1 - I_2) = 0.3365$, p-value$(I_1 - I_3) = 0.0004$ and p-value$(I_2 - I_3) = 0.0105$. At an $\alpha = 0.0167$, there are significant differences between I_1 and I_2 as well as I_2 and I_3 meaning that image# 137518 has preferred direction toward I_3 which is also shown in previous test that p-value of I_3 has high autocorrelation. This significant direction is also shown on image# 137518 where the the green spots tend to be autocorrelated. The bivariate test for $(I_1 - I_2, I_1 - I_3)$ was also applied for image# 137518. The single p-value of this test shows that there is a preferred direction in the image. The final test for detecting if the spot clustering in the lumen direction differs from the other two directions, the p-value of this test is small indicating that the anisotropy in I_1 direction differs from the combination of I_2 and I_3.

The same tests have been performed on image# 138763 figure with results in Table 3.3. None of the tests, however, show a significant result except the final test with p-value equal to 0.0401. This means that we have enough evidence to reject the null hypothesis and accept the alternative which claims that the property of being anisotropic in the lumen direction differs from the other two directions. This difference was not detected by the other statistical tests.

The main and final test statistic is also applied for all independent gastric cancer images in Table 2.6 (59 images) after they have been rotated toward the lumen surface. Only 7% of images (4 images), which are shown in Figure 3.8, have a significant p-value (< 0.05) so we reject the null hypothesis $2I_1 - I_2 - I_3 = 0$ in these cases. A half of those patients had chemotherapy and have Lauren classification equals two ($LS = 2$), but with various Japanese classification (JS). More interesting, however, is that the pathological tumor stage (pT) for all these patients is 5 which is the highest stage of tumor. These patients are survived approximately between one to three years and their status recored as died. This may indicate that the tumor is more deep into nearby tissue and more aggressive. However, this result has not been checked clinically as well as not enough significant directional images to check if directional image can be aggressive. In fact, if H_0 was true for all patients, we would expect 5% of the images to be rejected. Hence the rejection of H_0 for 4 images is consistent with $\alpha = 0.05$ level of significance.

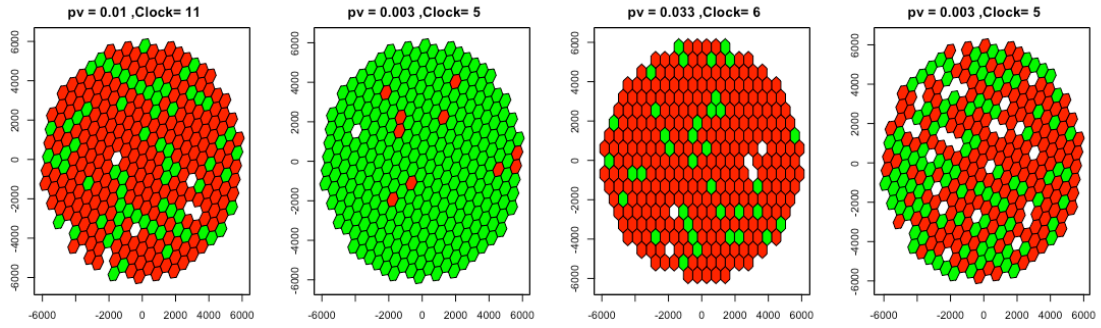


Figure 3.8: Four images rotated toward the lumen surface (12 o'clock) and p-value of statistical test for detecting if anisotropy in lumen direction is different to the other two directions.

3.4 The power of the test for detecting anisotropy toward lumen

The statistical power of the non-directional I test was investigated in Section 2.5 with a parameter κ which controlled autocorrelation in the image. The power of the directional version for lumen direction detection is now defined with an additional parameter in the covariance matrix, say ψ, which controls the autocorrelation in a specific direction. This matrix measures the joint variability of two random spots which are close and the position of spots is in the same direction. The first goal is to define an appropriate parametrisation of the covariance matrix that can be used to simulate directional images from the alternative hypothesis ($2I_1 - I_2 - I_3 \neq 0$). Then we will be able to find the power of test, which is the proportion of rejections of H_0 when it is false.

The process of simulating autocorrelated images from a multivariate normal was explained in Section 2.5. The same approch is used in this section but we need to adjust the definition of covariance matrix ($\mathbf{\Sigma}_{n\times n}$, where $\sigma_{ij} \neq 0, i \neq j$) which previously only took distance, not direction, into account. The extra parameter (ψ) is added to the definition of this matrix to also give spots that are allocated in the same direction more weight. To do this, suppose we have a preferred direction, say m, and an angle matrix using formula (3.4), where $B_{ij} \in (-\pi, \pi]$. This angle matrix is subtracted from the m to give $\theta_{ij} = B_{ij} - m$ which is a new angle matrix relative to the direction of interest m. For given κ and ψ, the variance-covariance matrix is defined as

$$\sigma_{ij} = e^{-\kappa D_{ij} \times \psi(1.2 - \cos(2\theta_{ij}))}, \tag{3.21}$$

where D_{ij} is defined in Equation (2.1), $\cos(2\theta_{ij}) \in (-1, 1]$ and we take double the angle, as each angle has a symmetric angle at the opposite side. Also we set 1.2 because the maximum value of $\cos(2\theta_{ij})$ is one, and we add an extra decimal place, say 0.2. This small ratio avoids $\cos(2\theta_{ij})$ to be zero when it is equal 1, otherwise the power of exponential function equals zero. In parametrisation, κ and ψ control the correlation in the image overall and in a direction m direction respectively.

For a given hexagonal grid using a real image, the positive angle of the favoured direction m, when it is lined up exactly on the hexagonal axes, can be either 1.9794

(direction of I_1), 1.1622 (direction of I_2) or 3.1416 (direction of I_3). These angles can be used to simulate images with high autocorrelation in different directions. Under an independent configuration, for instance, we select κ to be 0.1 from Table 2.5 and set ψ as 0.3 in the covariance definition function to simulate directional autocorrelated images. Figure 3.9 displays examples of directional images, containing 300 spots, with various m angles.

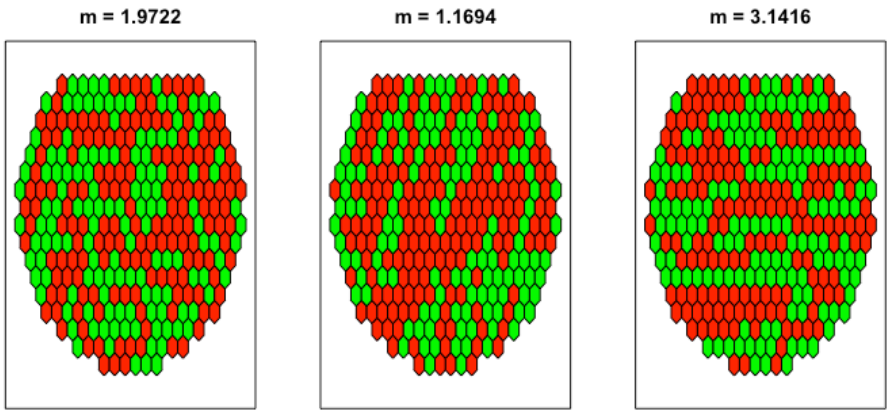


Figure 3.9: Simulated images with various directional autocorrelated using $\kappa = 0.01$ and $\psi = 0.3$ in the covariance matrix.

Now, a power function is calculated where the angles of directions I_2 and I_3 are interchangeable in power evaluation, but the main direction is I_1, where we expect that the statistical test has more power. Thus, the angle of I_1 is mainly considered in addition to either the angle of I_2 or I_3. Let us consider the angles of I_1 and I_3, however, the angles in between these two needed to be evaluated. Hence, an interval of angles, which are allocated between the angles of I_2 or I_3, are considered. This interval of angles is approximately symmetric in evaluation about its midpoint. Even though some angles can not be lined up exactly on hexagonal axes, they are essential to evaluate the power of the test with the high directionality of spots is between axes of the hexagon.

To determine the angles, the main angle is for I_1, which is $m = 1.9794$, is considered. Also we consider the angle of I_3, which is $m = 3.1416$ equivalent to $m = 0$. Now, we select some angles far away from the main hexagonal axes, e.g. $0.2, 0.3, 0.5$ and 0.6, where $m = 0.6$ illustrates a midpoint between the angles of I_2 and I_3. To compute the power, for each angle, we simulate 500 images of 300 spots under H_1 that is from a multivariate normal with zero mean vector and covariance matrix ($\mathbf{\Sigma}_{n\times n}$) which is defined

from Equation (3.21) with $\kappa = 0.1$ and various value of ψ. The κ value is determined from Section 2.5 to have independent images. Then, the test statistic in Section 3.3.3 is calculated for each simulated image to find the probability of rejecting the null hypothesis when it is false, $1 - \beta$, where β represent the probability of type II error. Figure 3.10 shows the power function for different angles. We can see that the test is most powerful when the maximum autocorrelation direction is lined up exactly with the direction of I_1. The statistical test is also still powerful when the highest autocorrelation is lined up with direction of both I_1 and I_3. This test, however, has very low power if the maximum autocorrelation lies in between hexagonal axes.

Basically, the pathological technique which generates a systematic grid of spot locations by RandomSpot system is a completely random process. Also, the locations of spots are in a continuous space and thus the direction of the lumen direction is arbitrary. That means the direction of maximum autocorrelation can occur between two axes of the hexagon and it may be difficult to detect. If the direction of lumen is precisely lined up with one of the axes of the hexagon grid it may increase the chance of detecting the maximum autocorrelation.

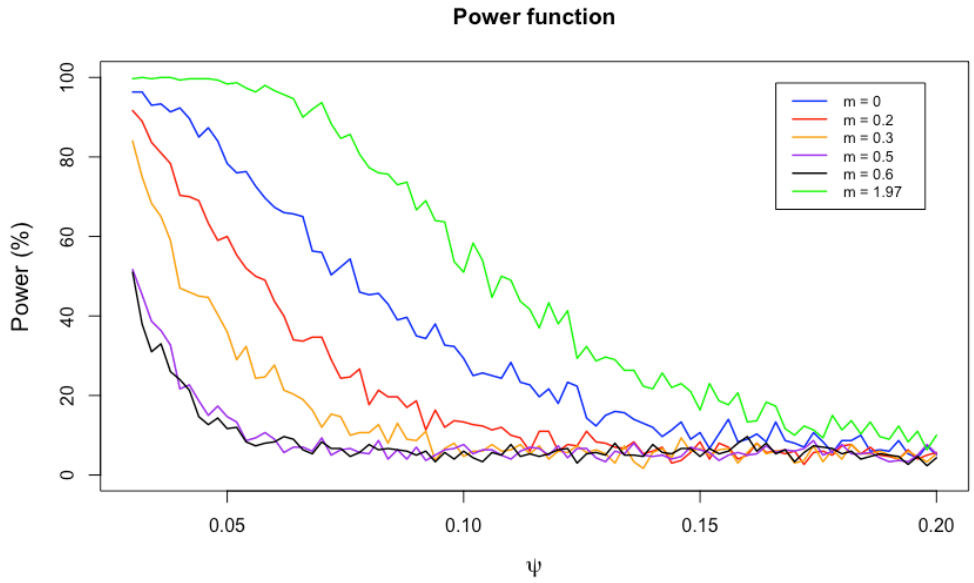


Figure 3.10: Estimated power function from 500 simulated directional images with $\kappa = 0.1$ and different ψ using various preferred direction m, where $m = 1.96$ shows the angle of directional I_1, and $m = 0$ represents the angle of directional I_3.

3.5 Discussion

Detecting anisotropy in biomedical images helps us to understand the behaviour of cancer movement, especially if there is a difference between the direction to the lumen surface and other directions. Many statistical tests have been performed, but the most interesting tests are the bivariate test in Section 3.3.2 and the z-test in Section 3.3.3. The first test can investigate if the image has a preferred direction in general, but, the second test is more specific for detecting if the autocorrelation parallel to the lumen direction, differs from the other directions. There were two sources of dependency in the process of detecting direction tests: the first one is between pairs of directional I statistics which was already covered by considering covariance between them. The second source is the dependency between spots which has been limited under the assumption of independent images. The two tests of detecting direction (Sections 3.3.2 and 3.3.3) are not the same as the region of rejections differ. However, the test of detecting autocorrelation parallel to the lumen direction is more accurate and powerful which had proved in Section 3.4.

The statistical test for detecting if anisotropy in the lumen direction equals that in the other directions can accurately identify the preferred direction under the assumption of independent images. In fact, no information is provided regarding to the reflection of images. When the reflection occurs, however, I_1 still has the same meaning for all images, but the I_2 and I_3 are swapped. We are indeed detecting if I_1 differs from a combination of I_2 and I_3, thus the statistical test still works effectively. When the angles are divided into three groups after rotation, the sides of the hexagons may not be perfectly lined up with the clock (just approximately), therefore the statistical test for investigating the direction in the lumen direction has less power.

By simulating clustered images, we checked the conservatively of the non-directional I test statistic in Chapter 2. As a result, the p-value of the non-directional I test statistic is conservative as it is affected by standard deviation when we have a simulated cluster image. That means the true probability of incorrectly rejecting the null hypothesis is never greater than the nominal level for a given significance level. The ratio of standard deviation for high autocorrelated images tends to be 50% smaller than the independent one. This result has been confirmed by calculating 100 I statistics from 100 random and clustered images, and then the sample standard deviation of both groups of I statistics

are found.

Figure 3.11: Example of a whole tumor and a subsample of a single region.

Based in our findings, it is recommended to pathologists that it is better to cover the whole tumor instead of considering a subsample to be able predict if the image has a preferred direction. This may help to predict the tumur spread direction and next affected target area in the body. Another reason is that the directional statistic I_1 is not statistically different to I_2 and I_3 in most images. This non-significant result may be expected in the area close to the lumen surface which is more likely to be homogeneous. An example is shown in Figure 3.11 for a whole tumor and a subsampled single region. Moreover, a square grid may be better and easier in detecting directionality as each direction is orthogonal to each other. To keep the power of the statistical test for any arbitrary rotation, we can either adjust the indicator of lumen direction to be six directions only, or when the pathologists generate the images, they need to make sure that the direction of lumen is lined up exactly with one of the three axes.

Chapter 4

Parameter Estimation in *BMRF* Models Using an Iterative Method

4.1 Overview

In Chapter 3, Moran's I statistic is modified to allow for calculation in various directions on a hexagonal grid in order to detect isotropy. Further, statistical tests were derived which considered dependency between directions. However, the statistical inference from these tests is only valid under the assumption that spots are independent; otherwise the distribution of I is unknown. Note that both the I and the directional I statistics are based on non-parametric summaries of data and are not parameters in any model. Now, in contrast, we consider a model based approach; where we can carry out inference regarding the parameters of a model rather than non-parametric summaries of data.

Further, modelling of biomedical images may help describe the spatial relationships of spots which may help in clinical problem solving and suggesting treatment plans. Other advantages of modelling is that it can help in future prediction for dependent biomedical image structure and for simulating patterns that occur in reality. In spite of these possible advantages, the modelling of biomedical images has not been considered previously.

Most spatial models are complex containing parameters that refer to spatial features and which can not be measured or quantified directly. The basic idea in this chapter is to model the connections between spots as parameters, where if there is dependence

between spot colour, then there are non-zero parameters. One of the most widely used mathematical models on regular grids of binary variables is the Binary Markov Random Field (*BMRF*) (Besag, 1975). The Ising model is an example of a *BMRF* which is used in statistical physics for modelling the behaviour of magnets in a square lattice.

The *BMRF* image models can also, more importantly, be a motivation for investigating a parameter estimation approach. The objective of this chapter is to propose a new method, called the iterative method (*IM*), for estimating parameters. This method allows maximisation of any given probability function $p(\boldsymbol{x};\boldsymbol{\theta})$ where parameter estimation is based on the data only and avoids explicit computation of the likelihood function. We do this by comparing randomly generated data with the observed data and iterating over parameter choices so that the parameter values become better over time. In this chapter, for computational convenience and stabilisation of *IM*, the spot labels take the values -1 and 1 rather than 0 and 1.

This chapter starts by presenting background on *BMRF* models and existing approaches to estimating its parameters in Section 4.2. It also includes the practical problem of maximum likelihood estimation for more than one parameter and a pseudo-likelihood method for estimating *BMRF* parameters is described in detail. A motivating example of *IM* is presented in Section 4.3 and a general description is given in Section 4.4. The mathematical idea behind the Markov chain Monte Carlo (*MCMC*) method will be presented in Section 4.5, which is our "simulator box", including the assessment of its convergence. The output of *MCMC* is modified to consider and test replicates of design points in Section 4.6. The components of *IM* are explained in Section 4.7. This section includes the process of adding and removing design points and determining the stopping criteria. A general description of *IM* for any parameter setting is then illustrated. Section 4.9 presents statistical inference for estimated parameters for detecting clustering, along with a hypothesis testing approach and examples. The inference is also generalised for detecting directionality. The next section includes the *IM* evaluation for different parameter settings and comparison with existing methods of parameter estimation. Finally some discussion is given in Section 4.11.

4.2 Background to the Binary Markov Random Field

Besag (1974) provided a general formulation for MRF models for pattern recognition based on the exponential family class. A general proof for the construction of these joint distributions is provided by Kaiser and Cressie (2000). Even though other modelling approaches are available, such as, fractal image models (Dubes and Jain, 1989) and grey-level variation models based on variograms (Matheron, 1963), the BMRF model seems best suited for our purpose and we adopt this model also for a hexagonal grid. In this section, we present the basic definitions of the *BMRF*.

Suppose an image contains a finite number of binary spots, $\boldsymbol{x} = (x_1, \ldots, x_n)$, in a hexagonal grid, where $x_i \in \{-1, 1\}$. The neighbourhood system of spots is determined through the matrix δ, which was defined in Section 2.2, and in addition the neighbourhood system for the three directions: $\delta^{(1)}$, $\delta^{(2)}$ and $\delta^{(3)}$.

Let $p(\boldsymbol{x}; \boldsymbol{\theta})$ define the *BMRF* model, where $\boldsymbol{\theta} \in \{\theta_1, \ldots, \theta_k\}$ is a vector of clustering parameters, defined on a grid which is a collection of spots. These spots correspond to the sites of the grid, for which the probability of a given site value, conditional on the values of all other sites in the grid, is equal to the probability of the site value conditional on the values at a small subset of the grid sites (Waks et al., 1990). The Markov random field can be written as

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{\theta})} \exp\Big\{\theta_1 \sum_{i=1}^{n} x_i + \sum_{k=1}^{3} \theta_{k+1} \sum_{i \neq j} \delta_{ij}^{(k)} x_i x_j\Big\}, \tag{4.1}$$

where

$$z(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{x}} \in \Omega_{\boldsymbol{x}}} \exp\Big\{\theta_1 \sum_{i=1}^{n} \tilde{x}_i + \sum_{k=1}^{3} \theta_{k+1} \sum_{i \neq j} \delta_{ij}^{(k)} \tilde{x}_i \tilde{x}_j\Big\}, \tag{4.2}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ with $\theta_i \in \mathbb{R}$, and, $x_i \in \{-1, 1\}$ are the spot values. If the x_i's are the results of independent Bernoulli trials then the sum of the x_i's has the expectation $n(2p-1)$ and variance $4np(1-p)$. The normalising constant $z(\boldsymbol{\theta})$ is obtained by summing over $\Omega_{\boldsymbol{x}}$, which is the set of all possible configurations for $\boldsymbol{x}$.

In the non-directional version of Equation (4.1), θ_2, θ_3 and θ_4 are equal and the dependence between spots x_i and x_j is defined only by parameter θ_2. If this parameter has a positive value, neighbouring spots tend to have the same colour; the opposite happens

if θ_2 is negative. When the size of this parameter is increased, the dependency between neighbouring spots is also increased. Whereas in the directional version, when all the parameters are considered, the image is said to be anisotropic if θ_2, θ_3 and θ_4 are not equal.

The *BMRF* model, which is a flexible stochastic process, is frequently used as a prior distribution in Bayesian statistics (Cressie, 1993). In many settings, computational issues can arise when, for example, we have a complex model with many parameters, or the likelihood is unavailable, either because it is not provided as a function of the parameters or it contains an unknown normalising constant which can not be quickly evaluated. In the latter case, Mller et al. (2006) used the auxiliary variable method to eliminate the unknown normalising constant. The auxiliary variable method can consider only the data ($\boldsymbol{x}$) or parameters ($\boldsymbol{\theta}$), as one of them should be fixed. Even though the normalising constant can be estimated, the computation of the normalising constant is not feasible in a large lattice (Mller et al., 2006; Reeves and Pettitt, 2004). The normalising constant makes it challenging to evaluate the maximum likelihood estimate because of mathematical reasons making it too expensive to calculate.

Some existing methods for estimating the *BMRF* parameters are the coding method (Besag, 1974), maximum pseudo-likelihood estimation (Besag, 1975, 1977) and maximum likelihood estimation (Cressie, 1993). Maximum likelihood estimation is described in Sections 4.2.1 for one and two parameter settings and maximum pseudo-likelihood estimation is described in Section 4.2.2 for one and two parameter settings, which are used for comparison with our new method in Section 4.10.

4.2.1 Maximum likelihood estimation of *BMRF*

The *BMRF* is a complex model where the exact likelihood function evaluation is a doubly intractable problem. The complexity comes from the normalisation constant, which is a sum over an exponentially large number of possible configurations, hence usually hard to compute.

In this section, we will estimate the parameters of the *BMRF* by maximum likelihood estimation for simple cases where we have only one and two parameters. The aim here is to diagnose the difficulty of using the standard method of parameter estimation and to

suggest ideas to overcome the limitations of *BMRF* parameter estimation.

One parameter case

To start thinking about parameter estimation of the *BMRF*, a simple case is first explained where there is no interest in the neighbourhood system. This model can be defined as

$$p(\boldsymbol{x};\theta_1) = \frac{1}{z(\theta_1)}\exp\Big\{\theta_1 \sum_{i=1}^{n} x_i\Big\} \tag{4.3}$$

where

$$z(\theta_1) = \sum_{\tilde{\boldsymbol{x}}\in\Omega_{\boldsymbol{x}}} \exp\Big\{\theta_1 \sum_{i=1}^{n} \tilde{x}_i\Big\}. \tag{4.4}$$

In order to estimate the unknown θ_1, the first approach is maximum likelihood estimator (*MLE*) using the joint density function in Equation (4.3),

$$\hat{\theta}_1 = \arg\max_{\theta_1} p(\boldsymbol{x};\theta_1). \tag{4.5}$$

For $\boldsymbol{x}$, the log-likelihood is

$$\begin{aligned}\mathcal{L}(\theta_1) =& \log\left\{\frac{\exp(\theta_1 \sum_{i=1}^{n} x_i)}{z(\theta_1)}\right\} \\ =& \theta_1 \sum_{i=1}^{n} x_i - \log(z(\theta_1)).\end{aligned} \tag{4.6}$$

The maximum likelihood estimator (*MLE*) can be found by finding the derivative of $\mathcal{L}(\theta_1)$ with respect to θ_1:

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = \sum_{i=1}^{n} x_i - \frac{z^{'}(\theta_1)}{z(\theta_1)},$$

and setting to zero, giving

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = 0, \Rightarrow \sum_{i=1}^{n} x_i = \frac{z^{'}(\theta_1)}{z(\theta_1)}, \tag{4.7}$$

where

$$z^{'}(\theta_1) = \sum_{\tilde{\boldsymbol{x}}\in\Omega_{\boldsymbol{x}}} \left(\sum_{i=1}^{n} \tilde{x}_i\right) \exp\Big\{\theta_1 \sum_{i=1}^{n} \tilde{x}_i\Big\}. \tag{4.8}$$

Here $\frac{z^{'}(\theta_1)}{z(\theta_1)}$ can not be written as an explicit equation for the parameter estimate.

From Equation (4.7), there is only one summary statistic, $t_1 = \sum_{i=1}^{n} x_i$, as we have one parameter to estimate. From the right hand side of Equation (4.7), which contains the normalising constant and its derivative, we can compute precisely θ_1 for a small image, for example for $n = 15$ which takes 7 minutes. In fact if a sequence of θ_1 (eg. 10 values) is considered to find the parameter value that maximises the log-likelihood function in Equation (4.6), the total computation time is 70 minutes. An image of 300 spots takes approximately 2 hours for a single value of θ_1 and 20 hours for a sequence of values. Since $z(\theta_1)$ is definitely an expensive computation, it is infeasible to cover a sequence of all possible θ_1.

Nevertheless, the one parameter *BMRF* model is not dependent on the spatial arrangement, and so $z(\theta_1)$ can be computed exactly using a binomial expression. Let us consider an experiment of n independent Bernoulli trials, each with probability of success p. As clarification, suppose that the record values $x_1, \ldots, x_n$ have $x_i = 1$ if the i^{th} spots is black and $x_i = -1$ otherwise. The sum of the x_i's, $t_1 = \sum_{i=1}^{n} x_i$, has expectation $n(2p-1)$ and variance $4np(1-p)$. Considering $t_1 \in \{-n, -n+2, -n+4, \ldots, n\}$ denoted by $2s - n, s = 0, \ldots, n$, which is the set of all positive values taken by $\sum_i x_i$ leads to

$$z(\theta_1) = \exp\Big\{ -n\theta_1 \Big\} \sum_{s=0}^{n} \binom{n}{s} \exp\Big\{ 2\theta_1 s \Big\}. \tag{4.9}$$

Here s corresponds to a binomial distribution, and there are $\binom{n}{s}$ different ways of distributing s successes in a sequence of n trials. Thus a simplified formula of Equation (4.9) is

$$z(\theta_1) = \frac{(1 + e^{2\theta_1})^n}{e^{n\theta_1}},$$

and thus, after the derivative of the exact constant has been found, the righthand side of Equation (4.7) can be written as

$$\frac{z^{'}(\theta_1)}{z(\theta_1)} = n \left[\frac{e^{2\theta_1} - 1}{e^{2\theta_1} + 1} \right]. \tag{4.10}$$

Here $\frac{z^{'}(\theta_1)}{z(\theta_1)}$ is replaced by $\sum_{i=1}^{n} x_i$ using Equation (4.7), so we can write Equation (4.10)

as

$$\frac{t_1}{n} = \left[\frac{e^{2\theta_1} - 1}{e^{2\theta_1} + 1}\right].$$

Hence, the probability of success, denoted by $p = t_1/n$, can be formulated as a function of θ_1

$$p = \frac{e^{2\theta_1}}{e^{2\theta_1} + 1}, \tag{4.11}$$

and similarly θ_1 can be written either as a function of p

$$\hat{\theta}_1 = \frac{1}{2} \log \left\{ \frac{p}{1-p} \right\}, \tag{4.12}$$

or as a function of t_1

$$\hat{\theta}_1 = \frac{1}{2} \log \left\{ \frac{t_1 + n}{-t_1 + n}, \right\} \tag{4.13}$$

where p, which is the proportion of tumor and t_1, which is a summary statistic, are calculated from the given image. Also, n refers to the total number of spots and $\hat{\theta}_1$ is the estimated parameter. Here, in the estimation of θ_1, no spatial information is included, and hence the given image is considered to be spatially independent. Therefore, when we have a completely independent structure image, either p or t_1 can be directly calculated and θ_1 can be then estimated exactly.

Two parameter case

The two-parameter setting for the *BMRF* model reflects the spatial dependence, as the non-directional I statistic does, because the model contains δ. The *BMRF* of two parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, and the data, with joint density function, can be written as

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{\theta})} \exp\Big\{ \theta_1 \sum_{i=1}^{n} x_i + \theta_2 \sum_{i \neq j} \delta_{ij} x_i x_j \Big\}, \tag{4.14}$$

where

$$z(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{x}} \in \Omega_{\boldsymbol{x}}} \exp\Big\{ \theta_1 \sum_{i=1}^{n} \tilde{x}_i + \theta_2 \sum_{i \neq j} \delta_{ij} \tilde{x}_i \tilde{x}_j \Big\}. \tag{4.15}$$

Similarly to the one parameter case, we estimate the unknown $\boldsymbol{\theta}$ using maximum likelihood based on the joint density function $p(\boldsymbol{x};\boldsymbol{\theta})$ as:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{x};\boldsymbol{\theta}). \tag{4.16}$$

The log-likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) =& \log\left\{\frac{\exp(\theta_1\sum_{i=1}^{n}x_i + \theta_2\sum_{i\neq j}\delta_{ij}x_ix_j)}{z(\boldsymbol{\theta})}\right\}\\ =&\theta_1\sum_{i=1}^{n}x_i + \theta_2\sum_{i\neq j}\delta_{ij}x_ix_j - \log(z(\boldsymbol{\theta})).\end{aligned}$$

We obtain the maximum by first finding the partial derivatives of $\mathcal{L}(\boldsymbol{\theta})$ with respect to θ_1 and θ_2, respectively. Starting with

$$\frac{\partial\mathcal{L}}{\partial\theta_1} = \sum_{i=1}^{n}x_i - \frac{\partial z/\partial\theta_1}{z(\theta_1)}, \tag{4.17}$$

and setting to zero gives

$$\frac{\partial\mathcal{L}}{\partial\theta_1} = 0, \Rightarrow \sum_{i=1}^{n}x_i = \frac{\partial z/\partial\theta_1}{z(\boldsymbol{\theta})}, \tag{4.18}$$

where

$$\partial z/\partial\theta_1 = \sum_{\tilde{\boldsymbol{x}}\in\Omega_{\boldsymbol{x}}}\left(\sum_{i=1}^{n}\tilde{x}_i\right)\ \exp\Big\{\theta_1\sum_{i=1}^{n}\tilde{x}_i + \theta_2\sum_{i\neq j}\delta_{ij}\tilde{x}_i\tilde{x}_j\Big\}. \tag{4.19}$$

Similarly, the derivative of the log-likelihood with respect to θ_2 is

$$\frac{\partial\mathcal{L}}{\partial\theta_2} = \sum_{i\neq j}\delta_{ij}x_ix_j - \frac{\partial z/\partial\theta_2}{z(\boldsymbol{\theta})},$$

and setting to zero gives

$$\frac{\partial\mathcal{L}}{\partial\theta_2} = 0, \Rightarrow \sum_{i\neq j}\delta_{ij}x_ix_j = \frac{\partial z/\partial\theta_2}{z(\boldsymbol{\theta})}, \tag{4.20}$$

where

$$\partial z/\partial\theta_2 = \sum_{\tilde{\boldsymbol{x}}\in\Omega_{\boldsymbol{x}}} \left(\sum_{i\neq j} \delta_{ij}\tilde{x}_i\tilde{x}_j \right) \exp\Big\{ \theta_1 \sum_{i=1}^{n} \tilde{x}_i + \theta_2 \sum_{i\neq j} \delta_{ij}\tilde{x}_i\tilde{x}_j \Big\}. \tag{4.21}$$

From Equation (4.14), there are two summary statistics: $t_1 = \sum_{i=1}^{n} x_i$ and $t_2 = \sum_{i\neq j}\delta_{ij}x_ix_j$ corresponding to parameters θ_1 and θ_2 respectively. Although each summary statistic can be calculated directly for a given image, it is expensive to evaluate the normalising constant in Equations (4.18) and (4.20). The image can be thus summarised with these statistics, which related to unknown parameters. These summarisations of data motivate estimation of the parameters in the *BMRF*.

The $\hat{\theta}_2$ in Equation (4.14) contains similar information to the I statistic, for instance, $\theta_2 \neq 0$ means that the spots are not independent, as does $I \neq 0$. The inference related to θ_2 is explained in detail in Section 4.9.

4.2.2 Pseudo-likelihood equations for *BMRF* parameter estimation

As the likelihood maximisation for the *BMRF* is typically intractable, this problem may be solved using an approximate inference method. Besag (1974) described and developed an approximation approach using pseudo-likelihood (*PL*). He replaced the likelihood by a more tractable object using conditional dependencies present among a finite set of binary random variables for first-order neighbours with spots labelled 0 and 1. In this section, the conditional distribution of the *BMRF*, with two parameters, for site x_i given all other site values is derived. Followed by the maximisation of the log-likelihood to find the parameter estimates.

Recall the probability density function of the *BMRF* in Equation (4.14), here we will simplify some notation as follows

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{z(\boldsymbol{\theta})}\exp\Big\{ h(\boldsymbol{x};\boldsymbol{\theta}) \Big\}, \tag{4.22}$$

where

$$z(\boldsymbol{\theta}) = \sum_{\check{\boldsymbol{x}}\in\Omega_{\boldsymbol{x}}} \exp\Big\{ h(\check{\boldsymbol{x}};\boldsymbol{\theta}) \Big\},$$

with

$$h(\boldsymbol{x};\boldsymbol{\theta}) = \theta_1 \sum_{i=1}^{n} x_i + \theta_2 \sum_{i \neq j} \delta_{ij} x_i x_j,$$

and

$$h(\check{\boldsymbol{x}};\boldsymbol{\theta}) = \theta_1 \sum_{i=1}^{n} \check{x}_i + \theta_2 \sum_{i \neq j} \delta_{ij} \check{x}_i \check{x}_j.$$

To drive the *PL*, we begin by calculating the conditional distribution of x_r given $\check{\boldsymbol{x}}$, $p(x_r|\check{\boldsymbol{x}})$, where $\check{\boldsymbol{x}}$ includes all spots except x_r. Let $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ be two spot configurations obtained from $\check{\boldsymbol{x}}$ by setting $x_r = 1$ or $x_r = -1$ respectively. The $h(\boldsymbol{x};\boldsymbol{\theta})$ function is additive over spot-spot pairs, and hence the two configurations can be written as

$$\begin{aligned} h(\boldsymbol{x}_r^+;\boldsymbol{\theta}) &= \theta_1 + \theta_2 \sum_j \delta_{rj} x_j + h(\check{\boldsymbol{x}};\boldsymbol{\theta}), \text{ and} \\ h(\boldsymbol{x}_r^-;\boldsymbol{\theta}) &= -\theta_1 - \theta_2 \sum_j \delta_{rj} x_j + h(\check{\boldsymbol{x}};\boldsymbol{\theta}), \end{aligned} \tag{4.23}$$

where $h(\check{\boldsymbol{x}};\boldsymbol{\theta})$ involves summing over spots when $r \neq i$. The probabilities of these two configurations are

$$\begin{aligned} p(\boldsymbol{x}_r^+;\boldsymbol{\theta}) &= \frac{1}{z(\boldsymbol{\theta})} \exp\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^+;\boldsymbol{\theta})\Big\}, \text{ and} \\ p(\boldsymbol{x}_r^-;\boldsymbol{\theta}) &= \frac{1}{z(\boldsymbol{\theta})} \exp\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^-;\boldsymbol{\theta})\Big\}. \end{aligned} \tag{4.24}$$

Actually the probability of partial configuration $\check{\boldsymbol{x}}$ is just the summation of the two equations in (4.24) that is $p(\check{\boldsymbol{x}};\boldsymbol{\theta}) = p(\boldsymbol{x}_r^+;\boldsymbol{\theta}) + p(\boldsymbol{x}_r^-;\boldsymbol{\theta})$.

Now the condition distribution of of x_r given $\check{\boldsymbol{x}}$ is

$$p(x_r|\check{\boldsymbol{x}}) = \frac{p(\boldsymbol{x};\boldsymbol{\theta})}{p(\check{\boldsymbol{x}};\boldsymbol{\theta})}. \tag{4.25}$$

Here the normalising constant cancels and the partial sum is only over neighbours. The $p(x_r|\check{\boldsymbol{x}})$ contains the condition distribution of event $x_r = s$, where s can be either -1 or 1, given $\check{x}$ which is

$$p(x_r = s|\check{\boldsymbol{x}}) = \frac{\left(\frac{s+1}{2}\right) \exp\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^+;\boldsymbol{\theta})\Big\} + \left(\frac{-s+1}{2}\right) \exp\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^-;\boldsymbol{\theta})\Big\}}{\exp\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^+;\boldsymbol{\theta})\Big\} + \exp\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^-;\boldsymbol{\theta})\Big\}}. \tag{4.26}$$

The *PL* is then

$$L(\boldsymbol{\theta};\boldsymbol{x}) = \prod_{r=1}^{n}\Big(p(x_r=1|\check{\boldsymbol{x}})\Big)^{\frac{x_r+1}{2}}\Big(p(x_r=-1|\check{\boldsymbol{x}})\Big)^{\frac{-x_r+1}{2}}$$
$$= \prod_{r=1}^{n}\left(\frac{\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})\}}{\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})\}+\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})\}}\right)^{\frac{x_r+1}{2}}\left(\frac{\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})\}}{\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})\}+\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})\}}\right)^{\frac{-x_r+1}{2}}. \tag{4.27}$$

The pseudo log-likelihood is then

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{r=1}^{n}\Bigg[\frac{x_r+1}{2}\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})-\log\Big(\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})\}+\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})\}\Big)\Big\}$$
$$+\frac{-x_r+1}{2}\Big\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})-\log\Big(\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})\}+\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})\}\Big)\Big\}\Bigg]. \tag{4.28}$$

Here we can substitute $h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})$ and $h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})$ into Equations (4.23) and the simplified expression is as follows

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{r=1}^{n}\Bigg[\frac{x_r}{2}\Big(\theta_1+\theta_2\sum_j\delta_{rj}x_j\Big)+\frac{1}{2}h(\check{\boldsymbol{x}};\boldsymbol{\theta})-\frac{x_r}{2}\Big(-\theta_1-\theta_2\sum_j\delta_{rj}x_j\Big)+\frac{1}{2}h(\check{\boldsymbol{x}};\boldsymbol{\theta})$$
$$+\frac{1}{2}\log\Big(\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{+};\boldsymbol{\theta})\}+\exp\{h(\boldsymbol{x}_{\boldsymbol{r}}^{-};\boldsymbol{\theta})\}\Big)\Bigg].$$

Expanding the pseudo log likelihood we obtain:

$$\sum_{r=1}^{n}\left[\theta_1\frac{x_r}{2}+\theta_2\sum_j\delta_{rj}x_jx_r-\log\Big(\exp\{\theta_1+\theta_2\sum_j\delta_{rj}x_j\}+\exp\{-\theta_1-\theta_2\sum_j\delta_{rj}x_j\}\Big)\right].$$

The simplest formulation of $\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})$ we can have is

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \theta_1\sum_{r=1}^{n}\frac{x_r}{2}+\theta_2\sum_{j\neq r}\delta_{rj}x_jx_r-\sum_{r=1}^{n}\log\Big(\exp\{\theta_1+\theta_2\sum_j\delta_{rj}x_j\}$$
$$+\exp\{-\theta_1-\theta_2\sum_j\delta_{rj}x_j\}\Big). \tag{4.29}$$

It is not possible to write the maximum log-likelihood estimator as an explicit function of the data. Therefore, the `optim` function in `R` is used to solve the maximisation problem by suppling functions multiplied by -1 as optim is written to minimise a function. The `optim` function uses a starting value for the parameters to be optimised and outputs the

estimated parameters. Some examples of using the pseudo-likelihood estimation method are shown in Section 4.10 and the estimated parameters are then compared with those from the *IM*.

Note that the *PL* parameter estimation method has problems at the boundary of the image where we have fewer than 6 neighbours per spot. Such problems may be solved by considering the joint distribution of only internal image spots (Besag, 1974).

4.3 A motivating example of the iterative method

In this section a motivating example of a simple distribution is considered to explain briefly the method of iterative parameter estimation for one parameter. Suppose that we have $x_i \sim Bin(1, p), i = 1, \ldots, n$ which follows a binomial distribution with sample size n and unknown parameter p. Let t denote a summary statistic related to p which can be computed from the data, $t = \sum x_i$, where $x_i \in \{0, 1\}$. We already know that $\hat{p} = t/n$ is an unbiased estimator of p using maximum likelihood estimator *MLE*, but for the sake of illustration, we suppose this estimate is not available.

The general process of the iterative method *IM* to estimate a single parameter, $\hat{p}$, is as follows:

1. We create a initial grid of three values of the parameter, $p_1^* < p_2^* < p_3^*$, these values are called design points with sample size $N = 3$ which is regularly increased through the *IM*.

2. We suppose that we have a simulator box which can simulate data from the binomial distribution for a given parameter.

3. For each value of p_j^*, $j = 1, \ldots, N$, with given data size n, we simulate a realisation x^* from the simulator box and then compute a summary statistic $t^*(p_j^*)$ which can be written mathematically as follows

$$t^*(p_j^*) = \sum_{i=1}^{n} x_i^*$$

where $x_i^* \sim Bin(1, p_j^*), \;\; i = 1, \ldots, n.$

4. For the given grid of initial values and related summary statistics, we can calculate the first estimate of $\hat{p}$ by solving a simple linear regression as we assume locally a linear relationship between p_j^* and $t^*(p_j^*),\ j = 1, \ldots, N$ given by

$$t^*(p_j^*) = \beta_0 + \beta_1 p_j^* + \varepsilon_j, \quad j = 1, \ldots, N. \tag{4.30}$$

After the model is fitted, we will have the estimated model parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$). We then estimate $\hat{p}$ by prediction to obtain

$$\hat{p} = \frac{t - \hat{\beta}_0}{\hat{\beta}_1}, \tag{4.31}$$

where t is the observed summary statistic of our data.

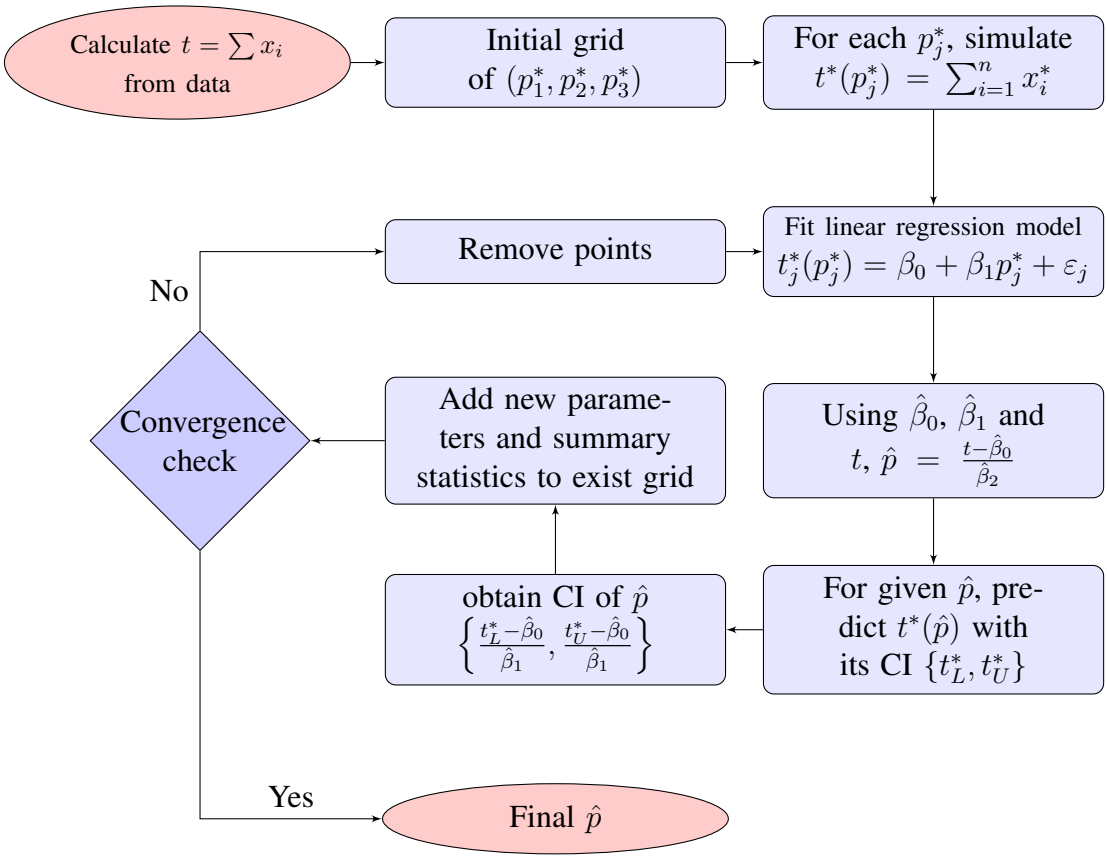


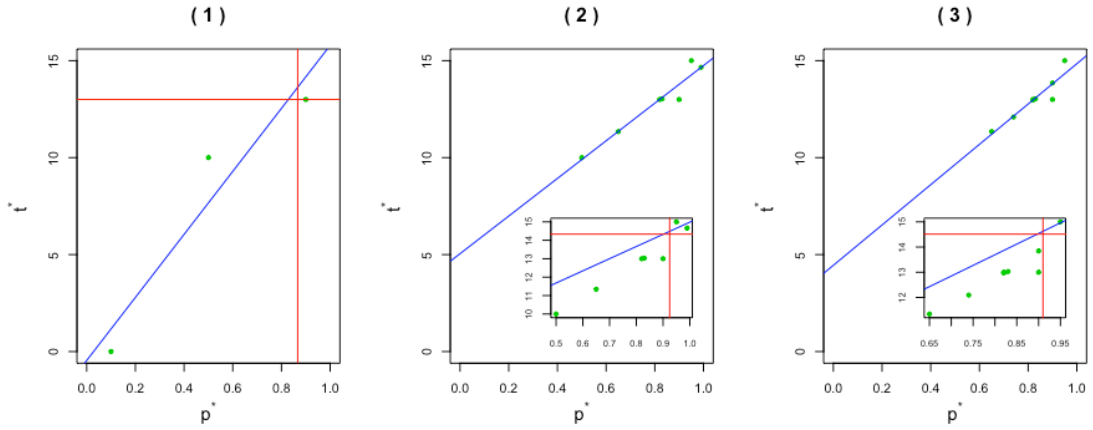
Figure 4.1: The steps of the iterative method (*IM*) for a single parameter using binomial distrubution.

5. Once we have $\hat{p}$, this is used to predict a corresponding summary statistic from Equation (4.30) and to obtain the lower and upper boundaries of a 95% confidence

interval (CI). We now have an additional set of summary statistics $(t_L^*, t^*(\hat{p}), t_U^*)$, each of which leads to additional values of p using Equation (4.31). The new points and corresponding summary statistics are added to the existing p_j^* and $t^*(p_j^*)$, $j = 1, \ldots, N$ respectively. The number of design points is now increased by 3.

6. After adding three new design points to the old ones, the point p_j^* that is furthest from the current $\hat{p}$ is removed with its corresponding $t^*(p_j^*)$.

7. Repeat the same steps, starting from 4, until the absolute difference of the current estimate of $\hat{p}$ and previous one has been minimised below a threshold of, say, 0.001. As this ratio decreases, the parameter estimate becomes more accurate.



N	$\hat{p} = \frac{t-\hat{\beta}_0}{\hat{\beta}_1}$	$\hat{p}_L$	$\hat{p}_U$	$t^*(\hat{p})$	t_L^*	t_U^*
3	0.83	0.03	0.95	13.03	0.00	15.00
4	0.82	0.65	0.99	13.00	11.34	14.66
5	0.82	0.74	0.90	12.97	12.09	13.85

Figure 4.2 & Table 4.1: Figure 4.2 shows, in order, the steps of the *IM* parameter estimation technique using $X_i \sim Bin(1, 0.8)$ and $n = 15$ by plotting design points for the whole parameter space and summary statistic space and the internal figure windows are a zoomed in version of the current local space of the $\hat{p}$ estimator. The horizontal red line shows the observed summary statistic $t = 13$, from data, the vertical red line shows $\hat{p} = 0.87$ using *MLE* and the blue line shows the fitted regression line. Each row in Table 4.1 illustrates the current parameter estimate and summary statistic value at each step of the *IM* with their CI and the number of design points N. The last row presents the last step with the final value of $\hat{p}$.

Figure 4.2 shows the steps of the iterative method for estimating a single parameter of given $X_i \sim Bin(1, p)$. The mathematical justification for adding points using a

confidence interval is to add stability to the regression fit. The danger, however, is that if this interval is too big then the convergence is more time-consuming.

To illustrate the *IM* on a real image, suppose we have the following data $X_i \sim Bin(1, 0.8)$ with $n = 15$ points and $t = 13$, which is the summary statistic from observed data. To estimate $\hat{p}$ by the *IM*, we initialise the first grid of values which is $(0.1, 0.5, 0.9)$ with $N = 3$ design points. For given these parameters and n, we simulate from the simulator box, to produce corresponding t^*, $(t^*(0.1), t^*(0.5), t^*(0.9))$. A simple linear regression is next fitted, using Equation (4.30), and the first extra design point is estimated using Equation (4.31). Through the *IM*, the number of design points N increases. The *IM* is stopped using the convergence condition that the absolute difference between the current and previous estimate of $\hat{p}$ is less than 0.001, then we lastly determine the optimal and final estimate of $\hat{p} = 0.82$.

Figure 4.2 illustrates the design points, for the same example, in the initial, middle and final steps of *IM*. Equivalently, Table 4.1 includes all steps of the *IM* where the last row of the table includes the final estimate. This table shows how the number of design points gradually increases, in addition to how the estimation of $\hat{p}$ converges. Note that the estimate of p using *MLE* equals 0.87 which is close to the estimation from *IM* ($\hat{p} = 0.82$).

4.4 General description of the iterative method

In Section 4.3, the iterative method for estimating parameters is illustrated in a simple framework with a single parameter. In this section, the *IM* is explained in general. Some differences and similarities between the *IM* and Approximate Bayesian computation *ABC* are then considered in Section 4.4.1.

The main idea of this method depends on a sequential simulation approach where we can simulate data $\boldsymbol{x}^*$ from the model $p(\boldsymbol{x}; \boldsymbol{\theta})$, which is a probability function of data $\boldsymbol{x}$ with given parameter $\boldsymbol{\theta}$. One of the main problems is that the simulation-based approach has a random output even when the simulation uses the same parameters. The optimal parameter values are unknown, and the parameter space is very large, and so the process can take a long time to run especially as it has a stochastic component. Therefore we want to get to the right area (homing-in on the parameter estimates) in the space by doing a sequential process where we move around the space in a clever way. This method is

explained in more detail for any model with a high-dimensional parameter space.

The method is initialised by setting a grid of values for each parameter that we need to estimate. Given the availability of a simulator box, which may, as here, produce a Markov chain Monte Carlo (*MCMC*) realisation, with given $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, we can generate independent and identically distributed data $\boldsymbol{x}^*$ ($\boldsymbol{x}^* \sim p(\boldsymbol{x}; \boldsymbol{\theta})$) from which we can calculate summary statistics $\boldsymbol{t}^* = (t_1^*, t_2^*, \ldots, t_k^*)$. Our aim is to update $\boldsymbol{\theta}$ sequentially, in such a way that $\boldsymbol{t}^*$ converges to the observed $\boldsymbol{t}$ from the given data.

For given summary statistics $\boldsymbol{t}_i^*, i = 1, \ldots, k$ and $\boldsymbol{\theta}$ with k parameters, we fit a local multiple linear regression model (MRM) in an adaptive local manner in which the summary statistic is modelled in terms of the parameters. The MRM, which is explained in Section 4.7.1, is a generalised version for any parameter setting.

After the model is fitted, the next design points are obtained and some design points are removed according to previous simulation realisations. The method is sequentially adaptive, whereby new design points are chosen which we think are closer to the part of the parameter space where the true estimate is located. We keep adding and removing design points sequentially homing-in on the right part of the parameter space and the simulated data becomes closer to the observed data. The process stops after convergence is achieved.

In a local region we can approximate the relationship between the parameters and the data summaries as linear. Of course, the relationship between $\boldsymbol{t}^*$ and $\boldsymbol{\theta}$ generally will not be linear, but that does not mean we will have a bad result as the design-points space is much smaller than the whole space, so the linearity should be reasonable in the local region, and this will potentially make improvement without any complication. We have two main spaces in the *IM*:

1. the space of parameters $\boldsymbol{\theta}$.

2. the space of summary statistics $\boldsymbol{t}^*$

In high dimensional space, it is challenging and complex to take into account both parameters and summary statistics spaces. Thus it is simple to compare and concentrate our approach with the parameter space as we are interested in estimating parameters and this space gives a similar view to the sample space. Here the main assumption in the parameter space is a linear relationship between parameters.

4.4.1 The *IM* and *ABC*

The *IM* shares some characteristics with Approximate Bayesian Computation (*ABC*) (Beaumont, 2010; Marin et al., 2012). Here we give some comparisons between the two methods. Both methods are likelihood-free methods, stochastic processes, less expensive for computational reasons and have the same idea of simulating data samples from the given model the value of a parameter (or parameter vector). However, the *IM* uses *MCMC* to simulate data, whereas the *MCMC* can use the *ABC* to estimate the acceptance probability without likelihoods.

For given observed data, the *IM* aims to estimate model parameters, whereas the *ABC* is used to estimate the posterior distributions of model parameters. In *ABC*, the parameter values are sampled from the prior which can be problematic if the data is very informative. This leads to the simulated data being far away from the observed data with no control over what data is simulated. The *IM*, however, starts by choosing initial values for each parameter, and new data is simulated from the given model in a sequential manner so that it will eventually resample close to the observed data.

Sometimes in *ABC* data can be close to the observed data but the determination of parameters is challenging (Beaumont, 2010). In the *IM*, design points, which are parameters, are added and old ones removed in an adaptive manner, with iterative steps until convergence. Our assumption is that the relationship between the parameters and the summary statistics from simulated data is linear locally to the current estimate.

The rejection technique in *ABC* is quite similar to the *IM* where both methods accept the close values. The *ABC* takes the nearest neighbours, whereas in the *IM* all parameters (design points) are accepted and then we remove those far away from the optimal values. Here the *IM* can accept more values than the *ABC*.

We can conclude that the *IM* can avoid the step of choosing the prior distribution, so that it is enough to have initial parameter values to start the method until we reach the convergence with optimal estimation of parameters. We can say that the *IM* is an extended and improved version of *ABC* approach.

Algorithm 3: The modified Metropolis-Hastings algorithm (*MCMC* simulator box) for generating image $\boldsymbol{x}$.

MCMC $(n, M, \boldsymbol{\theta})$;
Input: Number of spots n, number of iterations M and parameter value $\boldsymbol{\theta}$
Output: $\boldsymbol{x}$
Generate a binary $\boldsymbol{x}$;
for $j = 1$ *to* M **do**
  Set $\boldsymbol{x}^* = \boldsymbol{x}$;
  Choose random i in $(1, \ldots, n)$;
  Consider proposal $x^*[i] = -x^*[i]$;
  Calculate the acceptance ratio $q = \frac{p(\boldsymbol{x}^*;\boldsymbol{\theta})}{p(\boldsymbol{x};\boldsymbol{\theta})}$;
  if $q > 0$ **then**
    Accept proposal $\boldsymbol{x}^*$;
    $\boldsymbol{x} = \boldsymbol{x}^*$;
  else
    Generate $u \sim \text{Uniform}(0, 1)$;
    if $q > u$ **then**
      Accept proposal $\boldsymbol{x}^*$;
      $\boldsymbol{x} = \boldsymbol{x}^*$;
    else
      Reject $\boldsymbol{x}^*$;
    end
  end
end
return $\boldsymbol{x}$

4.5 *MCMC* simulator box

Our simulator box is a tool to draw independent Markov chain Monte Carlo (*MCMC*) samples from target distributions. This mechanism is an essential component of the iterative method that generates samples of size n spots depending on given parameters in order to calculate summary statistics. The Metropolis-Hastings (*M-H*) algorithm, proposed by Metropolis et al. (1953), is one of the best known of such methods for generating a sequence of random samples from a probability distribution especially when direct sampling is challenging. This algorithm allows us to indirectly sample from the *BMRF* in Equation (4.1) which is a complex distribution. This section briefly discusses how the algorithm can be used in an acceptance-rejection scheme when we have an available target distribution, and in addition assesses if the algorithm generates independently distributed data. As the Markov chain is aperiodic even for the same value of parameters,

the number of steps M for convergence is determined for the *BMRF* in Section 4.5.1 whatever the initial configuration ($\boldsymbol{x}$).

Hastings (1970) provided a more general description of the algorithm, see also Chib and Greenberg (1995), as follows: we start the *M-H* algorithm by generating a set of binary spot labels ($\boldsymbol{x}$) of length n. In the main loop of Algorithm 3, with M iterations, we pick a random location and flip the value of spot, then compute the acceptance probability q based upon the proposal distribution $p(\boldsymbol{x}^*; \boldsymbol{\theta})$ and the full joint density $p(\boldsymbol{x}; \boldsymbol{\theta})$, where the normalising constant cancels out. We accept the new candidate sample with probability q if $q > 0$. Algorithm 3 provides the detail of the *M-H* algorithm. In step 2, $\boldsymbol{x}$ can be any generated starting point, but the chosen value of M should be sufficiently large. More detail are provided in Section 4.5.1.

Now the data generated using *MCMC* is checked to see if it is independently distributed in the two parameter setting. To choose M, Ripley (1979) suggests $M = 4n$ is sufficient to ensure that samples are approximately independent. We have verified this (results not shown) for small values of θ_1 and θ_2 ($\theta_1 = -0.16, \theta_2 = 0.05$). However, for large values of $|\theta_i|$, we have found larger values of M are needed.

4.5.1 Convergence assessment

Ripley (1979) proposed that M in the simulator box should be equal to $4n$ when a spatial pattern is simulated with dependent samples, however, by experiment, after this number of steps the *MCMC* does not stabilise. In this section, the number of iterations needed is determined by simulation using the two parameter setting.

The *M-H* algorithm is a *MCMC* method for obtaining a sequence of random samples from a target distribution. Thus it should be run for a large number of iterations (M) and must be monitored for approximate convergence to its stationary distribution. This means, as the number of iterations increases, the distribution remains the same and is stable. The summary statistics ($\boldsymbol{t}^*$) are used to check convergence for given parameter values which influences the required number of steps (M).

To do the experiment, the *BMRF* is our target distribution which contains two parameters, θ_1 and θ_2, and we would like to investigate the appropriate M. If we consider the one parameter setting, this means θ_2 is assumed to be zero. In this case, no iterations

are needed, but for more than one parameter and $\theta_2 \neq 0$, the *MCMC* should be tested for positive and negative values of θ_2. In fact, when $\theta_1 = 0$ this means we do not care about if the spots are black or white, but when, for instance, θ_2 equals 0.4 this means we would like to have either black or white clustering, thus we have two possible outcomes (all black or all white).

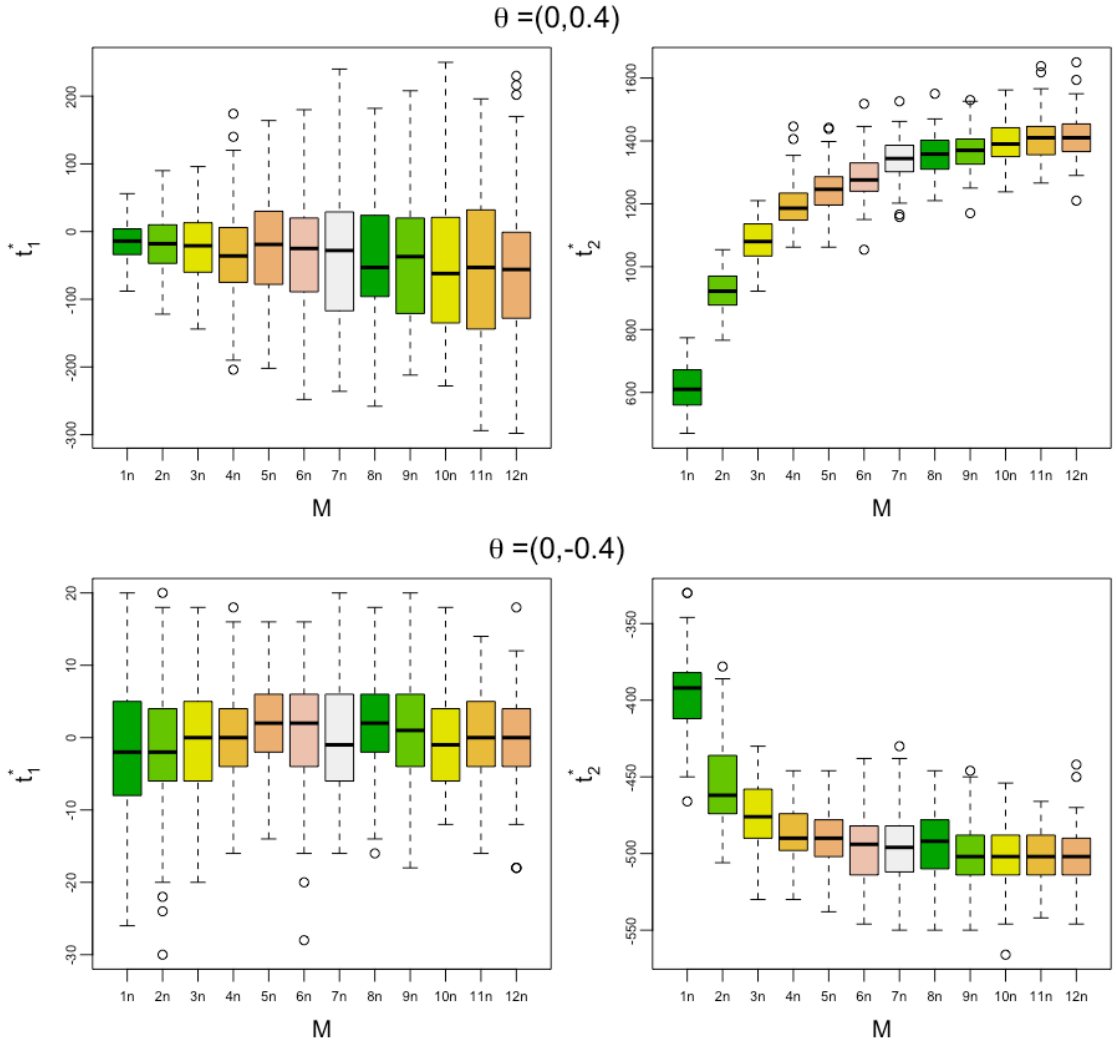


Figure 4.3: Each box-plot gives either the t_1^* or t_2^* summary statistic over various numbers of iterations M for 100 simulated images with $n = 300$ and $p = 0.5$ and for the given different parameter values $\boldsymbol{\theta} = (\theta_1, \theta_2)$

We start by considering an initial configuration by generating an image with $n = 300$ and $p = 0.5$. The first parameter value is calculated from Equation (4.12) which equals zero. We set $\theta_2 = \pm 0.4$ to see how the changing of the second parameter value effects convergence. Suppose we have various iteration numbers $M = (1n, 2n, 3n, 4n, \ldots, 10n, 11n, 12n)$. A 100 replicate MCMC runs using Algorithm 3, are then used to compute

summary statistics (t_1^* and t_2^*) from each number of iterations. These values are summarised by the box-plots in Figure 4.3 to see the stability of t_2^* across M when it is increased. In fact, θ_1 can be computed precisely for given p, thus the expected value of t_1^* equals zero in the $-1/1$ value setting. It is clear that as the number of iterations increases, t_2^* is visibly stabilised. To test the stability, we compare between pairs of t_2^* observations for various iterations to test for a difference in the mean for the last six iterations $(7n, 8n, 9n, 10n, 11n, 12n)$. The two-sample Wilcoxon non-parametric test is performed to compare the means of two independent samples of summary statistics under the null hypothesis that the mean for the summary statistic of the two iterations are equal. The non-parametric test is used because some of the t_2^* observations with some iterations do not follow a normal distribution. Figure 4.3 (top) illustrates how $\boldsymbol{t}_2^*$ converged for $\theta_2 = 0.4$ as the iteration number (M) increases. Here there were significant differences between $\boldsymbol{t}_2^*$ for iteration numbers $7n, 8n$ and $9n$ with a small p-value (<0.05). Whereas the mean for t_2^* when $M = 9n$ is not significantly different from t_2^*'s mean for iterations $10n, 11n$ and $12n$. When $\theta_2 = -0.4$, however, (Figure 4.3 bottom), the p-values of the test between the pairs of t_2^* for the last four iterations are larger than the significance level $\alpha = 0.05$. The same experiments using $\pm\theta_2$ were repeated with different $p = (0.1, 0.2, \ldots, 0.9)$, and we can conclude that the *MCMC* output converges to its stationary distribution when $M = 10n$ is sufficient, especially when θ_2 is positive. This is $6n$ times bigger than the value that Ripley (1979) suggested.

4.6 Modified statistics using an average of replicates

For any given parameter setting, the MCMC simulator box in Section 4.5 is designed to return a single simulated data point from a *BMRF* to calculate corresponding summary statistics. However, if we simulate S replicates of data, with their corresponding summary statistics, what is the appropriate S that maximises the speed of *IM*? The one and two-parameter settings of the *BMRF* are used in this section to identify how many replicates we should use in the *IM*.

4.6.1 One parameter model

In this section, the appropriate choice for the number of replicates, S, based on the same given parameter using *MCMC* is theoretically confirmed using the one parameter setting, followed by experimental illustration.

In a single parameter setting, the *M-H* algorithm produces a single output ($\boldsymbol{x}^*$) which is then used to calculate the corresponding summary statistic. We would like, however, to check if the average over many simulated summary statistics in the *IM*, works effectively or if a single simulated summary statistic is better to use.

Suppose we have a value of θ_1 which is chosen correctly, and we can simulate summary statistic $t_1^* = \sum_{i=1}^{n} x_i^*$ from each of S replicates. The average of t_1^* over S is equal to the summary statistic from the observed data, $t_1 = \sum_{i=1}^{n} x_i$, and it can be written mathematically as

$$\widehat{E\left(t_1^*\right)} \simeq \frac{1}{S}\sum_{r=1}^{S}\left\{t_1^*\right\}_r \simeq t_1, \tag{4.32}$$

where $\widehat{E\left(t_1^*\right)}$ is the estimate of expected value for the sample mean. This leads us to rewrite Equation (4.7) as

$$\frac{z^{'}(\theta_1)}{z(\theta_1)} \simeq \frac{1}{S}\sum_{r=1}^{S}\left\{t_1^*\right\}_r. \tag{4.33}$$

Considering the mean of the simulated summary statistic over S replicates in Equation (4.33). This can also be proved theoretically by assuming a fixed function of data $g(\boldsymbol{x}) = \sum_i x_i$ which should occur with probability $p(\boldsymbol{x};\theta)$. Hence, for a single given θ, we can estimate the expected value of $g(\boldsymbol{x})$ as

$$E(g(X)|\theta) = \sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{x}}} p\left(g(\boldsymbol{x});\theta\right) g(\boldsymbol{x}). \tag{4.34}$$

Now suppose $A = \frac{1}{S}\sum_{r=1}^{S} g(\boldsymbol{x})_r$, thus Equation (4.34) can be expressed as

$$A = \sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{x}}} \frac{1}{z(\theta)} \exp\left(\theta g(\boldsymbol{x})\right) g(\boldsymbol{x}).$$

Here $z^{'}(\theta) = \sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{x}}} \exp\left(\theta_1 g(\boldsymbol{x})\right) g(\boldsymbol{x})$, thus A can be written as

$$A = \frac{z^{'}(\theta)}{z(\theta)},$$

where $\frac{z^{'}(\theta)}{z(\theta)}$ is approximately the average over S simulated summary statistics t^{*}.

By experiment, Equation (4.33) can be investigated for a single parameter. However, the choice of the number of replicates S is vital to simulate data closer to the observed data. The appropriate choice helps in speeding-up the *IM* and obtaining the parameter estimates accurately. There are different ways of incrementing S, such as linear, geometric or fixed increasing. In addition to the increment of S, say $\boldsymbol{r}$, with various levels choices for removing either no points or one design point, say $\boldsymbol{M}$, are also needed to run the experiment. To check the speed of *IM*, the last number of design points, say $\boldsymbol{N}$, after the *IM* is stopped with last parameter estimates, is recorded.

The objective in this section is to study the effect on *IM* of the last number of final design points for two methods of increment, five ratios for incrementing S and two ways of removing design points. As previously said, the binomial simulator is used in only one parameter setting because it is cheaper than the *BMRF* and provides similar output.

We start by explaining the experiment in detail. Suppose we use a given data set of $n = 15$ with the observed summary statistic $t_1 = \sum_{i=1}^{n} x_i = 6$ and $\theta_1 = 0$ as initial value. We have three possible methods of incrementing S: linearly (L), geometrically (G) and without change using fixed S (F). For each, we will consider five increments of S, say $\boldsymbol{r}$, which can be either $10, 20, 30, 40,$ or 50 as well as removing no points or one point which we donate $\boldsymbol{M}$, which can be either 0 or 1. For each combination, the experiment is designed to perform the *IM* 50 times, so that each method of increment S (L, G, F) uses each value of $\boldsymbol{r}$ 50 times using both removal of no or one point. The last number $\boldsymbol{N}$ of design points from the *IM* is record as soon as we have the final parameter estimates. When we used, for example, a linear method (L) in running the *IM*, each value of $\boldsymbol{r}$ iterates 50 times for each removal of no or one points. Thus, we will produce a vector $\boldsymbol{N}$ of length 500 including the last number of design points. This dataset is called M^L, and a subset dataset M_0^L refers to removal of no point and M_1^L refers to removal of one point where each subset contains 250 observations.

Methods of comparison:

Two statistical tests have been used: 1) when we compare between groups of $\boldsymbol{N}$ which is a discrete variable using a large sample but $\boldsymbol{N}$ has many outliers. In this case the Kruskal-Wallis non-parametric test is used. The null hypothesis states that there is no differences in means across groups versus the alternative hypothesis that at least one mean is different from the others, and 2) we have one independent variable ($\boldsymbol{N}$) and three explanatory variables, which are factors for different levels with a large sample size, a two-way analysis of variance (ANOVA), which is a parametric test, is used to describe a linear relationship between variables. The full mathematical two-way ANOVA model is given by

$$N_{jlk} = \mu + r_j * M_l * b_k + \epsilon_{jlk}$$

which can be expanded to include all main effect as

$$N_{jlk} = \mu + r_j + M_l + b_k + (rM)_{jl} + (rb)_{jk} + (Mb)_{lk} + (rMb)_{jlk} + \epsilon_{jlk},$$

where μ is the grand mean, $j = 1, 2, 3, 4, 5$ and $l = k = 1, 2$. The r_j is the additive main effect of level j from the first factor, M_l is the additive main effect of level l from the second factor, $r_j b_k$ is the interaction of level l and k from the first and second factor and so forth. The errors ϵ_{jlk} are assumed to be independent and follow a normal distribution with mean zero and variance σ^2.

By fitting the ANOVA model, we need to examine how different levels of three factors ($\boldsymbol{r}, \boldsymbol{M}, \boldsymbol{b}$) and their interactions effect $\boldsymbol{N}$. Using a 0.05 level of significant, the null hypothesis of any main effect is that the means of observations grouped by one factor are equal, however H_0 of any interaction term is that there is no interaction between the main effects. For example, consider the method of incrementing S which is factor $\boldsymbol{M}$ with two levels L and F, H_0 is $\mu_L = \mu_F$. The p-values of the main effects and interaction are then determined using the F distribution (Freund and Wilson, 1998).

Steps of comparison:

1. The geometric method of incrementing S is excluded from the first analysis as it is too time-consuming. Let us consider the last number of design points $\boldsymbol{N}$ of the

geometric method removing no points (M_0^G) and compare it with the linear method (M_0^L) and the fixed method (M_0^F). The range of $\boldsymbol{N}$ of M_0^G is 930-318877550 with mean 98476730, whereas the last number of design points for linear and fixed methods have the ranges 30-66120 and 40-89150 with means 17966 and 16035 respectively. The $\boldsymbol{N}$ of M_0^G is nearly 3577 times bigger than the fixed method.

Table 4.2: ANOVA summary table of response variable N_{jlk} with main effects (r_j, M_l, and b_k) and their interactions, where $\boldsymbol{N}$ is the last number of design points from the *IM*, r_j shows an indicator of the incrementing of S using one of 5 levels $(10, 20, 30, 40, 50)$, M_l which is an indicator of the method type of incrementing S using one of two levels L and F and b_k can include two levels of removing points, the total number of observations is 1000.

Anova model	Variable	Dof	F-value	P-value
$N_{jlk} = \mu + r_j * M_l * b_k$	r_j	4	0.879	0.476
	M_l	1	2.356	0.125
	b_k	1	2.588	0.108
	$r_j M_l$	4	0.893	0.468
	$r_j b_k$	4	0.885	0.472
	$M_l b_k$	1	2.317	0.128
	$r_j M_l b_k$	4	0.901	0.462
$N_{jlk} = \mu + r_j + M_l + b_k$	r_j	4	0.879	0.476
	M_l	1	2.356	0.125
	b_k	1	2.588	0.108

2. We consider next only the linear (M^L) and fixed (M^F) methods as one set and define an indicator variable called $\boldsymbol{M}$ which contains two levels L and F, therefore the total number of observations is now 1000. The ANOVA model is fitted with results in Table 4.2 which contains $\boldsymbol{N}$ as the response variable and main effects $\boldsymbol{M}$, $\boldsymbol{r}$ and $\boldsymbol{b}$ as well as the interaction between them. If the interaction terms were not significant, we then refitted the model without them to see if the p-values of the main effects were changed.

   Consider first the interaction terms, none of the p-values are significant (> 0.05). This indicates that different combinations of interactions have no affect on $\boldsymbol{N}$. Based on the same table, and a 0.05 level of significance, the p-values for all main effects are more than 0.05 and therefore we can not reject the null hypothesis. This means the last number of design points (N_{jlk}) does not affect by the method of increasing S (M_l), ways of removing points (b_k) and values of increasing S (r_j) which states that none of them affect the last number of design points. This means,

there are no significant differences between methods and ratios of incrementing S and the way of removing points.

3. However, removing no point takes approximately five times longer than removing one point. Thus we are going to compare the methods of removing points for each linear (M_0^L and M_1^L) and fixed method (M_0^F and M_1^F) separately where each dataset has 250 observations. Figure 4.4 shows that the process of *IM* takes longer when removing no points in both the fixed and linear methods but that removing a single point has smaller variance. To confirm these differences, a Kruskal-Wallis non-parametric test is used as we have many outliers. In Table 4.3 there was no significant differences in means across the number of design point groups of removing no or one points in the fixed method, whereas the means of the last number of design points in linear method were significant.

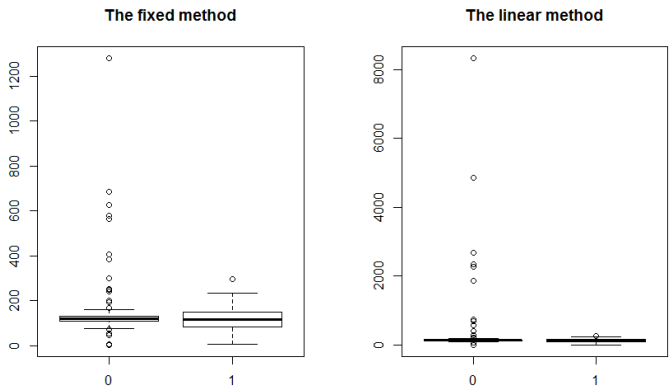


Figure 4.4: Box-plots of the last number of design points for fixed and linear methods including the two strategies of removing points (0 is removing no and 1 is removing one point).

Table 4.3: The non-parametric Kruskal-Wallis test comparing the last number of design points for removing no or one point in each L and F method.

Pairs of variable	Chi-squared	Dof	p-value
M_0^F and M_1^F	2.092	1	0.148
M_0^L and M_1^L	15.549	1	0.000

To sum up, there is no significant difference between the increments of S (r), and the method of incrementing S ($\boldsymbol{M}$) and the method of removing point ($\boldsymbol{b}$). However, a statistically significant difference was clear between removing no and one point, but

only for the linear method. In general, removing no points takes a longer time in the *IM* so it is better to remove points to move through the parameter space toward the true estimation and reduce the time-consumption. We can say that linear and fixed methods were equal so any method can be used. Our conclusion is that any r can do the same job in the *MCMC*, thus it is better to use only one observation ($r = 1$) which has lower time-consumption instead of considering the average over many observations. Also the way of incrementing S, therefore, is not needed and the output of *MCMC* is better to be a single output of simulated data ($\boldsymbol{x}^*$) in the one parameter setting.

4.6.2 Two parameter model

In this section, the output of *MCMC* as an average of S replicates is checked using the two parameter setting. A statistical experiment is only performed because it is challenging to prove it theoretically, as it was done for the one parameter setting in Section 4.6.1. To do the experiment, a range of $\theta_2 = (-1, -0.9, -0.8, \ldots, 0.8, 0.9, 1)$ is considered with a couple fixed values of $\theta_1 = (-1, 0, 1)$ as θ_2 effects the spots being black or white especially when $\theta_1 = 0$. Then we simulate data using 100 *MCMC* runs using Algorithm 3 for each given pairs of parameters and fixed $n = 300$. From the simulated data, the summary statistics ($\boldsymbol{t}_1^*$ and $\boldsymbol{t}_2^*$) are calculated. The box-plots of the summary statistics, over a range of θ_2 and fixed θ_1, are shown in Figure 4.5. It is clear that $\boldsymbol{t}_1^*$ versus θ_1 looks fine with small variances, but $\boldsymbol{t}_1^*$ for given θ_2 behaves unexpectedly. In the top part of Figure 4.5, the means of $\boldsymbol{t}_1^*$ with θ_2 should be unrelated and symmetric around $(-n, n)$, however t_1^* samples vary when θ_2 is increased positively. This variation indicates either completely black or white image.

There is a large variability in the *MCMC* output. This result was also confirmed by Aykroyd et al. (1996) who shows that the realisations from *BMRF* using the *MCMC* for appropriate combinations of parameters widely are different and the behaviour of certain parameter combinations are not straightforward and have big variances. Consequently the output of the simulator should not be the averaging over S replicates as it is meaningless, especially when we have very white and back images simulated from the same parameter value. This leads us to consider several outputs, say three, but without averaging as more sufficient for *IM*. Although a couple of outputs from *MCMC* are considered,

t_1^* could sometimes be extreme, especially with $\theta_2 > 0$. If the extreme t_1^* occurs, this leads us to add design points relatively far away in space from the current estimate and slows down the *IM*. However, the outlier design points are still controllable as these points are later removed in the *IM*.

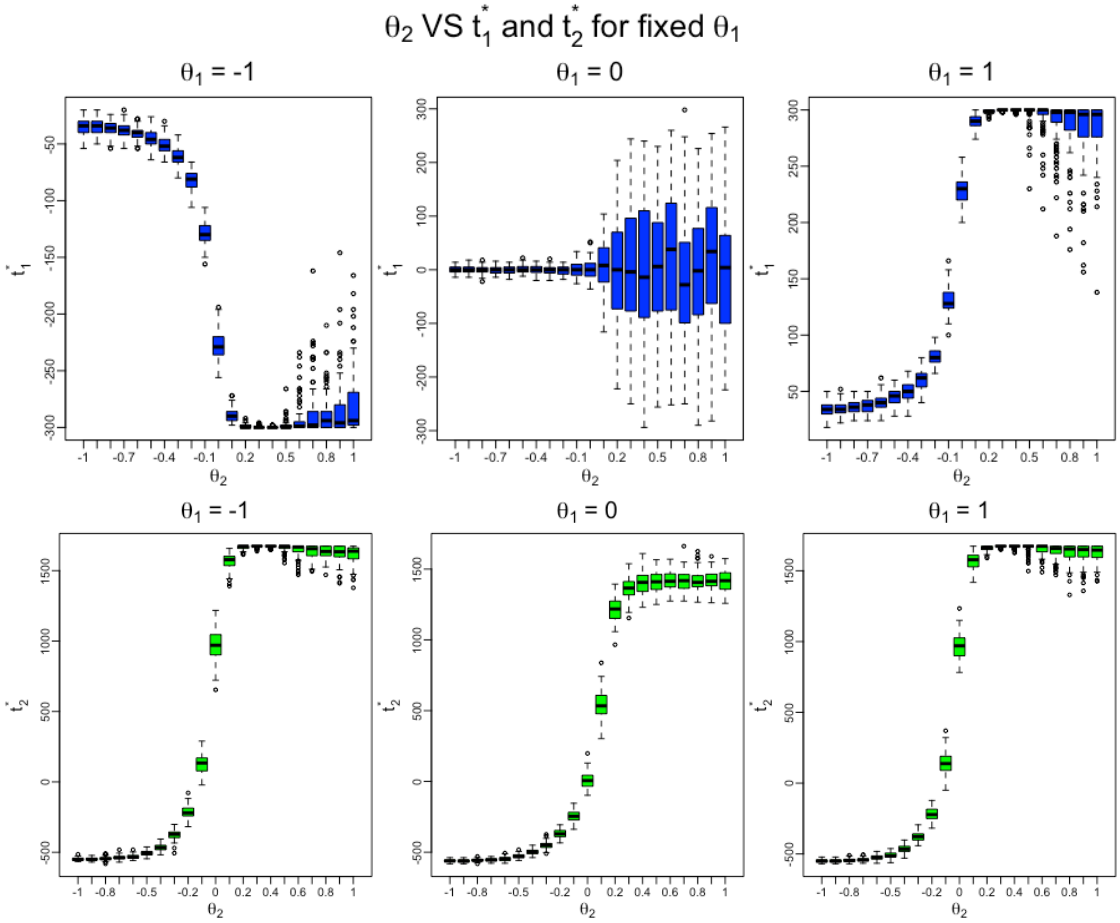


Figure 4.5: Each box-plot gives either $\boldsymbol{t}_1^*$ or $\boldsymbol{t}_2^*$ summary statistic from 100 simulated images using Algorithm 3 over a grid of θ_2 with fixed value of θ_1.

According to Section 4.6.1 and the experiment in this section, the *MCMC* is better to consider several values, eg. three values, as the output without averaging them because t^* has a large variability especially when θ_2 is positive.

4.7 The components of *IM*

Detail about *IM* components, which include how we can add/remove design points, as well as the appropriate criteria to stop the *IM*, are explained. As this estimation method is a sequential simulation-based approach, it is time-consuming. To reduce the computa-

tional burden, we introduce in Section 4.7.1 a way of sequentially removing and adding design points to reach the correct region of parameter space faster. In the meantime, the stopping criteria of the *IM* is explained in Section 4.7.2 which determines a convenient time to stop the method with accurate parameter estimates.

4.7.1 Sequentially adding/removing design points

Adding and removing design points at each iteration of the *IM* are essential parts of this method. The rationale is to move the points to the right place in the parameter space by adding new points close to the current estimate, and removing those far away. In each iteration of the *IM*, we add/remove more than one design point and compare with the current estimated design point that was obtained from previous step.

Let $\Theta_{N\times k}$ denote the existing design matrix of N design points with k parameters, and Θ_{new} with dimension $l \times k$ denote the current design points with l values. The last row of Θ_{new} contains a newly estimated design point, with k parameters that depend on previous fitted realisations (more detail in Section 4.8). To add points to existing ones, we do the following: $\Theta_{(N+l)\times k} = [\Theta^T, \Theta_{\text{new}}^T]^T$, here we are binding the two matrices and increasing the number of design points to be $N = N + l$ and the newly estimated design points are allocated in the last row of $\Theta_{(N+l)\times k}$. Algorithm 4 (Part 1) shows the steps of adding new design points to Θ. However, the process of removing points from Θ depends on evaluating the distance to the newly estimated design point from existing design points to detect where the differences are big. The evaluation is applied for each parameter to find the maximum difference. When the maximum differences for k parameters are allocated in the same row in Θ, one design point is removed, otherwise 2 to k points are removed. Suppose, for instance, we have the two parameter setting and the maximum differences between the new and all previous estimates of θ_1 and θ_2 are allocated in the same row of Θ, therefore one design point is removed, otherwise two points should be removed. The process of removing points is now explained mathematically.

For given $\Theta_{N\times k}$, where the last row contains the current estimate design point $\boldsymbol{\theta}_N = (\theta_{N1}, \theta_{N2}, \ldots, \theta_{Nk})$, we define

$$j_i = \arg\max_j |\theta_{ji} - \theta_{Ni}|, \quad i = 1, \ldots, k,$$

where j_i is an indicator of the maximum difference for the k^{th} parameter. The set of indicators for the parameters are defined as

$$J = \{j_1, \ldots, j_k\}.$$

The values of the index may be repeated when the maximum difference is allocated in the same row. Then, we define $\theta^{(J)}$, which is as Θ with rows corresponding to J removed. Finally the new design matrix will be $\Theta = \theta^{(J)}$. The steps of removing points are summrized in Algorithm 4 (Part 2).

Algorithm 4: The technique of adding/removing design points-Part 1.

ADD $(\Theta, \Theta_{\text{new}})$;
Input: Old and new matrices of design parameter points
Output: A combined matrix
$\Theta = [\Theta^T, \Theta_{\text{new}}^T]^T$;
return Θ;

Algorithm 4: The technique of adding /removing design points-Part 2.

REMOVE (Θ);
Input: A matrix of design parameter points
Output: Update matrix after removing some points
$j_i = \arg\max_j |\theta_{ji} - \theta_{Ni}|, \;\; i = 1, \ldots, k, \;\; J = \{j_1, \ldots, j_k\}$;
Define $\theta^{(J)}$;
$\Theta = \theta^{(J)}$;
return Θ;

The add/remove points step is vital to make sure the simulated data, for given parameters, is close to the local region of the observed one and moves points to the correct parameter space. When we remove points, we want to concentrate design points in the parameter space and reduce the variance. We then get the position required to be able to fit a linear regression model and then estimate the required parameters.

4.7.2 *IM* stopping criteria

Convergence of *IM* to the correct parameter estimates depends on the stopping criteria. Two possible criteria for stopping *IM* are considered, which depend on either the parameter estimates or their corresponding summary statistics, after which the parameter

estimates are taken as the values from the final iteration. When we actually reach the convergence stage in the *IM*, the parameter estimates become closer to the true values as well as the number of design points, N, increases and variance of estimated parameters/design points decreases. We begin by defining the two possible stopping criteria:

1. **Based on the parameter estimates:** stop if the summation of the absolute values of differences between the previous and current parameter estimates of any iteration is less than or equal to a small constant ratio, say $r = 0.01$. This can be written mathematically as

$$\text{Stop if} \quad |\hat{\theta}_1 - \hat{\theta}_1^o| + |\hat{\theta}_2 - \hat{\theta}_2^o| + \cdots + |\hat{\theta}_k - \hat{\theta}_k^o| \leq r \tag{4.35}$$

   where $\theta_1^o, \ldots, \theta_k^o$ are the parameter estimates from the previous iteration.

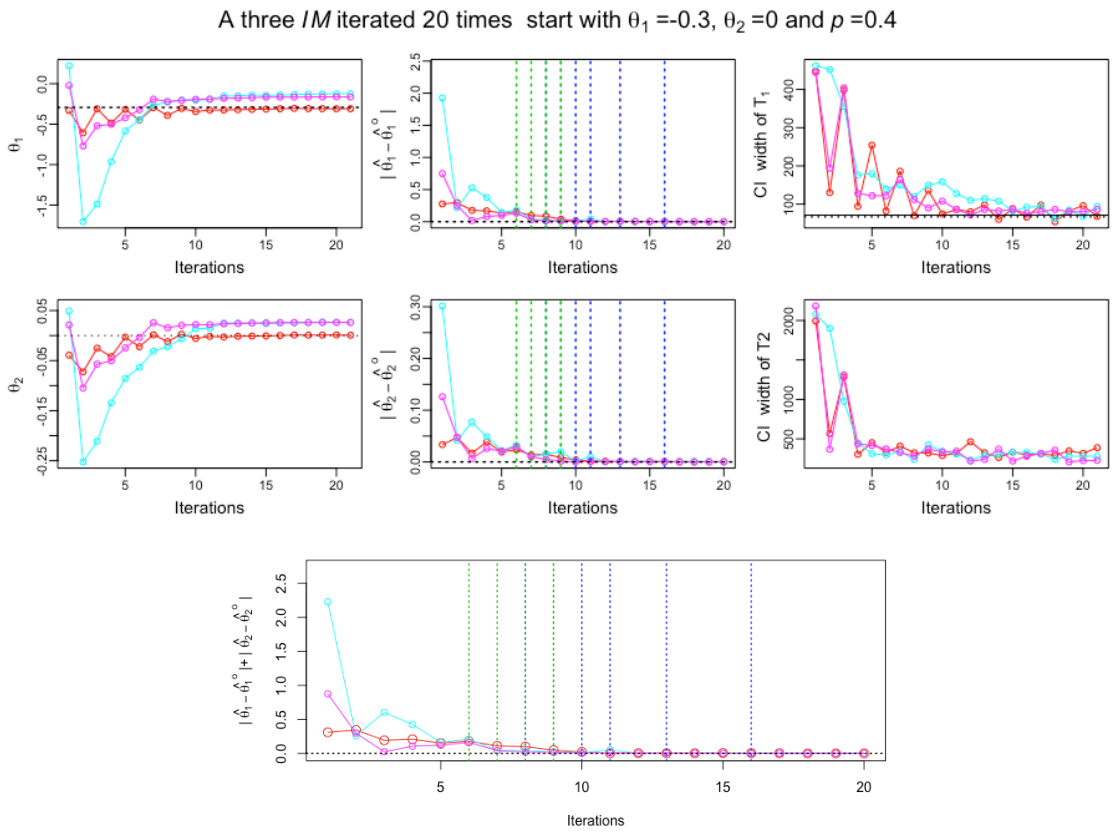


Figure 4.6: The iterative method repeated three times starting from an independent random configuration with $p = 0.358$, $\theta_1 = -0.293$ and $\theta_2 = 0$, where the blue vertical lines are $|\hat{\theta}_1 - \hat{\theta}_1^o| + |\hat{\theta}_2 - \hat{\theta}_2^o| \leq 0.01$, and the green vertical lines are the CI width of t_1 in the case of independence.

2. **Based on corresponding summary statistics:** select a threshold value for only the first summary statistic, t_1, using the width of its confidence interval (CI) to stop *IM*. The threshold is defined by calculating the width of the 95% confidence interval for t_1, where, for simplicity, we estimate the standard deviation on the assumption that the spots are independent to be able to calculate the confidence interval accurately otherwise the CI is unknown. As it has been defined in Section 4.2.1, the image belongs to a binomial distribution and the CI of t_1 is $\hat{t}_1 \pm 1.96\sqrt{4np(1-p)}$. To check the threshold in each iteration of *IM*, the current CI width of t_1 is calculated and determined and if it is less than or equal the threshold, then *IM* is stopped.

To investigate the optimal stopping criteria and convergence of *IM*, we design a simulation experiment using only the two parameter setting for simplification. The stopping criteria can then be generalised for any parameter setting. In order to investigate how long each stopping criteria takes, we consider an independent random image using $p = 0.4$, $\theta_1 = -0.2$ and $\theta_2 = 0$. The *IM* is iterated 20 times for each of three replicas, and then we highlight the two stopping methods as vertical lines in Figure 4.6 (the middle figure in the first and second rows). The second stopping method (green lines) stopped earlier, before the estimates of $\hat{\theta}_1$ and $\hat{\theta}_2$ had stabilised. Even though the second rule stopped earlier and consumed less time, the accuracy of estimated parameters was less than the first method.

In the same figure, we can see how the values of the estimated parameters of θ_1 and θ_2 as well as the CI of the summary statistics, converged when the number of iterations increased. Moreover, the summation of the absolute values of the differences between the previous and current parameter estimates in each iteration are shown in the bottom row of Figure 4.6.

The difference between the two stopping methods can also be shown using another experiment by iterating the *IM* 100 times using simulated independent image with, for instance, $p = 0.4$, $\theta_1 = -0.2$ and $\theta_2 = 0$. Then, we record the parameter estimates of the first and second stopping criteria in Figure 4.7. This plot displays the second stopping method (blue) has the highest variation, where parameter estimates are indeed far away from the given parameters. As a result, the first stopping criteria is used in the *IM*.

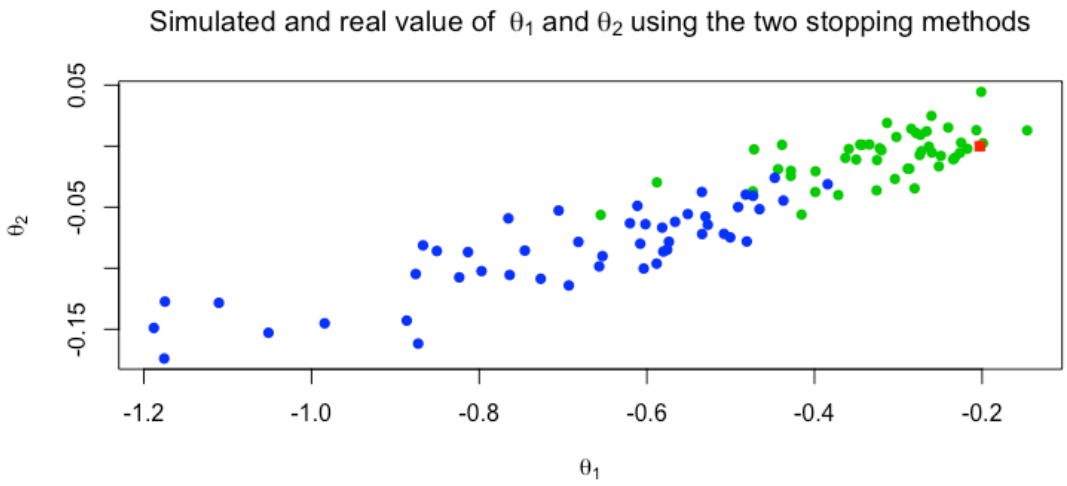


Figure 4.7: The iterative method repeated 100 times using the two stopping methods starting from an independent random image using $p = 0.4$, $\theta_1 = -0.2$ and $\theta_2 = 0$ in a red spot, where green spots denote the first method and blue spots denote the second method.

The ratio r in the fist criteria is chosen to be 0.01, which is good because the differences between current and previous estimated parameters are close to zero. To have unbiased estimators, the ratio r can be selected to be smaller but the time-consumption increases.

4.8 The sequential steps of *IM* for k parameters

The *IM* is a simulation-based optimisation method which can be applied to complex models. This method can also be beneficial when we have more than a one parameter setting which is hard to deal with. The main idea of the *IM* is similar to stochastic optimisation where the process works by minimising the value of a mathematical or statistical function when only simulated realisations are available. The general idea of the iterative method was explained in Section 4.4. The steps of *IM* for k parameters are now explained in detail.

The steps of *IM* for k parameters are as follows:

1. For a given real image, we calculate the observed summary statistics

$$\tilde{\boldsymbol{t}} = (\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_k).$$

2. Creating an initial design point of k parameters $\boldsymbol{\theta}^o = (\theta_1, \theta_2, \ldots, \theta_k)$, where θ_1 is calculated from Equation (4.13), where $\theta_2 = \theta_3 = \cdots = \theta_k = 0$. For each $\theta_i, i = 1, \ldots, k$, an interval of parameter values, $\theta_i \pm 0.1$, is also calculated, thus each parameter has three possible values and the number of design points is 3. Also, a central design point is considered as a midpoint, where all parameters are zeros. By considering the combination of k design points each with 3 values and the central point, we now have, in total, $N = 2k + 1$ design points. The design matrix $\Theta_{N\times k}$, which was defined in Section 4.7.1, contains all parameter values, where the last row, $(\theta_{N1}, \ldots, \theta_{Nk})$, contains the initial design point $\boldsymbol{\theta}^o$.

3. For the j^{th} set of parameters, $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jk}),\ j = 1, \ldots, N$, an image is simulated using Algorithm 3. Then, the j^{th} corresponding summary statistics, $\boldsymbol{t}_j = (t_{j1}, t_{j2}, \ldots, t_{jk})$, is calculated.

4. The relationship between design points, Θ, as explanatory variables and a response summary statistic, $\boldsymbol{t}_i$, is modelled by fitting a multiple regression model. Here, we assume the relationship between $\boldsymbol{t}_i$, where the i^{th} summary statistic contains N values, and Θ is locally linear and that there is no correlation between the corresponding summary statistics to simplify the calculations. The model can be written as

$$\boldsymbol{t}_i = \Theta\boldsymbol{\beta}^{(i)} + \boldsymbol{\epsilon}_i,\ \ i = 1, \ldots, k, \tag{4.36}$$

where Θ is the design matrix of dimensions $N \times (k+1)$ with the first column fixed to be 1, $\boldsymbol{t}_i$ is the i^{th} summary statistic of length N and $\boldsymbol{\beta}^{(i)}$ is the parameter of the model with dimensions $(k + 1) \times 1$. The error term $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \ldots, \epsilon_{Ni})$ for the i^{th} summary statistic has $E(\boldsymbol{\epsilon}_i) = 0$ and $Var(\boldsymbol{\epsilon}_i) = \sigma_i^2$. From the $\boldsymbol{t}_i$ observation of the i^{th} summary statistic, the least squares estimate $\boldsymbol{\beta}^{(i)}$ is then given by

$$\hat{\boldsymbol{\beta}}^{(i)} = (\Theta^T\Theta)^{-1}\Theta^T\boldsymbol{t}_i. \tag{4.37}$$

The fitted model with this estimate can be shown as

$$\boldsymbol{t}_i = \Theta\hat{\boldsymbol{\beta}}^{(i)}. \tag{4.38}$$

From Equation (4.38), $\hat{\boldsymbol{\beta}}^{(i)}$ is now treated as known and we seek to solve for $\boldsymbol{\theta}$ where $\boldsymbol{t}$ is replaced by unknown parameters $\boldsymbol{\theta}$ of length k that need to be estimated as follows

$$\underset{1\times 1}{\tilde{t}_i - \hat{\boldsymbol{\beta}}_0^{(i)}} = \underset{1\times k}{\boldsymbol{\theta}^T} \times \underset{k\times 1}{\hat{\boldsymbol{\beta}}^{*(i)}},\ i = 1, \ldots, k \tag{4.39}$$

where $\hat{\boldsymbol{\beta}}^{*(i)}$ contains $\hat{\boldsymbol{\beta}}^{(i)}$ of the i^{th} summary statistic, but without $\hat{\boldsymbol{\beta}}_0^{(i)}$. To estimate $\boldsymbol{\theta}$ of length k, all summary statistics, $\boldsymbol{t}_i, i = 1, \ldots, k$, are joined to Equation (4.39) to give

$$\underset{k\times 1}{\hat{\boldsymbol{\theta}}} = \underset{k\times k}{\hat{B}^*} \times (\underset{k\times 1}{\tilde{\boldsymbol{t}} - \hat{\boldsymbol{\beta}}_0}), \tag{4.40}$$

where $\hat{B}^*$ includes the model coefficients for all summary statistics. Here we have k linear equations and k unknown parameters. This is a system of linear equations which can be easily solved mathematically to give $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$.

5. Check for convergence. If Equation (4.35) holds, stop the *IM* and $\hat{\boldsymbol{\theta}}$ is the last parameter estimates, otherwise set $\boldsymbol{\theta}^o = \hat{\boldsymbol{\theta}}$ and continue the steps of *IM*.

6. For the given fitted model in Equation (4.38) and $\hat{\boldsymbol{\theta}}$, we predict the corresponding summary statistics $\hat{\boldsymbol{t}} = \hat{t}_1, \ldots, \hat{t}_k$. In this prediction, the lower and upper bounds of each summary statistic, $(\hat{t}_L, \hat{t}_U)$, are estimated. Each summary statistic now has three values and the number of summary statistics is 3. By considering the combination of k summary statistics with 3 values, we now have, in total, $2k + 1$ summary statistics. Similarly Step 2, but the other way around, the corresponding $\hat{\boldsymbol{\theta}}$ is predicted from Equation (4.40). In fact, the given $2k + 1$ parameters are replicated three times, but the corresponding summary statistics are regenerated from the *MCMC* to give various outputs of summary statistic for $2k+1$ parameters. The reason for replication is because the simulator box output can vary even for the same parameter setting (see Section 4.6.2). The final total number of design points is $l = 3(2k + 1)$. Define a new matrix, Θ_{new}, which includes the final set of design points with dimensions $l \times k$.

7. Add design points using Algorithm 4 (Part 1), and remove $4k$ design points using Part 2 of Algorithm 4 which has been repeated four times to remove approximately 50% of added design points.

8. If any parameter estimate goes far away, say $-5 < \hat{\theta}_i < 5,\ i = 1, \ldots, k$, these parameters are immediately removed using Algorithm 4.

9. Go to step 4.

Note that both design points (parameters) and corresponding summary statistics are added and removed. Furthermore, as the output of the simulator box can vary, sometimes it is difficult to control the extreme outlier design points in step 8. Hence, the range of parameter values is restricted to the range $(-5, 5)$. If we have design points outside this range, we should then remove 1 to $4k$ points to reduce the problem. Such extreme points mean that the *IM* takes a longer time to converge, and to obtain the last parameters estimates. Outlier design points can cause a completely prefect fitting of model (4.40), and thus the *IM* is stopped as the regression coefficients, in Equation (4.37), are not estimated.

Beaumont (2010) show the high correlation between the parameters of *BMRF* and the summary statistics. They also showed that the relationship between summary statistics and parameters is highly non-linear. The assumption of the multiple regression model (MRM) in the *IM* is that there is no correlation between dependent variables, but it could be broken as the fitting is local.

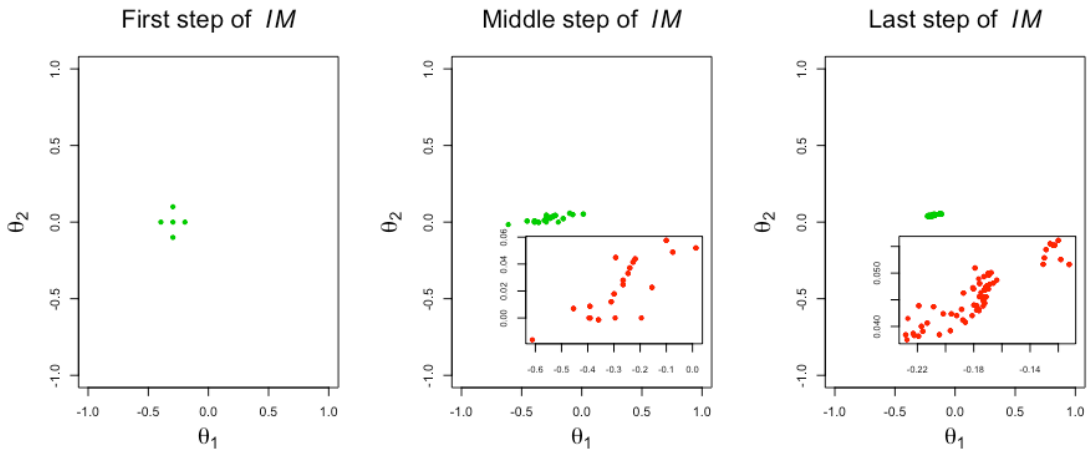


Figure 4.8: Snapshots from three stages of *IM* of the two parameter setting using a real image which contains 317 spots, where the big windows shows the whole parameter space and the internal figure windows are zoom-in versions of current estimated parameter space. By the last step of *IM* we determined $\hat{\theta}_1 = -0.16$ and $\hat{\theta}_2 = 0.05$.

An example of *IM* is shown in Figure 4.8 using the two parameter setting. A real image of 317 spots is used which has 0.36% proportion of tumor (POT). The initial

parameter values for θ_1 and θ_2 are -0.295 and 0, respectively. Here some snapshots are shown from first to final iterations in the *IM*. In the final stage of *IM*, we reach the optimal parameter estimates, $\hat{\theta}_1 = -0.16$ and $\hat{\theta}_2 = 0.05$, with total number of design points $N = 332$.

4.9 Statistical inference and hypothesis testing for $\hat{\boldsymbol{\theta}}$

The distribution of estimated parameters $\hat{\boldsymbol{\theta}}$ is unknown. Also, the calculation of the mean and standard deviation of $\hat{\boldsymbol{\theta}}$ through the *IM* can be quite challenging. Basically, the parameter space is unbounded during the *IM*, but the explored region is concentrating around the optimal estimate. The procedure for making statistical inference, which is based on simulation, is explained for two and k parameter settings. Some examples are given for the non-directional two-parameter setting. Statistical inference for more than two parameters is then explained in Section 4.9.1, including examples.

In order to make the inference, we need to test the hypotheses $H_0 : \theta_2 = 0$ which means there is no clustering in the image. To consider the alternative hypothesis ($H_1 : \theta_2 \neq 0$) means we must carry out the next step of the analysis with the estimated θ_2 from the *IM*.

The main idea of making the inference is estimating the confidence interval (CI) of the summary statistics under $\theta_2 = 0$, when the spots of the image are independently distributed, by simulation using *MCMC*. The distribution of the simulated summary statistics is then compared with the summary statistics of the observed image be able to accept or reject the null hypothesis. The steps of making the inference are as follows:

1. For a given real image, the observed summary statistics, $\tilde{\boldsymbol{t}} = (\tilde{t}_1, \tilde{t}_2)$, are calculated.

2. The optimal parameter estimates, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$, are found using the *IM*.

3. Under the null hypothesis $\theta_2 = 0$ and θ_1, which is estimated by Equation (4.13), simulate independently $\boldsymbol{t}_i^* = (t_{i1}^*, t_{i2}^*)$, $i = 1, \ldots, M$, where e.g. $M = 500$, using the *MCMC* from Algorithm 3. Here, however, the t_{i1}^* and t_{i2}^* of the i^{th} summary statistic are correlated.

4. Check if $\tilde{\boldsymbol{t}}$ is consistent with the distribution of $\boldsymbol{t}_i^*, i = 1, \ldots, M$. To do this, the Mahalanobis distance is calculated (Mardia et al., 1979), which measures the distance between the point $\tilde{\boldsymbol{t}}$ and the simulated summary statistics using the distribution of $\boldsymbol{t}_i^*$. But we first need to define the mean and variance-covariance matrix of a set of $\boldsymbol{t}_i^*$ including $\tilde{\boldsymbol{t}}$. The mean is calculated as

$$\bar{\boldsymbol{t}} = \frac{\sum_{i=1}^{M} \boldsymbol{t}_i^* + \tilde{\boldsymbol{t}}}{M+1},$$

where $\bar{\boldsymbol{t}} = (\bar{t}_1, \bar{t}_2)$ and we added one to the denominator as $\tilde{\boldsymbol{t}}$ is included. Now the variance-covariance matrix is

$$\Sigma_{2\times 2}^{1/2} = \frac{\sum_{i=1}^{M} (\boldsymbol{t}_i^* - \bar{\boldsymbol{t}})^2 + (\tilde{\boldsymbol{t}} - \bar{\boldsymbol{t}})^2}{M}. \tag{4.41}$$

The Mahalanobis distances of $\boldsymbol{t}_i^*$ and $\tilde{\boldsymbol{t}}$, with respect to Σ, are

$$\begin{aligned} \boldsymbol{d}_i^* &= \sqrt{\frac{(\boldsymbol{t}_i^* - \bar{\boldsymbol{t}})^2}{\Sigma}}, \quad i = 1, \ldots, M \\ d &= \sqrt{\frac{(\tilde{\boldsymbol{t}} - \bar{\boldsymbol{t}})^2}{\Sigma}}. \end{aligned} \tag{4.42}$$

5. Then, we compare and count how many d_i^* are larger than or equal to d ($d_i^* \geq d$) to calculate the p-value as follows

$$\text{p-value} = \frac{1 + \sum_{i=1}^{M} I\left[d_i^* \geq d\right]}{M+1}, \tag{4.43}$$

where $\frac{1}{1+M} \leq$ p-value ≤ 1.

If the p-value is less than or equal to $\alpha = 0.05$, we reject the null hypothesis ($H_0 : \theta_2 = 0$) at the 5% level of significance, which means there is clustering in the given image. Basically, the minimum value of the p-value can not be less than $\frac{1}{1+M}$, to increase the range of the p-value to include zero, we could increase M. Making inference in this section does not depend on the normal approximation but it is essential to have a large M. The Mahalanobis distance is an appropriate measure as there is a high correlation between the components of each summary statistic (t_1^* and t_2^*). Making statistical inference in this section can also be generalised for more than two parameters.

Table 4.4: The optimal parameter estimates of the non-directional parameter using the *IM* for 10 images with their corresponding p-values as well as the p-value of the I statistic.

Image #	Non-directional parameter estimates		$H_0 : \theta_2 = 0$	Non-Directional I	
	$\hat{\theta}_1$	$\hat{\theta}_2$	p-value	I	p-value
137507	-0.1573	0.0504	0.0045	0.1153	0.0004
137508	-0.0100	0.0984	0.0001	0.3464	0.0000
137509	-0.1501	0.0334	0.2052	0.0731	0.0304
137511	-0.0470	0.1221	0.0001	0.3446	0.0000
137513	-0.0761	0.1032	0.0001	0.2587	0.0000
137515	-0.0716	0.0614	0.0065	0.1538	0.0002
137516	0.1933	0.0761	0.0001	0.1680	0.0000
137517	-0.4809	0.0647	0.0210	0.1070	0.0012
137518	0.4640	0.0291	0.3683	0.0452	0.1565
137519	0.1056	0.0216	0.3518	0.0559	0.1393

The null hypothesis in this case will be $H_0 : \boldsymbol{\theta} = \mathbf{0}$, where $\mathbf{0}$ refers to all θ_2, θ_3 and θ_4 being zero which means the spots of an image are randomly distributed.

Examples:

From Table 4.4, two different scenarios have been chosen. Image# 137507 has $n = 316$ spots and estimates of $\hat{\theta}_1 = -0.1573$ and $\hat{\theta}_2 = 0.0504$ were determined using the *IM*. To calculate the p-value, we use the parameter value $\theta_1 = -0.29$, which is the estimated parameter when the spots of the image are independent, and $\hat{\theta}_2 = 0$. For these given parameter values, 500 images are simulated using *MCMC*. The summary statistics ($\boldsymbol{t}^*$) are then calculated, which are compared with the observed summary statistics, $\tilde{\boldsymbol{t}} = (-90, 332)$, from the real image. Using the Mahalanobis distance in Equation (4.42), the p-value is then calculated by Equation (4.43) which, in this case, is equal to 0.0045. As a result, the null hypothesis, that $\theta_2 = 0$, is rejected. This means $\tilde{\boldsymbol{t}}$ is not consistent with $\boldsymbol{t}^*$, which is also shown in Figure 4.9 (top). We can now say that there is clustering in image# 137507 and the estimated parameter values for $\hat{\boldsymbol{\theta}}$ are $\hat{\theta}_1 = -0.1573$ and $\hat{\theta}_2 = 0.0504$ from the *IM*.

Another example of calculating the p-value of image# 137509 is shown in Figure 4.9 (bottom). Here, the $\tilde{\boldsymbol{t}}$, $(-69, 180)$ (in a red point), is acutely consistent with $\boldsymbol{t}^*$. This means there is no evidence of clustering and the estimated parameter values are $\hat{\theta}_1 = -0.2366$ and $\hat{\theta}_2 = 0$. These parameters have been estimated when the spots of the

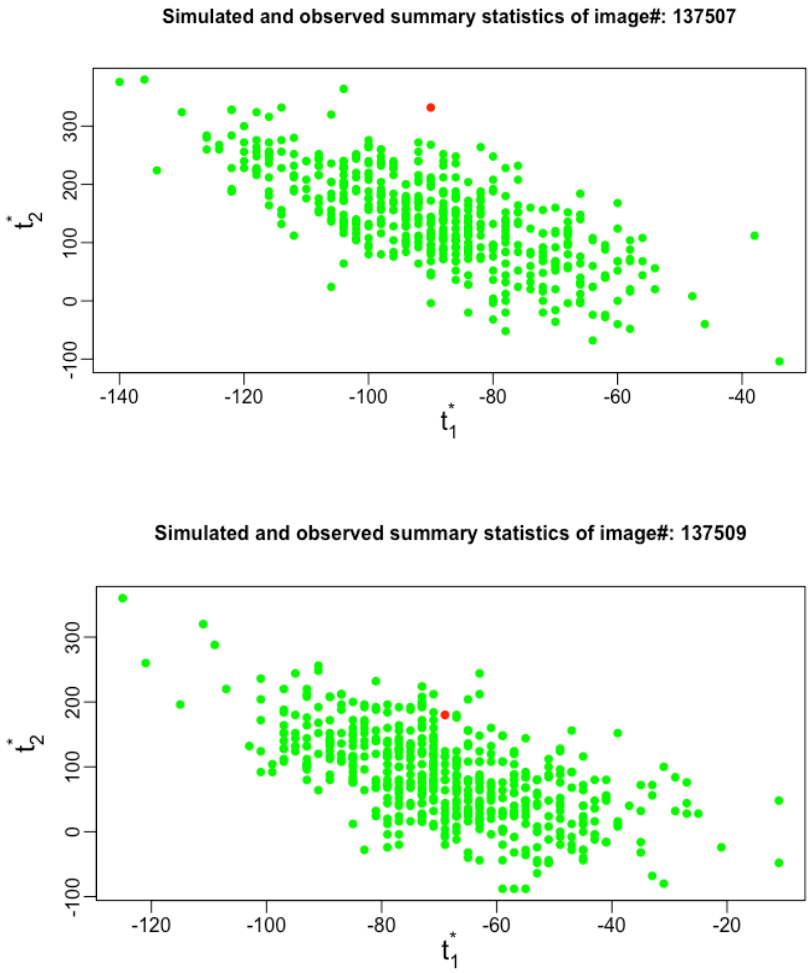


Figure 4.9: Making inference for $\hat{\boldsymbol{\theta}}$ for two images by comparing the observed $\tilde{\boldsymbol{t}}$ (red point) with the generating $\boldsymbol{t}^*$ (green) using independent images simulation using *MCMC* under $H_0 : \theta_2 = 0$ repeated 500 times.

image are considered independent.

The p-value in Equation (4.43) can be compared with the p-value of the I statistic to see if they both lead to the same conclusion of accepting or rejecting H_0 (the spots are independent). We used 10 images to show both p-values in Table 4.4. The p-values are similar, with $\alpha = 0.05$, except image# 137509 which shows no significant difference in the non-parametric test, whereas the parametric test was significant.

4.9.1 Statistical inference for directional $\hat{\theta}$

When we use statistical inference, this is an alternative to the likelihood ratio test which is difficult to evaluate. The procedure that we do is making a comparison between the

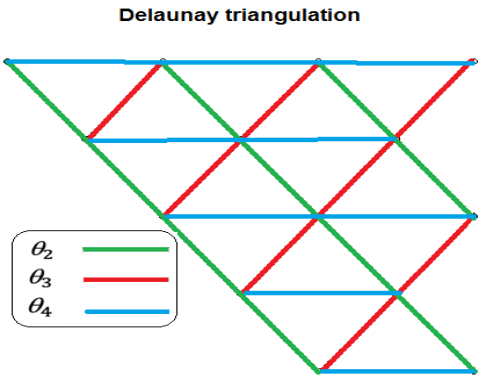


Figure 4.10: Location of the directional $\boldsymbol{\theta}$ on a hexagonal grid.

ratios of three likelihood functions and finding values of the parameters that maximise the likelihood functions. The three proposed likelihood functions are: 1) All spots of the image are independent, where the likelihood is unrestricted ($H_0 : \boldsymbol{\theta} = \mathbf{0}$); 2) isotropic where all $\boldsymbol{\theta}$'s being equal and not zero ($H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$); and 3) anisotropic where all $\boldsymbol{\theta}$'s are not equal ($H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$). At the beginning of this section, when $H_0 : \boldsymbol{\theta} = \mathbf{0}$ is rejected, there is an extra scenario of the hypothesis test that determines if the image also has a preferred direction. The directional parameters are an alternative compared to the directional I statistics. Figure 4.10 shows the direction of each parameter, where the direction of I_1 is equal to the direction of θ_2 and so forth. The hypothesis test for detecting direction is now stated and explained, including examples.

When $H_0 : \theta_2 = 0$ (or in a general form $H_0 : \boldsymbol{\theta} = \mathbf{0}$) is rejected in the previous section, we need to make inference about whether $\theta_2 = \theta_3 = \theta_4 \neq 0$ because we need to test for isotropy of being equal but not zero. The null hypothesis is $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, which states that the image has no preferred direction with all parameters being equal. Here $\boldsymbol{\theta}^0 = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_2, \hat{\theta}_2)$ where $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimated using *IM* when there is no directionality framework. If H_0, is rejected, this means $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$, which are estimated from the *IM* whilst considering directionality, is the correct parameter estimator and the image has a preferred direction.

The steps of calculating the p-value of the direction parameters is similar to the steps in Section 4.9. After rejecting $H_0 : \boldsymbol{\theta} = \mathbf{0}$, the estimated non-directional parameter $\hat{\theta}_2$, using the *IM*, is kept to create $\boldsymbol{\theta}^0$. The directional parameters, using the *IM*, are also estimated $\hat{\boldsymbol{\theta}}$. Then, for given $\boldsymbol{\theta}^0$, we generate a distribution of independent summary statistics $\boldsymbol{t}_i^* = (t_{i1}^*, t_{i2}^*, t_{i3}^*, t_{i4}^*)$, $i = 1, \ldots, M$, if $\tilde{\boldsymbol{t}} = (\tilde{t}_1, \tilde{t}_2, \tilde{t}_3, \tilde{t}_4)$, from real data,

is consistent with the distribution of $\boldsymbol{t}_i^*$, then there is no evidence to reject H_0. Here, the Mahalanobis distance is used to calculate the p-value using Equation (4.43). If the p-value is less than or equal to $\alpha = 0.05$, the null hypothesis, at the 5% level of significance, is rejected. This means that there is a preferred direction in the image and estimated parameter values are $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$. The biggest values from $\hat{\theta}_2, \hat{\theta}_3$ and $\hat{\theta}_4$ are then picked, where the spots of the chosen directions are highly autocorrelated.

Examples:

Some clustered images from Table 4.4 are picked, which have a significant p-value for the hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ to be able then to investigate if the image also has a preferred direction. Table 4.5 displays that the directional parameter estimates, of images in Figure 4.11, using *IM* with the extra hypothesis is applied when $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ being rejected.

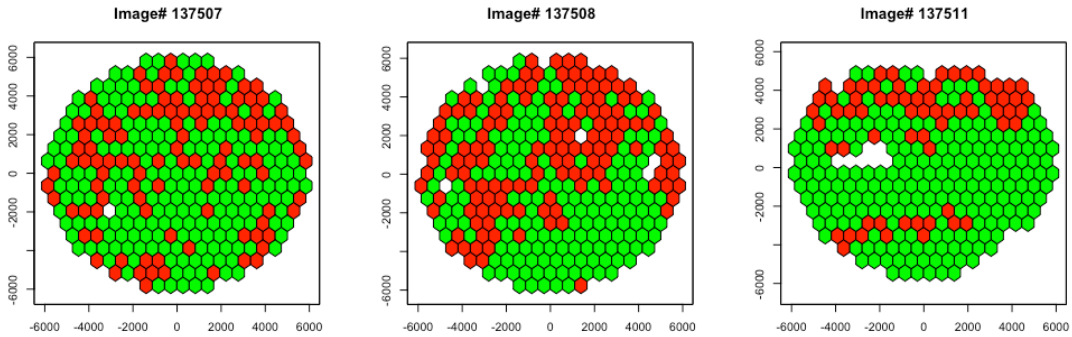


Figure 4.11: Example of three images that are used in Table 4.5.

Table 4.5: The optimal parameter estimates for directional parameters using the *IM* with their corresponding p-values as well as the p-values of the directional I statistic.

Image #	Directional parameter estimates				$H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$	$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$	Directional I
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	p-value	p-value	p-value
137507	-0.1600	0.0562	0.0392	0.0466	0.0045	0.9422	0.7868
137508	-0.0110	0.0284	0.2249	0.0901	0.0001	0.0100	0.0871
137511	-0.2225	0.1465	-0.0465	0.2092	0.0001	0.0020	0.9924

As soon as we decide that there is a preferred direction, we could also detect the location of this direction. To do this we just determine the parameter values are bigger compared to others in the same image. Image# 137507 has no preferred direction, whereas the other images have preferred directions. In image# 137508, the direction

of $\hat{\theta}_3$ stands out as being the largest. However, $\hat{\theta}_2$ and $\hat{\theta}_4$ of image# 137511 indicate preferred directions over $\hat{\theta}_3$ as they have bigger values. Thus, we have mainly high correlation in the directions of θ_4 but also θ_2 is not far behind. The directional I was also calculated for the same images to compare the result with directional parameters. However, recall the distributional assumption of the directional I is not valid when the spots are not independent, and therefore we can not trust the p-value for the directional I.

4.10 Accuracy of *IM* based on simulation

In order to inspect the performance and accuracy of the *IM* proposed in this chapter, the *MCMC* as in Algorithm 3, is used to sample images, using image sizes 50 and 300, with the specified parameters. Here, a fixed neighbourhood system on hexagonal grid is used. The performance and accuracy can be checked by simulating various spatial autocorrelations for given staring parameter values ($\boldsymbol{\theta}^0$) and we find out if the *IM* estimated the parameters correctly. A comparison of an existing estimation method with *IM* is also presented. The mean squared error (MSE) and the one-sample Hotelling's T^2 test statistic are used as the critera for comparisons.

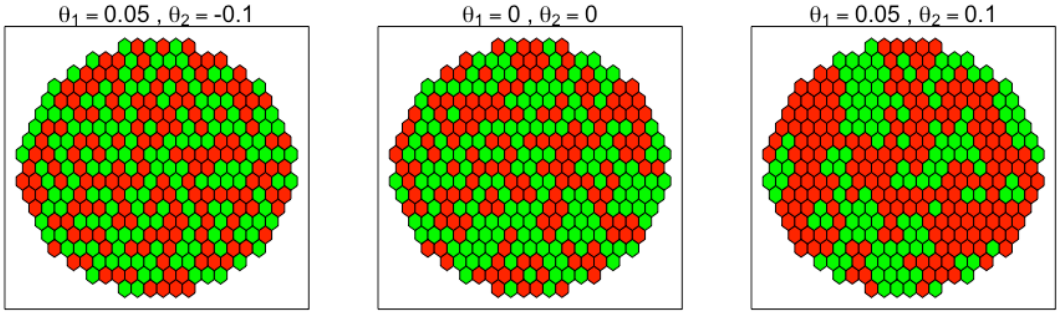


Image structure	Parameter values		Non-directional I	
	θ_1	θ_2	I	P-value
Regular	0.05	-0.1	-0.15	0.00
Random	0.00	0.0	0.00	0.90
Cluster	0.05	0.1	0.28	0.00

Figure 4.12 & Table 4.6: Three simulated images from *MCMC* for given non-directional parameters (θ_1 and θ_2), from the left regular, random and clustered images of 300 spots.

We start by defining briefly one-sample Hotelling's T^2 test statistic. This test is a multivariate generalisation of the t-test which compares between the sample mean vector

$\bar{\boldsymbol{\theta}} = (\bar{\theta}_1, \ldots, \bar{\theta}_k)$ of parameters and the hypothesised mean vector $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_k^0)$. Suppose Θ is a k-dimensional random variable which follows a multivariate normal distribution with $E(\Theta) = \bar{\boldsymbol{\theta}}$ and $V(\Theta) = \Sigma$. Each random variable $\boldsymbol{\theta}_i, i = 1, \ldots, k$, in Θ has n elements and the elements of Θ are not independent. The population variance-covariance matrix Σ is known and can be mathematically computed. The Hotelling's T^2 test statistic is

$$T^2 = n(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\Sigma^{-1}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^0).$$

Under the null hypothesis, $H_0 : \bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, the transformation of the test statistic T^2 is $T^2(k, n) = \frac{n-k}{k(n-1)}T^2$ which follows an F distribution with k and $n - k$ degrees of freedom. A one-sided p-value can be evaluated for $T^2(k, n)$ and we reject H_0 when the p-value is less than $\alpha = 0.05$. A sample size $n = 50$ is considered sufficient for the CLT to hold.

Different spatial autocorrelations can be determined by initialising non- and directional parameter settings in inclusive *IM* evaluation. Starting with the non-directional parameter setting, we choose $\boldsymbol{\theta}^0$ as a starting combination of θ_1^0 and θ_2^0 to generate regular, random and cluster images. Images are generated with two image sizes (50 and 300 spots). Figure 4.12 shows three images generated using *MCMC* for the starting parameter values using 300 spots. The parameter values are also listed in Table 4.6 with corresponding non-directional I statistics and corresponding p-values. The p-values of the I statistic confirmed the significance of regularity and clustered images, as appropriate.

For each combination of θ_1^0 and θ_2^0, 50 iteration images are generated using image sizes 50 and 300 spots, then the *IM* is run and the parameter estimates are recorded. Figure 4.13 shows 50 estimated parameters, here the θ_1^0 and θ_2^0 are highlighted as red cross points. For determination of the accuracy, the MSE and Hotelling's T^2 test are calculated in Table 4.7. The iterative method can accurately estimate the parameter when the image size is 300 because the MSE is quite small as well as the p-values, for all parameter combinations, being big (p-value> 0.05), thus there is no significant difference between the sample means $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$. This means the estimated parameters from the *IM* are consistent with the true parameter values, however, the *IM* does not work well when the image size is 50 spots. It is clear that the estimated parameter values

are not consistent because the MSE of θ_1 is big, although the MSE of θ_2 is quite small. Thus, we reject the null hypothesis and conclude that there is a significant difference between $\bar{\boldsymbol{\theta}}$ in the sample and $\boldsymbol{\theta}^0$. Hence, the image size 50 is not considered in the next simulation studies.

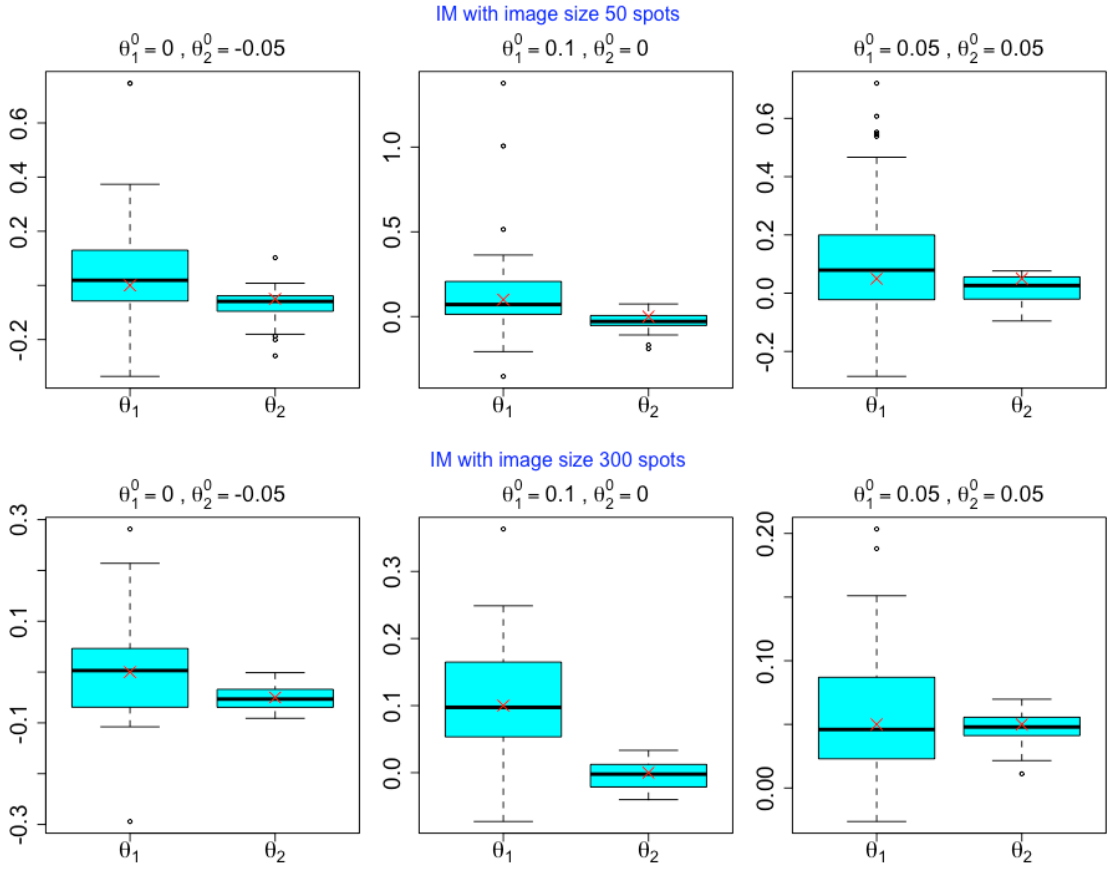


Figure 4.13: Box-plots of 50 estimated $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ from *IM* using simulated images from MCMC for given θ_1^0 and θ_2^0 which are shown as red cross points.

Table 4.7: The mean square error (MSE), standard deviation (Sd) and the p-value of Hotelling's T^2 multivariate test of 50 estimated parameters using the *IM* from simulated images (regular, random and cluster) for given parameters $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)$ with an image size of 50 and 300 spots.

Image simulated using 50 spots image							
Image structure	θ_1^0	$Sd(\boldsymbol{\theta}_1)$	MSE($\boldsymbol{\theta}_1$)	θ_2^0	$Sd(\boldsymbol{\theta}_2)$	MSE($\boldsymbol{\theta}_2$)	Hotelling's T^2 test p-value
Regular	0.00	0.2119	0.0473	−0.05	0.0609	0.0041	0.0432
Random	0.10	0.2715	0.0744	0.00	0.0533	0.0035	0.0049
cluster	0.05	0.2173	0.0510	0.05	0.0473	0.0034	0.0000
Image simulated using 300 spots image							
Image structure	θ_1^0	$Sd(\boldsymbol{\theta}_1)$	MSE($\boldsymbol{\theta}_1$)	θ_2^0	$Sd(\boldsymbol{\theta}_2)$	MSE($\boldsymbol{\theta}_2$)	Hotelling's T^2 test p-value
Regular	0.00	0.0895	0.0080	−0.05	0.0222	0.0005	0.9627
Random	0.10	0.0776	0.0062	0.00	0.0196	0.0004	0.2587
cluster	0.05	0.0521	0.0027	0.05	0.0120	0.0002	0.1624

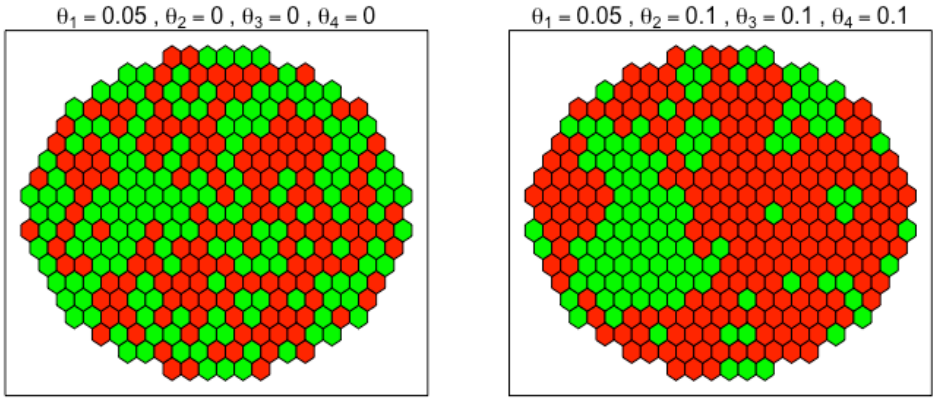


Image structure	Parameter values				Directional I			
	θ_1	θ_2	θ_3	θ_4	I_1	I_2	I_3	p-value
Random	0.05	0.0	0.0	0.0	-0.01	0.09	-0.07	0.86
Directional	0.05	0.1	0.1	0.1	0.37	0.39	0.38	0.00

Figure 4.14 & Table 4.8: Two simulated images of 300 spots fromm *MCMC* for given directional parameters ($\theta_1, \theta_2, \theta_3$ and θ_4), from the left non-directional and directional images.

Similarly, we consider the directional parameter setting. The $\boldsymbol{\theta}^0$ is chosen as a combination of θ_1^0, θ_2^0, θ_3^0 and θ_4^0 that can generate independent/random and dependent directional images. Images are generated with image sizes 300 spots. Figure 4.14 illustrates two images generated from *MCMC* for given parameter values. The parameter values are also shown in Table 4.8 with corresponding directional I statistics and corresponding p-values. Here the p-value in red indicates the detection of directionality in the image.

A similar experimence as for non-directional is applied here where we test the consistency of estimated parameters $\bar{\boldsymbol{\theta}}$ with the true parameter values ($\boldsymbol{\theta}^0$) using Hotelling's T^2 test. Figure 4.15 shows the box-plots of 50 sets of estimated parameters using independent and various dependent directional images where $\boldsymbol{\theta}^0$ is shown at the top of the plots as well as red cross points on the box-plots. From Table 4.9, it is clear that the MSE for all parameters are small as well as the p-values of the multivariate test are not significant. At the 5% level of significance, there is no evidence to reject $H_0 : \bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$ that means the true parameter values ($\bar{\boldsymbol{\theta}}$) are consistent with estimated values ($\boldsymbol{\theta}^0$). As a result, the *IM* works effectively in estimating the parameter values of *BMRF* in the case of directional and non-directional images.

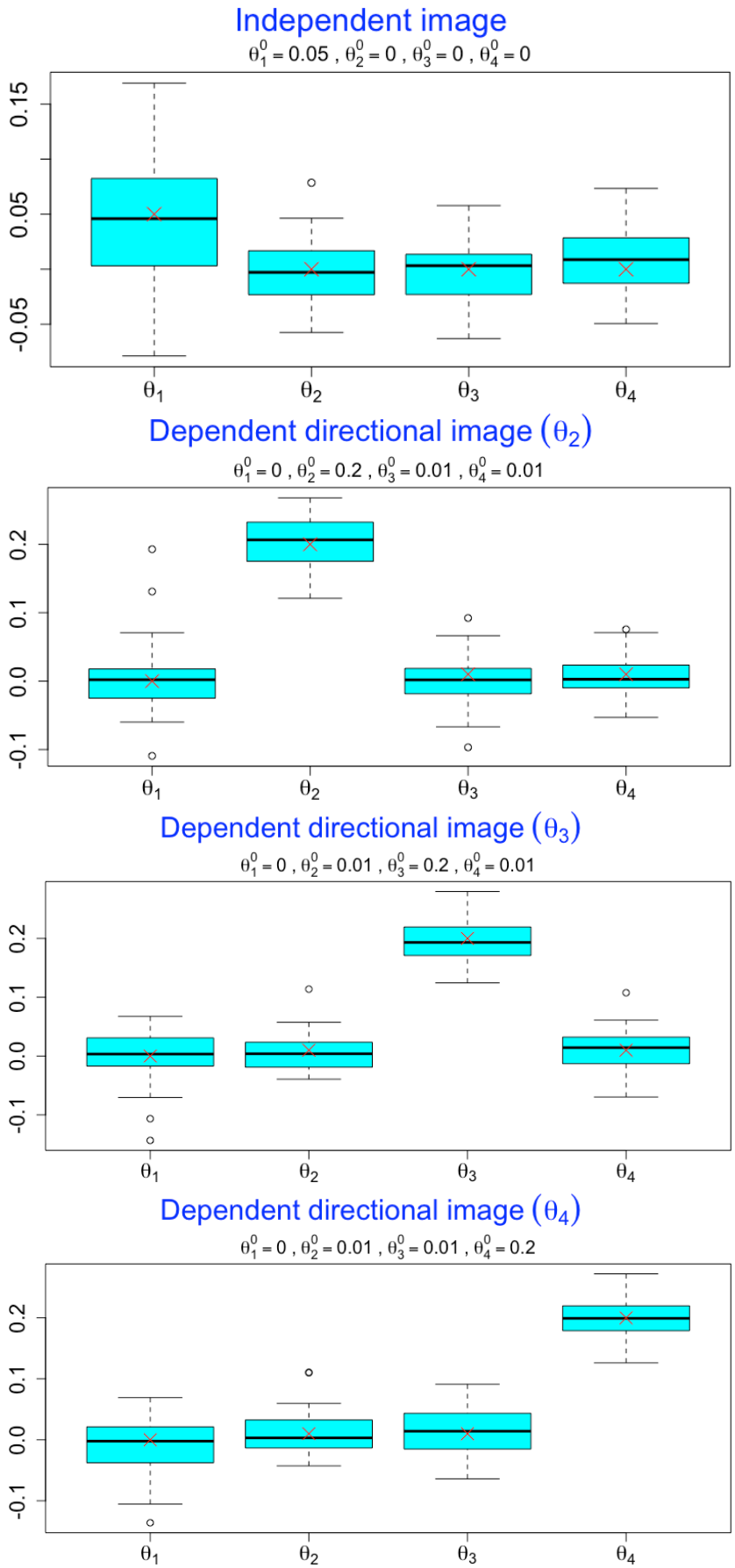


Figure 4.15: Box-plots of 50 estimated $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_4$ from *IM* using simulated images from MCMC for given θ_1^0,θ_2^0, θ_3^0 and θ_4^0 which show as red cross points.

Table 4.9: The mean square error (MSE), standard deviation (Sd) and the p-value of Hotelling's T^2 test of 50 estimated parameters using the *IM* from simulated images (independent and dependent) for given parameters (θ_1^0,θ_2^0,θ_3^0,θ_4^0) with image sizes 300 spots.

Independent image						
θ_1^0	$Sd(\boldsymbol{\theta}_1)$	MSE($\boldsymbol{\theta}_1$)	θ_2^0	$Sd(\boldsymbol{\theta}_2)$	MSE($\boldsymbol{\theta}_2$)	Hotelling's T^2 test p-value
0.05	0.052	0.003	0.00	0.029	0.001	
θ_3^0	$Sd(\boldsymbol{\theta}_3)$	MSE($\boldsymbol{\theta}_3$)	θ_4^0	$Sd(\boldsymbol{\theta}_4)$	MSE($\boldsymbol{\theta}_4$)	0.329
0.00	0.027	0.001	0.00	0.033	0.001	
Dependent Directional image (θ_2)						
θ_1^0	$Sd(\boldsymbol{\theta}_1)$	MSE($\boldsymbol{\theta}_1$)	θ_2^0	$Sd(\boldsymbol{\theta}_2)$	MSE($\boldsymbol{\theta}_2$)	Hotelling's T^2 test p-value
0.00	0.048	0.002	0.20	0.038	0.001	
θ_3^0	$Sd(\boldsymbol{\theta}_3)$	MSE($\boldsymbol{\theta}_3$)	θ_4^0	$Sd(\boldsymbol{\theta}_4)$	MSE($\boldsymbol{\theta}_4$)	0.161
0.01	0.036	0.001	0.01	0.030	0.001	
Dependent Directional image (θ_3)						
θ_1^0	$Sd(\boldsymbol{\theta}_1)$	MSE($\boldsymbol{\theta}_1$)	θ_2^0	$Sd(\boldsymbol{\theta}_2)$	MSE($\boldsymbol{\theta}_2$)	Hotelling's T^2 test p-value
0.00	0.044	0.002	0.01	0.030	0.001	
θ_3^0	$Sd(\boldsymbol{\theta}_3)$	MSE($\boldsymbol{\theta}_3$)	θ_4^0	$Sd(\boldsymbol{\theta}_4)$	MSE($\boldsymbol{\theta}_4$)	0.391
0.20	0.033	0.001	0.01	0.036	0.001	
Dependent Directional image (θ_4)						
θ_1^0	$Sd(\boldsymbol{\theta}_1)$	MSE($\boldsymbol{\theta}_1$)	θ_2^0	$Sd(\boldsymbol{\theta}_2)$	MSE($\boldsymbol{\theta}_2$)	Hotelling's T^2 test p-value
0.00	0.045	0.002	0.01	0.036	0.001	
θ_3^0	$Sd(\boldsymbol{\theta}_3)$	MSE($\boldsymbol{\theta}_3$)	θ_4^0	$Sd(\boldsymbol{\theta}_4)$	MSE($\boldsymbol{\theta}_4$)	0.291
0.01	0.037	0.001	0.20	0.034	0.001	

Now, we compare the effectiveness of the iterative method (*IM*) with the existing pseudo-likelihood method (*PL*), which is defined in Section 4.2.2, using only non-directional parameters and image sample size 300 spots. Different combinations of θ_1^0, θ_2^0, which are listed in Table 4.10, are considered to generate 100 images and then the parameters of these images are estimated by both (*IM*) and (*PL*) and the mean square error is used for comparisons. The results in Table 4.10 of the MSE of *PL* and *IM* parameter estimations show that our method predicts the parameters equally well or even better and the *IM* gives a better agreement between the prediction and the truth.

To sum up, the *IM* works well for a variety of image structures which cover regular, random, cluster and directional images but using only image size 300 spots. Fur-

thermore, the iterative method predicts the parameters better than the pseudo-likelihood method.

Table 4.10: The mean square error (MSE) of 100 estimated parameters using the iterative method (*IM*) and pseudo-likelihood method (*PL*), where the images are simulated for specified non-directional parameters, θ_1^0 and θ_2^0, with image size 300 spots.

Image simulate from		*IM*		*PL*	
θ_1^0	θ_2^0	MSE($\boldsymbol{\theta}_1$)	MSE($\boldsymbol{\theta}_2$)	MSE($\boldsymbol{\theta}_1$)	MSE($\boldsymbol{\theta}_2$)
-0.05	-0.05	0.0057	0.0005	0.0991	0.0037
-0.05	0.00	0.0040	0.0003	0.0974	0.0016
-0.05	0.05	0.0026	0.0002	0.0992	0.0048
0.00	-0.05	0.0059	0.0005	0.0750	0.0040
0.00	0.00	0.0033	0.0003	0.0734	0.0021
0.00	0.05	0.0025	0.0002	0.0733	0.0052
0.05	-0.05	0.0058	0.0005	0.0495	0.0034
0.05	0.00	0.0039	0.0004	0.0479	0.0018
0.05	0.05	0.0026	0.0002	0.0500	0.0055

4.11 Discussion

In this chapter, we established the *IM* simulation approach based on a new method for estimating the model parameters. This method has no assumption about test statistics, nor about the distribution of either data or parameters. The emphasis in this chapter is firstly, estimating in the two-parameter setting for a first-order system in a hexagon grid, which detects clustering. The *IM* is then generalised and extended to four parameters which detect directionality in images. The parameter estimation tends to be more accurate when the ratio of stopping criterion r, decreases but the computation time increases. The *IM* can also work effectively using any $n \times n$ square grid in $\mathbb{R}^2$.

The *IM* is a general technique that can be used in any application when the likelihood is difficult to evaluate in a complex model and a simulator box is available to simulate data easily from given parameter values. When a spatial pattern is simulated, M in the simulator box should be equal to $10n$ to stabilise the *MCMC*. Simulation-based statistical inference is effectively obtained as an alternative approach to the likelihood ratio test. Here there are three scenarios to test, either random, cluster (heterogeneous) or directional images.

The *BMRF* model captures similar information to the I statistic and is particularly

useful for large image sizes ($n = 300$). The *BMRF* has parameters, where significant positive values refer to clustered images, whereas negative values refer to a regular pattern. It is challenging to explain the mathematical relationship between the formula of the I statistic and the *BMRF* model, but this connection can be determined by the p-value of each test.

Directional *BMRF* parameters are more flexible than the directional I statistic and work effectively in detecting directions with fewer assumptions. The directional parameters are applicable for any image structure. The rotation of an image to the lumen direction has not been considered in this chapter. The reason for avoiding rotation is that when we do rotation, the direction of the lumen is not lined up exactly on one of the hexagon axes, and the power of the test is expected to be less, as we discussed for the I statistic in Chapter 3. However, if the orientation of the largest parameter value is allocated in the same axis as the lumen surface, we can say that there is a preferred direction in the direction of the lumen.

The pseudo-likelihood method of parameter estimation has a limitation of not considering the boundary spots in its calculation. In addition to this limitation, the result of parameter estimation will not be accurate when we have missing spots inside the image. Hence the IM is more flexible and more accurate. Nevertheless, the IM has one limitation, which is when we have extreme images, either very black or white, with $p = 0.1$ or 0.9, the method is less effective. This is simply because extreme images produce extreme design points which lead to perfect fitting of the linear regression model.

Chapter 5

Prediction of Biomedical Images

5.1 Motivation and introduction

The work in this chapter is an exploratory analysis motivated by the fact that pathologists tend to collect different samples from the whole tumor, where each patient can have more than one sampling image with different areas and resolutions.

The applications in this chapter use the rectal cancer dataset, which was described in Section 1.3.2. This dataset contains low-resolution spot classification of the whole cancer image W, a high-resolution biopsy Bx, which is sampled from the luminal site before surgery and a high-resolution subset from the whole tumor image which can be one or two disjoint sampling areas. Two sampling areas are G, which contains the highest proportion of tumor, and L, which is closest to the luminal site. A single sampling area is LG which is the area which is closest to the luminal site and in the meantime has the highest proportion of tumor. There are 202 images of pairs W and LG, 66 images of three sets W, G and L and 158 images of pairs W and Bx.

Nevertheless, there is no obvious criteria or method to follow in choosing sampling strategies and we need to find ways, if possible, to compare the information from the samples to see if efficiencies can be made in the collection of data. We focus on image prediction to determine the consistency of low and high resolution images. If the low-resolution images are correctly predicted, this means both images contain the same information, otherwise we need to either sample from both images or to increase the sampling frequency of low-resolution images as we loose information. The consistency

between sampling areas can be considered either from the overall differences of proportions of spot class distributions or the similarity of their spatial features. Here, a fundamental question is: "Can we gain much more information from doing more sampling of the same image or it is not worthwhile?"

By considering the difference in proportions of spot class distributions, another statistical question is "Do the high-resolution images, either Bx, G, L or LG, contain the same information as the whole (W)?" The purpose here would be to estimate the density of cell in the whole tumor ($TCD(W)$) without sampling the whole area, as this measurement is widely used by pathologists. Otherwise, it is better to sample the whole area of cancer. Moreover, pathologists tend to do two high-resolution samples which are L and G, the questions here are "Is it worth sampling both G and L?" or is it enough to consider the corresponding low-resolution areas in W, also "Do both G and L contain the same information?"

This chapter introduces a method for spatial prediction of spot class of low-resolution images from high-resolution images, L, G and LG, where these images overlap with W. Here we are investigating the consistency of the images by attempting to predict the spot class in one of the images from the information in the other one. As the biopsy images do not overlap with the others, the consistency of this type with others will be only checked by spot class distributions. Previously, a binary image (tumor vs. stroma) was considered, but now the excluded spots, which were described in Table 1.5, are kept as there are lots of missing spots which can affect the prediction evaluation. Now we have three classifications of spots: 1 refers to tumor, 2 denotes stroma and 0 to others.

This chapter begins with some notation and definitions of images in Section 5.2 which are used in the prediction process. The distribution of image class proportions for all patients is in Section 5.3. The prediction process is defined for predicting low from high resolution images. This approach is described in Section 5.4 which includes many cases of prediction depending on a smoothing parameter. A better way of prediction is then determined in Section 5.4.2. Finally some discussion is given in Section 5.5.

5.2 Notation and definitions

The notation used for our rectal cancer biomedical images is explained for images with various resolutions. The theoretical definitions are then used to define new methods of predicting spot classes:

1. A `whole tumor image`, W, is a set of coordinates of spots with the corresponding classes two-dimensional, where the delineation of tumor and here the boundary of W is usually a convex polygon. The i^{th} spot has coordinate $\boldsymbol{w}_i = (w_{i1}, w_{i2})$, with class $c_i(W) \in \{0, 1, 2\}$; $i = 1, \ldots, n$, where 1 refers to tumor, 2 denotes stroma, 0 to others and n is the number of spots in W. The set of spot indices in W is $S(W) = \{1, \ldots, n\}$.

2. A `high resolution image`, Y, is a set of coordinates of spots with the corresponding classes in two dimensions, where the delineation of tumor is usually a square region. The j^{th} spot in Y has coordinate $\boldsymbol{y}_j = (y_{j1}, y_{j2})$, with class $c_j(Y) \in \{0, 1, 2\}$; $j = 1, \ldots, m$ where m is the number of spots in Y. The set of spot indices in Y is $S(Y) = \{1, \ldots, m\}$.

The high resolution image Y can be displayed in the whole image W, but there are no coincident locations. A new image is defined which is a subset of W, say $W^{(Y)}$, containing the elements of W close to Y. Specifically $W^{(Y)}$ is the union of all spots in W that are a nearest neighbour to at least one spot in Y.

Table 5.1: The general notation of spots, classes of W, Y and $W^{(Y)}$ images, and the distances between pairs of images.

	W	Y	$W^{(Y)}$
Set of spot indices	$S(W)$	$S(Y)$	$S(W^{(Y)})$
Class of a spot	$c_i(W)$	$c_j(Y)$	$\tilde{c}_i(W)$
Coordinates of a spot	$\boldsymbol{w}_i = (w_{i1}, w_{i2})$	$\boldsymbol{y}_j = (y_{j1}, y_{j2})$	$\tilde{\boldsymbol{w}}_i = (\tilde{w}_{i1}, \tilde{w}_{i2})$
Number of spots	n	m	$<< n \,\&\, m$
Distance between pairs of spots in W and Y	D_{ij}		
Distance between pairs of spots in $W^{(Y)}$ and Y		$\tilde{D}_{ij}$	

To obtain the image $W^{(Y)}$, we need to calculate the distance matrix, D, between two sets of spots in the W and Y images with dimensions $(n \times m)$, where n and m are the

number of spots in W and Y respectively, with elements

$$D_{ij} = ||\boldsymbol{w}_i - \boldsymbol{y}_j||, \quad i \in S(W); \;\; j \in S(Y). \tag{5.1}$$

From this matrix we can find the minimum value over i in order to see which w_i is the nearest spot to y_j with related index I_j, where

$$I_j = \arg\min_i D_{ij}, \quad j \in S(Y). \tag{5.2}$$

Here $I_j \in S(W)$ are indices of W corresponding to the j^{th} element of Y. So the set of indices for the subset image $W^{(Y)} \subset W$ is

$$S(W^{(Y)}) = \bigcup_{j=1}^{m} \{I_j\}, \quad S(W^{(Y)}) \subset S(W), \tag{5.3}$$

where the spots are included in $W^{(Y)}$ if their index is in $S(W^{(Y)})$. The number of spots in subset images $W^{(Y)}$ is less than or equal to the number of spots in W. From here we can also define the indices of Y that are close to $S(W^{(Y)})$ as follows. Suppose the i^{th} spot in $S(W^{(Y)})$ has *neighbours* in $S(Y)$ defined by those spots in $S(Y)$ being closer to the i^{th} spot than any other spots. We denote this set by $\mathscr{I}_i$. In our data the size of $\mathscr{I}_i$ is in the range 1 to 10. This range represents the ratio of low and high resolution images. We refer to this set as the "immediate neighbours" for each i and this can be defined mathematically as

$$\mathscr{I}_i = \{j : I_j = i\}, \quad j \in S(Y); \;\; i \in S(W^{(Y)}). \tag{5.4}$$

Here $\mathscr{I}_i \subset S(Y)$ includes the indices of Y that are immediate spots to $W^{(Y)}$. Here if $\min_{i,j}\{D_{ij};\; i \in S(W^{(Y)});\; j \in S(Y)\} > 0$, this means none of the spots in Y are clearly stated in the same location as those in $W^{(Y)}$.

As $W^{(Y)} \subset W$, the spots and classes of $W^{(Y)}$ can be easily obtained. The set of spots indices $S(W^{(Y)}) \subset S(W)$, so any spot i in $S(W^{(Y)})$ is also a spot in $W^{(Y)}$, and its class can be defined as $\tilde{c} = \{c_i(W); i \subset S(W^{(Y)})\} \in \{0, 1, 2\}$ with coordinates

$\tilde{\boldsymbol{w}}_i = \boldsymbol{w}_i; i \in S(W^{(Y)})$. The distance matrix between $W^{(Y)}$ and Y, is defined as

$$\tilde{D} = \{D_{ij};\ i \in S(W^{(Y)});\ \ j \in S(Y)\}. \tag{5.5}$$

The $\tilde{D}$ helps in predicting $W^{(Y)}$ from Y. The general notations of the spots and classes in W, Y and $W^{(Y)}$ are summarised in Table 5.1, including the distance between pairs of images.

Figure 5.1 shows an example of image# 105420 which has a low-resolution whole tumor image W with two high-resolution layout images G and L with corresponding subset images $W^{(G)}$ and $W^{(L)}$ from W.

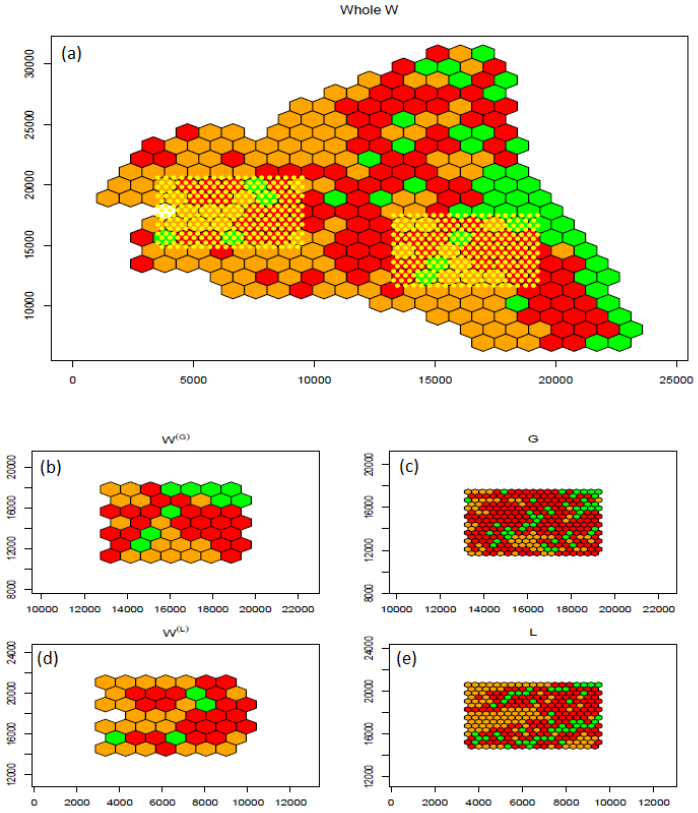


Figure 5.1: Image# 105420. (a) the whole tumor image, where the yellow dots show the locations of the spots on the high resolution images (G and L), (b) a subset image $W^{(G)}$ from the whole, (c) the high resolution image G, (d) a subset image $W^{(L)}$ from the whole and (e) the high resolution image L.

5.3 Consistency of distributions for class proportions

The consistency of images are investigated by comparing the overall distribution of spot class proportions. The questions here are: 1) Is Bx consistent with W; 2) Are any G, L or LG consistent with W; 3) Are high-resolution G and L consistent with low-resolution $W^{(G)}$, $W^{(L)}$ and 4) Are G and L consistent. The statistical test of assessing consistency is explained theoretically and then an example of a single image W, which has two corresponding images L and G, is illustrated. All images are then included for each question.

In general, the consistence of c images can be checked by firstly consider the differences in class distribution for c of images. Fisher's exact test is based on the hypergeometric distribution under the null hypothesis is that the frequency of classes in c images are the same which would be true if the c images are consistent. A dataset like this is summarised in an $r \times c$ table where r is the number of rows, which represents the class frequency, and c is the number of columns. The Fisher's test is more appropriate than the Pearson's χ^2 test as we expect at least one expected frequency in the table to be less than five when the small images are considered. Hogg and Tanis (1977) explained the Fisher's test for multiple groups as follows. Suppose we have $n_1, n_2, \ldots, n_c$ objects in each class, and $n_1 + n_2 + \cdots + n_c = N$, then the probability in Fisher's exact test is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_c = x_c) = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}\cdots\binom{n_c}{x_c}}{\binom{N}{m}}, \tag{5.6}$$

where X_1 is the number of successes in the first class of size n_1 and so forth, $\binom{x}{y} = \frac{x!}{y!(x-y)!}$ and $x_1 + x_2 + \cdots + x_c = m$ which is the total number of successes. A two-tailed p-value is then obtained directly by summing of the tails using the hypergeometric distribution.

Example:

The proportion distributions of classes are compared for image# 105420 in Figure 5.1. The frequency of the classes per image (W, LG, $W^{(G)}$ and $W^{(L)}$) is calculated, for instance for image W, by

$$f_k = \sum_{i=1}^{n} I[c_i(W) = k], \quad k = 0, 1, 2. \tag{5.7}$$

Here f_k is the frequency of class k. The frequencies of classes for each image version corresponding to Equation (5.7) are summarised in Table 5.2.

Table 5.2: The class summaries of image# 105420.

Frequency	W	G	L	$W^{(G)}$	$W^{(L)}$
f_0	132 (42%)	58 (20%)	104 (35%)	16 (33%)	24 (48%)
f_1	131 (42%)	191 (64%)	145 (49%)	24 (49%)	22 (44%)
f_2	52 (16%)	47 (16%)	47 (16%)	9 (18%)	4 (8%)
Total	315 (100%)	296 (100%)	296 (100%)	49 (100%)	50 (100%)

Now pairs of images are compared. From Table 5.2, pairs of images (corresponding to two columns) are considered to create a 3×2 table. The corresponding p-values of Fisher's exact test are displayed in Table 5.3. Here, the small p-values (< 0.01) are highlighted as a red, where $\alpha = 0.05/5 = 0.01$ is the level of significance using a Bonferroni correction for multiple testing (Bland and Altman, 1995). These significant p-values show inconsistency between the class distributions of G and W as well as between G and L. This means the distribution of classes of the G image is not consistent with W and L. However, the distributions of classes for the other image pairs may be consistent.

Table 5.3: The p-values of the Fisher's test for comparing the proportion distributions of classes using all possible pairs of images which are plotted in Figure 5.1. Bonferroni correction is used for significant p-values.

Pairs of images	P-value
G vs. W	0.000
L vs. W	0.157
G vs. $W^{(G)}$	0.075
L vs. $W^{(L)}$	0.156
G vs. L	0.000

The same comparisons are applied to all images to answer the questions at the beginning of this section. Under the null hypothesis, we would expect 5% of images to be rejected to confirm pairs of images are consistent, otherwise there is not enough evidence of consistency. This is a binomial problem, thus we can also consider the confidence interval to get an overall view of consistency. If we have consistency, we expect 95% of images to have large p-values (p-value> 0.05). The confidence interval using the binomial distribution for the 66 images is $[87\%, 99\%]$. Likewise the confidence interval of accepting consistency for the set of LG and W images is $[92\%, 98\%]$ and for the set of Bx and W images is $[90\%, 98\%]$.

Table 5.4: The percentage of the Fisher's test p-values being not rejected (> 0.05) for comparing the proportion distributions of class for all patient images where the percentages in red show the consistent pairs of images.

Pairs of images	p-value> 0.05
66 images	
G vs. W	12%
L vs. W	23%
G vs. $W^{(G)}$	97%
L vs. $W^{(L)}$	98%
G vs. L	17%
202 images	
LG vs. W	16%
LG vs. $W^{(LG)}$	92%
158 images	
Bx vs. W	4%

In Table 5.4, the set of W, L and G, only G and L are consistent with $W^{(G)}$ and $W^{(L)}$ respectively. Similarly, the LG and $W^{(LG)}$ have 92% of pairs of images which are consistent. This means the proportion distributions of classes for low-resolution images are consistent with high-resolution images when they are overlapping. These pairs of images can be then used in the following section to consider the prediction of spots spatially. However, the proportion distributions of classes for the whole tumour W is not consistent with any high-resolution images (G, L, LG, Bx). The proportion of classes for high-resolution images G and L are also not consistent.

To investigate the differences between the class distributions of inconsistent pairs of images, the box plot for each pair of images is plotted. Figure 5.2 shows box plots for the distributions of classes for each pair of images that were compared in Table 5.4. Here the median of class 0 for each pair of images is quite similar with lots of outliers, except Bx and W. The median of class 1 for W, which is the proportion of tumor in the whole image, tends to be lower than the others. Whereas the proportion of stroma on W has higher median than the other images.

To sum up, the low-resolution images are consistent with high-resolution images when they overlap. However, none of the high-resolution images can represent W and the high-resolution of G and L are not consistent. Therefore, it is important to sample the whole image, but there is less need to sample high-resolution images as they contain the same proportion of classes as the low-resolution images.

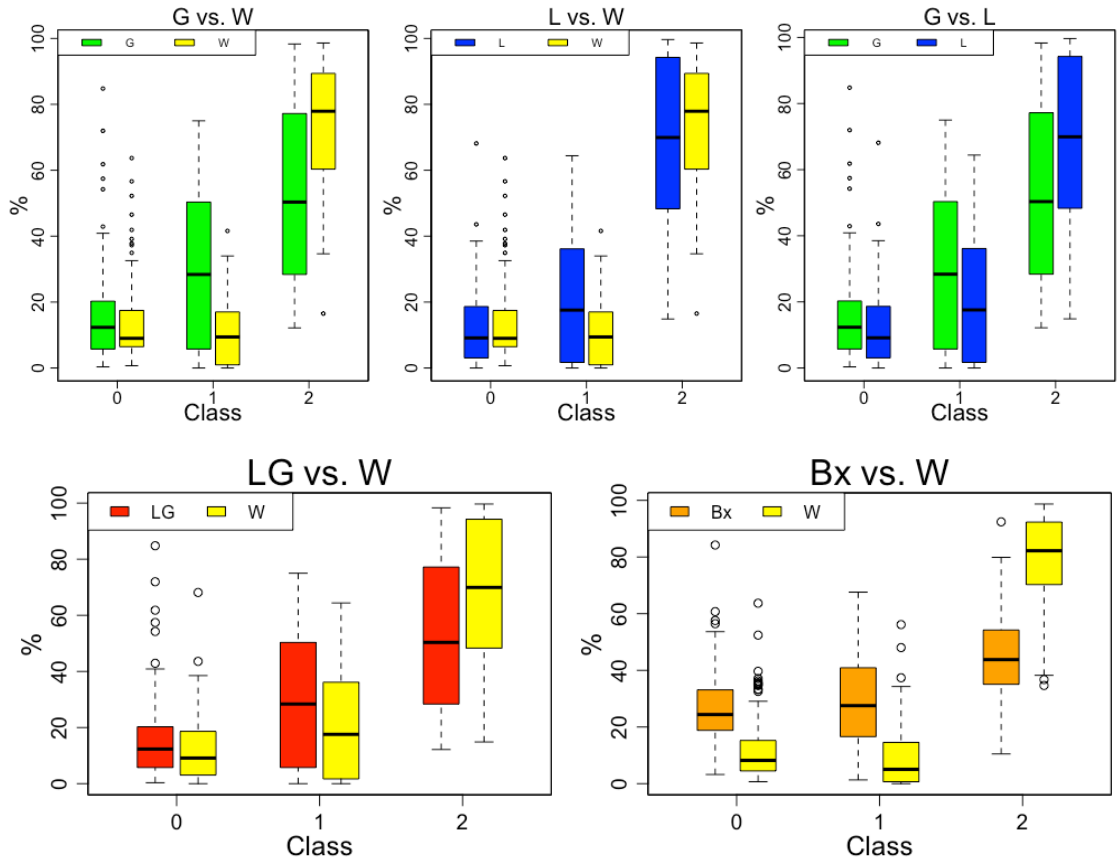


Figure 5.2: Box plots of class distributions for pairs of inconsistent images from Table 5.4, where 1 refers to the proportion of tumor, 2 denotes the proportion of stroma and 0 to the proportion of other classes.

5.4 Spatial class prediction of $W^{(Y)}$ using Y with distance-weighting

This section presents some methods of spatial prediction for the spot classes of low-resolution images from high-resolution images, and then tries to compare methods. Pairs of images, which share the same location but with different resolutions, are considered, e.g. $W^{(Y)}$ and Y. The distributions of classes for these pairs of images were confirmed in Section 5.3 that they are consistent. The objective now is to predict the class of each spot spatially, and find if there is a good matching agreement between the predicted and observed spot classification to determine then if pairs of images are spatially consistent.

The spots classes are predicted spatially using a weighted-distance mode, with weights dependent on distance, to be defined later. To assess the prediction, a correctly predicted ratio (*CPR*) is calculated which considers how many spots in the predicted image have

been correctly classified. This statistic allows us to compare the predicted and observed images to see if they are consistent.

To define this statistic, suppose we have observed and predicted $W^{(Y)}$ with a number of spots N.

$$CPR = \frac{\text{\# of spots correctly predicted}}{N} \times 100. \tag{5.8}$$

The distance matrix $\tilde{D}$ between the pairs of spots in $W^{(Y)}$ and Y images, which was defined in Equation (5.5), is used to find the weight of each spot in $W^{(Y)}$. This can be implemented by predicting the spot classes using a weighted sum of any neighbouring setting which is important in the prediction process. Two neighbouring settings are considered: a weighted-distance mode of either the immediate neighbours or all spots in the image. In addition to the neighbouring setting, a smoothing parameter, α, controls the relative weights.

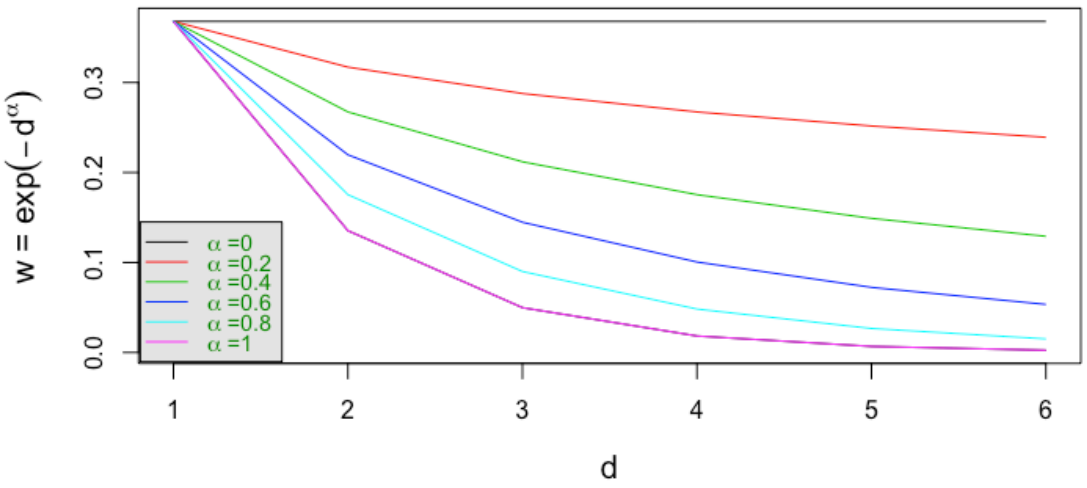


Figure 5.3: Plot of the relationship between the distance and weight for different α, where each line displays different values of α, where $\alpha = 0$ shows all spots with different distances have the same weight.

The distance matrix, $\tilde{D}$, is used to weight the spots depending on their distances from the original image Y. For instance, either all spots can contribute equal weights across the neighbourhood system, or the weight of immediate spots can be increased. This allows us to either weight all spots, or restrict to the immediate neighbourhood. To predict a spot class, the weighted-distance mode predicts the spot classification which depends on the weighting of this spot. The weight of spots can be calculated as an

exponential function of the negative distance between $W^{(Y)}$ and Y images raised to the power parameter α. Suppose that the i^{th} spot $\in S(W^{(Y)})$, which includes all spots in $W^{(Y)}$, needs to be predicted. Then the weight can be written as

$$w_{ij} = \exp\left(-\tilde{D}_{ij}^{\alpha}\right), \quad j \in S(Y); \;\; i \in S(W^{(Y)}). \tag{5.9}$$

Here there are $j = 1, \ldots, m$ spots $\in S(Y)$ with their distances and different weights depending on α. When this smoothing parameter increases, the weight decreases for far away spots. Figure 5.3 shows the relationship between the weight function and different smoothing parameters α. As α decreases the distance of spots tends to be equally weighted. Thus the determination of the spot class can vary depending on the smoothing parameter which affects the weight of spots. Different values of α are consider in prediction methods. In Section 5.4.1, methods of predicting low-resolution images $W^{(Y)}$ from high-resolution Y are explained, and then all methods are assessed in Section 5.4.2.

5.4.1 Predicting $W^{(Y)}$

In this section the steps of predicting $W^{(Y)}$ from Y are explained including all special cases of different α. Suppose that for the i^{th} spot, the class $\tilde{c}_i(W)$ can be predicted by considering the class of the spots in Y and their distance from spots in $W^{(Y)}$. This information contained in $c_j(Y)$ and $\tilde{D}_{ij}$ which were introduced in Section 5.2. A general formula for predicting the $W^{(Y)}$ classes from Y as a function of the weighted-distance cold be written as

$$\hat{\tilde{c}}_i(W)(w) = \arg\max_k \sum_{j \sim i} I\left[c_j(Y) = k\right] w_{ij}, \quad i \in S(W^{(Y)}); \;\; j \in S(Y), \tag{5.10}$$

where

$$w_{ij} = \exp\left(-(\psi_i \tilde{D}_{ij})^{\alpha}\right), \quad i \in S(W^{(Y)}); \;\; j \in S(Y), \tag{5.11}$$

where $j \sim i$ can be either immediate neighbours of i, which defined in Equations (5.4), or can include all $S(Y)$, $k \in \{0, 1, 2\}$, ψ_i is a constant which helps to scale the distance matrix and α controls distance priority. Equation (5.10) predicts the class of the i^{th} spot in which the classes of either immediate or all spots are weighted according to the

distance to the i^{th} spot. When $\alpha = 0$ for any neighbouring setting, all weights are equal, thus we only count how many times we take a certain class by considering the agreement over all classes k which maximises Equation (5.10).

Table 5.5: Methods of spatial prediction for classes of $W^{(Y)}$ from Y using Equation (5.10) using immediate neighbouring and all settings.

Prediction method	Method description	α	Neighbourhood setting
M_1	Maximum class	0	All
M_2	Unweighted local mode	0	Immediate
M_3	Weighted mode	$0 < \alpha < \infty$	All
M_4	Weighted mode	$0 < \alpha < \infty$	Immediate
M_5	Nearest spot	$\alpha \longrightarrow \infty$	Both immediate and all

Appropriate values of α and ψ_i in Equation (5.10) need to be determined. To choose the optimal α, the distance matrix is firstly scaled to find how well this smoothing parameter performs using *CPR*. The reason behind scaling is standardising the distance, and the α is then commensurable. The scaling of the distance matrix can be applied according to 1) nearest distance as minimum $\left(\psi_i = \frac{1}{\min_{j \in \mathscr{I}_i} \tilde{D}_{ij}}\right)$, 2) farthest distance $\left(\psi_i = \frac{1}{\max_{j \in \mathscr{I}_i} \tilde{D}_{ij}}\right)$ or 3) fixed scaling ($\psi_i = 1$). To compare between different scaling methods to be used later in spatial prediction, the *CPR*(%) was tested for an image with different values of α. The scaling by minimum distance had the behaviour that we expected, as it shows some special cases of Equation (5.10) when $\alpha = 0$ and $\alpha \longrightarrow \infty$, which are explained below.

There are five methods of spatial prediction that can be applied using Equation (5.10). An example of image# 105420, which is shown in Figure 5.1, is used for predicting low- from high-resolution images using all methods. The prediction methods are summarised in Table 5.5, which can be explained as follows:

M_1: Predicting $W^{(Y)}$ when $\alpha = 0$ over all spots

In Equation (5.10), when $\alpha = 0$ using all spots, all spots are equally weighted and every spot is predicted to have the most common class in Y. This method is a totally naive approach, but most spots are classified correctly. This case takes the most frequent class in Y as a prediction of $W^{(Y)}$ because we know nothing about the location of spots but at

least the most common spot classes are correctly estimated.

Table 5.6: Tables of agreements between the original classes ($\tilde{c}$) and predicted classes ($\hat{\tilde{c}}$) of $W^{(G)}$ and $W^{(L)}$ images predicted by M_1 with corresponding *CPR* using the image# 105420.

$W^{(Y)}$	$\hat{\tilde{c}}$	$\tilde{c}$			Total	*CPR*
		0	1	2		
$W^{(G)}$	1	16	24	9	49	49%
$W^{(Y)}$	$\hat{\tilde{c}}$	$\tilde{c}$			Total	*CPR*
		0	1	2		
$W^{(L)}$	1	24	22	4	50	44%

Table 5.6 illustrates the agreement of the classes between the original image and predicted image for both $W^{(G)}$ and $W^{(L)}$. Here, only class 1 is correctly predicted in both $W^{(G)}$ and $W^{(L)}$ with *CPR* equal 49% and 44% respectively.

M_2: Predicting $W^{(Y)}$ when $\alpha = 0$ over immediate neighbours

Again when $\alpha = 0$ using immediate neighbours, all spots are equally weighted. The prediction process for a spot is just counting how many times a certain class has been repeated in its immediate neighbourhood. This case can be called the "unweighted local mode".

Table 5.7: Tables of agreements between the original classes ($\tilde{c}$) and predicted classes ($\hat{\tilde{c}}$) of $W^{(G)}$ and $W^{(L)}$ images predicted by M_2 with corresponding *CPR* using the image# 105420.

$W^{(Y)}$	$\hat{\tilde{c}}$	$\tilde{c}$			Total	*CPR*
		0	1	2		
$W^{(G)}$	0	8	1	0	9	69%
	1	8	22	5	35	
	2	0	1	4	5	
	Total	16	24	9	49	
$W^{(Y)}$	$\hat{\tilde{c}}$	$\tilde{c}$			Total	*CPR*
		0	1	2		
$W^{(L)}$	0	18	1	0	19	76%
	1	6	19	3	28	
	2	0	2	1	3	
	Total	24	22	4	50	

Table 5.7 shows the agreement of classes between the original image and predicted image with corresponding *CPR* for each $W^{(G)}$ and $W^{(L)}$ using image# 105420. The the

diagonal part of the table shows the correctly predicted classes. For example, the second class of the $W^{(G)}$ image has the highest frequency which means 45% of spots with class 1 are correctly predicted using the unweighted method, but only 8% of the class 1 is well predicted. This method has a better prediction with higher *CPR* for spot classes more than the maximum class method.

M_3&M_4: Predicting $W^{(Y)}$ when $0 < \alpha < \infty$

This method contains two special cases which consider a range of the smooth parameter α with immediate neighbours called M_3 and with all spots settings called M_4. These methods described as "weighted mode".

To explain how the weighted mode method works, suppose we would like to predict the i^{th} spot in $W^{(Y)}$ using only the immediate neighbours $\mathscr{I}_i$, so the $\hat{\tilde{c}}_i(W)$ for given α can be estimated as the weighted majority of its neighbours. Here the i^{th} spot has a list of immediate neighbours in $\mathscr{I}_i \in S(Y)$. For instance, for the i^{th} spot, suppose $\mathscr{I}_i = \{10, 60, 54\}$ with corresponding classes $\{c_{10}(Y) = 1, c_{60}(Y) = 0, c_{54}(Y) = 1\}$ and weights $w_{j1} = \{w_{10,1} = 0.22, w_{60,1} = 0.33, w_{54,1} = 0.12\}$. Here, there are three immediate neighbours. The weighted frequency of class zero is 0.33 and class 1 is 0.22+0.12=0.34, thus $\hat{\tilde{c}}_i(W) = 1$ is the right estimate of the i^{th} spot as class 1 has the higher weight. Sometimes the classes are equally weighted, in this case the first element in the classes set is considered for predicting $\hat{\tilde{c}}_i(W)$. The same procedure of spot prediction also works when we consider all spots.

Using the weighted mode method, the *CPR*(%) is calculated when $0 < \alpha < 5$ in Figure 5.4 using $W^{(G)}$ and $W^{(L)}$ from image# 105420 considering the immediate neighbouring and all spots orders. This method of prediction considers different weights of spots by changing α, then we calculate *CPR* which depends on the mode of the highest weighted distance using the immediate neighbouring and all spots settings. It is clear that the *CPR* of $W^{(G)}$ and $W^{(L)}$ change with α and when the parameter value increases the *CPR* settles down. In the same figure, we also highlighted the *CPR* for both $W^{(G)}$ and $W^{(L)}$ using the maximum class method (M_1) as a green line, and the unweighted local mode method (M_2) in a blue line. Sometimes the *CPR* of the weighted mode method is identical to the *CPR* of either the maximum class or unweighted local mode method. For example, when $\alpha \geqslant 2$ in the $W^{(L)}$ image, the *CPR* of the weighted mode (M_3 and

M_4) and maximum class methods (M_1) are identical, where *CPR* is equal to 44%.

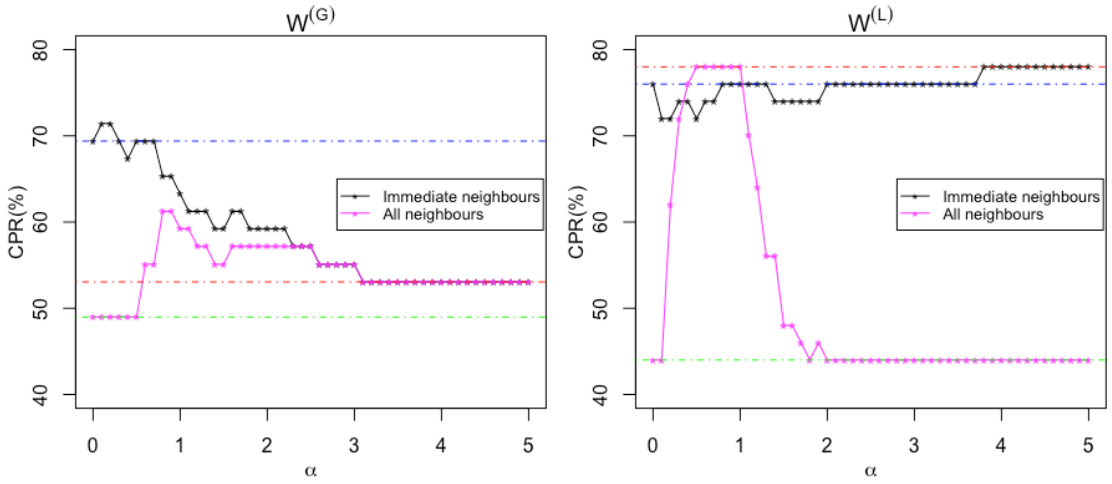


Figure 5.4: The *CPR* of $W^{(G)}$ and $W^{(L)}$ for image# 105420 using the weighted mode method for immediate neighbours (M_3) in black line and all spots (M_3) in pink line over α. Also, the *CPR* of M_1 in green lines, M_2 in blue lines and M_5 in red lines.

M_5: Predicting $W^{(Y)}$ when $\alpha \longrightarrow \infty$ using immediate neighbour

When $\alpha \longrightarrow \infty$ in Equation (5.10), w_{ij} tends to zero (see also Figure 5.3). Thus the prediction of spot classes are the same if we consider either immediate neighbourhood or all spots setting. For the i^{th}, we consider only the nearest spot in Y. Table 5.8 shows the

Table 5.8: Tables of agreements between the original classes ($\tilde{c}$) and predicted classes ($\hat{\tilde{c}}$) of $W^{(G)}$ and $W^{(L)}$ images predicted by M_5 with corresponding *CPR* using the image# 105420.

$W^{(Y)}$	$\hat{\tilde{c}}$	$\tilde{c}$			Total	*CPR*
		0	1	2		
$W^{(G)}$	0	8	4	1	13	53%
	1	8	15	5	28	
	2	0	5	3	8	
	Total	16	24	9	49	
$W^{(Y)}$	$\hat{\tilde{c}}$	$\tilde{c}$			Total	*CPR*
		0	1	2		
$W^{(L)}$	0	20	2	0	22	78%
	1	3	17	2	22	
	2	1	3	2	6	
	Total	24	22	4	50	

prediction by nearest spots method of image# 105420 for $W^{(G)}$ and $W^{(L)}$ images, where the *CPR* of $W^{(L)}$ image is better than $W^{(G)}$, by 25%. In $W^{(G)}$ image, the class 1 is the

best predicted spot class by 36% out of 53%, whereas in $W^{(L)}$ the better predicted class is class 0 with 40%. Figure 5.4 also shows the *CPR* of the immediate neighbours method in red line. In the following section all images are considered to equality assessment of spatial prediction methods.

5.4.2 Comparisons of spatial prediction methods

The five spatial prediction processes for low-resolution images from high-resolution images are explained with an application on a single image in Section 5.4.1. The equality of these methods is now assessed using all provided images. However, the optimal values of α for the weighted mode prediction method of immediate (M_3) and all spots settings (M_4) should firstly be determined. For each low-resolution image ($W^{(G)}$, $W^{(L)}$ and $W^{(LG)}$), a cross-validation technique is used to choose values of α for M_3 and M_4. After choosing α, a pairwise t-test is used to assess the equality of all prediction methods.

We start by defining the statistical methodologies, which are a cross-validation technique and a pairwise t-test, that are used followed by applications. The cross-validation method is a standard resampling technique which bases on leave out an image, whereby an image is excluded, the value of α which maximises the *CPR* is estimated and then this value of α is used to predict the *CPR* for excluded image (Bro et al., 2008). The steps of this technique are illustrated in Algorithm 5.

To explain how the algorithm works, let us consider the set of $W^{(G)}$ images called $\boldsymbol{x}$ of length 66 and we would like to estimate the smoothing parameters of the two neighbourhood settings and then calculate the corresponding *CPR* using M_3. We start by considering a range of α values, say $0 < \alpha < 3$, of length m. Each time an image is left out called x_{out} and the remaining number of images is then 65 images. For the remaining set of images, the *CPR* is calculated for each image using all α values. The output is $B_{65\times m}$, where each column is the *CPR* values of all images for a specific α. From this matrix, the mean of the *CPR* values for each column of α are calculated. Then we find which value of α has maximum *CPR* called $\hat{\alpha}$. Now using this $\hat{\alpha}$, the *CPR* of x_{out} is calculated called cpr. This process is repeated for all 66 images of $W^{(G)}$ to generate $\boldsymbol{cpr}$ using the prediction method M_3. We follow the same step to generate $\boldsymbol{cpr}$ of $W^{(G)}$ set using M_4. Likewise the $\boldsymbol{cpr}$ of $W^{(L)}$ and $W^{(LG)}$ are calculated for both M_3 and M_4.

Algorithm 5: Cross validation method for estimating $\boldsymbol{cpr}$ from a list of $\boldsymbol{x}$ images for given neighbouring setting.

Cross-validation $(\boldsymbol{x}, N)$;
Input : List of $W^{(Y)}$ images $\boldsymbol{x}$ of length n and neighbouring setting N
output: $\boldsymbol{cpr}$
Set $\boldsymbol{\alpha} =: (0, 0.1, 0.2, \ldots, 3)$;
$m =: |\boldsymbol{\alpha}|$;
for $i = 1$ *to* n **do**
    Set $x_{\text{out}} =: x_i$;
    Set $x_{\text{in}} =: x_1, \ldots, x_{i-1}, x_{i+1}, x_n$;
    Define $B_{(n-1)\times m}$;
    for $k = 1$ *to* $(n-1)$ **do**
        $W^{(Y)} = x_{\text{in}}[k]$;
        for $j = 1$ *to* m **do**
            For given α_j and N, predict the class of $W^{(Y)}$ from Equation (5.10);
            For given observed and predicted $W^{(Y)}$;
            Calculate CPR_j from Equation (5.8);
        end
        $B[k,] \leftarrow \boldsymbol{CPR}$;
    end
    $\hat{\alpha} =: \alpha\left[\arg\max \frac{\sum_{k=1}^{n-1} B[k,]}{n-1}\right]$;
    For given $\hat{\alpha}$ and N, predict the classes of x_{out} from Equation (5.10);
    For given observed and predicted x_{out};
    Calculate cpr_i from Equation (5.8);
end
$\boldsymbol{cpr} =: cpr_1, \ldots, cpr_n$;
return $\boldsymbol{(cpr)}$;

Figure 5.5 shows the mean of *CPR* from a single step of the cross-validation technique for $W^{(G)}$, $W^{(L)}$ and $W^{(LG)}$ images.

As a result from the cross-validation method with considering all images, the optimal parameter values ($\hat{\alpha}$) of $W^{(G)}$ image for M_3 are found to be one of 0.8, 0.9 and 1, and for M_4 are 0.7 and 0.8. For $W^{(L)}$, the $\hat{\alpha}$ of M_3 equals 0.3, but $\hat{\alpha}$ is in the range 0.8-2.5 when M_4 is used for prediction. The estimated smoothing parameters for $W^{(LG)}$ are 0.8 and 0.9 when we predict by M_3 and 0.7 when M_4 is used.

The $\boldsymbol{cpr}$ of spatial prediction methods M_1, M_2 and M_3, which have a fixed setting of α, is also calculated. The distributions and means of each $\boldsymbol{cpr}$ for each method and each image type are shown in Figure 5.6 and Table 5.9 respectively. In this table, the highest means are highlighted in red.

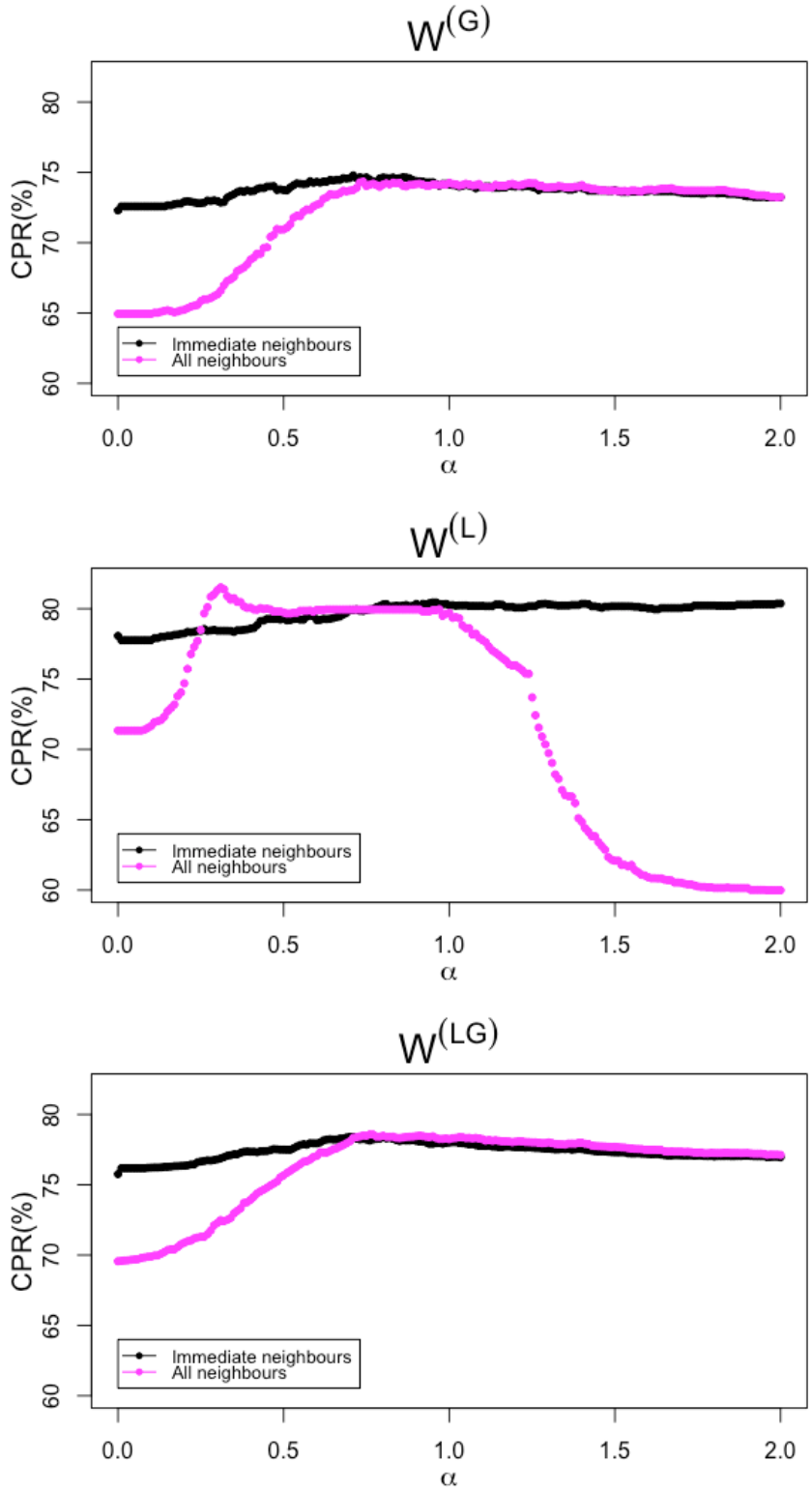


Figure 5.5: The mean of *CPR* for $W^{(G)}$, $W^{(L)}$ and $W^{(LG)}$ including all images over α except one excluded image using the weighted mode method for immediate neighbours (M_3) in black line and all spots (M_3) in pink line, where the number of images for pairs $W^{(G)}$ and $W^{(L)}$ is 65 and for $W^{(LG)}$ is 201.

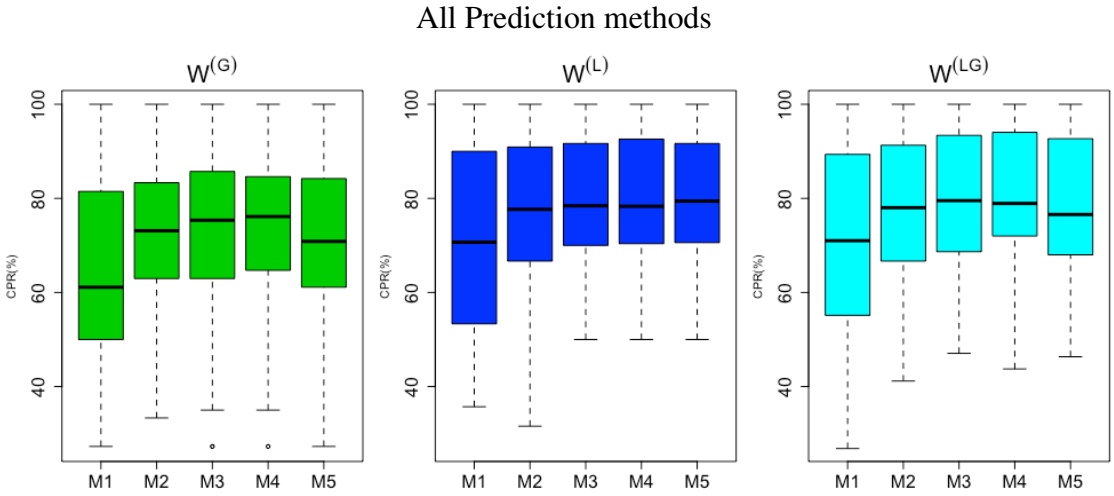


Figure 5.6: Boxplots of *CPR* for $W^{(G)}$, $W^{(L)}$ and $W^{(LG)}$ images using five prediction methods, where the number of images in pairs of $W^{(G)}$ and $W^{(L)}$ is 66 images and 202 images of $W^{(LG)}$.

Table 5.9: Means of *CPR* for different prediction methods for $W^{(G)}$, $W^{(G)}$ and $W^{(G)}$ images, where the highest means are in red.

$W^{(G)}$				
$\overline{\boldsymbol{cpr}}(M_1)$	$\overline{\boldsymbol{cpr}}(M_2)$	$\overline{\boldsymbol{cpr}}(M_3)$	$\overline{\boldsymbol{cpr}}(M_4)$	$\overline{\boldsymbol{cpr}}(M_5)$
65.95%	72.30%	72.67%	73.92%	71.80
$W^{(L)}$				
$\overline{\boldsymbol{cpr}}(M_1)$	$\overline{\boldsymbol{cpr}}(M_2)$	$\overline{\boldsymbol{cpr}}(M_3)$	$\overline{\boldsymbol{cpr}}(M_4)$	$\overline{\boldsymbol{cpr}}(M_5)$
71.02%	77.86%	79.57%	79.22%	79.76%
$W^{(LG)}$				
$\overline{\boldsymbol{cpr}}(M_1)$	$\overline{\boldsymbol{cpr}}(M_2)$	$\overline{\boldsymbol{cpr}}(M_3)$	$\overline{\boldsymbol{cpr}}(M_4)$	$\overline{\boldsymbol{cpr}}(M_5)$
71.51%	77.67%	79.79%	80.01%	78.36%

To assess all spatial prediction methods, pairwise comparisons between all possible paired methods, with corrections for multiple testing to obtain adjusted p-values, are performed. The used adjustment method is Bonferroni correction, which simply divides the Type I error rate (0.05) by the number of tests (McDonald, 2009).

A single paired t-test is explained to assess whether pairs of prediction methods are equally predicting low-resolution images. For $W^{(LG)}$, let $cpr_i(M_1)$ denote the *CPR* values of prediction method M_1 and $cpr_i(M_2)$ denote the *CPR* values of prediction method M_2, where $i = 1, \ldots, n$. The null hypothesis states that the true mean difference is zero, the differences are calculated $d_i = cpr_i(M_1) - cpr_i(M_2)$. Then, the mean ($\bar{d}$) and standard deviation (d_{sd}) of the differences, are calculated. The statistical test under H_0 is

Table 5.10: P-values of pairwise t-test for comparing *CPR* of pairs of prediction methods in $W^{(G)}$, $W^{(G)}$ and $W^{(G)}$ images, where the significant p-values are in red.

$W^{(G)}$				
	M_1	M_2	M_3	M_4
M_2	0.00	-	-	-
M_3	0.00	1.00	-	-
M_4	0.00	0.61	0.67	-
M_5	0.00	1.00	1.00	0.12

$W^{(G)}$				
	M_1	M_2	M_3	M_4
M_2	0.00	-	-	-
M_3	0.00	1.00	-	-
M_4	0.00	1.00	1.00	-
M_5	0.00	0.47	1.00	1.00

$W^{(LG)}$				
	M_1	M_2	M_3	M_4
M_2	0.00	-	-	-
M_3	0.00	0.11	-	-
M_4	0.00	0.01	1.00	-
M_5	0.00	1.00	0.37	0.03

defined as

$$T = \frac{\bar{d}}{SE(\bar{d})},$$

where $SE(\bar{d}) = d_{sd}/\sqrt{n}$. This test follows a t-distribution with $n-1$ degrees of freedom. A two-sided p-value of the single paired t-test is then calculated by comparing T to the t_{n-1} distribution from tables. Table 5.10 shows the pairwise t-test for all possible pairs of prediction methods with considering the adjustment of p-values. For predicting $W^{(G)}$ and $W^{(L)}$, on average, all methods are equally predicted except for M_1 which is is different. Similarly, the *CPR* average of M_1 is different than other methods in addition to the *CPR* average of M_4 which is different than the *CPR* averages of both M_2 and M_5. Whereas, M_2 and M_3, on average, are equally predicting images like M_5.

As a result, the prediction methods of $W^{(G)}$ and $W^{(L)}$ are the same with slightly relative parameter values, in particular in M_3 and M_4. This differences due to the the structure of images are different, and hence the smoothing parameter can also be differ. The prediction methods in $W^{(LG)}$ is different than $W^{(G)}$ and $W^{(L)}$, this occurs because the sample size of this type of image is approximately 4 times bigger than the other

images. Therefore, all low-resolution images can be predicted by high-resolution images with different prediction methods. This leads to the these two image types being considered to be consistent spatially.

5.5 Discussion

In this chapter the rectal cancer dataset, which contains different sampling area of the whole tumor, is only used to assess the consistency of pairs of images. Two ways of consistent assessment for pairs of images have been consider. The consistency of the distribution of spot classes is investigated for any pairs of images to find if they are consistent. More interestingly, the consistency of pairs of images sharing the same location can be checked by spatial prediction.

Different methods of spatial prediction for pairs of overlapping images are explained and their predictions are assessed for various images. We found that the distribution of classes for the low-resolution images are consistent with high-resolution images. These pairs of images were also spatially smooth, where nearby points in image tend to have the same classes. This means, sampling the the whole image is essential, however there is less need to sample high-resolution images.

Chapter 6

Applications in Pathology

6.1 Introduction and motivation

Response to cancer treatment, which is the body's reaction to a specific treatment regimens, is varied and it could be determined by, for instance, the tumor spot density in the whole tumor ($TCD(W)$) (West et al., 2010b), or the proportion of tumor (POT) (West et al., 2010a), but they suggested that more investigation is needed for POT. Similarly, we would like to investigate whether the spatial information encapsulated in I helps to classify patients into different treatment groups. Also, Hale et al. (2016) showed that POT of the biopsy (Bx), can be used to predict the chemotherapy benefit for patients, however, chemotherapy treatment, in our datasets, has been randomly given. It is of no interest to predict a random event, whereas treatment might be predicted if it has been allocated based on a diagnosis. This chapter covers the pathologists statistical questions through the project by investigating the usefulness of the I statistic. An exploratory analysis for all clinical covariates (also called variables) is performed in Sections 1.3.1 and 1.3.2, however, the I statistic was not included.

Now the frequently asked questions are highlighted for both gastric and rectal cancer datasets with specific aims. An essential question is: "Is the I statistic associated with the survival time of patients?" as we need to determine if the I statistic can be beneficial as a prognostic tool. Moreover, "Is I associated with any clinical variables, in particular, the proportion of tumor (POT)?" which is commonly used by pathologists. Finding any associations with I may help to find which variable affects tumor heterogeneity.

Moreover, the classification of spots, into tumor and stroma, is determined by the pathologist as described in Section 1.4, yet this has changed a couple of times through the project. Therefore, we perform an investigation, using only the gastric cancer dataset, to explore whether the changing of spot classification affects the significance of survival curves of the I statistic. Finally, the tumor spot density in the whole tumor ($TCD(W)$), which has been defined in Section 1.3.2, is an essential measurement for pathologists, which requires sampling the whole tumor image (W) and then calculating the TCD. Hence, another clinical question is "Can the $TCD(W)$ be predicted from I, as well as all clinical variables among patients?" The aim here is to find if there is any relationship that could decrease tumor heterogeneity and might improve targeted treatment.

The I statistic is a continuous covariate but can be made discrete by grouping patients into subcategories (e.g., two I subgroups classified relative to the median). The aim of this division is to allow the use of various standard analyses, in particular survival analysis, and answering the pathologists' questions. In addition to make a guide for pathologists to use the I statistic as a diagnostic tool. There are three different cutoffs that have been used to divide I into different subgroups. The first partitions I according to the median value classifying I into two sets, where $I_M = 0$ if I is less than or equal to the median, otherwise $I_M = 1$. The second division, divides the sorted values of I into three equal groups, called I_T, so that each group contains the same number of patients. Another classification divides I into three groups depending on its significance using the statistical test in Section 2.3.2. If we have a significant negative I with regular pattern, $I_S = 0$, if we have a random image, $I_S = 1$ and otherwise $I_S = 2$ when we have positive I with a significant clustering pattern. The proportion of tumor is also partitioned by the median which gives the binary variable POT_D.

Several survival techniques for modelling are considered, which are parametric, non-parametric and semi-parametric, in order to investigate the benefit of the I statistic for gastric and rectal cancers datasets. The main objective from modelling is investigating if the I statistic, I_M, I_T,or I_S, helps to predict the survival time of the patients. All clinical variables are also considered, and a variable selection procedure is applied. If the model included the I statistic, the model is highlighted and then its goodness of fit is assessed, otherwise the model is not relevant. To predict I and $TCD(W)$, we can simply fit a multiple regression model using only the main effects. The best model is

checking based on Akaike Information Criterion (AIC) (Akaike, 1973). Thus we choose the model that has the smallest AIC value using a stepwise selection method. Then, the goodness-of-fit is assessed by testing the randomness of residuals for the fitted model.

This chapter has been divided into four main sections. Some background about survival analysis and model selection and diagnostic is reviewed in Section 6.2. Sections 6.3 and 6.4 contain the analysis which is related to gastric cancer and rectal cancer, respectively. The conclusions and findings from this chapter are given in Section 6.5.

6.2 Survival analysis and model selection

Survival analysis can be defined as modelling of the time to death. The probability distribution of survival time can be either assumed to follow a particular form or to be distribution-free. Our aim is not predicting survival, but to use various survival models and compare the survival curves statistically to assess, in particular, whether there is a significant association between the I statistic and time to event. Finding any survival modes in which the I statistic is involved may help the pathologists in patient diagnosis. However, all covariates will also be considered in the assessment and the best model is selected based on a stepwise selection procedure. The diagnosis of the goodness for the fitted model is also obtained, however, the comparison between different survival models for the same dataset is not checked. If the model included the I statistic and well-fitted, the interpretation of the model is presented. Most definitions and models are drawn from Chatterjee and Chatterjee (2010); Collett (1994); Klein and Moeschberger (1997); Lee (2003); Parmar and Machin (1995); Rodríguez (2010).

Suppose T is a continuous non-negative random variable which is the survival time. Suppose the random variable T follows a distribution with a probability density function $f(t)$. Let $F(t)$ be the cumulative distribution function, i.e. $F(t) = P(T < t)$. The survival function $S(t)$ is given by

$$S(t) = P(T > t) = 1 - F(t). \tag{6.1}$$

Equation (6.1) gives the probability that a subject will survive past time t. A hazard function, $h(t)$, which is the instantaneous rate at which events occur, is defined mathe-

matically by

$$h(t) = \frac{f(t)}{S(t)}. \tag{6.2}$$

A functional form from the hazard function can be an alternative approach which can, alternatively, be determined from

$$S(t) = \exp\big\{ - H(t)\big\}, \tag{6.3}$$

where $H(t) = \int_0^t h(u)du$ denotes the cumulative hazard, which can be obtained from the survivor function, since $H(t) = -\log S(t)$.

From Equation (6.2) it is clear that if one of $h(t)$,$f(t)$ or $S(t)$ is known, the others can be calculated. These functions can be estimated using three classes of survival analysis models: parametric, non-parametric and semi-parametric. All models used will now be briefly described.

6.2.1 Parametric survival models

Parametric approaches are methods in which we make distributional assumptions about the survival times. Suppose ε is a random variable with a specific distribution on $(-\infty, \infty)$. For different individual, this random variable is assumed to be independent and identically distributed with known forms of density function $g(\varepsilon; \boldsymbol{d})$ and survivorship function $G(\varepsilon; \boldsymbol{d})$ but unknown parameters $\boldsymbol{d}$. The $G(\varepsilon; \boldsymbol{d})$ can be generated by introducing location and scale of the form

$$\log T = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \eta\varepsilon, \tag{6.4}$$

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients, x_j, $j = 1, \ldots, p$, are the covariates, η is an unknown scale parameter and T is the survival time (Lee, 2003; Rodríguez, 2010). Equation (6.4) is the general form of accelerated failure time (AFT) which describes contraction of survival time as a function of independent variables. The model (6.4) can also be expressed in term of survival time as

$$T = \exp\Big\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \eta\varepsilon\Big\}, \tag{6.5}$$

where the respective alternative density and survival functions are $f(t;\boldsymbol{b})$ and $S(t;\boldsymbol{b})$, respectively. The coefficients to be estimated and regression parameters are $\boldsymbol{b}=(\beta_0,\boldsymbol{\beta},\eta)$. The MLE of $\boldsymbol{b}$ is a set of $b_1,\ldots,b_k$, that maximise $l(\boldsymbol{b})$. The MLE of $\hat{\boldsymbol{b}}$ must be obtained by a numerical method as there is no closed form solution. The commonly used numerical method is the Newton-Raphson iterative procedure (also known as Newton's method). For more information see Lee (2003).

We now review five different distributional assumptions, which have been used, for the error term ε in model (6.4). This term can be, for instance, assumed to follow standard normal distribution which is similar to assume that T has a log-normal distribution in Equation 6.5. The five distributions are Weibull, exponential, log-logistic, logistic and log-normal. The parametric model can be fitted using the survival package `survreg` for any distribution. The distributions used are briefly defined with their $h(t)$,$f(t)$ or $S(t)$ functions in Table 6.1.

Table 6.1: Commonly used distributions for parametric survival models with corresponding probability density functions, survival functions, hazard rates and model parameters.

Distribution T	$f(t)$	$S(t)$	$h(t)$	Parameters
Weibull	$\lambda\gamma t^{\gamma-1}\exp\{-\lambda t^{\gamma}\}$	$\exp\{-\lambda t^{\gamma}\}$	$\lambda\gamma t^{\gamma-1}$	λ,γ
	$\gamma,\lambda>0,t\geq 0$			
Exponential	$\lambda\exp\{-\lambda t\}$	$\exp\{-\lambda t\}$	λ	λ
	$\lambda>0,t\geq 0$			
Log-logistic	$\frac{\lambda\gamma t^{\gamma-1}}{(1+\lambda t^{\gamma})^2}$	$\frac{1}{1+\lambda t^{\gamma}}$	$\frac{\lambda\gamma t^{\gamma-1}}{1+\lambda t^{\gamma}}$	λ,γ
	$\gamma,\lambda>0,t\geq 0$			
Logistic	$\frac{\exp\{t\}}{1+\exp\{t\}}$	$\frac{1}{1+\lambda t^{\gamma}}$	$\frac{(1+\lambda t^{\gamma})\exp\{t\}}{1+\exp\{t\}}$	λ,γ
	$\gamma,\lambda>0,t\geq 0$			
Log-normal	$\frac{\exp\left\{-\frac{1}{2\sigma^2}(\log t-\mu)^2\right\}}{t\sigma\sqrt{2\pi}}$	$1-\Phi\left(\log\frac{at}{\sigma}\right)$	$\frac{f(t)}{S(t)}$	σ
	$\sigma>0,t\geq 0$			

where $a=\exp\{-\mu\}$ and $\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_0^y\exp\{-\frac{u^2}{2}\}du$.

First of all the Weibull distribution, with parameters λ and γ (both of them greater than zero), is the most popular assumption. The λ is known as a scale parameter, while the parameter γ is the shape parameter. The hazard function, $h(t)$, is increasing over time

if $\gamma > 1$, constant if $\gamma = 1$, and decreasing if $\gamma < 1$. The second model is the exponential distribution, which is the simplest model of hazard function, $h(t) = \lambda$, which is assumed to be constant over time. The λ parameter is a positive constant which can be estimated by fitting the model to the observed data. The third model is the log-logistic with λ and γ parameters, and similarly the logistic. Finally, if $\varepsilon \sim N(0, 1)$, T has a log-normal distribution with σ parameter. Survival time is a continuous response in all distributions, but if the survival time is a discrete variable, the logistic distribution can also accept discrete response times.

6.2.2 Non-parametric survival models

Non-parametric survival models can be explained by the empirical probability of surviving past certain times obtained in the sample. This model has no distributional assumption required but it is a univariate method which requires categorical covariates, thus the discretised I (I_M, I_T and I_S) are used. In this section, the Kaplan-Meier (KM) method (Kaplan and Meier, 1958) is used to illustrate and plot the survival curves from lifetime data for each individual variable, but the log-rank test (Harrington, 1982) is used to compare between KM survival curves to detect if they are statistically different. The survival curves of the KM estimator are plotted using the survival package `survfit` function, and the log-rank test is applied using the survival package `survdiff` function to compare survival curves between specified groups. To compare statistically between curves, the log-rank test is used.

Table 6.2: At the j^{th} death time, number of deaths in each of two groups (Collett, 1994).

Group	Number of deaths at t_j	Number surviving beyond t_j	Number at risk just before t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

A log-rank test is used to compare between survival functions from different groups. Considering two groups of treatment, group 1 and group 2, the log-rank test is constructed as follows. We assume death times are independent in both groups and r are distinct death times recorded to the nearest time of death, $t_1 < t_2 < \cdots < t_r$, across the

two groups, so at time t_j, d_{ij} individuals in the i^{th} group, $j = 1, 2, \ldots, r$ and $i = 1, 2$. Suppose also that there are n_{ij} individuals at risk of death in the i^{th} group before time t_j. As a consequence, at time t_j, there are $d_j = d_{1j} + d_{2j}$ deaths out of n_j, where $n_j = n_{1j} + n_{2j}$ individuals are at risk (see also Table 6.2). Sometimes it is possible to have two patients die at the same time. This rarely occurs in the gastric cancer dataset as the time is recorded by day, but we could have a multiple event in the rectal cancer dataset because the time of death is recorded by the nearest month of death.

Now we consider the null hypothesis, H_0, that there is no difference between two survival functions. In order to assess the validity of this hypothesis we consider the difference between the observed number of dead individuals in the two groups at each of the death times. Collett (1994) explained that d_{1j} in Table 6.2 has a hypergeometric distribution, according to which the probability that the random variable associated with the number of death in group 1 takes the value d_{1j} is

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}, \tag{6.6}$$

where

$$\binom{x}{y} = \frac{x!}{y!(x-y)!}.$$

The mean of the hypergeometric random variable d_{ij} is

$$e_{ij} = n_{ij} d_j / n_j.$$

Next, we sum the differences $d_{1j} - e_{1j}$ over all r death times, in the first and second group of treatments

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}).$$

The variance of U_L is the sum of the variances of the d_{1j}, because the death times are independent of one another. Now, as d_{1j} has a hypergeometric distribution, the variance of d_{1j} is

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)},$$

therefore, the variance of U_L is

$$\text{var}(U_L) = \sum_{j=1}^{r} v_{1j} = V_L.$$

When the number of death times is large, U_L is approximately normally distributed. It then follows that, when the null hypothesis is true

$$U_L/\sqrt{V_L} \sim N(0,1),$$

the square of a standard normal random variable has a chi-squared distribution on one degree of freedom (dof$= 2-1=1$) under H_0

$$W_L = \frac{U_L^2}{V_L} \sim \chi_1^2.$$

The test based on this statistic is called the log-rank test. The larger the value of this statistic, the greater the evidence against the null hypothesis in favour of the alternative that the two treatment groups are not equally effective. The corresponding p-value of this statistic can be obtained from the distribution function of a chi-squared random variable.

6.2.3 Semi-parametric survival models

The Cox proportional hazard (PH) model (Cox, 1972) is one of predominant semi-parametric survival models, where the distribution of survival times is unknown. This model investigates the association between the survival time of patients and one or more independent variables. The Cox PH model works for quantitative and categorical variables and in addition can be extended to assess the effect of several risk factors on survival time. The Cox PH model is fitted using the survival package `coxph` function. The standard Cox proportional hazards model can be written as

$$h(t) = h_0(t)\exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}, \tag{6.7}$$

where $h(t)$ is the hazard function at time t, which is determined by a set of p variables, $h_0(t)$ is called the baseline hazard which illustrates the hazard when $x_1 = x_2 = \cdots =$

$x_p = 0$ and the coefficients $\beta_1, \ldots, \beta_p$ measure the impact of the variables. The $\exp(\beta_1)$, $\ldots$, $\exp(\beta_p)$ are also called hazard ratios (or relative risks). Coefficients and hazard ratios interpretations will be compared for each dataset. The coefficients estimation can be obtained without specifying the baseline hazard $h_0(t)$ by maximising the partial likelihood using the Newton-Raphson algorithm.

6.2.4 Model selection and diagnosis

In modelling, the essential variables need to be selected and then the optimal model is assessed. This section explains the process of selecting variables to determine the optimal model, in addition to diagnosing its goodness-of-fit.

Choosing the optimal model is determined by Akaike Information Criterion (AIC) which is a measure of goodness-of-fit. This statistical process of model selection is based on the log-likelihood $l(\hat{\boldsymbol{b}})$ for the fitted model, where $\hat{\boldsymbol{b}}$ refers to the parameters of the model. The AIC is computed as

$$\mathrm{A}IC = -2l(\hat{\boldsymbol{b}}) + 2k, \tag{6.8}$$

where k is the number of parameters in the model. A lower AIC value indicates a better model fit. The computation of the AIC statistics is difficult to obtain for all possible models with various variable settings due to computational efficiency. Thus the stepwise regression method using both forward and backward elimination is applied to compute the AIC statistics. The stepwise algorithm estimates the quality of each model, relative to each of the other models in order to choose which model has the best fit.

To assess the appropriateness of the linear regression model, residuals are defined and their plot examined. The residuals are defined as the difference between the observed value of the dependent variable, say y, and the predicted value ($\hat{y}$), which can be written mathematically as $e_i = \hat{y}_i - y_i$,$i = 1, \ldots, n$, where each observation (or patient) has one residual. In the case where the points in a residual plot are randomly dispersed around the horizontal axis, the model is well-fitting.

Nevertheless the standard residual-based measures of multiple regression are inappropriate for checking the survival time in parametric and semi-parametric models. To diagnose the survival model, after the model selection step, Cox-Snell residuals (Cox

and Erricker, 1968) are used. The procedure can be summarised as follows: Let $\hat{S}(t_i)$ denote the estimated survival function of the i^{th} individual. The Cox-Snell residuals are calculated as $r_i = -\log\{\hat{S}(t_i)\}, i = 1, \ldots, n$. Then we need to apply the Kaplan-Meier method to estimate the survival function $\hat{S}_R(r_i)$ of the Cox-Snell residual r_i, and calculate $-\log\{\hat{S}_R(r_i)\}, i = 1, \ldots, n$, which is the estimated cumulative hazard. Finally, we plot r_i against $-\log\{\hat{S}_R(r_i)\}$, and if the plot is close to a straight line with zero intercept and unit slope, the model is well-fitting. For more information see Collett (1994).

6.3 Gastric cancer

This section includes the analysis which is related to the gastric cancer dataset containing 223 patients. Variables in this dataset, which are used in this section, are defined in Table 1.1, and are as follows: pT of four stages is pathological tumor stage, JS of seven stages is the Japanese classification of tumor, LS of two stages is Lauren Classification of tumor and $chemo$ is a received chemotherapy indicator, where $chemo = 1$ indicates the patients who had no chemotherapy and $chemo = 2$ otherwise. The POT and the I statistic are also used as well as their partitioned versions POT_D, I_M, I_T and I_S, where the median of POT is 0.384 and I is 0.127. As the direction of lumen, defined in Section 3.3.3, was provided for this dataset, we can use the directional versions of the I statistic (I_1, I_2 and I_3). Only 218 images, however, have the indicator of direction to the lumen site, so only these images are considered.

This section begins with a survival time analysis using parametric, non-parametric and then semi-parametric models in Section 6.3.1. Then, we find if there are any associations between the I statistic and clinical variables in Section 6.3.2. Classification of spots, previously explained in Section 1.4, is also adjusted to include other possible classification of spots into tumor and stroma and then each option of classification is used to calculate the classified I statistics (I_M, I_T and I_S) in order to find if their significance of survival curves are changed.

6.3.1 Survival analysis

Three models of survival times are considered. We start with a parametric model, then non-parametric and semi-parametric models.

Firstly, in the parametric model, we need to determine the appropriate survival time distribution for the gastric cancer dataset, which was described in Table 6.1. To choose the appropriate distribution, we fit all five parametric models for a fixed covariates, where each variable is the only one in the model, and then the AIC is calculated. The AIC procedure is similar to those based on the likelihood function. Now all AIC values are shown in Table 6.3. By comparing the Weibull, exponential, log-logistic, logistic and log-normal models, we found that the log-normal has the smallest AIC for all variables. This distribution is now used to fit a parametric survival model for all clinical variables ($pT, JS, LS, chemo, POT$) in addition to including I, I_M, I_T, and I_S, which are separately added to the model.

Table 6.3: Comparison of survival models using Akaike Information Criterion (AIC) for each variable in turn, where lower AIC values indicate a better fit.

Variables	Models				
	Weibull	**Exponential**	**Log-logistic**	**Logistic**	**Lognormal**
pT	525	526	519	589	514
JS	531	531	526	595	521
LS	533	534	528	598	523
$chemo$	527	529	523	591	520
POT	531	532	526	595	521
I	532	532	527	597	522

We now have four possible parametric models with different versions of I. After the stepwise selection method, the I statistic and I_T and I_M were dropped from the survival models. We are left with only one model, which contains I_S as shown in Table 6.4 with AIC$= 511$. Some p-values of parameters ($pT = 2, 3$ and 4), however, are not significant, thus levels 1 and 2 as well as levels 3 and 4 of the pT covariate are merged as showing in the same table with AIC=514, but most of parameters are significant. The best parametric survival model can be expressed as

$$\log T = \beta_0 + \beta_1 I[pT = 2] + \beta_2 I[I_S = 2] + \beta_3 I[chemo = 2] + \eta\varepsilon, \qquad (6.9)$$

Table 6.4: The estimated coefficients with corresponding standard deviation, p-values and estimated scale parameter of the log-normal model for the gastric cancer dataset after a stepwise selection method.

Covariate	Estimated parameters	Sd	P-value
Intercept	7.886	943.664	0.993
$I[pT = 2]$	-5.843	943.664	0.995
$I[pT = 3]$	-5.606	943.664	0.995
$I[pT = 4]$	-6.319	943.664	0.995
$I[I_S = 2]$	0.385	0.217	0.076
$I[chemo = 2]$	-0.351	0.211	0.096
Scale(η)	0.190	0.086	
Intercept	2.103	0.316	0.000
$I[pT = 2]$	-0.434	0.286	0.129
$I[I_S = 2]$	0.437	0.216	0.043
$I[chemo = 2]$	-0.418	0.211	0.047
Scale(η)	0.197	0.087	

where $I[.]$ indicates a particular level of a discrete variable and η is a scale parameter. The estimated coefficients for the log-normal survival model for each variable are shown in Table 6.4, together with standard error and p-values to test the null hypothesis H_0 : $\beta_j = 0$. The significant coefficients have an important effect on the survival time, but none of p-values are significant. We can also check the model in Equation (6.9), using the Cox-Snell residuals plotted in Figure 6.1 showing that the model fit is unacceptable as there is serious deviation from the central line. Hence, the interpretation of this model is not included.

Secondly, the Kaplan-Meier non-parametric survival function is used to investigate the differences in survival curves. This model is fitted for each discrete variable individually where some examples are shown in Figure 6.2. There were significant differences between some survival curves within each variable, for instance, I_S, pT $chemo$. Let's compare the survival curves of two I_S groups. From Figure 6.2 (top-right plot), the horizontal axis represents time in years, and the vertical axis gives the probability of surviving. The two lines show survival curves of the two groups of I_S. The survival probability for patients is 100% at time zero. At year 2, the probability of survival is approximately 73% for $I_S = 1$ and 81% for $I_S = 2$. The median survival is approximately 4 years for $I_S = 2$ and zero years for $I_S = 1$. That means patients in group $I_S = 2$ had better survival than those for $I_S = 1$.

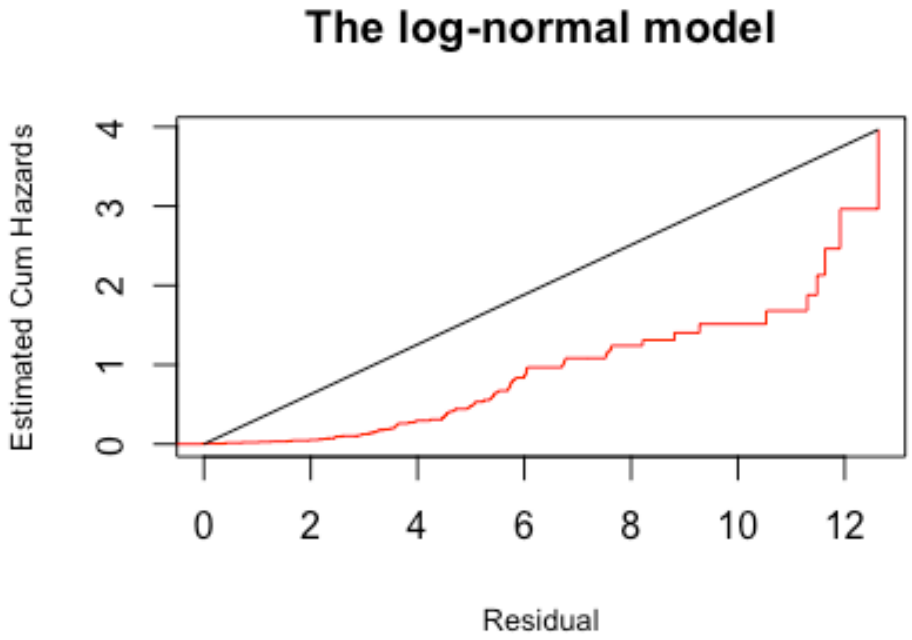


Figure 6.1: Cox-Snell residuals to assess the fit of the log-normal regression model in Equation (6.9) for gastric cancer dataset using, where the red line shows r_i against $-\log\{\hat{S}_R(r_i)\}$.

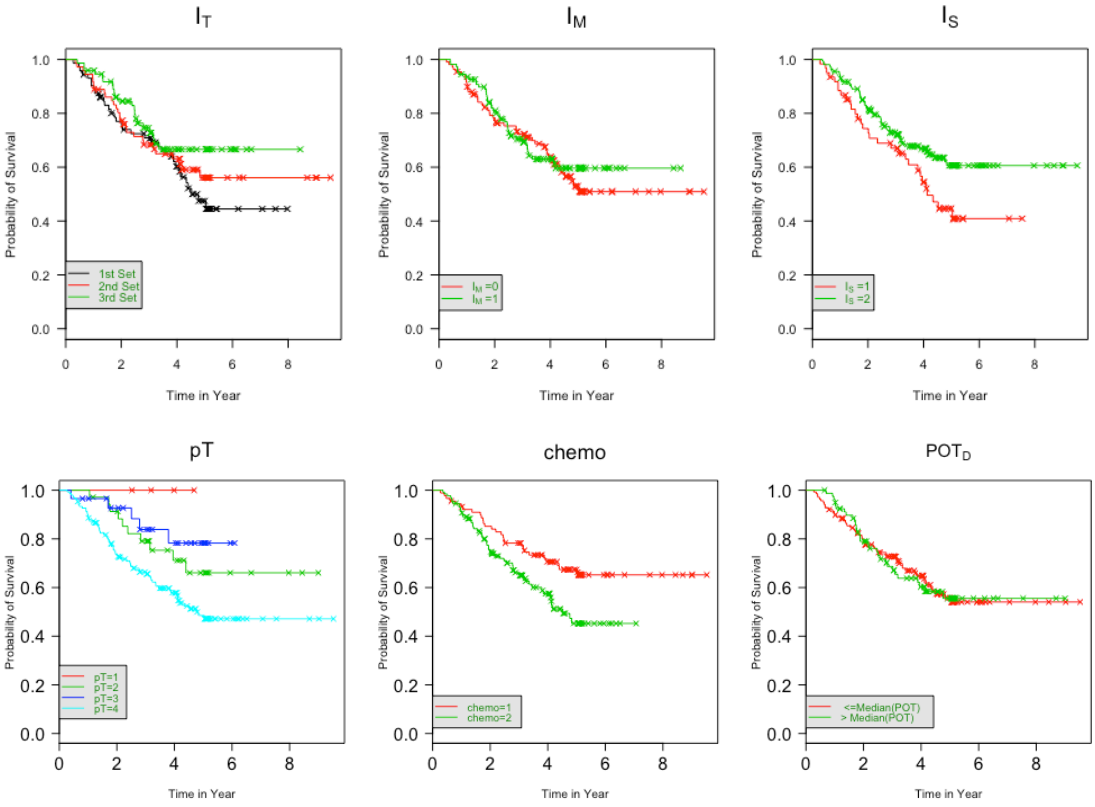


Figure 6.2: The Kaplan-Meier survival curves in the gastric cancer images for the classified I statistic (I_M, I_T and I_S), tumor stage pT, treatment type $chemo$ and classified POT (POT_D).

The significant differences between survival curves in Figure 6.2 can be confirmed using the log-rank test in Table 6.5, where the p-values less than 0.05 indicating that we can reject the null hypothesis H_0 : there is no difference between two survival curves. However, I_T, I_M, JS, LS and POT_D have similar survival curves.

West et al. (2010a) also found that the p-value of pT was significant (p-value= 0.0001). However, they found that there is a significant difference in survival times between low-POT and high-POT, but their median of POT value was 57.1%, whereas the median of POT in our data set is relatively low, 38.3%.

Table 6.5: The chi-squared statistic of the log-rank test with corresponding degrees of freedom and p-values for each discrete variable for the gastric cancer dataset.

Variables	Chi-square	Dof	p-value
I_T	2.9	2	0.20
I_M	2.9	1	0.60
I_S	4.4	1	0.04
pT	10.3	3	0.02
JS	8.0	6	0.20
LS	0.4	1	0.50
$chemo$	5.9	1	0.02
POT_D	0.0	1	0.90

In addition to the I statistic, the directional versions (I_1, I_2 and I_3) in Section 3.3.3 are also used in survival analysis. A new measurement, called $I(R)$, is calculated for 218 patients defined as

$$I(R) = \max(I_1, I_2, I_3) - \min(I_1, I_2, I_3).$$

The objective from calculating I_R measurement using the directional I statistics is investigating if heterogeneity affects the survival time and useful in patient dignosis. The I_R measurement is then classified using the same method as I which we call $I_M(R)$ and $I_T(R)$. The long-rank test is applied for the division of I_R shown in Table 6.6. Here there is no significant difference between survival curves of $I_M(R)$ and $I_T(R)$.

Table 6.6: The chi-square statistic of the log-rank test with their degrees of freedom and p-values for the divisions of the $I(R)$ statistic for directions.

Original I stat	classified I	Chi-square	Dof	p-value
$I(R)$	$I_T(R)$	0.3	2	0.9
	$I_M(R)$	0.5	1	0.9

Finally, the Cox PH regression is fitted, where all variables pT, JS, LS, $chemo$, POT are included in addition to the I statistic and its partitioned versions I_M, I_T and I_S which have been separately added. However, level 1 and 2 in pT covariate are merged as the number of patients in level 1 is low and the Cox PH model does not accept a low number of patients in any group, and thus the pT now has three levels: $2, 3, 4$, where $pT = 2$ is the default group. The best model is then selected by the stepwise method, which has the lowest AIC$= 807$, and can be expressed as

$$h(t) = h_0(t) \exp\Big\{\beta_1 I[pT = 3] + \beta_2 I[pT = 4] + \beta_3 I[I_S = 2] + \beta_4 I[chemo = 2]\Big\}. \tag{6.10}$$

Table 6.7: Cox PH model for the gastric cancer dataset shown in Equation (6.10).

Covariate	$\hat{\beta}$	$\exp\{\hat{\beta}\}$	**Sd($\hat{\beta}$)**	**P-value**
$I[pT = 3]$	-0.380	0.683	0.549	0.489
$I[pT = 4]$	0.579	1.784	0.343	0.091
$I[I_S = 2]$	-0.466	0.627	0.232	0.044
$I[chemo = 2]$	0.522	1.686	0.239	0.029

The estimated coefficients with their corresponding exponential, standard error and p-values, to test the null hypothesis that $H_0 : \beta_j = 0$, are shown in Table 6.7. From this table the significant coefficients on the survival time are $I_S = 2$ and $chemo = 2$. The estimation of the baseline hazard function $h_0(t)$ is the estimated hazard of death at time T for an individual whose I is random and who has not had chemotherapy treatment. To assess the goodness-of-fit for model (6.10), Cox-Snell residuals are plotted in Figure 6.3. It is clear that the Cox PH model is appropriate as the plot of the residual is close to the straight line.

The interpretation is that holding the other covariates constant, being a patient with clustered image ($I_S = 2$) reduces the hazard by a factor of 0.627 or 37.3%. While patients who had chemotherapy ($chemo = 2$) have a hazard ratio 1.686 indicating an increased risk of death by 68.6% compared with patients who had no treatment. As a result, the survival time for the gastric cancer dataset can be significantly predicted using I_S and non- and semi-parametric survival models, where the clustered images are associated with higher survival time and we can say that being clustered images are associated with a good prognostic.

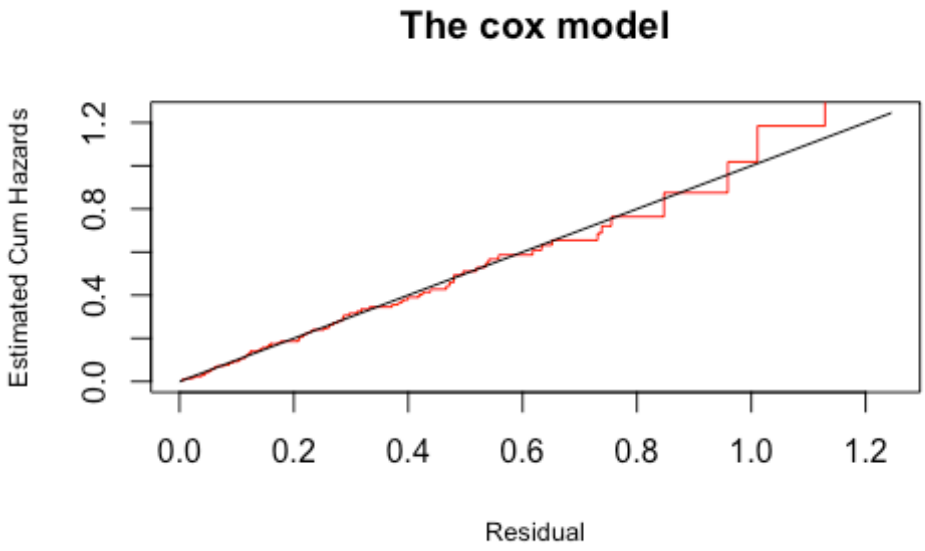


Figure 6.3: Cox-Snell residual plot for the gastric cancer dataset using the parametric model in Equation (6.9), where r_i against $-\log\{\hat{S}_R(r_i)\}$ is red line.

6.3.2 Predicting the *I* statistic

The I statistic may be considered as a measure of image heterogeneity. The aim for this section is to find how I is associated to clinical variables including POT. To describe the relationship between a set of clinical predictors and response I, we simply fit a multiple regression model. The survival variables are excluded as they have already been considered in Section 6.3.1. The clinical variables are pT, JS, LS, $chemo$ and POT.

Table 6.8: The estimated coefficients with their corresponding standard error and p-values of the fitted multiple regression model in Equation (6.11) for the gastric cancer dataset.

Covariate	$\hat{\beta}$	**Sd**($\hat{\beta}$)	**P-value**
Intercept	0.179	0.090	0.047
$I[pT = 2]$	0.030	0.066	0.644
$I[pT = 3]$	0.025	0.067	0.706
$I[pT = 4]$	−0.017	0.064	0.780
$I[JS = 2]$	−0.021	0.063	0.737
$I[JS = 3]$	0.020	0.057	0.719
$I[JS = 4]$	0.237	0.115	0.041
$I[JS = 5]$	0.125	0.116	0.283
$I[JS = 6]$	0.257	0.119	0.032
$I[JS = 7]$	0.057	0.073	0.436
$I[LS = 2]$	−0.194	0.099	0.052
$I[chemo = 2]$	0.031	0.017	0.084
POT	−0.107	0.046	0.022

After fitting the regression model and then applying the stepwise selection procedure, the best model with the smallest AIC value, equals -282, as follows

$$\begin{aligned} I = \beta_0 &+ \beta_1 I[pT = 2] + \beta_2 I[pT = 3] + \beta_3 I[pT = 4] + \beta_4 I[JS = 2] + \beta_5 I[JS = 3] \\ &+ \beta_6 I[JS = 4] + \beta_7 I[JS = 5] + \beta_8 I[JS = 6] + \beta_9 I[JS = 7] + \beta_{10} I[LS = 2] \\ &+ \beta_{11} I[chemo = 2] + \beta_{12} POT + \varepsilon, \end{aligned} \tag{6.11}$$

where ε is a random error which is assumed to be normally distributed with mean 0 and variance σ^2, and $I[.]$ refers to a particular level in a covariate.

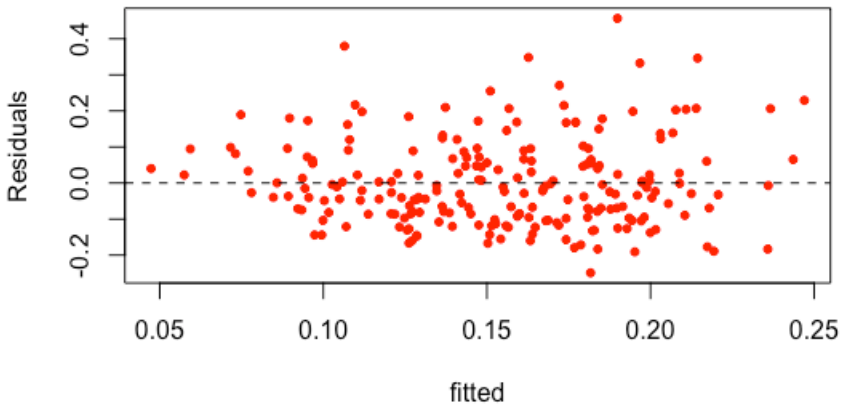


Figure 6.4: The residuals distribution of model in Equation (6.11)

However, none of the variables are dropped. The estimated regression coefficients β_i's together with their corresponding standard error and p-values are shown in Table 6.8. The null hypothesis of coefficient are $H_0 : \beta_i = 0, \ldots, 11$. A low p-value ($< 0.05$), implies that the null hypothesis can be rejected. To confirm the quality of the model, we look at the residual plot of this model in Figure 6.4. There is not much change in the overall pattern and residuals are randomly distributed. When none significant parameters are removed, for example pT, the AIC increased to -280 and the significant parameters are become not significant.

The interpretation of model (6.11), from Table 6.8, is as follows: the covariate that has a low p-value is likely to be meaningfully related to changes in the I statistic. For example, as POT increases by 1%, the I on average, decreased by 10.7%. Also, the

coefficients of JS for level 4 and 6 are positive which means that these levels have a higher I statistic than the default group ($JS = 1$). As a result, it is clear that there is a relationship between the I statistic and Japanese classification of tumor and proportion of tumor.

6.3.3 Sensitivity analysis of alternative allocation of spots

As discussed in Section 1.4, the pathologists recommended the way of grouping the spot types into two sets. However, the way of groping had been changed couple of times during the project. Using sensitivity analysis, several different ways of grouping the spots into two sets are considered to investigate the impact of spot allocation could affect the difference of the survival distributions. This section considers the classification of spots in Section 1.4 as well possible alternative allocation for the gastric cancer dataset.

Table 6.9: Different options of spot classification, where each of spot types 1, 2, 4, 5, 6 and 8 are defined as S (stroma) and T (tumor) and the highlighted grey column is the pathologists recommended classification.

Spot type	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_{16}
1	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
2	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
4	T	S	T	S	S	S	S	T	T	T	T	T	S	S	T	S
5	T	S	S	T	S	S	T	S	T	T	T	S	T	S	S	T
6	T	S	S	S	T	S	T	T	S	T	S	T	S	T	S	T
8	T	S	S	S	S	T	T	T	T	S	S	S	T	T	T	S

Pathologists have no doubt about both tumor and stroma spot classifications, which are spot types 1 and 2, in addition to excluding some spot classifications (0, 3 and 7). However, they sometimes reallocated other spot types, which are 4, 5, 6 and 8 spot classification, to be either a tumor or stroma. Considering every possible allocation of 4, 5, 6 and 8 that means there are 2^4 ways can be allocated to tumor or stroma. Here, we will investigate and experiment how different allocations could affect the log-rank test. Suppose the spot classification is denoted by O, so the different options of classification can be written as $O_1, \ldots, O_{16}$ (see Table 6.9), where O_1, for instance, means we consider the classification number 2 as stroma and the rest are tumor.

Table 6.9 displays all possible allocations, given that the tumor and stroma spot types are fixed, where O_5, highlighted in grey, is the allocation of spots which has been confirmed by pathologists in Section 1.4. All these different option sets of allocation have

Table 6.10: The p-values of log-rank test for different divisions of I statistics for 218 patients, where $O_k, k = 1, \ldots, 16$, are from Table 6.9, where the highlighted grey row is the pathologists recommended classification, where I_M refers to dividing the I statistic by median, I_T refers to dividing the I statistic into three equal groups and I_S refers to divide I into three groups depending on its significance

Classification option	P-value		
	I_M	I_T	I_S
O_1	0.662	0.104	0.008
O_2	0.232	0.237	0.121
O_3	0.346	0.243	0.116
O_4	0.188	0.083	0.002
O_5	0.599	0.298	0.044
O_6	0.530	0.366	0.157
O_7	0.630	0.232	0.004
O_8	0.959	0.604	0.062
O_9	0.064	0.137	0.002
O_{10}	0.484	0.103	0.005
O_{11}	0.070	0.171	0.002
O_{12}	0.637	0.451	0.029
O_{13}	0.164	0.120	0.004
O_{14}	0.761	0.271	0.060
O_{15}	0.326	0.335	0.208
O_{16}	0.654	0.216	0.003

been used to calculate the I statistic for 218 patients in Table 6.10, and we then applied the long-rank test to determine whether there is a significant difference between survival curves, partitioning I as before.

As a result, there are some option sets which have no significant differences in survival curves which are O_2, O_3, O_6, O_8, O_{14}, O_{15} and O_{16}. The I_M and I_T division of the I statistic show no significant values for all options of allocations. The significant p-values were only in I_S using O_1, O_4, O_5, O_7, O_9, O_{10}, O_{11}, O_{12}, O_{13} and O_{16} allocation options. As a result, we can say that the allocation options that are close in the result of log-rank test for all I_M, I_T and I_S are O_7, O_{12}, O_{16} which is close to O_5. Therefore, we can say that if the pathologists used either O_7, O_{12} and O_{16} or O_5 of allocation of the spots, they could have a similar result of log-rank test.

6.4 Rectal cancer dataset

The rectal cancer dataset includes multiple images per patient which are biopsy images (Bx), whole tumor images (W) and luminal site images (L). The clinical variables, defined in Section 1.3.2, are used in two types of survival times: follow-up (FU) and disease-free (DF) survival times. The clinical variables of 113 patients are $Pr.Tstage$ of three stages, which indicates pre-operative tumor assessment, pT of five stages, which shows the tumor stage, pN of two stages, which indicates lymph nodes stage, pM of two stages, which shows distant metastasis stage, $therapy$ of three types, which indicates the chemotherapy type, $Gender$, indicates the gender of patient, where $Gender = 1$ refers to male and 2 otherwise and Age, which is the age of patient. In addition to these variables, the tumour cell density of image type, which are $TCD(Bx)$, $TCD(W)$ and $TCD(L)$, are also defined. Extra variables, which we have calculated, are also defined in Table 6.11 to be used in the analysis of this section. This table includes the notation of proportion of tumor and the I statistic with different divisions for the three image types.

Table 6.11: Extra variables description of rectal cancer dataset, where Bx refers to biopsy image, W the whole tumor image and L lumen site image.

Variable name	Description
$POT(Bx)$	The proportion of tumor from Bx
$POT(W)$	The proportion of tumor from W
$POT(L)$	The proportion of tumor from L
$POT_D(Bx)$	Divide $POT(Bx)$ by median, where $POT_D(Bx) = 0$ if $POT(Bx) \leq$ median($POT(Bx)$) and 1 otherwise
$POT_D(W)$	Divide $POT(W)$ by median, where $POT_D(W) = 0$ if $POT(W) \leq$ median($POT(W)$) and 1 otherwise
$POT_D(L)$	Divide $POT(L)$ by median, where $POT_D(L) = 0$ if $POT(L) \leq$ median($POT(L)$) and 1 otherwise
$I(Bx)$	The I statistic of BX
$I(W)$	The I statistic of W
$I(L)$	The I statistic of L
$I_M(Bx)$	Divide $I(Bx)$ by median, where $I_M(Bx) = 0$ if $I(Bx) \leq$ median($I(Bx)$) and 1 otherwise
$I_M(W)$	Divide $I(W)$ by median, where $I_M(W) = 0$ if $I(W) \leq$ median($I(W)$) and 1 otherwise
$I_M(L)$	Divide $I(L)$ by median, where $I_M(L) = 0$ if $I(L) \leq$ median($I(L)$) and 1 otherwise
$I_T(Bx)$	Classify sorted $I(Bx)$ into three equally groups
$I_T(W)$	Classify sorted $I(W)$ into three equally groups
$I_T(L)$	Classify sorted $I(L)$ into three equally groups
$I_S(Bx)$	Classify $I(Bx)$ into three groups $I_S(Bx) = 0$ refers to significant regular image of Bx, $I_S(Bx) = 1$ denote a random image and $I_S(Bx) = 2$ show significant clustered image
$I_S(W)$	Classify $I(W)$ into three groups $I_S(W) = 0$ refers to significant regular image of W, $I_S(W) = 1$ denote a random image and $I_S(W) = 2$ show significant clustered image
$I_S(L)$	Classify $I(L)$ into three groups $I_S(L) = 0$ refers to significant regular image of L, $I_S(L) = 1$ denote a random image and $I_S(L) = 2$ show significant clustered image

Table 6.12: Best logistic models with corresponding estimated coefficients, standard error, p-values, estimated scale parameter and AIC value for the rectal cancer dataset using FU survival time after variable selection.

FU survival time			
Model 1: $\log(T) \sim Age + pT + pN + TCD(W) + TCD(L) + POT(L) + I(Bx)$ AIC=300.4			
Covariate	**Estimated parameters**	**Sd**	**P-value**
(Intercept)	143.537	28.371	0.000
Age	−0.575	0.347	0.098
$I[pT = 2]$	−37.364	18.628	0.045
$I[pT = 3]$	−30.866	16.783	0.066
$I[pT = 4]$	−18.068	19.001	0.342
$I[pN = 1]$	−20.414	8.643	0.018
$I[pN = 2]$	−37.591	12.100	0.002
$TCD(W)$	1.659	0.936	0.076
$TCD(L)$	−5.234	1.391	0.000
$POT(L)$	3.908	1.203	0.001
$I(Bx)$	−48.555	30.364	0.109
Scale(η)	2.646	0.157	
Model 2: $\log(T) \sim Age + pT + pN + TCD(W) + TCD(L) + POT(L) + I_M(Bx)$ AIC= 299.9			
Covariate	**Estimated parameters**	**Sd**	**P-value**
Intercept	139.575	28.444	0.000
Age	−0.561	0.358	0.116
$I[pT = 2]$	−36.111	18.805	0.055
$I[pT = 3]$	−32.243	17.016	0.058
$I[pT = 4]$	−22.689	19.619	0.247
$I[pN = 1]$	−18.852	8.925	0.035
$I[pN = 2]$	−36.247	12.326	0.003
$TCD(W)$	1.750	0.957	0.067
$TCD(L)$	−5.337	1.406	0.000
$POT(L)$	3.984	1.213	0.001
$I[I_M(Bx) = 1]$	−14.550	8.724	0.095
Scale(η)	2.659	0.157	
Model 3: $\log(T) \sim Age + pN + TCD(W) + TCD(L) + POT(L) + I_T(W)$ AIC= 299.6			
Covariate	**Estimated parameters**	**Sd**	**P-value**
Intercept	132.674	28.324	0.000
Age	−0.841	0.420	0.045
$I[pN = 1]$	−21.900	9.362	0.019
$I[pN = 2]$	−38.015	12.357	0.002
$TCD(W)$	1.797	1.054	0.088
$TCD(L)$	−5.256	1.434	0.000
$POT(L)$	4.224	1.259	0.001
$I[I_T(W) = 1]$	−21.170	10.828	0.050
$I[I_T(W) = 2]$	−29.976	14.325	0.036
Scale(η)	2.701	0.155	
Model 4: $\log(T) \sim Age + pN + TCD(Bx) + TCD(W) + TCD(L) + POT(Bx) + POT(L) + I_S(Bx)$ AIC= 299.8			
Covariate	**Estimated parameters**	**Sd**	**P-value**
Intercept	316.823	26.836	0.000
Age	−0.694	0.375	0.064
$I[pN = 1]$	−17.569	9.162	0.055
$I[pN = 2]$	−33.466	12.255	0.006
$TCD(Bx)$	1.713	1.032	0.097
$TCD(W)$	1.271	0.926	0.169
$TCD(L)$	−4.651	1.302	0.000
$POT(Bx)$	−1.390	0.901	0.123
$POT(L)$	3.483	1.141	0.002
$I[I_S(Bx) = 2]$	−209.109	0.000	0.000
Scale(η)	2.694	0.154	

Table 6.13: Best logistic models with corresponding estimated coefficients, standard error, p-values, estimated scale parameter and AIC value for the rectal cancer dataset using DF survival time after variable selection.

DF survival time			
Model 1: $\log(T) \sim pT + pN + TCD(W) + POT(L) + I_S(Bx)$ AIC= 367.8			
Covariate	**Estimated parameters**	**Sd**	**P-value**
Intercept	131.033	29.671	0.000
$I[pT = 2]$	−51.068	21.776	0.019
$I[pT = 3]$	−32.120	20.235	0.112
$I[pT = 4]$	−60.340	22.981	0.008
$I[pN = 1]$	−35.303	10.400	0.001
$I[pN = 2]$	−38.840	13.984	0.005
$TCD(W)$	6.359	3.883	0.101
$TCD(L)$	−2.323	1.476	0.115
$POT(W)$	−4.415	3.040	0.146
$POT(L)$	1.680	1.222	0.169
$I[I_S(Bx) = 2]$	−27.749	20.890	0.184
Scale(η)	2.873	0.147	

The $Pr.Tstage$ and pT have low numbers of patients in their levels, for example there are only two patients when $Pr.Tstage = 1$ and one patient when $pT = 2$. Very low numbers of patients in any level is problematic in fitting some survival models. To solve this problem, the levels, with a low number of patients, are grouped with the next level. In pT, we now have 4 levels instead of 5, where $pT = 0$ and $pT = 1$ are combined to be 17 patients. Similarly, $Pr.Tstage = 1$ and $Pr.Tstage = 2$ are joined, to form 46 patients in the first level ($pT = 1$) and 67 patients otherwise. The clinical variable pM is removed from the analysis as it has only one patient when $pM = 1$.

When survival models (parametric and semi-parametric) and multiple regression models are fitted, all defined variables at the beginning of the section are included in addition to the I statistic and its partitioned versions which are included individually, for example adding $I(Bx), I(W)$ and $I(L)$ or $I_M(Bx), I_M(W)$ and $I_M(L)$ and so forth. The best model is then selected by the stepwise selection procedure, assessed by checking the residual plots and finally interpreted if it is well-fitting and includes the I statistic.

In this section, different survival models, which have been defined in Section 6.2, are applied in Section 6.4.1 in order to find if the survival time can be predicted by the I statistic or its classified versions (I_M, I_T or I_S) for each image type including the clinical variables. Tumor heterogeneity may be detected by $TCD(W)$ or $I(W)$. Thus the relationship between these covariates, as explanatory variables, and the rest of the

clinical variables are investigated by fitting multiple regression models in Section 6.4.2.

6.4.1 Survival Analysis

The survival analysis of the rectal cancer dataset is considered using parametric, then non-parametric and finally semi-parametric models. Firstly, as the survival times for the rectal dataset is a discrete variable (time in months), the appropriate distribution for survival time is the logistic distribution from Table 6.1. For the FU survival time, the I statistic and its different divisions are included individually, along with all clinical variables, and thus we have four possible models.

The best logistic models which include any version of I statistic are shown in Table 6.12, where each method includes the corresponding estimated coefficients, standard error, p-values, estimated scale parameter and AIC value. Here, each model includes different version of I_M, I_T or I_S for all image types. Similarly, the same process is used for DF survival time and the best logistic models are shown in Table 6.13. The residuals of each of these fitted models are plotted in Figure 6.5. Even though I was included in all models, none of their plots are close to a straight line and thus all models using FU and DF survival times are not well fitted. Thus, the interpretation of these models is not included.

Secondly, the survival function of FU and DF is estimated by non-parametric models. Figures 6.6 displays the Kaplan-Meier curve for groups of some variables. Patients with $pN = 0$, using FU and DF survival times, have significantly higher survival than those from other levels (p-value equals 0.04 and 0.01 respectively). At month 16, the probability of survival is approximately 93% for $pN = 0$, 86% for $pN = 1$ and 64% for $pN = 2$. Similarly, the tumor stage (pT) has significant differences in the survival curves using DF survival time, whereas the same variable produce no significant differences in FU survival curves. Also the log-rank test is applied for all covariates in Table 6.14, where the significant p-values are highlighted in red to test the null hypothesis that $H_0 : \beta_j = 0$. Neither of $Pr.Tstage$, $therapy$, $Gender$, partitioned I for all images classified POT nor divided TCD have significant p-values (<0.05) for either FU or DF survival times which means that there is no significant evidence to reject the null hypothesis (no difference between two survival functions). Note that, to save space,

none of the survival curves nor the log-rank test results were included.

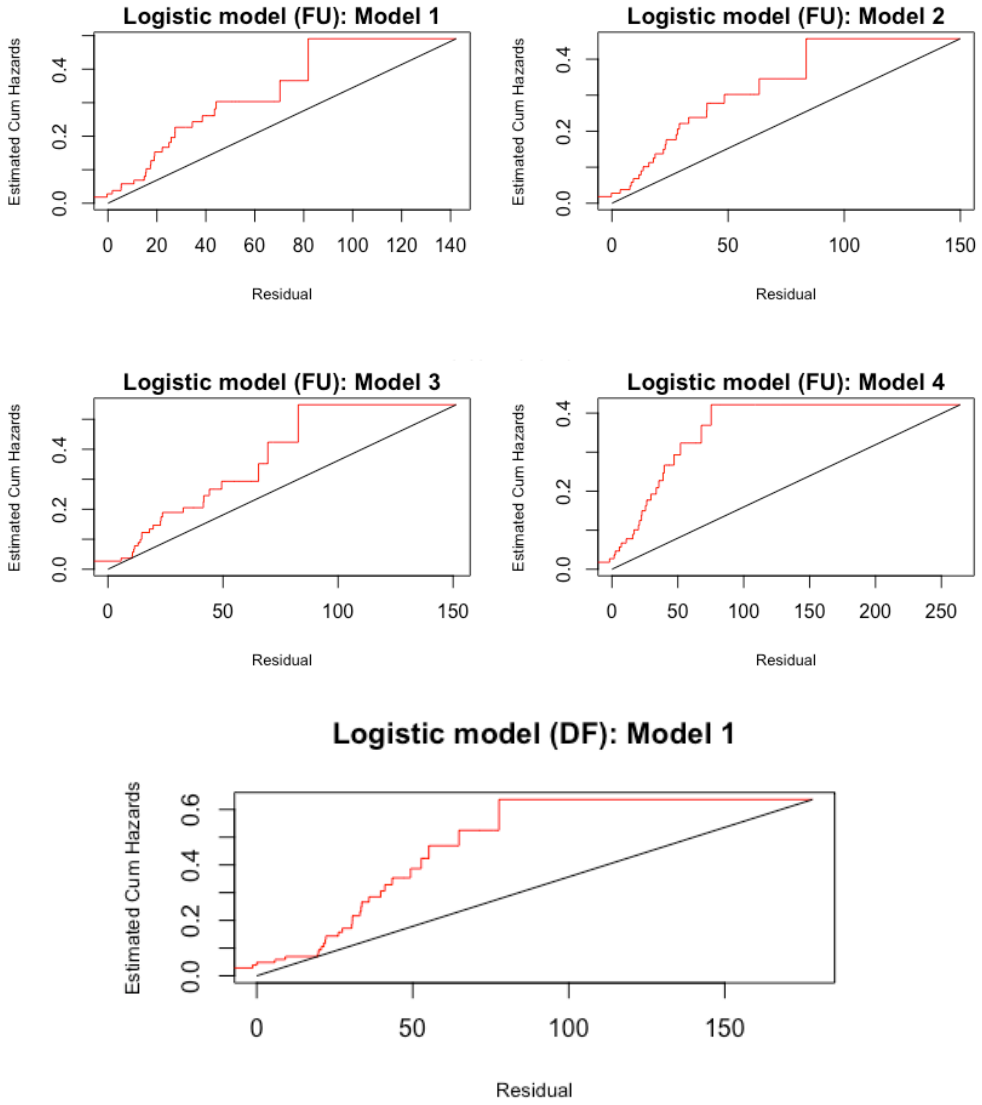


Figure 6.5: Cox-Snell residuals to assess the fit of logistic models in Table 6.12 and 6.13 for rectal cancer dataset using FU and DF survival times, where the red line shows r_i against $-\log\{\hat{S}_R(r_i)\}$.

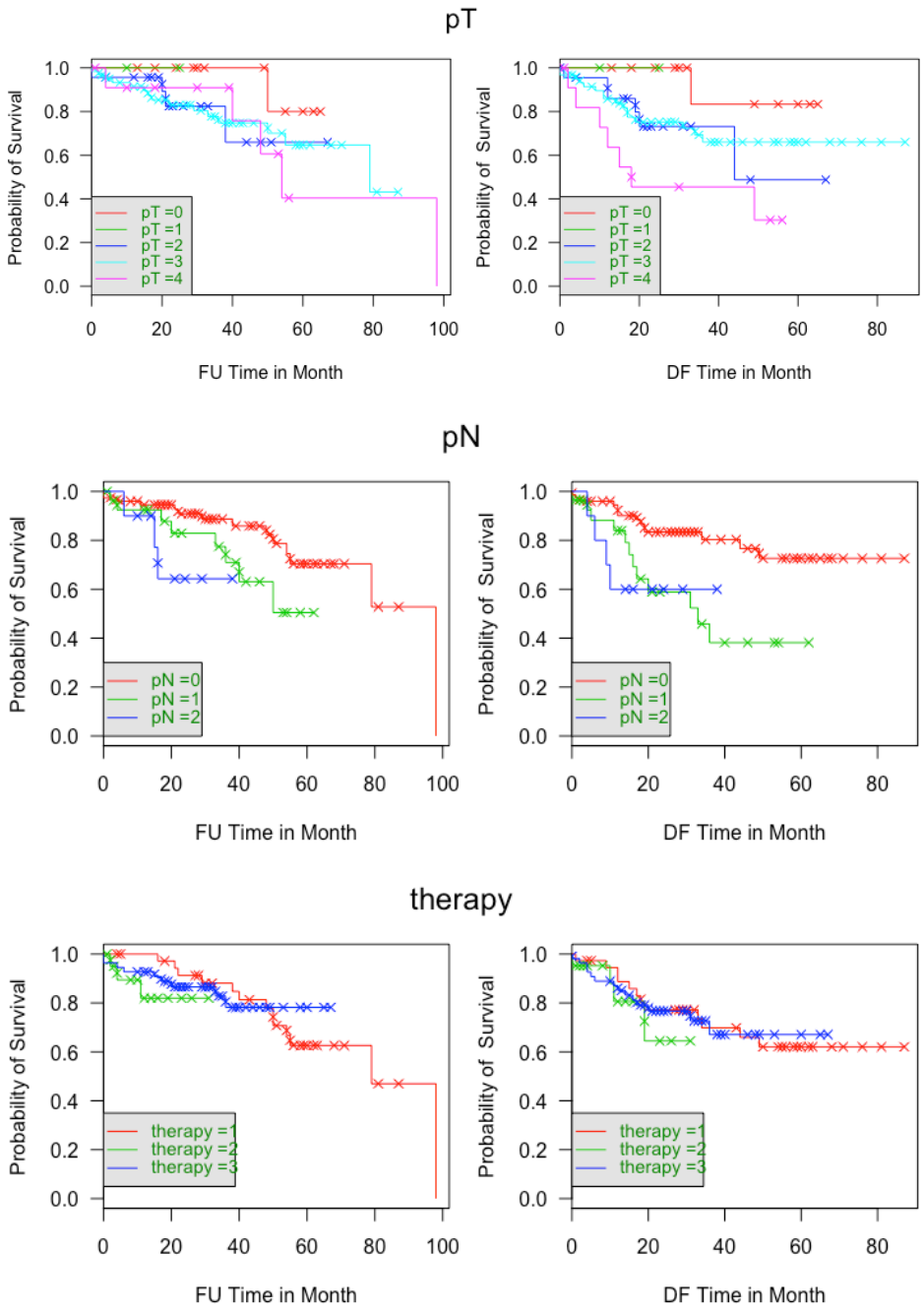


Figure 6.6: The Kaplan-Meier survival curves for the rectal cancer dataset for lymph node stage pN, chemotherapy type $therapy$ and tumor stage pT, where the first column shows follow-up (FU) and the second column presents disease-free (DF) survival times.

Finally, the Cox PH model is also applied using FU and DF survival times. Each version of I statistic is added individually with all clinical variables as we did in the parametric models. We have in total four models for each survival time. After the step-wise variable selection procedure, we only selected the models that included a division of the I statistic. The best models that includes I are three models using FU survival

Table 6.14: The chi-squared statistic of the log-rank test with corresponding, degrees of freedom and p-values for the variables of the rectal cancer dataset using FU and DF survival times.

FU survival time			
Variable	**Chi-square**	**Dof**	**P-value**
$Pr.Tstage$	0.4	2	0.80
pT	2.8	4	0.60
pN	6.6	2	0.04
pM	20.4	1	0.00
$therapy$	1.8	2	0.40
$Gender$	1.7	1	0.20
DF survival time			
Variable	**Chi-square**	**Dof**	**P-value**
$Pr.Tstage$	0.5	2	0.80
pT	10.6	4	0.03
pN	10.6	2	0.01
pM	55.5	1	0.00
$therapy$	0.5	2	0.80
$Gender$	0.1	1	0.80

time and only one model using DF survival time.

The best models with their estimated coefficients, exponential, standard error, p-values and AIC values are shown in Table 6.15. The goodness of fit assessment for the Cox PH best models is checked by the Cox-Snell residuals in Figure 6.7. The only graph which is close to a 45% line, is Model 3 Cox PH using FU survival time, indicating that this model provides a reasonable fit to the rectal cancer dataset. The best fitted model can be expressed as

$$h(t) = h_0(t)\exp\{0.034Age + 0.694I[pN = 1] + 1.622I[pN = 2] - 0.075TCD(W) + 0.274TCD(L) - 0.227POT(L) + 1.019I[I_T(W) = 1] + 1.578I[I_T(W) = 2]\}. \tag{6.12}$$

Table 6.15: The Cox PH model for the I statistic of various images from the rectal cancer dataset using the FU survival information after variable selection.

FU survival time				
Covariate	$\hat{\beta}$	$\exp\{\hat{\beta}\}$	**Sd**($\hat{\beta}$)	**P-value**
Model 1: $\log(T) \sim Age + Gender + pN + TCD(W) + TCD(L) + POT(L) + I(Bx)$ AIC= 179.9				
Covariate	$\hat{\beta}$	$\exp\{\hat{\beta}\}$	**Sd**($\hat{\beta}$)	**P-value**
Age	0.034	1.035	0.019	0.082
$I[Gender = 2]$	−0.767	0.464	0.524	0.143
$I[pN = 1]$	0.758	2.135	0.502	0.131
$I[pN = 2]$	1.436	4.205	0.734	0.050
$TCD(W)$	−0.083	0.920	0.054	0.126
$TCD(L)$	0.281	1.324	0.089	0.002
$POT(L)$	−0.206	0.813	0.078	0.009
$I(Bx)$	2.579	13.189	1.547	0.095
Model 2: $\log(T) \sim Age + Gender + TCD(W) + TCD(L) + POT(L) + I_M(Bx)$ AIC= 178.6				
Covariate	$\hat{\beta}$	$\exp\{\hat{\beta}\}$	**Sd**($\hat{\beta}$)	**P-value**
Age	0.041	1.042	0.019	0.032
$I[Gender = 2]$	−0.733	0.480	0.527	0.1647
$TCD(W)$	−0.078	0.924	0.049	0.111
$TCD(L)$	0.307	1.359	0.085	0.000
$POT(L)$	−0.223	0.799	0.074	0.003
$I[I_M(Bx) = 1]$	0.914	2.494	0.463	0.048
Model 3: $\log(T) \sim Age + pN + TCD(W) + TCD(L) + POT(L) + I_T(W)$ AIC= 180.0				
Covariate	$\hat{\beta}$	$\exp\{\hat{\beta}\}$	**Sd**($\hat{\beta}$)	**P-value**
Age	0.034	1.035	0.020	0.091
$I[pN = 1]$	0.694	2.003	0.501	0.166
$I[pN = 2]$	1.622	5.067	0.717	0.024
$TCD(W)$	−0.075	0.927	0.054	0.167
$TCD(L)$	0.274	1.316	0.084	0.001
$POT(L)$	−0.227	0.796	0.075	0.003
$I[I_T(W) = 1]$	1.019	2.772	0.592	0.085
$I[I_T(W) = 2]$	1.578	4.845	0.794	0.047
DF survival time				
Model 1: $\log(T) \sim pT + pN + TCD(W) + TCD(L) + I_T(Bx) + I_T(L)$ AIC= 249.0				
Covariate	$\hat{\beta}$	$\exp\{\hat{\beta}\}$	**Sd**($\hat{\beta}$)	**P-value**
$I[pT = 2]$	2.225	9.257	1.141	0.051
$I[pT = 3]$	1.465	4.328	1.135	0.197
$I[pT = 4]$	2.689	14.730	1.157	0.020
$I[pN = 1]$	1.305	3.688	0.437	0.003
$I[pN = 2]$	1.570	4.809	0.668	0.018
$TCD(W)$	−0.064	0.937	0.042	0.129
$TCD(L)$	0.024	1.025	0.016	0.135
$I[I_T(Bx) = 1]$	−1.161	0.312	0.628	0.064
$I[I_T(Bx) = 2]$	−3.526	0.029	1.351	0.009
$I[I_T(L) = 1]$	1.077	2.936	0.584	0.065
$I[I_T(L) = 2]$	3.559	35.141	1.378	0.009

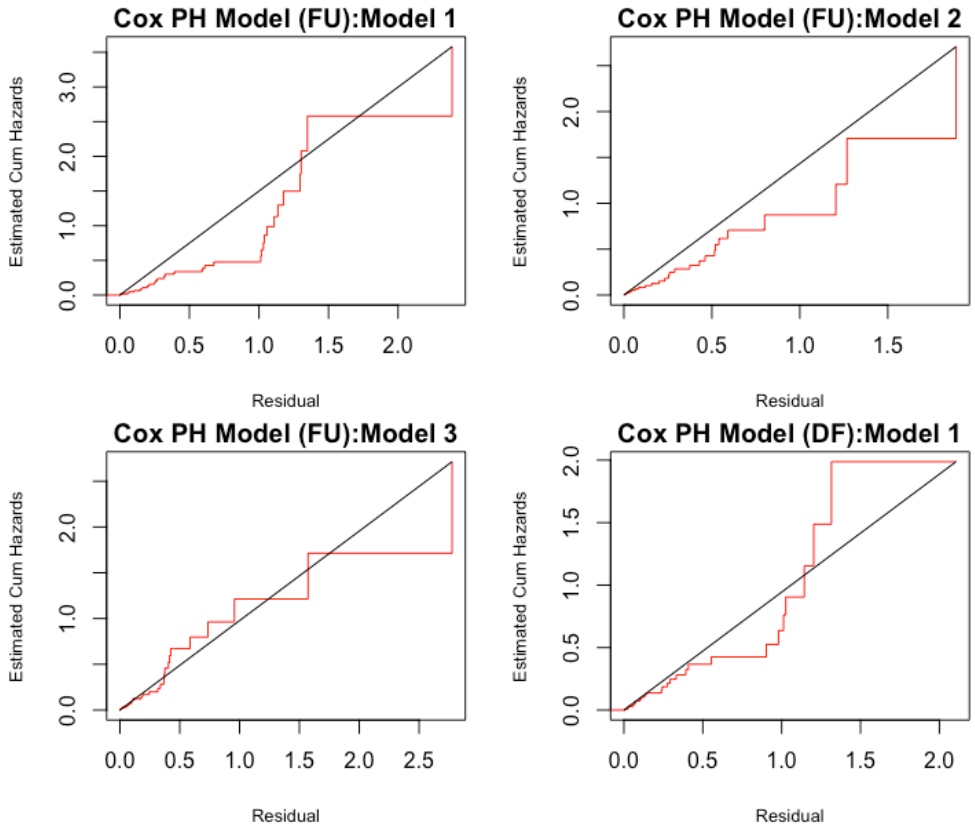


Figure 6.7: Cox-Snell residuals to assess the fit of Cox PH models in Table 6.15 for rectal cancer dataset using FU and DF survival times, where the red line shows r_i against $-\log\{\hat{S}_R(r_i)\}$.

Table 6.15 shows, in red, the p-values which are significantly different from zero under $H_0 : \beta_j = 0$ (p-value<0.05), whereas the others are not. However, if there is one of two levels in the same variable which is not significant, the insignificant level is still considered in the interpretation of the model. The interpretation of the model is that the positive estimate of $I[pN = 2]$ means that the higher stage of lymph nodes will increase the hazard level, the hazard for patients have $I[pN = 1]$ and $I[pN = 2]$ are 2 and 5 times that for those of who are in $I[pN = 0]$. Similarly, the positive estimate of $I[I_T(W) = 2]$ (more likely to be clustered images) illustrates that the structured images increase the hazard risk, thus those patients who have $I[I_T(W) = 1]$ and $I[I_T(W) = 2]$ have a hazard approximately 2.78 and 4.85 times those who are in $I[I_T(W) = 0]$, which is more likely to be unstructured images. For all patients, a 1% ratio increases in $TCD(W)$, the hazard increased by 27.4%, but a 1% ratio increase in $POT(W)$, the hazard decreased by 22.7%. In conclusion, the I statistic and its divisions were included in the logistic parametric models, but none of models were well-fitted for the rectal cancer dataset. In the non-parametric model, none of the classified I were significant. The only significant

I was when $I_T(W) = 2$ (more likely to be clustered images), which showed an increase of FU survival time.

6.4.2 Predicting $I(W)$ and $TCD(W)$

The aim for this section is to find the association between both the I statistic and tumor cell density for whole tumor only and explanatory variables. Both $I(W)$ and $TCD(W)$ can reflect tumor heterogeneity, but we need to find which clinical variables are associated with them. For simplicity in analysis, the I statistic for all images, as continuous variables, is only used without considering the partitional versions of I images, I_M, I_T and I_S. To model the relationship between more than two explanatory variables and each response variable, a multiple regression model for the main effects is fitted. Essential predictor variables are then selected by the stepwise procedure, the best model is assessed by checking the distribution of residuals and we use this to also interpret if the model is well-fitted. No survival time variables have been included because they have already been considered in Section 6.4.1.

Table 6.16: The estimated coefficients with their corresponding standard error and p-values of multiple regression model for rectal cancer dataset after variable selection.

$I(W) \sim POT(Bx) + POT(L) + I(L)$			
Covariate	$\hat{\beta}$	**Sd**($\hat{\beta}$)	**P-value**
Intercept	-0.049	0.023	0.037
$POT(Bx)$	0.122	0.049	0.015
$POT(L)$	0.422	0.041	0.000
$I(L)$	0.236	0.053	0.000
$TCD(W) \sim TCD(Bx) + TCD(L) + POT(Bx) + POT(W) + POT(L) + I(W)$			
Covariate	$\hat{\beta}$	**Sd**($\hat{\beta}$)	**P-value**
Intercept	0.887	0.650	0.175
$TCD(Bx)$	-0.158	0.071	0.029
$TCD(L)$	0.240	0.065	0.000
$POT(Bx)$	11.539	5.986	0.057
$POT(W)$	74.507	3.365	0.000
$POT(L)$	-13.606	5.576	0.016
$I(W)$	-5.736	2.582	0.028

To find the association between $I(W)$ and all variables, a multiple regression model is fitted. By using the stepwise selection procedure, we select the best model with lower

AIC$= -213$,

$$I(W) = \beta_0 + \beta_1 POT(Bx) + \beta_2 POT(L) + \beta_3 I(L) + \varepsilon. \tag{6.13}$$

The same set of variables is used to predict the $TCD(W)$. The best model is selected by the lowest AIC value using the stepwise selection procedure, which is as follows

$$\begin{aligned} TCD(W) = \beta_0 + \beta_1 TCD(Bx) + \beta_2 TCD(L) + \beta_3 POT(Bx) + \beta_4 POT(W) \\ + \beta_5 POT(L) + \beta_6 I(W) + \varepsilon. \end{aligned} \tag{6.14}$$

The AIC of this model equals 553 and the I statistic of the whole tumor contributes to the model. The parameter estimates for models 6.13 and 6.14, with their corresponding standard error and p-values, are shown in Table 6.16. The plots of residuals of both models are shown in Figure 6.8. It is clear that the residuals of model 6.14 (right figure) are not randomly distributed which means the model is not well fitted and one of the observation is identified as an outlier. In Figure 6.8 (left figure), however, the visualisation of the residuals from model 6.13 is well dispersed which means the model is well-fitted.

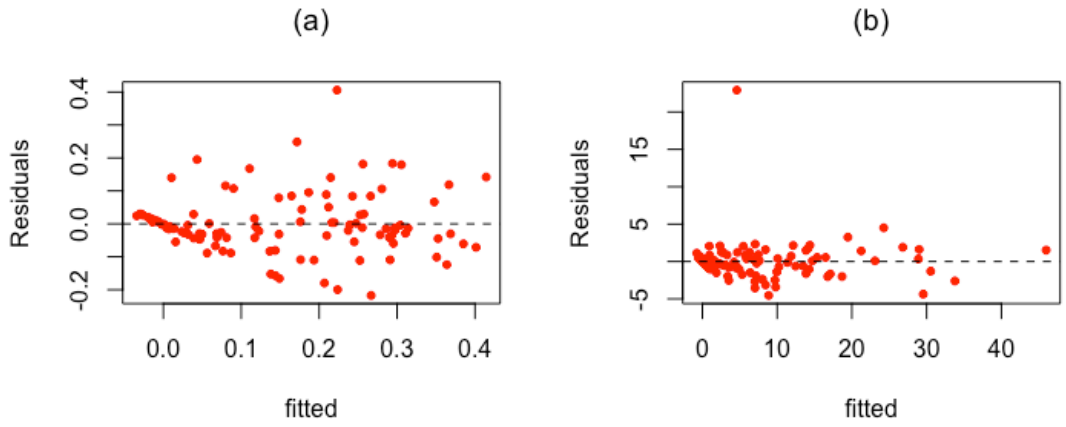


Figure 6.8: (a) Residuals versus fitted values plot of model (6.13) and (b) Residuals versus fitted values plot of model (6.14).

Therefore, only model (6.13) is interpreted. From Table 6.16, all covariates of model (6.13) have a low p-value (< 0.05) which means they are meaningfully related to changes in the I statistic. For instance, the coefficients of both proportions of tumor for Bx and L images indicate that for every additional 1% in POT we can expect the I statistic

of the whole tumor image to increase by an average of 12.2% and 42.2% respectively. Similarly, as the I statistic of the lumen site image rises by 1%, the I statistic of the whole tumor image on average, increased by 23.6%. As a result all covariates in model 6.13 are contributing, on average, to increase the I statistic of the whole tumor image that means more clustered image. More precisely the proportion of tumor for the lumen site image reflects more on the heterogeneity of the whole tumor image.

6.5 Conclusion

Clinical questions were answered in this chapter. The advantages of using non-parametric and Cox PH model are that there are no restriction about the distribution of survival times which can be either continuous or discrete. Although the survival time was not directly predicted by the I, when this statistic was divided by different cutoffs, we found there was significant difference between survival curves.

Regarding to the gastric cancer dataset, only the partitioned versions of I, I_S, has shown that there is significant difference between survival curves using non- and semi-parametric survival models. Patients who have clustered images tend to survive longer. This finding suggests that more structured tissue tends to be better than random ones. We also tested whether proportion of tumor (POT) is related to I, we showed that when POT has been increased, the I statistic, on average, is decreased.

In terms of the rectal cancer dataset, we found that using Cox PH model, when the images tends to be clustered ($I_T(W) = 2$), the survival time was increased. Furthermore, we investigated that $TCD(W)$ was not predicted by any clinical variables. As proportion of Bx and (L) as well as the I statistic of lumen site image were increased, the I statistic of whole tumour, on average, increased.

The affects of cluster in images on survival time are, in general, consistent which have been obtained in both gastric and rectal cancer datasets.

Chapter 7

Discussion, Future Work and Recommendations for the Pathologist

7.1 Discussion

In this thesis we used statistical methods to explore the spatial features of biomedical images on a hexagonal grid for stomach and rectum cancers. The analysis focussed on detecting local heterogeneity, detecting anisotropy and spatial consistency of images. We used statistical tests which were based on both derived asymptotic distributions and simulation-based methods. This project is the first one to look at pathological images spatially, and we found that objective numerical summaries of heterogeneity are more informative than only comparing the overall proportion of tumor (POT).

In the first part of this thesis, traditional pathological methods of biomedical image analysis were discussed. The gastric and rectal cancer datasets were also described using exploratory analysis. The spots classification of images were ascertained as the preferred classification of spots by pathologists.

In Chapter 2, we considered spatial statistical measurements, under a normal approximation of distribution, including the black-white join-count, Moran's I and Geary's C statistics. These statistics were compared and examined using extensive simulation studies. Moran's I was the most powerful measurement of spatial analysis when we had 300 or more spots. The I statistic was then used to assess the heterogeneity of images. To compute spatial statistics, a neighbouring system of the hexagonal grid was defined

effectively for single- and multiple-regions which also allowed for missing spots.

In the following part of this thesis, the I statistic was modified to measure the heterogeneity/clustering in different directions. Neighbouring systems for different directions were also defined to consider single and multi-region images in addition to the neighbouring system of rotated images. A statistical test for determining the heterogeneity in the direction of the lumen was established. However, the statistical test for detecting directionality was only valid under the null hypothesis that the spots are independently distributed (rather than isotropically clustered). Obviously when the spots in an image are independent, we mostly have no direction. Here it is meaningless to detect direction in an independent framework, but this was a limitation of testing the directional I statistic.

In Chapter 4, we overcame the limitation of directional I and investigated a more flexible simulation-based statistical test for detecting directions by parameter estimation. The parameters of the Markov random field model were another way to investigate the clustering which gives similar information to the I statistic. Here, we introduced a new simulation-based iterative method (*IM*) for the estimation of parameters in *BMRF* as the exact likelihood function is intractable. The statistical test of *IM* is distribution-free and effective for detecting heterogeneity either in the overall image or in different directions without any restrictions needed. We only need to use 300 spots to make the *IM* work effectively. Based on simulation, the accuracy of *IM* was compared with existing methods, and it was found that our method had a better performance and less error.

After that, the consistency for pairs of images with different resolutions is checked by either considering the overall distribution of spot classifications, or by considering the spot spatial features. The spatial consistency for pairs of images was checked by spot prediction, where we predicted low-resolution images from the high-resolution version. We investigated whether the images can be spatially predicted, which would mean the pairs of images were consistent. Finally, we addressed several pathology questions in both gastric and rectal cancer datasets in addition to relating the I statistic to patient survival. The I statistic displayed a difference in the survival curves for patients in the gastric cancer dataset. This showed that the patients can be classified into two groups depending on their image structure, where patients with heterogeneous images had higher survival times.

7.2 Future work

In this section possible future work is described for each chapter. In Chapter 2, the distribution of the directional I statistic is only valid under the null hypothesis that the spots are independently distributed. Future work would define the theoretical statistical test under $H_0 : I_1 = I_2 = I_3$ when the spots are autocorrelated.

In Chapter 4, future work would be to determine the output of *MCMC*, which has been checked with different settings by simulation, without large number of simulations. If we extend the *MCMC* to the extra parameter setting, can we determine the answer without simulation using theoretical methods. Future work coming from a combination of Chapters 4 and 5 could be to determine if pairs of images are consistent in terms of their parameters. For example, patients for before and after operations might have different estimated parameter values.

In Chapter 5, there are different ways of sampling images using low- and high-resolution spot classifications, here we would like to investigate how the standard error depends on the number of sampling spots. Pathologists could then decide what density of spots is better to use.

From Chapter 6, future work could be to further investigate the findings relating the heterogenous tumour to higher survival by considering more images for both gastric and rectal cancer datasets, where we found that more structured tissues tends to have better survival than random patterns. This features needs a pathological review.

7.3 Pathologist Recommendations

Pathologists should consider the following recommendation in which the heterogeneity of a tumor can be better measured numerically. Statistical tests are more effective when images of size 300 spots or more are provided and to avoid sample size 50 spots. When the pathologist allocating the hexagonal grid in the digitised histological slides and before sampling, it is important to do a rotation of the grid to make the direction of the lumen line up exactly with one of the three hexagonal axes to be able to measure the heterogeneity in the direction of the lumen accurately. Moreover, it is better to sample the whole tumor, in particular in the gastric cancer dataset, as we observed clustering in the

digital slide of the whole tumor, but with nothing showing in the sampling area which is close to the tumor site. In digital image sampling in rectal cancer dataset, we recommend to sample the whole image, but there is less need to sample high-resolution images as they are consistent with the low-resolution images. We would also recommend that the hexagon axes of the sampling grid should be exactly equal to simplify the mathematical computation in the future.

References

Akaike, H. (1973). Information theory and an extention of the maximum likelihood principle. pages pages 267–281.

Almohri, W. A. (2012). Strategies for establishing the number of clusters. Master's thesis, School of mathematics, University of Leeds.

American Cancer Society (2017). What Is Stomach Cancer? [Online; accessed May 21, 2018].

American Joint Committee on Cancer (2009). Colon and rectum cancer staging. Technical report.

Aoyama, T., Hutchins, G., Arai, T., Sakamaki, K., Miyagi, Y., Tsuburaya, A., Ogata, T., Oshima, T., Earle, S., Yoshikawa, T., and I. Grabsch, H. (2018). Identification of a high-risk subtype of intestinal-type japanese gastric cancer by quantitative measurement of the luminal tumor proportion. *Cancer Medicine*, 7: pages 4914–4923.

Aurello, P., Berardi, G., Giulitti, D., Palumbo, A., Tierno, S. M., Nigri, G., Angelo, F. D., Pilozzi, E., and Ramacciato, G. (2017). Tumor-stroma ratio is an independent predictor for overall survival and disease free survival in gastric cancer patients. *The Surgeon*, 15(6): pages 329–335.

Aykroyd, R., Haigh, J., and Zimeras, S. (1996). Unexpected spatial patterns in exponential family auto models. *Graphical Models and Image Processing*, 58(5): pages 452–463.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Longman Group Limited.

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics*, 41(1): pages 379–406.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages pages 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24.

Birnbaum, Z. W. and Tingey, F. H. (1951). One-sided confidence contours for probability distribution functions. *The Annals of Mathematical Statistics*, 22(4): pages 592–596.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer Science and Business Media, Inc.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *BMJ*, 310(6973):170.

Bro, R., Kjeldahl, K., Smilde, A., and Kiers, H. (2008). Cross-validation of component models: A critical look at current methods. *Analytical and bioanalytical chemistry*, 390: pages 1241–51.

Cancer Research UK (2018). Bowel cancer mortality statistics. [Online; accessed July, 2018].

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall.

Chatterjee, D. and Chatterjee, A. (2010). Binary logistic regression using survival analysis. *SSRN Electronic Journal*.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4): pages 327–335.

Cliff, A. D. and Ord, J. K. (1973). *Spatial autocorrelation*. Pion Limited, London.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes*. Pion Limited, London.

Cochran, W. G. (1952). The Chi-squre Test of Goodness of Fit. *The Annals of Mathematical Statistics*, 23(3): pages 315–345.

Collett, D. (1994). *Modelling Suevival Data in Medical Research*. Chapman and Hall.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society,Series B (Methodological)*, 34(2): pages 187–220.

Cox, D. R. and Erricker, B. C. (1968). A general definition of residuals.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley and Sons, Inc.

Cserni, Gabor (2011). Colorectal Cancer epidemiology and pathology. Bacs-Kiskun County Teaching Hospital.

Diggle, P. J. (1981). *Statistical Analysis of spatial point patterns*. Academic Press Inc, London UK.

Dubes, R. C. and Jain, A. K. (1989). Random field models in image analysis. *Journal of Applied Statistics*, 16:131–164.

Dworak, O., Keilholz, L., and Hoffmann, A. (1997). Pathological features of rectal cancer after preoperative radiochemotherapy. *International Journal of Colorectal Disease*, 12: pages 19–23.

Ebdon, D. (1985). *Statistics in Geography*. Blackwell publishers Inc, United State of America.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2013). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5): pages 359–386.

Freund, R. J. and Wilson, W. J. (1998). *Regression analysis*. Academic Press Inc, London.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5: pages 115–146.

Grabsch, H. (2013). Grabsch yokohama gastric cancer dataset @ONLINE.

Graham, A., Atkinson, P., and Danson, F. (2004). Spatial analysis for epidemiology. *Acta Tropica*, 91(3): pages 219 – 225.

Greene, F., Page, D., Fleming, I., Fritz, A., Balch, C., Haller, D., and Morrow, M. (2002). *AJCC Cancer Staging Manual*. Lippincott Raven Publishers, Philadelphia, PA.

Guidelines for the Management of Colorectal Cancer (2007). Technical report, London UK.

Gullo, I., Carneiro, F., Oliveira, C., and Almeida, G. (2017). Heterogeneity in gastric cancer: From pure morphology to molecular classifications. *Pathobiology : journal of immunopathology, molecular and cellular biology*, 85: pages 50–63.

Hale, M. D., Nankivell, M., Hutchins, G. G., Stenning, S. P., Langley, R. E., Mueller, W., West, N. P., Wright, A. I., Treanor, D., Hewitt, L. C., Allum, W. H., Cunningham, D., Hayden, J. D., and Grabsch, H. I. (2016). Biopsy proportion of tumour predicts pathological tumour response and benefit from chemotherapy in resectable oesophageal carcinoma: results from the uk mrc oe02 trial. *Oncotarget*, 7(47): pages 77565–77575.

Harrington, D. P. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69: pages 553–566.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1): pages 97–109.

Hogg, R. and Tanis, E. (1977). *Probability and Statistical Inference*. Macmillan Publishing Company, Inc., New York.

Hu, B., Hajj, N. E., Sittler, S., Lammert, N., Barnes, R., and Meloni-Ehrig, A. (2012). Gastric cancer: Classification, histology and application of molecular pathology. *Journal of Gastrointestinal Oncology*, 3(3).

Huijbers, A., Tollenaar, R. A. E. M., v Pelt, G. W., Zeestraten, E. C. M., Dutton, S., McConkey, C. C., Domingo, E., Smit, V. T. H. B. M., Midgley, R., Warren, B. F.,

Johnstone, E. C., Kerr, D. J., and Mesker, W. E. (2013). The proportion of tumor-stroma as a strong prognosticator for stage ii and iii colon cancer patients: validation in the victor trial. *Annals of Oncology*, 24(1): pages 179–185.

Japanese Gastric Cancer Association (2011). Japanese classification of gastric carcinoma: 3rd english edition. *Gastric Cancer*, 14(2): pages 101–112.

Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis*, 73(2): pages 199 – 220.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282): pages 457–481.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis Techniques for Censored and Truncated Data*. Springer, New York, Inc.

Lee, D., Ham, I.-H., Son, S. Y., Han, S.-U., Kim, Y.-B., and Hur, H. (2017). Intratumor stromal proportion predicts aggressive phenotype of gastric signet ring cell carcinomas. *Gastric Cancer*, 20(4): pages 591–601.

Lee, D. T. and Schachter, B. J. (1980). Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3): pages 219–242.

Lee, E. T. (2003). *Statistical Methods for survival data Analysis*. John Wiley and Sons, Inc.

Lee, J. and Wong, D. W. (2001). *Statistical Analysis with Arcview GIS*. John Wiley and Sons, Inc, United State of America.

Logan, J. R. (2012). Making a place for space: Spatial thinking in social science. *Annual review of sociology*, 38.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press Inc, London.

Marin, J. M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing*, 22(6): pages 1167–1180.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8): pages 1246–1266.

McDonald, J. H. (2009). *Handbook of Biological Statistics*.

Mesker, W., M C Junggeburt, J., Szuhai, K., de Heer, P., Morreau, H., Tanke, H., and Tollenaar, R. (2007). The carcinoma-stroma ratio of colon carcinoma is an independent fact of survival compared to lymph node status and tumor stage. 29: pages 387–98.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): pages 1087–1092.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10: pages 243–251.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37: pages 17–23.

Mouliere, F., Messaoudi, S. E., Gongora, C., Guedj, A.-S., Robert, B. G. M., Rio, M. D., Molina, F., Lamy, P.-J., Lopez-Crapez, E., Mathonnet, M., Ychou, M., Pezet, D., and Thierry, A. R. (2013). Circulating cell-free dna from colorectal cancer patients may reveal high kras or braf mutation load. *Translational oncology*, 6(3): pages 319–28.

Mller, J., Pettitt, A. N., Berthelsen, K. K., and Reeves, R. W. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2): pages 451–548.

Parmar, M. and Machin, D. (1995). *Survival Analysis A Practical Approach*. John Wiley and Sons Ltd.

Peng, C., Liu, J., Yang, G., and Li, Y. (2018). The tumor-stromal ratio as a strong prognosticator for advanced gastric cancer patients: proposal of a new tsnm staging system. *Journal of gastroenterology*, 53(5): pages 606–617.

plantmedicine (2018). Homoeopathic remedies for colorectal cancer. [Online; accessed June 1, 2018].

Reeves, R. and Pettitt, A. N. (2004). Efficient recursions for general factorisable models. *Biometrika*, 91(3): pages 751–757.

Ripley, B. D. (1979). Algorithm as 137: Simulating spatial patterns: Dependent samples from a multivariate density. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): pages 109–112.

Rodríguez, G. (2010). Parametric survival models. Technical report.

Royston, P. (1993). A pocket-calculator algorithm for the shapiro-francia test for non-normality: An application to medicine. *Statistics in Medicine*, 12(2): pages 181–184.

Rdel, C., Martus, P., Papadoupolos, T., Fzesi, L., Klimpfinger, M., Fietkau, R., Liersch, T., Hohenberger, W., Raab, R., Sauer, R., and Wittekind, C. (2005). Prognostic significance of tumor regression after preoperative chemoradiotherapy for rectal cancer. *Clinical Oncology*, 23(34): pages 8688–8696.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall.

Sen, A. (1976). Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical Analysis*, 8(2): pages 175–184.

Shamudheen Rafiyath, M. (2018). Gastric cancer staging @ONLINE.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4): pages 591–611.

Stone, R. S. (2017). Deep learning with sub-optimal ground truth: rectal cancer segmentation on whole slide images. Master's thesis, School of Computing, University of Leeds.

Treanor, D., Dattani, M., Quirke, P., and Grabsch, H. (2008). Systematic random sampling with virtual slides: A new software tool for tissue research. *Annual Meeting of Pathology*.

Turner, R. (2018). *Delaunay Triangulation and Dirichlet (Voronoi) Tessellation*.

Underwood, J. C. E. and Cross, S. S. (2009). *General and Systematic Pathology*. Elsevier limited.

Venables, W. N. and Ripley, B. D. (2010). *Modern Applied Statistics with S*. Springer Publishing Company, Incorporated.

Waks, A., Tretiak, O., and Gregoriou, G. (1990). Restoration of noisy regions modeled by noncausal markov random fields of unknown parameters. volume 2, pages pages 170 – 175.

West, N., Dattani, M., McShane, P., Hutchins, G., Grabsch, J., Mueller, W., Treanor, D., Quirke, P., and Grabsch, H. (2010a). The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients. *Br J Cancer*, 102: pages 1519–1523.

West, N., Grabsch, H., Treanor, D., Sebag-Montefiore, D., Thorpe, H., Jayne, D., Rutten, H., Swellengrebel, H. A., Nagtegaal, I. D., and Quirke, P. (2010b). Quantitative assessment of tumor cell density in rectal cancer following three different preoperative therapies compared to surgery alone. *Journal of Clinical Oncology*, 28(15): pages 3651–3651.

Wheeler, J., Warren, B., Mortensen, N., Ekanyaka, N., Kulacoglu, H., Jones, A., George, B., and Kettlewell, M. (2002). Quantification of histologic regression of rectal cancer after irradiation: a proposal for a modified staging system. *Diseases of the Colon & Rectum*, 8(45): pages 1051–1056.

Wright, A., Grabsch, H., and Treanor, D. (2015). RandomSpot: A web-based tool for systematic random sampling of virtual slides. *Journal of Pathology Informatics*, 6(1): pages 8.

Yamada, T., Yoshikawa, T., Taguri, M., Hayashi, T., Aoyama, T., Sue-Ling, H. M., Bonam, K., Hayden, J. D., and Grabsch, H. I. (2016a). The survival difference between gastric cancer patients from the UK and japan remains after weighted propensity score analysis considering all background factors. *Gastric Cancer*, 19(2): pages 479–489.

Yamada, T., Yoshikawa, T., Taguri, M., Hayashi, T., Aoyama, T., Sue-Ling, H. M., Bonam, K., Hayden, J. D., and Grabsch, H. I. (2016b). The survival difference between gastric cancer patients from the uk and japan remains after weighted propensity score analysis considering all background factors. *Gastric Cancer*, 19(2): pages 479–489.