Some spatial statistical techniques with applications to cellular imaging data

Thomas Honnor

A thesis presented in fulfilment of the requirements for the degree of

Doctor of Philosophy in Statistics

Department of Statistics
University of Warwick
April 2017

WARWICK
THE UNIVERSITY OF WARWICK

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors, Dr. Julia Brettschneider and Dr. Adam Johansen; collaborators, Dr. Stephen Royle and his research group; department colleagues; friends and family for all of their help and support over the course of my studies. This thesis would not have been possible without you.

Particular thanks go to Dr. Stephen Royle and his research group within the Centre for Mechanochemical Cell Biology at the University of Warwick for discussions of the biological and statistical problems of interest to them and for providing the data that serves to illustrate the ability of the methodologies proposed within this thesis.

I would also like to thank the Engineering and Physical Sciences Research Council for funding these studies through Doctoral Training Grants EP/K503204/1, EP/P50578X/1 and EP/L505110/1.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented was carried out by the author, with real biological data made available by Dr. Stephen Royle, Department of Mechanochemical Cell Biology, University of Warwick.

Material from Chapters 3 and 4 respectively has been made available by the author as working papers:

Thomas R. Honnor*, Julia A. Brettschneider, Adam M. Johansen,* **Differences in spatial point patterns with application to subcellular biological structures**, *CRiSM Working Paper 17-01.*

Thomas R. Honnor*, Adam M. Johansen, Julia A. Brettschneider,* **A non-parametric test for dependency between estimated local bulk movement patterns**, *CRiSM Working Paper 17-03.*

Material from Chapter 5 has been published by the author in the following paper:

F. M. Nixon, **T. R. Honnor**, N. I. Clarke, G. P. Starling, A. J. Beckett, A. M. Johansen, J. A. Brettschneider, I. A. Prior, and S. J. Royle. Microtubule organization within mitotic spindles revealed by serial block face scanning electron microscopy and image analysis. *Journal of Cell Science*, 130(10):1845-1855, April 2017.

Abstract

The aim of this thesis is to provide techniques for the analysis of a variety of types of spatial data, each corresponding to one of three biological questions on the function of the protein TACC3 during mitosis. A starting point in each investigation is the interpretation of the biological question and understanding of the form of the available data, from which a mathematical representation of data and corresponding statistical problem are developed.

The thesis begins with description of a methodology for application to two collections of (marked) point patterns to determine the significance of differences in their structure, achieved through comparison of summary statistics and quantification of the significance of such differences by permutation tests. A methodology is then proposed for application to a pair spatio-temporal processes to estimate their individual temporal evolutions, including ideas from optimal transportation theory, and a test of dependence between such estimators. The thesis concludes with a proposed model for line data, designed to approximate the mitotic spindle structure using trajectories on the surface of spheroids, and a comparison score to compare model fit between models and/or observations.

The results of methodologies when applied to simulated data are presented as part of investigations into their validity and power. Application to biological data indicates that TACC3 influences microtubule structure during mitosis at a range of scales, supporting and extending previous investigations.

Each of the methodologies is designed to require minimal assumptions and numbers of parameters, resulting in techniques which may be applied more widely to similar biological data from additional experiments or data arising from other fields.

Chapter 1

Introduction

State of the art microscope technology allows the collection of large numbers of high resolution images. Specialised preparation of biological samples and choice of imaging approach can lead to the identification of subcellular structures and the localisation of biomolecular species within such images. The purpose of collecting such images may be observational, to ascertain the typical behaviour within the cell, or experimental, to ascertain the impact on the cell of applied external conditions. It is commonly the case that images are collected from multiple samples, the analysis of which may be improved in terms of accuracy and reliability by statistical techniques. This thesis combines three investigations of spatial data arising from images of biological samples during mitosis and formulates statistical methodologies for their analysis to answer related biological questions of interest.

1.1 Biological background

Mitosis is the procedure through which eukaryote cells (those within fungi, plants and animals) replicate, with one cell dividing into two. Chromosomes encode the genetic material within cells and during mitosis the collection of chromosome pairs are divided such that one chromosome from each pair makes its way into each of the two resulting cells. Errors in the division of chromosome pairs can result in the death of resulting cells or mutations that may be harmful to the organism (Holland and Cleveland, 2009). A key research topic for cell biologists is therefore the process of mitosis, the action of biomolecular species during mitosis and the impact of applied external conditions on the process.

Division of chromosome pairs during mitosis is effectively a mechanical process. During the prometaphase of mitosis connecting fibers, kinetochore fibers or K-fibers, grow to connect two anchor points within the cell, centrosomes, to connection points on the chromosomes, kinetochores. Each K-fiber is made up of a bundle of cylindrical

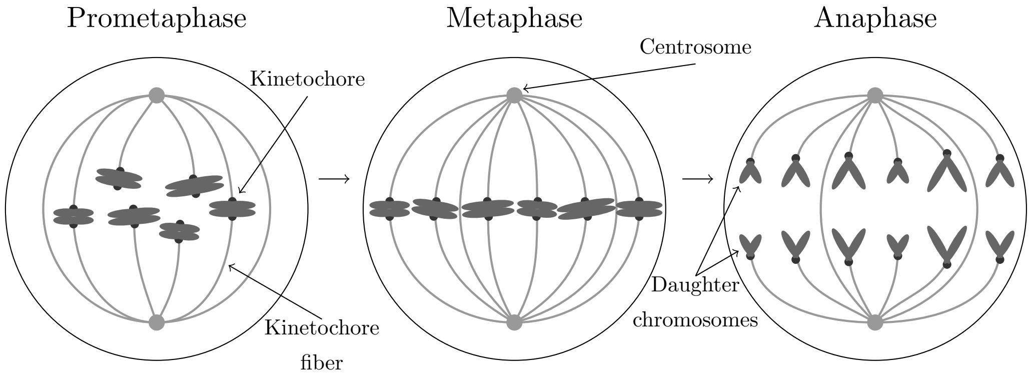


Figure 1.1: Diagram of the intermediate stages of mitosis.

structures, microtubules, which are believed to be held together by a mesh-like structure (Booth et al., 2011). Prometaphase is followed by metaphase during which chromosome pairs are pulled into alignment along the metaphase plate by the action of the K-fibers. Following this, during the anaphase, chromosome pairs are pulled apart into different halves of the cell before the cell divides. An illustration of this process may be seen in Figure 1.1. The structure of those microtubules which separate chromosomes during mitosis is referred to in combination as the mitotic spindle.

1.2 Biological questions of interest

Three related biological questions of interest are considered as part of this thesis, with each providing different types of spatial data and resulting in the development of corresponding statistical techniques. The theme linking the three problems is the action of TACC3 (Transforming Acidic Coiled-coil Containing protein 3) during mitosis.

Investigations are carried out using imaging data at differing scales to investigate different influences on the mitotic spindle structure. We refer to the micro scale as that which considers microtubules within a single K-fiber. The macro scale on the other hand considers microtubules within the entire mitotic spindle structure. Differentiation is necessary because different imaging techniques are used — it is not possible to determine every microtubule within a K-fiber from macro scale images, similarly the field of view of micro scale images is not large enough to make inference on the whole spindle structure.

1.2.1 Localisation of TACC3 and EB3

EB3 (End Binding protein 3) is a protein known to localise on the tips of growing microtubules during mitosis (Akhmanova and Steinmetz, 2010). There is some evidence that TACC3 also localises on the tips of microtubules (Gutierrez-Caballero et al., 2015). As a key purpose of microtubules is the division of chromosome pairs during mitosis, localisation of TACC3 at the tip of microtubules may be used as important evidence that TACC3 has some function during mitosis. We propose to investigate data provided by Dr. Stephen Royle (previously investigated as part of work by Gutierrez-Caballero et al. (2015)) comprised of images displaying localisation of both TACC3 and EB3, with evidence of dependence between the localisation patterns further supporting the belief that TACC3 is localised on the tips of microtubules and that it may have a function during mitosis.

The data available for this analysis is fluorescence microscopy images of both TACC3 and EB3 captured for multiple cells across a number of time points. The two proteins are each tagged with a fluorophore which emits light at a specific range of wavelengths when excited by incident light of particular wavelengths. Provided emission wavelengths are suitably distinct, the emitted light may be filtered and recorded by a camera to result in a pair of images for each sample at each time point. Variation of light intensity within each image may be interpreted as a surrogate for the spatial distribution of each biomolecular species. Examples of such images may be seen in Figure 4.5 in Chapter 4.

An existing technique for the comparison of localisation between two images is colocalisation analysis, described in more detail in Chapter 2. Due to dissatisfaction with existing approaches and the additional information contained within time series of images an alternative methodology for analysis is presented in Chapter 4. We consider time series of images to be representative observations of a spatio-temporal process, give a methodology for approximation of the temporal evolution of such processes and provide a test for dependence between temporal evolutions. This methodology is presented in detail in Chapter 4.

1.2.2 Impact of TACC3 on K-fiber structure

Given the evidence from Gutierrez-Caballero et al. (2015) and our investigations (Honnor et al., 2017b), that TACC3 is located at the tip of microtubules, and the results of investigations by Booth et al. (2011) and Nixon et al. (2015) it is proposed that TACC3 impacts the structure of microtubules within K-fibers. Interpreting mitosis as a mechanical process, on a micro scale differences in the structure of microtubules within K-fibers may impact the ability to transfer force to chromosomes and achieve accurate separation. We propose to investigate data provided by Dr. Stephen Royle (previously investigated as part of work by Nixon et al. (2015)) comprised of microtubule locations within K-fibers under control conditions and conditions where TACC3 is overexpressed, with evidence of significant differences supporting the belief that TACC3 has an impact on microtubule structure within K-fibers.

The data available for this analysis is obtained by electron microscopy of individual K-fibers within multiple cells under both experimental conditions — control and TACC3 overexpression. K-fiber cross-section images are obtained from parallel imaging planes approximately perpendicular to what is assumed to be the K-fiber axis, within which microtubules are distinguishable as dark circles. In some cases only single images are taken from each sample, from which microtubule centre locations are reported which we choose to model as point patterns. In other cases two images are taken from slices through a single sample at different distances along a K-fiber, from each of which microtubule centre locations are reported. Additional information is also provided on how microtubule centre locations, one from each image frame, are paired as ends of a common microtubule. We choose to model paired centre locations as paired point patterns, which we then re-express as a marked point pattern. Diagrams of both types of pattern may be seen in Figures 3.2 and 3.3 in Chapter 3.

There exists a large literature on the theory and application of point processes, an introduction to which is provided in this thesis in Chapter 2. We approach the problem by considering a number of summary statistics of the (marked) point patterns and comparing summary statistics between experimental groups, with permutation testing used to obtain a significance level for the difference. This methodology is presented in detail in Chapter 3.

1.2.3 Mitotic spindle structure modelling

Investigation of the previous problem is carried out at the micro scale, that of microtubules within individual K-fibers. There may be an extension of this or an additional impact of TACC3 visible on the macro scale of the whole mitotic spindle, irregularities in the structure of which may reduce its capability to separate chromosome pairs. We propose to investigate data provided by Dr. Stephen Royle comprised of microtubule endpoint pair locations under combinations of two temperature levels and three levels of TACC3 expression, with evidence of differences in microtubule structure indicating that TACC3 and/or temperature changes result in structural differences of the mitotic spindle.

The data available for this analysis is a collection of microtubule endpoint pairs obtained by serial block face scanning electron microscopy through multiple cells under a variety of experimental conditions, a detailed discussion of which is given by Nixon et al. (2017). Additionally, for each sample a pair of points are provided as an approximation to the centrosomes or poles of the mitotic spindle to permit orientation of the structure. We choose to investigate the data through the collection of straight lines connecting microtubule endpoints. Illustrations of resulting sets of lines may be seen in Figures 5.3 and 5.4 in Chapter 5.

We make an exploratory approach to modelling the mitotic spindle using idealised microtubules trajectories on the surface of a spheroid proposed based upon arguments of symmetry and parsimony, to which the deviation of observed lines may be determined. We then suggest a formula for comparison of model deviations between two models and/or samples to quantify the difference in the degree to which they deviate from the model. This methodology is presented in detail in Chapter 5.

1.3 Thesis outline

This chapter has provided a summary of three questions of interest arising from cellular imaging studies, statistical methodologies to investigate which are the focus of this thesis, alongside an overview of the biological knowledge necessary to put the problems in context. Chapter 2 introduces the background theory necessary to illuminate the methodologies presented in further chapter, including current approaches that have been applied to similar problems and others which we make use of.

Chapters 3 to 5 each present methodologies for the questions introduced in this chapter. Chapter 3 introduces a methodology for assessing the significance of the difference between collections of (marked) point patterns. The methodology is first applied to simulated (marked) point patterns, before application to patterns arising from real subcellular images. Chapter 4 introduces a methodology for estimating local bulk movement patterns and quantifying the significance of the dependence between pairs of such patterns. The methodology is applied to simulated image data, before application to real cellular images. Chapter 5 proposes an exploratory model for the mitotic spindle and statistic for the comparison of model fit between two models and/or samples. The methodology is applied to line patterns arising from real cellular images.

The final chapter of this thesis, Chapter 6, provides an overview of the conclusions of each of the previous chapters after which some directions in which each of the proposed methodologies may be extended are described.

1.4 Novel contributions

This thesis brings together theory from statistics, mathematics and physics in order to develop statistical techniques for the analysis of a particular range of biological data, but which are designed with application to more general data sets in mind. The novel contributions of this thesis are the statistic summarising the degree to which a collection of lines are oriented in the same direction, the effective force transference, introduced in Chapter 3; the methodology of estimating the temporal evolution of a spatio-temporal process by methods of optimal transportation and the procedure for testing for dependence between two such spatio-temporal processes, introduced in Chapter 4; and the mathematical model for microtubule direction in the mitotic spindle and a corresponding model fit comparison score, introduced in Chapter 5. Application to biological data sets is either completely novel, Chapters 4 and 5, or expands significantly on previous analysis of the same data set, Chapter 3.

Chapter 2

Background

2.1 Point processes

A summary of the evolution of the study of point processes including state of the art approaches may be found in works by Møller (2003), Gaetan and Guyon (2009) and Diggle (2013). This section defines some of the terms and notation used in reference to point processes in the rest of this thesis, including a description of the Poisson point process, the foundational tractable point process model. Point processes are denoted by underlined capital letters $\underline{X}, \underline{Y}, \ldots$, point patterns by underlined lower case letters $\underline{x}, \underline{y}, \ldots$, while lower case letters with subscripts $x_j, x_k, \ldots$ are used to denote points in patterns and lower case letters without subscripts $x, y, \ldots$ are used to denote points in the general space $\mathbb{R}^d$.

2.1.1 Definition

A spatial point process $\underline{X}$ is a random countable subset of a space S. The focus of this thesis is point processes on subsets of $\mathbb{R}^2$ and marked point processes on subsets of $\mathbb{R}^2 \times \mathbb{R}^2$, but the theory of spatial point processes is introduced in this and the following sections in the more general case of $S \subseteq \mathbb{R}^d$. In practice the entire process $\underline{X}$ is typically not visible, rather we observe $\underline{X}$ restricted to some bounded observation window $W \subseteq S$.

For any subset $\underline{x} \subset S$, let $n(\underline{x})$ denote the cardinality of $\underline{x}$, with $n(\underline{x}) = \infty$ if $\underline{x}$ is not finite. Denote by $\underline{x}_B = \underline{x} \cap B$ the restriction of a point configuration $\underline{x}$ to B. The set $\underline{x}$ is then said to be locally finite if $n(\underline{x}_B) < \infty$ whenever $B \subseteq S$ is bounded. The majority of the literature on point patterns and our investigations are restricted to point processes $\underline{X}$ whose realisations are locally finite subsets of S.

As with the notation $\underline{x}_B$, $\underline{X}_B$ will be used to denote the restriction of the random subset $\underline{X}$ to B. The locally finite restriction on $\underline{X}$ ensures that $\underline{X}$ takes values in the space N_{lf} defined by

$$N_{lf} = \{\underline{x} \subseteq S : n(\underline{x}_B) < \infty \text{ for all bounded } B \subseteq S\}.$$

Elements of N_{lf} are referred to as locally finite point configurations. The empty point configuration is denoted by $\emptyset$.

For a point process $\underline{X}$ on S the count function is the random variable given by $N(B) = n(\underline{X}_B)$.

2.1.1.1 Marked point processes

Consider $\underline{X}$, a point process on $T \subseteq \mathbb{R}^d$. Given some space V, if a random mark $v_i \in V$ is assigned to each point $x_i \in \underline{X}$, then $\underline{Y} = \{(x_i, v_i) : x_i \in \underline{X}\} \subset S = T \times V$ is called a marked point process with points in T and mark space V. Typically, the mark space V is a finite set or a subset of $\mathbb{R}^p, p \geq 1$.

A disc process is an example of a marked point process with mark space $V = (0, \infty)$, for which the marked point (x_i, v_i) is understood to represent the disc with centre x_i and radius v_i (Stoyan and Penttinen, 2000). A marked point process of this type may be produced in the case where $\underline{X}$ models a forest, with v_i recording the radius of the tree located at x_i. Association of point processes with geometric objects which can be identified with points in $\mathbb{R}^p$, for example ellipses or line segments, may be more broadly classified as germ-grain models (Heinrich, 1992) in which x_i, the germ, specifies the location of v_i, the grain.

A further example of a marked point process is the multitype point process, where the discrete marks, $V = \{1, 2, \ldots, k\}$, specify k different types of points (Lotwick and Silverman, 1982). One of the most studied multitype point process data sets is the amacrine cell data (Diggle, 1986) which records the locations of different light detecting cells within the eyes of a rabbit. Multitype point processes with k types of points may alternatively be considered as k-dimensional multivariate point process, that is a tuple $(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_k)$ of point processes $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_k$ corresponding to k different types of points. Multitype and multivariate point processes are equivalent, with the preferred choice of specification potentially dependent upon the application.

2.1.1.2 Formal treatment of spatial point processes

In the previous section marked point processes on $S = T \times V$ with points in $T \subseteq \mathbb{R}^d$ were distinguished from point processes on $S \subseteq \mathbb{R}^d$. However, it is possible to formalise discussion of both of these types of point processes and point processes on non-Euclidean spaces through a unified measure theoretic framework, where S is a general metric space.

More formal treatment requires the specification of σ-algebras for the space S and the set of locally finite point patterns N_{lf}. As the focus of this thesis is on point patterns on spaces $S \subseteq \mathbb{R}^d$ and not on proving theorems related to point patterns, the treatment of point patterns will be less formal, with $B \subseteq S$ used instead of the statement that B is a member of a σ-algebra over S, $F \subseteq N_{lf}$ used instead of the statement that F is a member of a σ-algebra over N_{lf} and all sets assumed measurable with respect to the appropriate Borel σ-algebra.

2.1.2 Properties

2.1.2.1 Stationarity and isotropy

A point process $\underline{X}$ on $\mathbb{R}^d$ is defined to be stationary if its distribution is invariant under translations. In other words, the distribution of $\underline{X} + x = \{x_i + x : x_i \in \underline{X}\}$ must be the same of that of $\underline{X}$ for any $x \in \mathbb{R}^d$.

A point process $\underline{X}$ on $\mathbb{R}^d$ is defined to be isotropic if its distribution is invariant under rotations about the origin. In other words, the distribution of $R\underline{X} = \{Rx_i : x_i \in \underline{X}\}$ is the same as that of $\underline{X}$ under the action of any matrix R from the special orthogonal group of dimension d, $R \in SO(d)$.

2.1.2.2 First order properties

The first order properties of the random count variables $N(B)$ for $B \subseteq S$ are described by the intensity measure. The intensity measure μ on $\mathbb{R}^d$ is given by

$$\mu(B) = \mathbb{E}N(B) \qquad\qquad B \subseteq \mathbb{R}^d.$$

If the intensity measure μ can be written as

$$\mu(B) = \int_B \rho(x)dx \qquad\qquad B \subseteq \mathbb{R}^d,$$

for some function $\rho : S \to [0, \infty)$, then ρ is referred to as the intensity function.

If $\rho(x) = \rho$ is constant over $x \in S$, then $\underline{X}$ is said to be homogeneous or first order stationary with intensity ρ. (A process $\underline{X}$ may be first order stationary or homogeneous, but not stationary as in the description of the previous section if $\mathrm{Var}(N(B))$ varies with the location of B.) The intensity ρ of a homogeneous point process may then be interpreted as the mean number of points per unit volume.

If $\rho(x)$ is not constant over $x \in S$, then $\underline{X}$ is said to be inhomogeneous. In the inhomogeneous case, $\rho(x)dx$ may be thought of as the probability of the occurrence of a point in an infinitesimally small ball with centre x and volume dx.

2.1.2.3 Second order properties

The second order properties of the random count variables $N(B)$ for $B \subseteq S$ are described by the second order factorial moment measure. The second order factorial moment measure $\alpha^{(2)}$ on $S \times S$ is given by

$$\alpha^{(2)}(C) = \mathbb{E} \sum_{x_i \neq x_j \in \underline{X}} \mathbb{1}[(x_i, x_j) \in C] \qquad\qquad C \subseteq S \times S,$$

where the sum is taken over all distinct pairs of points x_i and x_j.

The intensity measure μ and the second order factorial moment measure $\alpha^{(2)}$ together determine the second order moments of the random count variable $N(B), B \subseteq \mathbb{R}^d$ through

$$\begin{aligned}
\mathbb{E}[N(B_1)N(B_2)] &= \mathbb{E}\left[\sum_{x_i \in \underline{X}} \mathbb{1}\{x_i \in B_1\} \times \sum_{x_j \in \underline{X}} \mathbb{1}\{x_j \in B_2\}\right] \\
&= \mathbb{E}\left[\sum_{x_i \neq x_j \in \underline{X}} \mathbb{1}\{x_i \in B_1\}\mathbb{1}\{x_j \in B_2\}\right. \\
&\qquad\qquad \left. + \sum_{x_i = x_j \in X} \mathbb{1}\{x_i \in B_1\}\mathbb{1}\{x_j \in B_2\}\right] \\
&= \mathbb{E}\left[\sum_{x_i \neq x_j \in \underline{X}} \mathbb{1}\{x_i \in B_1\}\mathbb{1}\{x_j \in B_2\}\right] + \mathbb{E}\left[\sum_{x_i \in \underline{X}} \mathbb{1}\{x_i \in B_1 \cap B_2\}\right] \\
&= \alpha^{(2)}(B_1 \times B_2) + \mu(B_1 \cap B_2) \qquad B_1, B_2 \subseteq \mathbb{R}^d,
\end{aligned}$$

where the second summation term over all pairs of equal points $x_i = x_j$ with $x_i \in B_1$ and $x_j \in B_2$ reduces to the sum over the individual points in $B_1 \cap B_2$.

If the second order factorial moment measure $\alpha^{(2)}$ can be written as

$$\alpha^{(2)}(C) = \int\int \mathbb{1}\{(x, y) \in C\}\rho^{(2)}(x, y)dxdy \qquad\qquad C \subseteq \mathbb{R}^d \times \mathbb{R}^d,$$

where $\rho^{(2)}$ is a non-negative function then $\rho^{(2)}$ is called the second order product density. Intuitively, $\rho^{(2)}(x, y)dxdy$ is the probability of observing a pair of points from $\underline{X}$ occurring jointly in each of the two infinitesimally small balls with centres x, y and volumes dx, dy.

2.1.2.4 Complete spatial randomness

A point process $\underline{X}$ on S satisfies the independent scattering property, also referred to as complete spatial randomness, if $\underline{X}_{B^{(1)}}, \underline{X}_{B^{(2)}}, \ldots$ are independent for disjoint sets $B^{(1)}, B^{(2)}, \ldots$. As a result, if $\underline{X}$ satisfies the complete spatial randomness property the second order product density is simply the product of the intensity functions at the corresponding locations

$$\rho^{(2)}(x,y) = \rho(x)\rho(y) \qquad\qquad x,y \in S.$$

The concept of complete spatial randomness is important as a baseline for point processes to be compared to. Violation of complete spatial randomness may be caused by interactions between points which lead to clustering or regularity of point locations via attraction or repulsion respectively.

The only stationary point process on $S \subseteq \mathbb{R}^d$ which satisfies the property of complete spatial randomness is the homogeneous Poisson point process. Formal definition of the Poisson point process and discussion of its properties are introduced later in Section 2.1.4.

2.1.3 Summary statistics

Exploratory analysis for spatial point patterns and the validation of fitted models is often based upon the nonparametric estimation of summary statistics. These summary statistics provide information on the distribution of observed points and may be compared between observations and to theoretical reference values to illustrate how these distributions differ.

First and second order summary statistics are described in the following sections. Higher order summary statistics can also be considered, but the corresponding nonparametric estimators may be less stable if the number of points observed is not sufficiently large.

2.1.3.1 First order summary statistics

For $\underline{x}$ a realisation of a homogeneous point process $\underline{X}$ on the observation window W, obtained by first fixing W and then reporting all point locations within W, a natural unbiased estimator of the intensity ρ is given by

$$\hat{\rho} = \frac{n(\underline{x})}{|W|},$$

where $|W|$ denotes the volume of the observation window W.

In the case of inhomogeneous point processes $\underline{x}$ observed on the window W, a nonparametric kernel estimator of the intensity function is given by

$$\hat{\rho}_b(x) = \sum_{x_i \in \underline{x}} \frac{k_b(x - x_i)}{c_{W,b}(x_i)} \qquad\qquad x \in W. \tag{2.1}$$

In this expression k_b is a kernel with bandwidth $b > 0$, i.e. $k_b(x) = k(x/b)/b^d$ for some density function k in the case of $S \subseteq \mathbb{R}^d$. The term in the denominator,

$c_{W,b}(x_i)$, is an edge correction factor introduced by Diggle (1985) taking the value

$$c_{W,b}(x_i) = \int_W k_b(x - x_i)dx. \tag{2.2}$$

Nonparametric kernel estimators of the form presented in (2.1) are usually sensitive to the choice of bandwidth, b, while the choice of kernel, k, is less important (Diggle, 1985). Regardless of the choice of kernel and bandwidth, it can be shown that $\int_W \hat{\rho}_b(x)dx$ is an unbiased estimator of $\mu(W)$ (Møller and Waagepetersen, 2003).

2.1.3.2 Pair correlation function

If both the intensity, ρ, and second order product density, $\rho^{(2)}$, exist then the pair correlation function, g, is defined by

$$g(x,y) = \frac{\rho^{(2)}(x,y)}{\rho(x)\rho(y)} \qquad\qquad x, y \in S,$$

where it is taken that $g(x,y) = 0$ if either or both of $\rho(x)$ and $\rho(y)$ equal zero (Diggle, 2013).

The pair correlation function compares the joint probability of observing a pair of points to the marginal probabilities of observing each point in the pair. For a homogeneous Poisson point process the pair correlation function is equal to one for all $x, y \in S$ due to satisfaction of the property of complete spatial randomness. Values of $g(x,y) > 1$ indicate that pairs of points are more likely to occur jointly at the locations x, y than for a homogeneous Poisson point process and the converse is true for $g(x,y) < 1$.

If the point process $\underline{X}$ is stationary then g becomes translation invariant. If g is both stationary and isotropic, that is $g(x,y) = g(||x-y||) = g(r)$, then at least for small values of r, $g(r) > 1$ indicates aggregation or clustering at distances of r while $g(r) < 1$ indicates regularity at distances of r. Over larger values of r the conclusion to be drawn is less clear as there may be a mixture of aggregation and clustering on scales less than r.

Under the assumption that the pair correlation function is stationary and isotropic, g can be estimated from another summary statistic, the K-function, which is introduced alongside a description of the estimating procedure in the following section. An edge corrected kernel estimate of the pair correlation function is also given by Doguwa (1990).

2.1.3.3 K-function

The K-function for a stationary point process $\underline{X}$ on the space S is given by

$$K(r) = \frac{1}{\rho}\mathbb{E}\left[\frac{1}{N(S)} \sum_{x_i \neq x_j \in \underline{X}} \mathbb{1}\{||x_i - x_j|| < r\}\right]. \tag{2.3}$$

In this case $\rho K(r)$ is the expected number of further points within a distance of r from a randomly selected point in $\underline{X}$ (Diggle, 2013). For this reason, and to aid future understanding, we refer to the K-function as the scaled neighbourhood count function.

For a homogeneous Poisson point process $\underline{X}$ on $S \subseteq \mathbb{R}^2$ with intensity ρ and $x_i \in \underline{X}$ we have that

$$\begin{aligned}\mu(b(x_i, r)) &= \int_{b(x_i,r)} \rho dx \\ &= \rho|b(x_i, r)| \\ &= \rho\pi r^2,\end{aligned}$$

for $b(x_i, r)$ the two-dimensional ball of radius r centred at x_i. As such, the scaled neighbourhood count function for $\underline{X}$ is given by $K(r) = \pi r^2$. Values of $K(r) > \pi r^2$ are evidence for aggregation of points within $\underline{X}$ at distances of less than r. Values of $K(r) < \pi r^2$ are evidence of regularity of points within $\underline{X}$ at distances of less than r.

The transformation of the scaled neighbourhood count function named the L-function, and given by

$$L(r) = \left(\frac{K(r)}{\pi}\right)^{1/2},$$

in the case of $S \subseteq \mathbb{R}^2$, is commonly considered as an alternative to the K-function as the transformation is variance stabilising when estimated for stationary point processes using nonparametric methods (Møller and Waagepetersen, 2003). For a homogeneous Poisson point process $\underline{X}$ with intensity ρ we have $L(r) = r$ and as a result when plotting the L-function, plots of $L(r) - r$ are often considered. Values of $L(r) - r > 0$ are evidence for aggregation of points within $\underline{X}$ at distances of less than r. Values of $L(r) - r < 0$ are evidence for regularity of points within $\underline{X}$ at distances of less than r.

The K- and L-functions are cumulative functions and therefore summarise information across distances of less than or equal to r. As such, care must be taken when making inferences based upon $K(r)$ at a single distance of r. In cases where the point pattern is non-stationary, deviation from the theoretical values proposed in this section may be caused by inhomogeneous intensity rather than specific interaction between points. Furthermore, similarity in K- and L-functions for different point processes is not necessarily an indication that they are identical as very different point process models can share the same K-function (Baddeley and Silverman, 1984).

Realisations of point processes, $\underline{x}$, are typically restricted to observation windows $W \subset S$ and as such modification to the form of (2.3) is necessary to account for edge corrections. For example, an edge corrector estimator in the case where $\underline{x}$ is assumed to be a realisation of a stationary point process is given by

$$\hat{K}_{ecf}(r) = \frac{1}{\hat{\rho}} \sum_{x_i \neq x_j \in \underline{x}} \frac{\mathbb{1}\{||x_i - x_j|| < r\}}{\hat{\rho}|W \cap W_{x_j - x_i}|},$$

where $\hat{\rho}$ is an estimator of the intensity and $W_x = \{x + y : y \in W\}$ is a translation of the observation window W by $x \in \mathbb{R}^d$. The term $|W \cap W_{x_j - x_i}|$ is then an edge correction factor (Møller and Waagepetersen, 2003).

An alternative edge correction is given by minus sampling as

$$\hat{K}_{rs}(r) = \frac{1}{\hat{\rho}^2 |W_{\ominus r}|} \sum_{x_i \neq x_j, x_i \in \underline{x}, x_j \in \underline{x} \cap W_{\ominus r}} \mathbb{1}\{||x_i - x_j|| < r\},$$

where $W_{\ominus r} = \{x \in W : b(x, r) \subseteq W\}$ is the set of points in W whose distance to the boundary of W is greater than r. This is known as the simple border correction or reduced sample estimator of the K-function (Møller and Waagepetersen, 2003).

There is a loss of information in the reduced sample estimate as some pairs of points are excluded from the sum, while the edge correction factor estimate sums over all pairs of observed points. On the other hand, if the number of observed points is sufficiently large then ignoring some pairings for large values of r may be more acceptable than the potential for very large weights $1/|W \cap W_{x_j - x_i}|$ in the edge correction factor estimation.

If the pair correlation function g is isotropic, depending only upon $r = ||x - y||$, then it can be related to the derivative of the K function by

$$g(r) = \frac{K'(r)}{2\pi r},$$

in the case when $S \subseteq \mathbb{R}^2$, where the prime denotes differentiation of $K(r)$ with respect to r. However, estimators $\hat{K}$ of K are typically the sums of indicator functions, making it problematic to estimate K' from $\hat{K}$.

2.1.3.4 Empty space function

If the point process $\underline{X}$ on S is stationary then the empty space function, denoted by F, is the distribution function of the distance from any fixed point in S to the nearest point in $\underline{X}$ (Diggle, 2013).

$$F(r) = \mathbb{P}(\underline{X} \cap b(x,r) \neq \emptyset),$$

for x any point in S. The empty space function is also referred to as the spherical contact distribution function.

As in the case of the scaled neighbourhood count function, K, the empty space function may be estimated using a reduced sampling estimator based upon minus sampling. Define by $d(x,B) = \inf\{||x-y|| : y \in B\}$ the shortest distance from a point $x \in \mathbb{R}^d$ to the set $B \subset \mathbb{R}^d$. Further, let $I \subset W$ denote a finite regular grid of points chosen independently of $\underline{X}$ and $\#I_r$ denote the number of elements in the set $I_r = I \cap W_{\ominus r}$ for $r > 0$, where $W_{\ominus r}$ has previously been defined as $W_{\ominus r} = \{x \in W : b(x,r) \subseteq W\}$.

An unbiased, reduced sampling estimator of F is then given by

$$\hat{F}_{RS}(r) = \frac{1}{\#I_r} \sum_{x \in I_r} \mathbb{1}\{d(x, \underline{x}) \leq r\},$$

for $\#I_r > 0$.

A more efficient estimator of F is given by the Kaplan-Meier estimate (Baddeley and Gill, 1997)

$$\hat{F}_{KM}(r) = 1 - \prod_{s \leq r} \left(1 - \frac{\#\{x \in I : d(x,\underline{x}) = s, d(x,\underline{x}) \leq d(x,\partial W)\}}{\#\{x \in I : d(x,\underline{x}) \geq s, d(x,\partial W) \geq s\}}\right),$$

for values of $r > 0$, where ∂W denotes the boundary of the observation set W and the convention $0/0 = 0$ is used.

2.1.3.5 Nearest neighbour function

The nearest neighbour function, G, for a stationary point process $\underline{X}$ with intensity ρ is given by

$$G(r) = \frac{1}{\rho|A|} \mathbb{E}\left[\sum_{x_i \in \underline{X} \cap A} \mathbb{1}\{(\underline{X} \setminus x_i) \cap b(x_i, r) \neq \emptyset\}\right] \qquad r > 0,$$

for an arbitrary set $A \subset \mathbb{R}^d$ with $0 < |A| < \infty$. As $\underline{X}$ is assumed to be stationary, the nearest neighbour function does not depend upon the chosen set A. The G-function may be interpreted as the distribution function of the distance from a typical point in $\underline{X}$ to its nearest neighbour in $\underline{X}$.

A combination of the empty space function and nearest neighbour function is the J-function, given by

$$J(r) = \frac{1 - G(r)}{1 - F(r)},$$

for values of $F(r) < 1$. The J-function is the ratio of the probability of observing a point $x_j \in \underline{X}$ within a distance of r of any point $x \in S$ to that of observing a point $x_j \in \underline{X}$ within a distance of r of a randomly selected point $x_i \in \underline{X} \setminus x_j$. For a point process satisfying the property of complete spatial randomness, given $x_i \in \underline{X}$ the distribution of $\underline{X} \setminus x_i$ on $S \setminus x_i$ is unchanged. As a result, for such point processes the J-function satisfies $J(r) = 1$ for $r > 0$.

Both the spherical contact distribution function and nearest neighbour function are cumulative functions and thus care should be given to their interpretation at single values of r, but in general $F(r) < G(r)$ and correspondingly $J(r) < 1$ indicates clustering while $F(r) > G(r)$ and correspondingly $J(r) > 1$ indicates regularity.

Similarly to the case of the empty space function, an estimate of G can be obtained through a reduced sampling estimator

$$\hat{G}_{RS}(r) = \frac{1}{\hat{\rho}|W_{\ominus r}|} \sum_{x_i \in \underline{x} \cap W_{\ominus r}} \mathbb{1}\{d(x_i, \underline{x} \setminus x_i) \leq r\}$$

over the range of values of r for which $|W_{\ominus r}| > 0$.

A Kaplan-Meier estimate of G is given by

$$\hat{G}_{KM}(r) = 1 - \prod_{s \leq r} \left(1 - \frac{\#\{x_i \in \underline{x} : d(x_i, \underline{x} \setminus x_i) = s, d(x_i, \underline{x} \setminus x_i) \leq d(x_i, \partial W)\}}{\#\{x_i \in \underline{x} : d(x_i, \underline{x} \setminus x_i) \geq s, d(x_i, \partial W) \geq s\}}\right)$$

for values of $r > 0$.

Estimates of J may be produced as the ratio of estimators of F and G.

2.1.3.6 Comparison of nonparametric summary statistic estimates

Estimates of the nonparametric summary statistics presented in this section, for example $K(r)$, may be bounded by confidence intervals for each value of r. Confidence intervals allow the comparison of nonparametric summary statistic estimates and the testing of null hypotheses H_0 under which it is possible to simulate $\underline{X}$.

Confidence intervals for the estimator $\hat{K}$ may be obtained via a bootstrapping simulation procedure under H_0, provided that it is possible to simulate realisations of point patterns $\underline{X}$ under H_0 (Davison and Hinkley, 1997). For a given distance $r > 0$ define by $T_0(r) = T(\underline{x}, r)$ the estimate of the scaled neighbourhood count function for the observed point pattern $\underline{x}$ on the observation window W. Further, let $T_i(r) = T(\underline{X}_i, r), i \in \{1, 2, \ldots, n\}$ be the estimates of $K(r)$ for independent and identically distributed simulations $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ under H_0. The empirical distribution of $T_1(r), T_2(r), \ldots, T_n(r)$ can then be used to estimate any quantile of the distribution of $T_0(r)$ under H_0, where the precision of the estimator is limited only by the number of simulations n.

Although the $T_i(r)$ are independent of each other, vectors $(T_1(r), \ldots, T_n(r))$ are dependent for different values of $r > 0$. Caution should therefore be taken when comparing to confidence intervals across different values of r.

2.1.4 Poisson point process model

The homogeneous Poisson point processes has previously been introduced in Section 2.1.2.4 as the only stationary point process on $S \subseteq \mathbb{R}^d$ satisfying the requirements for complete spatial randomness. In this section a formal definition of the Poisson point process is given and in the following sections its properties are explored in more detail.

Let f be a density function on a set $B \subseteq S$ and let $n \in \mathbb{N} = \{1, 2, 3, \ldots\}$. A point process consisting of n independent and identically distributed points with density f is called a binomial point process of n points in B with density f, denoted by $\underline{X} \sim \text{binomial}(B, n, f)$.

A point process $\underline{X}$ on S is defined to be a Poisson point process with intensity function $\rho(x)$ if the following properties are satisfied (Møller and Waagepetersen, 2003):

1. For any $B \subseteq S$ with $\mu(B) = \mathbb{E}(N(B)) < \infty$, $N(B) \sim \text{po}(\mu(B))$, the Poisson distribution with mean $\mu(B)$.

2. For any $n \in \mathbb{N}$ and $B \subseteq S$ with $0 < \mu(B) < \infty$, conditional on $N(B) = n$, $\underline{X}_B \sim \text{binomial}(B, n, f)$ with $f(x) = \rho(x)/\mu(B)$.

We then write $\underline{X} \sim \text{Poisson}(S, \rho)$.

The process $\text{Poisson}(S, \rho)$ is called a homogeneous Poisson point process on S with rate or intensity ρ if ρ is constant. For $\rho(x)$ which varies as a function of $x \in S$ the process $\text{Poisson}(S, \rho)$ is called an inhomogeneous Poisson point process on S. The homogeneous Poisson point process on S with constant unit intensity, $\text{Poisson}(S, 1)$, is referred to as the standard or unit rate Poisson point process.

For all further discussion we restrict attention to Poisson point processes defined on spaces $S \subseteq \mathbb{R}^d$, with locally integrable intensity functions $\rho : S \to [0, \infty)$, that is $\int_B \rho(x)dx < \infty$ for all bounded $B \subseteq S$. Under this restriction, the intensity measure is locally finite, that is $\mu(B) < \infty$ for bounded $B \subseteq S$ and diffuse, that is $\mu(\{x\}) = 0$ for $x \in S$.

2.1.4.1 Superposition

A union $\cup_{i=1}^{\infty} \underline{X}_i$ of independent point processes $\underline{X}_1, \underline{X}_2, \ldots$ is called a superposition. If Poisson point processes $\underline{X}_i \sim \text{Poisson}(S, \rho_i)$, $i = 1, 2, \ldots$, are mutually independent and the intensity function $\rho = \sum \rho_i$ is locally integrable, then with probability one, point locations in $\underline{X} = \cup_{i=1}^{\infty} \underline{X}_i$ are unique, and $\underline{X} \sim \text{Poisson}(S, \rho)$ (Kingman, 1993).

2.1.4.2 Independent thinning

Let $p : S \to [0, 1]$ be a function and $\underline{X}$ be a point process on a space S. The point process $\underline{X}_{\text{thin}} \subseteq \underline{X}$ obtained by independently including each point $x_i \in \underline{X}$ in $\underline{X}_{\text{thin}}$ with probability $p(x_i)$, is said to be an independent thinning of $\underline{X}$ with retention probabilities $p(x), x \in S$. Furthermore, if $\underline{X} \sim \text{Poisson}(S, \rho)$ is subject to independent thinning with retention probabilities $p(x), x \in S$, and we define $\rho_{\text{thin}}(x) = \rho(x)p(x), x \in S$ then $\underline{X}_{\text{thin}}$ and $\underline{X} \setminus \underline{X}_{\text{thin}}$ are independent Poisson processes with intensity functions ρ_{thin} and $\rho - \rho_{\text{thin}}$ respectively (Møller and Waagepetersen, 2003).

2.1.4.3 Independent, random displacement

Consider the point process $\underline{X}$ on $T \subseteq \mathbb{R}^d$. Let $\underline{Y} = \{(x_i, v_i) : x_i \in \underline{X}\}$ be the marked point process with points in $T \subseteq \mathbb{R}^d$ and mark space $V \subseteq \mathbb{R}^d$. In the case that, conditional on $\underline{X}$, the marks v_i are independent and each distributed according to a density f_{x_i} on $\mathbb{R}^d$ which does not depend upon $\underline{X} \backslash x_i$, we may define the point pattern $\underline{X}^*$ obtained by independent, random displacements of $\underline{X}$ as $\underline{X}^* = \{x_i + v_i : x_i \in \underline{X}\}$.

If $\underline{X}$ is a homogeneous Poisson point process with constant intensity ρ and the distribution of marks f_{x_i} is independent of location x_i then $\underline{X}^*$ is also a homogeneous Poisson point process with intensity $\rho^* = \rho$, identical to that of $\underline{X}$ (Kingman, 1993).

2.1.4.4 Simulation

Taken in combination, the definition of Poisson point processes as binomial point processes and the independent thinning property of Poisson point processes provide a straightforward method for simulating Poisson point processes on bounded sets $B \subset \mathbb{R}^d$.

To simulate a homogeneous Poisson point process $\underline{X}$ on the bounded set $B \subset S \subseteq \mathbb{R}^d$ with constant intensity $\rho(x) = \rho_0$ the procedure begins by determining a box $B_0 = [-a_1, a_1] \times [-a_2, a_2] \times \ldots \times [-a_d, a_d]$ containing B, $B \subseteq B_0$. The number of points in B_0, $N(B_0) \sim \text{po}(\rho_0 2^d a_1 a_2 \ldots a_d)$, may then be sampled from the appropriate Poisson distribution. The location of the points are then sampled uniformly over B_0 by sampling each i^{th} coordinate uniformly over the appropriate box dimension, Uniform$[-a_i, a_i]$. The realisation of $\underline{X}_B$ is then obtained by disregarding those points whose locations lie outside B in the set $B_0 \setminus B$.

To simulate an inhomogeneous Poisson point process $\underline{X}$ on the bounded set B

with intensity $\rho(x)$ bounded above by a constant $\rho_0 \geq \rho(x), \forall x \in B$ with $\rho_0 > 0$ it is convenient to first simulate $\underline{Y}$, the homogeneous Poisson process on B with constant intensity ρ_0. The independent thinning of $\underline{Y}$ with retention probabilities $p(y_i) = \rho(y_i)/\rho_0$ is then a realisation of $\underline{X}$ by the independent thinning property.

2.1.4.5 Summary statistics

For $\underline{x}$ a realisation of a homogeneous Poisson point pattern with constant intensity ρ on the observation window W, the estimator

$$\hat{\rho} = \frac{n(\underline{x})}{|W|},$$

is both unbiased and the maximum likelihood estimator of ρ.

Simply due to the fact that the homogeneous Poisson point process satisfies the property of complete spatial randomness, we obtain the previously stated results that

$$\begin{aligned} g(x, y) &= 1 \\ K(r) &= \pi r^2 \\ L(r) &= r \\ J(r) &= 1, \end{aligned}$$

for $\underline{X}$ a homogeneous Poisson point process on $S \subseteq \mathbb{R}^d$.

The empty space function and nearest neighbour function are both given by

$$\begin{aligned} F(r) &= \mathbb{P}(n(b(0, r) \cap \underline{X}) > 0) \\ &= 1 - \mathbb{P}(n(b(0, r) \cap \underline{X}) = 0) \\ &= 1 - \exp(-\rho\pi r^2) \\ \Rightarrow G(r) &= 1 - \exp(-\rho\pi r^2), \end{aligned}$$

for $\underline{X}$ a homogeneous Poisson point process with intensity ρ on $S \subseteq \mathbb{R}^d$.

The tractability of summary statistics for the homogeneous Poisson process and their relative ease of simulation makes them a common reference process when studying summary statistics. Summary statistics for more advanced point process models are typically intractable.

2.1.5 Point process density

If $\underline{X}_1$ and $\underline{X}_2$ are two point processes defined on the same space S, then the distribution of $\underline{X}_1$ is said to be absolutely continuous with respect to the distribution of $\underline{X}_2$ if there exists a function $f : N_{lf} \to [0, \infty]$ such that

$$\mathbb{P}(\underline{X}_1 \in F) = \mathbb{E}[\mathbb{1}\{\underline{X}_2 \in F\} f(\underline{X}_2)] \qquad F \subseteq N_{lf}.$$

Should such a function f exist, it is referred to as the density of $\underline{X}_1$ with respect to $\underline{X}_2$. Poisson processes are not always absolutely continuous with respect to each other, but they are always absolutely continuous with respect to the standard (unit rate) Poisson point process in cases where S is bounded (Møller and Waagepetersen, 2003)

The Papangelou conditional intensity for a point process $\underline{X}$ on S with density f with respect to the standard Poisson point process is defined by

$$\lambda^*(\underline{x}, x) = \frac{f(\underline{x} \cup x)}{f(\underline{x})} \qquad\qquad \underline{x} \in N_f, x \in S \setminus \underline{x},$$

where $N_f = \{\underline{x} \subset S : n(\underline{x}) < \infty\}$ is the set of finite point configurations contained within S and it is taken that $a/0 = 0$ for $a \geq 0$ (Gaetan and Guyon, 2009). Heuristically, $\lambda^*(\underline{x}, x)dx$ may be interpreted as the conditional probability of $\underline{X}$ having a point in an infinitesimal region containing x and of size dx given that the rest of $\underline{X}$ is $\underline{x}$.

For a Poisson point process with intensity ρ the Papangelou conditional intensity is given by $\lambda^*(\underline{x}, x) = \rho(x)$, which is independent of $\underline{x}$. For each of the other Markov point processes introduced in Section 2.1.6.2, the density f is known only up to proportionality through $h \propto f$ with $h : N_{lf} \to [0, \infty)$. The Papangelou conditional intensity is therefore a particularly useful method of describing a point process because its particular formulation does not depend upon the normalising constant of f.

2.1.6 Alternative point process models

A single observation of a point process may indicate deviation from homogeneity, visible by an uneven distribution of points and detectable by the form of previously mentioned summary statistics or more straightforwardly via quadrat counts (Diggle, 2013). However, with the evidence of a single point pattern it is impossible to determine whether deviation from homogeneity is caused by an inhomogeneous underlying intensity or dependencies between point locations. Stronger evidence for a particular conclusion may be obtained by analysing replicated point patterns, repeated samples of the same process, but this data is not always available or the results conclusive.

The following sections introduce Cox processes, for which deviation from homogeneity is caused by inhomogeneous intensity, and Markov point processes, for which deviation from homogeneity is caused by dependencies between point locations.

2.1.6.1 Cox processes

Poisson point process models may be too simplistic for real data, but can form a foundation for the construction of more flexible model classes. Cox processes (Cox, 1955) are a natural extension of the Poisson point process, obtained by considering the intensity function to be the realisation of a random field.

Consider a point process $\underline{X}$ on the space S and suppose that $Z = \{Z(x) : x \in S\}$ is a non-negative random field, that is $Z(x)$ is a non-negative random variable for all $x \in S$, such that with probability one, $x \to Z(x)$ is a locally integrable function. If the conditional distribution of $\underline{X}$ given Z is a Poisson process on S with intensity function Z, then $\underline{X}$ is defined to be a Cox process driven by Z. In the case where Z is deterministic, $\underline{X}$ simply becomes a Poisson process with intensity function $\rho = Z$.

Further generalisations of the Cox process are given by Neyman-Scott processes (Neyman and Scott, 1958), log-Gaussian Cox processes (Møller et al., 1998) and shot noise Cox processes (Møller, 2003).

2.1.6.2 Markov point processes

Markov point processes form another large class of alternatives to the Poisson point process and are typically used to model interactions between points (van Lieshout, 2000). These interactions are incorporated through the specification of densities with respect to the standard Poisson point process, under conditions which ensure the Markov property. The Markov property requires that the conditional intensity $\lambda^*(\underline{x}, x)$ is dependent only upon $\underline{x} \cap b(x, R)$, those points in $\underline{x}$ which are within a distance of R of x, for some constant R. A focus on locally finite point processes means that they are often used to model repulsive behaviour, but it is also possible to model attraction through Markov point processes.

Pairwise interaction point processes (Ripley, 1977) form an introduction to the class of Markov point processes and are specified through their density with respect to the standard Poisson process

$$f(\underline{x}) \propto \prod_{x_i \in \underline{x}} \phi(x_i) \prod_{\{x_i, x_j\} \subseteq \underline{x}} \phi(\{x_i, x_j\}),$$

where ϕ is an interaction function, that is a non-negative function for which f is integrable with respect to the standard Poisson point process.

The range of interaction of the pairwise interaction point process is defined by

$$R = \inf\{r > 0 : \forall \{x, y\} \subset S, \phi(\{x, y\}) = 1 \text{ if } ||x - y|| > r\}.$$

The Poisson point process with intensity $\rho(x)$ is equivalent to the pairwise interaction point process with interaction function $\phi(x) = \rho(x)$, $\phi(\{x, y\}) = 1$ as there is no interaction between points. The range of interaction for the Poisson point process is therefore $R = 0$. Pairwise interaction point processes modelling point processes which are not Poisson are analytically intractable because of the unknown normalising constant.

Strauss processes (Strauss, 1975) are pairwise interaction point processes for which $\phi(x) = \beta > 0$ is constant and $\phi_2(\{x, y\}) = \phi_2(||x - y||) = \gamma^{\mathbb{1}\{||x-y|| \leq R\}}$ for $0 \leq \gamma \leq 1$. The parameter γ is the interaction parameter, with the strength of repulsion between points increasing as γ decreases. The extreme case in which $\gamma = 0$ is referred to as a hard core process with hard core R, as under this formulation points are prohibited from being closer than a distance of R. The other extreme case in which $\gamma = 1$ is simply the homogeneous Poisson point process with intensity β.

2.2 Comparison between pairs of spatial processes

2.2.1 Colocalisation analysis

Colocalisation analysis is a widely used technique for the analysis of fluorescence microscopy images (Adler and Parmryd, 2012). A number of colocalisation statistics have been proposed, formulated to quantify the degree to which biomolecules are deemed to interact based upon similarities in their location and evaluated using a pair of images, one for each biomolecular species. Although a commonly used term, colocalisation is poorly defined and may be used by different authors to refer to both co-occupation and correlation (Adler and Parmryd, 2012). Co-occupation is deemed to occur when sufficiently high intensity is observed in the same places for both images, while correlation occurs when there is a linear relationship between intensity values paired at the same locations.

Given pixel intensity values $m^0(x)$ and $m^1(x)$ across locations x within a region of interest χ^*, a subset of the image space $\chi^* \subseteq \chi = \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$, a number of the most commonly used colocalisation statistics are as follows:

Definition 1. Pearson's correlation coefficient (Pearson, 1895) is given by

$$r_\rho = \frac{\sum_{x \in \chi^*}(m^0(x) - \bar{m}^0)(m^1(x) - \bar{m}^1)}{\sqrt{\sum_{x \in \chi^*}(m^0(x) - \bar{m}^0)^2}\sqrt{\sum_{x \in \chi^*}(m^1(x) - \bar{m}^1)^2}} \in [-1, 1],$$

where

$$\bar{m}^0 = \frac{1}{|\chi^*|}\sum_{x \in \chi^*} m^0(x) \qquad \bar{m}^1 = \frac{1}{|\chi^*|}\sum_{x \in \chi^*} m^1(x).$$

As a measure of colocalisation, Pearson's correlation coefficient is clearly a measure of correlation. A variant on Pearson's correlation coefficient, formulated to highlight the implication of values of $m(x) = 0$ indicating absence of the biomolecule at x and $m(x) > 0$ indicating presence of the biomolecule at x, is Manders' overlap coefficient.

Definition 2. Manders' overlap coefficient (Manders et al., 1993) is given by

$$r = \frac{\sum_{x\in\chi^*} m^0(x)m^1(x)}{\sqrt{\left(\sum_{x\in\chi^*} m^0(x)^2\right)\left(\sum_{x\in\chi^*} m^1(x)^2\right)}} \in [0,1],$$

which in turn leads to the specification of the split overlap coefficients

$$k_0 = \frac{\sum_{x\in\chi^*} m^0(x)m^1(x)}{\sum_{x\in\chi^*} m^0(x)^2} \qquad\qquad k_1 = \frac{\sum_{x\in\chi^*} m^0(x)m^1(x)}{\sum_{x\in\chi^2} m^1(x)^2},$$

such that $r = \sqrt{k_0 k_1}$.

The split overlap coefficients quantify the degree of colocalisation using a pair of statistics, with k_0 quantifying the degree to which m^0 is colocalised with m^1 and k_1 quantifying the degree to which m^1 is colocalised with m^0. Such a distinction may be useful in cases where a biomolecular species is located everywhere that the species for comparison is located, while also being located in other regions. In such cases a single split overlap coefficient may be large enough to indicate a need for further investigation even if the overlap coefficient is not exceptionally large.

Manders' overlap coefficient quantifies a combination of co-occupation and correlation in unclear proportions, leading some to recommend against its use in favour of alternative statistics (Adler and Parmryd, 2010). For example, Manders' colocalisation coefficients which quantify colocalisation solely through co-occupation.

Definition 3. Manders' colocalisation coefficients (Manders et al., 1993) are given by

$$M_0 = \frac{\sum_{x\in\chi^*} m^0(x)\mathbb{1}\{m^1(x) > 0\}}{\sum_{x\in\chi^*} m^0(x)} \quad M_1 = \frac{\sum_{x\in\chi^*} m^1(x)\mathbb{1}\{m^0(x) > 0\}}{\sum_{x\in\chi^*} m^1(x)} \in [0,1].$$

A development of Manders' colocalisation coefficients which sets an automatic threshold, $t > 0$, to reduce the impact of background noise is given by Costes' approach.

Definition 4. Costes' approach (Costes et al., 2004) suggests coefficients

$$\tilde{M}_0 = \frac{\sum_{x\in\chi^*} m^0(x)\mathbb{1}\{m^0(x) > t, m^1(x) > at + b\}}{\sum_{x\in\chi^*} m^0(x)} \in [0,1]$$
$$\tilde{M}_1 = \frac{\sum_{x\in\chi^*} m^1(x)\mathbb{1}\{m^0(x) > t, m^1(x) > at + b\}}{\sum_{x\in\chi^*} m^1(x)} \in [0,1],$$

based upon a threshold value of t. The values of a and b are determined as the intercept and slope respectively of the orthogonal regression of $m^1(x)$ on $m^0(x)$, $x \in \chi^*$. Where orthogonal regression minimises the sum of the squares of the perpendicular distances to the regression line, in comparison to linear regression where the sum of the squares of the vertical distances to the regression line is minimised. The threshold is reduced from $\max\{m^0(x), (m^1(x) - b)/a\}$ to a critical value t at which the correlation of $\{(m^0(x), m^1(x)), x : m^0(x) < t \text{ or } m^1(x) < at + b\}$ is zero.

Costes' approach thresholds to ignore from statistic calculations those points x at which $m^0(x) < t$ or $m^1(x) < at + b$, across which the correlation between m^0 and m^1 is zero. A correlation of zero is deemed appropriate to consider at least one of the intensities at such locations to be representative only of noise and therefore contributing no evidence of colocalisation. Orthogonal regression is used to result in thresholds which are independent of the labelling of m^0 and m^1.

Differences in the quantity being measured between colocalisation statistics make them difficult to interpret and compare between experiments where different statistics have been used. Co-occupation based measures are typically easier to interpret as the proportion of each biomolecular species observed at shared locations, but ignore the fact that interacting biomolecules require a fixed number of biomolecules of each species and thus a linear relationship between intensities. As an alternative, correlation based measures do take into account the relationship between intensity values.

In the commonly expected presence of background noise co-occupation may be recorded at every pixel location, an issue which Costes' approach attempts to resolve through thresholding. Background noise may also impact correlation statistics, for which specification of a region of interest χ^* containing large numbers of pixels with intensity levels consistent with noise alone may mask the strength of any linear relationship between intensity values.

The majority of the presented colocalisation statistics take values in the range [0,1], with zero indicating absence of colocalisation and one indicating complete colocalisation. Pearson's correlation coefficient differs, taking values in the range [-1,1], with zero indicating absence of colocalisation and one indicating strong colocalisation. Negative values of Pearson's correlation coefficient are difficult to interpret in the biological context, although well understood in a statistical context. Fixed ranges of values provide some ability to interpret the strength of colocalisation, but there is no convincing method of analysing the significance of obtained colocalisation statistics. Instead, colocalisation statistic values may be classified into one of five categories from very weak to very strong based upon crude threshold values (Zinchuk and Zinchuk, 2008).

The biggest criticism of each of the proposed and often used colocalisation statistics is their ignorance of the spatial nature of the data within the region of interest χ^*. Techniques from other statistical fields which do take the spatial nature of the data into account are described in the following sections as potential inspirations for measures alternative to colocalisation.

In cases where a single pair of images are compared to analyse colocalisation, it is difficult to distinguish between coincidental co-occupation of biomolecular species and true interaction. When a sequence of images is collected over time, colocalisation may be quantified at each time point to provide a more reliable indicator of interaction. We approach the problem of determining interaction by estimating and comparing movement patterns between consecutive time points, on the basis that chance similar movements are less likely to occur than chance similar localisations.

On an experimental level FRET, Fluorescence (or Förster) resonance energy transfer (Clegg, 1995), is an alternative methodology which may more accurately determine interaction between biomolecular species. However, false negatives may be recorded by FRET due to the requirement that fluorophores be very closely separated, which may not be the case even for interacting biomolecules, and false positives may be recorded as a result of cross-talk or bleed-through between fluorophore colours (Piston and Kremers, 2007).

2.2.2 Metrics between probability distributions

In cases where observed spatial processes are non-negative and finite, observations may be normalised to be considered as probability densities over space,

$$\mu^0(x) = \frac{m^0(x)}{\sum_{y\in\chi} m^0(y)} \qquad\qquad \mu^1(x) = \frac{m^1(x)}{\sum_{y\in\chi} m^1(y)}.$$

which evolve over time. Similar to colocalisation statistics presented in the previous section, there are a number of methods for quantifying the distance between probability distributions including total variation distance and Hellinger distance, see for example the summary by Gibbs and Su (2002). A distance of particular interest is the Wasserstein metric (Givens and Shortt, 1984).

Definition 5. Let (χ, d) be a metric space. The Wasserstein metric between μ^0 and μ^1 on (χ, d) is $W(\mu^0, \mu^1) = \inf \mathbb{E}[d(X, Y)]$, taken over joint distributions of X and Y with marginals μ^0 and μ^1 respectively.

Importantly for our analysis, the Wasserstein metric takes into account the space on which the probability measures are defined, through d, in a manner that alternatives such as the total variation distance and Hellinger distance do not.

2.2.3 The earth mover's distance

For some investigations, it may undesirable to normalise observations m^0 and m^1 into probability densities μ^0 and μ^1 as the relative total mass of each observation is informative. In such cases an alternative, but closely related, measure of the distance between m^0 and m^1 with finite total masses is provided by the earth mover's distance (Rubner et al., 2000).

Definition 6. The earth mover's distance between two non-negative spatial processes m^0 and m^1 over the finite discrete space χ is given by

$$\begin{aligned} \mathrm{EMD}(m^0, m^1) &= \frac{\sum_{x\in\chi, y\in\chi} \hat{f}(x,y)d(x,y)}{\sum_{x\in\chi, y\in\chi} \hat{f}(x,y)} \\ \hat{f} &= \operatorname*{arg\,min}_{f\in\eta(m^0,m^1)} f(x,y)d(x,y), \end{aligned}$$

for $d(x,y)$ a cost function and $\eta(m^0, m^1)$ the set of f for which

$$\begin{aligned} f(x,y) &\geq 0 && \forall x \in \chi, y \in \chi \\ \sum_{x\in\chi} f(x,y) &\leq m^1(y) && \forall y \in \chi \\ \sum_{y\in\chi} f(x,y) &\leq m^0(x) && \forall x \in \chi \\ \sum_{x\in\chi, y\in\chi} f(x,y) &= \min\left(\sum_{x\in\chi} m^0(x), \sum_{y\in\chi} m^1(y)\right). \end{aligned}$$

If m^0 and m^1 are interpreted as spatial distributions of mass and the cost of moving unit mass from $x \in \chi$ to $y \in \chi$ is $d(x,y)$ then the earth mover's distance is the minimal total cost of rearranging m^0 into m^1 normalised by the total mass moved. The conditions on $\eta(m^0, m^1)$ ensure that only positive quantities of mass are moved, the total mass moved into $y \in \chi$ is no more than $m^1(y)$, the total mass moved out of $x \in \chi$ is no more than $m^0(x)$ and that the total amount of mass moved is the minimum of the total mass in m^0 and the total mass in m^1.

In the case where the total mass of m^0 and m^1 is identical, the earth mover's distance is equivalent to the Wasserstein distance (Levina and Bickel, 2001). The earth mover's distance is typically defined over a discrete space, but may be alternatively expressed for continuous spaces. In the case where $\sum_{x\in\chi} m^0(x) > \sum_{y\in\chi} m^1(y)$ the earth mover's distance may be equated to the Wasserstein distance by augmenting χ with the location z such that $d(x,z) = 0$ $\forall x \in \chi$ and specifying $m^1(z) = \sum_{x\in\chi} m^0(x) - \sum_{y\in\chi} m^1(y)$. A similar argument may be given for $\sum_{y\in\chi} m^1(y) > \sum_{x\in\chi} m^0(x)$.

Specification of the earth mover's distance and its use in practice has been motivated by problems in image analysis and in particular image comparison and retrieval

(Peleg et al., 1989; Rubner et al., 2000). Given a colour image, a histogram may be produced which summarises the number of pixels satisfying intensity constraints in each of the blue, red and green channels according to a proposed binning. Histograms summarising related images are expected to result in small values of the earth mover's distance when compared, allowing the single or multiple best matching images within a collection to be retrieved.

Calculation of the earth mover's distance is an optimal transportation problem, which with the augmentation of χ to equalise total mass and the restriction of $m(x) \in \mathbb{Q}$ is an assignment problem (Munkres, 1957). The computational complexity of solving such a problem is $O(n^3 \log n)$ for $n = |\chi|$, the total number of locations (Rubner et al., 2000). An implementation of the earth mover's distance is available in R (R Core Team, 2016) via the emdist package (Urbanek and Rubner, 2012), which returns the value of the earth mover's distance and, important in later considerations, the optimal set of flows $\hat{f}$.

2.2.4 Common component models

Spatial processes, $M(x)$, defined over a continuous space $x \in \chi$ may be be modelled using techniques from geostatistics (Diggle and Ribeiro, 2007; Fanshawe and Diggle, 2011). Observations of a realisation m of M at a discrete collection of locations $x \in \chi^*$ may be used to fit a Gaussian geostatistical model of the form

$$M(x) = \mu(x) + S(x) \qquad\qquad m(x) = \mu(x) + S(x) + Z(x),$$

for which μ is a mean trend term, S is a Gaussian random field and Z is an observation error term. Inference may then be made on the underlying mean trend, the spatial covariance structure and/or for predictions of $M(x)$ at unobserved locations $x \in \chi \setminus \chi^*$. In each case inference is typically carried out using Bayesian statistical techniques.

The mean trend term, $\mu(x)$, defines the average value of M at location x and may be expressed as dependent upon p spatial covariates $d_i(x)$ through the coefficients β_i

$$\mu(x) = \beta_0 + \sum_{i=1}^{p} \beta_i d_i(x).$$

The Gaussian random field, S, then specifies the spatial correlation structure of M. In cases where S may be assumed to be stationary and isotropic, the joint distribution of the collection $(S(x_1), S(x_2), \ldots, S(x_n))$ for any points $x_i \in \chi$ is an n-dimensional multivariate normal with

$$\mathbb{E}[S(x_i)] = 0 \qquad\qquad \mathrm{Cov}(S(x_i), S(x_j)) = \sigma^2 \rho(||x_i - x_j||),$$

for a standard deviation $\sigma > 0$ and suitable correlation function ρ. Examples of correlation functions include the Matérn and powered exponential (Diggle and Ribeiro, 2007), which themselves are dependent upon parameters that determine the form and scale of covariances. The observation error term is independently normally distributed at each location $x \in \chi$ with

$$\mathbb{E}[Z(x)] = 0 \qquad\qquad \mathrm{Cov}(Z(x_i), Z(x_j)) = \tau^2 \mathbb{1}\{x_i = x_j\},$$

for a standard deviation $\tau > 0$.

Gaussian geostatistical models describe real valued spatial processes M which take values in $\mathbb{R}^{\chi}$. In scenarios where M is a count process, taking values in $(\mathbb{Z}^+)^{\chi}$, an alternative formulation is provided by the Poisson generalised geostatistical model (Diggle and Ribeiro, 2007)

$$M(x) \sim \mathrm{po}(\lambda(x)) \qquad\qquad \log(\lambda(x)) = \mu(x) + S(x) + Z(x).$$

Under this formulation M is related to the Gaussian geostatistical model by the log link function, μ is an offset and Z provides the potential for extra-Poisson variation.

In cases where inference is to be made in relation to two spatial processes, M^1 and M^2, a class of common component models have been developed (Fanshawe and Diggle, 2011). Gaussian common component models consider processes M^1 and M^2 as a combination of independent Gaussian random fields, S^1 and S^2 corresponding to the respective spatial process and S^0 which is common to both processes

$$\begin{aligned} M^1(x) &= \mu^1(x) + \sigma_{01} S^0(x) + S^1(x) + Z^1(x) \\ M^2(x) &= \mu^2(x) + \sigma_{02} S^0(x) + S^2(x) + Z^2(x). \end{aligned}$$

The terms σ_{01} and σ_{02} incorporate the degree of influence of the common component to each of the processes. Equivalent common component models have been proposed for spatial count data and applied to investigate common causes of disease (Knorr-Held and Best, 2001).

Geostatistical models and in particular common component models were considered as one approach to make inference on the localisation of TACC3 and EB3. Image data could be taken as a surrogate for the quantity of each biomolecular species across the space of pixel locations, χ^*. The relative influence of the common component S^0 in comparison to the observation specific components could be summarised as a measure of colocalisation. Progress with this approach was limited by the requirement for assumptions on correlation structure, a lack of information with which to specify μ and the computational challenges of fitting the models, it is therefore not described further in later chapters.

2.2.5 Object tracking techniques

The approach presented in Chapter 4 determines the significance of the dependence between the movement patterns of EB3 and TACC3, estimated as the solution of an inverse problem the results of which are then summarised over subregions. The purpose of this is to take advantage of the information within the temporal evolution of each spatial distribution on the basis that coincidental dependence between movement patterns is less likely to occur by chance than coincidental similarity between spatial distributions at a single time point.

In cases of small to medium numbers of individually resolvable objects there exist a number of object tracking methodologies, some of which have been applied to cellular images (McFarlane and Schofield, 1995). An investigation of the effectiveness of a wide range of such methods for this application has been compiled by Chenouard et al. (2014). Each methodology may be broadly divided into two phases, the first of identifying objects in each image and the second of then connecting object locations between time points to propose movement trajectories. Performance at both steps may be improved by careful specification of parameters based upon prior information.

In each of the images that we wish to investigate there are vast numbers of each biomolecular species of interest, resulting in an approximately continuous spatial distribution. This precludes the use of specific object tracking methodologies and we instead introduce a methodology based upon estimation of aggregate or bulk movement patterns. As the proposed technique does not require specification of parameters it is expected that it will also be more widely applicable to alternative data sets arising from cell imaging and other fields.

2.3 Hypothesis testing

2.3.1 Statistical test

Given data, x, a realisation of a random variable, X, a null hypothesis, H_0, may be tested using a test statistic function, $t(x) \in \mathbb{R}$. The null hypothesis is rejected if the probability that $t(X)$ is as or more extreme than the observed $t(x)$ is less then a specified significance level, α. The definition of as or more extreme is included within the specification of the null hypothesis and may be represented by the binary relation $\succ$, with $a \succ b$ representing that a is as or more extreme than b. The binary relation must be reflexive, i.e. $a \succ a$, transitive, i.e. $a \succ b$ and $b \succ c$ implies $a \succ c$, and total, i.e. for all a and b at least one of $a \succ b$ and $b \succ a$. For example, a one-sided test at the upper tail uses the binary relation specified by $a \succ b \iff a \geq b$, while a two-sided test uses the binary relation specified by $a \succ b \iff |a| \geq |b|$.

The null hypothesis is rejected at significance level α if the observed value of $t(x)$ lies within the corresponding rejection region, $R(\alpha)$, satisfying

$$\mathbb{P}(t(X) \in R(\alpha)|H_0) = \alpha \qquad\qquad \forall a \in R(\alpha), b \in \mathbb{R} \setminus R(\alpha), a \succ b.$$

Equivalently, the null hypothesis is rejected at significance level α if the p-value, $p = \mathbb{P}(t(X) \succ t(x)|H_0)$, is less than α. In the case where the distribution function of $t(X)$, F_t, is known and invertible determination of $R(\alpha)$ or p may be carried out analytically. In cases where the distribution function is unknown or cannot be inverted to obtain an analytical expression for p or $R(\alpha)$ an alternative testing procedure must be used, a general and more specialised example of which are presented in the following sections.

A hypothesis testing methodology is valid if the probability of a type I error, that is the probability of rejecting H_0 when it is true, is no greater than the significance level, α, for any $\alpha \in [0, 1]$. The methodology described in this section is valid by construction, but the criteria is one against which other testing procedures may be measured.

2.3.2 Bootstrap test

Bootstrap methods allow the testing of a null hypothesis in cases where F_t is unknown or not invertible, by instead performing calculations using repeated samples, $\{t_1(x), t_2(x), \ldots, t_n(x)\}$, of $t(X)$ under H_0. There exist a number of bootstrap tests, each of which differ in the production of samples but share a common testing methodology given this information. For example, in the case where $X = (X_1, X_2, \ldots, X_n)$ is a collection of n independent, identically distributed random variables a collection of samples may be generated, with $t_i(x) = t(x^i)$ arising from x^i a random sample with replacement from $x = (x_1, x_2, \ldots, x_n)$ of size n. The permutation test is an additional example and is described in more detail in the following section.

The observed test statistic value, $t(x)$ denoted by t_0, may be taken in combination with simulated test statistics to produce the set $\{t_0, t_1, t_2, \ldots, t_n\}$. This set of statistics may then be ordered according to the binary relation $\succ$

$$t_{(n+1)} \succ t_{(n)} \succ \ldots \succ t_{(2)} \succ t_{(1)},$$

and in the case where simulated test statistics are independent a rejection region specified as

$$R(\alpha) = \{a \in \mathbb{R} : a \succ t_{(n+1-\lfloor (n+1)\alpha \rfloor)}, t_{(n+1-\lfloor (n+1)\alpha \rfloor)} \not\succ a\}.$$

This test is valid as the probability that $t(x) = t_0$ lies in any α proportion of the collection of t_i, and therefore by definition the most extreme α proportion, is less than or equal to α when the t_i are independent and identically distributed. The test may be expressed in terms of the p value

$$p = \frac{1}{n+1} \sum_{i=0}^{n} \mathbb{1}\{t_0 \succ t_i\},$$

rejecting H_0 at significance level α when $p \leq \alpha$.

A hypothesis testing procedure is exact if the probability of a type I error is identically the significance level α for all $\alpha \in [0, 1]$. Under the above formulation the bootstrap test is valid but not exact. To illustrate this consider $n = 9$ and $\alpha = 0.05$, the resulting set $R(0.05)$ are those values which are strictly more extreme than every t_i. As a result the probability of rejecting H_0 is zero, because $t(x) = t_0$ is one of the t_i, and thus the probability of a type I error is zero, strictly less than 0.05. However, the number of samples n may be chosen such that $(n+1)\alpha \in \mathbb{Z}$ to result in an exact bootstrap test when samples t_i are independent.

An alternative formulation of the bootstrap test which is both valid and exact is obtained by first specifying $k = n + 1 - \lfloor (n+1)\alpha \rfloor$ and counts c^+ and c^0 as follows

$$\begin{aligned} c^+ &= \sum_{i=0}^{n} \mathbb{1}\{t_i \succ t_{(k)}, t_{(k)} \not\succ t_i\} \\ c^0 &= \sum_{i=0}^{n} \mathbb{1}\{t_i \succ t_{(k)}, t_{(k)} \succ t_i\}. \end{aligned}$$

Intuitively, c^+ is the number of t_i which are strictly more extreme than $t_{(k)}$ and c^0 is the number of t_i which are exactly as extreme as $t_{(k)}$. A randomisation test function $\phi(x) = \mathbb{P}(\text{reject } H_0)$, given by

$$\phi(x) = \begin{cases} 1 & \text{if } t(x) \succ t_{(k)}(x) \text{ and } t_{(k)}(x) \not\succ t(x) \\ \frac{(n+1)\alpha - c^+(x)}{c^0(x)} & \text{if } t(x) \succ t_{(k)}(x) \text{ and } t_{(k)}(x) \succ t(x) \\ 0 & \text{if } t(x) \not\succ t_{(k)}(x) \text{ and } t_{(k)}(x) \succ t(x), \end{cases}$$

then determines rejection of H_0 (Chung and Romano, 2013).

The size of the test is

$$\begin{aligned} \mathbb{E}_X[\phi(X) | H_0 \text{ true}] &= \frac{c^+}{n+1} + \frac{(n+1)\alpha - c^+}{c^0} \times \frac{c^0}{n+1} \\ &= \alpha, \end{aligned}$$

as under H_0 the t_i are identically distributed and thus $t(x) = t_0$ may be considered to be a single random sample from the collection of $t_{(i)}$. This formulation of the bootstrap test is therefore valid and exact.

2.3.3 Permutation test

Under some formulations of H_0 in cases where X is a collection of, potentially dependent, random variables a permutation test may be carried out. The suitability of a permutation test is dependent upon the existence of a collection of operators, $\lambda \in \Lambda$, under which $t(X) \overset{\mathrm{d}}{=} t(\lambda X)$, with $\overset{\mathrm{d}}{=}$ used to denote equality in distribution. This is known as exchangeability under the action of all $\lambda \in \Lambda$, which we will refer to as Λ-exchangeability. An example is the scenario in which the variables within X are divided into two collections, those with distribution function F_1 and those with distribution function F_2, with H_0 specifying that $F_1 \equiv F_2$. The random variables X are then exchangeable under the set of operators Λ equivalent to relabelling of the random variables as members of the two collections.

Specifying $|\Lambda| = n + 1$ and the identity operator $\lambda_0 \in \Lambda$ such that $\lambda_0 X = X$, the bootstrap test conducted using the set of all samples generated by allowable permutations $\{t(\lambda_0 x), t(\lambda_1 x), \ldots, t(\lambda_n x)\}$ is referred to as the exact permutation test (Edgington, 1964). A benefit of the permutation test is that the distribution function of the test statistic, F_t, need not be specified. Permutation tests simply require structure within the data X under which there is Λ-exchangeability for some set of operators Λ.

In cases where the size of Λ and the time required to calculate each $t(\lambda x)$ combine to be infeasible an approximate permutation test may be carried out. The approximate permutation test is a bootstrap test evaluated using samples $t(\lambda x)$ for $\lambda \in \Lambda'$, a randomly sampled subset of Λ of size n' which includes λ_0 (Edgington, 1969). The resulting test function is dependent upon both X and Λ', $\phi(X, \Lambda')$, but provided Λ' is sampled from Λ independently of X the hypothesis test is both valid and exact, since

$$\begin{aligned}
\mathbb{E}_{\Lambda',X}[\phi(X,\Lambda')|H_0] &= \mathbb{E}_{\Lambda'}[\mathbb{E}_X[\phi(X,\Lambda')|H_0 \text{ true}]|H_0 \text{ true}] \\
&= \mathbb{E}_{\Lambda'}[\alpha|H_0 \text{ true}] \\
&= \alpha.
\end{aligned}$$

Although the approximate permutation test remains valid and exact, its power is reduced in comparison to the exact permutation test under the complete set of operators, Λ. When testing at the $\alpha = 0.05$ significance level with $n' = 1000$ the power of the approximate test is at least 94.5 percent of that for the exact test, rising to at least 98.3 percent when $n' = 10000$ (Jöckel, 1986).

2.3.4 Multiple hypothesis testing

As described in Section 2.3.1, an exact hypothesis test carried out at significance level α has a probability equal to α of rejecting the null hypothesis in cases where it is true, also known as the size of the test or the probability of a type I error. When a single test is carried out specification of the significance level therefore limits the probability of such an error. In cases where multiple tests are carried out, each has a corresponding probability of a type I error and the combination of these probabilities must be taken into account to avoid making misleading conclusions.

A measure of the error rate for a collection of hypothesis tests is the family wise error rate, the probability of making at least one type I error (Shaffer, 1995). Under the assumption that the results of each of n exact tests are independent and that each test is carried out at the same significance level, α, the family wise error rate is $1 - (1 - \alpha)^n$. As an illustration, consider $n = 16$ tests carried out at significance level $\alpha = 0.05$. The resulting false positive rate is then $1 - (1 - 0.05)^{16} = 0.56$, i.e. there is a greater than 50 percent chance of rejecting the null hypothesis for at least one of the tests when all of the null hypotheses are true.

A method for the control of the family wise error rate across n tests is the Bonferroni correction, for which each individual test is carried out at the reduced significance level of α/n (Dunn, 1961). The family wise error rate in the case of independent tests, $1 - (1 - \alpha/n)^n \leq \alpha$ by Boole's inequality, is controlled by the Bonferroni correction to be no more than the significance level α. However, for large numbers of comparisons the Bonferroni correction can result in tests which lack power and are conservative when the results of individual tests are positively correlated (Simes, 1986). There exist alternatives to the Bonferroni correction which control the family wise error rate and maintain greater statistical power, including the stepdown max-based permutation testing procedure of Romano and Wolf (2005).

Under the null hypothesis p-values are uniformly distributed on the interval $[0, 1]$. The empirical distribution of p-values may therefore be compared to the Uniform$[0, 1]$ distribution in a single hypothesis test, using for example the one-sample Kolmogorov-Smirnov test (Massey, 1951). In the case of repeated application of the same hypothesis test to multiple sets of identically distributed data, rejection of this omnibus test may then be taken as evidence against the null hypothesis.

2.3.5 Effect size

When comparing the difference or dependence between two sets of observations, it is often important to investigate effect size alongside corresponding significance tests. For example, in the case of samples $\underline{x} = \{x_1, x_2, \ldots, x_{n_x}\}$ observed under control conditions and samples $\underline{y} = \{y_1, y_2, \ldots, y_{n_y}\}$ observed under treatment conditions following an intervention, a statistical test may tell us if the change in observation mean, $\bar{y} - \bar{x}$, is significant but nothing about the magnitude of the effect size. In cases where n_1 and n_2 are very large, a very small treatment effect may result in rejection of the null hypothesis of no difference between the means of the populations from which x and y are sampled, but be of no practical importance.

In the case of comparing the difference between sets $\underline{x}$ and $\underline{y}$, Cohen's d statistic (Cohen, 1992) is a measure of effect size. The value of d depends upon sample means, $\bar{x}$ and $\bar{y}$, and standard deviations, s_x and s_y,

$$\begin{aligned} d &= \frac{\bar{x} - \bar{y}}{s} \\ s &= \sqrt{\frac{(n_x - 1)s_x^2 - (n_y - 1)s_y^2}{n_x + n_y - 2}}. \end{aligned}$$

The magnitude of d may be assessed by comparison to thresholds (Cohen, 1992), with $|d| \leq 0.2$ denoting a negligible effect size, $0.2 < |d| \leq 0.5$ a small effect size, $0.5 < |d| \leq 0.8$ a medium effect size and anything greater than 0.8 a large effect size. Alternatively, the magnitude of d may be assessed in comparison to its standard deviation, s_d, given by

$$s_d^2 = \left(\frac{n_x + n_y}{n_x n_y} + \frac{d^2}{2(n_x + n_y - 2)}\right)\left(\frac{n_x + n_y}{n_x + n_y - 2}\right).$$

Chapter 3

Differences between collections of point patterns

This chapter introduces in more detail the methodology proposed for testing for differences between collections of point patterns. The material has been made available as a working paper (Honnor et al., 2017a), and is presented in this chapter in a slightly reduced format as background information common to the whole thesis is presented in Chapter 2.

An introduction is first given to put the biological question in context, describe the data and formulate a statistical question which aims to answer the biological question with the data available. Following this the methodology is presented in Section 3.2. In Section 3.3 a validation study design is presented and the results analysed. Application of the methodology to a set of TACC3 biological data follows in Section 3.4 before a summary of the conclusions is presented in Section 3.5.

3.1 Introduction

Advances in sensor (Kanoun and Trankler, 2004) and storage technology (Grochowski and Hoyt, 1996) allow parallelisation of data collection across sensor networks and continuous monitoring. Improvements in communication networks have also made collected data more accessible. One result of this is the production of large, specialised spatial point pattern data sets, the analysis of which requires development of new statistical techniques. A particular area in which this is apparent is the imaging of large numbers of biological samples at high magnifications. The resulting images may be analysed computationally to determine the location of subcellular structures of interest. Further investigation can shed light on the inner workings of the cell and the effect of applied external conditions.

This chapter introduces and applies a methodology for comparing the structure of point patterns with a particular biological application in mind, the structure of microtubules within kinetochore fibers, but with further applicability to more general data sets. Analysis of microtubule structure is of particular importance as microtubules perform a vital role during chromosome separation in mitosis, where errors can lead to aneuploidy, a common cause of genetic disorders.

Point pattern data comprising observations from two populations such as those analysed in this chapter may arise in numerous ways. For example, Plant locations, divided into two populations based upon the species of plant (Mateu et al., 2014), and the location of neurons within the brain, divided based upon whether the individual suffers from mental illness (Diggle et al., 1991).

It may be desirable to determine if there is variation in the structure of point patterns to make inference on underlying differences between the two populations. Such variation may occur consistently, but at a small enough scale to make detection by eye impossible. This chapter describes a statistical methodology for application to point patterns and a class of marked point patterns, to test for the existence of structural differences between two collections of point patterns.

One modelling approach is to model each of the populations individually and compare the models. Due to the wide variety of models and difficulties fitting them to data, we instead compare the collections of point patterns directly using a number of nonparametric summary statistics, which are then combined across and compared between collections to produce test statistics. Nonparametric permutation testing is then be used to quantify the significance of reported test statistic values.

We introduce a variety of test statistics, such as the number of points and the distances between points, and a number of comparison methods, for example pointwise and functional comparisons. Decisions on which of the suggested testing procedures are most suitable will depend upon the type of pattern structure of interest, the required sensitivity of the testing procedure and the desired interpretability of the test statistic.

3.1.1 Mathematical representation of data

The methodology in the following section is designed for application to point patterns $\underline{x} \in \chi_2$, where χ_2 is the set of all finite point patterns on $\mathbb{R}^2$. That is $\underline{x} = (x_1, x_2, \dots, x_{n(\underline{x})})$ with $x_j \in \mathbb{R}^2$ and $n(\underline{x}) \in \mathbb{N}$. The point pattern $\underline{x}$ is represented as an ordered collection of point locations for identifiability purposes. To summarise,

$$\chi_2 := \{(x_1, x_2, \dots, x_{n(\underline{x})}) : n(\underline{x}) \in \mathbb{N}, x_i \in \mathbb{R}^2 \text{ for } i = 1, 2, \dots, n\}.$$

We will consider a collection of point patterns $\underline{x}^i \in \chi_2$ indexed by the set of

$i \in I$. The notation $\underline{x}^J = (\underline{x}^i : i \in J)$ is used to represent the collection of point patterns indexed by the set J, where ordering is again simply for identifiability. Two collections of point patterns, indexed by I_0 and I_1 and denoted by $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$, are obtained by partitioning the set I. Under this notation, x^i_j is the location of the j^{th} point in the i^{th} point pattern, $\underline{x}^i$.

Additional methodologies are also designed for application to marked point patterns $\underline{y} \in \chi_2^+$. The point space χ_2^+ is the set of all finite point patterns on $\mathbb{R}^2$ augmented with a third coordinate equal to zero, with each point x_j assigned a corresponding mark v_j which is itself a vector in $\mathbb{R}^2$ augmented with a third coordinate equal to h. That is $\underline{y} = ((x_1, v_1), (x_2, v_2), \ldots, (x_{n(\underline{y})}, v_{n(\underline{y})}))$ with $x_j \in \mathbb{R}^2 \times \{0\}$ and $v_j \in \mathbb{R}^2 \times \{h\}$. Again, the points of $\underline{y}$ are ordered for identifiability purposes. The reason for the particular formulation of χ_2^+ will become clear in light of the particular application described in Section 3.4, but in brief is chosen to represent paired point patterns in parallel planes separated by a distance of h by a marked point pattern. To summarise,

$$\begin{aligned}\chi_2^+ := \{((x_1, v_1), &(x_2, v_2), \ldots, (x_{n(\underline{y})}, v_{n(\underline{y})})) : n(\underline{y}) \in \mathbb{N}, x_i \in \mathbb{R}^2 \times \{0\} \\ &\text{and } v_i \in \mathbb{R}^2 \times \{h\} \text{ for } i = 1, 2, \ldots, n(\underline{y})\}. \qquad (3.1)\end{aligned}$$

We again consider a collection of marked point patterns $\underline{y}^i \in \chi_2^+$ indexed by the set I, with $\underline{y}^J = (\underline{y}^i : i \in J)$ the collection of marked point patterns indexed by the set J. The sets of indices I_0 and I_1 are defined analogously to those for the unmarked point patterns.

3.1.2 Statistical problem

Point patterns $\underline{x}^{I_0}$ may be considered to be independent realisations of a point process $\underline{X}^0$ and point patterns $\underline{x}^{I_1}$ to be independent realisations of a point process $\underline{X}^1$. The aim of the proposed methodology for point patterns is to make inference on the existence and form of a difference between the processes $\underline{X}^0$ and $\underline{X}^1$ using the available data $\underline{x}^I$.

Similarly, marked point patterns $\underline{y}^{I_0}$ may be considered to be independent realisations of a marked point process $\underline{Y}^0$ and marked point patterns $\underline{y}^{I_1}$ to be independent realisations of a marked point process $\underline{Y}^1$. The aim of the proposed methodology for marked point patterns is to make inference on the existence and form of a difference between the processes $\underline{Y}^0$ and $\underline{Y}^1$ using the available data $\underline{y}^I$.

3.2 Quantification of differences in point pattern structure

3.2.1 Summary statistics

Summary statistics provide information on the distribution of observed points and may be compared between observations. Comparison may also be made to the homogeneous Poisson process for which frequently used summary statistics are tractable. Summary statistics for more advanced point process models are typically intractable. This section introduces a number of summary statistics, each of which summarises a property of a single (marked) point pattern, with the following section combining these values over the index sets I_0 and I_1 to produce test statistics.

The data which is the focus of our study, and is described in more detail in Section 3.4, is comprised of measurements of multiple observed point patterns which may only be compared according to the relative location of points, due to the absence of a common fixed location and orientation with which to specify a consistent co-ordinate system. As a result, while intensities could be estimated for each individual point pattern, it is not possible to combine these separate estimates into intensity profiles for the generating processes. An alternative to determining intensity profiles of the generating processes is the comparison of features of the estimated intensity of each observation, for example the smoothness of the intensity quantified via estimation of the parameters of the correlation function of a fitted Gaussian process. We instead investigate the patterns through summary statistics which may be accurately estimated using only the information obtained from relative point locations. Such summary statistics can also be more intuitive for explanation to non-statisticians, such as biologists for whom our analysis aims to inform.

For $\underline{X}$ a homogeneous point process on the observation window W, the intensity is given by

$$\rho = \frac{\mu(W)}{|W|},$$

where $|W|$ denotes the volume of the observation window W. For $\underline{x}$ a realisation of a homogeneous point process, obtained by first fixing the observation window W and then reporting all point locations within W, the intensity may be estimated by

$$\hat{\rho} = \frac{n(\underline{x})}{|W|}, \tag{3.2}$$

as introduced earlier in Section 2.1.3.1.

The nearest neighbour of point x_j in pattern $\underline{x}$ may be denoted by $\mathrm{nn}(x_j)$

$$\mathrm{nn}(x_j) = \left\{ x_k : k = \arg\min_{l} ||x_l - x_j|| \right\},$$

where the set definition accounts for the possible existence of more than one nearest neighbour. The nearest neighbour distance of point x_j may then be denoted by $\text{nnd}(x_j)$ with

$$\text{nnd}(x_j) = \inf_{x \in \text{nn}(x_j)} \{||x_j - x||\},$$

the unique infimum over $\text{nn}(x_j)$ which gives the Euclidean distance between x_j and its nearest neighbours in $\underline{x}$. The mean nearest neighbour distance for the point process $\underline{X}$ is then

$$\overline{\text{nnd}}(\underline{X}) = \mathbb{E}[\text{nnd}(x), x \sim \text{Uniform}\{\underline{X}\}],$$

an estimator of which for the point pattern $\underline{x}$ is given by

$$\overline{\text{nnd}}(\underline{x}) = \frac{1}{n(\underline{x})} \sum_{j=1}^{n(\underline{x})} \text{nnd}(x_j). \tag{3.3}$$

The graph produced by including only those edges which represent a nearest neighbour relationship is unlikely to be connected. As a result, the disjoint connected subgraphs may be translated to produce clearly different point patterns which retain the same nearest neighbour properties. The minimum spanning tree is a connected alternative graph for which the translation of subgraphs typically results in a difference in a corresponding mean minimum spanning distance.

Considering the weighted graph associated with the point pattern $\underline{x}$ denoted by

$$\mathcal{G}(\underline{x}) = (V(\underline{x}), E(\underline{x}), D(\underline{x})),$$

the vertex set of $\mathcal{G}(\underline{x})$, denoted by $V(\underline{x})$, is the set of point locations $x_j \in \underline{x}$. The graph $\mathcal{G}(\underline{x})$ is the complete graph on $V(\underline{x})$, meaning that the edge set $E(\underline{x})$ is the set of edges joining every vertex to every other vertex. Graph edges are weighted by the distances between points, with the edge joining the vertices at x_j and x_k being attached the weight $||x_j - x_k||$, producing the set of edge weights $D(\underline{x})$. The minimum spanning tree of the graph $\mathcal{G}(\underline{x})$, denoted by $\mathcal{G}^*(\underline{x})$, is the spanning subgraph of $\mathcal{G}(\underline{x})$ with minimum weight. Let $E^*(\underline{x}) \subset E(\underline{x})$ denote the edge set of $\mathcal{G}^*(\underline{x})$ and $D^*(\underline{x}) \subset D(\underline{x})$ denote the corresponding set of edge weights. The expected edge length of the minimum spanning tree of the point process $\underline{X}$, referred to as the mean minimum spanning distance $\overline{\text{msd}}(\underline{X})$, is then

$$\overline{\text{msd}}(\underline{X}) = \mathbb{E}[d, d \sim \text{Uniform}\{D^*(\underline{X})\}].$$

An estimate of the mean minimum spanning distance for a point pattern $\underline{x}$ is given by

$$\overline{\text{msd}}(\underline{x}) = \frac{1}{n(\underline{x}) - 1} \sum_{d \in \mathcal{D}^*(\underline{x})} d, \tag{3.4}$$

where the divisor is $n(\underline{x}) - 1$ as the minimum spanning tree of n points is made up of $n-1$ edges.

The K-function, or scaled neighbourhood count function, is described previously in Section 2.1.3.3 for point processes, with examples of edge corrected estimators for point patterns. As part of this analysis we consider the edge corrected estimator $\hat{K}(\underline{x}, r)$ given by

$$\hat{K}(\underline{x}, r) = \frac{|W|}{n(\underline{x})^2} \sum_{j \neq k} e_{j,k} \mathbb{1}(||x_j - x_k|| \leq r), \tag{3.5}$$

where $e_{j,k}$ is an edge correction term such that $e_{j,k}$ is the proportion of the circumference of the circle with centre x_j and radius $||x_j - x_k||$ which is contained within the observation window W (Diggle et al., 2000).

The G-function, or nearest neighbour function, is described previously in Section 2.1.3.5 for point processes, with examples of edge corrected estimators for point patterns. As part of this analysis we consider the uncorrected estimator $\hat{G}(\underline{x}, r)$ given by

$$\hat{G}(\underline{x}, r) = \frac{1}{n(\underline{x})} \sum_{j=1}^{n(\underline{x})} \mathbb{1}\left[\mathrm{nnd}(x_j) \leq r\right], \quad r \in [0, \infty), \tag{3.6}$$

for reasons described in more detail in Section 3.4.4.

The location of points within marked point patterns may be summarised by the previous statistics described in this section. We propose a further summary statistic for the marked point pattern $\underline{y}$ which may be formulated using only the marks $v_j \in \mathbb{R}^2 \times \{h\}$, each of which represents the vector connecting related points in paired point patterns lying in parallel planes separated by a distance of h. Let $\hat{v}_j$ denote the unit vector in the direction of the mark v_j and β be given by

$$\beta = \sum_{j=1}^{n(\underline{y})} \hat{v}_j. \tag{3.7}$$

We then propose the effective force transference summary statistic for the marked point process $\underline{Y}$ as

$$\mathrm{EFT}(\underline{Y}) = \mathbb{E}[\cos(\theta(v, \beta)), (x, v) \sim \mathrm{Uniform}\{\underline{Y}\}],$$

where $\theta(v, \beta)$ is the angle between vectors v and β. An estimate of the effective force transference summary statistic for a marked point pattern $\underline{y}$ is given by

$$\mathrm{EFT}(\underline{y}) = \frac{1}{n(\underline{y})} \sum_{j=1}^{n(\underline{y})} \cos(\theta(v_j, \beta)).$$

This approximation may alternatively be expressed as

$$\begin{aligned}\mathrm{EFT}(\underline{y}) &= \frac{1}{n(\underline{y})}\sum_{j=1}^{n(\underline{y})}\frac{\hat{v}_j \cdot \beta}{||\beta||}\\ &= \frac{||\beta||}{n(\underline{y})}. \end{aligned} \tag{3.8}$$

A further expression of the effective force transference estimator is given by

$$\begin{aligned}\mathrm{EFT}(\underline{y}) &= \frac{1}{n(\underline{y})}\sqrt{\beta \cdot \beta}\\ &= \frac{1}{n(\underline{y})}\sqrt{\sum_{j=1}^{n(\underline{y})}\hat{v}_j \cdot \sum_{k=1}^{n(\underline{y})}\hat{v}_k}\\ &= \frac{1}{n(\underline{y})}\sqrt{\sum_{j=1}^{n(\underline{y})}\sum_{k=1}^{n(\underline{y})}\cos(\theta(v_j, v_k))}. \end{aligned} \tag{3.9}$$

From this formulation it is clear that the effective force transference is dependent only upon the angles between mark vectors, independently of point locations and the length of mark vectors.

As β is the sum of $n(\underline{y})$ unit length vectors, $0 \leq ||\beta|| \leq n(\underline{y})$ and $\mathrm{EFT}(\underline{y}) \in [0, 1]$. The expression of effective force transference given in (3.9) makes it clear that $\mathrm{EFT}(\underline{y}) = 1$ if and only if all mark vectors are parallel and the greater the angles between lines, the smaller the value of the effective force transference. As a result $\mathrm{EFT}(\underline{y})$ is a measure of the degree to which mark vectors v_j are similarly oriented. Further to the general interpretation of the effective force transference test statistic for any $\underline{y} \in \chi_2^+$, Section 3.4.4 introduces a specific interpretation in the case of the TACC3 data which motivates the name.

There are a number of directional statistics which could have been considered instead of the effective force transference test statistic, such as the concentration parameters of the Von Mises-Fisher distribution (Mardia, 1975) or Kent distribution (Kent, 1982). Our observed unit directions are akin to locations on the unit 2-hemisphere (half of the 2-sphere in $\mathbb{R}^3$), while both of these distributions may be specified over the 2-sphere. This difference and the physical interpretation of the effective force transference in light of the TACC3 data lead us to favour investigation of the effective force transference over the potential alternatives.

3.2.2 Test statistics

The first order summary statistic, $n(\underline{x})$, may be compared between point pattern collections using the pattern size test statistic

$$\delta_N(I) = \frac{1}{|I_0|}\sum_{i\in I_0} n(\underline{x}^i) - \frac{1}{|I_1|}\sum_{i\in I_1} n(\underline{x}^i).$$

In some cases, and specifically in the case of the data we investigate in Section 3.4, point patterns are reported by methods other than first fixing the observation window W and then reporting all point locations within W. In such cases the area of the observation window, $|W|$, may be considered to be random and compared between point pattern collections using the pattern area test statistic

$$\delta_W(I) = \frac{1}{|I_0|}\sum_{i\in I_0} |W^i| - \frac{1}{|I_1|}\sum_{i\in I_1} |W^i|.$$

The estimated homogeneous intensity of point patterns, $\hat{\rho}(\underline{x})$, defined in (3.2), may be compared between collections using the intensity test statistic

$$\delta_\rho(I) = \sum_{i\in I_0} \omega_0(\underline{x}^i)\hat{\rho}(\underline{x}^i) - \sum_{i\in I_1} \omega_1(\underline{x}^i)\hat{\rho}(\underline{x}^i),$$

where weights ω_0 and ω_1 may be formulated to produce unweighted test statistics with

$$\omega_0(\underline{x}^i) = \frac{1}{|I_0|}, \qquad \omega_1(\underline{x}^i) = \frac{1}{|I_1|},$$

or weighted test statistics with

$$\omega_0(\underline{x}^i) = \frac{n(\underline{x}^i)}{\sum_{j\in I_0} n(\underline{x}^j)}, \qquad \omega_1(\underline{x}^i) = \frac{n(\underline{x}^i)}{\sum_{j\in I_1} n(\underline{x}^j)}. \tag{3.10}$$

The notation δ_ρ is used to denote unweighted test statistics, with $\delta_{\rho,\omega}$ used to denote weighted test statistics. Uniform weighting takes into account the evidence of each point pattern equally, while weighting according to the number of points attaches equal weight of evidence to each point.

Pattern size, area and intensity test statistics summarise point pattern properties through single numbers dependent only upon first order properties, independently of the location of observed points x_j within the observation window W. In cases where point patterns are expected to be realisations of homogeneous Poisson point patterns, specification of the intensity completely describes the distribution of the number of points and their locations, while specification of the number of points completely describes the distribution of point locations. However, knowledge of the number of points or the estimated density under the assumption of homogeneity is inadequate for specification of the distribution of point locations for inhomogeneous

Poisson point processes and alternative classes of point processes. As a result, test statistics based upon the second order property of distances between points may be more suitable when the patterns are not believed to be realisations of homogeneous Poisson processes.

We propose comparison of $\overline{\text{nnd}}(\underline{x})$, defined in (3.3), between collections using the mean nearest neighbour test statistic

$$\delta_{\text{nnd}}(I) = \sum_{i \in I_0} \omega_0(\underline{x}^i)\overline{\text{nnd}}(\underline{x}^i) - \sum_{i \in I_1} \omega_1(\underline{x}^i)\overline{\text{nnd}}(\underline{x}^i). \tag{3.11}$$

Similarly, we propose comparison of the mean minimum spanning distance $\overline{\text{msd}}(\underline{x})$, defined in (3.4), between collections using the mean minimum spanning distance test statistic

$$\delta_{\text{msd}}(I) = \sum_{i \in I_0} \omega_0(\underline{x}^i)\overline{\text{msd}}(\underline{x}^i) - \sum_{i \in I_1} \omega_1(\underline{x}^i)\overline{\text{msd}}(\underline{x}^i).$$

The mean nearest neighbour distance and mean minimum spanning distance test statistics compare average separation distances, ensuring that the statistics summarise information only on the distances between points and not on the number of points in each pattern. While nearest neighbour and minimum spanning distances both summarise point separation distances, they are expected to differ in particular for cluster processes containing multiple clusters. For a cluster process the mean nearest neighbour distance summarises within-cluster point separation distances, while the mean minimum spanning distance tends to summarise both within-cluster point separation distances and between-cluster distances.

Summarising the information contained within a point pattern by a single number before comparison between collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ leads to a large loss of information. An alternative approach summarises the information in each observation by a function to allow the preservation of a greater amount of information, with comparison between functions revealing differences which may be unobservable when comparing single number summary statistics. The additional information can be carried forwards even in cases where the comparison of functions results in a single number.

Diggle et al. (2000) propose a test statistic for the comparison of the scaled neighbourhood count function, $K(r)$, an estimator of which is defined in (3.5). An estimate of the average scaled neighbourhood count function evaluated at distance r for the collection of point patterns $\underline{x}^J$ indexed by the set J, denoted by $\bar{K}(\underline{x}^J, r)$, is given by

$$\bar{K}(\underline{x}^J, r) = \frac{1}{\sum_{i \in J} n(\underline{x}^i)} \sum_{i \in J} n(\underline{x}^i)\hat{K}(\underline{x}^i, r). \tag{3.12}$$

The proposed test statistic, which will in future be referred to as the scaled neighbourhood count test statistic, is given by

$$\delta_K(I) = \int_0^{r_0} \frac{1}{r^2} \left(\sum_{i \in I_0} n(\underline{x}^i) \right) \left[\bar{K}(\underline{x}^{I_0}, r) - \bar{K}(\underline{x}^I, r) \right]^2 dr$$

$$+ \int_0^{r_0} \frac{1}{r^2} \left(\sum_{i \in I_1} n(\underline{x}^i) \right) \left[\bar{K}(\underline{x}^{I_1}, r) - \bar{K}(\underline{x}^I, r) \right]^2 dr. \qquad (3.13)$$

The integration in (3.13) is carried out across the range $r = 0$ to $r = r_0$, where r_0 is chosen to be large enough to summarise the information present in the point patterns, but small enough in comparison to the dimensions of W that the impact of edge corrections is not too great. The dominant term in the sampling variance of $\hat{K}(\underline{x}, r)$ is of order $1/n(\underline{x})$ for a homogeneous Poisson process (Diggle et al., 2000), motivating the weighted average in the expression of $\bar{K}(\underline{x}^J, r)$ in (3.12). Furthermore, for a homogeneous Poisson point process and values of r which are small in relation to the dimensions of the observation window W, the dominant term in the sampling variance of $\hat{K}(\underline{x}, r)$ with r is of order r^2 (Diggle et al., 2000). As a result a multiplier of $1/r^2$ is included in the expression of the scaled neighbourhood count test statistic given by (3.13).

Estimated nearest neighbour functions, defined in (3.6), may be averaged over the collection of point patterns $\underline{x}^J$ indexed by the set J, to produce $\hat{G}(\underline{x}^J, r)$ given by

$$\hat{G}(\underline{x}^J, r) = \sum_{i \in J} \omega_J(\underline{x}^i) \hat{G}(\underline{x}^i, r), \qquad (3.14)$$

where ω_J is defined for the set J as ω_0 is defined for the set I_0 in (3.10). We propose two test statistics for comparison of $\hat{G}(\underline{x}^{I_0}, r)$ and $\hat{G}(\underline{x}^{I_1}, r)$ across the range of distances r. These will be referred to as the L_1 nearest neighbour distribution test statistic, denoted by $\delta_{G,1}(I)$, and the L_∞ nearest neighbour distribution test statistic, denoted by $\delta_{G,\infty}(I)$, given by

$$\begin{aligned}
\delta_{G,1}(I) &= ||\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{I_1}, r)||_1 \\
&= \int_0^\infty |\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{I_1}, r)| dr. & (3.15) \\
\delta_{G,\infty}(I) &= ||\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{(I_1}, r)||_\infty \\
&= \sup_r |\hat{G}(\underline{x}^{I_0}, r) - \hat{G}(\underline{x}^{I_1}, r)|. & (3.16)
\end{aligned}$$

Both the L_1 and L_∞ nearest neighbour distribution test statistics propose comparison of the nearest neighbour function, $G(r)$, over an infinite range of r values. In practise, there is a finite value of $r = r_1$, equal to the largest observation window diameter, which all observed nearest neighbour distances are less than and, as a result, $\hat{G}(\underline{x}, r) = 1, \forall r \geq r_1$. The upper limit r_1 could then replace the upper limit of integration in the L_1 nearest neighbour distribution test statistic and limit the set over which the supremum is taken in the L_∞ nearest neighbour distribution test statistic to $r \in [0, r_1]$.

The L_∞ and L_1 norms are the two extreme L_p norms which respectively attribute all weight to the maximum difference between functions across the argument r and equal weight to the difference between functions across r. As a result, the L_1 distance may be preferred as it explicitly compares across the entire range of values of r. The form of the comparison in the scaled neighbourhood count function test statistic is more reminiscent of a weighted combination of L_2 norms.

We propose comparison of the effective force transference test statistic between collections using the effective force transference test statistic

$$\delta_{\mathrm{EFT}}(I) = \sum_{i \in I_0} \omega_0(\underline{y}^i)\mathrm{EFT}(\underline{y}^i) - \sum_{i \in I_1} \omega_1(\underline{y}^i)\mathrm{EFT}(\underline{y}^i).$$

3.2.3 Significance quantification

In the formulation of the statistical problem, we propose that observations $\underline{x}^{I_0}$ are independent realisations of the point process $\underline{X}^0$ and observations $\underline{x}^{I_1}$ are independent realisations of the point process $\underline{X}^1$. It is not possible that any single one-dimensional test statistic totally summarises all possible differences between $\underline{X}^0$ and $\underline{X}^1$. In the previous section we therefore proposed a number of test statistics, each of which compares particular properties of $\underline{X}^0$ and $\underline{X}^1$. Despite the variety of proposed test statistics our null hypothesis is unchanged for each test statistic, H_0: $\underline{X}^0$ and $\underline{X}^1$ are equal in distribution, with straightforward extension to $\underline{Y}^0$ and $\underline{Y}^1$.

Without making further assumptions on the properties of $\underline{X}^0$ and $\underline{X}^1$, the distribution of proposed test statistics under the null hypothesis cannot be analytically determined. Analysis of the significance of calculated test statistics is therefore carried out using permutation testing, introduced previously in Section 2.3.3. Under the null hypothesis the labelling of point patterns as members of the collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ is exchangeable because they are assumed to be independent realisations of the same point process, $\underline{X} \overset{\mathrm{d}}{=} \underline{X}^0 \overset{\mathrm{d}}{=} \underline{X}^1$, say.

Exchangeability of allocation to collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ is theoretically described by the application of permutations $\lambda \in S_{|I|} = \{\lambda_0, \lambda_1, \dots, \lambda_m\}$, the symmetric group of degree $|I|$, to the set of indices I. Application of the permutation λ_j to the set of indices I results in subsets $I_0^{(j)}$ and $I_1^{(j)}$ satisfying

$$\begin{aligned} I_0^{(j)} \cup I_1^{(j)} &= I, & I_0^{(j)} \cap I_0^{(j)} &= \emptyset, \\ |I_0^{(j)}| &= |I_0|, & |I_1^{(j)}| &= |I_1|, \end{aligned}$$

between which point patterns are compared. The observed test statistic under the identity permutation, $\delta(\lambda_0(I))$ with $\lambda_0 : \lambda_0(I) = I$, is then be compared to the set $\{\delta(\lambda_0(I)), \delta(\lambda_1(I)), \ldots, \delta(\lambda_m(I))\}$ in the case of the exact permutation test.

In practice, the total number of permutations is too large to feasibly calculate for both simulated data and the real data examples considered in the following sections. An approximate permutation test is therefore carried out using a subset of $S_{|I|}$ of size m' equal to 10 000 to ensure good power (Jöckel, 1986). In the case where $|I_0| = k$, $I_0^{(j)}$ is obtained by randomly sampling k elements of I without replacement such that $I_0^{(j)} \neq I_0^{(l)}$ $\forall l \in \{0, 1, \ldots, j-1\}$. The corresponding set $I_1^{(j)}$ is then simply $I \setminus I_0^{(j)}$ and the first permutation is chosen specifically to be the identity permutation such that $I_0^{(0)} = I_0$ and $I_1^{(0)} = I_1$.

Across all proposed test statistics $\delta(I) = 0$ indicates no difference between $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$. Scaled neighbourhood count and nearest neighbour distribution test statistics take values in $[0, \infty)$, as a result one-sided p-values are calculated at the upper tail. All other proposed test statistics take values in $(-\infty, \infty)$, as a result two-sided p-values are calculated.

3.3 Validation study

To confirm the suitability of the methods described in the previous section, we first apply them to simulated data which models the features that we wish to differentiate between.

3.3.1 Simulation description

Let HPPP(ρ, W) denote the homogeneous Poisson point process with intensity ρ on the window $W \subset \mathbb{R}^2$. Similarly, let IPPP$(\rho(x), W)$ denote the inhomogeneous Poisson point process with intensity $\rho(x)$ on the window $W \subset \mathbb{R}^2$.

Let CPP(d, W) denote the cluster point pattern containing $n+1$ points distributed over the observation window $W \subset \mathbb{R}^2$, according to $d = \{d_1, d_2, \ldots, d_n\}$, an ordered set of n fixed strictly positive separation distances arranged in ascending order. Point location x_{j+1} is determined by rejection sampling using as a proposal distribution Uniform$\{x \in \mathbb{R}^2 : \exists k \in \{1, 2, \ldots, j\} \text{ st } ||x - x_k|| = d_j\}$, with acceptance probability $\mathbb{1}\{x \in W, \inf_{k \in 1,2,\ldots,j} ||x - x_j|| \geq d_j\}$. The result of this procedure is

```
Data: $d$, $W$
Result: Point pattern $\underline{x}$ on $W$ with nearest neighbour distances $\{d_1\} \cup d$
$x_1 \leftarrow$ centre of $W$;
for $i$ in $1 : n$ do
    repeat
        $j \sim \text{Uniform}\{1, \dots, i\}$;
        $\theta \sim \text{Uniform}[0, 2\pi]$;
        $x_{i+1} \leftarrow x_j + d_i(\cos\theta, \sin\theta)$;
    until $\min_{k \in \{1,2,\dots,i\}} ||x_{i+1} - x_k|| \geq d_i$ and $x_{i+1} \in W$;
end
$\underline{x} \leftarrow (x_1, x_2, \dots, x_{n+1})$;
```

Algorithm 1: CPP(d, W) generation

point locations distributed according to the conditional distribution

$$x_{i+1}|x_1, x_2, \dots, x_i \sim \text{Uniform}\left(W \cap \left\{x \in \mathbb{R}^2 : \min_{j \in \{1,2,\dots,i\}} ||x - x_j|| = d_i\right\}\right).$$

An algorithmic description for the generation of a realisation of a CPP(d, W) process is given in Algorithm 1. From Algorithm 1, it is clear that $\text{nnd}(x_1) = d_1$ and $\text{nnd}(x_i) = d_{i-1}, i \in \{2, 3, \dots, n+1\}$. Similarly, the minimum spanning tree of $\underline{X} \sim \text{CPP}(d, W)$ has weight $\sum_{i \in \{1,2,\dots,n\}} d_i$ by construction.

Let MPP(n, u, ϕ) denote the marked point process on χ_2^+, defined in 3.1, with a fixed number of points n and mark directions v_j deviating from $u \in \mathbb{R}^3$ by angles of up to ϕ. As the effective force transference summary statistic depends only upon the distribution of marks v_j, point locations x_j are fixed at the origin. Marks are then simulated for $u^{'} = (0, 0, 1)$ with

$$\phi_j \sim \text{Uniform}[0, \phi], \quad \theta_j \sim \text{Uniform}[0, 2\pi], \quad v_j^{'} = (\sin\phi_j \cos\theta_j, \sin\phi_j \sin\theta_j, \cos\phi_j).$$

Marks are transformed by the rotation R for which $Ru^{'} = u$ and scaled by the constant η_j such that $v_j \cdot \hat{z} = h$, producing $v_j = \eta_j R v_j^{'}$.

3.3.2 Study design

For each simulation type, the required collection of point patterns $\underline{x}^I$ or $\underline{y}^I$ is simulated before permutation testing is carried out for each of the appropriate test statistics, with the resulting p-value being recorded. One hundred independent replicates of each simulation and testing procedure are carried out to provide information on the sensitivity and specificity of proposed tests and the variability of these properties. In the case of point patterns, $\underline{x}$, collection sizes are $|I_0| = |I_1| = 30$ and for marked point patterns, $\underline{y}$, the collection sizes are $|I_0| = |I_1| = 14$ to approximately match the observed data analysed in Section 3.4. A realisation of each simulation type is displayed in Figure 3.1.

Homogeneous intensity simulations are made up of point patterns $\underline{x}$ simulated according to HPPP(ρ_0, W) for $\underline{x} \in \underline{x}^{I_0}$ and HPPP$(\alpha\rho_0, W)$ for $\underline{x} \in \underline{x}^{I_1}$. The base intensity $\rho_0 = 10^{-4}$ is chosen to approximately match that of the observed data, with each W the square window with area sampled independently from Uniform$[500^2 - 10^5, 500^2 + 10^5]$ to produce approximately the same number of points per pattern as the observed data. Tested values of α vary across the set $\{1, 1.1, 1.2, 1.5\}$, resulting in various strengths of difference between collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$.

Inhomogeneous intensity simulations are made up of point patterns $\underline{x}$ simulated according to HPPP(ρ_0, W_0) for $\underline{x} \in \underline{x}^{I_0}$ and IPPP(ρ, W_0) for $\underline{x} \in \underline{x}^{I_1}$. The observation window W_0 is fixed as the square window with side length 500. The inhomogeneous intensity at the point with coordinate location $(x, y) \in \mathbb{R}^2$ is given by

$$\rho((x,y)) = \begin{cases} \frac{3\rho_0 x}{250} & x + y < 500, x < y \\ \frac{3\rho_0 y}{250} & x + y < 500, x > y \\ \frac{3\rho_0(500-x)}{250} & x + y > 500, x > y \\ \frac{3\rho_0(500-y)}{250} & x + y > 500, x < y. \end{cases}$$

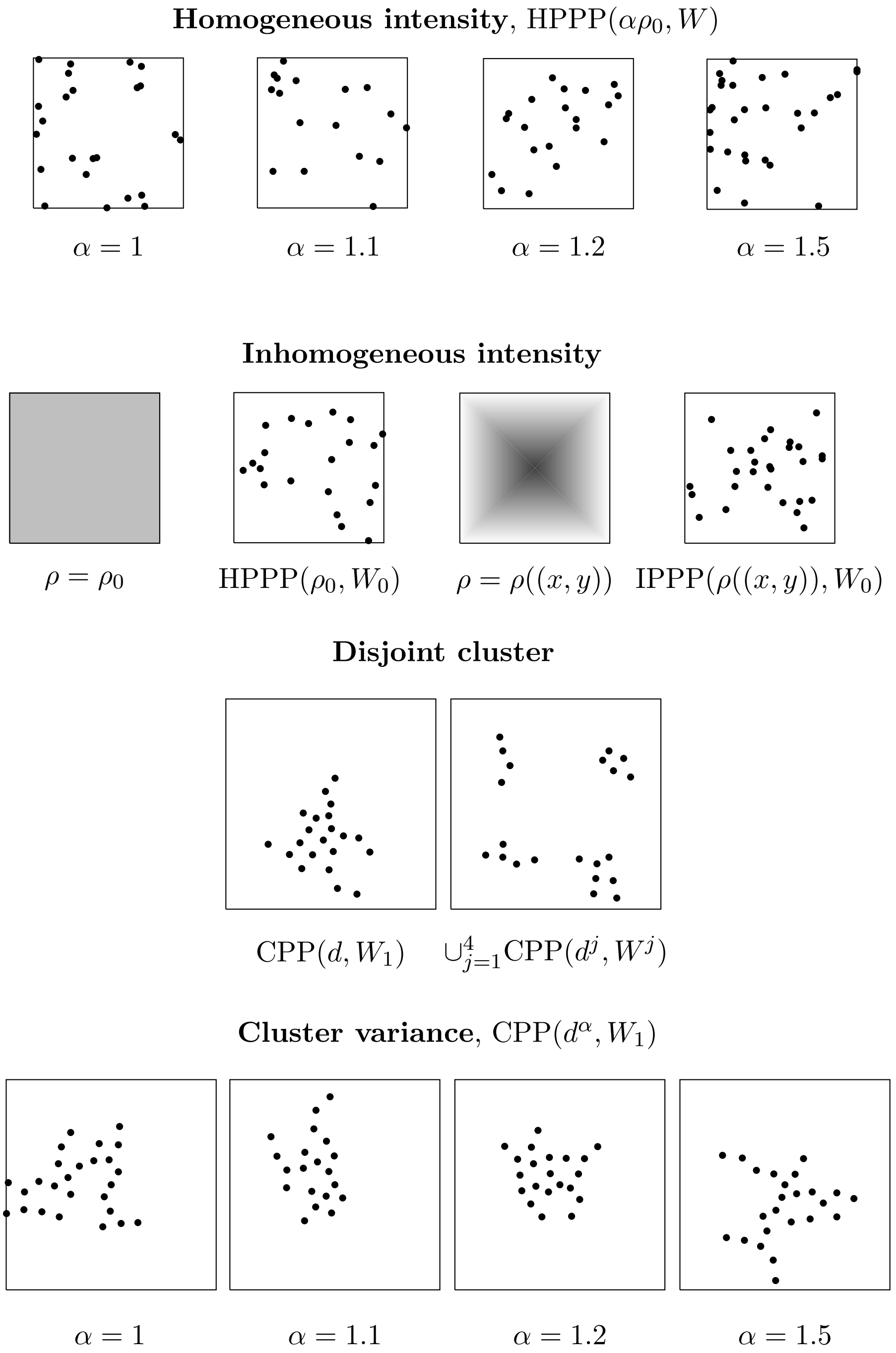
The shape of the inhomogeneous intensity is then a square based pyramid with maximum height $3\rho_0$, chosen such that

$$\int_{W_0} \rho(x)dx = \int_{W_0} \rho_0 dx.$$

Under this formulation the first order properties of $\underline{x}$ are identical for both collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$, while the second order properties differ.

Disjoint cluster simulations aim to compare point patterns made up of a single cluster with point patterns made up of four clusters, one in each quadrant of the square observation window. For $\underline{x} \in \underline{x}^{I_1}$ point patterns $\underline{x} = \cup_{j=1}^4 \underline{x}^j$ are the union of four simulated realisations $\underline{x}^j$ of CPP(d^j, W_1^j). The W_1^j divide W_1, the square window with area $|W_1| = 2|W_0| = 2 \times 500^2$, into four square quadrants each with area $|W_0|/2$. The total number of points $n(\underline{x}) \sim$ po$(\rho_0|W_0|)$, is truncated to be greater than eight such that each quadrant contains at least two points. Point separation distances are determined from $\tilde{d}$, simulated as an independent sample of size $n(\underline{x}) - 4$ from Normal$(55, 10^2)$, truncated to be non-negative and arranged in increasing order. Once sampled, elements of $\tilde{d}$ are partitioned into four subsets d^j with

$$\tilde{d}_j \in d^j, \quad j \in \{1, 2, 3, 4\},$$
$$\mathbb{P}(\tilde{d}_i \in d^j) = 1/4, \quad i \in \{5, 6, \ldots, n(\underline{x})\}, \quad j \in \{1, 2, 3, 4\}.$$



Homogeneous intensity, $\mathrm{HPPP}(\alpha\rho_0, W)$

$\alpha = 1$ $\alpha = 1.1$ $\alpha = 1.2$ $\alpha = 1.5$

Inhomogeneous intensity

$\rho = \rho_0$ $\mathrm{HPPP}(\rho_0, W_0)$ $\rho = \rho((x, y))$ $\mathrm{IPPP}(\rho((x, y)), W_0)$

Disjoint cluster

$\mathrm{CPP}(d, W_1)$ $\cup_{j=1}^{4}\mathrm{CPP}(d^j, W^j)$

Cluster variance, $\mathrm{CPP}(d^\alpha, W_1)$

$\alpha = 1$ $\alpha = 1.1$ $\alpha = 1.2$ $\alpha = 1.5$

Figure 3.1: Realisations of each of the simulation types.

Point patterns $\underline{x}$ are simulated according to $\text{CPP}(\tilde{d}, W_1)$ for $\underline{x} \in \underline{x}^{I_0}$ to produce point patterns with the same nearest neighbour distribution as for $\underline{x} \in \underline{x}^{I_1}$. The number of points $n(\underline{x})$ is simulated from $\text{po}(\rho_0|W_0|)$ and again truncated to produce values greater than eight. Separation distances d are again determined by $\tilde{d}$ a sample of size $n(\underline{x}) - 4$ from $\text{Normal}(55, 10^2)$, truncated at zero to produce only non-negative separation distances and arranged in increasing order, with

$$d = \{\tilde{d}_1, \tilde{d}_2, \tilde{d}_2, \tilde{d}_3, \tilde{d}_3, \tilde{d}_4, \tilde{d}_4, \tilde{d}_5, \tilde{d}_6, \ldots, \tilde{d}_{n(\underline{x})-4}\}.$$

All point patterns $\underline{x} \in \underline{x}^I$ have identical first order properties, due to common observation window sizes and distribution of number of points, and identically distributed nearest neighbour distances. However, $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ differ in that $\underline{x} \in \underline{x}^{I_0}$ are made up of a single cluster while $\underline{x} \in \underline{x}^{I_1}$ are made up of four clusters.

Cluster variance simulations are made up of point patterns $\underline{x}$ simulated by $\text{CPP}(d, W_1)$ for $\underline{x} \in \underline{x}^{I_0}$ and $\text{CPP}(d^\alpha, W_1)$ for $\underline{x} \in \underline{x}^{I_1}$. For all observations $n(\underline{x})$ is simulated from $\text{po}(\rho_0|W_0|)$. Separation distances d are an independent sample of size $n(\underline{x})$ from $\text{Normal}(55, 10^2)$, truncated to be non-negative, while d^α are an independent sample of size $n(\underline{x})$ from $\text{Normal}(55, 10^2\alpha)$, again truncated to be non-negative, for α varying across the set $\{1, 1.1, 1.2, 1.5\}$. Point patterns therefore have identical first order properties and identical mean nearest neighbour distances, but different nearest neighbour distance variances between collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$.

Varying angle simulations are made up of marked point patterns $\underline{y}$ simulated according to $\text{MPP}(n, u, \phi)$ for $\underline{y} \in \underline{y}^{I_0}$ and $\text{MPP}(n, u, \alpha\phi)$ for $\underline{y} \in \underline{y}^{I_1}$. The number of points, n are sampled independently from $\text{po}(30)$ to approximately match observed data. Direction u is fixed at $(0, 0, 1)$ as results are independent of its value. Maximum angle ϕ is fixed at 10 degrees, the approximate average for observed data, with α varying across the set $\{1, 1.05, 1.1, 1.2, 1.5\}$. There is therefore a greater variance in line trajectories for $\underline{y} \in \underline{y}^{I_1}$ than for $\underline{y} \in \underline{y}^{I_0}$ with increasing α.

3.3.3 Study results

Simulated point patterns are tested using each of the relevant test statistics using a total of 10 000 permutations and this procedure is repeated for 100 independent sets of simulations, resulting in 100 p-values for each test statistic for each parameter value. The distribution of p-values is compared to the Uniform distribution on $[0, 1]$ via the Kolmogorov-Smirnov test (Massey, 1951) as an omnibus test to determine if p-values differ from their expected distribution under the null hypothesis of $\underline{X}^0 \overset{\text{d}}{=} \underline{X}^1$. Omnibus tests are carried out at a corrected five percent significance level, using the Bonferroni correction to account for multiple testing of the same data using a number of different test statistics. The proportion of p-values in the interval $[0, 0.05]$ is also presented as the proportion of individual tests which would have rejected the null hypothesis at the five percent level. These values are presented for simulated point patterns $\underline{x}$ in Table 3.1 and simulated marked point patterns $\underline{y}$ in Table 3.2.

Test statistics based upon first order statistics, δ_N, δ_ρ and $\delta_{\rho,\omega}$, successfully report a difference for homogeneous intensity simulations where the difference in intensity is at least ten percent, but fail to report a difference in any other case. This behaviour is to be expected as the alternative classes of simulations produce differences in second order properties, while keeping first order properties constant. There is negligible difference between unweighted and weighted intensity test statistics, δ_ρ and $\delta_{\rho,\omega}$.

The pattern area test statistic, δ_W, consistently fails to reject the uniform distribution of p-values. This is the expected behaviour as for homogeneous intensity simulations the pattern area, $|W|$, is identically distributed for both collections of point patterns $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$. For all other simulations the pattern area is exactly identical for all simulated point patterns and it is unnecessary to test the pattern area test statistic.

Test statistics based upon single number second order summary statistics, δ_{nnd}, $\delta_{\mathrm{nnd},\omega}$, δ_{msd} and $\delta_{\mathrm{msd},\omega}$, successfully report a difference for both homogeneous and inhomogeneous intensity simulations. Mean minimum spanning distance test statistics, δ_{msd} and $\delta_{\mathrm{msd},\omega}$, are additionally able to report the difference between disjoint cluster simulations. This is expected by construction of the mean minimum spanning distance summary statistic which summarises both within- and between-cluster distances, in comparison to the mean nearest neighbour distance summary statistic which summarises only within-cluster distances. As these test statistics compare the means of separation distances they are unable to report a difference when only the variance of separation distances changes, as in the case of the cluster variance simulations. The difference between weighted and unweighted test statistics is small, but in general weighted test statistics are more accurate at detecting differences. Test statistics based upon the mean minimum spanning distance generally outperform those based upon the mean nearest neighbour distance.

Test statistics based upon functional summary statistics of nearest neighbour distances, $\delta_{G,1}$, $\delta_{G,1,\omega}$, $\delta_{G,\infty}$ and $\delta_{G,\infty,\omega}$, successfully report a difference in every case except for disjoint cluster simulations. This behaviour is expected as by design the nearest neighbour distribution depends upon within-cluster separation distances that are unchanged for disjoint cluster simulations, while the between-cluster separation distance does change. The difference between weighted and unweighted test statistics is again small, with weighted test statistics performing slightly better for homogeneous and inhomogeneous intensity simulations and no clear difference for cluster variance simulations. Test statistics comparing nearest neighbour functions

		δ_N	δ_W	δ_ρ	$\delta_{\rho,\omega}$	δ_{nnd}	$\delta_{\text{nnd},\omega}$	δ_{msd}	$\delta_{\text{msd},\omega}$	δ_K	$\delta_{G,1}$	$\delta_{G,1,\omega}$	$\delta_{G,\infty}$	$\delta_{G,\infty,\omega}$
Homogeneous intensity, α	1.0	0.07	0.07	0.02	0.04	0.05	0.04	0.04	0.04	0.08	0.07	0.06	0.07	0.05
	1.1	0.23†	0.11	0.46†	0.44†	0.17†	0.21†	0.30†	0.30†	0.01	0.17†	0.18†	0.14†	0.14†
	1.2	0.65†	0.05	0.96†	0.96†	0.63†	0.71†	0.80†	0.87†	0.06	0.67†	0.72†	0.51†	0.54†
	1.5	1.00†	0.04	1.00†	1.00†	1.00†	1.00†	1.00†	1.00†	0.04	1.00†	1.00†	0.99†	1.00†
Inhomogeneous intensity		0.07	-	0.07	0.08	0.98†	0.99†	1.00†	1.00†	1.00†	0.97†	0.99†	0.95†	0.97†
Disjoint cluster		0.06	-	0.06	0.08	0.06	0.05	1.00†	1.00†	1.00†	0.09	0.09	0.09	0.09
Cluster variance, α	1.0	0.04	-	0.04	0.05	0.05	0.04	0.05	0.04	0.04	0.05	0.03	0.05	0.05
	1.1	0.03	-	0.03	0.03	0.05	0.05	0.04	0.05	0.06	0.32†	0.33†	0.18†	0.17†
	1.2	0.07	-	0.07	0.07	0.06	0.07	0.07	0.08	0.07	0.93†	0.92†	0.43†	0.50†
	1.5	0.05	-	0.05	0.06	0.05	0.09	0.02	0.02	0.08†	1.00†	1.00†	1.00†	1.00†

Table 3.1: Validation study results from testing for differences between simulated point patterns for a number of different simulation types, simulation parameters and test statistics. Proportion of p-values in the range $[0, 0.05]$, $\dagger$ indicates non-uniformity of p-values under the Kolmogorov-Smirnov test at the Bonferroni corrected (Dunn, 1961) $5/13 = 0.38$ (Homogeneous intensity) or $5/12 = 0.42$ (Inhomogeneous intensity, Disjoint cluster, Cluster variance) percent significance level.

		δ_{EFT}	$\delta_{\text{EFT},\omega}$
Varying angle, α	1.00	0.02	0.01
	1.05	$0.32^{\dagger}$	$0.32^{\dagger}$
	1.10	$0.84^{\dagger}$	$0.83^{\dagger}$
	1.20	$1.00^{\dagger}$	$1.00^{\dagger}$
	1.50	$1.00^{\dagger}$	$1.00^{\dagger}$

Table 3.2: Validation study results from testing for differences between marked point patterns for a number of different simulation parameters and test statistics. Proportion of p-values in the range $[0, 0.05]$, † indicates non-uniformity of p-values under the Kolmogorov-Smirnov test at the Bonferroni corrected (Dunn, 1961) $5/2 = 2.5$ percent significance level.

using the L_1 distance outperform those which compare using the L_∞ distance, likely because they directly summarise the difference over the entire range of distances r.

The scaled neighbourhood count test statistic, δ_k, accurately reports a difference in the case of inhomogeneous intensity and disjoint cluster simulations, but is largely ineffective for alternative simulations. As the scaled neighbourhood count summary statistic is normalised by the estimated intensity, $\hat{\rho}$, it is not expected to report a difference for homogeneous intensity simulations. Accurate detection in the case of disjoint cluster simulations may result from the fact that the scaled neighbourhood count function summarises more point separation distances than simply nearest neighbour distances.

The effective force transference test statistics, δ_{EFT} and $\delta_{\text{EFT},\omega}$, accurately report a difference in mark orientation in close to one third of simulations when the difference in maximum deviation is increased by at least five percent, consistently so when the increase is at least 20 percent.

As a result of the simulation study we have confirmed the scenarios in which each of the test statistics are effective, supporting the use of specific test statistics when it is desired to test for particular differences in structure. We have also confirmed the suitability of the permutation testing approach, with an absence of evidence to reject uniformity of p-values when there is no difference between simulated point pattern collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$.

3.4 Investigation of changes in K-fiber microtubule organisation following TACC3 overexpression

3.4.1 Biological background

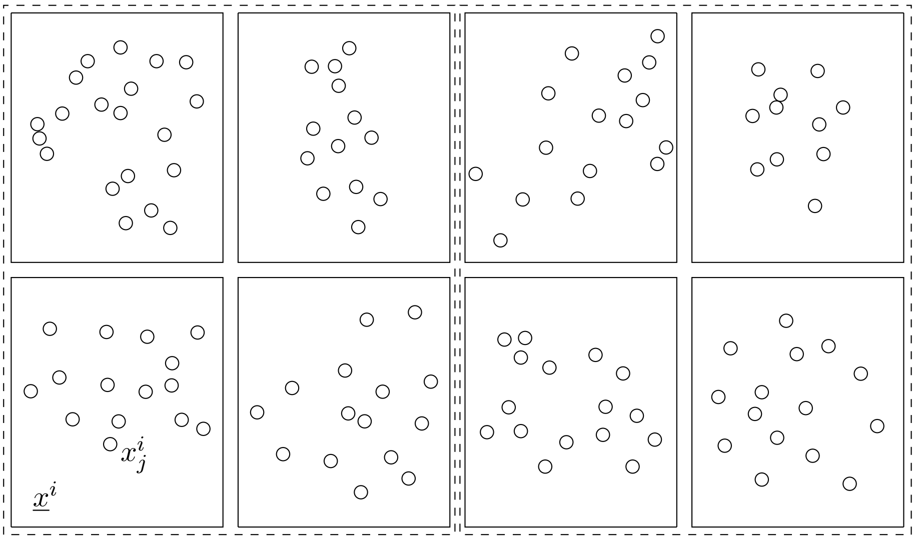
During mitosis, subcellular structures known as kinetochore fibers (K-fibers) connect two anchor points, centrosomes, within the cell to each of the chromosome pairs (Booth et al., 2011). The chromosomes contain the genetic information and successful mitosis requires the pairs to be evenly divided, such that each of the two daughter cells contains one of each chromosome (Holland and Cleveland, 2009). K-fibers are believed to apply the force necessary to separate chromosome pairs and as a result they are important for successful cell division. An illustration of the intermediate stages in the process of mitosis including the function of the K-fibers can be seen in Figure 1.1.

Each K-fiber is made up of a number of microtubules, approximately cylindrical structures which are bound together by a mesh structure to form the rigid K-fiber. Our collaborators, Dr. Stephen Royle and his research group within the Centre for Mechanochemical Cell Biology at the University of Warwick, are interested in the effect that overexpression of TACC3, Transforming acidic coiled-coil containing protein 3, may have on the structure of microtubules within K-fibers. The structure may be visualised by microscopy imaging of cells at the correct point in the cell cycle under a control regime and a treatment regime for which there is overexpression of TACC3.

Example diagrams of point locations arising from microtubule locations may be seen in Figure 3.2. Note that these are not the microscope images from which point locations are recorded, rather they illustrate the interpretation of the results with circles centred at point locations with radius equivalent to the expected radius of microtubules. Point patterns are presented in boxes to separate them, but these boxes are not observation windows, due to the method in which point locations are reported (described in more detail in the following section). Dashed boxes divide point patterns into $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$, again for illustrative purposes as the total number of point patterns is too large to present them all.

We choose to investigate 2D coordinate locations obtained from imaging biological samples as point patterns. This results in $|I| = 63$, I_0 indexing observations under the control regime with $|I_0| = 26$ and I_1 indexing observations under the treatment regime with $|I_1| = 37$.

Also available is a data set comprising paired 2D microtubule coordinate locations obtained from two parallel image slices through the same sample approximately perpendicular to microtubule directions, including information on which coordinate



$$\underline{x}^{I_0} = \{\underline{x}^i : i \in I_0\} \qquad \underline{x}^{I_1} = \{\underline{x}^i : i \in I_1\}$$

$$\underline{x}^i \sim \underline{X}^0 \qquad \underline{x}^i \sim \underline{X}^1$$

Figure 3.2: Diagrams of sample microtubule location data. Each pattern represents a single K-fiber with microtubules represented as circles, the centres of which are analysed as point locations and the radii of which are consistent with the expected microtubule radius.

locations represent ends of the same microtubule in each of the slices. Paired coordinate locations are reported as

$$\underline{x} = \{(x_{1,0}, x_{1,1}), (x_{2,0}, x_{2,1}), \ldots, (x_{n(\underline{x}),0}, x_{n(\underline{x}),1})\}.$$

We choose to investigate paired 2D coordinate locations obtained by imaging biological samples as marked point patterns $\underline{y}$ with $x_j = x_{j,0} \times \{0\}$ and $v_j = (x_{j,1} - x_{j,0}) \times \{h\}$.

An example diagram of a single marked point pattern may be seen in Figure 3.3. This is again not a microscope image but rather a schematic diagram using the information reported from analysis of the images themselves. Arrows are used to represent each mark direction, v_j, oriented such that point locations, x_j, are at the tail of each arrow. Three diagrams in combination present the marked point pattern as projections into perpendicular frames, perpendicular to and parallel to the K-fiber axis.

Paired coordinate observations are similarly collected for multiple samples under each of the two experimental regimes. This results in $|I| = 28$, I_0 indexing observations under the control regime with $|I_0| = 13$ and I_1 indexing observations under the treatment regime with $|I_1| = 15$.

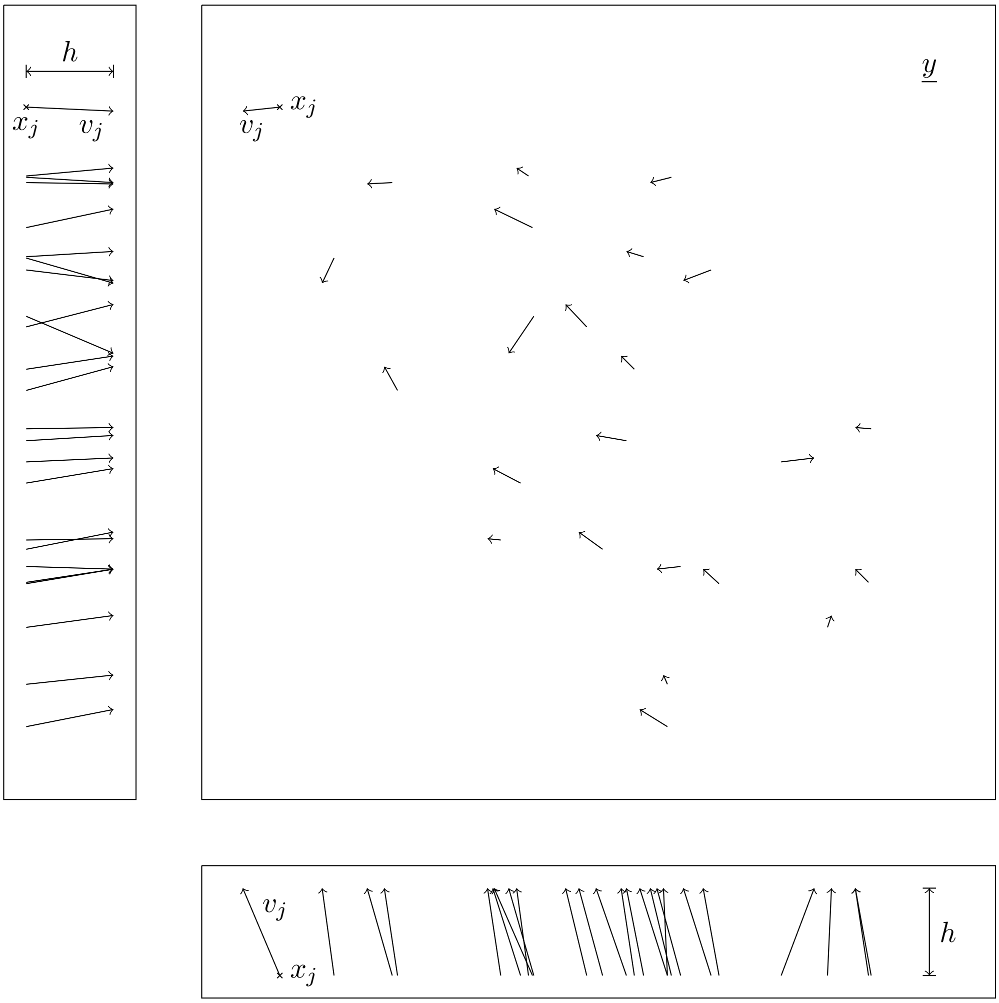


Figure 3.3: Plots of sample marked point pattern data. Point locations and mark orientations representing the location and direction of microtubules when viewed parallel to the K-fiber axis (left and bottom) and perpendicular to the K-fiber axis (top right).

3.4.2 Observed data features

Using the notation introduced in Section 2.1.3.1, point patterns $\underline{x}$ are realisations of point processes, $\underline{X}$, obtained by reporting the locations of all points contained within a fixed observation window W. Point patterns produced from observed data are instead made up of points added to the pattern using an iterative process. The experimenter first selects a single microtubule cross-section within the microscope image, taken to be located within the K-fiber of interest, with the coordinates of its centre recorded as the first point location, x_1. The coordinates corresponding to any microtubule cross-section whose centre lies within a distance of 105nm of x_1 are then added to the point pattern as point locations $x_2, x_3, \ldots$. This procedure is iterated, searching in a neighbourhood of size 105nm from each point in the pattern until there exists no further microtubule cross-sections whose centres satisfy this requirement. The specific distance of 105nm has been suggested by previous biological investigations to be the maximum centre-to-centre distance at which K-fiber microtubules may be connected by mesh, as microtubules have an average radius of 12.5nm and the mesh forms edge-to-edge connections at distances up to 80nm (Nixon et al., 2015). All locations and distances are on the scale of nanometres and as a result we omit explicit reference to units in the rest of this chapter, describing instead points within a distance of 105.

The purpose of the iterative procedure is to isolate and report the locations of only those microtubules which belong to a single K-fiber. An estimate to the observation window may be recreated using the observed data, resulting in $W(\underline{x})$ given by

$$W(\underline{x}) = \cup_{x_j \in \underline{x}} b(x_j, 105), \tag{3.17}$$

where $b(x_j, 105)$ is the two dimensional ball centred at x_j with radius 105. Under this formulation $W(\underline{x})$ is a realisation of $W(\underline{X})$, a random observation window. As a result, estimators which are unbiased in the deterministic case of $W(\underline{X}) = W$ are not expected to be unbiased when the estimator $W(\underline{x})$ is used instead.

Paired coordinate locations for observed data represent the centres of endpoints of microtubules which may be modelled as straight lines at the scale of our observations. Point locations x_j specify microtubule locations, while marks v_j specify microtubule directions. Combining these pieces of information, a parametric specification of the j^{th} microtubule is given by

$$L_j(t) = x_j + tv_j \qquad\qquad t \in [0,1]. \tag{3.18}$$

The purpose of the K-fiber is the separation of chromosome pairs during mitosis, with each K-fiber forming a physical link between a chromosome and the mitotic spindle pole, transferring the force required to physically pull apart the chromosome pair. Under the simplifying assumption that force is applied through the K-fiber via the application of forces of identical magnitude, F, in the direction of each of the microtubules, v_j, the resultant force is given by

$$\sum_{j=1}^{n(\underline{y})} F\hat{v}_j = F\beta.$$

The magnitude of the resultant force is $F||\beta||$ and, if we further assume the K-fiber to be optimally directed such that all of this resultant force is used to separate the chromosome pair, the proportion of input force effectively transferred is

$$\frac{F||\beta||}{F \times n(\underline{y})} = \frac{||\beta||}{n(\underline{y})} = \mathrm{EFT}(\underline{y}).$$

Therefore, the effective force transference summary statistic intuitively relates to the performance of the K-fiber, with larger values of effective force transference indicating that the K-fiber is more effective at transferring force as its component microtubules are better aligned.

Observed paired point pattern data arises from parallel imaging planes which are oriented approximately perpendicular to the axis of the K-fiber. As the K-fiber axis is not uniquely defined and its features may be obscured before imaging takes place, there is the possibility of variation in the orientation of the imaging planes between samples. Imaging a single sample at different orientations results in changes in the length of marks, $||v_j||$, but does not affect the trajectory of marks relative to each other. The effective force transference test statistic has been proposed with this particular property in mind and, as it is calculated using only relative mark trajectories, it is independent of the orientation of the parallel imaging planes (provided imaging planes are not themselves parallel to any of the microtubule directions).

3.4.3 Assumption checking

Coordinate locations reported for observed data represent the centres of microtubule cross-sections. Under the biological assumption that microtubules have radii of 12.5 and are connected by a mesh structure which extends a distance of no further than 80 (Nixon et al., 2015), we expect

$$\mathrm{nnd}(x_j) \in [25, 105], \tag{3.19}$$

for all points x_j in all point patterns $\underline{x}$. In practice 1811 of the 1824 total points satisfy (3.19), with two values of $\mathrm{nnd}(x_j) = 23.5$ and 11 values of $\mathrm{nnd}(x_j) > 105$, the largest of which is 127.3.

Extending the idea of a nearest neighbour distance to marked point patterns through the expression of lines $L_j(t)$ in Equation 3.18, we have

$$\text{nnd}_{\min}(L_j) = \min_{k\in\{1,2,\ldots,j-1,j+1,\ldots,n(\underline{y})\}} \min_{t_1,t_2\in[0,1]} ||L_j(t_1) - L_k(t_2)||,$$

and expect that $\text{nnd}_{\min}(L_j) \in [25, 105]$. In practice, for all but two patterns values of $\text{nnd}_{\min}(L_j^i)$ lie within the slightly extended range $[20, 110]$.

Marked point pattern $\underline{y}^{13} \in \underline{y}^{I_0}$ has a large number of points with $\text{nnd}_{\min}(L_j^{13}) > 110$ and pattern $\underline{y}^{17} \in \underline{y}^{I_1}$ has a large number of points with $\text{nnd}_{\min}(L_j^{17}) < 20$. Individual investigation of these patterns indicates that the discrepancy in $\underline{y}^{13}$ may be explained by incorrect scaling from pixel coordinates to physical coordinates and the discrepancy in $\underline{y}^{17}$ may be caused by imaging planes significantly far from perpendicular to the K-fiber axis. Despite these deviations from the assumptions, we continue with analysis of the entire data set as the observations are only slightly outside of the expected range and because the amount of data is already limited due to the time and expertise required to collect the images. A sensitivity analysis of the data after the removal of these two marked point patterns (not shown) indicates that although there are marginal changes in p-values, the significance of results remains unchanged.

Each of the proposed summary statistics is calculated under the assumption that point patterns are stationary. Visualisation of the estimated intensity of point patterns, obtained using the approach described by Diggle (1985), supports this assumption. Point patterns are further assumed to be generated by a process other than the homogeneous Poisson point process, as a result of which we propose alternative methods. Consider x_j a randomly selected point from a homogeneous Poisson point process with intensity ρ. We then have that

$$\mathbb{P}(25 \le \text{nnd}(x_j) \le 105) = \exp(-\rho\pi 25^2) - \exp(-\rho\pi 105^2),$$

which is maximised by the value of $\rho = \rho^*$ given by

$$\rho^* = \frac{2\log 105 - 2\log 25}{\pi(105^2 - 25^2)} = 8.8 \times 10^{-5}, \tag{3.20}$$

at which its value is 0.79. As over 99 percent of observed points x_j do satisfy $\text{nnd}(x_j) \in [25, 105]$, there is strong evidence to support the assumption that point patterns are not realisations of homogeneous Poisson point processes.

3.4.4 Exploratory data analysis

Under the reconstruction of observation windows given by (3.17), the area of the observation window, $|W(\underline{x})|$, may be taken as a surrogate for the cross sectional area

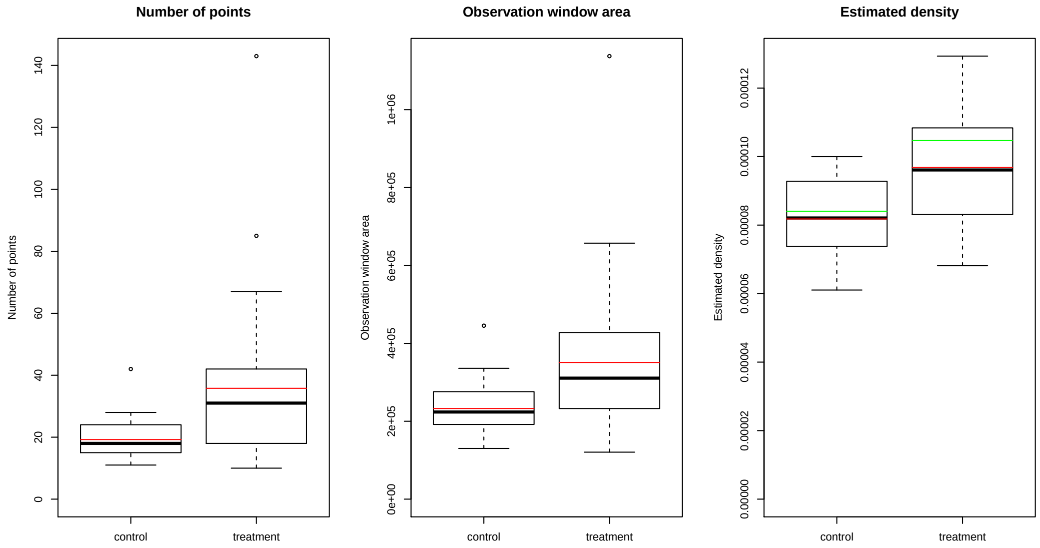


Figure 3.4: Boxplots of first order summary statistics, $n(\underline{x})$, $|W(\underline{x})|$ and $\hat{\rho}$, for observed TACC3 point patterns divided into controls, indexed by I_0, and treatments, indexed by I_1. Horizontal red lines indicate means, horizontal green lines indicate weighted means.

of the K-fiber. There is also a clear dependence between the observation window area $|W(\underline{x})|$ and the number of points $n(\underline{x})$. This dependence impacts the accuracy of the estimated density $\hat{\rho} = n(\underline{x})/|W(\underline{x})|$.

Figure 3.4 displays boxplots of the number of points, observation window area and estimated density for the observed point patterns separated into collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ according to whether they were obtained under the control or treatment experimental regime. Means and medians are greater for treatment observations for each of the first order test statistics, indicating that K-fibers are made up of a greater number of microtubules which are more closely separated within thicker K-fibers when the TACC3 protein is overexpressed. Weighted mean densities over the collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ are greater than unweighted means, indicating that K-fibers with greater numbers of microtubules are more tightly packed.

Figure 3.5 displays boxplots of the mean nearest neighbour distance and mean minimum spanning distance for the observed point patterns divided into two collections for control, $\underline{x}^{I_0}$, and treatment, $\underline{x}^{I_1}$, observations. In each case the average separation distance is reduced for treatment observations, indicating that TACC3 may limit the distance at which microtubules can be bound together into the K-fiber structure. Each weighted mean is also less than the corresponding unweighted mean, providing further evidence that K-fibers with greater numbers of microtubules are

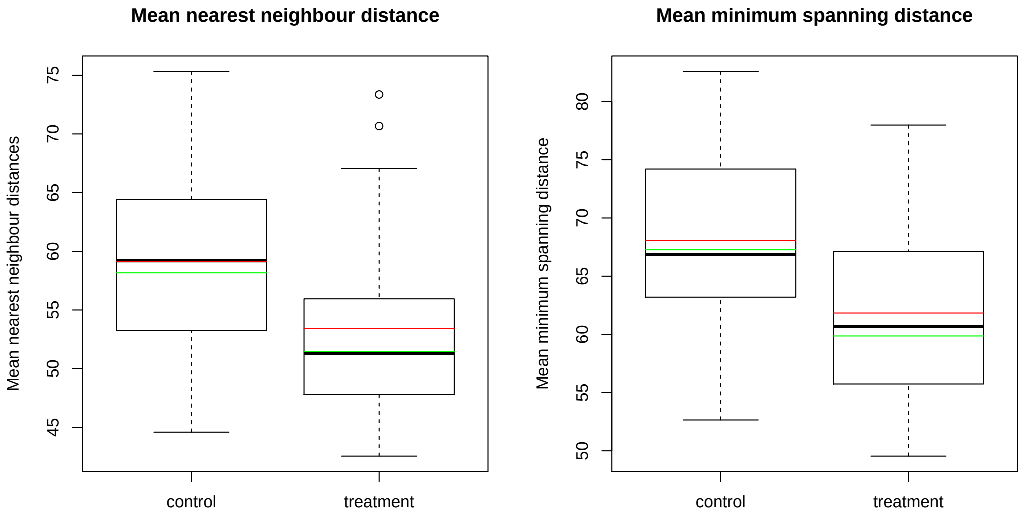


Figure 3.5: Boxplots of second order summary statistics, $\overline{\mathrm{nnd}}(\underline{x})$ and $\overline{\mathrm{msd}}(\underline{x})$, for observed TACC3 point patterns divided into controls, indexed by I_0, and treatments, indexed by I_1. Horizontal red lines indicate means, horizontal green lines indicate weighted means.

more tightly packed.

The assumption that $\mathrm{nnd}(\underline{x}) \leq 105$, verified in Section 3.4.3, in conjunction with the specification of observation windows $W(\underline{x})$ in (3.17), ensures that $\mathrm{nn}(x_j) \in \underline{x}$ for all points $j \in \{1, 2, \ldots, n(\underline{x})\}$ in each of the patterns. The nearest neighbour function, G, may therefore be calculated without the need for edge correction.

Plots of the estimated nearest neighbour function, $\hat{G}(\underline{x}, r)$, for TACC3 point patterns typically lie below the theoretical form for the Poisson point process with intensity ρ^* from (3.20), $G(r) = 1 - \exp(-\rho^* \pi r^2)$, for small values of r less than approximately 50. There is therefore evidence of regularity on short length scales, although it is unclear the extent to which this is caused by the restriction of $\mathrm{nnd}(x_j) > 25$ and the cumulative nature of the nearest neighbour function obscures the exact scale on which there is regularity. At values of r greater than approximately 50, plots of the nearest neighbour function exceed their expected theoretical form. There is therefore evidence of clustering at large length scales, although it is once again unclear the extent to which this is caused by the restriction of $\mathrm{nnd}(x_j) < 105$.

Evidence of clustering and regularity can be observed in Figure 3.6 which includes plots of $\hat{G}(\underline{x}^{I_0}, r)$ for control observations and $\hat{G}(\underline{x}^{I_1}, r)$ for treatment observations. Across the entire range of distances r we observe $\hat{G}(x^{I_0}, r) \leq \hat{G}(x^{I_1}, r)$, indicating that $\mathbb{P}(\mathrm{nnd}(x_j) < r)$ is greater for treatment observations than control observations,

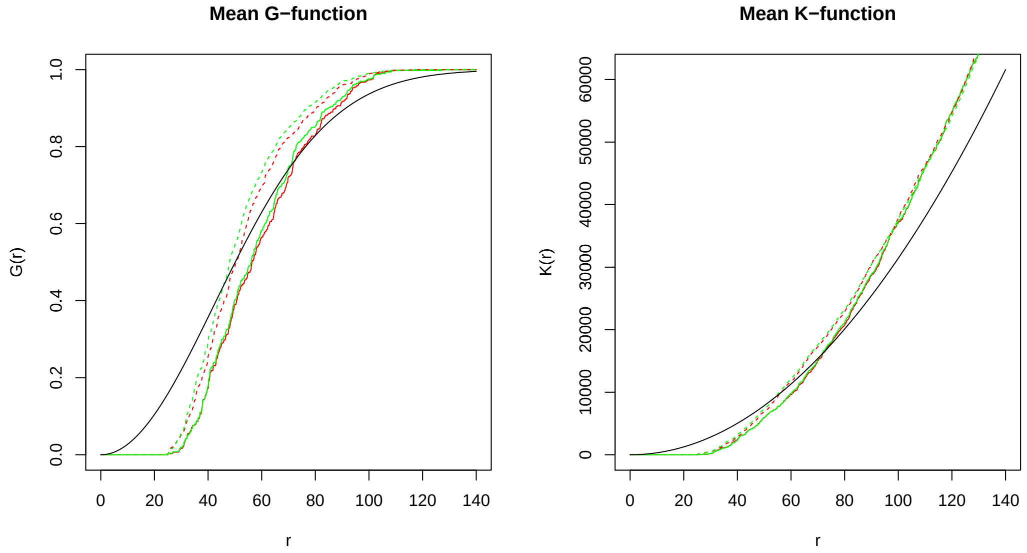


Figure 3.6: Plots of second order summary statistic functions, average nearest neighbour functions, G (left), and scaled neighbourhood count functions, K (right), for observed TACC3 point patterns. Red lines indicate means, green lines indicate weighted means. Solid lines indicate the average over controls, indexed by I_0, dashed lines indicate average over treatments, indexed by I_1. Black lines are theoretical values for Poisson point processes with intensity ρ^*.

further supporting the previous assertion that microtubules are more tightly packed for observations collected under the treatment regime. It can also be observed that weighted mean estimates of the nearest neighbour function lie below unweighted mean estimates across the entire range of distances r, again suggesting that K-fibers with greater numbers of microtubules are packed more tightly.

Figure 3.6 also includes plots of estimates $\bar{K}(\underline{x}^{I_0}, r)$ for control observations, $\bar{K}(\underline{x}^{I_1}, r)$ for treatment observations and the theoretical value πr^2. Over short distances r, $\bar{K}(\underline{x}^{I_0}, r) < \bar{K}(\underline{x}^{I_1}, r)$ indicating that microtubules are more tightly packed when TACC3 is overexpressed even when normalising for differences in point density. The difference between the weighted and unweighted means of estimated scaled neighbourhood count functions is negligible.

Boxplots of the effective force transference for control observations and treatments observations are presented in Figure 3.7. Across all observations the effective force transference is above 0.9, suggesting that microtubules are generally consistently aligned. The effective force transference is on average greater for control observations, indicating greater agreement in microtubule directions in comparison to treatment observations. Application of the testing methodology is required to

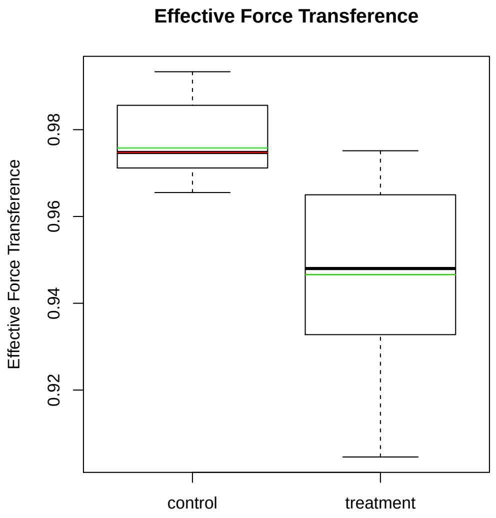


Figure 3.7: Boxplots of marked point pattern summary statistics EFT($\underline{y}$) for observed TACC3 marked point patterns divided by controls, indexed by I_0, and treatments, indexed by I_1. Horizontal red lines indicate means, horizontal green lines indicate weighted means.

δ_N	0.0005	δ_{nnd}	0.0057	δ_K	0.1092	δ_{EFT}	0.0011
δ_W	0.0018	$\delta_{\text{nnd},\omega}$	0.0005	$\delta_{G,1}$	0.0061	$\delta_{\text{EFT},\omega}$	0.0005
δ_ρ	0.0001	δ_{msd}	0.0019	$\delta_{G,1,\omega}$	0.0005		
$\delta_{\rho,\omega}$	0.0002	$\delta_{\text{msd},\omega}$	0.0005	$\delta_{G,\infty}$	0.0087		
				$\delta_{G,\infty,\omega}$	0.0013		

Table 3.3: Results (p-values) of testing for differences between treatment and control TACC3 (marked) point pattern data for a number of different test statistics.

assess the significance of the observed difference.

3.4.5 Permutation testing results

Table 3.3 summarises the p-values for approximate permutation testing of observed data using 10 000 permutations for each of the 15 proposed test statistics. In every case other than the scaled neighbourhood count test statistic, δ_K, the difference between $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ or $\underline{y}^{I_0}$ and $\underline{y}^{I_1}$ is reported as significant at the five percent significance level, with some reported as significant at the 0.5 and even 0.05 percent significance levels.

Point pattern size, area and density are significantly greater for $\underline{x}^{I_1}$ than for $\underline{x}^{I_0}$, with little difference between the results for weighted and unweighted intensity statistics. Biologically, this may be taken as evidence that overexpression of TACC3 causes K-fibers to contain significantly greater numbers of microtubules, have significantly larger cross-sectional area and significantly higher density of microtubules than K-fibers under control conditions. The difference between weighted and unweighted density test statistics is negligible.

Mean nearest neighbour distance and mean minimum spanning distance are significantly greater for $\underline{x}^{I_0}$ than for $\underline{x}^{I_1}$. Biologically, this may be taken as evidence that overexpression of TACC3 results in closer spacing of microtubules within K-fibers than under control conditions. Weighted mean nearest neighbour distance and mean minimum spanning distance test statistics result in noticeably smaller p-values than their unweighted alternatives.

Tighter packaging of microtubules within K-fibers following overexpression of TACC3 is further supported by all of the nearest neighbour function test statistics. There is some evidence that the average difference between $\hat{G}(\underline{x}^{I_0}, r)$ and $\hat{G}(\underline{x}^{I_1}, r)$, measured by the L_1 nearest neighbour distribution test statistic, is more significant than the maximum difference, measured by the L_∞ nearest neighbour distribution test statistic. Although both weighted and unweighted test statistics are significant at the five percent level, weighted test statistics are also significant at the 0.1 percent level while the unweighted test statistics are not.

The scaled neighbourhood count test statistic, δ_K, is the only test statistic which fails to reject the null hypothesis at the five percent level. This result may be explained by the formulation of the scaled neighbourhood count function or the accuracy of its estimation. Failure to reject the null hypothesis for δ_K, combined with rejection of the null hypothesis for density test statistics, δ_ρ and $\delta_{\rho,\omega}$, could be evidence that differences in the generating processes $\underline{X}^0$ and $\underline{X}^1$ are limited to the density of points. A difference in point density would also be expected to result in differences in point separation distances, which have been detected for the observed data. Alternatively, difficulties accurately estimating the scaled neighbourhood count function are caused by the requirement of edge correction and uncertain estimation of the density due to the non-standard specification of the observation windows. These difficulties limit the reliability of results obtained using the scaled neighbourhood count function in comparison to the presented alternatives.

Both the weighted and unweighted effective force transference test statistics result in rejection of the null hypothesis at the upper five percent level, and would also report significance at the stricter 0.5 percent level, indicating that the effective force transference is significantly reduced in treatment samples in comparison to control samples. Biologically, this may be taken as evidence that increased expression of TACC3 causes microtubules in K-fibers to be less well aligned, potentially reducing

Statistic	Effect size	d	Classification	s_d
δ_N	16.6 microtubules	0.84	Large	0.27
δ_W	1.19×10^5 μm^2	0.77	Medium	0.27
δ_ρ	1.50×10^{-5} microtubules/μm^2	1.03	Large	0.28
δ_{nnd}	-5.71μm	0.72	Medium	0.27
δ_{msd}	-6.25μm	0.81	Large	0.27
δ_{EFT}	-0.03	1.50	Large	0.47

Table 3.4: Reported treatment effect sizes for TACC3 data.

their capacity to accurately divide chromosome pairs.

Taken in combination, the results for point patterns $\underline{x}$ indicate that overexpression of TACC3 results in larger K-fibers which are made up of more tightly packed microtubules. The results for marked point patterns $\underline{y}$ indicate further that overexpression of TACC3 results in K-fibers comprised of less well aligned microtubules. The protein TACC3 is believed to influence changes in the microtubule structure indirectly, by altering properties of the mesh structure binding microtubule into K-fibers. On the evidence of our analysis an interpretation may be that increased expression of TACC3 alters the mesh structure by limiting the distance at which it may bind microtubules. Limiting mesh connection distances may require microtubules to be closer and pull microtubules out of a natural alignment, resulting in reduced force transference properties which are then compensated for by increased numbers of microtubules per K-fiber. All of these effects have been observed in the data set studied. Further biological experimentation would clearly be needed to confirm or contradict this hypothesis.

Table 3.4 summarises the treatment effect sizes for those test statistics based upon differences in means. The effect sizes support the results of the statistical testing procedure, indicating that the treatment increases the number of microtubules per K-fiber and K-fiber area. The treatment increases the packing density of microtubules within K-fibers and correspondingly reduces the distance between nearest neighbours and within the minimum spanning tree. Finally, the treatment reduces the effective force transference corresponding to a less well-ordered structure. All of the effect sizes are non-negligible in comparison to reference category values (Cohen, 1992) and their standard errors. More work would be necessary to determine effect size for the proposed functional comparisons.

3.4.6 Sensitivity analysis

3.4.6.1 Sensitivity analysis design

The are multiple possibilities for errors when imaging biological samples and reporting the coordinate locations of microtubule cross-section centres. Errors can be minimised, but are difficult to completely remove. As a result, the proposed methodology should be robust to perturbations which are small on a scale determined by the application. To ensure that this is the case we carry out a sensitivity analysis. Each of the perturbations are simulated and tested using each of the appropriate test statistics using 10 000 random permutations. Repeating this process 100 times produces 100 p-values for each test statistic.

Relabelling perturbations consider scenarios in which observations are incorrectly assigned to collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$, potentially due to human error. Perturbed index sets are denoted by I_0^q and I_1^q, for $q \in [0, 0.5]$ the probability of incorrect assignment. That is

$$\mathbb{P}(i \in I_0^q) = \begin{cases} 1-q & i \in I_0 \\ q & i \in I_1 \end{cases} \qquad \mathbb{P}(i \in I_1^q) = \begin{cases} q & i \in I_0 \\ 1-q & i \in I_1. \end{cases}$$

Relabelling perturbations are tested over values of $q \in \{0, 0.05, 0.1, 0.2, 0.5\}$ where $q = 0$ leaves collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ unchanged and $q = 0.5$ results in completely random allocation of point patterns to collections, expected to produce uniformly distributed p-values.

Additional points perturbations consider scenarios in which there exist additional microtubule centres which are not reported, potentially due to poor image quality. For each point pattern $\underline{x}$, a homogeneous Poisson point pattern $\underline{x}'$ is simulated according to HPPP$(\alpha\rho_0, W(\underline{x}))$ and $\underline{x} \cup \underline{x}'$ reported. The parameter value $\rho_0 = 10^{-4}$ is an approximation to the density of TACC3 point patterns. For each marked point pattern $\underline{y}$, a marked point pattern $\underline{y}'$ is simulated according to MPP(n, u, ϕ) and $\underline{y} \cup \underline{y}'$ reported. The number of marked points n is simulated as po(30α), with reference direction $u = \beta$ and maximum angle $\phi = 20$ degrees. The parameter α is varied over the set $\{0.2, 1, 3\}$ to represent increases in the number of points of approximately 20 percent, 100 percent and 300 percent. As the additional points in $\underline{x}'$ and $\underline{y}'$ are identically distributed across I_0 and I_1, increasing values of α are expected to reduce the reported significance of differences between the collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$.

Point location perturbations consider scenarios in which the location of microtubule cross-section centres are recorded with some degree of error. Point locations x_j ($x_{j,1}$ in the case of marked point patterns) are each shifted in a random direction by a distance uniformly sampled between zero and ϵ. The parameter ϵ takes values in the set $\{5, 20\}$, in comparison to the typical microtubule radius of 12.5.

3.4.6.2 Sensitivity analysis results

A summary of the 100 p-values obtained for each test statistic under each perturbation is provided by Table 3.5.

For relabelling perturbations the proportion of p-values reported significant at the five percent level reduces with increasing probability of mislabelling, α. The impact when $\alpha = 0.05$ is generally small, indicating that minor errors in allocation to collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ do not render the results useless. For $\alpha = 0.5$ a Kolmogorov-Smirnov test at the five percent significance level across the 100 p-values reports insufficient evidence to reject the uniform distribution of p-values on the interval $[0, 1]$ across all test statistics, as expected. Increasing values of α for point patterns, $\underline{x}$, appear to have a greater impact when unweighted test statistics are used, while the converse is true for marked point patterns, $\underline{y}$.

In the case of additional points perturbations, the proportion of p-values reported as significant at the five percent level reduces as the number of additional points increases with α. The impact when $\alpha = 0.2$ is generally small, indicating that the methodology is robust so long as a large proportion of points are identified. First order statistics report very little change as more points are added because original and perturbed observation window areas and number of points are strongly correlated. Increasing values of α again have a greater impact on unweighted test statistics for point patterns, $\underline{x}$.

Point location perturbations have a negligible impact on the proportion of p-values reported significant at the five percent level, indicating that the methodology is robust to small errors in point locations. For first order statistics this behaviour is expected as they are calculated independently of point locations. While perturbations to point locations does impact the value of second order summary statistics, the random perturbations do not make collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ more similar and thus the difference between them remains consistently significant.

3.5 Conclusions

3.5.1 Statistical methodology

This chapter presents a number of test statistics for the comparison of point patterns and marked point patterns of a particular form based upon existing and novel summary statistics. The suitability of each for a variety of scenarios has been highlighted by application to simulated data sets, the results of which indicate that the

		δ_N	δ_W	δ_ρ	$\delta_{\rho,\omega}$	δ_{nnd}	$\delta_{\mathrm{nnd},\omega}$	δ_{msd}	$\delta_{\mathrm{msd},\omega}$	δ_K	$\delta_{G,1}$	$\delta_{G,1,\omega}$	$\delta_{G,\infty}$	$\delta_{G,\infty,\omega}$	δ_{EFT}	$\delta_{\mathrm{EFT},\omega}$
Relabelling, α	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.05	0.95	0.92	1.00	0.91	0.88	0.90	0.94	0.93	0.03	0.86	0.90	0.80	0.88	0.96	0.87
	0.10	0.79	0.71	0.89	0.72	0.69	0.72	0.81	0.76	0.14	0.68	0.72	0.58	0.65	0.83	0.69
	0.20	0.49	0.44	0.61	0.47	0.38	0.38	0.39	0.43	0.10	0.33	0.39	0.28	0.39	0.40	0.33
	0.50	$0.06^{\ddagger}$	$0.06^{\ddagger}$	$0.05^{\ddagger}$	$0.04^{\ddagger}$	$0.06^{\ddagger}$	$0.06^{\ddagger}$	$0.06^{\ddagger}$	$0.04^{\ddagger}$	$0.08^{\ddagger}$	$0.06^{\ddagger}$	$0.06^{\ddagger}$	$0.06^{\ddagger}$	$0.08^{\ddagger}$	$0.04^{\ddagger}$	$0.04^{\dagger}$
Additional	0.2	1.00	1.00	1.00	1.00	0.67	0.97	0.92	1.00	0.02	0.72	0.98	0.66	0.97	1.00	1.00
points, α	1.0	1.00	1.00	0.96	1.00	0.35	0.53	0.56	0.76	0.03	0.40	0.62	0.22	0.50	0.90	0.84
	3.0	1.00	1.00	0.92	0.99	0.23	0.26	0.37	0.52	0.51	0.22	0.29	0.12	0.17	$0.06^{\ddagger}$	$0.05^{\ddagger}$
Point	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
location, ϵ	20	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.02	1.00	1.00	0.91	1.00	1.00	1.00

Table 3.5: Sensitivity study results from testing for differences between treatment and control TACC3 (marked) point pattern data for a number of different perturbation types, perturbation parameters and test statistics. Proportion of p-values in the range $[0, 0.05]$, ‡ indicates failure to reject uniformity of observed p-values under the Kolmogorov-Smirnov test at the Bonferroni corrected (Dunn, 1961) $5/13 = 0.38$ (for point patterns) or $5/2 = 2.5$ (for marked point patterns) percent significance level.

methodology is adept at detecting minor differences, smaller than those which may be detected by eye. Application to a biological data set has provided positive results, indicating significant differences between control and treatment observations which matches and extends a previous analysis of the same data (Nixon et al., 2015). A sensitivity analysis further reports that the testing procedure is not unduly affected by variation in the data on a scale consistent with measurement error.

We have proposed a total of 13 test statistics for application to 2D point pattern data and two further statistics for application to marked point pattern data for the comparison of collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ (respectively $\underline{y}^{I_0}$ and $\underline{y}^{I_1}$) assumed to be independent realisations of (marked) point processes $\underline{X}^{I_0}$ and $\underline{X}^{I_1}$ ($\underline{Y}^{I_0}$ and $\underline{Y}^{I_1}$). Alongside these statistics we propose the use of a nonparametric permutation hypothesis testing procedure to determine p-values of the proposed test statistics under the null hypothesis that $\underline{X}^0$ and $\underline{X}^1$ are identically distributed, given a collection of observed data $\underline{x}^I$.

The test statistics are designed to test for differences in various features of the point patterns under comparison, for example the number of points or the average distance between nearest neighbouring points. The particular feature tested will depend upon the data set being analysed and the questions of interest, but our suggestions cover a variety. Test statistics also differ in the format of the comparison, for example the difference between average nearest neighbour distances or through comparisons between functions which quantify nearest neighbour distances. Some of the proposed comparisons are more straightforward to present and interpret, while others are more sensitive to differences between the collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$.

A detailed simulation study has been used to highlight the various scenarios in which each of the proposed test statistics are effective. Manipulating the size of the difference between simulated collections $\underline{x}^{I_0}$ and $\underline{x}^{I_1}$ allows the sensitivity of each of the proposed approaches to be quantified, the results of which support the use of the proposed techniques to be able to detect even minor differences between collections. In the case of no difference between simulated data sets, there is insufficient evidence to reject the uniform distribution of p-values on the interval $[0, 1]$, indicating that the testing approach has good specificity.

3.5.2 Biological conclusions

The proposed methodologies are also applied to a collection of real biological data in which points represent microtubule locations under control, $\underline{x}^{I_0}$, and treatment, $\underline{x}^{I_1}$, conditions where TACC3 has been overexpressed. For this data set, investigation in particular of the distances between points is of interest as it is understood that nearby microtubules are bound together in K-fibers by a mesh-like structure and that TACC3 may have an impact on this mesh. Investigation by eye is insufficient for this application as the data sets appear visually very similar, so we apply the proposed methodology.

Testing of the observed data using all of the proposed test statistics results in a number of rejections of the null hypothesis. Summarising these results, we learn that K-fibers contain significantly more microtubules, the constituent microtubules are more tightly packed and variation in the orientation of microtubules is greater following overexpression of TACC3. A possible explanation for this may be that overexpression of TACC3 limits the distance over which the mesh may connect microtubules, bringing them closer together and disrupt their organisation. Due to the vital importance of K-fibers during mitosis, the results obtained by the analysis support more targeted investigation of the impact of TACC3 through further experimentation.

The observed data analysed in this chapter is also analysed by Nixon et al. (2015), for which discussions with us are acknowledged. There is an agreement in results between our works, although ours extends further by introducing test statistics based upon different features and with different measures of comparison. The largest difference between our work comes in the testing procedure, with Nixon et al. (2015) quantifying the significance of observed differences using t-tests reliant upon the assumption of normality, while we carry out permutation tests that require no similar assumptions. A significant new contribution of our work is to consider the analysis of marked point pattern data to make comment on the 3D structure of microtubules within K-fibers, with results indicating that overexpression of TACC3 has a significant effect on this structure.

The results of an in-depth sensitivity analysis of the proposed methodology suggest it to be robust to the small errors in data recording which are expected to be most likely. This is particularly valuable for our application, as images are produced through a human controlled imaging process from which coordinate locations are manually reported, with the possibility for errors at each step.

Chapter 4

Dependency between estimated local bulk movement patterns

This chapter introduces in more detail the methodology proposed for estimation and testing for dependence between local bulk movement patterns. The material has been made available as a working paper (Honnor et al., 2017b), and is presented in this chapter in a slightly reduced format as background information common to the whole thesis is presented in Chapter 2.

An introduction is first given to put the biological question in context, describe the data and formulate a statistical question which aims to answer the biological question with the data available. Following this the methodology is presented in Section 4.2. Section 4.3 describes the formulation of a validation study, the results of which are then presented and analysed. Application of the methodology to a set of TACC3 and EB3 biological data follows in Section 4.4 before summary of the conclusions in Section 4.5.

4.1 Introduction

For many phenomena in which recordings of a physical process are made across various times and locations, a question of interest is the evolution in location over time and the comparison of this evolution between observations. In cases of small numbers of clearly defined individual units, estimates of movement patterns may be obtained via object tracking algorithms. Such procedures may be used for application to the movement of a species of animal (McFarlane and Schofield, 1995) or the movement of specific subcellular structures (Chenouard et al., 2014).

Our interest extends further to physical processes for which individual units are not resolvable or for which a very large number of individual units precludes use of more standard object tracking procedures. The example considered in this chapter is the location of biomolecules of a particular species within a cell, as observed by microscope imaging. In this case each biomolecule is a discrete unit, but their large number and limits on microscope resolution make it impossible to identify and attempt to track each biomolecule. Observations are therefore interpreted as a density of biomolecules across space which evolves in time, from which we propose an estimation of movement patterns. An alternative example may be the distribution of water molecules, observed via the depth of water within a tank. The problem of determining movement patterns using a sequence of observations is an ill-posed inverse problem and as a result we provide an estimate of movement patterns averaged over local regions, which we term local bulk movement patterns.

Once movement patterns have been estimated, it may be desirable to compare between observations to determine whether the degree of dependence between movement patterns is statistically significant. In this chapter the purpose of determining dependence is to investigate whether two biomolecular species of interest interact, a process which typically requires biomolecules from each species to become conjoined, at which times they are similarly located and undergo similar movements. Comparison of movement patterns may also be of interest in other settings, for example for animal species to understand predator-prey relationships (Mitchell and Lima, 2002). We introduce a methodology which summarises and compares the high dimensional data resulting from estimation of movement patterns in a manner which is both meaningful and takes into account the spatial nature of the observations.

Proposed methodologies are applied to simulated processes with independent temporal evolutions, the results of which support the theoretical validity of the permutation testing procedure. Further testing on a number of different types of simulation with varying degrees of dependence indicate that the testing procedure is able to identify such scenarios with good power.

The methodology is then applied to an observed data set comprised of fluorescence microscopy images of TACC3, Transforming Acidic Coiled-Coil Containing Protein 3, and EB3, End-Binding protein 3. As an end binding protein, EB3 is known to locate at the growing end of microtubules during mitosis (Mimori-Kiyosue et al., 2000). It is further believed that TACC3 influences microtubule structure during mitosis (Booth et al., 2011), with our investigation considering whether EB3 and TACC3 undergo dependent movements in order to make inference on the location of TACC3, from which biologists may infer a function of the TACC3 protein.

4.1.1 Mathematical representation of data and statistical problem

We wish to analyse the spatio-temporal process M, that is $M_t(x) \in \mathbb{Q}^+$ for locations $x \in \chi$ and times $t \in \tau$, under a minimum of modelling assumptions. The requirement that M is non-negative and rational valued across all locations and times is necessary such that M may be likened to the distribution of a collection of basic units and realistic as recording and storage of M is digital, at least in the case of images collected over time. In situations where the number of basic units is very large, M may be likened to a density and tracking individual units becomes theoretically and computationally very difficult. We therefore propose a methodology to investigate bulk movement patterns on a scale greater than that of individual units for application in such cases.

The data available for the analysis of M is a collection of observation values $m_t(x) \in \mathbb{Q}^+$ across locations $x \in \Psi$ and times $t \in \Upsilon$. Observed values may be obtained from a realisation of the process M via projection or averaging and are expected to be subject to noise. Motivating examples have been provided in the introduction, with our focus in this chapter on M the distribution of biomolecules of a single species and $m_t(x)$ the light intensity emitted by biomolecules at the centre of pixel locations $x \in \Psi = \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$ at discrete time points $t \in \Upsilon = \{t_1, t_2, \ldots, t_{n_3}\}$ as recorded in the presence of background noise by a digital camera attached to a confocal microscope.

In future it will be necessary to refer to the collection of values across a set of locations $\psi \subseteq \Psi$ and times $\upsilon \in \Upsilon$ which will be denoted by $m_\upsilon(\psi)$. Our aim is to make inference on the dependency between the temporal evolution of two processes M^0 and M^1 via the comparison of two collections of observations, denoted by $m^0_\Upsilon(\Psi)$ and $m^1_\Upsilon(\Psi)$, over identical location, Ψ, and time, Υ, spaces.

4.2 Estimating movement patterns and a test for dependence

4.2.1 Approximation of movement

Under the specification in the previous section as a scaling of the number of basic units at each location at each time point, M may be interpreted as a spatial distribution of mass which evolves over time. The process describing the quantity of such mass moving from location x at time s to location y at time t may then be denoted by $F_{s,t}(x, y)$. We specifically consider direct dependency in movement patterns F^0 and F^1, corresponding to processes M^0 and M^1, such that $F^0_{s,t}(x, y)$ is positively associated with $F^1_{s,t}(x, y)$. That is, the quantity of mass moving from location x to y between time points s and t is positively associated between processes M^0 and M^1, across all pairs of locations and times. Although $F_{s,t}$ is described as a movement pattern, it also includes a description of those masses which remain fixed in place through $F_{s,t}(x, x)$, considered as movements which both start and end at the same location.

Determination of $F_{s,t}(x, y)$ using the information available in $m_{\{s,t\}}(\psi)$ may be formulated as a solution of the inverse problem

$$\left|\left|m_t(x) - \left(m_s(x) - \sum_{y\in\psi^*} F_{s,t}(x, y) + \sum_{y\in\psi^*} F_{s,t}(y, x)\right)\right|\right|_2 = 0, \tag{4.1}$$

of reducing a measure of the discrepancy between $m_t(\psi)$ and the result of applying all movements $F_{s,t}$ to $m_s(\psi)$ to zero, where ψ^* is the augmentation of ψ with the additional location z such that

$$\begin{aligned} m_s(z) &= \max\left\{\sum_{x\in\psi} m_t(x) - m_s(x), 0\right\} \\ m_t(z) &= \max\left\{\sum_{x\in\psi} m_s(x) - m_t(x), 0\right\}, \end{aligned}$$

to account for differences in total mass at times s and t. The solution of (4.1) is obtained by any $F_{s,t}$ satisfying the constraints

$$\begin{aligned} F_{s,t}(x, y) &\geq 0 && \forall x, y \in \psi^* \\ \sum_{y\in\psi^*} F_{s,t}(x, y) &= m_s(x) && \forall x \in \psi^* \\ \sum_{x\in\psi^*} F_{s,t}(x, y) &= m_t(y) && \forall y \in \psi^*. \end{aligned}$$

The first constraint ensures that only positive masses are moved, the second ensures that the total mass moving out of each location x at time s is $m_s(x)$ and the third ensures that the total mass moving into each location y at time t is $m_t(y)$.

The problem of solving (4.1) for $F_{s,t}$ satisfying these constraints is an ill-posed problem as the solution is typically not unique. We therefore consider $\hat{F}^{\psi}_{s,t} = \cup_{x,y\in\psi}\hat{F}_{s,t}(x, y)$, the solution to the regularised problem of

$$\hat{F}_{s,t} = \underset{\{F_{s,t}:\left|\left|m_t(x)-\left(m_s(x)-\sum_{y\in\psi^*} F_{s,t}(x,y)+\sum_{y\in\psi^*} F_{s,t}(y,x)\right)\right|\right|_2=0\}}{\arg\min} ||F_{s,t}||_c$$

according to the same set of constraints as previously and where

$$||F_{s,t}||_c = \sum_{x,y\in\psi} F_{s,t}(x, y)c(x, y)$$

for some cost function c. The solution of this regularised problem is again not necessarily unique (see further discussion in the following section), but the problem is closer to being well-posed and later steps in the methodology are designed to result in identical outcomes for a majority of $m_{\{s,t\}}(\psi)$.

The resulting estimator, $\hat{F}^{\psi}_{s,t}(x, y)$, may be interpreted as the set of movements which minimise the total cost of rearranging the distribution of mass $m_s(\psi)$ into $m_t(\psi)$ according to a cost function $c(x, y)$. This can be calculated using the earth mover's distance with

$$\hat{F}^{\psi}_{s,t} = \underset{f \in \eta^{\psi}_{s,t}}{\arg\min} \sum_{x,y \in \psi} f_{s,t}(x, y) c(x, y)$$

where $\eta^{\psi}_{s,t}$ is the set of all functions f satisfying

$$\begin{aligned} f_{s,t}(x, y) &\geq 0 && \forall x, y, \in \psi \\ \sum_{y \in \psi} f_{s,t}(x, y) &\leq m_s(x) && \forall x \in \psi \\ \sum_{x \in \psi} f_{s,t}(x, y) &\leq m_t(y) && \forall y \in \psi \\ \sum_{x,y \in \psi} f_{s,t}(x, y) &= \min \left\{ \sum_{x \in \psi} m_s(x), \sum_{y \in \psi} m_t(y) \right\}. \end{aligned}$$

For the analyses proposed we focus on a single cost function, $c(x, y) = ||x - y||_2$, equal to Euclidean distance. This cost function is chosen to be homogeneous and isotropic across $x, y \in \psi$ and penalises proposed movements only according to the distance moved, as we wish to avoid further assumptions. Returning to the interpretation of m as a distribution of mass, if we assume that applied forces are constant between s and t then $c(x, y) = ||x - y||_2$ is exactly the energy required to move unit mass from x to y. The estimated collection of movements $\hat{F}^{\psi}_{s,t}$ is then that which minimises the total energy required to rearrange $m_s(\psi)$ into $m_t(\psi)$. Some further discussion of the choice of cost function follows in Sections 4.2.2 and 4.2.6.

4.2.2 Movement summary statistic

The collection of movements $\hat{F}^{\psi}_{s,t}$ is an estimator of F^{ψ}, taking values in the high dimensional set $(\mathbb{Q}^{+})^{\psi \times \psi}$. To facilitate comparison between $\hat{F}^{\psi,0}$ and $\hat{F}^{\psi,1}$ resulting from observations m^0 and m^1 we first summarise the information in $\hat{F}^{\psi}$ via the summary statistic $\hat{S}^{\psi} \in (\mathbb{R}^{+})^8$.

Defining $a(v) \in (0, 2\pi]$ to be the anticlockwise angle between the vector $(1, 0)$ and the vector $v \in \mathbb{R}^2$, and sets $\theta_1 = (15\pi/8, 2\pi] \cup (0, \pi/8]$ and $\theta_j = ((2j-3)\pi/8, (2j-1)\pi/8]$ for $j \in \{2, 3, \ldots, 8\}$, elements of the summary statistic are given by

$$(\hat{S}^{\psi}_{s,t})_j = \sum_{x,y \in \psi : a(y-x) \in \theta_j} \hat{F}^{\psi}_{s,t}(x, y) ||x - y||_2 \qquad j \in \{1, 2, \ldots, 8\}.$$

An illustration of this summary is presented in Figure 4.1. To elaborate on the formulation of the summary statistic, $(\hat{S}^{\psi}_{s,t})_j$ is equal to $(t - s)$ multiplied by

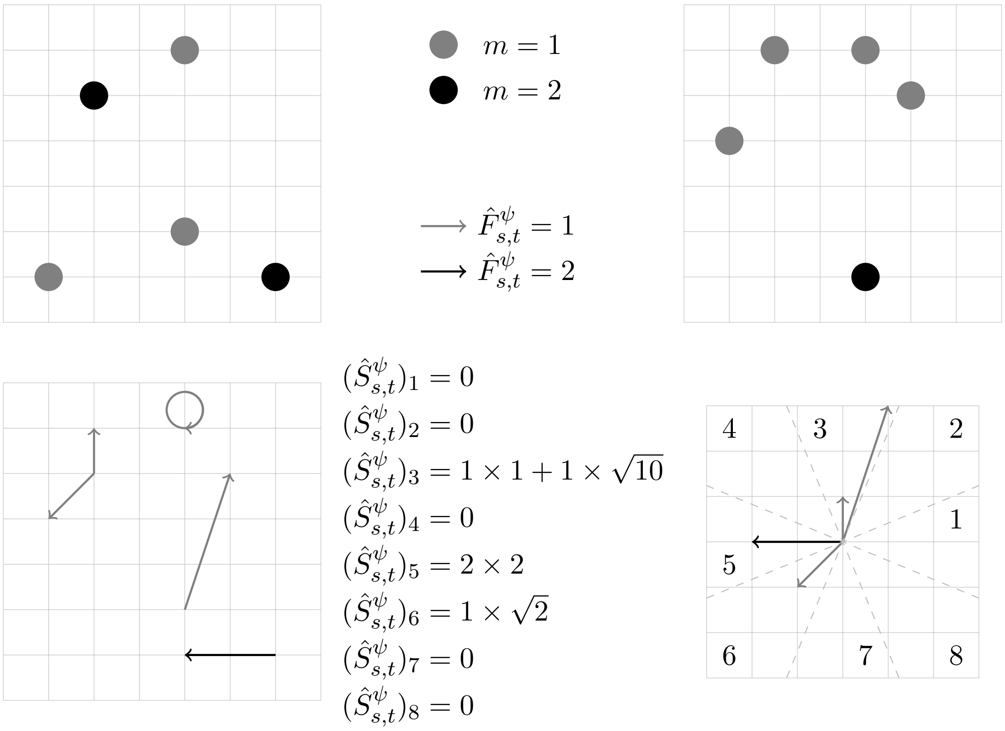

$m = 1$
$m = 2$
$\longrightarrow \hat{F}^{\psi}_{s,t} = 1$
$\longrightarrow \hat{F}^{\psi}_{s,t} = 2$
$(\hat{S}^{\psi}_{s,t})_1 = 0$
$(\hat{S}^{\psi}_{s,t})_2 = 0$
$(\hat{S}^{\psi}_{s,t})_3 = 1 \times 1 + 1 \times \sqrt{10}$
$(\hat{S}^{\psi}_{s,t})_4 = 0$
$(\hat{S}^{\psi}_{s,t})_5 = 2 \times 2$
$(\hat{S}^{\psi}_{s,t})_6 = 1 \times \sqrt{2}$
$(\hat{S}^{\psi}_{s,t})_7 = 0$
$(\hat{S}^{\psi}_{s,t})_8 = 0$
4
3
2
1
5
6
7
8

Figure 4.1: Distributions of mass $m_s(\psi)$, top left, and $m_t(\psi)$, top right, estimated movements $\hat{F}^{\psi}_{s,t}$, bottom left, and contributions to the calculation of the summary statistic $\hat{S}_{s,t}$, bottom right.

the total momentum in directions within an angle of $\pi/8$ of the cardinal or ordinal direction (E,NE,N,NW,W,SW,S,SE) corresponding to j. This formulation is specific to $\psi \subseteq \mathbb{R}^2$ as is the focus of this chapter, however, generalisation may be made to spaces of dimension other than two as $\hat{S}^{\psi}$ is intended to summarise the collection of movements or the bulk movement pattern within the space ψ.

The division of movements into eight different directions provides more detail than if they were for instance separated into the four cardinal directions. In the particular instance where ψ is a regular grid, $\psi = \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$, the proposed division also avoids peculiarities at the boundaries of sets θ_j as there exists no direction $y - x$ lying exactly on the border of any of the sets θ_j. To establish this, note that $a(y-x) = \pi/8$ if and only if $(y-x)_2/(y-x)_1 = \tan(\pi/8) = 1+\sqrt{2} \in \mathbb{R}\backslash\mathbb{Q}$, whereas ψ restricts $(y-x)_2/(y-x)_1 \in \mathbb{Q}$. A similar argument holds for all boundaries between θ_i and θ_{i+1}, each of which occurs at angles which are odd multiples of $\pi/8$.

Particular spatial distributions $m_s(\psi)$ and $m_t(\psi)$ may lead to non-uniqueness of $\hat{F}^{\psi}_{s,t}$ when $c(x, y) = ||x - y||_2$. For example, consider three colinear locations $\psi = \{x, y, z\}$ with $||x - y||_2 + ||y - z||_2 = ||x - z||_2$ and $\Upsilon = \{0, 1\}$ with $m_0(\psi) = (1, 1, 0)$ and $m_1(\psi) = (0, 1, 1)$. In such a scenario $\hat{F}^{\psi}_{0,1}(a, b) = \mathbb{1}\{a = x, b = z\}$ and

$\hat{F}^{\psi}_{0,1}(a,b) = \mathbb{1}\{a = x, b = y \text{ or } a = y, b = z\}$ both satisfy the regularised inverse problem presented in the previous section. However, in both cases the total cost

$$||\hat{F}^{\psi}_{0,1}||_c = \sum_{a,b\in\psi} \hat{F}^{\psi}_{0,1}(a,b)||a-b||_2 = 2,$$

is identical. The specification of $\hat{S}^{\psi}$ in terms of momentum, moving masses multiplied by distances moved, resolves this most common scenario of non-unique $\hat{F}^{\psi}$ into a consistent value of $\hat{S}^{\psi}$. Alternative scenarios in which $\hat{F}^{\psi}_{s,t}$ is not unique that result in different values of $\hat{S}^{\psi}_{s,t}$ do exist, but require particular values of m_s and m_t at more than three locations and complex interactions with values of m_s and m_t at the remaining locations in ψ, which are expected to be unlikely and increasingly so for larger spaces ψ.

4.2.3 Summary statistic comparison

Our interest is in the degree to which movement patterns differ between observations m^0 and m^1 and the significance of any difference. In order to quantify this we require a method of comparing summary statistics $\hat{S}^{\psi,0}_{s,t}$ and $\hat{S}^{\nu,1}_{s,t}$ between subsets $\psi, \nu \in \Psi$. We propose to make this comparison using the function $\hat{G}^{\psi,\nu}_{s,t}$.

Before specifying $\hat{G}^{\psi,\nu}_{s,t}$ we first define the intermediate quantity

$$\mu^{\psi,0}_{s,t}(x) = \sum_{j=1}^{8} \mathbb{1}\left\{x = e_j\left(\hat{S}^{\psi,0}_{s,t}\right)_j\right\} \qquad\qquad x \in \mathbb{R}^2,$$

where e_j is the unit length vector such that $a(e_j) = (j-1)\pi/4$. The intermediary $\mu^{\psi,0}_{s,t}$ may then be seen as the spatial distribution of eight unit masses, each of which is at a distance $(\hat{S}^{\psi,0}_{s,t})_j$ from the origin in the direction e_j.

The value of the comparison is then given by

$$\hat{G}^{\psi,\nu}_{s,t} = \mathrm{EMD}(\mu^{\psi,0}_{s,t}, \mu^{\nu,1}_{s,t}),$$

the minimal cost required to rearrange the eight unit masses with Euclidean distance as cost function. Small values of $\hat{G}$ imply close agreement of bulk movement patterns, with large values of $\hat{G}$ indicating differences.

Comparison of summarised movement patterns $\hat{S}$ is proposed using this method rather than a more straightforward alternative because the $\hat{S}$ summarise spatial information. The first element of $\hat{S}$ quantifies momentum in easterly directions (between east north east and east south east), with the second in north-easterly directions and the fifth in westerly directions. Intuitively, $\hat{S} = (0,1,0,0,0,0,0,0)$ is therefore more similar to $\hat{S} = (1,0,0,0,0,0,0,0)$ than to $\hat{S} = (0,0,0,0,1,0,0,0)$. Comparison using the earth mover's distance takes this into account, while approaches which treat $\hat{S}$ as a vector generally do not.

In the case where M is the distribution of a physical quantity the elements of $\hat{S}$ are proportional to the momentum of the movements, with $\hat{G}^{\psi,\nu}$ then proportional to the impulse required to transform $\hat{S}^{\psi,0}$ into $\hat{S}^{\nu,1}$.

4.2.4 Combination of summary statistic comparisons

As previously stated, the high dimensionality of estimated movement patterns, $\hat{F}$, makes them difficult to compare. We have therefore proposed summary statistics $\hat{S}$ and a method to compare them between observations m^0 and m^1 via $\hat{G}$.

A comparison of bulk movement patterns across the entire space Ψ between two consecutive time points is given by $\hat{G}^{\Psi,\Psi}_{s,t}$. However, aggregating movements over Ψ for large image spaces may result in unintuitive values of $\hat{G}^{\Psi,\Psi}_{s,t}$. For example, if movement in the upper half of Ψ is in easterly directions and in the lower half of Ψ is in westerly directions for m^0 and vice versa for m^1 then $\hat{S}^{\Psi,0}_{s,t}$ and $\hat{S}^{\Psi,1}_{s,t}$ are expected to be very similar and the value of $\hat{G}^{\Psi,\Psi}_{s,t}$ very small.

A more effective comparison may be obtained by comparing bulk movement patterns over smaller subregions of Ψ and then combining these values. We therefore consider the partition of Ψ into w subregions denoted by $\Psi^w = \{\psi_1, \psi_2, \ldots, \psi_w\}$ and the combined comparison statistic

$$\hat{H}^{\Psi^w,\Psi^w}_{s,t} = \sum_{j=1}^{w} \hat{G}^{\psi_j,\psi_j}_{s,t}.$$

Choice of the number, sizes and organisation of subregions clearly has an effect on the value of $\hat{H}$. In future examples we consider the regular grid location space $\Psi = \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$ which may be partitioned into identically sized square subregions which tessellate Ψ, an illustration of which may be seen in Figure 4.2. This restricts the number of subregions to $w = u_1 u_2$ for $u_1 = n_1/k$ and $u_2 = n_2/k$ for $u_1, u_2, k \in \mathbb{N}$. Selection of a larger value of w results in comparisons over smaller subregions, resulting in a value of $\hat{H}^{\Psi^w,\Psi^w}$ which is sensitive to differences in bulk movement patterns on a smaller local scale.

On the other hand, $\hat{H}^{\Psi^w,\Psi^w}_{s,t}$ is dependent upon estimated movement patterns $\hat{F}^{\psi_j}_{s,t}$ which are calculated using only the information in $m_{\{s,t\}}(\psi_j)$. Movements estimated using all available data, $\hat{F}^{\Psi}_{s,t}$, are expected to be the closest estimator of the true $F_{s,t}$, with the combination of regional estimates $\cup_{i=1}^{w} \hat{F}^{\psi_j}_{s,t}$ expected to decrease in accuracy with w as movements between subregions are not taken into consideration. There is therefore a balance in choosing w large enough that the local comparisons are sensitive, but small enough that the local movement patterns are accurate.

Partition of Ψ into Ψ^w has computational benefits, discussed in more detail in Section 4.2.6.

Assuming mass is evenly distributed over the space and moves at a consistently distributed speed, the magnitude of $\hat{S}^{\psi_j}$ is approximately proportional to subregion size which is in turn inversely proportional to the number of subregions. Values of $\hat{H}^{\Psi^w,\Psi^w}$ may therefore be considered to be on the same scale regardless of the choice of w. Despite this, we do not compare $\hat{H}^{\Psi^w,\Psi^w}$ for different values of w as we are more interested in tests of the significance of $\hat{H}^{\Psi^w,\Psi^w}$ through comparison to $\hat{H}^{\Psi^w,\lambda\Psi^w}$ for some rearrangement operator λ on the collection of subregions Ψ^w.

4.2.5 Significance quantification

Our aim is to determine whether or not the evolutions of M^0 and M^1 over time are dependent. The formulation of $\hat{H}^{\Psi^w,\Psi^w}$ gives an insight into this, with small values indicating similarity in local bulk movement patterns and evidence of dependence and the converse for large values. However, without making further assumptions on the evolution of processes M^0 and M^1 it is not possible specify a parametric distribution of $\hat{H}$ under which its significance may be quantified. We therefore consider nonparametric testing of the significance of the observed value $\hat{H}^{\Psi^w,\Psi^w}_{s,t}$.

In particular, we consider a permutation test (previously introduced in Section 2.3.3) under the action of $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_w\} \in \Lambda$ on the collection of subregions $\Psi^w = \{\psi_1, \psi_2, \ldots, \psi_w\}$, with $\lambda\Psi^w = \{\lambda_1\psi_1, \lambda_2\psi_2, \ldots, \lambda_w\psi_w\}$ chosen such that under a specified null hypothesis H_0 there exists exchangeability of the set $\{S^{\psi_1}_{s,t}, S^{\psi_2}_{s,t} \ldots, S^{\psi_w}_{s,t}\}$ under the action of any $\lambda \in \Lambda$. That is

$$\{S^{\psi_1}_{s,t}, S^{\psi_2}_{s,t} \ldots, S^{\psi_w}_{s,t}\} \overset{\text{d}}{=} \{S^{\lambda_1\psi_1}_{s,t}, S^{\lambda_2\psi_2}_{s,t} \ldots, S^{\lambda_w\psi_w}_{s,t}\}.$$

In practice, the set Λ is typically too large to calculate $\hat{H}^{\Psi^w,\lambda\Psi^w}_{s,t}$ for all $\lambda \in \Lambda$ and so an approximate permutation test is carried out using a random subset Λ' of Λ which includes the identity operator $\lambda^0 : \lambda^0\Psi^w = \Psi^w$. The resulting lower tail p-value is then given by

$$p = \frac{1}{|\Lambda'|} \sum_{\lambda \in \Lambda'} \mathbb{1}\{\hat{H}^{\Psi^w,\lambda\Psi^w}_{s,t} \leq \hat{H}^{\Psi^w,\Psi^w}_{s,t}\}.$$

Each null hypothesis tested according to this method is comprised of three statements. The first statement is that there is between-sample independence of local bulk movement patterns, that is $\{S^{\psi_1,0}_{s,t}, S^{\psi_2,0}_{s,t}, \ldots, S^{\psi_w,0}_{s,t}\}$ is independent of $\{S^{\psi_1,1}_{s,t}, S^{\psi_2,1}_{s,t}, \ldots, S^{\psi_w,1}_{s,t}\}$. The second statement is required to specify the set of operations Λ under which permutation testing is carried out, in particular those operations for which the marginal distributions are identical, $S^{\psi_j}_{s,t} \overset{\text{d}}{=} S^{\lambda_j\psi_j}_{s,t}$, examples of which are provided in Section 4.2.5.2. The final statement is that there is within-sample independence of local bulk movement patterns. That is $S^{\psi_j}_{s,t}$ is independent of $S^{\psi_k}_{s,t}$ for $j \neq k$, required to ensure exchangeability under the action of $\lambda \in \Lambda$.

Within-sample dependence between $S_{s,t}^{\psi_j}$ and $S_{s,t}^{\psi_k}$ is caused by movements between subregions, $F_{s,t}(x, y) > 0$ for $x \in \psi_j, y \in \psi_k$. In cases where the interval between time points, $t - s$, is small in comparison to the speed at which mass moves in M, $F_{s,t}(x, y)$ is expected be non-zero only for those locations x and y close to the border between adjacent subregions ψ_j and ψ_k. In the rest of this chapter we consider observations from designed experiments for which the time points are chosen close together to satisfy this. The proportion of movements which are inter-subregion may be further limited by reducing the number of subregions w, limiting the proportion of the space Ψ which is adjacent to a boundary between subregions.

Variation in the second statement of the null hypothesis allows for the testing of independence under differing assumptions. In Section 4.2.5.2 four example null hypotheses are given with corresponding sets of operators Λ. However, the methodology is by no means restricted to these four null hypotheses, but rather is more broadly applicable to any null hypothesis for which a suitable set of operations Λ may be determined under which the S^{ψ_j} are exchangeable. Sets of operations Λ are typically created based upon null hypothesis statements that permit reflection or rotational symmetry.

4.2.5.1 Operator definitions

In the following sections we will repeatedly refer to particular rearrangements of the sets ψ_j which we therefore define here.

Let R_e define a set of rearrangements of ψ_j based upon reflections. Firstly, $\rho_1\psi_j$ is the rearrangement based upon the reflection of locations $x \in \psi_j$ across the line passing through the centre of ψ_j in the direction $(1, 0)$. Similarly, $\rho_2\psi_j$ across the line in direction $(1, 1)$, $\rho_3\psi_j$ across the line in direction $(0, 1)$ and $\rho_4\psi_j$ across the line in direction $(-1, 1)$. Further, let $\rho_0\psi_j = \psi_j$ be the identity rearrangement.

Let R_o define a set of rearrangements of ψ_j based upon rotations. Firstly, $\varrho_1\psi_1$ is the rearrangement based upon the rotation of locations $x \in \psi_j$ anticlockwise about the centre of ψ_j by angle $\pi/2$. Similarly, $\varrho_2\psi_j$ by an angle of π and $\varrho_3\psi_j$ by an angle of $3\pi/2$. Further, let $\varrho_0\psi_j = \psi_j$ be the identity rearrangement.

A diagram illustrating $\rho \in R_e$ and $\varrho \in R_o$ for Ψ may be seen in Figure 4.2.

Applied in combination as $\rho\varrho$ or $\varrho\rho$ with $\rho \in R_e = \{\rho_0, \rho_1, \rho_2, \rho_3, \rho_4\}$ and $\varrho \in R_0 = \{\varrho_0, \varrho_1, \varrho_2, \varrho_3\}$ a number of permutations are identical. For example, $\rho_3\varrho_1\psi_j = \rho_2\varrho_0\psi_j$. There are a total of eight unique transformations of ψ_j of this form, one listing of which is $\{\rho_0\varrho_0, \rho_1\varrho_0, \rho_2\varrho_0, \rho_3\varrho_0, \rho_4\varrho_0, \rho_0\varrho_1, \rho_0\varrho_2, \rho_0\varrho_3\}$.

The value of $\hat{S}_{s,t}^{\rho\varrho\psi_j}$ for any of the $\rho \in \{\rho_0, \rho_1, \rho_2, \rho_3, \rho_4\}$ and $\varrho \in \{\varrho_0, \varrho_1, \varrho_2, \varrho_3\}$ is obtained by a straightforward rearrangement of $\hat{S}_{s,t}^{\psi_j}$. For example, $\hat{S}_{s,t}^{\rho_3\varrho_1\psi_j} = ((\hat{S}_{s,t}^{\psi_j})_3, (\hat{S}_{s,t}^{\psi_j})_2, (\hat{S}_{s,t}^{\psi_j})_1, (\hat{S}_{s,t}^{\psi_j})_8, (\hat{S}_{s,t}^{\psi_j})_7, (\hat{S}_{s,t}^{\psi_j})_6, (\hat{S}_{s,t}^{\psi_j})_5, (\hat{S}_{s,t}^{\psi_j})_4)$.

4.2.5.2 Example hypotheses

Definition 7. The isotropic null hypothesis, H_0^I, specifies on a scale according to w between-sample independence of local bulk movement patterns, that local bulk movement patterns, S^{ψ_j}, are homogeneous and isotropic across the whole space Ψ and within-sample independence of local bulk movement patterns. The corresponding set of operators for which there exists exchangeability under H_0^I is denoted by Λ^I.

Under the assumption of isotropy, local bulk movement patterns S^{ψ_j} are identically distributed under rearrangement of ψ_j by rotation and reflection represented by the action of $\rho\varrho, \rho \in R_e, \varrho \in R_o$ on ψ_j. Further, under the assumption of homogeneity, local bulk movement patterns S^{ψ_j} are identically distributed under any rearrangement of subregions, realised as a reordering of $\Psi^w = \{\psi_1, \psi_2, \ldots, \psi_w\}$. The set of operators Λ^I is therefore

$$\Lambda^{H,I} = \{(\lambda_1, \lambda_2, \ldots, \lambda_w) : \lambda_j \psi_j = \rho\varrho\psi_{\sigma(j)}, \rho \in R_e, \varrho \in R_o, \sigma \in \mathcal{S}_w\},$$

where $\mathcal{S}_w$ is the symmetric group of size w. The total number of unique operators is $|\Lambda^I| = 8^w \times w!$.

Definition 8. The homogeneous null hypothesis, H_0^H, specifies on a scale according to w between-sample independence of local bulk movement patterns, that local bulk movement patterns, S^{ψ_j}, are homogeneous across the whole space Ψ and within-sample independence of local bulk movement patterns. The corresponding set of operators for which there exists exchangeability under H_0^H is denoted by Λ^H.

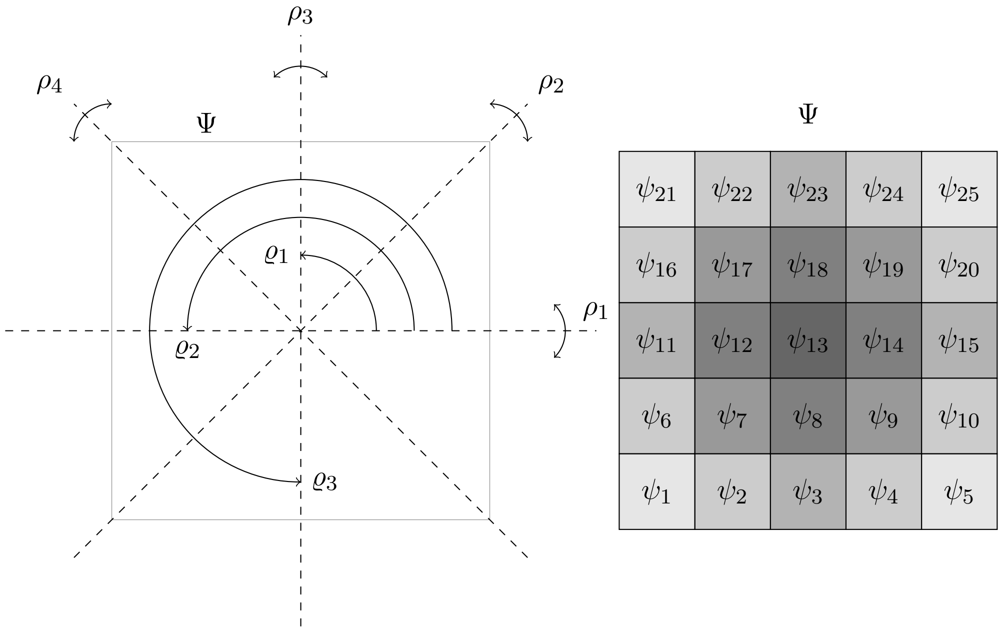
Under the sole assumption of homogeneity the set of operators Λ^H is given by

$$\Lambda^H = \{(\lambda_1, \lambda_2, \ldots, \lambda_w) : \lambda_j \psi_j = \psi_{\sigma(j)}, \sigma \in \mathcal{S}_w\}.$$

The total number of unique operators is $|\Lambda^H| = w!$. Note that $\Lambda^H \subset \Lambda^I$.

Definition 9. The symmetric null hypothesis, H_0^S, specifies on a scale according to w between-sample independence of local bulk movement patterns, that local bulk movement patterns, S^{ψ_j}, are symmetric across the whole space under the application of $\rho\varrho, \rho \in R_e, \varrho \in R_o$ to Ψ and within-sample independence of local bulk movement patterns. The corresponding set of operators for which there exists exchangeability under H_0^S is denoted by Λ^S.

Symmetry in the distribution of S^{ψ_j} under a limited set of rotations and reflection of Ψ allows us to partition Ψ^w into classes for which ψ_j and ψ_k are members of the same class if and only if $S^{\psi_j} \overset{\mathrm{d}}{=} S^{\rho\varrho\psi_k}$ for some suitable choice of ρ and ϱ.



ψ_{21}	ψ_{22}	ψ_{23}	ψ_{24}	ψ_{25}
ψ_{16}	ψ_{17}	ψ_{18}	ψ_{19}	ψ_{20}
ψ_{11}	ψ_{12}	ψ_{13}	ψ_{14}	ψ_{15}
ψ_{6}	ψ_{7}	ψ_{8}	ψ_{9}	ψ_{10}
ψ_{1}	ψ_{2}	ψ_{3}	ψ_{4}	ψ_{5}

Figure 4.2: Illustration of rotations and reflections of Ψ, left, and the division of Ψ into subregions $\psi_j \in \Psi^{25}$ with shading according to the exchangeability classes under the symmetric null hypothesis, Definition 9, right.

The specification of symmetry under the application of all rotations and reflections $\rho\varrho, \rho \in R_e, \varrho \in R_o$ to Ψ requires Ψ to be square.

An example in the case of $w = 25$ is illustrated in Figure 4.2, for which the exchangeability classes are $\{\psi_{13}\}$, $\{\psi_8, \psi_{12}, \psi_{14}, \psi_{18}\}$, $\{\psi_7, \psi_9, \psi_{17}, \psi_{19}\}$, $\{\psi_3, \psi_{11}, \psi_{15}, \psi_{23}\}$, $\{\psi_2, \psi_4, \psi_6, \psi_{10}, \psi_{16}, \psi_{20}, \psi_{22}, \psi_{24}\}$ and $\{\psi_1, \psi_5, \psi_{21}, \psi_{25}\}$. Examples of allowed transformations are

$$\begin{aligned} S^{\rho\varrho\psi_{13}} &\overset{\mathrm{d}}{=} S^{\psi_{13}} && \forall \rho \in R_e, \varrho \in R_o \\ S^{\rho_3\psi_2} &\overset{\mathrm{d}}{=} S^{\psi_4} \\ S^{\rho_4\psi_8} &\overset{\mathrm{d}}{=} S^{\varrho_1\psi_8} \overset{\mathrm{d}}{=} S^{\psi_{14}}. \end{aligned}$$

The total number of unique operators is

$$|\Lambda^S| = \begin{cases} 8 & \text{if } w = 1 \\ 2^4 \times 4! & \text{if } w = 4 \\ 8 \times (2^4 \times 4!) & \text{if } w = 9 \\ (k \times 2^4 \times 4!) \times \left(\frac{k(k-1)}{2} \times 8!\right) & \text{if } w = (2k)^2, k = 2, 3, \dots \\ 8 \times (2k \times 2^4 \times 4!) \times \left(\frac{k(k-1)}{2} \times 8!\right) & \text{if } w = (2k+1)^2, k = 2, 3, \dots. \end{cases}$$

Note that $\Lambda^S \subset \Lambda^I$ as Λ^I includes all rearrangements, rotations and reflections and Λ^S includes only those which preserve the classes of Ψ^w. Further, $\Lambda^S \nsubseteq \Lambda^H$

as Λ^H includes only rearrangements of subregions and Λ^S requires rotations and reflections of subregions in combination with rearrangements, and $\Lambda^H \not\subseteq \Lambda^S$ as Λ^H includes all rearrangements of subregions and Λ^S restricts rearrangements within the classes partitioning Ψ^w.

Definition 10. The horizontal reflection null hypothesis, H_0^R, specifies on a scale according to w between-sample independence of local bulk movement patterns, that local bulk movement patterns, S^{ψ_j}, are horizontally symmetric across the whole space under the application of ρ_3 to Ψ and within-sample independence of local bulk movement patterns. The corresponding set of operators for which there exists exchangeability under H_0^R is denoted by Λ^R.

Under H_0^R subregions ψ_j are each paired with the corresponding subregion located at the same position as $(\rho_3\Psi^w)_j$ for which rearrangement is allowed under the application of ρ_3 to both ψ_j and $(\rho_3\Psi^w)_j$.

In the case where Ψ^w is u_1 subregions wide and u_2 subregions high, with $w = u_1u_2$, the total number of unique operators is

$$|\Lambda^R| = 2^{\lceil \frac{u_1}{2} \rceil u_2},$$

where $\lceil u \rceil$ is the value of u rounded up to the nearest integer.

4.2.6 Computational considerations

Estimation of local bulk movement patterns $\hat{S}^{\psi}_{s,t}$ first requires estimation of movement patterns $\hat{F}^{\psi}_{s,t}$ resulting from calculation of the earth mover's distance between $m_s(\psi)$ and $m_t(\psi)$. Calculation of the earth mover's distance is based upon solution of an assignment problem, the computational cost of which is superlinear in the number of origin and destination locations $x, y \in \psi$, requiring $O(|\psi|^3 \log|\psi|)$ operations (Rubner et al., 2000). Partitioning Ψ into w subregions, ψ_j, therefore results in a collection of $\hat{S}^{\psi_j}_{s,t}$ which may be determined at less overall computational cost than $\hat{S}^{\Psi}_{s,t}$, by a factor of $1/w^2$ as

$$\begin{aligned} w \times O(|\psi_j|^3 \log|\psi_j|) &= O(w|\psi_j|^3 \log|\psi_j|) \\ &= O\left(w\left(\frac{|\Psi|}{w}\right)^3 \log\frac{|\Psi|}{w}\right) \\ &= O\left(\frac{1}{w^2}|\Psi|^3 \log\Psi - \frac{1}{w^2}|\Psi|^3 \log w\right) \\ &= \frac{1}{w^2} O(|\Psi|^3 \log|\Psi|). \end{aligned}$$

Furthermore, as calculation of $\hat{S}^{\psi_j}_{s,t}$ is independent of calculation of $\hat{S}^{\psi_k}_{s,t}$ for disjoint sets ψ_j and ψ_k, partitioning of Ψ into subregions ψ_j permits parallelisation of earth mover's distance calculations.

The cost function for estimating movements, $\hat{F}^{\psi}_{s,t}(x,y)$, is specified as $c(x,y) = ||x-y||_2$, the Euclidean distance between locations x and y. This cost function satisfies the triangle inequality, that is for locations x, y and z we have $c(x,z) \leq c(x,y) + c(y,z)$. An interpretation of the triangle inequality for our application is that it is always as or more expensive to move mass from x into y and an equal quantity of mass from y to z than it is to move mass directly from x to z. As a result, provided c satisfies the triangle inequality, we can state before calculation of the earth mover's distance that $\hat{F}^{\psi}_{s,t}(x,x) = \min\{m_s(x), m_t(x)\}$. Calculation of $\hat{F}^{\psi}_{s,t}(x,y)$ may therefore be based upon the collection of data $\{m_{s,+}(\psi), m_{t,-}(\psi)\} = \{m_s(x) - \min\{m_s(x), m_t(x)\}, m_t(x) - \min\{m_s(x), m_t(x)\}, x \in \psi\}$. However, for every pair $m_{s,+}(x), m_{t,-}(x)$ at least one is zero. Therefore, the total number of origin and destination locations is reduced by at least half, resulting in a large computational saving when c satisfies the triangle inequality.

As stated in Section 4.2.5.1, the value of $\hat{S}^{\rho\varrho\psi_j}_{s,t}$ may be obtained without further calculation from $\hat{S}^{\psi_j}_{s,t}$ by rearrangement if $\rho \in R_e$ and $\varrho \in R_o$. In cases where the null hypothesis permits operations of rearrangement, reflection and rotation dramatic savings can therefore by made by avoiding repeated recalculation of the earth mover's distance.

4.3 Validation study

To confirm the suitability of the methods described in the previous section, we first apply them to simulated data modelling varying degrees of dependence which our methodology is designed to detect.

4.3.1 Simulation description

A general simulation may be made up of a total of k objects, where the centre of object i at time t is denoted by $c_i(t)$ and its intensity by b_i. The evolution over time of the process is determined by the relationship between $c_i(s)$ and $c_i(t)$ for time points s and t. The shape of object i is specified by $d_i(x)$, the set of points belonging to shape i whose centre is located at x. In this framework simulations may be thought of as sequences of germ-grain models, with point locations or germs $c_i(t)$ associated with sets corresponding to object shapes or grains $d_i(x)$. Simulations may include an observation error term $\epsilon(x,t)$ at locations x at times t. Under such a formulation we may specify

$$m(x,t) = \sum_{i=1}^{k} b_i \mathbb{1}\{x \in d_i(c_i(t))\} + \epsilon(x,t)$$

$$F_{s,t}(x,y) = \sum_{i=1}^{k} b_i \mathbb{1}\{x \in d_i(c_i(s)), y - x = c_i(t) - c_i(s)\}.$$

For the validation study investigated in this section we consider a discrete location space $\Psi = \{1, 2, \ldots, 60\} \times \{1, 2, \ldots, 60\}$ and a discrete time space $\Upsilon = \{1, 2, 3, 4, 5\}$. The number of objects, k, differs between simulation classes but objects are all of the same intensity $b_i = 30$. Objects are further all the same size and shape, that of a Greek cross of the four locations directly adjacent to $x = (x_1, x_2)$ and x itself

$$d_i(x) = \{x, (x_1, x_2 - 1), (x_1, x_2 + 1), (x_1 - 1, x_2), (x_1 + 1, x_2)\},$$

and initially independently distributed uniformly over Ψ, $c_i(1) \sim \text{Uniform}\{\Psi\}$. The evolution of object centres over time also differs between simulations and is therefore described separately.

For each class of simulation 15 replicates are produced, allowing 105 unique comparisons between simulations of the same class. Dependence between movement patterns in observations of the same class is introduced by the inclusion of a proportion of identical objects in each simulation.

Definition 11. Noise simulations contain no objects, $k = 0$, and therefore represent only Poisson distributed white noise with mean three, that is $\epsilon(x,t) \sim \text{Poisson}(3)$ independently for all times $t \in \Upsilon$ and locations $x \in \Psi$.

There is no dependence between noise simulations. The white noise specified for noise simulations is also included in all further simulations.

Definition 12. Isotropic simulations contain 100 objects, $k = 100$, for which

$$c_i(t+1) = c_i(t) + 3(\cos(\theta_{i,t}), \sin(\theta_{i,t})) \qquad \mod 60,$$

for $\theta_{i,t} \sim \text{Uniform}[0, 2\pi)$ simulated independently for each object and each time point. Centre locations are calculated under the specification that 60 mod 60 = 60 and with rounding of $c_i(t)$ to the nearest location in Ψ.

The evolution of object centres for isotropic simulations is independent of location and uniform across all directions, resulting in movement patterns which are both isotropic, homogeneous and reflection and rotationally symmetric under combinations $\rho\varrho, \rho \in R_e, \varrho \in R_o$ applied to Ψ. Taking object centres modulo 60 ensures that all objects remain within the location space Ψ, appearing to wrap around from each edge to the opposite edge.

Isotropic 10 simulations each contain the same 10 objects and 90 objects simulated independently for each observation. Similarly, isotropic 30 simulations each contain the same 30 objects and 70 objects simulated independently for each observation.

Definition 13. Homogeneous simulations contain 100 objects, $k = 100$, for which

$$c_i(t+1) = c_i(t) + 3\left(\cos\left(\frac{\pi}{4}\right), \sin\left(\frac{\pi}{4}\right)\right) \qquad \bmod 60,$$

again under the specification that 60 mod 60 = 60 and with rounding of $c_i(t)$ to the nearest location in Ψ.

The evolution of object centres for homogeneous simulations is independent of location, resulting in movement patterns which are homogeneous. Evolutions are also symmetric under reflection of Ψ by ρ_2 as centres move in the direction of the vector $(1, 1)$. Movement patterns in homogeneous simulations are neither isotropic nor reflection or rotationally symmetric in any other way. Taking object centres modulo 60 again ensures that all objects remain within the location space Ψ.

Homogeneous 10 simulations each contain the same 10 objects and 90 objects simulated independently for each observation. Similarly, homogeneous 30 simulations each contain the same 30 objects and 70 objects simulated independently for each observation.

Definition 14. Symmetric simulations contain 100 objects, $k = 100$, for which

$$c_i(t+1) = c_i(t) + 3\frac{z - c_i(t)}{||z - c_i(t)||},$$

where $z = (30.5, 30.5)$ is the point at the centre of Ψ and locations $c_i(t)$ are rounded to the nearest location in Ψ.

The evolution of object centres for symmetric simulations produces movement patterns which are reflection and rotationally symmetric under combinations $\rho\varrho, \rho \in R_e, \varrho \in R_o$ applied to Ψ. Movement patterns in symmetric simulations are neither homogeneous nor isotropic, additionally differing from all other described simulations in that the distribution of the collection of objects is expected to vary over time as they aggregate towards the centre of Ψ.

Symmetric 10 simulations each contain the same 10 objects and 90 objects simulated independently for each observation. Similarly, symmetric 30 simulations each contain the same 30 objects and 70 objects simulated independently for each observation.

Object speeds have been fixed at three as a compromise, with speeds smaller than three resulting in movements which may be accurately determined by eye alone and speeds larger than three expected to limit accuracy of $\hat{F}$ as an estimator of F. The following section provides the results of testing the described classes of simulations under a variety of hypotheses for a range of subregion sizes and thus a range of values of w. With square subregions of width 10, 12, 15 and 20 (corresponding to $w = 36$, 25, 16 and 9), object speeds of three ensures that the proportion of objects moving between subregions is small.

Illustrative examples of each class of simulation are presented in Figure 4.3.

4.3.2 Study results

4.3.2.1 Movement pattern estimation

Figure 4.4 displays a single example of the movements obtained via estimation of $\hat{F}_{s,t}$ and the true movements given by $F_{s,t}$. Data is a single isotropic simulation of 10 objects on a region of size 20 by 20, with the additional condition that objects must remain within the observation window.

Generally good agreement is observed between $\hat{F}_{s,t}$ and $F_{s,t}$, in particular when objects are well separated. In cases where objects overlap and trajectories cross there is a reduction in accuracy which is to be expected based upon the formulation of the estimator $\hat{F}_{s,t}$. Investigation of this example and others suggest that the proposed method of determining $\hat{F}_{s,t}$ results in a reasonable estimator of $F_{s,t}$, the accuracy of which may be increased by reducing the size of the time step, $t - s$, to restrict the potential for object trajectories to cross.

4.3.2.2 Independent simulations

The first set of tests carried out are for each of the independent simulations under each of the hypotheses H_0^I, H_0^H and H_0^S. These tests are carried out across all four pairs of consecutive time points $1 \to 2$, $2 \to 3$, $3 \to 4$ and $4 \to 5$, and across each of the four considered subregion sizes 20, 15, 12 and 10 corresponding to $w = 9$, 16, 25 and 36. Table 4.1 summarises the results of these approximate permutation tests under 10 000 permutations, presenting the results as the proportion of the 105 returned between-sample p-values which are less than 0.05. Furthermore, † is used to denote those collections of between-sample p-values which reject the omnibus null hypothesis of Uniform[0,1] distribution of between-sample p-values according to the Kolmogorov-Smirnov test (Massey, 1951) at the five percent level. Issues of multiple testing specific to this investigation are discussed in Section 4.3.2.4.

Noise simulations contain no objects and therefore no structured movements. As

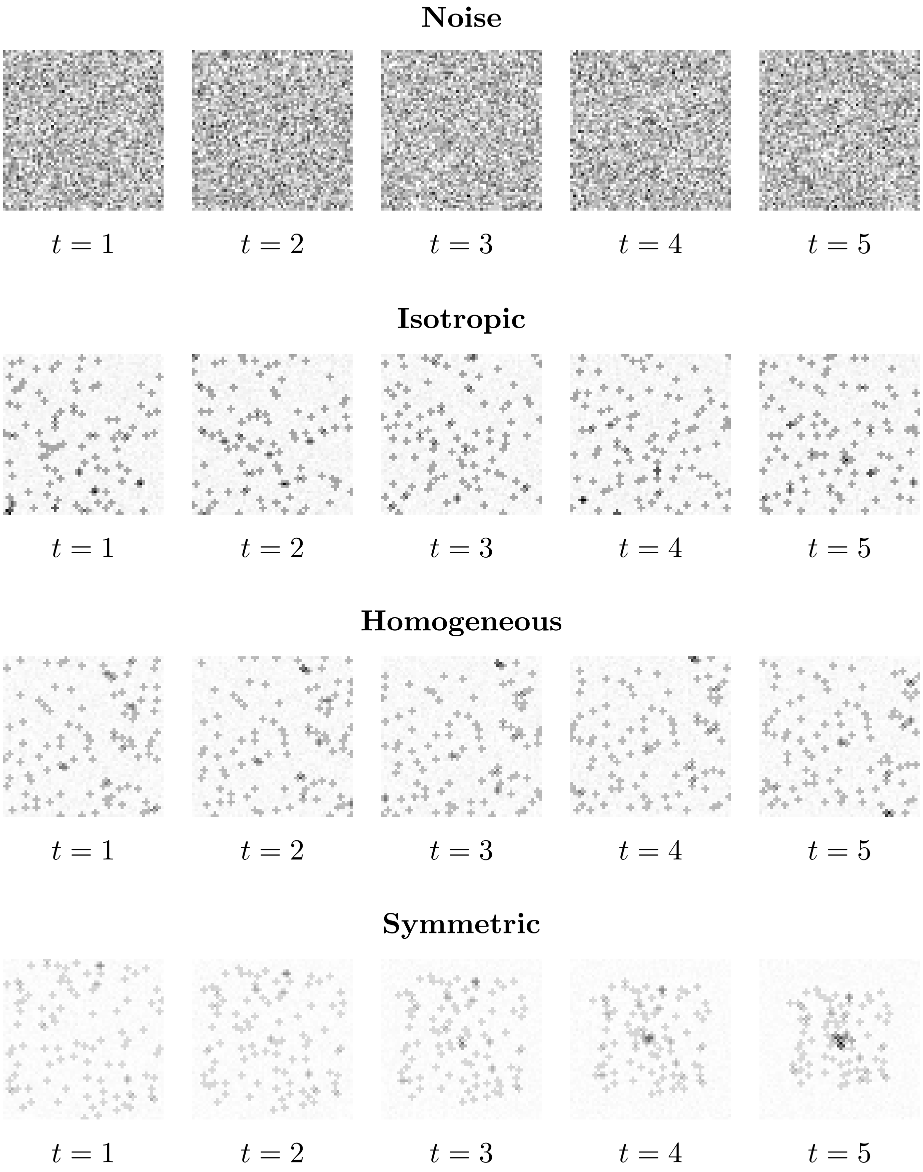


Figure 4.3: Examples of simulated spatio-temporal data. Pixel intensities correspond to values of m, linearly scaled such that the maximum value of m across all time points is black and the value $m = 0$ is white.

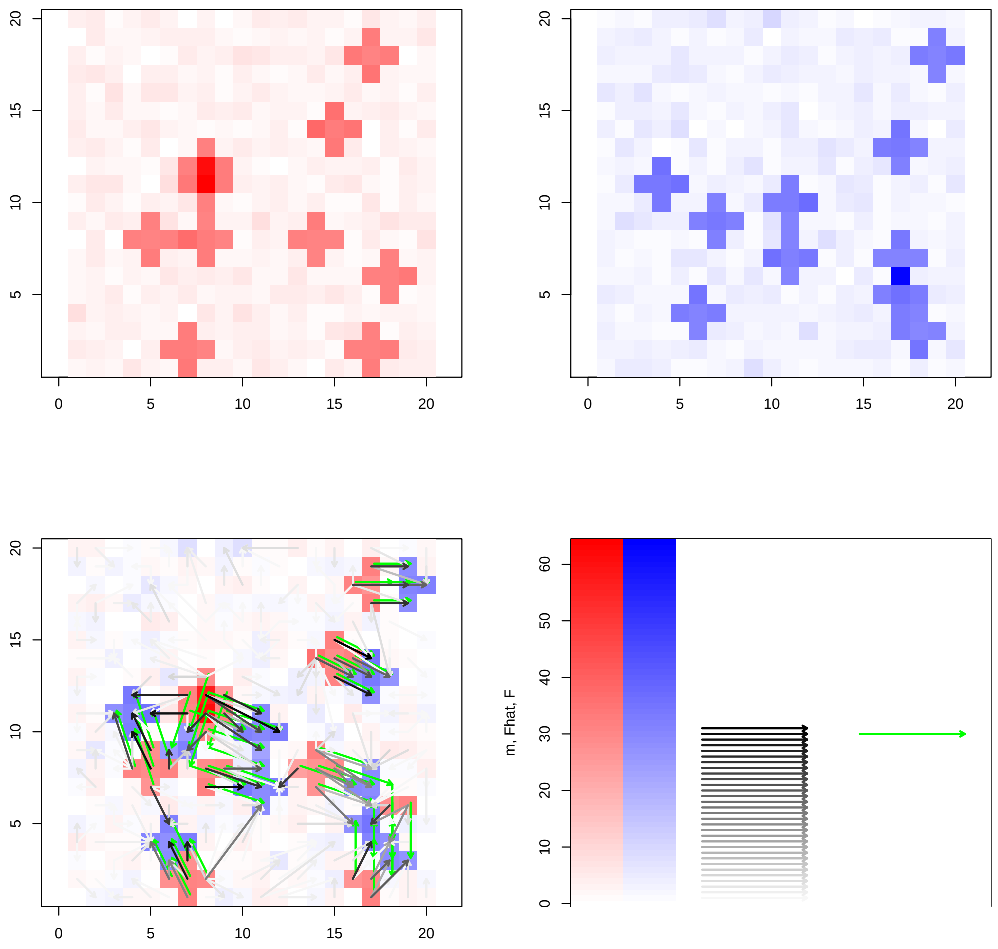


Figure 4.4: Illustration of observed mass distributions m_s, top left, m_t, top right, estimated movements $\hat{F}_{s,t}$ and true movements $F_{s,t}$, bottom left, for a single simulation. A scale is provided in the bottom right. Simulated movements are represented by green arrows, with grey arrows used to represent estimated movements of varying quantities of mass and mass distributions displayed as shades of red for m_s and blue for m_t. The background image of the bottom left plot displays $m_s - \min\{m_s, m_t\}$ in red and $m_t - \min\{m_s, m_t\}$ in blue, the information on which movements are estimated.

a result, local bulk movement patterns, $\hat{S}^{\psi_j}$, are expected to be identically distributed under the action of all operators $\lambda \in \Lambda^I, \Lambda^H$ and Λ^S. Between-sample independence of simulations and within-sample independence of local bulk movement patterns $\hat{S}^{\psi_j}$ induced by construction therefore suggests that the null hypotheses should be satisfied, reflected in Table 4.1 by values close to 0.05 and a general absence of markers †. This is the observed behaviour, with the two marked occasions rejecting the omnibus null hypothesis doing so with p-values of 0.013 and 0.009.

Isotropic simulations are also independent and expected to produce local bulk movement patterns $\hat{S}^{\psi_j}$ which are isotropic, homogeneous and symmetric. A difference between noise and isotropic simulations is the movement of objects between subregions, which introduce a within-sample dependence between local bulk movement patterns. The results in Table 4.1 indicate that there is not enough evidence to reject within-sample independence, as the proportion of between-sample p-values less than 0.05 remains approximately 0.05 and the single rejection of the omnibus null hypothesis at the five percent level occurs with p-value 0.024. We therefore conclude that within simulations local bulk movement patterns are approximately independent, an important result for the interpretation of future test results and one that has been shown to hold across all considered subregion sizes.

Homogeneous simulations are independent, producing local bulk movement patterns which are homogeneous but neither isotropic nor symmetric. The result of this is rejection of null hypotheses H_0^I and H_0^S as local bulk movement patterns $\hat{S}^{\psi_j}$ are not identically distributed under the action of operators $\lambda \in \Lambda^I$ and Λ^S. Under H_0^I and H_0^S a generally greater proportion of between-sample p-values are less than 0.05 for larger subregion sizes, potentially because movement patterns $\hat{F}^{\psi_j}$ are more accurately estimated for larger subregions. Homogeneous simulations are particularly susceptible to inaccuracy in the estimation of $\hat{F}^{\psi_j}$ for small subregion sizes as an object exiting the northeast of the subregion may be replaced by an object entering the southwest of the subregion, resulting in an estimated movement in the opposite direction to that which actually occurred. An increase in significant between-sample p-values for increasing subregion size is counterintuitive in terms of rejection of the null hypothesis due to within-sample dependence, as smaller subregions are expected to increase the proportion of objects crossing boundaries, increasing the within-sample dependence. Homogeneous simulations typically fail to reject the omnibus null hypothesis under H_0^H at the five percent level as expected, with the one exception doing so with p-value 0.027.

Symmetric simulations produce similarly expected results, consistently rejecting H_0^I and H_0^H as local bulk movement patterns $\hat{S}^{\psi_j}$ are not identically distributed under the action of $\lambda \in \Lambda^I$ and Λ^H. When testing against H_0^S, failure to reject the omnibus null hypothesis at the five percent level is also commonly seen, with the two exceptions rejecting with p-values 0.032 and 0.022.

The results of testing independent simulations across a number of hypotheses indicate that if a null hypothesis is chosen which is suitable for the movement patterns of the process, the testing procedure generally returns between-sample p-values whose distribution is indistinguishable from Uniform[0,1]. That is to say, local bulk movement patterns are between-sample independent (by construction), equal in distribution under the action of $\lambda \in \Lambda$ for an appropriate specification of Λ (also by construction) and within-sample independent. This suggests that the proposed testing procedure is valid under considered specification of the null hypothesis, following which rejection of future tests may be taken as evidence against between-sample independence of local bulk movement patterns.

4.3.2.3 Dependent simulations

The power of the proposed testing procedure is assessed by testing dependent simulations under the most appropriate choice of null hypothesis. That is, isotropic 10 and isotropic 30 simulations under H_0^I, homogeneous 10 and homogeneous 30 simulations under H_0^H and symmetric 10 and symmetric 30 simulations under H_0^S, where each test is an approximate permutation test using 10 000 random samples from the corresponding set of operators Λ. Table 4.2 summarises the results of these tests, presenting the proportion of the 105 between-sample p-values which are less than 0.05 and marking with ‡ those which fail to reject the omnibus null hypothesis of Uniform[0,1] distribution of between-sample p-values under the Kolmogorov-Smirnov test at the five percent level. Tests are again carried out over four pairs of consecutive time points and four subregion sizes varying between 10 and 20.

Over all comparisons the testing procedure is generally able to detect dependence when it exists, observed as generally consistent rejection of the omnibus null hypothesis at the five percent level. This is always the case for the simulations sharing 30 out of the 100 total objects, but there are some failures to reject the omnibus null hypothesis at the five percent level when the degree of dependence is weaker and only 10 objects are shared. For isotropic 10 simulations the failure to reject the omnibus null hypothesis occurs with a p-value of 0.739. For homogeneous 10 simulations there are more cases, failing to reject the omnibus null hypothesis with p-values 0.201, 0.142, 0.145 and 0.439. For symmetric 10 simulations we again see a few failures to reject the omnibus null hypothesis, with p-values 0.935, 0.627, 0.081, 0.273 and 0.116.

Although the omnibus null hypothesis is typically rejected at the five percent level, it is by no means the case that every between-sample test rejects the respective

		H_0^I				H_0^H				H_0^S			
		20	15	12	10	20	15	12	10	20	15	12	10
noise	1 → 2	0.03	0.06	0.06	0.04†	0.02	0.08	0.06	0.02†	0.03	0.06	0.05	0.05
	2 → 3	0.06	0.03	0.08	0.06	0.06	0.02	0.06	0.06	0.07	0.02	0.06	0.03
	3 → 4	0.04	0.04	0.04	0.10	0.06	0.04	0.05	0.09	0.05	0.07	0.04	0.05
	4 → 5	0.04	0.04	0.03	0.05	0.04	0.04	0.04	0.04	0.02	0.02	0.04	0.04
isotropic	1 → 2	0.06	0.10	0.07	0.03	0.08	0.09	0.07	0.03	0.07	0.04	0.03	0.03
	2 → 3	0.05	0.04	0.03	0.07	0.05	0.01	0.04	0.06	0.08	0.02	0.04	0.06
	3 → 4	0.06	0.07	0.10	0.08	0.05	0.06	0.08†	0.06	0.06	0.04	0.08	0.05
	4 → 5	0.05	0.03	0.06	0.05	0.04	0.03	0.06	0.05	0.05	0.02	0.04	0.05
homogeneous	1 → 2	0.88†	0.73†	0.62†	0.44†	0.05	0.02	0.05	0.06	0.90†	0.71†	0.67†	0.52†
	2 → 3	0.53†	0.40†	0.39†	0.41†	0.04	0.03	0.09	0.05	0.6†	0.40†	0.36†	0.39†
	3 → 4	0.80†	0.70†	0.56†	0.55†	0.03	0.06	0.03	0.03	0.79†	0.70†	0.65†	0.60†
	4 → 5	0.70†	0.57†	0.57†	0.33†	0.04†	0.03	0.05	0.03	0.70†	0.57†	0.54†	0.33†
symmetric	1 → 2	0.00†	1.00†	0.98†	0.94†	1.00†	1.00†	0.98†	0.94†	0.02	0.08	0.07	0.06
	2 → 3	1.00†	1.00†	0.91†	0.98†	1.00†	1.00†	0.93†	0.98†	0.08	0.07†	0.07	0.05
	3 → 4	0.99†	0.96†	0.99†	1.00†	1.00†	0.96†	0.99†	1.00†	0.08	0.03	0.06	0.06
	4 → 5	0.98†	1.00†	1.00†	1.00†	1.00†	1.00†	1.00†	1.00†	0.09	0.07	0.10	0.05†

Table 4.1: The results of testing each class of independent spatio-temporal simulations under each of the three hypotheses H_0^I, H_0^H and H_0^S for a range of consecutive time points and subregion sizes. Comparison between the 15 simulations in each class produces 105 between-sample p-values for each test, with the table presenting the proportion of p-values less than 0.05 and † used to denote sets of p-values which reject the omnibus hypothesis of Uniform[0,1] distribution of between-sample p-values under the Kolmogorov-Smirnov test at the five percent level.

null hypothesis at the five percent level, with the proportions of tests that do so varying between 0.04 and 0.90. This is to be expected due to the difficulty of the problem and the general solution that we propose, requiring only minor assumptions on movement patterns to determine exchangeability of local bulk movement patterns. Despite this, when the optimal choice of subregion size is made and the dependence is 30 percent of all objects the proportion of between-sample tests reported significant at the five percent level is on average 0.77, which we take as an indicator that the proposed testing procedure has an acceptable level of power.

There is almost universal improvement in detection ability, measured by an increase in the proportion of between-sample p-values which are less than 0.05, with decreasing subregion size. As subregion size reduces, the expected number of independent simulated objects in each subregion is reduced. The relative contribution to $\hat{S}^{\psi_j}$ by shared simulated objects in the subregions in which they are located is therefore increased, making their existence more easily detectable under the testing procedure. Despite this, there is expected to be a limit below which further reduction of subregion sizes will result in a reduction in power, as the accuracy of the estimator $\hat{F}^{\psi_j}$ is reduced as the probability of objects moving between subregions increases.

In general the proportion of between-sample p-values less than 0.05 is greater for isotropic simulations than for both homogeneous and symmetric simulations, for which the proportions are broadly comparable. A possible explanation for this may be that for homogeneous and symmetric simulations movements are determined solely by the location of objects, with closely separated objects undergoing very similar movements. This could result in greater numbers of coincidentally similar movements, making dependencies more difficult to identify for homogeneous and symmetric simulations in comparison to isotropic simulations for which movement direction is independent of object location and coincident similarities are as a result rarer.

There is expected to be some dependence between tests at different pairs of consecutive time points, but the form of this dependence is difficult to quantify. For isotropic and homogeneous simulations local bulk movement patterns are expected to be identically distributed in time and observed results are similarly consistent over time. This is not the case for symmetric simulations, as objects aggregate at the centre of Ψ over time. Despite this, there is no discernible difference in results for symmetric simulations across different pairs of consecutive time points, suggesting that whether movements are spread over Ψ or congregated closer to its centre has no impact on the performance of the testing procedure.

The power of the proposed testing procedure is further assessed by testing de-

		20	15	12	10
isotropic 10	$1 \to 2$	0.19	0.11	0.23	0.26
	$2 \to 3$	0.10	0.12	0.10	0.17
	$3 \to 4$	$0.07^{\ddagger}$	0.20	0.10	0.23
	$4 \to 5$	0.10	0.11	0.18	0.23
isotropic 30	$1 \to 2$	0.37	0.47	0.84	0.90
	$2 \to 3$	0.47	0.40	0.83	0.88
	$3 \to 4$	0.49	0.51	0.93	0.90
	$4 \to 5$	0.57	0.33	0.84	0.88
homogeneous 10	$1 \to 2$	$0.03^{\ddagger}$	$0.12^{\ddagger}$	0.13	0.13
	$2 \to 3$	0.15	0.16	0.22	0.30
	$3 \to 4$	$0.05^{\ddagger}$	0.06	$0.10^{\ddagger}$	0.13
	$4 \to 5$	0.10	0.11	0.09	0.21
homogeneous 30	$1 \to 2$	0.37	0.40	0.62	0.80
	$2 \to 3$	0.24	0.63	0.50	0.90
	$3 \to 4$	0.16	0.45	0.50	0.73
	$4 \to 5$	0.20	0.51	0.67	0.69
symmetric 10	$1 \to 2$	$0.05^{\ddagger}$	0.11	0.09	0.16
	$2 \to 3$	$0.07^{\ddagger}$	0.10	0.16	0.17
	$3 \to 4$	$0.06^{\ddagger}$	0.10	0.10	0.09
	$4 \to 5$	$0.04^{\ddagger}$	0.12	$0.10^{\ddagger}$	0.06
symmetric 30	$1 \to 2$	0.21	0.39	0.53	0.60
	$2 \to 3$	0.24	0.48	0.37	0.53
	$3 \to 4$	0.19	0.26	0.77	0.70
	$4 \to 5$	0.27	0.30	0.40	0.69

Table 4.2: The results of testing each class of dependent simulations against the most appropriate hypothesis for a range of consecutive time points, subregion sizes and degrees of dependence. Comparison between the 15 simulations in each class produces 105 between-sample p-values for each test, with the table presenting the proportion of p-values less than 0.05 and ‡ used to denote sets of p-values which fail to reject the omnibus null hypothesis of Uniform[0,1] distribution of between-sample p-values under the Kolmogorov-Smirnov test at the five percent level.

pendent isotropic simulations under the three null hypotheses H_0^I, H_0^H and H_0^S. Rejection of any of these hypotheses may be taken as evidence of between-sample dependence because local bulk movement patterns for isotropic simulations are exchangeable under operators $\lambda \in \Lambda^I$ and both sets of operators $\Lambda^H \subset \Lambda^I$ and $\Lambda^S \subset \Lambda^I$ by construction. Table 4.3 summarises the results of these tests, presenting the proportion of the 105 between-sample p-values which are less than 0.05 and marking with ‡ those which fail to reject the omnibus null hypothesis of Uniform[0,1] distribution of between-sample p-values under the Kolmogorov-Smirnov test at the five percent level. These test are again carried out over the four pairs of consecutive time points and four subregion sizes varying between 10 and 20.

Across all null hypotheses there is largely consistent rejection of the omnibus null hypothesis at the five percent level. The exceptions to this fail to reject the omnibus null hypothesis with p-values of 0.739 for H_0^I, 0.824 for H_0^H and 0.624 for H_0^S. The proportion of between-sample tests reporting significance at the five percent significance level is greatest under H_0^I, but only mildly greater than under H_0^H which is in turn mildly greater than under H_0^S. These results support the effectiveness of the testing procedure using any suitably large valid set of operations Λ under which local bulk movement patterns are exchangeable, with a minor reduction in power in comparison to the test carried out under the maximal set of operations Λ under which local bulk movement patterns are exchangeable.

4.3.2.4 Multiple testing

Both the testing of simulations in the previous section and the testing of real data in the following section raise questions of multiple comparisons, previously introduced in more detail in Section 2.3.4. Focusing on a single example, the results presented in Table 4.1 for the testing of noise simulations under H_0^I are based upon tests of 105 between-sample comparisons across four subregion sizes and four pairs of consecutive time points, for a total of 1 680 statistical tests. Considering the testing of the 16 omnibus null hypotheses via the Kolmogorov-Smirnov test at the five percent level, we observe one rejection of the null hypothesis. However, this individual result must be considered in the wider picture of all sixteen omnibus tests.

In all tests of the omnibus null hypothesis we expect strong positive correlation between test results, the exact form of which is difficult to quantify. This is because tests across different subregion sizes are applied to the same simulation data each time and for all but one class of simulations data is identically distributed across all time points, indicating that tests between $s = 1, t = 2$ and $s = 2, t = 3$ are expected to produce very similar results. The Bonferroni correction and alternative corrections are therefore not applied, with tests carried out at the nominal five

		isotropic				homogeneous				symmetric			
		20	15	12	10	20	15	12	10	20	15	12	10
isotropic 10	1 → 2	0.19	0.11	0.23	0.26	0.20	0.11	0.24	0.26	0.17	0.15	0.18	0.22
	2 → 3	0.10	0.12	0.10	0.17	0.13	0.14	0.10	0.16	0.08	0.11	0.06	0.15
	3 → 4	$0.07^{\ddagger}$	0.20	0.10	0.23	$0.06^{\ddagger}$	0.19	0.10	0.22	$0.05^{\ddagger}$	0.20	0.09	0.19
	4 → 5	0.10	0.11	0.18	0.23	0.10	0.13	0.22	0.26	0.10	0.10	0.15	0.17
isotropic 30	1 → 2	0.37	0.47	0.84	0.90	0.27	0.48	0.81	0.89	0.30	0.46	0.76	0.90
	2 → 3	0.47	0.40	0.83	0.88	0.47	0.43	0.84	0.89	0.41	0.39	0.83	0.76
	3 → 4	0.49	0.51	0.93	0.90	0.30	0.54	0.95	0.84	0.44	0.50	0.75	0.79
	4 → 5	0.57	0.33	0.84	0.88	0.41	0.30	0.83	0.79	0.56	0.32	0.73	0.78

Table 4.3: The results of testing dependent isotropic spatio-temporal simulations against the three appropriate hypotheses for a range of consecutive time points, subregion sizes and degrees of dependence. Comparison between the 15 simulations in each class produces 105 p-values for each test, with the table presenting the proportion of p-values less than 0.05 and ‡ used to denote sets of p-values which fail to reject the omnibus null hypothesis of Uniform[0,1] distribution under the Kolmogorov-Smirnov test at the five percent level.

percent level but resulting conclusions made in light of the presence of multiple testing.

4.4 Investigation of EB3 and TACC3 data

4.4.1 Biological background

As stated in the introduction to this chapter, we aim to make inference on dependence between the local bulk movement patterns of biomolecular species TACC3 and EB3. The protein EB3 is known to localise at the tip of growing microtubules during mitosis (Mimori-Kiyosue et al., 2000) and the biological question of interest is whether TACC3 is similarly located, with the results potentially shedding light on its impact on the process of mitosis.

The available data is comprised of confocal fluorescence microscopy images collected across seven samples at a total number of between 47 and 57 time points. Images are collected of live cells during mitosis with TACC3 tagged with a green fluorescing protein and EB3 tagged with a red fluorescing protein. Microscope resolution is such that each pixel is 68.9nm square and images are collected at a rate of one per second. Green and red intensities are recorded at the same time by two different digital cameras, resulting in two greyscale images for each sample at each time point, one corresponding to EB3 and one to TACC3. Digital cameras are used for image acquisition, meaning that the data contained in each image is accessible as a matrix of integer valued intensity levels, one for each pixel location. We treat intensity levels as a surrogate measure for quantity of the relevant biomolecular species located at the centre of each pixel. Example images from five consecutive time points for three of the samples may be seen in Figure 4.5, illustrating the structure seen, variability within samples between time points and between samples more generally.

The proposed testing methodology requires an assumption under which the subregions ψ^j may be rearranged and remain identically distributed. As EB3 is located at the end of microtubules which grow in a spindle structure during mitosis, we make the assumption that movement patterns are symmetric across the line connecting microtubule organising centres, the poles of the approximately ellipsoid spindle structure. The mitotic spindle is a three dimensional structure, with the resulting two dimensional images a projection into a single plane. Original image sequences are therefore rotated and cropped to focus on only the spindle region, with the imagined line connecting what is estimated to be the spindle poles vertical. The result is image sizes of 180 pixels wide and 240 pixels high across all samples, permitting exact tessellation by square subregions ψ^j of side lengths 20, 15 and 12.

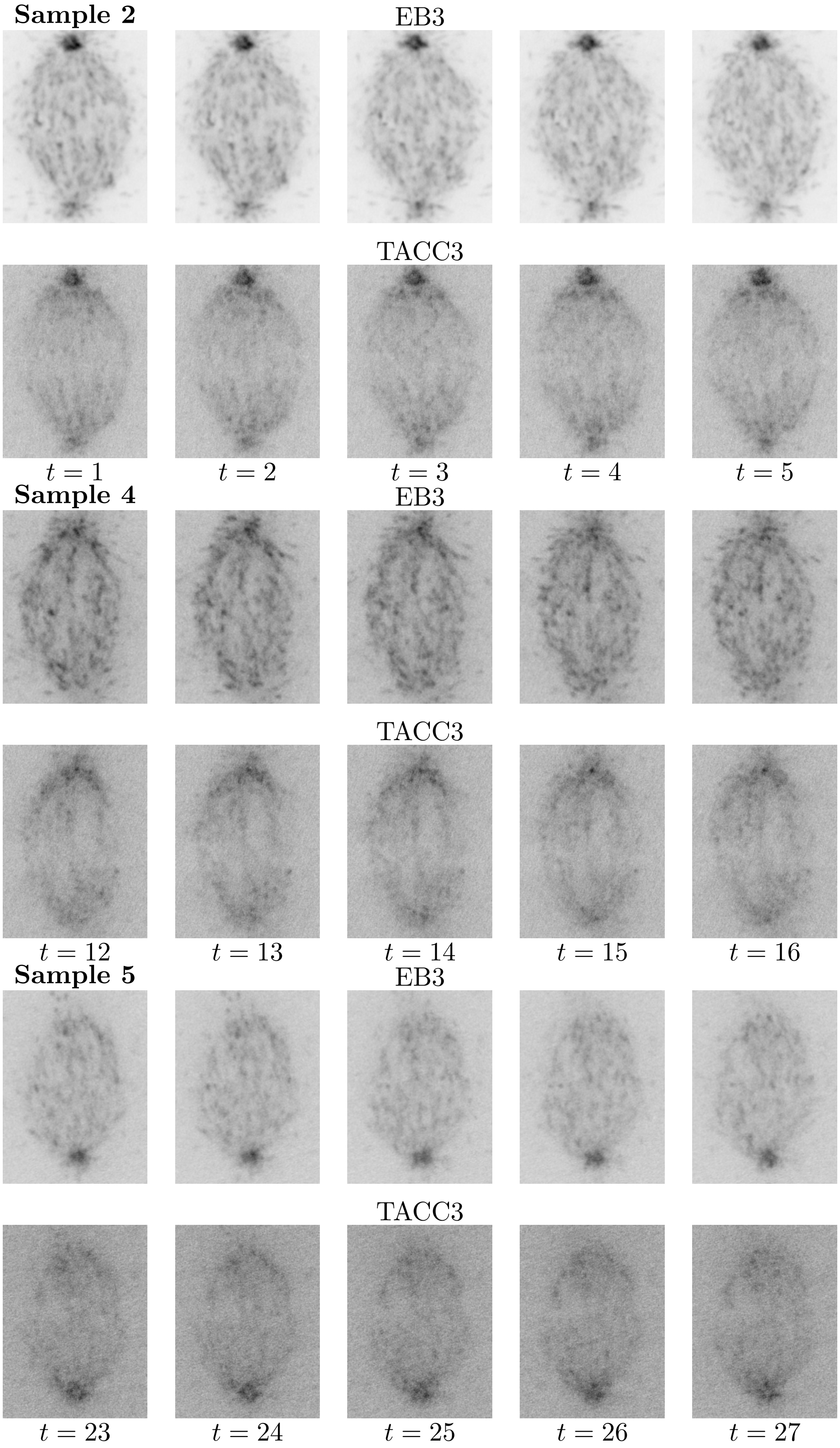


Figure 4.5: EB3 and TACC3 biological images from three samples across five consecutive time points. Pixel intensities correspond to values of m with the maximum value of m across all time points black and $m = 0$ white.

Collection of images every second is expected to capture the location of biomolecules at a great enough time resolution that their movement patterns may be estimated. In investigation of the same data (Gutierrez-Caballero et al., 2015) biomolecule clusters were located and tracked using an automatic object tracking process requiring specification of a number of parameters tuned based upon the observed images and prior beliefs about the biology underpinning the process (Applegate et al., 2011). Results indicated that both biomolecular species are located at the end of growing microtubules, separated by an average distance of 229nm, with TACC3 closest to the growing tip. We propose to investigate the same data set without specification of parameters, other than the assumption of reflection symmetry across the imagined vertical line connecting the spindle poles which divides images in half, and further statistically test for dependence between TACC3 and EB3 bulk movement patterns.

4.4.2 Exploratory data analysis

Intensities in images of EB3 are in general greater than those for TACC3 due to expression of TACC3 at a lower level. Greater expression of TACC3 is avoided, as it results in aggregation of TACC3 away from microtubule tips (Gutierrez-Caballero et al., 2015), and consequently away from EB3, which could obscure the ability to investigate dependence between TACC3 and EB3 movements. Estimation of movement patterns for TACC3 is therefore expected to be more challenging as intensities are in some cases on a similar scale to background noise.

Scaling of intensity in a subregion, $m(\psi)$, by a positive constant across time points s and t consistent with changing the brightness of images results in an identical scaling of the summary statistic, $\hat{S}^{\psi}_{s,t}$, but a potentially non-linear change to the comparison score $\hat{G}^{\psi,\psi}_{s,t}$ between observations. However, if the scaling is consistent across all subregions and all time points within an observation then the impact of differences in brightness on test results should be minimal.

As an exploratory investigation, colocalisation between image pairs may be quantified using Pearson's correlation coefficient, the results of which are displayed in Figure 4.6. Calculations are made for comparison pairs EB3 × TACC3, for which we are interested in the degree of similarity, and EB3 × TACC3* and TACC3 × TACC3*, where TACC3* is the vertical reflection of TACC3 for which we expect to see only coincidental similarity. For each of the seven samples correlation values are greater for EB3 × TACC3 than the alternatives, indicating that there is more than coincidental similarity between EB3 and TACC3 localisation. However, without a methodology for quantifying the significance of obtained correlation values it is impossible to conclude that there is dependence between the distribution of EB3 and

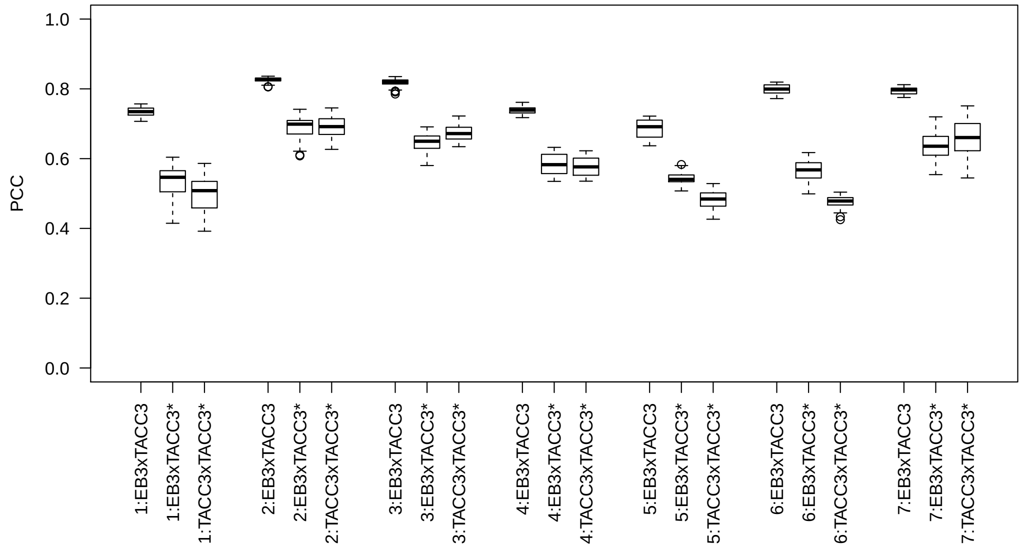


Figure 4.6: Boxplots to display the distribution of values of Pearson's correlation coefficient between pixel intensity values across all pairs of consecutive time points for each of the biological sample comparisons considered.

TACC3.

A major foundation of the estimator $\hat{F}^{\psi}_{s,t}$ is that $m_s(\psi)$ and $m_t(\psi)$ represent distributions of approximately the same masses at two different time points. In the context of the EB3 and TACC3 image data this equates to minimal changes in intensity on a subregion by subregion basis, caused by biomolecules moving between subregions, and on a whole image basis, caused by photobleaching over time. To investigate the validity of these assumptions we plot changes in pixel average intensity between consecutive time points

$$\frac{1}{|\psi|}\sum_{x\in\psi} m_{t+1}(x) - m_t(x),$$

for subregions, ψ_j, of sizes 20, 15 and 12 and for the whole image space, Ψ, the results of which may be seen for sample four in Figure 4.7.

Average intensity plots indicate that there is a consistent but minor reduction in intensity across the whole space between consecutive time points, consistent with photobleaching, but not believed to be significant enough to render the estimation of $\hat{F}$ as unreliable. There are sometimes large changes in subregion intensity between consecutive time points, corresponding to particularly large concentrations of biomolecules moving between subregions, but the vast majority of changes on an individual pixel scale are small in comparison to average pixel intensity, indicating that movements are largely contained within subregions. Predictably, larger

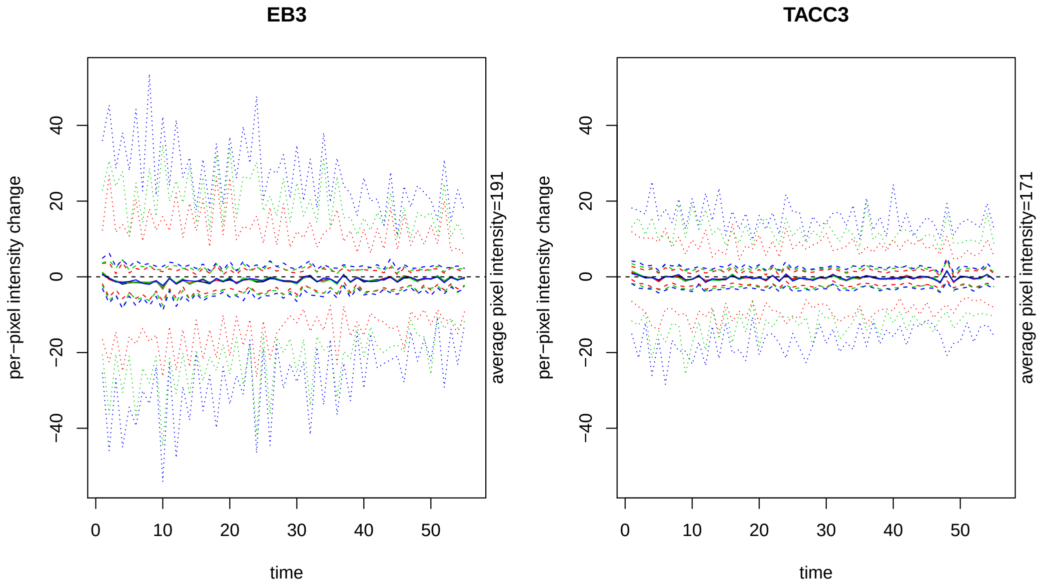


Figure 4.7: Variation in subregion total pixel intensity over time on a per-pixel scale for EB3, left, and TACC3, right. Solid black line denotes the mean intensity change over Ψ, with solid red, green and blue lines corresponding to medians over subregions ψ_j of size 20, 15 and 12 respectively. Coloured dashed lines denote upper and lower quartiles over the collection of subregions ψ_j, with dotted lines denoting maximum and minimum differences. Horizontal black dashed line denotes no change in average intensity.

relative intensity changes are observed for smaller subregion sizes as the proportion of locations $x \in \psi_j$ which are close to the subregion boundary increases, meaning biomolecules are more likely to move between subregions.

4.4.3 Permutation testing results

As described in the previous section, we make three comparisons between EB3, TACC3 and TACC3*, the vertical reflection of TACC3 across the horizontal line through the centre of Ψ. Testing is carried out as an approximate permutation test of 10 000 random permutations under H_0^R, which specifies on a scale determined by the number of subregions, w, between-sample independence of local bulk movement patterns, identical distribution of the collection of $S_{s,t}^{\psi_j}$ under the action of any $\lambda \in \Lambda^R$ and within-sample independence of local bulk movement patterns.

Test results are presented in Table 4.4 as the proportion of between sample comparisons across all consecutive time points reporting p-values of less than 0.05, with † used to denote those collections of p-values which reject the omnibus null hypothesis of Uniform[0,1] distribution of between-sample comparisons under the Kolmogorov-Smirnov test at the five percent level. Note the difference between presentation of EB3 and TACC3 results in comparison to simulation results — here collections of p-values tested against the omnibus hypothesis are combined across the between 46 and 56 pairs of consecutive time points, while for simulation data they were collected across the 105 pairs of between-sample comparisons at a single pair of consecutive time points.

Rejection of the null hypothesis is expected to occur if any of the three components of H_0^R are not met, however, the biological question of interest relates only to the first condition of between-sample independence of local bulk movement patterns. It is for this reason that the comparison between EB3 and TACC3* is made, for which between-sample independence is expected by construction and rejection of H_0^R may be attributed to improper specification of Λ^R or within-sample dependence of local bulk movement patterns. The results indicate that the omnibus null hypothesis is not rejected at the five percent level for comparison between EB3 and TACC3* for any sample, suggesting that there is insufficient evidence to refute the assumption of horizontally symmetric bulk movement patterns and within-sample independence. The omnibus null hypothesis is rejected in some instances of comparison between TACC3 and TACC3* but these may be attributed to between-sample dependencies, in particular for subregions along the vertical centre of Ψ which may be compared to reflections of themselves under $\lambda \in \Lambda^R$.

For each of the seven samples the omnibus null hypothesis comparing EB3 and TACC3 is rejected at the five percent level, taken as strong evidence of between-sample dependence in local bulk movement patterns between EB3 and TACC3. The proportion of between-sample tests reporting p-values less than 0.05 varies between 0.36 and 1.00, rising to between 0.58 and 1.00 when subregions of the smallest size, 12, are considered. This is evidence of regular rejection of H_0^R, improving with reduction in subregion size but not detectable at only a single subregion size which must be accurately specified. The effect of varying subregion sizes is similar to that observed for simulated data sets.

Investigation of the distribution of between-sample p-values across the range of consecutive time points shows that non-significant values are interspersed within significant values at the five percent level. This suggests that rather than periods of dependence and periods of independence of local bulk movement patterns, between some pairs of time points movement estimation may be particularly inaccurate as objects of considerable intensity move between subregions, resulting in an insignificant test result.

		20	15	12
Sample 1	EB3 × TACC3	$0.48^{\dagger}$	$0.55^{\dagger}$	$0.68^{\dagger}$
	EB3 × TACC3*	0.05	0.07	0.04
	TACC3 × TACC3*	0.09	0.14	$0.12^{\dagger}$
Sample 2	EB3 × TACC3	$0.85^{\dagger}$	$0.98^{\dagger}$	$1.00^{\dagger}$
	EB3 × TACC3*	0.11	0.07	0.07
	TACC3 × TACC3*	0.11	0.20	$0.22^{\dagger}$
Sample 3	EB3 × TACC3	$0.53^{\dagger}$	$0.60^{\dagger}$	$0.69^{\dagger}$
	EB3 × TACC3*	0.02	0.05	0.04
	TACC3 × TACC3*	0.18	$0.25^{\dagger}$	$0.25^{\dagger}$
Sample 4	EB3 × TACC3	$0.36^{\dagger}$	$0.67^{\dagger}$	$0.71^{\dagger}$
	EB3 × TACC3*	0.05	0.07	0.07
	TACC3 × TACC3*	0.11	0.09	$0.22^{\dagger}$
Sample 5	EB3 × TACC3	$0.47^{\dagger}$	$0.51^{\dagger}$	$0.58^{\dagger}$
	EB3 × TACC3*	0.02	0.07	0.05
	TACC3 × TACC3*	$0.11^{\dagger}$	0.13	0.11
Sample 6	EB3 × TACC3	$0.78^{\dagger}$	$0.85^{\dagger}$	$0.93^{\dagger}$
	EB3 × TACC3*	0.05	0.07	0.07
	TACC3 × TACC3*	0.09	0.18	0.16
Sample 7	EB3 × TACC3	$0.47^{\dagger}$	$0.53^{\dagger}$	$0.67^{\dagger}$
	EB3 × TACC3*	0.00	0.05	0.11
	TACC3 × TACC3*	$0.09^{\dagger}$	0.11	$0.09^{\dagger}$

Table 4.4: The results of testing for dependency between spatio-temporal biological data for seven samples, each of which comprises two sets of images representing the locations of EB3 and TACC3 over between 47 and 57 time points and a resulting third set, TACC3*, which is the vertical reflection of the TACC3 data. Table values are the proportion of between-sample p-values under H_0^R which are significant at the five percent level for subregion sizes varying between 20 and 10. † is used to indicate the collections of between-sample p-values which reject the omnibus null hypothesis of Uniform[0,1] distribution at the five percent level according to a Kolmogorov-Smirnov test. Testing is carried out using approximate permutation tests with 10 000 permutations.

4.5 Conclusions

4.5.1 Statistical methodology

We have proposed a method for the estimation of local bulk movement patterns within two samples and a testing procedure for quantifying the significance of the dependence between these patterns. The procedure relies upon a minimal set of assumptions, namely consistency of total mass within the whole space Ψ, the ability to define identically sized and shaped subregions $\psi_j \subseteq \Psi$ for which movements of mass between subregions is minimal in comparison to movements within subregions and symmetry or otherwise of subregion movement patterns such that they may be transformed and rearranged and remain identically distributed. These assumptions are generally easier to satisfy for observations derived from symmetric physical phenomena where observations have been collected at a sufficiently high time resolution.

Application of the proposed methodology to simulated data for which movement patterns are independent produces results which reject correctly specified null hypotheses at a rate consistent with that of the size of the test, supporting the theoretical validity of the testing procedures. In cases of incorrectly specified null hypotheses, tests may reject the null hypothesis when movement patterns are independent as the null hypothesis is a composite of three statements and failure to satisfy any one is sufficient for rejection of the null hypothesis.

Application of the proposed methodology to simulated data where movement patterns are partially dependent often results in rejection of correctly specified null hypotheses, but due to the complexities of the problem and the minimal set of assumptions power may be limited. In the case of 10 percent dependence between movement patterns, correctly specified null hypotheses are rejected at the five percent level typically between 10 and 30 percent of the time. In the case of 30 percent dependence between movement patterns, correctly specified null hypotheses are rejected at the five percent level between 20 and 90 percent of the time. In almost all cases, the collection of between-sample p-values rejects the omnibus null hypothesis of Uniform[0,1] distribution according to the Kolmogorov-Smirnov test at the five percent level, indicating that the omnibus hypothesis may be useful to detect dependence when comparing over multiple samples or time points.

The statistical testing procedure compares local bulk movement patterns, where local is on a scale defined by the size of subregions. In application to simulated data the test is more powerful for smaller subregion sizes, indicating that dependencies between movement patterns may be insignificant when summaries are made over larger areas.

4.5.2 Biological conclusions

Analysis of a biological data set to compare movement patterns of EB3 and TACC3 biomolecule species during mitosis reports consistent rejection of the omnibus null hypothesis at the five percent level. Further, there is a general failure to reject the omnibus hypothesis when EB3 observations are compared to a transformation of TACC3 data by vertical reflection. Taken in combination, these results suggest that null hypotheses are in this instance rejected on the basis of dependence between movement patterns rather than due to a lack of exchangeability within each observation under the specified set of transformations and rearrangement of subregions. As in the case of simulated data, an increase in the proportion of significant between-sample p-values is seen as subregion size decreases. Our analysis therefore supports the previous work by Gutierrez-Caballero et al. (2015) that the movement patterns of EB3 and TACC3 are dependent, potentially through their localisation on the tips of growing microtubules.

In cases where the degree of dependence between two variables is of interest there exist measures of effect size for Pearson's correlation coefficient (Cohen, 1992). The analysis in this Chapter endeavours to ascertain the existence of dependence between two spatio-temporal processes, for application to images of biomolecule distributions. However, due to the novel approach used to quantify dependence more work would need to be carried out to determine the magnitude of the effect size and its variability. We therefore rely solely upon the results of the statistical testing procedure which we have proposed.

Chapter 5

Mitotic spindle modelling and comparison

This chapter introduces a preliminary model of the structure of the mitotic spindle on a macro, whole-spindle scale alongside a score to compare model fit between models and/or observations. The material forms part of a paper published in the Journal of Cell Science (Nixon et al., 2017) written alongside our collaborator Dr. Stephen Royle and his research group. The R code required to implement the proposed methods has been made available online at `www.github.com/quantixed/VolumeFinder` alongside notes for its use.

An introduction is first given to highlight specific relevant biological and imaging background information and describe mathematically the form of the data. Following this the model is presented and a comparison score described in Section 5.2. Section 5.3 presents an exploratory analysis of samples of observed spindle data and comparisons between model fit for each of the samples. A summary of conclusions and directions for further investigation is then presented in Section 5.4 to end the chapter.

5.1 Introduction

The collection of microtubules which bind the centrosomes in each half of a cell undergoing mitosis to the chromosome pairs is known as the mitotic spindle. The purpose of these microtubules is to apply the force necessary to separate the chromosome pairs, with accurate separation required to ensure the health of the daughter cells (Holland and Cleveland, 2009). To quantify the ability of the spindle to perform this task accurately we investigate the degree to which observed microtubules deviate from a proposed model of the spindle.

Information on microtubules within the spindle is available via an imaging technique known as Serial Block Face Scanning Electron Microscopy (Nixon et al., 2017). This procedure images the sample in sections, imaging the surface of the sample via scanning electron microscopy before slicing off the top layers and imaging the newly revealed surface. Repeated imaging and slicing results in a collection of images recorded in parallel planes at a sequence of depths through the sample. Microtubules may be determined from within images using automated image analysis techniques, the result of which is a collection of straight lines.

Imaging data at the metaphase state of mitosis has been provided by our collaborator Dr. Stephen Royle for a total of 11 samples across six different experimental conditions. Control samples are unmodified and express normal levels of TACC3. Mutant samples are engineered to express S558A, a TACC3 mutant which is known to be unable to localise to the mitotic spindle. The final set of samples are engineered to overexpress TACC3. Samples of each type are assessed after being grown at 37 degrees Celsius, warm conditions, or grown at 37 degrees Celsius and then cooled to zero degrees Celsius for 10 minutes before fixing and imaging, cold conditions. A small number of replicated experiments results in 11 data sets across the six different sample and condition combinations.

The aim of our work is to present a model which describes the organisation of microtubules within the mitotic spindle using the available samples and biological knowledge. Model fit for different observations may then be compared to determine conditions under which the mitotic spindle structure deviates from our expectation, the result of which may be reduced performance during mitosis and an increased possibility of errors. Due to the limited number of samples and the inherent complexity of the structure we limit our investigation to an exploratory analysis, presenting a simple but effective model and a comparison score by which model fit may be compared.

5.1.1 Mathematical representation of data

For a given sample observed straight lines may be expressed parametrically as $L_i(t) \in \mathbb{R}^3$ for the parameter $t \in [0,1]$ and index $i \in \{1, 2, \ldots, n\}$. Under this formulation the direction of lines may be associated with an increasing value of t as $L_i(1) - L_i(0)$, however, they may be also associated with a decreasing value of t as $L_i(0) - L_i(1) = -(L_i(1) - L_i(0))$. In Section 5.2.2 we describe how each line may be associated with a unique direction.

Alongside lines L_i, observations are provided with the location of two fixed points, $p_1, p_2 \in \mathbb{R}^3$. These fixed points represent the location of centrosomes within the cell, the two fixed points from which microtubules are understood to originate. When not observable directly from images the locations of p_1 and p_2 are estimated by the biologist conducting the experiments based upon the observed structure of lines. Points p_1 and p_2 are referred to in future discussion as the poles of the spindle, with the line connecting them referred to as the axis of the spindle.

The density of observed lines varies in space for each sample and in some cases observations do not contain the entire spindle structure. As a result, we limit our model to specification of the expected orientation of lines within the sample at any location $v \in \mathbb{R}^3$. Following specification of such a model, deviation between observations and the model may be summarised by the angles between observed and model directions, collated to produce the set $\Theta \in [0, \pi]^n$ from which model fit is quantified.

Automatic detection of microtubules from images is by no means a perfect procedure, with possibilities of failure to detect microtubules, reporting of microtubules which do not exist and errors in the specification of lines. We therefore wish to propose a statistic Π comparing Θ_1 and Θ_2 arising from different models and/or samples which is robust to such potential errors.

5.2 Model formulation and comparison of model fit

5.2.1 Spheroid model

We have a collection of assumptions based upon the purpose of the biological spindle which guide our creation of a model. Firstly, lines are believed to represent sections of trajectories which originate at the poles, p_1 and p_2, and travel towards the equatorial plane which is perpendicular to and bisects the spindle axis connecting the poles (Jaqaman et al., 2010). This is because microtubules originate at centrosomes and travel towards chromosomes aligned approximately equidistant from the centrosomes.

Secondly, trajectories are expected to vary continuously in space. That is, the angle $\theta(\rho(v_1), \rho(v_2))$ between proposed trajectories $\rho(v_1), \rho(v_2)$ at locations v_1 and v_2 is expected to be small when the distance between v_1 and v_2, $||v_1 - v_2||_2$, is small. This is under the assumption that there is a consistent ordering of microtubules, that they are locally similarly aligned and do not cross each other at large angles.

Finally, we expect reflection symmetry across the equatorial plane under the assumption of a force balance between the microtubules in each hemisphere and rotational symmetry about the spindle axis, as a parsimonious assumption in the absence of further information.

A model which satisfies each of these constraints is based upon spheroids whose surface is described by the vectors v satisfying

$$\frac{v_x^2}{a^2} + \frac{v_y^2}{a^2} + \frac{v_z^2}{b^2} = 1,$$

for varying values of a, where the poles are located at $p_1 = (0,0,b)$ and $p_2 = (0,0,-b)$. Model trajectories $\rho(v)$ are then tangents to geodesic paths on the surface of the spheroid which pass through v and and both poles. The derivation of these model directions is provided in Sections 5.2.3 and 5.2.4.

5.2.2 Data alignment

Lines L_i are initially reported in the coordinate system of the imaging procedure, for which they lie in planes perpendicular to the z-direction at heights $z \in \{\zeta_1, \zeta_2, \ldots, \zeta_m\}$. Fixed points p_1 and p_2 are also provided in this coordinate system, but the spheroid model is specified for a coordinate system in which the spindle axis is aligned with the z-direction and centred at the origin, for which spindle poles are located at $\tilde{p}_1 = (0,0,b)$ and $\tilde{p}_2 = (0,0,-b)$.

Data is therefore transformed to produce $\tilde{L}_i(t) = R(L_i(t) - c)$ for $c = (p_1 - p_2)/2 \in \mathbb{R}^3$ and $R \in \mathrm{SO}(3)$ the matrix such that Rv results in the rotation of the vector v by an anticlockwise angle of ϕ about the vector u for

$$\phi = \cos^{-1}\left(\frac{(p_1 - c).(0,0,1)}{||p_1 - c||_2}\right)$$
$$u = \frac{((p_1 - c)_y, -(p_1 - c)_x, 0)}{\sqrt{(p_1 - c)_x^2 + (p_1 - c)_y^2}}.$$

The vector u is therefore perpendicular to both the z-direction and the original spindle axis $p_1 - p_2$, resulting in a rotation which aligns the two.

An illustration of the model and the required transformation is displayed in Figure 5.1.

Following transformation from $L_i(t)$ to $\tilde{L}_i(t)$ the direction of lines may be clarified. Lines are directed such that with increasing t they approach the equatorial plane to match the growth of microtubules from centrosomes at the poles to chromosomes at the equatorial plane. Taking the midpoint as a reference point to determine which hemisphere the trajectory lies in, they are oriented such that $|\tilde{L}_i(0)_z| \geq |\tilde{L}_i(1)_z|$.

5.2.3 Model direction

Specification of p_1 and p_2 determines b, the semi-principal axis length of the spheroid in the z-direction. For a given location $v \in \mathbb{R}^3$ such that $|v_z| < b$ there is a unique spheroid centred at the origin that passes through v with semi-principal axis lengths

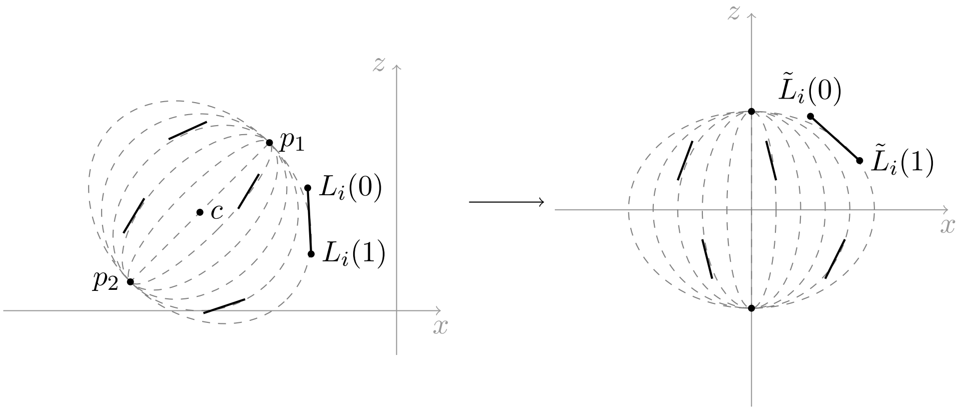


Figure 5.1: Transformations required to align observed line data with the spindle model coordinate system.

b in the z-direction and $a(v)$ in both the x- and y-directions. The value of $a(v)$ is given by

$$a(v)^2 = b^2 \left(\frac{v_x^2 + v_y^2}{b^2 - v_z^2} \right).$$

The tangent plane to the spheroid with principal axis lengths a and b at any point w along its surface is perpendicular to the normal to the surface of the spheroid at w. This normal may be obtained as the gradient of the scalar field F given by

$$\begin{aligned} F(w) &= \frac{w_x^2}{a^2} + \frac{w_y^2}{a^2} + \frac{w_z^2}{b^2} \\ \nabla F(w) &= \left(\frac{\partial F}{\partial w_x}, \frac{\partial F}{\partial w_y}, \frac{\partial F}{\partial w_z} \right) \\ &= \left(\frac{2w_x}{a^2}, \frac{2w_y}{a^2}, \frac{2w_z}{b^2} \right). \end{aligned}$$

Once $a(v)$ has been determined its value is fixed and thus

$$\nabla F(v) = \left(\frac{2v_x}{a(v)^2}, \frac{2v_y}{a(v)^2}, \frac{2v_z}{b^2} \right).$$

The proposed trajectory at v, $\rho(v)$, is then that which is perpendicular to the spheroid surface normal at v, $\nabla F(v)$, and passes through the extended spindle axis at $(0, 0, h)$ for some $h \in \mathbb{R}$. To ensure that the proposed direction is towards the equatorial plane we require that $\text{sign}(h) = \text{sign}(\rho(v)_z)$. The result is

$$\begin{aligned}
\rho(v) &= v - (0,0,h) && h : \rho(v) \cdot \nabla F(v) = 0 \\
\Rightarrow 0 &= (v - (0,0,h)) \cdot \left(\frac{2v_x}{a(v)^2}, \frac{2v_y}{a(v)^2}, \frac{2v_z}{b^2} \right) \\
\Rightarrow h &= v_z + \frac{b^2}{a(v)^2} \frac{v_x^2 + v_y^2}{v_z} \\
\Rightarrow \rho(v) &= \left(v_x, v_y, -\frac{b^2}{a(v)^2} \frac{v_x^2 + v_y^2}{v_z} \right).
\end{aligned}$$

Substituting in the value of $a(v)$ simplifies this to

$$\rho(v) = \left(v_x, v_y, \frac{v_z^2 - b^2}{v_z} \right).$$

5.2.4 Imaging correction

Observed directions $\tilde{L}_i(1) - \tilde{L}_i(0)$ could be compared to model directions at their midpoint $\rho(L_i(0.5))$, with the angle between the two representing discrepancy from the model. However, lines L_i lie in parallel imaging planes, resulting in a structured discrepancy that is dependent upon the orientation of the imaging planes which may be corrected for to obtain a more accurate measure of model fit.

Imaging planes in the original coordinate system are perpendicular to the z-direction, $\hat{z} = (0,0,1)$. After undergoing transformation they are then perpendicular to $R\hat{z}$ for the previously stated rotation matrix R. Model directions $\rho(v)$ may be projected into these same planes before comparison, resulting in projected directions $\rho'(v)$ given by

$$\rho'(v) = \rho(v) - (\rho(v) \cdot R\hat{z})R\hat{z}.$$

Discrepancy between directions proposed by the model and those of observed lines are then summarised by the collection of angles $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ for

$$\theta_i = \cos^{-1} \left(\frac{\rho'(\tilde{L}_i(0.5)) \cdot (\tilde{L}_i(1) - \tilde{L}_i(0))}{||\rho'(\tilde{L}_i(0.5))||_2 ||\tilde{L}_i(1) - \tilde{L}_i(0)||_2} \right).$$

An illustration of the angle θ_i is displayed in Figure 5.2.

5.2.5 Model comparison

Collections of angles Θ_0 and Θ_1 corresponding to different models applied to the same set of data or the same model applied to two sets of data may be compared using a comparison statistic $\Pi(\Theta_0, \Theta_1)$ to quantify relative model fit.

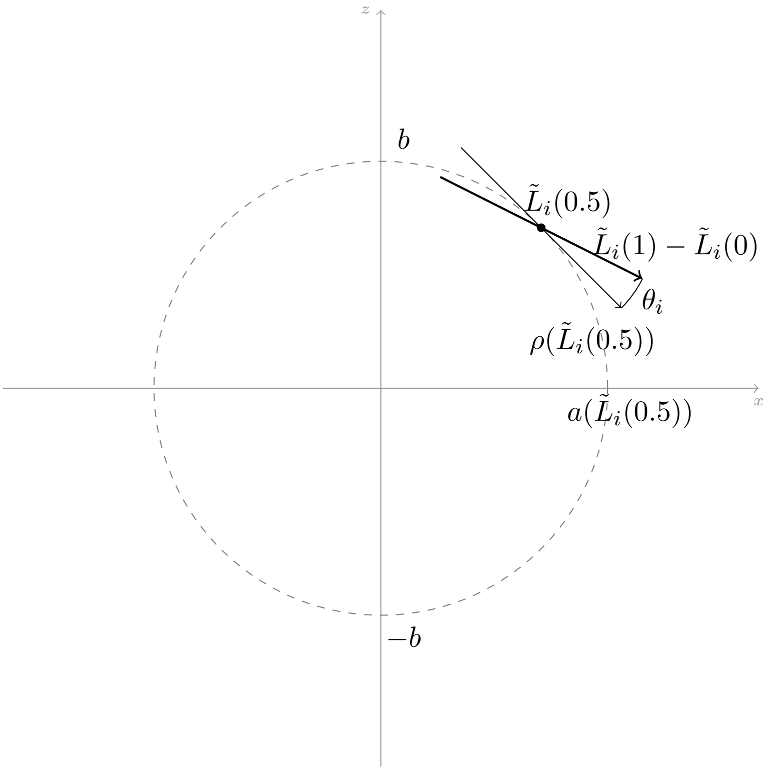


Figure 5.2: Discrepancy between observed line directions and fitted spindle model directions.

One approach could be to consider $\Pi(\Theta_0, \Theta_1) = \bar{\Theta}_0 - \bar{\Theta}_1$ the difference between the means of each collection. However, the collection Θ is in general not symmetrically distributed, making the mean a poor choice of summary statistic. Further, the mean may be heavily influenced by a small number of extreme deviations, potentially arising as a result of falsely specified lines at the stage of imaging and microtubule detection.

Instead, we propose the comparison statistic

$$\Pi(\Theta_0, \Theta_1) = \frac{1}{|\Theta_0||\Theta_1|} \sum_{\theta \in \Theta_0, \vartheta \in \Theta_1} \mathbb{1}\{\theta < \vartheta\},$$

taking values in the set $[0, 1]$. Under this formulation Π summarises comparison between every element of Θ_0 and Θ_1, resulting in a more informed comparison than that between a summary statistic of Θ_0 and a summary statistic of Θ_1. Comparison of θ and ϑ by an indicator function ensures that minor changes in either are expected to result in negligible changes in the value of $\Pi(\Theta_0, \Theta_1)$ and the impact of individual extreme values of $\theta \in \Theta_0$ or $\vartheta \in \Theta_1$ is limited to $1/|\Theta_0|$ or $1/|\Theta_1|$ respectively.

Values of $\Pi(\Theta_0, \Theta_1)$ close to 0.5 indicate little difference between sets Θ_0 and Θ_1, interpreted as similar fit for both models. Values greater than 0.5 are interpreted as better fit for the model producing Θ_0, with values less than 0.5 interpreted as better fit for the model producing Θ_1. Under the assumption that Θ_0 and Θ_1 are samples from continuous probability distributions exact equality of any pair $\theta \in \Theta_0, \vartheta \in \Theta_1$ occurs with probability zero and as a result we may specify that $\Pi(\Theta_0, \Theta_1) = 1 - \Pi(\Theta_1, \Theta_0)$.

5.3 Investigation of spindle data

5.3.1 Exploratory analysis

We have a collection of 11 observations for analysis ($1 : C^+, 2 : C^+, 3 : C^-, 4 : M^+, 5 : M^+, 6 : M^+, 7 : M^-, 8 : M^-, 9 : OE^+, 10 : OE^+, 11 : OE^-$) where C denotes a control sample with natural levels of TACC3, M denotes a mutant sample where TACC3 is removed and replaced by S558A which is known to be unable to localise on the mitotic spindle and OE denotes a sample engineered to overexpress TACC3. Superscripts $^+$ and $^-$ are used to indicate the temperature under which each sample is treated, hot and cold respectively. The number of lines in each sample varies between a minimum of 1891 and a maximum of 8800. The length of lines varies between 12 and 4650, with a median of 223 and quartiles at 107 and 430.

Our model specifies directions only for those lines which when transformed to the standard coordinate system lie vertically between the two poles, that is lines for which $|\tilde{L}_i(t)_z| < b \; \forall t \in [0, 1]$ where $\tilde{p}_1 = (0, 0, b)$. We further limit ourselves to analysis of only those lines with length greater than 60, chosen as a threshold to remove those very short lines which are expected to be least accurately specified. The number of such lines within samples varies between 1819 and 7688.

A plot of the observed lines for sample two is displayed in Figure 5.3. Note that this is not the microscope images but rather an illustration of the lines inferred from microscope images to represent microtubules. The structure of lines appears to be approximately consistent with that of the spheroid model, with poles p_1 and p_2 appearing to be reasonably specified. Plots of lines displayed in x-y planes show a strong parallel structure, an artefact of the parallel imaging planes which the final step of the methodology aims to correct for. Despite the generally good fit, there are still some lines observable which deviate significantly from what the model would propose.

The sample which by eye looks least like the idealised spheroid model is sample five, the resulting collection of lines of which are displayed in Figure 5.4. It is difficult to reconcile the lines in sample five with any sort of structured model, potentially due to influence of the experimental conditions on the structure and/or the imaging procedure appearing to include only a portion of the spindle.

Each of the collections of angles Θ summarising model fit may be plotted as a

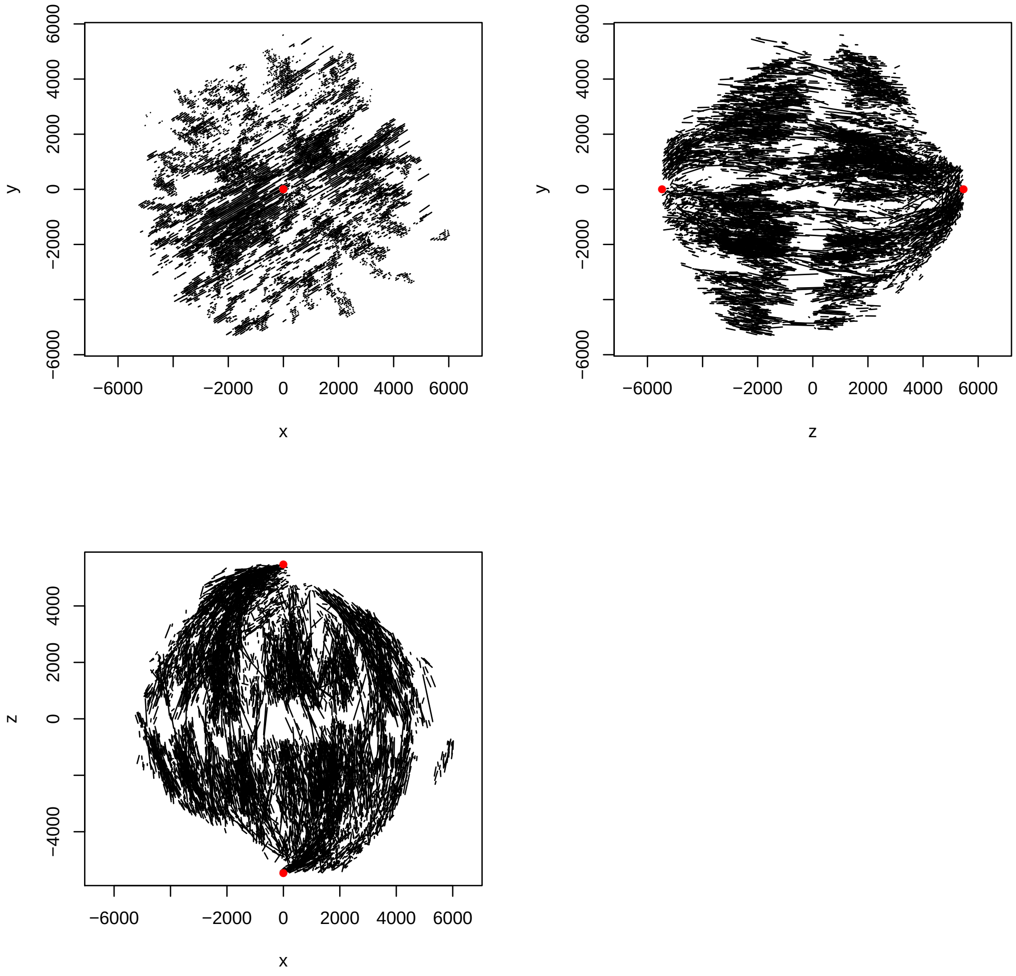


Figure 5.3: Plotted lines $\tilde{L}_i(t), t \in [0, 1]$ determined from imaging of biological sample two, a control sample treated under warm conditions. Poles $\tilde{p}_1$ and $\tilde{p}_2$ are plotted in red.

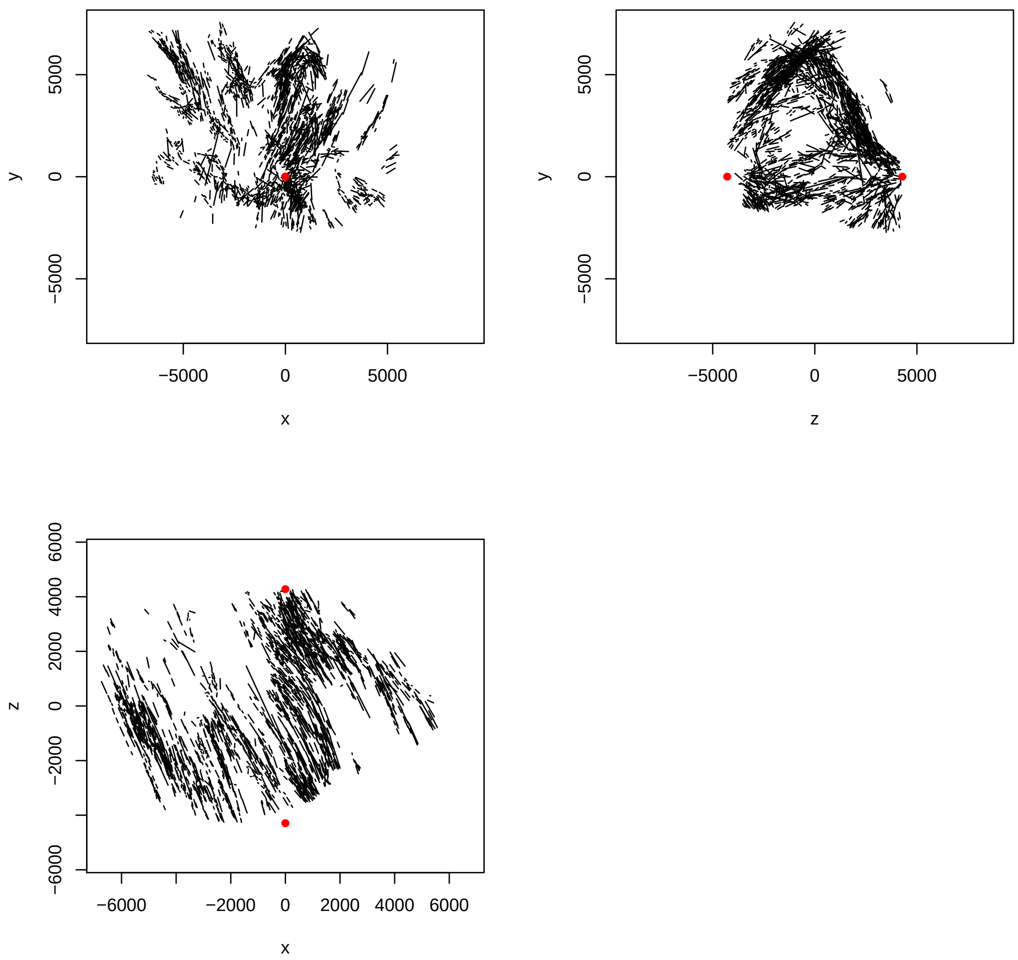


Figure 5.4: Plotted lines $\tilde{L}_i(t), t \in [0, 1]$ determined from imaging of biological sample five, a mutant sample treated under warm conditions. Poles $\tilde{p}_1$ and $\tilde{p}_2$ are plotted in red.

smoothed, normalised density to visualise their distribution and differences between samples and experimental conditions. The density of the angle distribution is estimated using a Gaussian kernel density estimate with bandwidth selection according to the suggestion of Scott (1992). An edge correction is made similar to that in Equation 2.2 described for point pattern density estimation, with each kernel divided by its integral over the interval $[0, \pi]$ on which angles θ take their values. This ensures that the resulting estimator is a normalised density on the interval $[0, \pi]$ which integrates to one. Density estimates for the 11 observations are plotted in Figure 5.5. Plotted densities display peaks at close to zero, indicating that the proposed model is in general a good fit. There is variation between observations, but the variation between samples (control, mutant, overexpressing) appears to be on the same scale as the variation between treatments (hot and cold). It is therefore very difficult to draw any conclusions on the impact of the conditions.

5.3.2 Model comparison results

The results of comparing model fit between each of the 11 observations by Π are presented in Table 5.1. The results are generally inconclusive, although a couple of general comments may be made.

For both control and TACC3 overexpressing samples there is no consistent difference between temperature treatments, with reported values of Π close to and either side of 0.5. For mutant samples there are larger deviations from 0.5 when comparing between temperature treatments, with cold treated observations deviating further from the model than warm observations four and six. However, warm sample five deviates further from the model than both cold treated observations. This may provide some indication that in the absence of TACC3 the mitotic spindle is more susceptible to disruption at low temperatures.

Comparing between samples, there appears to be little difference in model fit for control and TACC3 overexpressing samples. A difference is seen in comparison between both control and TACC3 overexpressing samples against mutant samples, with indication that model fit is superior for samples that express TACC3 to a greater or lesser degree rather than those which instead express the mutant S558A. As S558A is unable to locate to the spindle, this may provide some indication that TACC3 is important for stability of the mitotic spindle.

Given the relatively small number of samples of each type we are unable to make any conclusions as the result of statistical testing procedures, however, the results presented in this section may be used to inform future experiments and data collection, which may in turn be analysed by the proposed methodology.

The limited number of samples of each type and the novel comparison of model

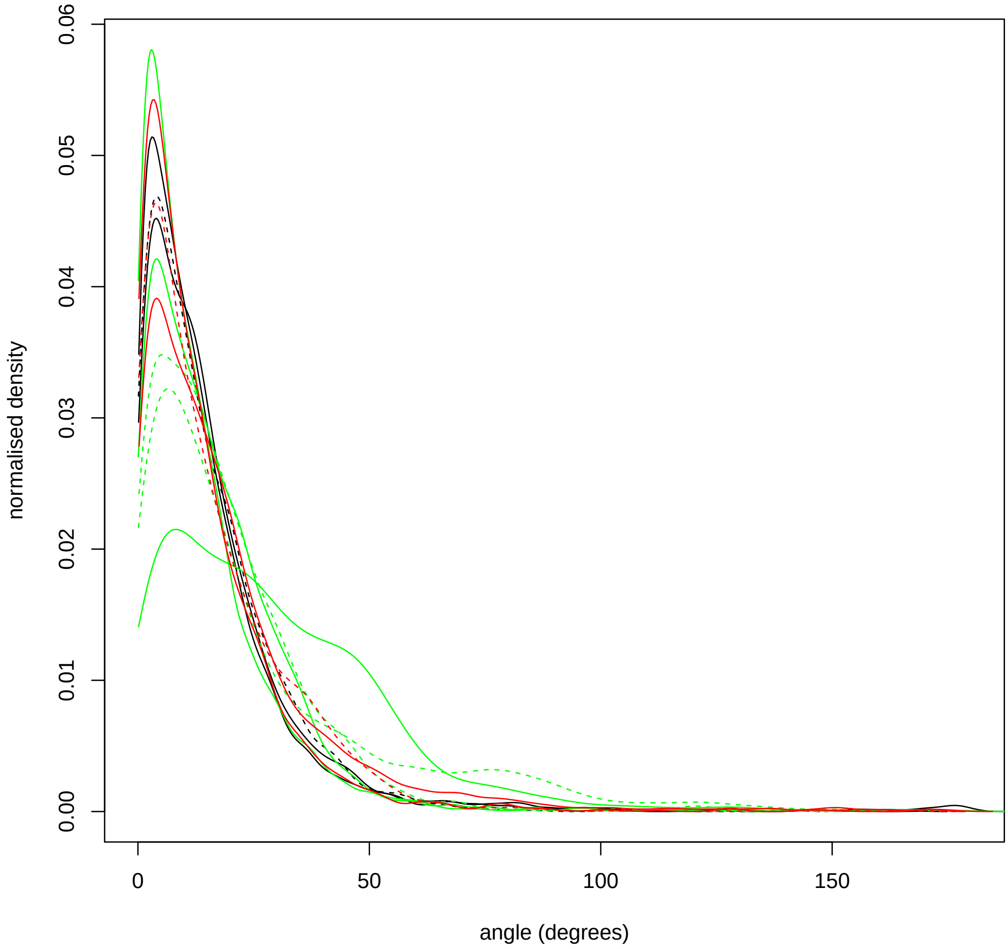


Figure 5.5: Smoothed density estimates of the model discrepancy angle, Θ, for each of the 11 biological observations. Black lines are control samples, green are mutant samples and red TACC3 overexpressing samples. Solid lines denote those warm treated, dashed lines those cold treated.

$\Theta_1 \setminus \Theta_0$	$1 : C^+$	$2 : C^+$	$3 : C^-$	$4 : M^+$	$5 : M^+$	$6 : M^+$	$7 : M^-$	$8 : M^-$	$9 : OE^+$	$10 : OE^+$	$11 : OE^-$
$1 : C^+$	0.50	0.53	0.51	0.56	0.30	0.48	0.44	0.39	0.54	0.45	0.49
$2 : C^+$	-	0.50	0.48	0.53	0.28	0.45	0.41	0.37	0.51	0.43	0.47
$3 : C^-$	-	-	0.50	0.55	0.29	0.47	0.43	0.39	0.54	0.45	0.49
$4 : M^+$	-	-	-	0.50	0.26	0.42	0.38	0.35	0.49	0.40	0.44
$5 : M^+$	-	-	-	-	0.50	0.69	0.65	0.58	0.73	0.65	0.69
$6 : M^+$	-	-	-	-	-	0.50	0.46	0.41	0.56	0.47	0.51
$7 : M^-$	-	-	-	-	-	-	0.50	0.45	0.60	0.51	0.55
$8 : M^-$	-	-	-	-	-	-	-	0.50	0.64	0.56	0.60
$9 : OE^+$	-	-	-	-	-	-	-	-	0.50	0.41	0.45
$10 : OE^+$	-	-	-	-	-	-	-	-	-	0.50	0.54
$11 : OE^-$	-	-	-	-	-	-	-	-	-	-	0.50

Table 5.1: Spindle model fit comparison scores, $\Pi(\Theta_0, \Theta_1)$, between each pairing, $\{\Theta_1, \Theta_2\}$, of the eleven observations. Lower triangle entries follow directly from $\Pi(\Theta_0, \Theta_1) = 1 - \Pi(\Theta_1, \Theta_0)$.

fit using the Π statistic mean that more work would be required to determine the magnitude of the effect size and its variability.

5.4 Conclusions

We have proposed a model for the direction of microtubules within the mitotic spindle structure which satisfies the limited number of conditions that we impose and is also straightforward to perform calculations with. The model is not expected to be the single structure to which all mitotic spindles adhere, and there is some doubt that there is indeed such a single structure, but it provides a parsimonious representation of our beliefs and the beliefs of our biologist collaborators while also fitting well to real observations in general. We further specify a statistic for comparison of model fit between two models and/or observations which is designed to explicitly compare fit across all observed lines and is expected to be robust to the types of outliers that may occur for cellular imaging data.

Application of the proposed model to a number of observations arising from cellular imaging indicates that the impact of temperature is negligible, except in the case of mutant cells expressing S558A instead of TACC3 for which cold treated observations appear to deviate further from the proposed model. Furthermore, mutant observations generally deviate further from the model than control observations and those engineered to overexpress TACC3. These interpretations arise from collections of observation comparison scores which may generally lie above or below 0.5, the indicator of equivalent model fit, but there are exceptions in every category and the variation between comparison scores is typically on the same scale as deviations from 0.5. This, combined with a limited number of observations under each condition means that any interpretations are indicative rather than the results of formal statistical tests. In future cases where larger numbers of samples are available it should be possible to apply a permutation testing methodology by relabelling samples to make more informed conclusions.

Chapter 6

Discussion

6.1 Overview

In conjunction with our collaborator Dr. Stephen Royle and members of his research group we have obtained a number of biological data sets and corresponding biological questions of interest. For each, we have modelled the data as an appropriate spatial mathematical object and formulated statistical problems from the biological questions of interest. Methodologies are presented which combine techniques from spatial statistics, optimal transportation and nonparametric hypothesis testing to make inference on the statistical questions of interest. In cases where a testing procedure is described validation has been carried out on simulated data and the sensitivity of the results to changes in the input data investigated.

Application of proposed methodologies to data sets from cells with varying levels of expression of the protein TACC3 has resulted in evidence of the influence of TACC3 on the process of mitosis. We have presented evidence that the movement of TACC3 in the mitotic spindle is directly dependent upon the movement of EB3, a protein known to localise on the tip of growing microtubules, suggesting that TACC3 has some function in microtubule creation. We have presented evidence that TACC3 overexpression results in greater numbers of microtubules packed more tightly and with a greater variety of directions within K-fibers. We have also presented preliminary evidence that when TACC3 is removed and replaced with a mutant type the mitotic spindle is less well structured.

Evidence provided in answer to the biological questions of interest support previous investigations where they have been carried out and provide new information to encourage further biological studies. Importantly, proposed methodologies are formulated with the aim of minimising assumptions such that the methods may be applied to further data sets, potentially arising from completely different scientific fields.

6.2 Possible extensions

Alongside application to alternative data sets there are a number of ways in which the methods presented in this thesis may be extended, a selection of which are described in this section.

In Chapter 3 we describe a methodology for testing for the presence of structural differences in (marked) point patterns. Marked point patterns are used to model point data recorded in each of two parallel imaging planes alongside additional information on how the points are connected. This is done under the assumption that microtubule trajectories are straight lines between the two planes. A more detailed investigation may be carried out using point data recorded in more than two parallel imaging planes (similar to the manner in which mitotic spindle data is produced) alongside additional data on the connections between points in neighbouring frames. The resulting data could be modelled as a marked point pattern where points are located in one of the two end planes, with each assigned a mark made up of an ordered set of direction vectors, one for each pair of adjacent imaging frames. Such data would require a new summary statistic, potentially an extension of the effective force transference statistic, and may require an extension of the described permutation testing approach for currently considered marked point pattern data.

For the analysis of existing marked point patterns it may be desirable to determine a summary statistic which summarises both the locations and marks of marked point patterns. The effective force transference summary statistic is currently constructed to be dependent only upon marks, as point locations are dependent upon the choice of imaging planes. While differences in the structure of point locations may be tested separately using any of the number of proposed test statistics it may be preferred to have a single statistic and correspondingly a single test to determine structural differences in marked point patterns, potentially through some transformation of the observations as in the specification of the mitotic spindle model to avoid issues with imaging plane specification.

In Chapter 4 we describe a methodology for estimating local bulk movement patterns and a test for dependence between such patterns. The movement estimation procedure is not exact and may be improved, in particular in the case of crossing object trajectories, potentially through alternative expression of the cost function, $c(x, y)$. The cost function as specified is fixed at $c(x, y) = ||x - y||_2$ across all pairs of time points and locations. Modifications to the cost function applied to calculate $\hat{F}_{s,t}$ could be made to take into account the estimated movement pattern between the preceding pair of time points, $\hat{F}_{r,s}, r < s$, under the belief that objects velocities change continuously in time. Such an approach might be similar to those proposed for tracking smaller numbers of objects, with simplifications to reduce the computational burden.

As currently formulated, tests of dependence between local bulk movement patterns are carried out separately between each pair of consecutive time points. It may be preferred to create a single test of dependence across all time points to account for existing dependencies, which could be strengthened by reformulation of the cost function to take into account information from prior time points. Such an approach could be used to support or contradict the evidence currently obtained via the omnibus test.

In Chapter 5 we describe a model for the mitotic spindle. As an exploratory analysis and initial model there are a number of ways in which the investigation could be extended. In cases where fixed point centrosomes are not directly observable within images, but instead estimated by the biologist conducting the imaging, they may instead be included in the model as parameters which are then be optimised over to obtain the best fit model, as determined by Π. Similarly, the spheroid model may be relaxed or deformed according to a number of parameters over which the best fitting model may be determined. For example, poles may be allowed to be unequal distances from the equatorial plane and trajectories may be considered to lie on the surface of ellipsoids with three unique semi-axis lengths, both of which relax the symmetry requirements upon which our model is based and might be observed in practice.

An additional extension may be made to the comparison score, Π, between models and/or samples. As presented Π compares all angles in Θ_0 to all angles in Θ_1, the result of which is potentially influenced by differences in model fit at distantly separated locations within the spindle. This issue could be overcome by including in Π only those comparisons between angles observed at similar locations within the mitotic spindle, with similar locations determined with the rotational and reflection symmetries assumed in mind.

Bibliography

J. Adler and I. Parmryd. Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry Part A*, 77A(8):733–742, March 2010. doi: 10.1002/cyto.a.20896.

J. Adler and I. Parmryd. Colocalization analysis in fluorescence microscopy. In *Methods in Molecular Biology*, pages 97–109. Springer Nature, 2012. doi: 10.1007/978-1-62703-056-4_5.

A. Akhmanova and M. O. Steinmetz. Microtubule +TIPs at a glance. *Journal of Cell Science*, 123(20):3415–3419, October 2010. doi: 10.1242/jcs.062414.

K. T. Applegate, S. Besson, A. Matov, M. H. Bagonis, K. Jaqaman, and G. Danuser. plusTipTracker: Quantitative image analysis software for the measurement of microtubule dynamics. *Journal of Structural Biology*, 176(2):168–184, November 2011. doi: 10.1016/j.jsb.2011.07.009.

A. J. Baddeley and B. W. Silverman. A cautionary example on the use of second-order methods for analyzing point patterns. *Biometrics*, 40(4):1089, December 1984. doi: 10.2307/2531159.

A. Baddeley and R. D. Gill. Kaplan-Meier estimators of distance distributions for spatial point processes. *The Annals of Statistics*, 25(1):263–292, February 1997. doi: 10.1214/aos/1034276629.

D. G. Booth, F. E. Hood, I. A. Prior, and S. J. Royle. A TACC3/ch-TOG/clathrin complex stabilises kinetochore fibres by inter-microtubule bridging. *The EMBO Journal*, 30(5):906–919, February 2011. doi: 10.1038/emboj.2011.15.

N. Chenouard, I. Smal, F. de Chaumont, M. Maška, I. F. Sbalzarini, Y. Gong, J. Cardinale, C. Carthel, S. Coraluppi, M. Winter, A. R. Cohen, W. J. Godinez, K. Rohr, Y. Kalaidzidis, L. Liang, J. Duncan, H. Shen, Y. Xu, K. E. G. Magnusson, J. Jaldén, H. M. Blau, P. Paul-Gilloteaux, P. Roudot, C. Kervrann, F. Waharte, J.-Y. Tinevez, S. L. Shorte, J. Willemse, K. Celler, G. P. van Wezel, H.-W. Dan, Y.-S. Tsai, C. O. de Solórzano, J.-C. Olivo-Marin, and E. Meijering. Objective comparison of particle tracking methods. *Nature Methods*, 11(3): 281–289, January 2014. doi: 10.1038/nmeth.2808.

E. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, April 2013. doi: 10.1214/13-aos1090.

R. M. Clegg. Fluorescence resonance energy transfer. *Current Opinion in Biotechnology*, 6(1):103–110, January 1995. doi: 10.1016/0958-1669(95)80016-6.

J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. doi: 10.1037/0033-2909.112.1.155.

S. V. Costes, D. Daelemans, E. H. Cho, Z. Dobbin, G. Pavlakis, and S. Lockett. Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal*, 86(6):3993–4003, June 2004. doi: 10.1529/biophysj.103.038422.

D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164, 1955. ISSN 00359246. URL `http://www.jstor.org/stable/2983950`.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge University Press (CUP), 1997. doi: 10.1017/cbo9780511802843.

P. Diggle. A kernel method for smoothing point process data. *Applied Statistics*, 34 (2):138, 1985. doi: 10.2307/2347366.

P. Diggle and P. J. Ribeiro. *Model-based Geostatistics.* Springer New York, 2007. ISBN 978-0-387-48536-2.

P. J. Diggle. Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. *Journal of Neuroscience Methods*, 18(1-2):115–125, October 1986. doi: 10.1016/0165-0270(86)90115-9.

P. J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* Taylor & Francis Inc, 2013. ISBN 1466560231.

P. J. Diggle, N. Lange, and F. M. Benes. Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association*, 86(415):618–625, 1991. ISSN 01621459. URL `http://0-www.jstor.org.pugwash.lib.warwick.ac.uk/stable/2290390`.

P. J. Diggle, J. Mateu, and H. E. Clough. A comparison between parametric and non-parametric approaches to the analysis of replicated spatial point patterns. *Advances in Applied Probability*, 32(02):331–343, June 2000. doi: 10.1017/s0001867800009952.

S. I. Doguwa. On edge-corrected kernel-based pair-correlation function estimators for point processes. *Biometrical Journal*, 32(1):95–106, 1990. doi: 10.1002/bimj.4710320117.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961. doi: 10.1080/01621459.1961.10482090.

E. S. Edgington. Randomization tests. *The Journal of Psychology*, 57(2):445–449, April 1964. doi: 10.1080/00223980.1964.9916711.

E. S. Edgington. Approximate randomization tests. *The Journal of Psychology*, 72(2):143–149, July 1969. doi: 10.1080/00223980.1969.10543491. URL `http://dx.doi.org/10.1080/00223980.1969.10543491`.

T. R. Fanshawe and P. J. Diggle. Bivariate geostatistical modelling: a review and an application to spatial variation in radon concentrations. *Environmental and Ecological Statistics*, 19(2):139–160, August 2011. doi: 10.1007/s10651-011-0179-7.

C. Gaetan and X. Guyon. *Spatial Statistics and Modeling*. Springer New York, 2009. ISBN 978-0-387-92257-7.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, December 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x.

C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984. doi: 10.1307/mmj/1029003026.

E. Grochowski and R. Hoyt. Future trends in hard disk drives. *IEEE Transactions on Magnetics*, 32(3):1850–1854, May 1996. doi: 10.1109/20.492876.

C. Gutierrez-Caballero, S. G. Burgess, R. Bayliss, and S. J. Royle. TACC3-ch-TOG track the growing tips of microtubules independently of clathrin and Aurora-A phosphorylation. *Biology Open*, 4(2):170–179, January 2015. doi: 10.1242/bio.201410843.

L. Heinrich. On existence and mixing properties of germ-grain models. *Statistics*, 23(3):271–286, January 1992. doi: 10.1080/02331889208802375.

A. J. Holland and D. W. Cleveland. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nature Reviews Molecular Cell Biology*, 10(7): 478–487, July 2009. doi: 10.1038/nrm2718.

T. R. Honnor, J. A. Brettschneider, and A. M. Johansen. Differences in spatial point patterns with application to subcellular biological structures. Working Paper 17-01, CRiSM, Centre for Research in Statistical Methodology, University of Warwick, January 2017a.

T. R. Honnor, A. M. Johansen, and J. A. Brettschneider. A nonparametric test for dependency between estimated local bulk movement patterns. Working Paper 17-03, CRiSM, Centre for Research in Statistical Methodology, University of Warwick, March 2017b.

K. Jaqaman, E. M. King, A. C. Amaro, J. R. Winter, J. F. Dorn, H. L. Elliott, N. Mchedlishvili, S. E. McClelland, I. M. Porter, M. Posch, A. Toso, G. Danuser, A. D. McAinsh, P. Meraldi, and J. R. Swedlow. Kinetochore alignment within the metaphase plate is regulated by centromere stiffness and microtubule depolymerases. *The Journal of Cell Biology*, 188(5):665–679, March 2010. doi: 10.1083/jcb.200909005.

K.-H. Jöckel. Finite sample properties and asymptotic efficiency of Monte Carlo tests. *The Annals of Statistics*, 14(1):336–347, 1986. ISSN 00905364. URL `http://www.jstor.org/stable/2241285`.

O. Kanoun and H.-R. Trankler. Sensor technology advances and future trends. *IEEE Transactions on Instrumentation and Measurement*, 53(6):1497–1501, December 2004. doi: 10.1109/tim.2004.834613.

J. T. Kent. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):71–80, 1982. ISSN 00359246. URL `http://www.jstor.org/stable/2984712`.

J. F. C. Kingman. *Poisson Processes (Oxford Studies in Probability)*. Clarendon Press, 1993. ISBN 9780198536932.

L. Knorr-Held and N. G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):73–85, February 2001. doi: 10.1111/1467-985x.00187.

E. Levina and P. Bickel. The earth mover's distance is the mallows distance: some insights from statistics. In *Proceedings Eighth IEEE International Conference* *on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE, Institute of Electrical and Electronics Engineers (IEEE), 2001. doi: 10.1109/iccv.2001.937632.

H. W. Lotwick and B. W. Silverman. Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):406–413, 1982. ISSN 00359246. URL `http://www.jstor.org/stable/2345499`.

E. M. M. Manders, F. J. Verbeek, and J. A. Aten. Measurement of co-localization of objects in dual-colour confocal images. *Journal of Microscopy*, 169(3):375–382, March 1993. doi: 10.1111/j.1365-2818.1993.tb03313.x.

K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3):349–393, 1975. ISSN 00359246. URL `http://www.jstor.org/stable/2984782`.

F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, March 1951. doi: 10.1080/01621459.1951.10500769. URL `http://dx.doi.org/10.1080/01621459.1951.10500769`.

J. Mateu, F. P. Schoenberg, D. M. Diez, J. A. González, and W. Lu. On measures of dissimilarity between point patterns: Classification based on prototypes and multidimensional scaling. *Biometrical Journal*, 57(2):340–358, December 2014. doi: 10.1002/bimj.201300150.

N. J. B. McFarlane and C. P. Schofield. Segmentation and tracking of piglets in images. *Machine Vision and Applications*, 8(3):187–193, May 1995. doi: 10.1007/bf01215814.

Y. Mimori-Kiyosue, N. Shiina, and S. Tsukita. The dynamic behavior of the APC-binding protein EB1 on the distal ends of microtubules. *Current Biology*, 10(14):865–868, July 2000. doi: 10.1016/s0960-9822(00)00600-x.

W. A. Mitchell and S. L. Lima. Predator-prey shell games: large-scale movement and its implications for decision-making by prey. *Oikos*, 99(2):249–259, November 2002. doi: 10.1034/j.1600-0706.2002.990205.x.

J. Møller. Shot noise Cox processes. *Advances in Applied Probability*, 35(03):614–640, September 2003. doi: 10.1017/s0001867800012465.

J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. CRC Press, 2003. ISBN 9781584882657.

J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, September 1998. doi: 10.1111/1467-9469.00115.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. ISSN 03684245. URL `http://0-www.jstor.org.pugwash.lib.warwick.ac.uk/stable/2098689`.

J. Neyman and E. L. Scott. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(1):1–43, 1958. ISSN 00359246. URL `http://www.jstor.org/stable/2983905`.

F. M. Nixon, C. Gutiérrez-Caballero, F. E. Hood, D. G. Booth, I. A. Prior, and S. J. Royle. The mesh is a network of microtubule connectors that stabilizes individual kinetochore fibers of the mitotic spindle. *eLife*, 4, June 2015. doi: 10.7554/elife.07635. URL `http://dx.doi.org/10.7554/eLife.07635`.

F. M. Nixon, T. R. Honnor, N. I. Clarke, G. P. Starling, A. J. Beckett, A. M. Johansen, J. A. Brettschneider, I. A. Prior, and S. J. Royle. Microtubule organization within mitotic spindles revealed by serial block face scanning electron microscopy and image analysis. *Journal of Cell Science*, 130(10):1845–1855, April 2017. doi: 10.1242/jcs.203877. URL `http://jcs.biologists.org/content/130/10/1845`.

K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. ISSN 03701662. URL `http://0-www.jstor.org.pugwash.lib.warwick.ac.uk/stable/115794`.

S. Peleg, M. Werman, and H. Rom. A unified approach to the change of resolution: space and gray-level. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):739–742, July 1989. doi: 10.1109/34.192468.

D. W. Piston and G.-J. Kremers. Fluorescent protein FRET: the good, the bad and the ugly. *Trends in Biochemical Sciences*, 32(9):407–414, September 2007. doi: 10.1016/j.tibs.2007.08.003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):172–212, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984796`.

J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469): 94–108, March 2005. doi: 10.1198/016214504000000539.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/a:1026543900054.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* JOHN WILEY & SONS INC, 1992. ISBN 0471547700.

J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1): 561–584, January 1995. doi: 10.1146/annurev.ps.46.020195.003021.

R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986. doi: 10.1093/biomet/73.3.751.

D. Stoyan and A. Penttinen. Recent applications of point process methods in forestry statistics. *Statistical Science*, 15(1):61–78, 2000. ISSN 08834237. URL `http://www.jstor.org/stable/2676677`.

D. J. Strauss. A model for clustering. *Biometrika*, 62(2):467–475, August 1975. doi: 10.1093/biomet/62.2.467. URL `http://dx.doi.org/10.1093/biomet/62.2.467`.

S. Urbanek and Y. Rubner. *emdist: Earth Mover's Distance*, 2012. URL `https://CRAN.R-project.org/package=emdist`. R package version 0.3-1.

M. N. M. van Lieshout. *Markov Point Processes and Their Applications.* World Scientific Pub Co Pte Lt, July 2000. doi: 10.1142/p060.

V. Zinchuk and O. Zinchuk. *Quantitative Colocalization Analysis of Confocal Fluorescence Microscopy Images*, chapter 4.19, pages 1–16. John Wiley & Sons, Inc., 2008. ISBN 9780471143031. doi: 10.1002/0471143030.cb0419s39. URL `http://dx.doi.org/10.1002/0471143030.cb0419s39`.