AN APPLICATION OF SPATIAL POINT PROCESS METHODS TO
RSV INFECTION EXPERIMENTS

A Thesis in
Statistics
by
Ivan Botev Simeonov

© 2010 Ivan Botev Simeonov

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2010
The thesis of Ivan Botev Simeonov was reviewed and approved* by the following:

John Fricks  
Assistant Professor of Statistics  
Thesis Co-Adviser  

Francesca Chiaromonte  
Professor of Statistics and Public Health Sciences  
Thesis Co-Adviser  

Bruce G. Lindsay  
Willaman Professor of Statistics  
Head of Department of Statistics  

*Signatures are on file in the Graduate School.
Abstract

Viral infections are a common occurrence among humans and other animals. Biologists are often interested in investigating the progression of infections and co-infections (more than one virus strain) within a host organism. Various studies point to the fact that prior infections may substantially affect susceptibility to subsequent infections – not just through responses due to the immune system, but through changes in the state of the cells exposed to infection within a tissue [Martinez et al., 2007, Lo et al., 2005, Ruggiero et al., 1989]. As part of the current investigation, *in vitro* studies are conducted where infections are observed in cultures of epithelial cells, since these cultures closely resemble the structure present in the human throat. Some of the key questions concern the spatial distribution of infections in these cultures, e.g. if there is a tendency for infected cells to increase or decrease the probability of a second infection. A spatial analysis allows us to characterize the infection status of cells as a function of their location and proximity to one another and thus to explore how susceptibility may be affected by local conditions (e.g. in terms of crowding of the cells in the culture) and signaling mechanisms among the cells (e.g. in response to infection).
# Table of Contents

List of Figures vi

List of Tables xi

Acknowledgments xii

Chapter 1
Introduction and Background 1
1.1 A Motivating Biological Problem .......................... 1
1.2 Introduction to Spatial Point Processes ................. 2

Chapter 2
K-function 8
2.1 K-function ........................................ 8
2.1.1 Cross K-function ................................ 11
2.2 Pair Correlation Function ............................ 12
2.3 Inference for the K-function .......................... 12
2.3.1 Null Envelopes for K-functions ..................... 18
2.3.2 Simulation Study for Evaluating Estimators of $K_X(r)$ ...... 19
2.3.3 Effect of Size of Available Observed Point Patterns on K-
function Estimation .................................. 21
2.3.4 Effect of Size of Available Observed Point Patterns on Cross
K-function Estimation .................................. 22

Chapter 3
Application of Exploratory Spatial Point Process Methods to
RSVA In Vitro Infection Experiments 32
3.1 Description of Experiments and Data ....................... 32
List of Figures

1.1 Microscopy imaging of a cell culture infected with two strains of RSV. Blue marks represent cell nuclei, red marks are located on the membrane of cells infected with RSV B, and green marks are located in the cytoplasm of cells infected with RSV A. ............................... 3

1.2 An observed point pattern generated from a homogeneous Poisson point process with \( \lambda = 30 \). ................................................................. 4

1.3 Left: A simulated Poisson point pattern from the intensity \( \lambda X_1(\xi) = 50 \) (the intensity is shown with background shading and the points are shown in red); Middle: A simulated Poisson point pattern from the intensity \( \lambda X_2(\xi) = e^{2.061704 + 3\xi_1} \); Right: A simulated Poisson point pattern from the intensity \( \lambda X_3(\xi) = e^{1+6.42897^*\xi_1^*\xi_2} \). In all three cases \( N(\cdot) \sim \text{Poisson}(50) \). ......................................................... 6

2.1 Simulated point pattern from an inhomogeneous Poisson process with intensity \( \lambda X(\xi) = 100 \frac{1}{\sqrt{2\pi0.21^2}} e^{-\frac{(\xi_1-0.5)^2}{2\times0.21^2}} \frac{1}{\sqrt{2\pi0.21^2}} e^{-\frac{(\xi_2-0.5)^2}{2\times0.21^2}} \) ................. 14

2.2 2D example of kernel smoothed intensity (the data is shown in Figure 2.1). The kernel is Gaussian, and the bandwidths used are from left to right: 0.07, 0.13 and 0.4. The true intensity is shown in on the right most graph. ................................................................. 14

2.3 2D example of a point pattern within and outside an observation window. The figure illustrates the need for edge correction in estimating the intensity of the process. ................................. 15

2.4 Left: Homogeneous intensity; Middle: Inhomogeneous intensity with low spatial variation; Right: Inhomogeneous intensity with high spatial variation. ......................................................... 20
2.5 Evaluating estimators of $K_X(r)$: Homogeneous Intensity. Upper panels: Estimated mean of $K$-functions from 39 simulations using formulas (1) (left), (2) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Estimated mean of $K$-functions from 39 simulations using formulas (3) (left), (4) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. Lower panel: Estimated mean of $K$-functions from 39 simulations using formula (5) from Table 2.1 with $R$ equal to 50, 100, 150, 200, 250, and 300.

2.6 Evaluating estimators of $K_X(r)$: Inhomogeneous Low Intensity. Upper panels: Estimated mean of $K$-functions from 39 simulations using formulas (1) (left), (2) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Estimated mean of $K$-functions from 39 simulations using formulas (3) (left), (4) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. Lower panel: Estimated mean of $K$-functions from 39 simulations using formula (5) from Table 2.1 with $R$ equal to 50, 100, 150, 200, 250, and 300.

2.7 Evaluating estimators of $K_X(r)$: Inhomogeneous High Intensity. Upper panels: Estimated mean of $K$-functions from 39 simulations using formulas (1) (left), (2) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Estimated mean of $K$-functions from 39 simulations using formulas (3) (left), (4) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. Lower panel: Estimated mean of $K$-functions from 39 simulations using formula (5) from Table 2.1 with $R$ equal to 50, 100, 150, 200, 250, and 300.

2.8 Effects of size of available observed point patterns on $K$-function Estimation: Homogeneous Intensity. The expected number of points generated from the point process are shown on the right column. The estimate of the $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
2.9 Effects of size of available observed point patterns on $K$-function Estimation: Inhomogeneous Low Intensity. The expected number of points generated from the point process are shown on the right column. The estimate of the $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.

2.10 Effects of size of available observed point patterns on $K$-function Estimation: Inhomogeneous High Intensity. The expected number of points generated from the point process are shown on the right column. The estimate of the $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.

2.11 Effects of size of available observed point patterns on cross $K$-function Estimation: Homogeneous Intensity. The expected number of points generated from the point processes are shown on the first and second columns to the right. The estimate of the cross $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.

2.12 Effects of size of available observed point patterns on cross $K$-function Estimation: Inhomogeneous Low Intensity. The expected number of points generated from the point processes are shown on the first and second columns to the right. The estimate of the cross $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.

2.13 Effects of size of available observed point patterns on cross $K$-function Estimation: Inhomogeneous High Intensity. The expected number of points generated from the point processes are shown on the first and second columns to the right. The estimate of the cross $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
3.1 Microscopy imaging of a cell culture infected with one strain of RSV. Blue marks represent cell nuclei and green marks are located in the cytoplasm of cells infected with RSV A.

3.2 Counts of nuclei (left), and A marks (right) as a function of lag. Different points represent different replicate experiments at each lag duration, and color coding is consistent across the two plots – thus, for instance, the light blue points represent well replicate 5 at lag $h = 12$, which has the highest number of nuclei, and the lowest number of A stains.

3.3 Kernel estimates of the nuclei intensities for well replicate 3 at $h = 0$ (left), well replicate 2 at $h = 9$ (middle), and well replicate 5 at $h = 15$ (right). The bandwidth was set to $b = 250$.

3.4 Upper panels: Estimated $K$-functions for observed nuclei, $\hat{K}_N(r)$ (black), and simulated cell centers, $\hat{K}_C(r)$ (red). The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Kernel estimates of observed nuclei intensities with bandwidth $b = 250$. Lower panels: Kernel estimates of simulated cell centers intensities with bandwidth $b = 250$. Well replicate 4 at $h = 0$ and well replicate 5 at $h = 15$ are shown on the left and the right, respectively.

3.5 Kernel smoothed density of estimated radii of A stains. Well replicate 4 at $h = 0$ and well replicate 5 at $h = 15$ are shown on the left and the right, respectively.

3.6 Upper panels: Estimated inhomogeneous $K$-functions for observed A stains, $\hat{K}_{A,A}(r)$ (black), along with average and min-max bands for the $99 \hat{\hat{K}}_{A,A}(r)$ from the simulations paradigm 1 (in colors) – the Poisson behavior ($\pi r^2$) is subtracted to each of the curves to help visualizing their variation. Well replicate 4 at $h = 0$ and replicate 5 at $h = 15$ are shown on the left and the right, respectively.

3.7 Upper panels: Estimated inhomogeneous $K$-functions for observed A stains, $\hat{K}_{A,A}(r)$ (black), along with average and min-max bands for the $99 \hat{\hat{K}}_{A,A}(r)$ from the simulations paradigm 2 with a 95% trim (in colors). Lower panels: Estimated inhomogeneous $K$-functions for observed A stains, $\hat{K}_{A,A}(r)$ (black), along with average and min-max bands for the $99 \hat{\hat{K}}_{A,A}(r)$ from the simulations paradigm 2 with 80% trimming (in colors) – the Poisson behavior ($\pi r^2$) is subtracted to each of the curves to help visualizing their variation. Well replicate 4 at $h = 0$ and replicate 5 at $h = 15$ are shown on the left and the right, respectively.
3.8 Upper panels: scatter plots of the counts of A marks $c_{A,g}$ vs the count of nuclei $c_{N,g}$ for a gridding of the wells in 213 units, with lowess smooths (in black) superimposed to visualize the relationship. Lines are also shown (in red and green) for the 2.5%-percentile and 9.5%-percentile vs count of nuclei, $c_{N,g}$, to provide a reference. Middle panels: Spatial plots of the well where the counts are outside of the bands from the upper panel. Lower Panels: Kernel smoothed intensities of Nuclei. Replicate 4 at $h = 0$ and replicate 5 at $h = 15$ are shown on the left and the right, respectively.

4.1 Counts of nuclei (left), B marks (middle) and A marks (right) as a function of lag for the co-infection experiments. Different points represent different replicate experiments at each lag duration, and color coding is consistent across the three plots – thus, for instance, the green points represent well replicate 3 at lag $h = 12$, which has the lowest number of nuclei, the lowest number of B stains, and the highest number of A stains.
List of Tables

2.1 Summary of the estimation approaches for the $K$-function that will be compared through a simulation study . . . . . . . . . . . . 18
Acknowledgments

I owe my deepest gratitude to my thesis advisors, Professors John Fricks and Francesca Chiaromonte, and to our close collaborator, Professor Mary Poss. Their encouragement, guidance and support at all levels were crucial to my understanding of the subject. Professor Poss also made data from her laboratory at the Pennsylvania State University available for this work. Lastly, I would like to thank all of those who supported me in any respect during the completion of the project.

Ivan Botev Simeonov
Chapter 1

Introduction and Background

In this chapter, we introduce a motivating biological problem, mathematical notation that will be used throughout this thesis, and some notions on spatial point process through the canonical example of spatial Poisson point processes.

1.1 A Motivating Biological Problem

Viral infections are a common occurrence among humans and other animals. Biologists are often interested in investigating the progression of infections and co-infections (more than one virus strain) within a host organism. Various studies point to the fact that prior infections may substantially affect susceptibility to subsequent infections – not just through responses due to the immune system, but through changes in the state of the cells exposed to infection within a tissue [Martinez et al., 2007, Lo et al., 2005, Ruggiero et al., 1989]. As part of the current investigation, in vitro studies are conducted where infections are observed in cultures of epithelial cells, since these cultures closely resemble the structure present in the human throat. Some of the key questions concern the spatial distribution of infections in these cultures, e.g. if there is a tendency for infected cells to increase or decrease the probability of a second infection. A spatial analysis allows us to characterize the infection status of cells as a function of their location and proximity to one another and thus to explore how susceptibility may be affected by local conditions (e.g. in terms of crowding of the cells in the culture) and signaling mechanisms among the cells (e.g. in response to infection). In particular, we are in-
Interested in detecting if the probability of an infection of the same or different virus increases or decreases depending on the location (absolute or relative) of the infected cells. A significant decrease in probability would suggest local conditions or signaling mechanisms leading to decreased susceptibility, and a significant increase in probability would conversely suggest local conditions or signaling mechanisms leading to increased susceptibility. The insight obtained from this type of analysis might lead to better development of vaccines and drugs for treating various infections. More details on an application of this type will be provided in Chapter 3.

Data from \textit{in vitro} studies are obtained from microscopy imaging of cell cultures in which cell nuclei as well as infections status are marked with fluorescent stains. An example is shown in Figure 1.1, where the blue stains represent nuclei, the red are membrane stains representing RSVB infections, and the green are cytoplasmic stains representing RSVA infections. RSV is a virus affecting the human respiratory system [Manoha et al., 2007, Jamaluddin et al., 2001], and A and B are two strains of the virus. Images such as the one shown in Figure 1.1 can be processed with specialized software to obtain counts as well as spatial coordinates of the different types of marks [Image-Pro Plus 6.3, 2007].

When investigating spatial features for data from \textit{in vitro} experiments, one has to ascertain and account for effects related to the cell culture structure, as well as the microscopy imaging techniques used to gather the data that are not of primary interest. For instance, cell cultures may be inhomogeneous in terms of cell density and grow more populated over the time frame of a given experiment. Moreover, microscopy imaging software requires careful tuning of parameters and even with careful tuning may produce some ambiguous results due to the nature of and the differences between different types of stains. More details on these facts and their implications are provided in Chapter 3.

\subsection{1.2 Introduction to Spatial Point Processes}

An important tool in detecting how susceptibility of the cells may be affected by local conditions (e.g. in terms of crowding) and signaling mechanisms among the cells (e.g. in response to infection) are spatial point processes, which can be used
Figure 1.1. Microscopy imaging of a cell culture infected with two strains of RSV. Blue marks represent cell nuclei, red marks are located on the membrane of cells infected with RSV B, and green marks are located in the cytoplasm of cells infected with RSV A.

to model the locations of cells and infected cells in the cultures. Of particular interest are the overall abundance of points, their locations, and any evidence of attraction or repulsion. Attraction or repulsion will be a proxi for a tendency of the cells surrounding an infected cell to have higher or lower susceptibility to infection. Throughout the rest of the thesis whenever we refer to point processes we will mean spatial point processes.

A point process $\mathbf{X}$ is a random countable subset of the space $\mathbb{R}^d$ [Moller and Waagepetersen, 2003], which is usually observed in a bounded observation window $W \subset \mathbb{R}^d$. This bounded observation window is assumed to be a Borel set. In most practical applications, the number of points from $\mathbf{X}$ in the observation window $W$ is finite; denoting by $N_{\mathbf{X}}(\cdot)$ the number of points in a given set, we can write $N_{\mathbf{X}}(W) < \infty$. The expression $N_{\mathbf{X}}(\cdot)$ is a random mapping from any reasonable subset in $\mathbb{R}^d$ to the natural numbers. We will denote the observed point patterns (point configurations) as $\mathbf{x} = \{x_1, ..., x_{n_\mathbf{x}}\}$, where $x_1, ..., x_{n_\mathbf{x}} \in \mathbb{R}^d$ are the unordered observed points and $n_\mathbf{x}$ is the number of observed points. If $d = 2$, then each of the points will have two coordinates indicating its location in $W$. Throughout the rest of the thesis we will use $d = 2$; however, the definitions are also true
for larger $d$.

To give some intuition of point processes, we consider here a simple but important example: the homogeneous Poisson point process. If we consider such a process in a 2D unit square $W$, the number of points that will populate the square is determined as a draw from a Poisson distribution with mean $\mu(W) = \lambda \times |W| = \lambda \times 1$, and their locations in the square are determined by independent draws from a uniform distribution. Another way of describing the same process is the following: if $A_i, i = 1, ..., k$ are disjoint bounded 2D subsets, then the number of points within each of the disjoint bounded subsets are independent Poisson random variables with means $\mu(A_1) = \lambda \times |A_1|, ..., \mu(A_k) = \lambda \times |A_k|$ for all $k$ and all reasonable $A_i, i = 1, ..., k$. An example of an observed point pattern generated by a Poisson point process with $\lambda = 30$ in the 2D unit-square $W$ is shown in Figure 1.2.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{poisson_pattern.png}
\caption{An observed point pattern generated from a homogeneous Poisson point process with $\lambda = 30$.}
\end{figure}

The intensity plays an important part in the definition of a point process. The intensity measure is defined as $\mu(A) = E[N_X(A)]$. In other words, the intensity measure is the expected number of points in any set $A \subset \mathbb{R}^2$. Note that $A$ needs to be a sufficiently regular (i.e. Borel) set. If there exists a non-negative function $\lambda_X(\cdot)$ such that $\mu(A) = \int_A \lambda_X(\xi) d\xi$ for any $A$, then $\lambda_X(\cdot)$ is called the intensity function of the point process $X$. If $\lambda_X(\cdot) = c$ for some constant $c$, then the process is said to be homogeneous. If $\lambda_X(\cdot)$ is a non-constant function, then the process
is said to be inhomogeneous with a spatially varying intensity. For the example shown in Figure 1.2, the process is a homogeneous Poisson point process with intensity $\lambda = 30$.

The Poisson point process is often used as a baseline to model uniform spatial randomness. If a point process exhibits clustering or repulsion, then the process cannot be a Poisson point process. Conditioned on the number of points, the point locations of a Poisson process are independent of one another. This can be interpreted as a form of absence of interaction, points neither repel nor attract one another. The simplest type of Poisson point process is the homogeneous Poisson point process. However, a Poisson point processes can be inhomogeneous in space, i.e. have a non-constant intensity. Inhomogeneous Poisson point processes are also used to model no interaction among points.

A Poisson point process $\mathbf{X}$ on $\mathbb{R}^2$ with intensity function $\lambda_\mathbf{X}(\cdot)$ is defined as follows [Moller and Waagepetersen, 2003]

1. For any $A \subset \mathbb{R}^2$ with $\mu(A) = \int_A \lambda_\mathbf{X}(\xi)d\xi < \infty$, $N_\mathbf{X}(A) \sim \text{Pois}[\mu(A)]$.

2. For any $A \subset \mathbb{R}^2$, such that $0 < \mu(A) < \infty$, conditional on $N_\mathbf{X}(A) = n_\mathbf{x}$, the points $\mathbf{x} = \{x_1, \ldots, x_{n_\mathbf{x}}\}$ are independently and identically distributed with density $f(\xi) = \frac{\lambda_\mathbf{X}(\xi)}{\int_A \lambda_\mathbf{X}(\omega)d\omega}$, $\xi \in A$.

The number of points is determined based on (1) with a Poisson distribution. The scattering of the points is defined by (2) with a density proportional to the intensity $\lambda_\mathbf{X}(\cdot)$ and with points distributed independently from each other. Notice that based on the definition there is no restriction on how close the points can be to each other. Figure 1.3 shows three point patterns simulated from the distributions of Poisson point processes with the following intensities on $\mathbb{R}^2$: $\lambda_{\mathbf{X}_1}(\xi) = 50$, $\lambda_{\mathbf{X}_2}(\xi) = e^{2.061704+3*\xi_1}$, $\lambda_{\mathbf{X}_3}(\xi) = e^{1+6.42897*\xi_1*\xi_2}$ in the bounded observation window $W = \text{square}((0,0), (1,0), (0,1), (1,1))$. The Poisson point process $\mathbf{X}_1$ is homogeneous (constant intensity), while the Poisson processes $\mathbf{X}_2$ and $\mathbf{X}_3$ are inhomogeneous. In all three cases $\mu(W) = \int_W \lambda_{\mathbf{X}_1}(\xi)d\xi = \int_W \lambda_{\mathbf{X}_2}(\xi)d\xi = \int_W \lambda_{\mathbf{X}_3}(\xi)d\xi = 50$, so we simulate $N(\cdot)$ from a Poisson distribution with mean 50.

We observe that in the left plot there is no tendency for the points to fall in a particular region. In the middle plot, the points have a higher likelihood of falling
Figure 1.3. Left: A simulated Poisson point pattern from the intensity $\lambda_{X_1}(\xi) = 50$ (the intensity is shown with background shading and the points are shown in red); Middle: A simulated Poisson point pattern from the intensity $\lambda_{X_2}(\xi) = e^{2.061704 + 3\xi_1}$; Right: A simulated Poisson point pattern from the intensity $\lambda_{X_3}(\xi) = e^{1 + 6.42897 \xi_1 \xi_2}$. In all three cases $N(\cdot) \sim \text{Poission}(50)$.

into regions where the $\xi_1$ coordinate is greater, and on the right plot the points have a tendency to locate where both of the coordinates $\xi_1$ and $\xi_2$ are large.

Consider now a probability surface $p(\xi) : \mathbb{R}^2 \rightarrow [0, 1]$, and a Poisson point process $X$ with intensity $\lambda_X(\xi)$. With any point in our point process we associate a Bernoulli random variable, with probability of success $p(\xi)$. If the outcome of this Bernoulli random variable is 1, then we retain the point. If the outcome is 0, then we delete the point. This procedure is referred to as thinning. The resulting point process $X_{\text{thinned}}$ will also be a Poisson process with an intensity function $\lambda_{X_{\text{thinned}}}(\xi) = p(\xi)\lambda_X(\xi)$ [Moller and Waagepetersen, 2003]. This leads us to an algorithm for fast generation of Poisson point processes with an arbitrary intensity $\lambda_X(\xi)$ on a specified bounded observation window $W$:

1. Find the maximum of the intensity in the observation window $W$, i.e. $c = \max_{\xi \in W}(\lambda_X(\xi))$.

2. Simulate a homogeneous Poisson process $Y$ with intensity $\lambda_Y(\xi) = c$.

3. Perform thinning on the simulated points, keeping each point, independently, with probability $p(\xi) = \frac{\lambda_X(\xi)}{c}$.

From the above discussion, the intensity of the resulting process will indeed be $\lambda_{Y_{\text{thinned}}}(\xi) = p(\xi) c = \frac{\lambda_X(\xi)}{c} c = \lambda_X(\xi)$. The simulated Poisson point patterns in Figure 1.3 were created using this algorithm.
Poisson point processes are just one class of such point processes. Many more examples have been introduced and described in the literature[Moller and Waagepetersen, 2003, Gelfand et al., 2010]. One such example of a point process is the Cox process. These are obtained by considering the intensity function of a Poisson process as a realization of a non-negative random field. Another example is a Markov point process. These are commonly used to model repulsion among points. One particular example of a Markov point process is the hard-core Strauss point process, where points are prohibited from being closer than a certain distance from each other.
Chapter 2

*K-function*

In this chapter, we introduce a function that is widely used as a descriptor of point processes and procedures for its estimation. We also present a simulation study evaluating the performance of standard and alternative estimators for this descriptor.

### 2.1 *K*-function

In addition to the intensity function, which captures the tendency of a point process $X$ to contain more points at specific locations, another important descriptor is the Ripley’s *K*-function. The *K*-function captures the tendency of points to repel or attract one another. Given a set $A \subset \mathbb{R}^2$ with finite area $0 < |A| < \infty$, for a point process with homogeneous intensity $\lambda_X(\cdot) = \lambda_X$, it is defined as

\[
K_X(r) = \frac{1}{|A|} E \left[ \sum_{x \in X \cap A} \sum_{\tilde{x} \in X \cap A, \tilde{x} \neq x} \frac{\mathbb{1}_{\{d(x, \tilde{x}) \leq r\}}}{\lambda_X} \right] \tag{2.1}
\]

\[
= \frac{1}{|A|\lambda_X^2} E \left[ \sum_{x \in X \cap A} \sum_{\tilde{x} \in X \cap A, \tilde{x} \neq x} \mathbb{1}_{\{d(x, \tilde{x}) \leq r\}} \right] \tag{2.2}
\]

where $E[\cdot]$ indicates the expected value, $\mathbb{1}_{\{\cdot\}}$ is an indicator function of an event, $d(\cdot, \cdot)$ is the distance between two points, and $r$ is a radius. The expression $\lambda_X K_X(r)$ can be interpreted as the expected number of points in a ball of radius $r$ from a “typical” point of the process.
For a point process \( X \) whose intensity \( \lambda_X(\cdot) \) is \textit{not} constant in space but still possesses a property of spatial regularity called second order intensity reweighted stationarity, the appropriate definition to use is an inhomogeneous version of Ripley’s \( K \) [Baddeley et al., 2000, Moller and Waagepetersen, 2003]. This is defined as

\[
K_X(r) = \frac{1}{|A|} E \left[ \sum_{x \in X \cap A} \frac{1}{\lambda_X(x)} \sum_{\tilde{x} \in X, \tilde{x} \neq x} \frac{1_{\{d(x, \tilde{x}) \leq r\}}}{\lambda_X(\tilde{x})} \right] \tag{2.3}
\]

Second order intensity reweighted stationarity implies that the function above does not depend on the region we select. It has the same value when computed for any \( A \). Intuitively, \( K_X(r) \) expresses the expected spatial accumulation of points in a neighborhood of radius \( r \) around a “typical” point of the process, adjusted for the intensity of the process itself.

Before discussing more properties of the \( K \)-function, we need to introduce another definition; namely, the second-order factorial moment measure of a point process \( X \) given by

\[
\alpha^2(B_1 \times B_2) = E \left[ \sum_{x, \tilde{x} \in X, x \neq \tilde{x}} 1_{\{(x, \tilde{x}) \in B_1 \times B_2\}} \right] \quad , \quad B_1 \times B_2 \subset \mathbb{R}^2 \times \mathbb{R}^2. \tag{2.4}
\]

Assume a non-negative function \( \lambda^2_{X,X}(\xi, \eta) \) exists, such that for any reasonable \( B_1 \subset \mathbb{R}^2 \) and \( B_2 \subset \mathbb{R}^2 \)

\[
\alpha^2(B_1 \times B_2) = \int_{B_1} \int_{B_2} \lambda^2_{X,X}(x, \tilde{x}) \, dx \, d\tilde{x} \quad , \quad B_1 \times B_2 \in \mathbb{R}^2 \times \mathbb{R}^2. \tag{2.5}
\]

Then, the function \( \lambda^2_{X,X}(\xi, \eta) \) is called the second-order product density of the process. The expression \( \lambda^2_{X,X}(\xi, \eta) d\xi d\eta \) can be interpreted as the joint probability of points falling in each of the infinitesimal regions centered at \( \xi \) and \( \eta \). Based on the above definitions, [Moller and Waagepetersen, 2003] proved the following: Suppose that the point process \( X \) has intensity function \( \lambda_X(\cdot) \) and second order product density \( \lambda^2_{X,X}(\cdot, \cdot) \), then for any functions \( h_1 : \mathbb{R}^2 \to [0, \infty) \) and \( h_2 : \mathbb{R}^2 \times \)
\(\mathbb{R}^2 \to [0, \infty)\), we have

\[
E \left[ \sum_{x \in \mathbf{X}} h_1(x) \right] = \int h_1(x) \lambda_{\mathbf{X}}(x) dx
\]  \hspace{1cm} (2.6)

\[
E \left[ \sum_{x, \tilde{x} \in \mathbf{X}, x \neq \tilde{x}} h_2(x, \tilde{x}) \right] = \int \int h_2(x, \tilde{x}) \lambda_{\mathbf{X}, \mathbf{X}}(x, \tilde{x}) dx d\tilde{x}.
\]  \hspace{1cm} (2.7)

We can use this fact to rewrite the \(K\)-function. Setting \(h_2(x, \tilde{x}) = \frac{1_{\{d(x, \tilde{x}) \leq r\}}}{\lambda_{\mathbf{X}}(x) \lambda_{\mathbf{X}}(\tilde{x})}\), we obtain

\[
K_{\mathbf{X}}(r) = \frac{1}{|A|} E \left[ \sum_{x \in \mathbf{X} \cap A} \frac{1}{\lambda_{\mathbf{X}}(x)} \sum_{\tilde{x} \in \mathbf{X}, \tilde{x} \neq x} \frac{1_{\{d(x, \tilde{x}) \leq r\}}}{\lambda_{\mathbf{X}}(\tilde{x})} \right]
\]  \hspace{1cm} (2.8)

\[
= \frac{1}{|A|} \int \int 1_{\{x \in A, d(x, \tilde{x}) \leq r\}} \frac{\lambda_{\mathbf{X}, \mathbf{X}}^2(x, \tilde{x})}{\lambda_{\mathbf{X}}(x) \lambda_{\mathbf{X}}(\tilde{x})} dx d\tilde{x}
\]  \hspace{1cm} (2.9)

Based on the extended Slivnyak-Mecke Theorem [Moller and Waagepetersen, 2003], it can be shown that for a Poisson point process \(\mathbf{X}\) with intensity function \(\lambda_{\mathbf{X}}(\cdot)\) the second order product density \(\lambda_{\mathbf{X}, \mathbf{X}}^2(\xi, \eta) = \lambda_{\mathbf{X}}(\xi) \lambda_{\mathbf{X}}(\eta)\). Using this result, we can compute the \(K\)-function for a Poisson process \(\mathbf{X}\)

\[
K_{\mathbf{X}}(r) = \frac{1}{|A|} E \left[ \sum_{x \in \mathbf{X} \cap A} \frac{1}{\lambda_{\mathbf{X}}(x)} \sum_{\tilde{x} \in \mathbf{X}, \tilde{x} \neq x} \frac{1_{\{d(x, \tilde{x}) \leq r\}}}{\lambda_{\mathbf{X}}(\tilde{x})} \right]
\]

\[
= \frac{1}{|A|} \int \int 1_{\{x \in A, d(x, \tilde{x}) \leq r\}} \frac{\lambda_{\mathbf{X}, \mathbf{X}}^2(x, \tilde{x})}{\lambda_{\mathbf{X}}(x) \lambda_{\mathbf{X}}(\tilde{x})} dx d\tilde{x}
\]

\[
= \frac{1}{|A|} \int \int 1_{\{x \in A, d(x, \tilde{x}) \leq r\}} \frac{\lambda_{\mathbf{X}}(\xi) \lambda_{\mathbf{X}}(\eta)}{\lambda_{\mathbf{X}}(x) \lambda_{\mathbf{X}}(\tilde{x})} dx d\tilde{x}
\]

\[
= \frac{1}{|A|} \int \int 1_{\{x \in A, d(x, \tilde{x}) \leq r\}} dx d\tilde{x} = \frac{|A|}{|A|} \pi r^2 = \pi r^2
\]

One of the important properties of the \(K\)-function is that it is invariant under independent thinning as first introduced in Section 1.2. In brief, we have a function \(p(\xi) : \mathbb{R}^2 \to [0, 1]\), which we can think of as a probability surface, and a point process \(\mathbf{X}\) with intensity \(\lambda_{\mathbf{X}}(\cdot)\) and second order product density \(\lambda_{\mathbf{X}, \mathbf{X}}^2(\cdot, \cdot)\). For any point in our point process, we draw a Bernoulli random variable with probability of success \(p(\xi)\). If the outcome is 1, then we retain the point. If
the outcome is 0, then we delete the point. The resulting $X_{\text{thin}}$ will be a point process with intensity $\lambda_{X_{\text{thin}}} (\xi) = p(\xi) \lambda (\xi)$ and second order product density $\lambda^2_{X_{\text{thin}}, X_{\text{thin}}} (\xi, \eta) = p(\xi)p(\eta) \lambda^2_{X, X}(\xi, \eta)$ [Moller and Waagepetersen, 2003].

It follows that the $K$-function for $X_{\text{thin}}$ is

$$K_{X_{\text{thin}}}(r) = \frac{1}{|A|} \int \int 1_{\{d(x, \tilde{x}) \leq r\}} \frac{\lambda^2_{X_{\text{thin}}, X_{\text{thin}}}(x, \tilde{x})}{\lambda_{X_{\text{thin}}}(x) \lambda_{X_{\text{thin}}}(\tilde{x})} \, dx \, d\tilde{x}$$

$$= \frac{1}{|A|} \int \int 1_{\{d(x, \tilde{x}) \leq r\}} \frac{p(x)p(\tilde{x}) \lambda^2_{\tilde{X}, X}(x, \tilde{x})}{p(x) \lambda_X(x) p(\tilde{x}) \lambda_X(\tilde{x})} \, dx \, d\tilde{x}$$

$$= \frac{1}{|A|} \int \int 1_{\{d(x, \tilde{x}) \leq r\}} \frac{\lambda^2_{\tilde{X}, X}(x, \tilde{x})}{\lambda_X(x) \lambda_X(\tilde{x})} \, dx \, d\tilde{x}$$

$$= K_X(r)$$

### 2.1.1 Cross $K$-function

So far we have discussed properties of one point process $X$. In this section we discuss the interaction of two point processes. Consider point processes $X$ and $Y$ on $\mathbb{R}^2$. If they obey cross second order intensity reweighted stationarity, we can describe the tendency of points in one process to repel or attract points in the other process by using the cross $K$-function [Moller and Waagepetersen, 2003]

$$K_{X, Y}(r) = \frac{1}{|A|} E \left[ \sum_{x \in X \cap A} \frac{1}{\lambda_X(x)} \sum_{y \in Y} \frac{1_{\{d(x,y) \leq r\}}}{\lambda_Y(y)} \right]$$

(2.10)

Note that if the two processes are independent we have

$$K_{X, Y}(r) = \frac{1}{|A|} E \left[ \sum_{x \in X \cap A} \frac{1}{\lambda_X(x)} \sum_{y \in Y} \frac{1_{\{d(x,y) \leq r\}}}{\lambda_Y(y)} \right]$$

$$= \frac{1}{|A|} E \left[ \sum_{x \in X \cap A} \frac{1}{\lambda_X(x)} \right] E \left[ \sum_{y \in Y} \frac{1_{\{d(x,y) \leq r\}}}{\lambda_Y(y)} \right]$$

$$= \frac{1}{|A|} \int 1_{\{x \in X \cap A\}} \frac{\lambda_X(x)}{\lambda_X(x)} \, dx \times \int 1_{\{d(x,y) \leq r\}} \frac{\lambda_Y(y)}{\lambda_Y(y)} \, dy = \frac{|A|}{|A|} \pi r^2 = \pi r^2$$
2.2 Pair Correlation Function

The second order product density \( \lambda^2_{X,X}(\cdot, \cdot) \) was defined in Section 2.1. It was shown that for a Poisson point process \( X \) with an intensity function \( \lambda_X(\cdot) \) the second order product density is \( \lambda^2_{X,X}(\xi, \eta) = \lambda_X(\xi)\lambda_X(\eta) \). In order to study whether a point process \( X \) deviates from a Poisson point process, it is useful to normalize the second order product density \( \lambda^2_{X,X}(\cdot, \cdot) \) by dividing by \( \lambda_X(\xi)\lambda_X(\eta) \) [Moller and Waagepetersen, 2003]. If both \( \lambda^2_{X,X}(\cdot, \cdot) \) and \( \lambda_X(\cdot) \) exist, then the pair correlation function is defined as

\[
g(\xi, \eta) = \frac{\lambda^2_{X,X}(\xi, \eta)}{\lambda_X(\xi)\lambda_X(\eta)}.\]

If \( X \) is a Poisson point process, then the pair correlation function \( g(\xi, \eta) = 1 \). If \( g(\xi, \eta) > 1 \), this is evidence that a pair of points are more likely to occur jointly at the locations \( \xi \) and \( \eta \) than for a Poisson process with the same intensity function indicating attraction. If \( g(\xi, \eta) < 1 \), this is evidence that pair of points are less likely to occur jointly at the locations \( \xi \) and \( \eta \) than for a Poisson process with the same intensity function indicating repulsion. The pair correlation function \( g(\xi, \eta) \) is related to the \( K \)-function. Using equation 2.8, we can re-write the \( K \)-function in the following way:

\[
K_X(r) = \frac{1}{|A|} E \left[ \sum_{x \in \mathbb{R} \cap A} \frac{1}{\lambda_X(x)} \sum_{\tilde{x} \in X, \tilde{x} \neq x} \frac{1\{d(x, \tilde{x}) \leq r\}}{\lambda_X(\tilde{x})} \right]
\]

\[
= \frac{1}{|A|} \int \int 1_{\{x \in A, d(x, \tilde{x}) \leq r\}} \frac{\lambda^2_{X,X}(x, \tilde{x})}{\lambda_X(x)\lambda_X(\tilde{x})} dx \, d\tilde{x}
\]

\[
= \frac{1}{|A|} \int \int 1_{\{x \in A, d(x, \tilde{x}) \leq r\}} g(x, \tilde{x}) \, dx \, d\tilde{x}
\]

2.3 Inference for the \( K \)-function

In section 2.1, we introduced the \( K \)-function. In this section, we discuss some well established methods for its estimation and a new method we propose. We will compare these methods in terms of their performance and investigate the effect of size the available observed point patterns on the estimation of the \( K \)-function.

Intuitively, \( K_X(r) \) expresses the expected spatial accumulation of points in a neighborhood of radius \( r \) around a “typical” point of the process \( X \) adjusted for the intensity of the process itself. Thus, estimation of \( K_X(r) \) for any fixed \( r \) requires visiting each point in the observed point pattern \( x \), counting the number of unique
observed points within a radius $r$ of it, and averaging over the observed pattern. In doing this, we have to adjust for the size of the observation window $W$, the lack of knowledge about points outside $W$ (edge effects), and the intensity $\lambda_X(\cdot)$. In order to adjust for the size of $W$, we divide the counts by $|W|$.

The estimation of the intensity can be accomplished as follows. If we are willing to assume that a process is homogeneous, then an unbiased estimate of the intensity is $\hat{\lambda}_X = \frac{N_X(W)}{|W|}$, where $|W| = \int_W d\xi$ denotes the area of $W$. If we additionally assume that the process $X$ is a homogeneous Poisson process, then this estimate is also the maximum-likelihood estimate of the constant intensity [Moller and Waagepetersen, 2003](See Section 1.2.). Nonparametric estimation of a non-constant but smooth intensity function can be accomplished by kernel smoothing

$$\hat{\lambda}_X(\xi) = \sum_{i=1}^{n} \frac{k_b(\xi - x_i)}{e_{W,b}(x_i)} , \xi \in W$$ (2.11)

where $k(\cdot)$ is a $d$-dimensional density, $b > 0$ is the bandwidth, and $k_b(\xi) = \frac{k(\xi)}{b^d}$ defines a kernel. Note that while the choice of $k(\cdot)$ obviously affects the estimation, the choice of bandwidth $b$ is by far more consequential – as it determines the degree of smoothness of $\hat{\lambda}_X(\cdot)$ [Moller and Waagepetersen, 2003]. Bandwidth selection balances a trade-off between bias and variance in the estimation of $\lambda_X(\cdot)$. To illustrate the effect of bandwidth selection in the nonparametric estimation of the intensity, we consider the data in Figure 2.1 and compute the estimated intensity with formula 2.11 on a grid of points in this window. Figure 2.2 shows the estimates using a Gaussian kernel and bandwidths of 0.07, 0.13, and 0.4.

When the bandwidth is small (0.07), the estimated intensity becomes more variable. When the bandwidth is large (0.4), the surface is over-smoothed. The second panel from left to right, which has a bandwidth of 0.13, seems to provide a good balance between the two. When decreasing bandwidth, the variability of the estimated intensity will increase across $W$. If we used a bandwidth close to zero, the estimated intensity would be zero at locations without an observed point and large at locations with observed points with infinitely high variability. On the other hand, increasing the bandwidth reduces variability but introduces bias
Figure 2.1. Simulated point pattern from an inhomogeneous Poisson process with intensity \( \lambda_X(\xi) = 100 \frac{1}{\sqrt{2\pi0.21^2}} \exp\left(-\frac{(\xi - 0.5)^2}{2 \cdot 0.21^2}\right) \) \( \exp\left(-\frac{(\xi - 0.5)^2}{2 \cdot 0.21^2}\right) \)

Figure 2.2. 2D example of kernel smoothed intensity (the data is shown in Figure 2.1). The kernel is Gaussian, and the bandwidths used are from left to right: 0.07, 0.13 and 0.4. The true intensity is shown in on the right most graph.

in the estimated intensity. If the bandwidth is large enough, the estimate of the intensity will be essentially the same for any location \( \xi \in W \), \( \hat{\lambda}_X(\xi) \approx \frac{N_X(W)}{|W|} \). In this situation, the variation of our estimate is zero; however, we have a substantial bias problem unless the true intensity was homogeneous.

Another element of the formula 2.11 is \( e_{W,b}(\cdot) \), which is an edge-correction factor. The edge-correction factor reduces bias in \( \hat{\lambda}_X(\xi) \) at locations \( \xi \) near the boundary of the observation window \( W \). The bias is due to the fact that we lack observed points outside the window. For a location near the boundary of \( W \), the estimate will sum over fewer values, thus causing a systematic underestimation of the intensity. A common formulation of the correction factor is

\[
e_{W,b}(x) = \int_W k_b(w - x)dw \quad (2.12)
\]
Figure 2.3. 2D example of a point pattern within and outside an observation window. The figure illustrates the need for edge correction in estimating the intensity of the process.

The problem is illustrated in Figure 2.3. Assume that the point process is defined on a support represented by the whole square area in the plot (black perimeter) but observed only in the smaller square (blue perimeter). The circles represent the area of practical importance for the chosen kernel and bandwidth specification at two example locations (i.e. the kernel is nearly zero outside of this area). When we estimate the intensity at a location in the center of our observation window, there are more points over which to sum (green circle). When we estimate the intensity at a location near the boundary of the observation window, there are fewer points to sum over (blue circle) — leading to systematic underestimation of the intensity.

Edge effects were illustrated in Figure 2.3 for the intensity but concern the $K$-function as well. When counting neighbors of a point near the boundary of our observation window, we will systematically undercount, because we lack any knowledge about points generated by the process outside $W$. In order to correct for this bias, different edge corrections have been introduced in the literature [Moller and Waagepetersen, 2003]. Most of them give higher weight to points near the
boundary of \( W \), thus making up for the lack of points outside.

One popular choice is the so-called translation edge-correction factor defined as

\[
|W \cap W_{x_i-x_j}|
\]

\( W_{x_i-x_j} \) is the window translated by \( x_i-x_j \), where \( x_i, x_j \) are different points in the observed point pattern \( x \). This factor is used in place of \(|W|\) when dividing the counts, so that the formula for estimation of \( \hat{K}_X(r) \) becomes

\[
\hat{K}_X(r) = \sum_{i=1}^{n_X} \frac{1}{\hat{\lambda}_X(x_i)} \sum_{j \neq i, 1}^{n_X} \frac{1_{\{d(x_i,x_j) \leq r\}}}{\hat{\lambda}_X(x_j)|W \cap W_{x_i-x_j}|}
\]  

(2.13)

In the above expression, the estimated intensity \( \hat{\lambda}_X(\cdot) \) also contains an edge correction factor as previously shown in equation 2.12.

An alternative approach to adjust for edge effects in the estimation of the \( K \)-function is to create a “buffer” away from the boundary of \( W \). The outer sum (averaging) required in the calculation of \( \hat{K}_X(r) \) will be restricted to points \( x_i \) that are at least \( R > 0 \) away from the boundary, i.e. that belong to a smaller set \( W_R \subset W \). The size of the buffer, and thus \( W_R \), is determined by how far we want to increase \( r \) in the estimation of the \( K \)-function. As an illustration, in Figure 2.3 we can think of the larger square as the original window of observation \( W \) and the smaller blue square as the buffered window \( W_R \). With this buffering we will not need to correct for edge effects whenever \( r \leq R \), because for all points in \( W_R \) we can observe a full neighborhood of radius \( r \). If we use this approach, then the formula for the estimation of the \( K \)-function becomes

\[
\hat{K}_X(r) = \frac{1}{|W_R|} \sum_{i: x_i \in W_R} \frac{1}{\hat{\lambda}_X(x_i)} \sum_{j \neq i, 1}^{n_X} \frac{1_{\{d(x_i,x_j) \leq r\}}}{\hat{\lambda}_X(x_j)}
\]  

(2.14)

with the estimated intensity \( \hat{\lambda}_X(\cdot) \) still containing its own edge correction factor.

As shown in the above formulas, the computation of \( \hat{K}_X(r) \) requires estimates of the intensity \( \lambda_X(\cdot) \) at all points \( x_i \) in the observed pattern, whatever the approach to edge correction. These may be obtained by kernel smoothing methods, as discussed above, but some caution should be exercised. First, intensity estimation is influenced by the selection of bandwidth; ultimately we want the estimation procedure for \( \hat{K}_X(r) \) to be robust to this. Also, [Baddeley et al., 2000] suggested using a leave-one-out approach when estimating the intensity at each of the ob-
served points, since it reduces bias. In this approach, the estimate at each observed point is calculated ignoring the point itself. The authors show that by not including the point the estimate of the intensity is less biased for inhomogeneous Poisson processes. They argue that this is also the case for other point processes with attractive forces (pair correlation function $\geq 1$), while it is not completely clear for processes with repulsive forces (pair correlation function $\leq 1$). Intuitively, if we consider the formula for intensity estimation (equation 2.11), we notice that including the point at which $\hat{\lambda}_X(\cdot)$ is calculated carries the largest term only because we are estimating the intensity at the point. Gaussian or other unimodal kernels will place the highest weight at the point itself.

The new method we propose for estimating the $K$-function is based on the assumption that our intensity function(s) can be thought of as approximately locally constant. We call it the circle method. Note that the calculation of $\hat{K}_X(r)$ additively accumulates “contributions” from each of the points in the $x$ pattern; in symbols, using the buffer approach for edge correction we can rewrite $\hat{K}_X(r) = \frac{1}{|W_R|} \sum_{i=1}^{n_X} s_i(r)$, where

$$s_i(r) = \frac{1}{\hat{\lambda}_X(x_i)} \sum_{j=1}^{n_X} \frac{1 \{d(x_i, x_j) \leq r\}}{\hat{\lambda}_X(x_j)}$$ (2.15)

If we focus on a ball $B_{R,i}$ of sufficiently small radius $R$ around $x_i$ and consider any radius $r \leq R$, we have that all of the $\lambda_X(x_j)$’s for $x_j$ such that $d(x_i, x_j) \leq r$ and $\lambda_X(x_i)$ itself are equal, since we assume that the intensity of $X$ is locally constant. Their common value can be estimated based on the number of $x$ points in the ball around $x_i$ as $\frac{1}{|B_{R,i}|} \sum_{k=1}^{n_X} 1 \{d(x_i, x_k) \leq R\}$, where $|B_{R,i}| = \pi R^2$.

In summary, the formula we will use is

$$\hat{K}_X(r) = \frac{\pi R^2}{|W_R|} \sum_{i:x_i \in W_R} \frac{1}{\left(\sum_{k=1}^{n_X} 1 \{d(x_i, x_k) \leq R\}\right)^2} \sum_{j \neq i,j=1}^{n_X} 1 \{d(x_i, x_j) \leq r\}$$ (2.16)

The drawback of this method is that it makes an additional assumption of local homogeneity of the intensity, which will not hold true in practice. On the other hand, the advantage of this method is speed of computation and intuitive appeal.
Table 2.1 provides a summary of all the formulas introduced above.1

<table>
<thead>
<tr>
<th>n: Type</th>
<th>Estimate K-function</th>
<th>Estimate Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) leave-one-out-No,Edge-Yes</td>
<td>$K_X(r) = \sum_{i=1}^{n} \frac{1}{\lambda_X(x_i)} \sum_{j \neq i, 1}^{\infty} \frac{1_{d(x_i, x_j) \leq r}}{\lambda_X(x_j)</td>
<td>W \cap W_{x_i} - x_j</td>
</tr>
<tr>
<td>2) leave-one-out-Yes,Edge-Yes</td>
<td>$K_X(r) = \sum_{i=1}^{n} \frac{1}{\lambda_X(x_i)} \sum_{j \neq i, 1}^{\infty} \frac{1_{d(x_i, x_j) \leq r}}{\lambda_X(x_j)</td>
<td>W \cap W_{x_i} - x_j</td>
</tr>
<tr>
<td>3) leave-one-out-No,Edge-No(R)</td>
<td>$K_X(r) = \sum_{i=1}^{n} \frac{1}{\lambda_X(x_i)} \sum_{j \neq i, 1}^{\infty} \frac{1_{d(x_i, x_j) \leq r}}{\lambda_X(x_j)</td>
<td>W \cap W_{x_i} - x_j</td>
</tr>
<tr>
<td>4) leave-one-out-Yes,Edge-No(R)</td>
<td>$K_X(r) = \sum_{i=1}^{n} \frac{1}{\lambda_X(x_i)} \sum_{j \neq i, 1}^{\infty} \frac{1_{d(x_i, x_j) \leq r}}{\lambda_X(x_j)</td>
<td>W \cap W_{x_i} - x_j</td>
</tr>
<tr>
<td>5) Circle(proposed)</td>
<td>$K_X(r) = \sum_{i=1}^{n} \frac{1}{\lambda_X(x_i)} \sum_{j \neq i, 1}^{\infty} \frac{1_{d(x_i, x_j) \leq r}}{\lambda_X(x_j)</td>
<td>W \cap W_{x_i} - x_j</td>
</tr>
</tbody>
</table>

2.3.1 Null Envelopes for K-functions

In this section we describe how to obtain a reference, or null interval, for $K_X(r)$ and $K_{X, Y}(r)$ at any fixed value of $r$. Note that we can use an analogous approach for obtaining confidence intervals after assuming a particular point process model and estimating its parameters. This would be based on simulations as described below, and thus akin to a parametric bootstrap. If we are able to simulate from the distribution of the null hypothesized point process $X_o$, we can proceed as follows. Simulate independent and identically distributed point patterns from the distribution of the point process $X_o$: $x_i \sim X_o; i = 1, \ldots, B$. For each point pattern $x_i$, estimate the $K$-function. Now for any fixed $r$ we have $B$ values, and we can take the spread between the 2.5% and the 97.5% percentiles of such values as a 5% level null interval. An estimated $K_{obs}(r)$ outside such interval is evidence against $X_o$, in the sense that $X$ does not behave consistently with $X_o$ in terms of repulsion/attraction among points at range $r$. The estimate on the original observed point pattern is given by $K_{obs}$. When it is computationally intensive.

1Note that when using a “buffer” to deal with edge effects (Edge-No) we consider fixing the size of the buffer in absolute terms (R; (3) and (4)). Also, in the Circle method, the size of the buffer and that of the ball constructed around each point are taken to be equal (R).
to calculate the $K$-function or to simulate from the distribution of $X_0$, we can limit the number of simulations based on the desired level. If we draw $B$ point patterns, we will have

$$P \left( \hat{K}_{\text{obs}}(r) < \min_{i=1,...,B} \hat{K}_i(r) \right) = P \left( \hat{K}_{\text{obs}}(r) > \max_{i=1,...,B} \hat{K}_i(r) \right) \leq \frac{1}{B+1}$$

[Moller and Waagepetersen, 2003] where $\hat{K}_i(\cdot)$ indicates the estimate on the i-th simulated point pattern. Thus, for $B = 39$ the minimum and maximum provide the 2.5%-lower and the 97.5%-upper limits for the null interval. When $B = 99$, the minimum and maximum provide the 1%-lower and 99%-upper limits, etc. Repeating these calculations for several values of $r$ provides a null envelope for the $K$-function. It should be stressed that we have many null intervals/tests (one for each $r$), and they are all computed on the same set of simulations. Therefore, the null envelope we compute for the $K$-function does not control the overall level at 5% or 2%.

Now, let us consider the cross $K$-function. If we are able to simulate from the joint distribution of the null hypothesized joint point process $\{X_0,Y_0\}$, we can proceed as follows. Simulate independent and identically distributed point patterns $\{x_i,y_i\} \sim \{X_0,Y_0\}$, $i = 1,...,B$. For each pair of point patterns $\{x_i,y_i\}$, estimate the cross $K$-function. Now for any fixed $r$, we have $B$ values and can take the spread between the 2.5% and the 97.5% percentiles of such values as a 5% level null interval; an estimated $\hat{K}_{\text{obs,obs}}(r)$ outside such interval is evidence against $\{X_0,Y_0\}$ in the sense that $X,Y$ does not behave consistently with $\{X_0,Y_0\}$ in terms of repulsion/attraction among points at range $r$.

### 2.3.2 Simulation Study for Evaluating Estimators of $K_X(r)$

In order to find which of the seven approaches listed in Table 2.1 performs best, we conduct a simulation study. This is accomplished by looking at the mean estimated $K$-function from 39 point patterns simulated from the distribution of a Poisson process and observing which approach produces mean estimates closest to the true value of $\pi r^2$. In particular, we simulate from the distributions of three different Poisson point processes with intensities shown in Figure 2.4. We choose
a homogeneous intensity, a bell shaped inhomogeneous intensity with a low level of inhomogeneity, and a bell shaped inhomogeneous intensity with a high level of inhomogeneity.\(^2\) On average all of the intensities produce around 6900 points in the observation window \(W\), which is a circle centered at \((1780, 1630)\) with a radius of 1350. These average abundancies and observation window are selected to match features of the actual data application described in Chapter 3. The maximum value of \(r\) we consider in the estimation is 50, so we buffer by \(R = 50\).

![Figure 2.4](image.png)

**Figure 2.4.** Left: Homogeneous intensity; Middle: Inhomogeneous intensity with low spatial variation; Right: Inhomogeneous intensity with high spatial variation.

Results are shown in Figures 2.5, 2.6, and 2.7 for the homogeneous, low inhomogeneous and high inhomogeneous cases, respectively. The different curves in each figure panel represent the mean \(K\)-function from 39 simulated point patterns, computed with different bandwidth for the estimation of the intensity. The chosen bandwidths are 50, 100, 150, 200, 250, and 300. We also subtract \(\pi r^2\) to better visualize the deviations; thus, good performance corresponds to curves close to the horizontal zero line. The number reported in parenthesis in the legend of each figure panel is the number of points from one of the 39 simulated point patterns in the first summation of the \(K\)-function estimation procedure (i.e. the number after the “buffering” has been applied; if no buffering is applied then it is just the number of points in the simulated point pattern).

After visually inspecting the graphs we notice that the approach that consistently performs best is the one displayed on the second row on the right. This

\[^2\text{Note that we estimate the } K\text{-function of a homogeneous process using an inhomogeneous assumption. This is done only for illustrative purposes. If we assume a homogeneous intensity then a better unbiased estimate of the intensity in the } K\text{-function estimator is } \hat{\lambda}_X^2 = \frac{N_X(W)(N_X(W)-1)}{|W|^2}\]
corresponds to equation (4) in Table 2.1, which employs a kernel estimate with
leave-one-out, a typical edge correction factor (equation 2.12) for the intensity,
and “buffering” to deal with edge effects in the K calculation. The edge correction
method for the K-function seems to perform well for the homogeneous case but
produces a noticeable bias (downward curvature) at all bandwidth considered for
the inhomogeneous low, the inhomogeneous high, for the leave-one-out case, and
the case without leave-one-out. The estimate based on the “buffering” method
performs well in the homogeneous case for almost all of the bandwidths considered
using the leave-one-out adjustment. This estimate does produce some upward cur-
vature (bias) for the inhomogeneous low case at some bandwidths and performs
well for the inhomogeneous high case. The leave-one-out adjustment does play
an important role in the estimation. Looking at each row, the estimate on the
right using the leave-one-out adjustment is closer to $\pi r^2$. The proposed circle
method performs well for the homogeneous case; however, it is highly sensitive to
the selection of $R$ and thus unreliable for inhomogeneous low and inhomogeneous
high.

### 2.3.3 Effect of Size of Available Observed Point Patterns
on K-function Estimation

After establishing which estimation approach to use we will investigate the effect of
having more, or fewer observed points in the estimation of the K-function. We will
rescale the intensities in Figure 2.4 as to generate on average 350 points, 750 points,
1000 points, 1500 points, 2500 points, 6900 points, and 69000 points. Once again,
these settings are relevant for the actual data application described in Chapter 3.
We will use bandwidths of 50, 150, 250 and will produce confidence envelopes from
99 simulations in each of the three intensity scenarios defined in Section 2.3.2. The
results are shown in Figures 2.8, 2.9, and 2.10. We observe that the mean from
the 99 simulations is close to $\pi r^2$ for sufficiently large bandwidth (150 and 250)
even when the number of points used in the estimation is small. A sufficiently large
bandwidth makes bias manageable even at small sample sizes; however, the width
of the confidence envelopes is large. The variability of the K-function estimator
decreases as the number of points available for estimation increases.
2.3.4 Effect of Size of Available Observed Point Patterns on Cross K-function Estimation

In order to estimate the cross $K$-function (multivariate $K$-function), we can use an estimate similar to the one selected in Section 2.3.2 for the univariate $K$-function. Given observed point patterns $x = \{x_1, \ldots, x_{n_x}\}$ and $y = \{y_1, \ldots, y_{n_y}\}$, we compute

$$
K_{x,y}(r) = \frac{1}{|W_R|} \sum_{i: x_i \in W_R} \frac{1}{\lambda_X(x_i)} \sum_{j=1}^{n_y} \frac{1_{\{d(x_i, y_j) \leq r\}}}{\lambda_Y(y_j)}
$$

(2.17)

where the estimates of the intensities are obtained using kernel smoothing methods with leave-one-out adjustment and edge correction factor, as defined in equation 2.12. For dealing with edge effects in the estimation of the $K$-function we use “buffering”.

In order to assess how the available number of points from the two processes affects this estimate, we proceed along the same lines as in Section 2.3.2. We simulate $X$ and $Y$ from the distribution of Poisson point processes, independently, with three different intensities observed on the window $W$, which is a circle centered at $(1780, 1630)$ with a radius of 1350. We scale these intensities to obtain a range of configurations for the expected number of points from $X$ and $Y$. The configurations are selected because of relevance to the actual data analysis described in Chapter 3. For each intensity scenario and configuration, we simulate 99 independent pairs of $X_i$ and $Y_i$, $i = 1, \ldots, 99$ point patterns on which to compute the cross-$K$ estimate. We use bandwidths of 50, 150, 250. Results for homogeneous, low inhomogeneous, and high inhomogeneous intensities are shown in Figures 2.11, 2.12, and 2.13. We observe that for a large enough bandwidth and for each underlying intensity the estimates perform well even with small sample sizes.
Figure 2.5. Evaluating estimators of $K_X(r)$: Homogeneous Intensity. Upper panels: Estimated mean of $K$-functions from 39 simulations using formulas (1) (left), (2) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Estimated mean of $K$-functions from 39 simulations using formulas (3) (left), (4) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. Lower panel: Estimated mean of $K$-functions from 39 simulations using formula (5) from Table 2.1 with $R$ equal to 50, 100, 150, 200, 250, and 300.
Figure 2.6. Evaluating estimators of $K_X(r)$: Inhomogeneous Low Intensity. Upper panels: Estimated mean of $K$-functions from 39 simulations using formulas (1) (left), (2) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Estimated mean of $K$-functions from 39 simulations using formulas (3) (left), (4) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. Lower panel: Estimated mean of $K$-functions from 39 simulations using formula (5) from Table 2.1 with $R$ equal to 50, 100, 150, 200, 250, and 300.
Figure 2.7. Evaluating estimators of $K_X(r)$: Inhomogeneous High Intensity. Upper panels: Estimated mean of $K$-functions from 39 simulations using formulas (1) (left), (2) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Estimated mean of $K$-functions from 39 simulations using formulas (3) (left), (4) (right) from Table 2.1 with bandwidths 50, 100, 150, 200, 250, and 300. Lower panel: Estimated mean of $K$-functions from 39 simulations using formula (5) from Table 2.1 with R equal to 50, 100, 150, 200, 250, and 300.
<table>
<thead>
<tr>
<th>$E[N_X(W)]$</th>
<th>Band = 50</th>
<th>Band = 150</th>
<th>Band = 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>750</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>1500</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
</tr>
<tr>
<td>2500</td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
<tr>
<td>6900</td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
<tr>
<td>69000</td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Figure 2.8.** Effects of size of available observed point patterns on $K$-function Estimation: Homogeneous Intensity. The expected number of points generated from the point process are shown on the right column. The estimate of the $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
<table>
<thead>
<tr>
<th>$E[N_X(W)]$</th>
<th>Band = 50</th>
<th>Band = 150</th>
<th>Band = 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>750</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>1500</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
</tr>
<tr>
<td>2500</td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
<tr>
<td>6900</td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
<tr>
<td>69000</td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Figure 2.9.** Effects of size of available observed point patterns on $K$-function Estimation: Inhomogeneous Low Intensity. The expected number of points generated from the point process are shown on the right column. The estimate of the $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
Figure 2.10. Effects of size of available observed point patterns on $K$-function Estimation: Inhomogeneous High Intensity. The expected number of points generated from the point process are shown on the right column. The estimate of the $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
<table>
<thead>
<tr>
<th>$E[N_X(W)]$</th>
<th>$E[N_Y(W)]$</th>
<th>Band = 50</th>
<th>Band = 150</th>
<th>Band = 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>350</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>350</td>
<td>17000</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>800</td>
<td>10000</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
</tr>
<tr>
<td>1000</td>
<td>9000</td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
<tr>
<td>1500</td>
<td>8000</td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
<tr>
<td>1600</td>
<td>9000</td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Figure 2.11.** Effects of size of available observed point patterns on cross $K$-function Estimation: Homogeneous Intensity. The expected number of points generated from the point processes are shown on the first and second columns to the right. The estimate of the cross $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
<table>
<thead>
<tr>
<th>$E[N_X(W)]$</th>
<th>$E[N_Y(W)]$</th>
<th>Band = 50</th>
<th>Band = 150</th>
<th>Band = 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>350</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>350</td>
<td>17000</td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>800</td>
<td>10000</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
</tr>
<tr>
<td>1000</td>
<td>9000</td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
<tr>
<td>1500</td>
<td>8000</td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
<tr>
<td>1600</td>
<td>9000</td>
<td><img src="image16" alt="Graph" /></td>
<td><img src="image17" alt="Graph" /></td>
<td><img src="image18" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Figure 2.12.** Effects of size of available observed point patterns on cross $K$-function Estimation: Inhomogeneous Low Intensity. The expected number of points generated from the point processes are shown on the first and second columns to the right. The estimate of the cross $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
The blue lines correspond to the minimum and maximum. The red line represents the estimate of the cross function with confidence envelopes from 99 simulations with

### Figure 2.13.

Effects of size of available observed point patterns on cross $K$-function Estimation: Inhomogeneous High Intensity. The expected number of points generated from the point processes are shown on the first and second columns to the right. The estimate of the cross $K$-function with confidence envelopes from 99 simulations with bandwidths 50, 150, 250 are shown on the left, middle and right panels, respectively. The blue lines correspond to the minimum and maximum. The red line represents the average.
Chapter 3

Application of Exploratory Spatial Point Process Methods to RSVA In Vitro Infection Experiments

3.1 Description of Experiments and Data

In this chapter, we present an application of the statistical methods introduced in the first two chapters to real data. Analyzing this data will shed light on the spatial dynamics of an RSVA infection. We start with a description of the data and then discuss results of our analysis.

While experiments were also run with two strains of RSV (A and B) to investigate co-infections, here we consider homologous infections with one strain of the virus. In these homologous infection experiments, which are referred to as 2A throughout the chapter, wells plated with epithelial cells are left standing for one hour plus a lag period of $h = 0, 3, 6, 9, 12$ or 15 hours then exposed to the RSVA virus at large concentrations (MOI=1, corresponding to infection of an expected 100% of the cells) for 24 hours. After this lag, the wells are cleared and imaged through microscopy. Several independent experiments are run on separate wells for each lag duration providing 6 replicates per lag.

Fluorescent tags and dyes in the specimen allow us to identify cell nuclei and cells infected with RSVA (through cytoplasmic stains) in the well images. For
Figure 3.1. Microscopy imaging of a cell culture infected with one strain of RSV. Blue marks represent cell nuclei and green marks are located in the cytoplasm of cells infected with RSV A.

Illustration, Figure 3.1 shows marks in a selected region of one well image. Cytoplasmic stains have a notably different appearance than the nuclei, and the underlying shape and size of cells and volume exclusion among them creates constraints on the location of the stains.

Using image analysis software [Image-Pro Plus 6.3, 2007] with appropriately tuned parameters, we can produce aggregate counts of marks (for nuclei and cytoplasmic stains) as well as 2D spatial coordinates for each mark—more precisely for the center of each distinct mark identified by the software.

Aggregate counts at various lags (Figure 3.2) illustrate some features of the experimental system we use and suggest notable changes in susceptibility. First, the number of nuclei (cells in the wells) systematically increases with lag (Figure 3.2, left panel) suggesting that cells in the tissue cultures continue to divide over the duration of our experiments—a longer lag implies a longer total experiment time. Second, the number of A marks (a proxy for cells infected with RSVA; right panel) decreases, suggesting a decrease in susceptibility. Importantly, the decrease in susceptibility may be due to intrinsic cell properties that change as the number of cells in a well increases.

One approach to further investigate the effects on susceptibility is to analyze
Figure 3.2. Counts of nuclei (left), and A marks (right) as a function of lag. Different points represent different replicate experiments at each lag duration, and color coding is consistent across the two plots – thus, for instance, the light blue points represent well replicate 5 at lag $h = 12$, which has the highest number of nuclei, and the lowest number of A stains.

...their spatial structure using the information we have on the locations of nuclei and A marks within each well.

Notably, nuclei marks reveal that cells populate wells quite unevenly, in addition to increasing in number over time. The observed inhomogeneities are likely a consequence of the way the tissue cultures are plated, and they vary in a non-systematic way from well to well. Figure 3.3 shows kernel estimates of the nuclei intensities (i.e. roughly speaking, the local average number of cells per area; see Section 1.2) for three wells as instances. Bull’s-eye patterns as well as concentration in regions near the rim of the wells are clearly visible. Inhomogeneous intensity in the tissue cultures, while not of interest per se, may indeed create diverse local environments for the cells as well as affect our ability to detect repulsion or attraction among our infection marks.

3.2 Modeling a Null Scenario

The microscopy imaging data at our disposal poses some challenges detecting spatial association among infected cells. The software does not identify cells in space and label them as infected (or not infected) with the RSVA strain; it only provides...
Figure 3.3. Kernel estimates of the nuclei intensities for well replicate 3 at $h = 0$ (left), well replicate 2 at $h = 9$ (middle), and well replicate 5 at $h = 15$ (right). The bandwidth was set to $b = 250$.

locations of nuclei and cytoplasmic stains (see Figure 3.1). One could pre-process the data to come closer to the ideal format by introducing ad hoc rules to attribute A marks to nuclei (e.g. to the closest nucleus within a given distance) and to deal with ambiguous cases (e.g. eliminating, or randomizing the attribution of, marks that are almost equally close to more than one nucleus). However, this type of pre-processing is necessarily arbitrary in nature and not guaranteed to reflect the cells’ shape or size and volume exclusion constraints which affect the locations of our marks. In addition, we have evidence that our tissue cultures create non-uniform cell supports upon which the infections act (see Figure 3.3). As an alternative to pre-processing the data we have into the format we would like, our approach here is to design simulation procedures that generate data in the same format as the ones we have (i.e. locations of A marks), but under a null scenario that serves as a benchmark to assess spatial association. In particular, these procedures need to reproduce non-uniformity in cell support and to reflect the constraints on locations of cytoplasmic stains imposed by the cells’ shape or size and volume exclusion.

We consider each well image separately; this is a disc-shaped observation window in 2D with center coordinates $(1780, 1630)$ and radius 1350 in pixels (1 pixel corresponding to 6.45 microns length). Within such a window, we have observed point patterns for two point processes, i.e. 2D coordinates of nuclei and A marks. To assess repulsion or attraction among the A marks, we contrast the observed A point patterns with “no-association” patterns simulated with the following procedures.
3.3 Simulation Procedures

Simulating the cell support: We start with the observed nuclei in a well – as produced by the imaging software [Image-Pro Plus 6.3, 2007]. For each nucleus in turn, we independently draw a random radius from a uniform distribution between 1.5 and 2.25 pixels and create a disc centered at the nucleus – resulting in a collection of spherical cells with diameters uniformly distributed between 19.35 and 29.025 microns (consistent with observed cell sizes). However, the cells may overlap – more often so the more densely populated is the well. When a newly generated cell (disc) induces one or more overlaps with existing cells (discs), we shift its center as to eliminate them. Note that simulating cells and shifting them as needed to avoid overlaps is crucial for subsequent simulation of cytoplasmic stains to ensure that any significant repulsion or attraction detected at small scales for A marks is not an artifact of the cells shape or size and volume exclusion in our cell cultures.

Figure 3.4 illustrates how this simulation approach results in a plausible support of spherical cells whose spatial arrangement reproduces closely the one of the original cells in a well. We compare $\hat{K}_N(r)$ for the observed nuclei and $\hat{K}_C(r)$ for the simulated cell centers (upper panels) and show kernel estimated intensities for both observed nuclei (mid panels) and simulated cell centers (lower panels) for two example wells (all other wells showed similar behaviors). We also explored the option of selecting a random location within each simulated cell (instead of the center) to provide simulated cell nuclei, but the simulations were indistinguishable. Note that the match is less accurate but still satisfactory for the well on the right, which has a much higher overall number of cells (this is a replicate for $h = 15$). In particular, inaccuracies may be due to the actual cells not being spherical, to the existence of cells with diameters in excess of the range covered in the simulations, etc. – factors that are more consequential in more densely populated wells.

Note also that the behavior of $\hat{K}_N(r)$ in both example wells supports the notion of not using a Poisson scenario as reference for our analyses. In the upper panels of Figure 3.4 the estimated $K$ functions are shown after subtracting $\pi r^2$, so if our cell nuclei behaved as an inhomogeneous Poisson process, we would see curves close to the horizontal line at 0. This is not the case. Our curves show a distinct “dip” at $r$
approximately equal to $3 - 4$, which may indeed capture the cells’ volume exclusion (resulting in a non-Poisson-like repulsion at such scales). However, at larger scales, our nuclei are consistent with an inhomogeneous Poisson point accumulation.

**Simulating A marks:** To account for volume exclusion phenomena, we used two alternative simulation paradigms for A marks, as described below:

1. **Paradigm 1:** Once a plausible cell support is in place, we go through the cells one by one and “infect” each, independently, with probability $\frac{n_A}{n_N}$ ($n_N$ and $n_A$ are, respectively, the number of observed nuclei and A marks). We then place an A mark in the center of each infected cell (disc). These are the centers of simulated cytoplasmic stains. We also explored the option of selecting a random point within each infected cell, but the simulations were indistinguishable from those with A marks placed in the centers. This procedure is repeated as to generate a total of 99 simulated A marks patterns

$$\tilde{a}^{(s)} = \{\tilde{a}_1^{(s)}, \ldots, \tilde{a}_{n_A}^{(s)}\}, \quad s = 1, \ldots, 99$$

each with an expected number of marks $E(\hat{n}_A^{(s)}) = n_A$ and lacking any systematic spatial structure except for that implicitly imposed by the cell support. These simulated patterns are then used to benchmark the observed A marks pattern $a = \{a_1, \ldots, a_{n_A}\}$. The simulated A marks represent an independent random thinning of the simulated cell nuclei, which reproduce quite closely the actual cell nuclei (see Figure 3.4).

2. **Paradigm 2:** While Paradigm 1 allows us to reproduce the dip in the $K$-function at ranges $r = 3 - 4$ (cell volume exclusion), our analysis showed a deeper dip at slightly larger ranges ($r = 7 - 8$) consistent across all replicates and lags (Figure 3.6). While this might indicate a short range repulsion effect (decrease in susceptibility to RSVA), we have reasons to suspect that it may be due to artifacts, in particular the imaging of cytoplasmic stains. Even with careful parameter tuning, blurring of stains and/or merging of two nearby blurred stains translates into detected cytoplasmic marks that tend to be larger than the typical cell size. In other words, limits in the imaging software may induce a “pseudo” exclusion at scales larger than the cell volume.
Figure 3.4. Upper panels: Estimated $K$-functions for observed nuclei, $\hat{K}_N(r)$ (black), and simulated cell centers, $\hat{K}_C(r)$ (red). The Poisson behavior ($\pi r^2$) is subtracted from each of these curves to help visualizing their variation. Mid panels: Kernel estimates of observed nuclei intensities with bandwidth $b = 250$. Lower panels: Kernel estimates of simulated cell centers intensities with bandwidth $b = 250$. Well replicate 4 at $h = 0$ and well replicate 5 at $h = 15$ are shown on the left and the right, respectively.
exclusion. The image analysis software produces areas for cytoplasmic stains from which we can approximate radii as $r = \sqrt{\frac{\text{Area}}{\pi}}$. Kernel smoothed densities of these radii for two example wells are shown in Figure 3.5. Simulation paradigm 2 follows the same steps as paradigm 1 except that instead of using the exclusion from the simulated cell boundaries we used randomly drawn radii from the kernel smoothed density of the actual A stain radii computed for a well. More specifically, we did not allow two simulated A marks, say $i$ and $j$, to have a distance less than $r_i + r_j$ (i.e. the sum of two randomly drawn radii) from one another. The kernel smoothed density of the actual A stain radii can and should be trimmed in order to remove anomalously large stains. Drawing radii from this trimmed distribution gives us better results as shown in Figure 3.7.

![Kernel smoothed density of estimated radii of A stains. Well replicate 4 at $h = 0$ and well replicate 5 at $h = 15$ are shown on the left and the right, respectively.](image)

**Figure 3.5.** Kernel smoothed density of estimated radii of A stains. Well replicate 4 at $h = 0$ and well replicate 5 at $h = 15$ are shown on the left and the right, respectively.

### 3.4 Results

Our simulations indicate that even fully accounting for the cell support presented by our cell cultures, A marks have both inhomogeneous intensity and attraction or repulsion trends among them. We will start with the latter. In order to investigate attraction or repulsion, we plot $\hat{K}_{A,A}(r)$ (observed A stains) for two example
wells, along with average and min-max bands from the $99 \tilde{K}_{A,A}^{(s)}(r)$ simulated with paradigm 1 in Figure 3.6. We observe that the bands do not entirely fit the initial dip in the plot as $r$ increases to $7 - 8$. Figure 3.7 upper panels show the estimated $K$-functions for the same example wells with min-max bands from paradigm 2 with a trim of 95%. We observe that the initial dip is now entirely within the bands, and this is more so for the lower panels, where paradigm 2 was implemented with a trim of 80%. However, $A$ stains significantly cluster together at larger ranges up to approximately $r = 20 - 30$, where the further accumulation of marks resembles that of a Poisson scenario, neither increasing nor decreasing. While we are still in the process of developing a careful biological interpretation for it, this attraction suggests that susceptibility to RSVA increases for medium-range neighbors of infected cells – in excess of any clustering that might be induced by inhomogeneity in the intensity of infected cells or by the spatial structure of the cell culture.

**Figure 3.6.** Upper panels: Estimated inhomogeneous $K$-functions for observed $A$ stains, $\hat{K}_{A,A}(r)$ (black), along with average and min-max bands for the $99 \hat{K}_{A,A}^{(s)}(r)$ from the simulations paradigm 1 (in colors)– the Poisson behavior ($\pi r^2$) is subtracted to each of the curves to help visualizing their variation. Well replicate 4 at $h = 0$ and replicate 5 at $h = 15$ are shown on the left and the right, respectively.

Next, we discuss intensity behavior, the patterns in the way $A$ marks are distributed on the cell support. We subdivided each well into a tessellation, and for each tile $g$ computed counts of nuclei, $c_{N,g}$, and counts of observed $A$ marks, $c_{A,g}$. Within the setting we implicitly assumed in the simulation procedures in Section 3.3, the number of $A$ marks in each of the different tiles is going to be distributed as a binomial random variable with $n = c_{N,g}$, and probability of success
Figure 3.7. Upper panels: Estimated inhomogeneous $K$-functions for observed A stains, $\hat{K}_{A,A}(r)$ (black), along with average and min-max bands for the 99 $\hat{K}^{(s)}_{A,A}(r)$ from the simulations paradigm 2 with a 95% trim (in colors). Lower panels: Estimated inhomogeneous $K$-functions for observed A stains, $\hat{K}_{A,A}(r)$ (black), along with average and min-max bands for the 99 $\hat{K}^{(s)}_{A,A}(r)$ from the simulations paradigm 2 with 80% trimming (in colors)– the Poisson behavior ($\pi r^2$) is subtracted to each of the curves to help visualizing their variation. Well replicate 4 at $h = 0$ and replicate 5 at $h = 15$ are shown on the left and the right, respectively.

$p = \frac{n_A}{w_N}$.

Figure 3.8 shows the scatter plots of $c_{A,g}$ vs $c_{N,g}$, with their lowess smooths, as well as the 2.5%-percentile and the 97.5%-percentile of the respective binomial distribution vs $c_{N,g}$ providing a reference, for the two example wells used in Figures 3.6, and 3.7. For tiles where the counts are small (sparsely populated areas of the well), there appears to be a direct relationship between the counts of observed A marks and nuclei – consistent with the behavior expected from a random thinning of the cells. However, in more densely populated areas, the relationship becomes negative in contrast to the expectation consistent with random thinning.
As long as we restrict attention to regions where there is a sufficient density of cells, we can detect locally within wells the same kind of inverse relationship between number of A infections and number of cells that we were able to detect for wells overall (Figure 3.3), and we have evidence that this inverse relationship is significant. This suggests a tendency for cells in densely populated areas not to be infected with A, or in other words to show a decrease in susceptibility to A infection for cells with close-by neighbors. We interpret this as a crowding effect, suggesting that sufficiently tightly packed cells change their state in ways that decrease their likelihood of becoming infected. On a technical note, the grids comprised 213 units for each well after excluding those near the rim of the disk for which less than 80% of the unit area is available – as we wanted to compare regions of the approximately same area. All other wells showed results similar to those in the examples of Figures 3.6, 3.7, and 3.8.
Figure 3.8. Upper panels: scatter plots of the counts of A marks $c_{A,g}$ vs the count of nuclei $c_{N,g}$ for a gridding of the wells in 213 units, with lowess smooths (in black) superimposed to visualize the relationship. Lines are also shown (in red and green) for the 2.5%-percentile and 9.5%-percentile vs count of nuclei, $c_{N,g}$, to provide a reference. Middle panels: Spatial plots of the well where the counts are outside of the bands from the upper panel. Lower Panels: Kernel smoothed intensities of Nuclei. Replicate 4 at $h = 0$ and replicate 5 at $h = 15$ are shown on the left and the right, respectively.
Summary and Future Work

We introduced some methods for exploratory spatial point process analysis and used the Poisson point process as an important class of such processes. We then conducted a simulation study to evaluate different approaches for the estimation of the $K$-function and the effects of sample size on this estimation. Finally, we applied these methods to real data from microscopy imaging of tissue cultures infected with the RSVA virus.

When studying the spatial features of A marks (indicators of RSVA infection) in our data, we are considering a dense infection (MOI=1) which is administered after increasing lags of time from the beginning of the experiment but is not given much time to act (there is no “duration of exposure” dimension). Thus, the effects we can detect are likely intrinsic and occurring on a short time scale (due to immediate changes in cells states, not responses that require time to mount). We observe the following:

1. An inverse relationship between number of infected cells and number of cells, indicates that crowding, through changes in the cells states, induces a decrease in susceptibility. This can be detected globally (overall counts in wells) but also locally (counts in grid units within the wells).

2. A tendency for infected cells to cluster (attract one another) at ranges of 129-193.5 microns is seen, even after accounting for the cell support and the non-homogeneous intensity of infected cells. This suggests increased susceptibility at certain ranges around infected cells.
Note that in our exploratory analyses we modeled and simulated from null scenarios that we used as benchmarks to assess observed behaviors. In the future, we will build simulation models that accurately reproduce the patterns observed for various marks in microscopy images and will be informative as explanatory models.

In addition to investigating one infection, biologists are often interested in the progression of co-infections (more than one virus stain) within a host organism. Various studies point to the fact that prior infections may substantially affect susceptibility to subsequent infections not just through responses due to the immune system, but through changes in the state of the cells exposed to infection within a tissue [Martinez et al., 2007, Lo et al., 2005, Ruggiero et al., 1989]. We also have data (again from Professor Poss’ laboratory) on co-infection experiments in which an infection with the RSVB virus at MOI=0.01 is followed by an infection with the RSVA virus at MOI=1 after a a lag of 0h, 3h, 6h, 9h, 12h, 15h hours. Due to the low MOI used for the initial RSVB infection, imaging from these experiments produces a low number of B marks (see Figure 4.1). In the future, new experiments will be run with a higher MOI for the RSVB infection as to obtain similar numbers of B and A marks (approximately 10000 for both). This will allow us to reliably investigate attraction or repulsion between the two types of marks and thus gain some insight on how one infection affects susceptibility to another. When dealing with co-infections, we will have to reconsider parameter tuning for the Image Pro Plus imaging software used in Professor Poss’ laboratory; RSVB and RSVA infections are marked with membrane and cytoplasmic stains, respectively, which present different characteristics. In particular, optimization of the software parameters for accurate detection of B marks has been problematic.

To avoid repetition and the associated chances of errors, we implemented programs to automate the use of Image Pro Plus software – so it can be applied with consistent parameter settings to multiple images. Our routines proved reliable and indeed eliminated mistakes in the repetition of parameter tuning steps. They will be modified and expanded to reflect the tuning procedures developed for the new data.

In order to detect any attraction or repulsion between A and B marks, we will use the cross $K$-function again simulating null hypothesized scenarios to create
appropriate benchmarks. For these simulations, in additions to the procedures already described for cell support and A marks, we will use the following procedures:

**Simulating B marks:** With the cell support in place, we go through the cells one by one and “infect” each, independently, with probability $\frac{n_B}{n_N}$ ($n_B$ is the number of observed B marks). We then place a B mark at a random location on the perimeter of each infected cell (disc). This procedure is repeated as to generate a total of 99 simulated B marks patterns

$$\widetilde{b}^{(s)} = \{\tilde{b}_1^{(s)}, \ldots, \tilde{b}_{\tilde{n}_B^{(s)}}^{(s)}\}, \quad s = 1, \ldots, 99$$

The expected number of marks in each simulated pattern is $E(\tilde{n}_B^{(s)}) = n_B$; no systematic spatial structure exists except for that created by the cell support.

**Joint simulation of A and B marks:** To assess repulsion or attraction between the two infections, we require a benchmark for the joint behavior of A and B marks. An intuitive one is given by considering spatially uncoupled processes for A and B – thus, we use the procedures we outlined in section 3.3 and above to generate simulated A and B marks patterns *independently* from each other. Note that in doing so the same cell can be “infected” with both A and B. The uncoupled simulated patterns $\tilde{a}^{(s)}, \tilde{b}^{(s)}, \quad s = 1, \ldots, 99$ are then used to benchmark the observed patterns $a, b$.

**Figure 4.1.** Counts of nuclei (left), B marks (middle) and A marks (right) as a function of lag for the co-infection experiments. Different points represent different replicate experiments at each lag duration, and color coding is consistent across the three plots – thus, for instance, the green points represent well replicate 3 at lag $h = 12$, which has the lowest number of nuclei, the lowest number of B stains, and the highest number of A stains.
Pilot runs of the new experiments suggest that improvements in the experimental protocols now produce lower levels of inhomogeneity in the cell cultures. This is an important step forward, since we detected that cell susceptibility to RSVA infection is affected by crowding among the cells. Removing these effects will facilitate our task when attempting to detect significant attraction or repulsion between A marks and B marks.


