DEPARTAMENT D’INFORMÀTICA

MODELLING, ESTIMATION AND ANALYSIS OF DYNAMIC PROCESSES FROM IMAGE SEQUENCES USING TEMPORAL RANDOM CLOSED SETS AND POINT PROCESSES WITH APPLICATION TO THE CELL EXOCYTOSIS AND ENDOCYTOSIS.

ESTER DÍAZ FERNÁNDEZ

UNIVERSITAT DE VALÈNCIA
Servei de Publicacions
2011

Aquesta Tesi Doctoral va ser presentada a València el dia 21 de desembre de 2010 davant un tribunal format per:

- Dr. Miguel López Díaz
- Dra. Amelia Simó Vidal
- Dra. Francisco Josep Ferri Rabassa
- Dr. Pedro García Sevilla
- Dr. Juan de Mata Domingo Esteve

Va ser dirigida per:
Dra. María Elena Díaz Fernández
Dr. Guillermo Ayala Gallego

Dipòsit legal: V-4124-2011
I.S.B.N.: 978-84-370-8075-8
Edita: Universitat de València
Servei de Publicacions
C/ Arts Gràfiques, 13 baix
46010 València
Spain
Telèfon:(0034)963864115

Department of Computer Science
Universitat de València

Modelling, estimation and analysis of dynamic processes from image sequences using temporal random closed sets and point processes with application to the cell exocytosis and endocytosis

Ester Díaz Fernández

Advisors
Dr. María Elena Díaz Fernández
Dr. Guillermo Ayala Gallego

Thesis submitted for the degree of Doctor of Philosophy (Ph.D)

November 2010

To my parents
To my daughter Ester

Acknowledgments

The author would like to thank her supervisors Dr. Elena Díaz, from the Department of Informatics, and Dr. Guillermo Ayala, from the Department of Statistics and Operative Research, Universitat de València, for their continuous encouragement, support, guidance and help during the course of this work, as well as their valuable advice and contributions to the work herein presented.

The author is also grateful to Dr. Derek Toomre, Dr. Liang Wei Gong and Dr. Roberto Zoncu from the Department of Cell Biology, Yale University, for their collaboration in this work.

This thesis has been supported by the Project "Novel imaging, modelling and analysis of exo-endocytosis coupling in neuroendocrine cells" supported by Human Frontier Science Organization (HFSP) organization, reference RGY40/2003. This work has been also supported by Project "Image sequence analysis based on spatio-temporal stochastic models. Applications to Cell Biology" supported by the Spanish Ministry of Science and Education, reference TIN2007-67587 (FEDER FUNDS).

Publications derived from this work

1. L. Gong, G. Di Paolo, **E. Díaz**, G. Cestra, M.E. Díaz, M. Lindau, P. De Camilli, and D. Toomre. *Phosphatidylinositol phosphate kinase type Iγ regulates dynamics of large dense-core vesicle fusion.* Proceedings of the National Academy of Sciences. Vol. 102, Num. 14, 5204-5209. April 2005.
2. R. Sebastián, **E. Díaz**, G. Ayala, M.E. Díaz, R. Zoncu and D. Toomre. *Studying endocytosis in space and time by means of temporal Boolean models.* Pattern Recognition, Vol. 39, Num. 11, pages 2175-2185. 2006.
3. G. Ayala, R. Sebastián, M.E. Díaz, **E. Díaz**, R. Zoncu and D. Toomre. *Analysis of spatially and temporally overlapping events with application to image sequences.* IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 28, Num. 10, pages 1707-1712. October 2006.
4. **E. Díaz**, R. Sebastián, G. Ayala, M.E. Díaz, R. Zoncu, D. Toomre and S. Gasman. *Measuring Spatio-Temporal Dependencies in Bivariate Temporal Random Sets with Applications to Cell Biology.* IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 30, Num. 09, pages 1659-1671. September 2008.
5. **E. Díaz**, G. Ayala, M.E. Díaz, L. Gong and D. Toomre. *Automatic detection of dense-core vesicles in secretory cells and statistical analysis of their intracellular distribution.* IEEE/ACM Transactions on Computational Biology and Bioinformatics. Vol. 7, Num. 01, pages 2-11. January-March 2010.
6. M.E. Díaz, G. Ayala, **E. Díaz**, *Estimating the Duration of Overlapping Events from Image Sequences Using Cylindrical Temporal Boolean Models.* Journal of Mathematical Imaging Vision. Vol. 38, Num. 2, pages 83-94. October 2010.

Collaborations with other groups generated by this work

The work presented in this thesis has generated collaborations with the following Universities and Institutes.:

1. Amelia Simó. Department of Statistics and Operative Research. Universitat Jaume I. Estimation of temporal Boolean models.
2. Derek Toomre. Department of Cell Biology, Yale University. Analysis of cellular endocytosis and exocytosis processes characterized by image sequences.
3. Antonio Miguel Garcia de Diego. Instituto Teófilo Hernando. Universidad Autónoma de Madrid. Automatic detection of exocytic vesicles.

Resumen

En esta tesis se presentan nuevos modelos y metodologías para el análisis de procesos dinámicos a partir de imágenes o secuencias de imágenes, en las cuales se produce solapamiento espacial y temporal de los objetos de análisis. Este solapamiento o coexistencia espacial y temporal es habitual en muchos fenómenos de la naturaleza. Por ello, se hace necesario modelizarlo adecuadamente en diversas disciplinas científicas tales como Microscopía, Ciencias de los Materiales, Biología, Geoestadística o Redes de Comunicaciones.

El trabajo realizado se enmarca en la teoría de Procesos Puntuales y Conjuntos Aleatorios Cerrados (RACS), dentro de la Geometría Estocástica. Los modelos propuestos son una extensión de la teoría de modelos booleanos en $\mathbb{R}^2$ incorporando una componente temporal.

La motivación de este trabajo fue originalmente su aplicación en un proyecto multidisciplinar que ha abarcado Estadística, Ciencias de la Computación, Biología y Microscopía, con el objetivo de analizar la exocitosis y la endocitosis celular. Los modelos y metodologías presentados en esta tesis responden a la necesidad de considerar adecuadamente el solapamiento espacio-temporal, necesidad que surgió durante este proyecto. La exocitosis y la endocitosis son dos procesos celulares mediante los cuales la célula segrega o absorbe sustancias a través de la membrana citoplasmática respectivamente. El estudio se realizó utilizando imágenes obtenidas por medio de las técnicas de microscopía electrónica y de reflexión total interna (TIRFM). En dichas secuencias, se observan las moléculas de proteínas participantes en el proceso de endocitosis como agrupaciones fluorescentes que se superponen en un número aleatorio de imágenes consecutivas. Estas imágenes se pueden modelizar como realizaciones de un proceso estocástico estacionario e isotrópico. No obstante, la metodología desarrollada permite analizar fenómenos reales en otros campos de la Ciencia en los que se observa superposición espacio-temporal de objetos con formas y duraciones aleatorias, como Geología, Química, Comunicaciones, etc.

En primer lugar, se introduce el modelo booleano temporal y se revisan los métodos de estimación de los parámetros del modelo. En segundo lugar, se presenta un método para la estimación de la función de distribución de la duración de los eventos en un modelo booleano temporal univariado basado en la función de covarianza espacio-temporal. Se ha realizado un estudio de simulación con diferentes

funciones de densidad para la duración, con resultados muy satisfactorios, incluso en imágenes con ruido. La metodología se ha aplicado al estudio particular de la endocitosis celular.

En tercer lugar, se ha desarrollado el modelo booleano temporal bivariado para estudiar la interrelación entre dos procesos espacio-temporales que coexisten en tiempo y espacio, y caracterizar su grado de solapamiento y dependencia espacial, temporal y espacio-temporal. Se presenta una generalización de las funciones $\mathbb{K}$ de Ripley, la covarianza espacio-temporal y la función de correlación para conjuntos aletorios bivariados. Se proponen como descriptores la función $\mathbb{K}$ de Ripley, la función $\mathbb{L}$ y la función de correlación. Para contrastar la hipótesis de independencia se realiza un test de Monte Carlo. La metodología desarrollada no es sólo un procedimiento de test sino que también permite cuantificar el grado y el intervalo espacio-temporal de las interacciones. Además, no requiere asumir hipótesis paramétricas. Se ha realizado un estudio de simulación analizando diferentes tipos de dependencia. La metodología desarrollada también ha sido aplicada con éxito al estudio de las interacciones de diferentes proteínas que participan en la endocitosis celular.

En cuarto lugar, a partir de imágenes de alta resolución de varias células obtenidas por microscopía electrónica, se ha modelizado la distribución en el citoplasma celular de vesículas exocíticas (gránulos) como una realización de un proceso puntual finito y los patrones asociados a varios grupos de células como réplicas de diferentes procesos puntuales. El objetivo es estudiar diferencias entre grupos de tratamientos que presumiblemente afectan a la localización espacial de los gránulos. Se ha caracterizado la distribución espacial de los gránulos respecto a la membrana plasmática mediante varios descriptores funcionales lo que ha permitido establecer diferencias significativas entre ambos grupos de células que hubiera sido imposible detectar mediante un enfoque convencional basado en la estadística clásica. Para segmentar las imágenes, se ha desarrollado una herramienta automática de detección de los gránulos, con resultados similares a los obtenidos manualmente para las mismas imágenes.

Finalmente, destacamos que hemos desarrollado una herramienta de software para la simulación y la estimación de análisis de modelos booleanos temporales (disponible en $http://www.uv.es/tracs/$), de manera que es posible para técnicos y científicos de cualquier disciplina la aplicación de los métodos desarrollados en esta tesis a otros problemas en los que exista superposición espacio-temporal de objetos con formas y duraciones aleatorias.

En conclusión, el desarrollo de estos nuevos modelos estocásticos espacio-temporales permitirán el modelado de procesos dinámicos a partir de secuencias de imágenes donde aparecen partículas con formas, tamaños y duraciones aleatorias que se superponen en el tiempo y el espacio. Hasta donde nosotros cononocemos, ésta es la primera vez que estas herramientas son aplicadas en el estudio de la exocitosis y endocitosis celular. Su aplicación contribuirá a un mejor conocimiento de estos procesos y facilitará la futura investigación en el campo de la Biología Celular, por ejemplo en el estudio de enfermedades asociadas a disfunciones en la secreción como la diabetes.

Abstract

In this thesis, new models and methodologies are introduced for the analysis of dynamic processes characterized by image sequences with spatial temporal overlapping. The spatial temporal overlapping exists in many natural phenomena and it is necessary to address it properly in several Science disciplines such as Microscopy, Material Sciences, Biology, Geostatistics and Communication Networks.

This work is related to the Point Process and Random Closed Set theories, within Stochastic Geometry. The proposed models are an extension of Boolean Models in $\mathbb{R}^2$ by adding a temporal dimension.

The study has been motivated for its application in a multidisciplinary project that combined Statistics, Computer Sciences, Biology and Microscopy, with the aim of analysing the cell exocytosis and endocytosis. The methods and models herein presented addressed the necessity of properly considering the spatial temporal overlapping, that arised during the project. Exocytosis is the process by which cells secrete vesicles outside the plasma membrane and endocytosis is the opposite mechanism. Our data were image sequences obtained by Electron Microscopy and Total Internal Reflection Fluorescence Microscopy (TIRFM). Fluorescent tagged-proteins are observed as overlapped clusters with random shape, area and duration (number of consecutive frames). We assume that they can be modelled as realizations of a stationary and isotropic stochastic process. The methodology herein proposed could be used to analyze similar phenomena in other Fields of Science such as Geostatistics, Chemistry, Medicine, Cellular communications, etc.

First, as State of the Art, the temporal Boolean model is introduced and some estimation methods for the parameters of the model are presented. Second, we proposed a method for the estimation of the event duration distribution function of a univariate temporal Boolean model based on spatial temporal covariance. A simulation study is performed with several duration probability density functions, with a very good performance even in noisy images. This methodology has been applied to the cell endocytosis.

Third, we introduce the bivariate temporal Boolean model to study interactions between two overlapped spatial temporal processes and to quantify their overlapping and dependencies. We propose a non-parametric approach based on a generalization of the Ripley $\mathbb{K}$-function, the spatial-temporal covariance and the pair correlation

functions for a bivariate temporal random closed set. The Ripley $\mathbb{K}$-function, the $\mathbb{L}$-function and the pair correlation function are proposed as descriptors. A Monte Carlo test was performed to test the independence hypothesis. This methodology is not only a test procedure but also allows us to quantify the degree and spatial temporal interval of the interaction. No parametric assumption is needed. A simulation study analysing several types of dependencies has been conducted and an application to the study of different proteins that mediate in cell endocytosis has been performed.

Fourth, from high spatial resolution EM images, we model the distribution of exocytic vesicles (granules) within the cell cytoplasm as a realization of a finite point process (a point pattern), and the point patterns of several cell groups are considered replicates of different point processes. Our aim was to study differences between two treatment groups in terms of granule location. We characterize the spatial distribution of granules with respect to the plasma membrane by means of several functional descriptors, that allowed us to detect significant differences between the two cell groups that would not be observed by a classical approach. In order to perform image segmentation, we developed an automatic granule detection tool with similar performance to that of the manual one-by-one marking.

Finally, it is important to point out that we have implemented a software toolbox for the simulation and analysis of temporal Boolean models (available at $http://www.uv.es/tracs/$), so scientists and technicians of any discipline can apply the proposed methods to any other problems where spatial temporal overlapping exists.

In summary, the new spatial temporal stochastic models herein proposed will allow modelling of dynamic processes from image sequences where several forms of random shape, size and duration overlap. To our knowledge, this is the first time that these tools have been applied to the study of cell exo and endocytosis, and they should contribute to a better understanding of these two complex processes. Our methodologies will help future research in Cell Biology, e.g. in the study of diseases related to secretion dysfunctions, such as diabetes.

Contents

Resumen **11**

Abstract **15**

List of Tables **v**

List of Figures **vii**

1 Introduction **1**

1.1 Motivation of this work 1

1.2 Outline of the thesis 3

2 State of the Art **5**

2.1 Biological background 6

2.1.1 Cell exocytosis 6

2.1.2 Cell endocytosis 10

2.1.3 Electron Microscopy 13

2.1.4 Total Internal Reflection Fluorescence Microscopy 14

2.2 Stochastic models 15

2.2.1 Point processes 15

2.2.2 Random Closed Sets (RACS) 17

2.2.3 The Boolean model 18

2.3 Temporal Boolean Models 21

2.3.1 Marginal analysis of the temporal dimension 23

2.3.2 Estimation of probability density function by aggregating temporal cross-sections 24

2.3.3 Estimation of probability density function by combining temporal cross-sections . . . 28

3 Estimation of the duration distribution in temporal Boolean models using the spatial-temporal covariance 33

3.1 Introduction . . . 33
3.2 The model . . . 36
3.2.1 The duration distribution of isolated events . . . 37
3.2.2 Evaluating the marginal duration distribution of isolated events 40
3.2.3 Error in the estimation of the mean duration using isolated events . . . 40
3.3 The covariance of a cylindrical temporal Boolean model . . . 42
3.3.1 Estimation of the distribution function of the duration . . . 45
3.3.2 Estimators . . . 45
3.4 A simulation study . . . 46
3.4.1 Estimating the cumulative distribution function . . . 47
3.4.2 Application to noisy images . . . 50
3.5 Application to cell Biology . . . 52
3.6 Conclusions . . . 54

4 The bivariate case: bivariate temporal random sets to model spatial-temporal dependencies 57

4.1 Introduction . . . 58
4.2 Bivariate Temporal Random set . . . 60
4.2.1 Notation and basic definitions . . . 60
4.2.2 Summary statistics for bivariate temporal random sets . . . 62
4.3 Estimators . . . 64
4.4 Testing spatial and spatial-temporal dependencies . . . 65
4.4.1 Independent temporal random sets . . . 65
4.4.2 Testing null hypotheses by using toroidal shift randomizations 67
4.4.3 Monte Carlo test . . . 69
4.4.4 Estimating the spatial-temporal interval of dependencies . . . 69
4.5 A simulation study . . . 70
4.6 Application to cell Biology . . . 76

4.7 Conclusions 80

5 Automatic detection of exocytic vesicles and statistical analysis of their intracellular distribution 83

5.1 Introduction 83

5.2 Data and image processing 86

5.2.1 Detection of granules 87

5.2.2 Shape and grey-level descriptors 89

5.2.3 Linear Discriminant Analysis 91

5.3 Statistical analysis 92

5.4 Results 96

5.4.1 Comparing counts of granules 96

5.4.2 Comparing distances to the plasma membrane 97

5.4.3 Comparing densities within a given distance to the plasma membrane 100

5.5 A simulation study 100

5.6 Conclusions 103

6 Conclusions and Future Work 105

A Software 109

A.1 Image segmentation 109

A.2 Statistical analysis of granules 111

A.3 Functional Data Analysis 112

A.4 Temporal Random Closed Sets (TRACS) 113

A.4.1 Temporal Boolean Models 113

A.4.2 Bivariate TRACS 117

B Supplementary Material 121

Bibliography 123

List of Tables

3.1 Description of the models simulated. 47
3.2 Volume fractions simulated. 47
3.3 Relative errors of the median of the duration (%). 51
3.4 Relative errors of the median of the duration using isolated clumps (%). 51
3.5 Relative errors in the estimation of the median of the duration in noisy images with a percentage of corrupted pixels R (%). 52
3.6 Description of biological image sequences. 53

4.1 Shifts applied to germs of type 1 to generate germs of type 2. 71
4.2 Monte Carlo p-values using $\mathbb{K}_{12}$-, g_{12}- and $\mathbb{L}_{12}$-functions. 72
4.3 Pairs of image sequences of fluorescently-labelled proteins. 77
4.4 p-values of Monte Carlo test applied to biological image sequences. . 80

5.1 Shape and grey-level descriptors. 91
5.2 Mean probabilities of belonging to a given class. 92

LIST OF TABLES

List of Figures

2.1 Constitutive and regulated vesicle exocytosis. 7

2.2 Stages in exocytosis and main proteins involved: (a) transport, (b) docking, (c) priming at the plasma membrane and (d) fusion. 9

2.3 Main stages of endocytosis. 11

2.4 Several snapshots of a fluorescent-tagged clathrin image sequence. . 12

2.5 An example of temporal Boolean model and three cross-sections. . . 23

2.6 Three consecutive cross-sections of a Temporal Boolean model and the three Boolean models derived. 29

3.1 Several consecutive frames of a simulated temporal Boolean model with cylindrical grains. Arrows point to an isolated episode. 35

3.2 A spatial-temporal reconstruction of a cylindrical temporal Boolean model. 36

3.3 Several snapshots of a fluorescent-tagged protein image sequence. . . 36

3.4 Estimated densities f^* of the random duration D^* for different remaining sample fractions. From top to bottom: uniform distribution, $U(4,8)$; Gamma distribution, $Ga(18, 0.3333)$, and normal distribution, $N(6,1)$. Durations are in unit time. 41

3.5 Relative errors in the estimation of the mean duration as a function of δ. From top to bottom: uniform distribution, $U(4,8)$; Gamma distribution, $Ga(18, 0.3333)$, and normal distribution, $N(6,1)$. 43

3.6 The spatial-temporal covariance of a temporal random set. 44

3.7 The first row corresponds to $ED = 6$ and the second row to $ED = 10$. From left to right, the columns correspond to $\lambda = 0.0001, 0.0002, 0.0003$. Uniform distribution for the durations. 48

3.8 The median (dashed line), the maximum and the minimum (solid lines) of the absolute error in the estimation of F_D. 49

3.9 The minimum and maximum deviations of the estimated cumulative distribution function F_D from the theoretical function. 50

3.10 Several consecutive snapshots of a fluorescent-tagged protein image sequence. The binary shapes correspond to the segmented areas of fluorescence. 52

3.11 Estimated cumulative distribution function of the duration of endocytic events. 54

4.1 A spatial-temporal plot of the fluorescence areas covered by two fluorescently-tagged proteins. 59

4.2 A realization of a simulated bivariate temporal random set. 61

4.3 A spatial-temporal reconstruction. 61

4.4 Graphical description of the spatial-temporal cross-covariance (a) and cross $\mathbb{K}$-function (b) for a bivariate temporal random set. The dark grey cylinders represent the i-th temporal random set and the light grey cylinders the j-th temporal random set. 63

4.5 An example of toroidal shift randomization. 68

4.6 Analysis of spatial dependencies for simulation 1. The empirical function $\mathbb{K}^{(1)}_{12,0}$ and its envelopes. 73

4.7 Analysis of spatial dependencies for simulation 1. Plot (a) corresponds to the pair correlation function, $g_{12,0}(s)$ and (b) to $\mathbb{L}^{(1)}_{12,0}(s) - s$. . . . 73

4.8 Analysis of spatial-temporal dependencies for simulation 3. 74

4.9 Analysis of spatial-temporal dependencies for simulation 5. 75

4.10 Estimating the interval of dependencies for simulation 3. 75

4.11 Estimating the interval of dependencies for simulation 4. 76

4.12 Plots of $g_{12}(s)$ and $\mathbb{L}_{12}(s)$ for pairs of fluorescent-labelled-proteins. Plots (a) and (b) correspond to pair 1 and plots (c) and (d) to pair 2. 78

4.13 Plots of $g_{12}(s)$ and $\mathbb{L}_{12}(s)$ for pair 3, Clathrin-RFP and Caveolin-GFP. 78

4.14 Plots of $L_{ij}(s,t) - st$ for pairs of fluorescent-labelled-proteins. (a) Pair 1, Clathrin-RFP and Hip1R-GFP. (b) Pair 2, Clathrin-RFP and Epsin-GFP. 79

4.15 Plots of $L_{ij}(s,t) - st$ for pair 3, Clathrin-RFP and Caveolin-GFP. . . 79

4.16 Estimating the spatial-temporal interval of dependencies. Plots (a) and (b) correspond to pair 1, Clathrin-RFP and Hip1R-GFP. Plots (c) and (d) to pair 2, Clathrin-RFP and Epsin-GFP. 81

5.1 A section of a wild-type adrenal chromaffin cell and the spatial point pattern associated. 85

5.2 Image processing. (a) Original image. (b) Negated image. (c) Background of the negated image. (d) Top-Hat image. (e) Binary image after thresholding. (f) Watershed lines over the binary image. (g) Watershed lines over the negated image. (h) Segmented granules. . . 88

5.3 Fitting a two-dimensional Gaussian. (a) Raw grey-level image. (b) Three-dimensional plot of the raw image. (c) Gaussian fit. 90

5.4 Intensity function $\lambda(x)$ of the process for two cells. (a) A mutant cell. (b) A control cell. Color is proportional to the intensity function value, from small values (green) to higher values (yellow, orange and white). 95

5.5 Comparison of ECDFs: (a) $\hat{F}_2$ function (mutant, solid line) against $\hat{F}_1$ (control, dashed line). (b) Bootstrap regions. 99

5.6 Bootstrap regions: (a) bootstrap regions of h_1 (dashed line) and h_2 (solid line); (b) bootstrap regions of the difference $h_1 - h_2$. 101

5.7 Several simulated point patterns. Realizations with (a) $\delta = 0.25$; (b) $\delta = 0.50$; (c) $\delta = 0.75$ and (d) $\delta = 1$. 102

5.8 The p-values provided by the bootstrap test as a function of δ. Plot (a) displays the cumulative distribution function of the distances from the granules to the plasma membrane and plot (b) the respective p-values associated with the h-function. Solid line corresponds to the median, dashed lines represent the lower and upper quartiles and dotted lines correspond to the 0.05 and 0.95 quantiles. 103

A.1 An electron micrograph of cell and the delineated membranes. 110

A.2 Parameter setting menu. 110

A.3 The results obtained: granule location and data script. 111

A.4 Two zooms of the granules with the granule membrane delineated. . 112

A.5 Several consecutive frames of a simulated TBM with cylindrical grains. 117

A.6 Parameter estimation: (a) Estimation of $\alpha(t)$, $\beta(t)$ and γ. (b) Estimated density of the duration. 117
A.7 A reconstruction of a TBM using function tracs3dplot. 118
A.8 An example of p-value map. 119

Chapter 1

Introduction

1.1 Motivation of this work

Studying complex dynamic processes in Medicine, Biology or Engineering requires an innovative and multi-disciplinary approach. For example, recent advances in live microscopy techniques allow the Biologists to extract new and very valuable information about some cellular processes from image sequences with hundreds of frames. A manual image processing or classical statistical analysis would not capture the benefits that such high-resolution image sequences provide. By contrast, the expertise in Stochastic Geometry and Computer Science could lead to outstanding results and solve most of the problems and drawbacks that Biologists find in their statistical analysis [59, 54, 18, 31, 41, 40] of such complex data.

Automatic image segmentation provides helpful tools that facilitate the image processing and increase the quality of the data obtained. The development of spatial temporal stochastic models allows us to model dynamic processes and to estimate with great accuracy the main parameters that characterize the phenomena.

In this thesis, different problems in Stochastic Geometry are addressed. In spite of the considerable number and variety of applications of Boolean models, the problem of analyzing overlapping events in space and time from image sequences remains elusive. More sophisticated mathematical models and statistical methods are needed i.e. hybrid models that capture both temporal and geometric properties while formalizing a configuration of independent randomly placed particles with random durations. First, we have modeled the image sequences of fluorescent-labeled proteins where random shapes overlap in space and time as a temporal Boolean model, to study the cell endocytosis.

Second, a bivariate case is also proposed to model the interactions between two overlapped spatial temporal processes and to quantify their dependencies. Although the problem of measuring interaction between types of events in bivariate spatial point processes has been reported and analyzed [17, 24, 80, 87], this is not the case for random sets. In point processes, the standard summary methods use bivariate counterparts of the empty space function, the nearest neighbor distance distribution, the Ripley $\mathbb{K}$-function and derived statistics such as the pair correlation function and the J-function. However, further extensions which capture dependencies associated with time and spatial properties of overlapping events are needed. An example of the problem addressed with a bivariate temporal random closed set is the joint study of the interactions of different proteins in the cell endocytosis.

Third, from high spatial resolution electron microscope images, we have modeled the distribution of exocytic vesicles (granules) within the cell cytoplasm as a realization of a finite point process (a point pattern), for which the point patterns of several cell groups associated with different cell treatments are considered as replicates of different point processes. Our aim was study differences between two treatment groups in terms of granule location. Each experiment implies the analysis of dozens of cells with hundred of granules each. Therefore, we developed an automatic granule detection tool for image segmentation with a similar performance to that of the manual one-by-one marking.

Problems similar to the biological ones herein studied that also involve the existence of random spatial patterns have been encountered in other fields, such as Ecology, Epidemiology and Geographical Information Systems. Spatial Statistics has provided several models to analyze any given set of points or processes which can generate random point data [24, 17, 66]. In some problems the temporal dimension should be explicitly considered. Several applications of spatial-temporal point processes in Biology are [71, 91].

1.2 Outline of the thesis

In Chapter 2, we introduce some basic concepts of Cell Biology and Microscopy, which allow the reader a better understanding of the applications of the proposed methodology. Second, Point Processes and Random Closed Sets are reviewed. We also present the-state-of-the-art in Temporal Boolean Models applied to the analysis of image sequences where spatial and temporal overlapping exists. In Chapter 3 the definition of the cylindrical temporal Boolean model and some theoretical results on the statistical analysis of isolated events is presented. We propose a semi-parametric estimator of the distribution function of the duration based on the spatial-temporal covariance of random sets. In Chapter 4, we model the binary image sequences of two different event types as a realization of a bivariate temporal random set and propose a non-parametric approach to quantify spatial and spatial-temporal interrelations using the pair-correlation, cross-covariance and the Ripley $\mathbb{K}$ functions. We propose a randomization procedure to test independence between event types by applying random toroidal shifts and a Monte Carlo test. In Chapter 5, we model the locations of granules of a given cell as a realization of a finite spatial point process and the point patterns associated with the cell groups as replicated point patterns of different spatial point processes. An algorithm to segment the granules on electron microscopy images is proposed. The relative locations of the granules with respect to the plasma membrane are characterized by two functional descriptors and the descriptors of the different cells for each group are compared using bootstrap procedures. Chapter 6 is devoted to the conclusions of this thesis and future extensions.

Chapter 2

State of the Art

Spatial Statistics has provided several valuable models to analyze sets of points or processes which can generate point sets following some stochastic law [24, 17, 66]. Problems that involve random spatial patterns have been encountered in various fields such as Biology, Ecology, Epidemiology, Geology, etc. In some cases the temporal dimension should be explicitly considered. Several applications of spatial-temporal point processes in Biology are [71, 91].

Approaches based on random sets have also been considered to study similar processes with overlapping in these areas of knowledge. In particular, the Boolean model is a well-known model when we study binary images composed by the union of several overlapping random shapes located independently in the 2D Euclidean space [79, 57, 42]. The use of such models allows us to estimate the mean number of objects per unit area or the mean size (area and perimeter) of the objects from an image. The Boolean model has been widely used since it was proposed by G. Matheron in the late 1960s [51, 52, 72], because it can provide a good description for irregular patterns observed in Microscopy, Material Sciences, Biology, Chemistry, Geostatistics or Cellular Communication Networks. A complete study of this kind of stochastic model is detailed in [56], focused on modelling (static) $2D$ or $3D$ overlapping objects.

Other interesting studies which to some extent introduce dynamics into the Boolean model are [27, 86]. In [86], van den Berg et al. let the grains change their positions and they characterized the properties of this modified Boolean model. In [27] a wireless sensor network is modelled by defining a blinking Boolean model in which grains switch in an uncoordinated way between an on and off mode to save energy, and the latency of the network is studied.

In most of the studies that deal with 3D data it is common to analyze each of the 2D sections separately, even with isotropic data. Dougherty et al. ([26]) present an application for counting illuminated randomly sized spheres in 3D regions by taking the cross-sections with spheres to form the estimate. A similar approximation is carried out in [68] to estimate the number of straight cylinders in a random system.

The definition of the temporal Boolean model to study an image sequence as a whole and quantify spatial-temporal overlapping is developed in [70, 9, 69]. An interesting result of non-homogenous temporal Boolean model is [36]. Fuzzy bivariate temporal random closed sets applied to the study of cell endocytosis are developed in [22].

First, we review the biological processes that have motivated this thesis. Second, we introduce basic concepts of point processes, random closed sets (RACS) and Boolean models. Third, we define the temporal Boolean model (TBM) and we present some estimation methods previously published, to which the author of this thesis has contributed.

2.1 Biological background

In eukaryotic cells, the communication with the outside occurs through the plasma membrane. The cell absorbs nutrients and other necessary substances from outside in a process called *endocytosis*. The molecules are progressively surrounded by a part of the plasma membrane, which finally invaginates to form a vesicle within the cytoplasm. During the inverse process, called *exocytosis*, the cell releases secretory vesicles, which contain the active molecules needed for the generation of some processes in other cells. Then, exocytosis is the basis of intercellular communication in multicellular organisms.

2.1.1 Cell exocytosis

The biological process consisting of the release of a wide array of extracellularly acting molecules by a cell is called *exocytosis*. The fusion of secretory vesicles with the plasma membrane occurs in the form of *constitutive exocytosis* that is required for the insertion of new plasma membrane in essentially all cells [2, 13, 74]. Some extracellular molecules (e.g., plasma proteins, antibodies, extracellular matrix components, etc.) are also secreted by a constitutive exocytotic pathway. In many cell types, a second pathway also exists in which exocytosis can be tightly regulated to allow the controlled release of vesicle contents or regulated insertion of new membrane components due to fusion of preformed secretory vesicles only *in response to a physiological signal.* This second pathway is known as *regulated exocytosis* [2, 13]. The two pathways for vesicle exocytosis are illustrated in Fig. 2.1.

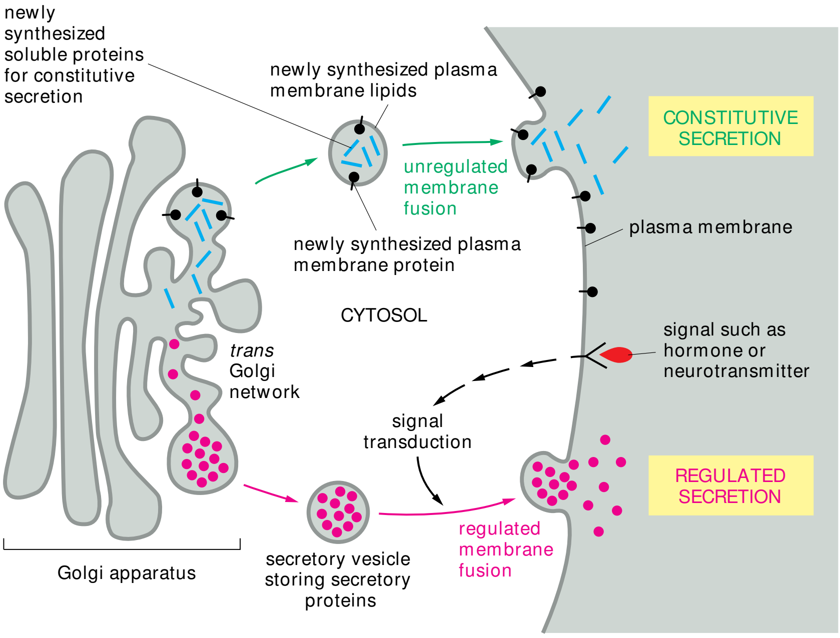


Figure 2.1: Constitutive and regulated vesicle exocytosis.

Vesicles are originated in the Golgi apparatus and travel across the cytoplasm to the plasma membrane. Modulation of regulated exocytosis is central to the regulation of cellular signalling. A variety of disorders (such as epilepsy, hypertension, diabetes and asthma) are closely associated with pathological modulation of exocytosis. Therefore, regulated exocytosis has been extensively studied in synapses where it is the mechanism by which neurotransmitters are very rapidly released in a controlled manner from synaptic vesicles to mediate neurotransmission. Certain neurons also possess *dense-core vesicles*, that can be triggered to undergo exocytosis independently of synaptic vesicles [13, 74].

A wide range of non-neuronal cell types contain also dense-core vesicles (also called *granules* or *secretory vesicles*), the contents of which serve a diverse range of physiological functions. These include cells specialized to secrete large amounts of secretory products. Besides neurons, regulated exocytosis of secretory granules has been extensively studied in many well-characterized cell types of neuroendocrine, endocrine and exocrine cells due to their practical advantages [13, 74].

Adrenal chromaffin cells is a certain experimentally favorable cell type to study regulated exocytosis. Chromaffin cells are neuroendocrine cells of the adrenal medulla and are widely used for the investigation of synaptic transmission, since they are of neuronal origin and release *catecholamine*, similarly to adrenal nerve terminals. The main advantage of chromaffin cells is that they can be isolated and purified in large quantities [46, 74].

Therefore, functional and structural studies of granules of adrenal chromaffin cells have been very important in the development of the current description of the various stages in regulated exocytosis [41, 43, 46, 47, 61, 62, 90]. Electron microscopy (EM) techniques provide images with very high spatial resolution and allow the detection and characterization of secretory granules.

Exocytosis in chromaffin cells undergoes a multistep cycle that includes: i) formation of vesicles; ii) transport to the active zone; iii) vesicle docking at the active zone; iv) priming (or the transformation of vesicles into fusion-competent vesicles); v) fusion and, finally, vi) recycling [2, 13, 90]. This process is illustrated in Fig. 2.2. Before exocytosis, the vesicle first moves from the cytoplasm to the plasma membrane. That is known as the *transport* phase. Second, the vesicle becomes close to the plasma membrane, a process often referred to as *docking*. Third, there is an activation involving metabolic energy, required to achieve fusion competence and known as *priming*. Finally, the vesicle and the plasma membrane merge in the *fusion* phase, and the vesicle content is released.

Each of these steps involves a multitude of proteins that operate by complex protein-protein and protein-phospholipid interaction networks. Transport of vesicles occurs along microtubular tracks with the aid of kinesin motor proteins. The initial contact between the vesicle and the plasma membrane is mediated by protein complexes that appear to be essential for ensuring that only appropriate membranes fuse. Apparently, each fusion step is controlled by a specific family member of proteins

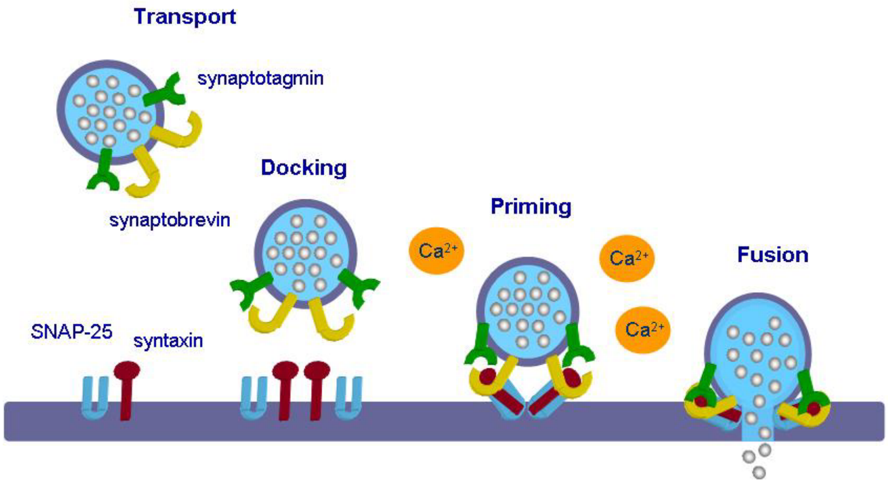


Figure 2.2: Stages in exocytosis and main proteins involved: (a) transport, (b) docking, (c) priming at the plasma membrane and (d) fusion.

that resides on the vesicle membrane (such as, synaptotagmin and synaptobrevin) and appears to play a key role in recruiting docking proteins to the vesicle surface.

Fig. 2.2 shows the main proteins involved in regulated exocytosis. *Syntaxin* and *SNAP*-25 form a complex on the plasma membrane that interacts with *synaptobrevin* on the vesicle membrane. The assembly of the resulting ternary complex forces the opposing membranes into close apposition and may be enough to induce exocytosis. Vesicle fusion requires Ca^{2+} ions that enters the cell via voltage-gated calcium channels. The near universal role of Ca^{2+} as the trigger for granule fusion predicts the existence of some proteins capable of activating the fusion machinery upon binding Ca^{2+}. Several different candidate proteins have been suggested to play this role, although *synaptotagmin* seems to be the major Ca^{2+} sensor to induce rapid exocytosis [12, 13, 88].

Regulated exocytosis could be affected by the over-expression or the inhibition of the specific proteins involved. In recent years, many of the proteins involved in the mechanism of secretory vesicle exocytosis have been identified and their *in vivo* importance has been established. The interactions between these proteins and the way in which a Ca^{2+} signal leads to secretory vesicle exocytosis are known in outline, although the full details of the processes involved still remain to be resolved [41, 74, 78, 81, 88, 90]. Therefore, studying the effect of the over-expression or inhibition of a given protein in the intracellular distribution of dense-core vesicles in chromaffin cells is central to the understanding of regulated exocytosis [13, 74].

Studies have traditionally classified vesicles in adrenal chromaffin cells according to their morphological or structural appearance. Several types of granules are described and different classification of granules are made in [61, 46, 47, 62].

In [61], two types of vesicles are distinguished: (1) *dense vesicles* with darker appearance, most of them spherical and with a mean diameter of 270 nm, and a few ones with ellipsoidal or elongated appearance; (2) *clear vesicles* with lighter appearance (almost white), which do not content functional substance.

A similar classification is given by [62], which distinguishes two types of granules: (1) *"stable" vesicles or stand-alone dense vesicles*, which are isolated electron-dense vesicles with dark appearance, most of them spherical, although some ones have elongated or irregular shapes. (2) *"active" vesicles or dense vesicles with "halos"*, which are electron-dense vesicles with irregular white vacuoles associated. The white vacuoles or zones seem to be deformations of vesicle membrane which remain empty, as if the dense vesicles were a dark ball floating in a bigger empty balloon. The location of the dense vesicle (the ball) is random and erratic. Then, white zones are absolutely irregular, both in size and shape. These granules are lighter than stand-alone vesicles and their pH is lower. Vesicles with halos are called active because the ratio of the number of these vesicles and the number of stable vesicles increased when cells are stimulated. In fact, when cell stimulation arises, at least half of the initial population of vesicles are exocytosed. Most of the remainder ones present halos. Isolated white vacuoles without a electron dense core are also observed, and they might be empty membranes.

2.1.2 Cell endocytosis

Endocytosis is the process by which cells traffic components from the plasma membrane into various intracellular compartments [53, 31]. This process regulates a wide range of physiological functions including cell migration, signaling, nutrient uptake, wound repair and neurotransmission.

Clathrin-mediated endocytosis is one of the best characterized endocytic routes [53, 44]. Endocytosis happens in discrete events, in which cargo-loaded vesicles detach from the plasma membrane and are trafficked inside the cell. The steps of this process are the following: (1) Clathrin assembly is initiated by binding cargo molecules to receptors on the plasma membrane and (2) a molecular lattice of clathrin molecules covers a portion of the plasma membrane containing the cargo-receptor complex [75]. Upon assembly, (3) the clathrin lattice curves into a basket (vesicle) which encases the cargo-laden endocytic vesicle [29]. Epsin helps to recruit the clathrin lattice, and it is thought to contribute to the acquisition of curvature [34]. Subsequently, (4) the basket pinches off the plasma membrane and the clathrin coat is shed from the vesicle, which is further trafficked to internal cellular compartments. Hip1R couples the clathrin lattice to the actin cytoskeleton, which participates in the pinching (5) and trafficking of the endocytic vesicle [10]. During endocytosis, invagination of the plasma membrane is used to bring nutrients into the cells and to recycle peripheral cellular components. The sequential recruitment of adaptor proteins, coat proteins and mechanoenzymes to sites of endocytosis, which is required for the formation of a complete endocytic vesicle, has been recently revealed by multi-colour TIRFM [54, 60, 91]. These steps are shown schematically in Fig. 2.3.

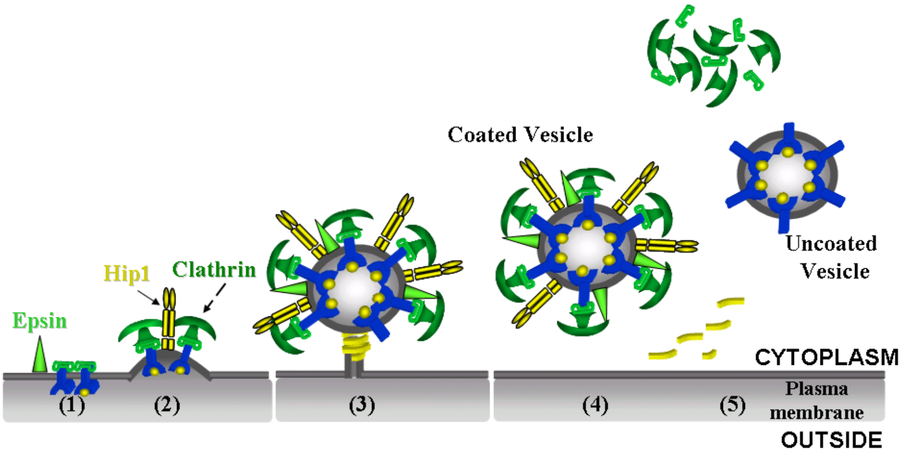


Figure 2.3: Main stages of endocytosis.

Several proteins mediate in this complex process, such as, GTP dynamin. Dynamin surrounds the neck of clathrin-coated vesicles, and hydrolyzes GTP, resulting in the fission of the vesicle. The exact mechanism of this release is still unknown. Dynamin may directly ”pinch”, or ”pop”, the vesicle from the plasma membrane, due to a conformational change resulting from the hydrolysis of GTP. Alternatively, dynamin may recruit additional factors which, in turn, operate the fission step [73]. It has been reported that a sudden burst of dynamin associated with the coated pit precedes the budding and disassembly of the vesicle [31, 53]. In any case, the GTPase dynamin has emerged as a crucial mediator of the endocytosis of clathrin-coated vesicles.

Some of the steps in the endocytosis process has been inferred from structural and biochemical studies [16]. However, the endocytosis process has been recently observed in live cells, due to fluorescent molecules that are incorporated to proteins such as clathrin or dynamin. This fluorescence could be measured with specialized microscope techniques such as Total Internal Reflection Fluorescence Microscopy (TIRFM) [82, 84, 83, 45]. This technique illuminates a thin section near the cell-coverslip interface and gives a very high signal-to-noise ratio, thus facilitating visualization of cellular processes near the plasma membrane. Viewed by TIRFM, the assembly of fluorescently-labeled clathrin molecules at a site of ongoing endocytosis results in the appearance and steady growth of a diffraction-limited spot. The time which elapses between the appearance and the disappearance of a fluorescent clathrin spot is defined as the duration, or lifetime, of a discrete endocytic event [31, 70, 91]. An image of a cell expressing clathrin fluorescent protein provided by TIRFM is shown in Fig. 2.4.

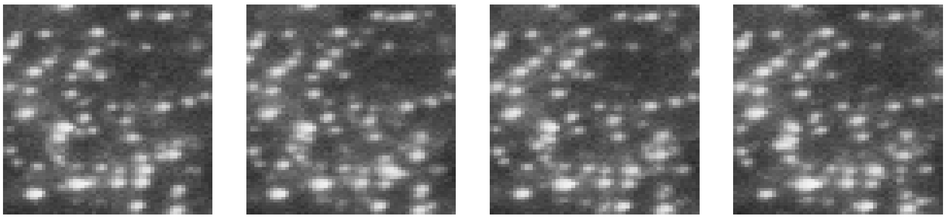


Figure 2.4: Several snapshots of a fluorescent-tagged clathrin image sequence.

From a biological point of view, it is key to characterize the areas where endocytic events occur (from now on, *endocytic spots*) by estimating parameters which allow knowing more about the underlying phenomenon of endocytosis, such as the number of endocytic spots per unit area, their mean perimeter or area, or their mean duration. An *endocytic spot* is defined as the associated area of the minimum cluster of fluorescent clathrin molecules that can be detected. It can vary in size and shape due to many factors, including: the intrinsic variability in the size of vesicles, the reversibility of the process, and the instrumental noise.

The different areas of fluorescence overlap with each other and vary in size and shape, forming relatively large and irregular regions with high fluorescence. From now on, these regions will be called *clumps*. Each connected component or clump may comprise one or an unknown number of very close and overlapped clathrin-coated vesicles.

Due to overlapping, classical techniques based on segmentation and labelling of the connected components would lead to an underestimation of the number of vesicles. It is a common practice in the endocytosis literature to use shape and size criteria from consecutive frames of a time-lapse movie, in order to select the clumps that are presumably composed by a single endocytic vesicle and to estimate from them key parameters such as the mean area, the mean perimeter or the mean duration [31].

This approach could lead to an underestimation since the selected clumps are biased to events with lesser duration and size. Given a frame, the endocytic spots corresponding to larger vesicles will have a greater probability of belonging to a non-isolated endocytic spot, i.e., a greater probability to touch other endocytic spots. If only the isolated endocytic spots are used to study the phenomenon then a biased sample is considered (one that includes smaller vesicles with higher probability) producing biased estimators, such as the mean number per unit area, the size of a single endocytic spot (area or perimeter) or the duration of a given event.

2.1.3 Electron Microscopy

Electron Microscopes (EMs) are scientific instruments that use a beam of highly energetic electrons to examine objects on a very fine scale. EMs were developed due to the limitations of Light Microscopes which are limited by the physics of light to 500× or 1000× magnification and a resolution of 0.2 micrometers. Seeing the fine details of the interior structures of organic cells requires at least 10000× magnification, which was just not possible using Light Microscopes. EMs function exactly as their optical counterparts except that they use a focused beam of electrons instead of light to “image” the specimen and gain information as to its structure and composition.

The basic steps involved in all EMs are: (a) a stream of electrons is formed (by the

Electron Source) and accelerated toward the cell using a positive electrical potential; (b) this stream is confined and focused using metal apertures and magnetic lenses into a thin, focused, mono-energetic beam; (c) this beam is focused onto the sample using a magnetic lens; and, (d) interactions occur inside the irradiated sample, affecting the electron beam. These interactions and effects are detected and transformed into an image. The above steps are carried out in all EMs regardless of type. There are two different types of EM: Transmission Electron Microscopes (TEM) and Scanning Electron Microscopes (SEM).

A TEM works much like a slide projector. TEMs shine a beam of electrons through the specimen. Whatever part is transmitted is projected onto a phosphor screen. The image strikes the phosphor image screen and light is generated, allowing the user to see the image. The darker areas of the image represent those areas of the sample that fewer electrons were transmitted through (they are thicker or denser). The lighter areas of the image represent those areas of the sample that more electrons were transmitted through (they are thinner or less dense).

2.1.4 Total Internal Reflection Fluorescence Microscopy

Total Internal Reflection Fluorescence Microscopy (TIRFM) is one of the most advanced technique of microscopy and is based in Snell's law: if light travelling in a dense medium strikes a less dense medium beyond a certain 'critical angle' θ_c, the light undergo Total Internal Reflection, TIR.

Cells are grown on glass coverslip or transparent materials of high refractive index, and a beam of light, usually from a laser, is optically coupled into the cover slip by a prism or the objective itself. If the light approaches the aqueous medium at greater angle than θ_c, it totally reflects into the glass; however, if the light 'rays' simply bounced off the interface like a mirror, this would never illuminate the cell. Then, the laser beam illuminates the sample beyond a critical angle to generate an evanescent wave. Namely, as the beam of light traveling through a high refractive index medium (e.g. glass; R.I. 1.51) encounters another media with a lower refractive index (i.e., water or cell cytosol; R.I. 1.3-1.4), the beam undergoes total internal reflection. As a consequence, a small excitation wave called *Evanescent Field* is generated.

The evanescent field is only $100 - 200$ nm thick, and it decays exponentially as it moves away from the coverslip. Therefore, only objects which are within $100-200$ nm of the bottom plasma membrane of the cell are illuminated, while the nucleus, inner cytosol and upper plasma membranes are left in the dark. In this way, it is possible to image membrane-associated events, such as endocytosis and exocytosis, with a very high signal-to-noise ratio, thus facilitating visualization of cellular processes near the plasma membrane [3, 45, 82, 84, 83].

Under TIRFM, the assembly of fluorescently-labeled proteins at a site of ongoing endocytosis results in the appearance and steady growth of a diffraction-limited spot, in such a way that the areas of fluorescence of endocytic events overlap, forming random clumps in space and time.

Electronic Microscopy has high spatial resolution, but only gives snapshots of the process. By contrast TIRFM offers a compromise in that good spatial-temporal resolution can be achieved. TIRFM microscopes allow the Biologist to observe how the vesicles fusion with the plasma membrane or how isolated molecules move near the plasma membrane [84, 83].

2.2 Stochastic models

Now, we briefly introduce basic concepts of the Point Processes and the Random Closed Sets. We also define the Boolean model.

2.2.1 Point processes

A spatial point process is a stochastic mechanism producing a countable set of points in the plane. A good reference about statistical analysis of spatial point processes is [24]. It is usual to assume stationarity and isotropy, i. e., invariance under spatial shifts and invariance under rotation respectively. Although the models are often defined as processes on the whole plane, in practice we only apply them to data from finite planar regions and it will be sufficient for our purposes if stationarity and isotropy hold to a reasonable approximation within the study region.

Let $x_i : i = 1, 2, 3, \ldots$ denote the points of a realization in $\mathbb{R}^2$, a point pattern. Let us denote $E[X]$ the expectation of a random variable X, $N(C)$ the number of events in the planar region C, $\nu_2(C)$ the area of C, dx an infinitesimal region which contain the point x, and $\|x_i - x_j\|$ the Euclidean distance between the points x_i and x_j.

We can now define the first-order and second-order properties of a spatial point process. The first-order intensity function measures the *uniformity* of the pattern, i.e., the way in which the expected value (mean or average) of the process varies across space. It is given by

$$\lambda(x) = \lim_{\nu_2(dx)\to 0} \frac{E[N(dx)]}{\nu_2(dx)}, \tag{2.1}$$

where dx is an infinitesimal region at location x with area $\nu_2(dx)$ and $N(dx)$ denotes the number of events in the region.

The second-order intensity function is a measure of the *dependency* structure of the events in a region C, i.e., the covariance (or correlation) between values of the process at different regions in space. It is given by,

$$\lambda_2(x_i, x_j) = \lim_{\nu_2(dx_i),\nu_2(dx_j)\to 0} \frac{E[N(dx_i)N(dx_j)]}{\nu_2(dx_i)\nu_2(dx_j)}. \tag{2.2}$$

For a stationary point process, $\lambda(x) = \lambda$, i.e., the expected number of events at an arbitrary location x is constant for all $x \in C$; and $\lambda_2(x_i, x_j) = \lambda_2(x_i - x_j)$ so that dependence between events depends only on the difference $h = x_i - x_j$ between x_i and x_j. When the point process is stationary and isotropic, $\lambda_2(x_i, x_j) = \lambda_2(\|x_i - x_j\|)$.

An alternative characterization of the second-order properties of a stationary isotropic process is provided by the function $\mathbb{K}(s)$ defined as

$$\mathbb{K}(s) = \frac{1}{\lambda} E[N_0(s)] \tag{2.3}$$

being $N_0(s)$ the number of further events within distance s from an arbitrary event.

In practice, this function is estimated by locating the center of a disk of radius s over each event and counting the number of events (without including the center) inside the disk. The count is done for every event. If events are regularly spaced, each one is likely to be surrounded by empty space and therefore, at small distances, $\mathbb{K}(s)$ will be relatively small.

An estimator of the $\mathbb{K}$-function is proposed by Ripley [24] and given by

$$\hat{\mathbb{K}}(s) = \frac{\nu_2(C)}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} I(d_{ij} \leq s), \tag{2.4}$$

where $I(\cdot)$ denotes the indicator function. A weight w_{ij} is included because of the edge effects.

The $\mathbb{K}$-function and the intensity $\lambda_2(s)$ of a stationary point process are related to the distribution function and probability density function of the distances between pairs of events in a point pattern. A useful property of the $\mathbb{K}$-function is that it is invariant under random thinning. In a random thinning operation, each point of a given point pattern X is randomly either deleted or retained (i.e. not deleted) with a retention probability p. The result is another point pattern, consisting of those points of X that were retained, with intensity $p\lambda$ and the same $\mathbb{K}$-function.

2.2.2 Random Closed Sets (RACS)

Let us introduce some basic notation. The Lebesgue measure in $\mathbb{R}^d$ will be denoted by ν_d (ν_3 is the volume in $\mathbb{R}^3$ and ν_2 the area in $\mathbb{R}^2$). In particular, for A a Borel subset of $\mathbb{R}^2$, $\nu_2(A)$ and $U(A)$ will denote the area and perimeter of the set A, respectively. $A \oplus B = \{a + b : a \in A, b \in B\}$ is the Minkowski addition of sets A and B. Finally, $\check{A} = \{-a : a \in A\}$ is the symmetric of A with respect to the origin.

Random closed sets are mathematical models for irregular random area patterns whose formal definition was provided by [52].

Let $\mathcal{F}$ be the class of closed subsets in the Euclidean space $\mathbb{R}^2$ and σ_f the σ-algebra generated by the sets $\mathcal{F}_K = \{F \in \mathcal{F} : F \cap K \neq \emptyset\}$ where K is a compact subset of $\mathbb{R}^2$. If $\mathcal{P}$ denotes a probability measure in $(\mathcal{F}, \sigma_f)$, then $(\mathcal{F}, \sigma_f, \mathcal{P})$ is a *random closed set*. Let Φ be a random closed set, i.e., a random element of this probability space. Note that the definition is given in such a way that $\{\Phi \cap K \neq \emptyset\}$ is a random event, i.e., a given compact subset K touches the random closed set, and its probability $P(\Phi \cap K \neq \emptyset)$ can be computed.

In fact, the following set function characterizes the probability distribution of the random set Φ

$$T(K) = \mathcal{P}(\Phi \cap K \neq \emptyset) \text{ with } K \text{ any compact subset.} \tag{2.5}$$

This set function is known as the *capacity functional* of Φ.

Intuitively, a random closed set Φ is a random mechanism producing random (closed) subsets in space.

We model each binary image of the sequence as a realization of a random set in $\mathbb{R}^2$ and the whole image sequence as a random set in $\mathbb{R}^2 \times \mathbb{R}_+$. Let us denote by $\Phi(t)$ the random set corresponding to the random binary image at the observation time t, with $t \geq 0$. If we stack the observed temporal cross-sections for every time t, we define the *temporal random set* in $\mathbb{R}^2 \times \mathbb{R}_+$ as

$$\Phi = \cup_{t \geq 0} \Phi(t). \tag{2.6}$$

This model can be considered as a particular case of a non-isotropic three dimensional random set. The lack of isotropy comes from the temporal dimension, since we assume that the temporal variation is different from the spatial variation.

2.2.3 The Boolean model

The *Boolean model* is a type of random closed set. Let $\Psi = \{y_1, y_2, \ldots\}$ be a stationary Poisson point process in $\mathbb{R}^d$ with intensity λ [17, 79, 24, 80].

Let $\Xi_1, \Xi_2, \ldots$ be a sequence of independent and identically distributed (as Ξ_0) random compact sets in $\mathbb{R}^d$ such that they are independent of the Poisson point process and $E\nu_d(\Xi_0 \oplus \check{K}) < +\infty$ for every compact set K.

A *Boolean model* is defined as

$$\Phi = \cup_{n=1}^{\infty} (\Xi_n + y_n). \tag{2.7}$$

The points y_n are called the *germs* and the sets Ξ_n are the *grains*. The random set Ξ_0 is called the *primary grain*.

The distribution of a random closed set is completely characterized by its *capacity functional* defined as $T(K) = P(\Phi \cap K \neq \emptyset)$ (with K an arbitrary compact subset). For a Boolean model, this functional is given by

$$T(K) = 1 - \exp\{-\lambda E\nu_d(\Xi_0 \oplus \check{K})\}. \tag{2.8}$$

When germs follow a stationary Poisson process, the model is stationary. If primary grains Ξ_0 are isotropic (e.g. discs with random radii), then the model is isotropic, i.e., its distribution is invariant against random rotations.

In particular, if $K = \{0\}$, the value $T(\{0\}) = P(0 \in \Phi)$ is known as the *area* *fraction*, the mean area in $\mathbb{R}^2$ (the mean volume in $\mathbb{R}^3$) covered by the random set per unit area (respectively, per unit volume). The area fraction in a d-dimensional Boolean model is given by

$$p = 1 - \exp\{-\lambda E\nu_d(\Xi_0)\}. \tag{2.9}$$

In general, the area fraction in a d-dimensional Boolean model is defined as the mean of the area of Ξ in a unitary region B

$$p = E(\nu_d(\Xi \cap B)), \quad \nu_d(B) = 1. \tag{2.10}$$

When Ξ is stationary, p doesn't depends on region B and then

$$p = P(0 \in \Xi). \tag{2.11}$$

If $K = 0$, then

$$P(0 \in \Xi) = 1 - exp\{-\lambda E(\nu_d(\Xi_0))\}. \tag{2.12}$$

Intuitively, a Boolean model in $\mathbb{R}^2$ could be seen as a set of points randomly distributed in the plane, where a random shape is located. Given an isotropic and stationary Boolean model in $\mathbb{R}^2$ we are interested in the following three unknown parameters, which allow us to characterize the process: the intensity λ of the germ process (number of germs per unit area); and parameters of the distribution of the primary grain, such as, the mean area of the primary grain, $a_0 = E\nu_2(\Xi_0)$, and the mean perimeter, $u_0 = EU(\Xi_0)$.

A more in-depth study of this model can be found in [72, 56, 79, 17, 4, 37]. In [5, 6, 7, 8] some estimation methods are introduced. Some previous applications of Boolean models within Image Processing are [39, 38, 23, 70, 9].

A robust method to estimate the intensity λ of the process, the mean area a_0 and the mean perimeter of the primary grain u_0 is the minimum contrast method. The minimum contrast method consists of the minimization of the difference between an estimated aggregate parameter (e.g. the capacity functional) and an approximation expressed in terms of the intensity, mean area and mean perimeter [72, 56]. An alternative is the method of moments based on the coverage fraction, mean area, boundary length and Euler-Poincare characteristic. This method is computationally easy but leads to biased estimators. Other methods have been presented to estimate the intensity, among them one based on Monte Carlo approximations of the likelihood function, and another on a stochastic version of the EM algorithm [58, 49].

The minimum contrast method has been used successfully with irregular sampling windows (see [70]). See [56, 79] for a detailed statistical analysis of Boolean models. See [37, 70] for computational details.

Let K be a convex and compact subset of $\mathbb{R}^2$ containing the origin and t a positive real number. If $tK = \{tk : k \in K\}$ (i.e., the homothetic of K) then let us define the following function

$$H_K(t) = 1 - \frac{1 - T(tK)}{1 - p}, \tag{2.13}$$

where $p = P(0 \in \Phi) = 1 - T(\{0\})$, is the area fraction of Φ (or mean area covered per unit area by the stationary random set Φ). This function is a probability distribution function known as *the contact distribution function.*

It follows that,

$$H_K(t) = 1 - exp\Big\{ - \lambda E[\nu_2(A_0 \oplus tK) - \nu_2(A)]\Big\}, \tag{2.14}$$

where K is the unit disk on the plane and A_0 the primary grain.

Let us define

$$H_K^l(t) = -\log(1 - H_K(t)). \tag{2.15}$$

Then, it is well-known [56, 79] that

$$H_K^l(t) = \lambda E[\nu_2(A_0 \oplus K) - \nu_2(A_0)]. \tag{2.16}$$

The generalized Steiner Formula establishes ([52]) that

$$E\nu_2(A_0 \oplus \breve{K}) = E\nu_2(A_0) + \frac{U(K)EU(A_0)}{2\pi} + \nu_2(K). \tag{2.17}$$

In particular, if we consider $K = B(0, t)$, the disk centered at the origin with radius t then we have

$$E\nu_2(A_0 \oplus B(0, t)) = E\nu_2(A_0) + EU(A_0)t + \pi t^2. \tag{2.18}$$

Thus,

$$\frac{H^l_{B(0,1)}(t)}{t} = \lambda\pi t + \lambda u_0. \tag{2.19}$$

Similarly, if K is taken equal to a unit square centered at the origin, $S(0,1)$, then

$$\frac{H^l_{S(0,1)}(t)}{t} = \lambda t + \frac{2}{\pi}\lambda u_0. \tag{2.20}$$

The contact distribution function can be estimated from the estimators of the capacity functional and the area fraction from Eq. 2.13. The capacity functional $T(K)$ can be estimated from a given realization (i.e. from each binary image of the sequence) by using the minus-sampling estimator given by

$$\hat{T}_W(K) = \frac{\nu_2((\Phi \oplus \check{K}) \cap (W \ominus K))}{\nu_2(W \ominus K)}, \tag{2.21}$$

where W is the sampling window, K is a disk and $A \ominus \check{K} = \{x \in A : x + K \subset A\}$.

Using the estimator given in (2.21), the left-hand side of (2.19) and (2.20) for different t values can be estimated. A linear fit using these estimated values provides us with estimates for λ and u_0. If $c_0 + c_1 t$ is the fitted function then the parameters can be estimated as

$$\hat{\lambda} = \frac{c_1}{\pi},\ \hat{u}_0 = \frac{c_0}{\hat{\lambda}}. \tag{2.22}$$

2.3 Temporal Boolean Models

Temporal Boolean models are a particular case of a non-isotropic $3D$ Boolean model. Now, we briefly introduce some interesting properties of temporal Boolean models and some parameter estimation methods. Details on proofs, methodology, simulation studies and results could be found in [9, 70]. The estimation methods proposed in this section are not used in the following studies presented in this thesis. However, they have been used in other works previously published, to which the author of this thesis has contributed, and they form the State-of-the-Art in respect of temporal Boolean models.

Let $\Psi = \{x_1, x_2, \ldots\}$ be a stationary Poisson point process in $\mathbb{R}^3$ with intensity λ. Let $\{\Xi_i\}_{i\geq 1}$ be a sequence of independent and identically distributed (as Ξ_0) random compact sets in $\mathbb{R}^3$ which are independent of the Poisson point process Ψ. Let us assume that $E\nu_3(\Xi_0 \oplus \check{K}) < +\infty$ for any K compact subset of $\mathbb{R}^3$. The 3*D Boolean model* Φ is the random set

$$\Phi = \cup_{i\geq 1}(\Xi_i + x_i), \tag{2.23}$$

where the points x_i are called *germs* and the sets Ξ_i are called *grains*, with Ξ_0 being *the primary grain.*

The distribution of Φ is determined by the *capacity functional* defined as $T(K) = \mathcal{P}(\Phi \cap K \neq \emptyset)$, with K being any compact subset. For a 3D Boolean model we have

$$T(K) = 1 - \exp\{-\lambda E\nu_3(\Xi_0) \oplus \check{K})\}. \tag{2.24}$$

Particularly, $T(0) = p = 1 - \exp\{-\lambda E\nu_3(\Xi_0)\}$ is the probability that a given point is covered by Φ, also known as the *volume fraction.* Under the hypothesis of stationarity (the distribution of the random set is invariant against translation), the intensity of the germ process corresponds to the mean number of points per unit volume and is denoted by λ.

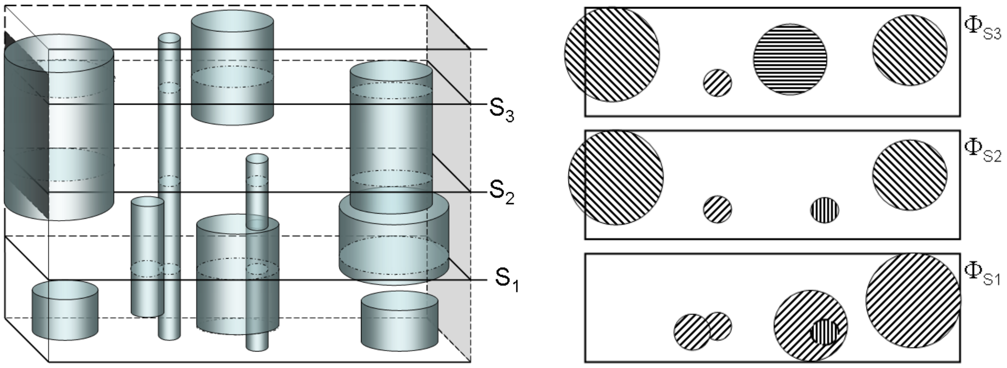
Now, let $\{(x_i, t_i)\}$ be a realization of a Poisson point process in $\mathbb{R}^2 \times \mathbb{R}_+$, where x_i denotes the location and t_i the birth time of the i-germ. The ith-grain is the rectangle $\Xi_i = A_i \times [0, d_i]$, where $\{A_i\}_{i\geq 1}$ is a sequence of independent and identically distributed (as A_0) random compact sets in $\mathbb{R}^2$ and $\{d_i\}_{i\geq 1}$ is a sequence of independent and identically distributed (as D) positive random variables independent of the Poisson point process Ψ and independent of the sequence $\{A_i\}_{i\geq 1}$. The *temporal Boolean model* Φ is defined as

$$\Phi = \cup_{i\geq 1}(A_i + x_i) \times [t_i, t_i + d_i]. \tag{2.25}$$

A shape A_i and a duration d_i are associated with each point (x_i, t_i). The set $(A_i + x_i) \times [t_i, t_i + d_i]$ is a rectangle in $\mathbb{R}^2 \times \mathbb{R}_+$, the i-th event.

Fig. 2.5 (a) displays a realization of a temporal Boolean model where A_0 is a disc of random radius. The vertical axis corresponds to time, whereas the horizontal plane is the image plane. The union of these cylinders is a realization of the temporal Boolean model. It is important to note that there is temporal and spatial overlapping. The degree of overlapping depends on both the number of grains per unit area and time and their durations. Due to the time discretization, a given grain could be born between two consecutive temporal cross sections but it would not be observed (if it was still alive) until the next cross section (frame).

Given a temporal Boolean model Φ, we will denote $\Phi_s = \Phi \cap \mathbb{R}^2 \times \{s\}$, i.e. Φ_s is a *temporal cross-section* of Φ at time s. Fig. 2.5 (b) displays some consecutive temporal cross-sections.



(a) (b)

Figure 2.5: An example of temporal Boolean model and three cross-sections.

Temporal cross-sections are (static) 2D Boolean models ([52, 79]). In fact, we developed the following proposition in [9].

Proposition 1 *If Φ is a temporal Boolean model with intensity λ and primary grain $A_0 \times [0, D]$, then the temporal cross-section Φ_s is a 2D Boolean model with primary grain A_0. The random sets Φ and Φ_s have the same area fraction. Therefore the intensity of Φ_s is $\gamma = \lambda ED$ since*

$$1 - \exp\{-\lambda E\nu_2(A_0)ED\} = 1 - \exp\{-\gamma E\nu_2(A_0)\}. \tag{2.26}$$

2.3.1 Marginal analysis of the temporal dimension

Let $\Psi = \{t_n\}_{n\geq 1}$ be a Poisson point process in $\mathbb{R}$ with intensity θ and let $\{d_n\}_{n\geq 1}$ be a sequence of independent and identically distributed non-negative random variables. If the n-th grain is born at time t_n and dies at time $t_n + d_n$, then this grain will be alive during the time interval $[t_n, t_n + d_n]$. Let $s_0 = -\infty < s_1 < s_2 < \ldots < s_m < s_{m+1} = +\infty$ be a set of fixed (previously specified) points in $\mathbb{R}$, the sampling times, where m is the length of the image sequence. Let $N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$ denote the number of points in Ψ such that $t_n \in (s_{i-1}, s_i]$ and $t_n + d_n \in (s_{j-1}, s_j]$ with $j > i$. In short, the random variable $N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$ counts the number of events in Ψ being born in $(s_{i-1}, s_i]$ and dying in $(s_{j-1}, s_j]$. Let us consider the point process in $\mathbb{R}$ defined as $\Psi_{(-\infty,s_1]}^{[s_{j-1},s_j)} = \{t_n \in \Psi : t_n \leq s_1 < s_{j-1} \leq t_n + d_n < s_j\}$. In particular, we define $\Psi_s = \{t_n : t_n \leq s < t_n + d_n\}$, i.e. the events alive at time s. Note that the point process $\Psi_{(-\infty,s_1]}^{[s_{j-1},s_j)}$ is an inhomogeneous Poisson point process with intensity function

$$r(t_n) = \theta P(s_{j-1} \leq t_n + d_n < s_j) \; if \; t_n \leq s_1 \; and \; 0 \; if \; t_n > s_1, \tag{2.27}$$

because it is a thinning of the original point process Ψ [79, 17].

Proposition 2 *The total number of points in* $\Psi_{(-\infty,s_1]}^{[s_{j-1},s_j)}$, $N_{(-\infty,s_1]}^{(s_{j-1},s_j)}$, *is a Poisson variable with mean given by,*

$$EN_{(-\infty,s_1]}^{[s_{j-1},s_j)} = \theta(p(s_{j-1} - s_1) - p(s_j - s_1)), \tag{2.28}$$

being

$$p(s) = \int_s^{+\infty} P(D \geq t)dt. \tag{2.29}$$

Note that, $p(0) = \int_0^{+\infty} P(D \geq t)dt = ED$.

As a corollary, we see that the mean number of points born before s and dead after s is equal to $EN_{(-\infty,s]}^{(s,+\infty)} = \theta p(0) = \theta ED$.

In general, the mean number of points, $N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$, which were born in the time interval $(s_{i-1}, s_i]$ and died in the time interval $(s_{j-1}, s_j]$, with $j > i$ is a Poisson variable with mean $EN_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = EN_{(-\infty,s_i]}^{(s_{j-1},s_j]} - EN_{(-\infty,s_{i-1}]}^{(s_{j-1},s_j]} = \theta(p(s_{j-1} - s_i) - p(s_j - s_i) - p(s_{j-1} - s_{i-1}) + p(s_j - s_{i-1}))$.

2.3.2 Estimation of probability density function by aggregating temporal cross-sections

The individual parameters of a temporal Boolean model are not observable due to overlapping. We can only estimate them through estimates of aggregate parameters, such as the capacity functional, which characterizes the set Φ as a whole and can be directly observed [56]. The ergodicity property ensures that the corresponding aggregate estimators converge with probability of one to the theoretical values if the sampling window expands to the whole plane.

In [9, 70], a non-parametric approach based on the accumulation of different frames is proposed to estimate the dynamic key parameters such as the probability density of durations and the intensity λ of the germ process. By accumulating different frames of the image sequence, new random sets with higher intensity are obtained. As we have mentioned, temporal cross-section of Φ, Φ_s, is also a $2D$ Boolean model. The union of Boolean models are also realizations of a Boolean model (see [56]). Then, we can construct different accumulated sequences and for each we can estimate their spatial intensity and the mean perimeter for each temporal cross-section by using the minimum contrast method ([56, 72]).

Let $\cup_{i=1}^{k}\Psi_{s_i}$ be the point process obtained by the union of all events alive in some of the temporal cross-sections $s_1, ..., s_k$. We are interested in the spatial intensity of $\cup_{i=1}^{k}\Psi_{s_i}$. The (random) number of points in $\cup_{i=1}^{k}\Psi_{s_i}$ is

$$\psi(s_1, s_2, \ldots, s_k) = EN_{(-\infty,s_1]}^{[s_1,+\infty)} + \sum_{i=2}^{k} EN_{(s_{i-1},s_i]}^{(s_i,+\infty)} = \theta\left[kp(0) - \sum_{i=2}^{k} p(s_i - s_{i-1})\right]. \quad (2.30)$$

In particular, if the times s_i are equally spaced, as usual, with $s_i - s_{i-1} = \delta$, then Eq. (2.30) becomes

$$\psi(s_1, s_1 + \delta, \ldots, s_1 + (k-1)\delta) = \theta\left[kp(0) - (k-1)p(\delta)\right]. \quad (2.31)$$

Given the original sequence $\{\Phi_{s_i} \cap W\}_{i=1,\ldots,m}$ with $s_i - s_{i-1} = \delta$, we consider the accumulated sequence defined as $\tilde{\Phi}_{s_i} = \cup_{j=i}^{i+k}\Phi_{s_j}$ with $i = 1, \ldots, m-k+1$ and k an integer, and will denote its intensity as $\lambda_s(k, \delta)$.

Note that this intensity only depends on the time between consecutive frames δ and the number of frames accumulated, k. In particular, if $\delta = 0$ and $k = 1$, we see from Eq. (2.31) that the spatial intensity for any temporal cross-section of the original sequence is $\lambda_s(0, 1) = \lambda p(0) = \lambda ED$.

For each frame in an aggregated sequence, we could estimate the intensity λ_s, applying the following methodology. Let K be a convex and compact subset of $\mathbb{R}^2$ containing the origin. If $tK = \{tk : k \in K\}$ (i.e. the homothetic of K) with $t \geq 0$, then let us consider the contact distribution function

$$H_K(t) = 1 - \frac{1 - T(tK)}{1 - p} = 1 - exp\{-\lambda_s(E[\nu_2(A_0 \oplus t\check{K})] - E\nu_2(A_0))\}, \tag{2.32}$$

with $t \geq 0$, p being the area fraction of a temporal cross-section Φ_s and λ_s the intensity of the germ process. The capacity functional $T(K)$ defined in Eq. (2.24) can be estimated from a given realization (i.e. from each binary image of the sequence) by using the minus-sampling estimator given by

$$\hat{T}_W(K) = 1 - \frac{\nu_2((W \cap \Phi^c) \ominus \check{K})}{\nu_2(W \ominus \check{K})}, \tag{2.33}$$

where W is the sampling window.

Note, that due to the minus-sampling edge-correction applied to the denominator of Eq. (2.33), the estimator $\hat{H}_K(t)$ may not be monotone. Them $H^l_K(t) = -\log(1 - H_K(t))$ should be used, yielding

$$\frac{H^l_{S(0,1)}(t)}{t} = \lambda_s t + \frac{2}{\pi}\lambda_s u_{s,0}. \tag{2.34}$$

with K being the unit square centered at the origin, $S(0, 1)$, and $u_{s,0}$ the mean perimeter of the events in a temporal cross-section. Using this estimator, $H^l_K(t)$ can be calculated for different values of t. A least square fit using these estimated values provides us with estimates for λ_s and $u_{s,0}$.

Afterwards, the batch-mean method could be applied over this set of estimates so as to calculate the mean spatial intensity of the cross-sections, $\hat{\lambda}_s(k, \delta)$ ([48, 70]). This procedure is repeated for different values of k and δ. Then, these estimates could be used to obtain new estimators which give us information about the dynamics of the process.

Eq. (2.31) becomes

$$\psi(s_1, s_1 + \delta, \ldots, s_1 + (k-1)\delta) = \lambda\nu_2(W)\Big[kp(0) - (k-1)p(\delta)\Big]. \tag{2.35}$$

From the definition of $\lambda_s(k,\delta)$ it is clear that

$$\lambda_s(k,\delta) = \frac{\psi(s_1, s_1+\delta, \ldots, s_1+(k-1)\delta)}{\nu_2(W)} = \lambda\Big[kp(0) - (k-1)p(\delta)\Big]. \tag{2.36}$$

Let us rewrite Eq. (2.36) as

$$\lambda_s(k,\delta) = (\lambda p(0) - \lambda p(\delta))k + \lambda p(\delta). \tag{2.37}$$

If we consider the right hand side of Eq. (2.37) as a function of the number of frames k, then the coefficient associated with k and the constant have simple interpretations. Let us denote

$$\alpha(\delta) = \lambda p(0) - \lambda p(\delta), \tag{2.38}$$

$$\beta(\delta) = \lambda p(\delta). \tag{2.39}$$

$\alpha(\delta)$ represents the mean number of new grains which are born between two frames δ apart. Note that due to the stationarity of the process, it can also be interpreted as the mean number of dead grains between these two frames. The function $\beta(\delta)$ represents the mean number of germs which remain alive between two frames separated by δ frames. Then Eq. (2.37) can be reformulated as $\lambda_s(k,\delta) = \alpha(\delta)k + \beta(\delta)$.

We can fix δ and calculate the estimates of $\hat{\lambda}_s(k,\delta)$ under different $k's$. A weighted least square fit could be performed where the predictor is k and the dependent variable is $\hat{\lambda}_s(k,\delta)$. That fit is repeated for different $\delta \in \{\delta_1, \ldots, \delta_n\}$.

Finally, estimates of $\hat{\alpha}(\delta_i)$ and $\hat{\beta}(\delta_i)$ with $i = 1, \ldots, n$ will be used, in turn, to obtain estimates of the spatial temporal intensity $\hat{\lambda}$ as follows,

$$\alpha'(0) = -\lambda p'(0) = \lambda P(D \geq 0) = \lambda, \tag{2.40}$$

where $\alpha'(0)$ denotes the first derivative of the function $\alpha(\delta)$ at the origin. Furthermore, since $\lambda_s(0,1) = \lambda p(0)$ can be obtained from the data, $p(\delta)$ can be found from Eq. (2.47), as we will see later. Finally the density of durations $\hat{f}_D$ can be estimated as,

$$
\begin{aligned}
&p'(\delta) = -P(D \geq \delta) = -(1 - P(D \leq \delta)) = F_D(\delta) - 1, \\
&p''(\delta) = f_D(\delta),
\end{aligned} \tag{2.41}
$$

where $p''(\delta)$ is the second derivative of $p(\delta)$.

2.3.3 Estimation of probability density function by combining temporal cross-sections

Now, we will present another method to estimate the probability density of the event duration D of a temporal Boolean model, f_D. It could be estimated by combining two discrete temporal cross sections and their negated images. (See [9]). Let us denote the temporal cross-section at time s as $\Phi_s = \Phi \cap (\mathbb{R}^2 \times \{s\})$, for the times $s_1 \leq \ldots \leq s_k$. Each Φ_{s_i} is contained in the product space $W \times [0, T]$, where W is the observation window and $[0, T]$ is the total time interval observed.

Let us denote Φ_s also the projection of Φ_s over $\mathbb{R}^2$, i.e., the binary image observed at time s. It is given by:

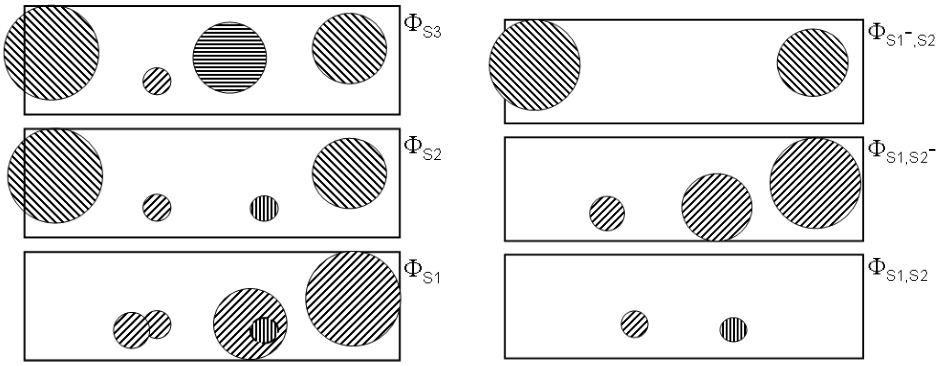
$$
\Phi_s = \cup_{\{i: t_i \leq s \leq t_i + d_i\}} (A_i + x_i). \tag{2.42}
$$

For $s_1 \leq s_2$, we will consider the following three random sets:

$$
\begin{aligned}
\Phi_{s_1, s_2} &= \cup_{i: t_i \leq s_1 \leq s_2 \leq t_i + d_i} A_i + x_i, && (2.43) \\
\Phi_{s_1, s_2^-} &= \cup_{i: t_i \leq s_1 \leq t_i + d_i < s_2} A_i + x_i, && (2.44) \\
\Phi_{s_1^-, s_2} &= \cup_{i: s_1 < t_i \leq s_2 \leq t_i + d_i} A_i + x_i. && (2.45)
\end{aligned}
$$

Φ_{s_1,s_2} corresponds to the union of the grains which are in Φ_{s_1} and Φ_{s_2}, Φ_{s_1,s_2^-} to the grains in Φ_{s_1} but not in Φ_{s_2}, and finally $\Phi_{s_1^-,s_2}$ to the grains in Φ_{s_2} but not in Φ_{s_1} (see Fig. 2.6 (b)). This approximation combines two cross-sections of the 3D Boolean model to obtain new information of the underlying model, in contrast with other studies in which sections are analyzed separately [68, 26].

In [9], the following theorems and propositions are introduced. Detailed proofs can be found in [9].



(a) (b)

Figure 2.6: Three consecutive cross-sections of a Temporal Boolean model and the three Boolean models derived.

Theorem 1 *Let* Φ *be a temporal Boolean model. The random sets* Φ_{s_1,s_2}*,* Φ_{s_1,s_2^-} *and* $\Phi_{s_1^-,s_2}$ *are three independent 2D Boolean models. Let us denote* $\beta(s_2-s_1)$ *the intensity of* Φ_{s_1,s_2} *(the mean number of germs that are alive in the time interval* $[s_1,s_2]$*), which is given by*

$$\beta(s_2-s_1)=\lambda p(s_2-s_1), \tag{2.46}$$

where $p(s)=\int_s^{+\infty}P(D\geq v)dv$*. Moreover, under stationarity,* Φ_{s_1,s_2^-} *and* $\Phi_{s_1^-,s_2}$ *have the same intensity,* $\alpha(s_2-s_1)$ *(the mean number of germs that died in the time interval* $[s_1,s_2]$*) given by*

$$\alpha(s_2-s_1)=\lambda ED-\lambda p(s_2-s_1)=\gamma-\beta(s_2-s_1). \tag{2.47}$$

It holds that $\gamma=\alpha(s)+\beta(s)$, i.e. the mean number of grains per unit area that appear and remain through a time interval of length s is constant and is equal to the mean number of germs per unit area in each frame. The sets Φ_{s_1,s_2}, Φ_{s_1,s_2^-} and $\Phi_{s_1^-,s_2}$ cannot be observed due to the overlapping between the different events. However, note that we can observe $\Phi_{s_1}=\Phi_{s_1,s_2}\cup\Phi_{s_1,s_2^-}$, $\Phi_{s_2}=\Phi_{s_1,s_2}\cup\Phi_{s_1^-,s_2}$, $\Phi_{s_1}\setminus\Phi_{s_2}=\Phi^c_{s_1,s_2}\cap\Phi_{s_1,s_2^-}\cap\Phi^c_{s_1^-,s_2}$ and $\Phi_{s_2}\setminus\Phi_{s_1}=\Phi^c_{s_1,s_2}\cap\Phi_{s_1^-,s_2}\cap\Phi^c_{s_1,s_2^-}$, where $A\setminus B=A\cap B^c$ denotes the set substraction.

Proposition 3

$$P(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2}) = P(0 \in \Phi_{s_2} | 0 \notin \Phi_{s_1}) = 1 - \exp\{-\alpha(s_2 - s_1)a_0\}. \tag{2.48}$$

Then $P(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2})$ will be estimated as

$$\hat{P}(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2}) = \frac{\hat{P}(0 \in \Phi_{s_1} \setminus \Phi_{s_2})}{\hat{P}(0 \notin \Phi_{s_2})} = \frac{\nu_2(\phi_{s_1} \cap \phi_{s_2}^c \cap W)}{\nu_2(\phi_{s_2}^c \cap W)}, \tag{2.49}$$

and, by using (2.48) and (2.49), we will estimate $\delta(s_2 - s_1) = \alpha(s_2 - s_1)a_0$ solving the equation

$$1 - \exp\{-\hat{\delta}(s_2 - s_1)\} = \frac{1}{2}\left(\frac{\nu_2(\phi_{s_1} \cap \phi_{s_2}^c \cap W)}{\nu_2(\phi_{s_2}^c \cap W)} + \frac{\nu_2(\phi_{s_2} \cap \phi_{s_1}^c \cap W)}{\nu_2(\phi_{s_1}^c \cap W)}\right). \tag{2.50}$$

The mean area (a_0) and the mean perimeter (u_0) of A_0 and γ were estimated from each image in the sequence by using the minimum contrast method [56]. The different estimates are correlated values of the same parameters since the model is stationary. The different estimates over the image sequence were then combined by using the batch-mean method [48].

Let $\hat{a}_0$ and $\hat{\gamma}$ be the estimates. The functions $\alpha(s)$ and $\beta(s)$ can be estimated as

$$\hat{\alpha}(s) = \frac{\hat{\delta}(s)}{\hat{a}_0} \text{ and } \hat{\beta}(s) = \hat{\gamma} - \hat{\alpha}(s). \tag{2.51}$$

From 2.47, it follows that $\alpha'(s) = \lambda(1 - F_D(s))$ and $\alpha''(s) = -\lambda f_D(s)$ where α' and α'' are the first and second derivatives of α, whereas F_D and f_D are the cumulative distribution function and the density function of the random variable D, respectively.

The function $\alpha(s)$ can be estimated at the observation points $\{s_1, \ldots, s_n\}$. Its first and second derivatives at any point could be estimated by making use of functional data analysis. A functional datum is a set of discrete measured values $\{(s_j, y_j)\}_{j=1,\ldots,n}$. First, it is necessary to convert these values to a function which is computable for any value. In [9], an interpolation process was not used because the discretization might include some observational error. Instead, a smoothing technique was used to transform the raw data $\{(s_j, y_j)\}_{j=1,\ldots,n}$ to a function $y(t) = \sum_{k=1}^{K} c_k \Phi_k(t)$, being

$\{\Phi_k(t)\}_{k=1,\ldots,K}$ a basis functions possessing a certain number of derivatives. A polynomial spline basis was chosen, where each $\Phi_k(t)$ is a piecewise cubic function. The coefficients c_k of the expression $y(t)$ were obtained by minimizing the least squares criterion $SMSSE(y|c) = \sum_{j=1}^{n}(y_j - \sum_{k=1}^{K} c_k\Phi_k(t_j))^2$. A detailed presentation of this method is given in [64]. This procedure has been applied to the estimated values $\{(s_j, \hat{\alpha}(s_j))\}_{j=1,\ldots,n}$ to obtain an estimated function $\hat{\alpha}(s)$ and to calculate the first and second derivatives of the fitted values. The spatial-temporal intensity λ was estimated by taking into account that $\alpha'(0) = -\lambda p'(0) = \lambda P(D \geq 0) = \lambda$, yielding

$$\hat{\lambda} = \hat{\alpha}'(0). \tag{2.52}$$

The probability density of D, f_D, was estimated as

$$\hat{f}_D(s) = -\frac{1}{\hat{\lambda}}\hat{\alpha}''(s). \tag{2.53}$$

Chapter 3

Estimation of the duration distribution in temporal Boolean models using the spatial-temporal covariance

In order to study the shape-size features and durations of the events, it is a usual practice to analyze only isolated episodes. However, this sample is biased because faster and smaller events tend to be isolated. We model the images as a realization of a cylindrical temporal Boolean model. We evaluate the bias introduced when ruling out non-isolated episodes. We propose an estimator of the duration distribution and perform a simulation study to assess its accuracy. The method is applied to fluorescent-tagged proteins image sequences. Results show that this procedure is effective for analyzing dynamic processes where spatial and temporal overlapping occurs.

3.1 Introduction

The problem of estimating shape-size characteristics of overlapping objects from images is common in many real applications. Boolean models explicitly consider and assume this overlap, providing a good description for very irregular patterns observed in Microscopy, Material Sciences, Biology, Medicine, Chemistry, Geostatistics, Cellular Communications Networks or Image Processing, since they formalize the configuration of independent and randomly located particles. What it is observed is a pattern of overlapping random shapes, as shown in Fig. 3.1. More precisely, a Boolean model is a random closed set consisting of a Poisson point process producing the *germs* coupled with an independent random shape process i.e. a sequence of independent and identically distributed random closed sets, the *grains*. The connected components made of overlapping sets are called *clumps*. Definitions and statistical analysis of these models can be found in [51, 52, 57, 72, 79]. Previous applications of Boolean models within Image Processing, in particular for texture analysis and classification, are [23, 39, 33].

In spite of the considerable number and variety of applications of Boolean models, the problem of analyzing overlapping events in space and time from image sequences remains elusive. More sophisticated mathematical models and statistical methods are needed i.e. hybrid models that capture both time and geometric properties while formalizing a configuration of independent randomly placed particles with random durations. Fig. 3.1 displays several consecutive frames of a simulated model where discs are randomly located in space and time with random radii and random durations (see Video1 in supplementary material). Fig. 3.2 depicts a spatial-temporal reconstruction of the model. These sophisticated data require specialized methods for effective analysis, in particular for extracting useful information and for the interpretation of results.

In applied fields, such as Cell Biology, the estimation of shape-size features and durations of overlapping events from image sequences is often confined to visual inspection or limited statistical analysis typically done manually on a one-by-one basis of isolated events, those that can be completely observed from the beginning to the end of their lifetimes (see Fig. 3.1). However, this procedure presents several drawbacks. First, each experiment involves dozens of sequences with thousands of frames each, making manual analysis impossible in large image sequences. Second, manual marking of isolated events is somewhat subjective. Third, the use of only isolated events leads to a biased sample i.e. one that includes smaller and shorter events with higher probability. The evaluation of this bias is not possible without the assumption of a stochastic model about the (random) mechanism generating the locations, times of occurrence and durations of the events. In this study we define a stochastic model, the cylindrical temporal Boolean model, which will allow us to assess the bias introduced when excluding non-isolated events. We will propose an estimator of the cumulative distribution function of the duration of the events based on the spatial-temporal covariance of random sets.

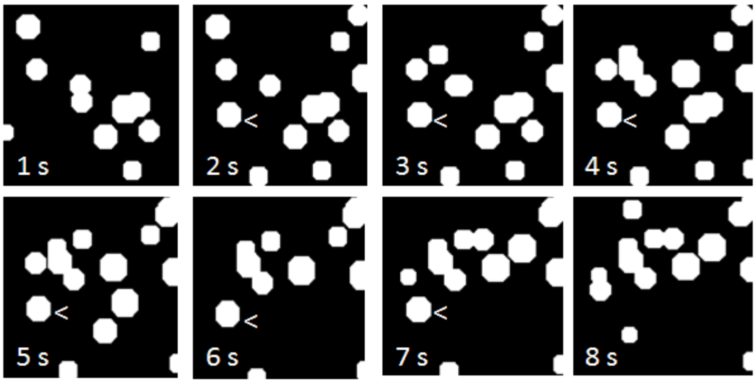


Figure 3.1: Several consecutive frames of a simulated temporal Boolean model with cylindrical grains. Arrows point to an isolated episode.

The procedure will be applied to fluorescent-tagged endocytic proteins image sequences in order to estimate the duration of cell endocytosis. TIRFM illuminates a thin section near the cell-coverslip interface and gives a very high signal-to-noise ratio, thus facilitating visualization of cellular processes near the plasma membrane ([83]). Under TIRFM, the assembly of fluorescently-labeled proteins at a site of ongoing endocytosis results in the appearance and steady growth of a diffraction-limited spot, in such a way that the areas of fluorescence of endocytic events overlap, forming random clumps in space and time (see Fig. 3.3 and Video Cell1 in supplementary material).

Section 3.2 contains the definition of the cylindrical temporal Boolean model and some theoretical results on the statistical analysis of isolated events. In Section 3.3, a semi-parametric estimator of the distribution function of the duration based on the spatial-temporal covariance of random sets is introduced. In Section 3.4 a simulation study is performed. Section 3.5 shows an application to biological image sequences of cell endocytosis. Finally, conclusions and further developments are summarized in Section 3.6.

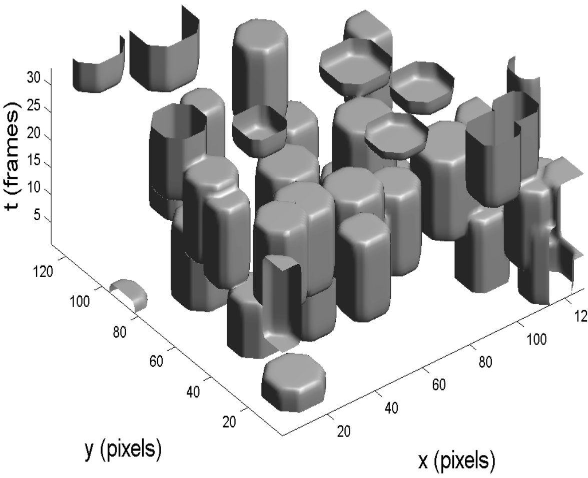


Figure 3.2: A spatial-temporal reconstruction of a cylindrical temporal Boolean model.

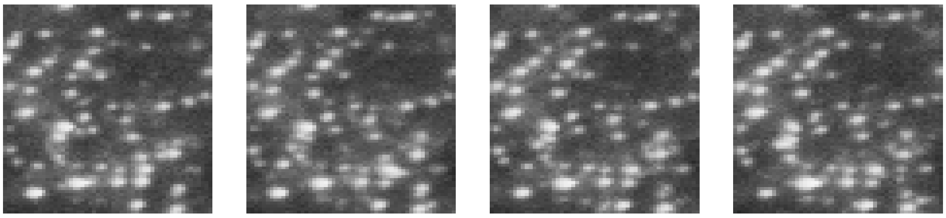


Figure 3.3: Several snapshots of a fluorescent-tagged protein image sequence.

3.2 The model

This model is a particular case of the temporal Boolean model ([9, 70]). Graphically, a temporal Boolean model consists of a set of events with location x_i and occurrence time t_i, with random duration d_i and random shape Ξ_i, which overlap in space and time. If the spatial grains are disks of random radii, then the events form cylinders in $\mathbb{R}^2 \times \mathbb{R}_+$. We can observe these kind of random cylinders in a TIRFM image sequence of a fluorescent-tagged protein, since the snapshots remain alive several frames. Fig. 3.1 displays some consecutive temporal cross-sections of a simulated model. The union of these cylinders is a realization of the model (see Fig. 3.2). Let us now introduce some notation and definitions.

Definition 1 (Cylindrical Temporal Boolean Model) *Let* $\Psi = \{(x_i, t_i)\}_{i \geq 1}$ *be a stationary Poisson point process in* $\mathbb{R}^2 \times \mathbb{R}_+$ *with intensity* λ*. Let* $\{(R_i, D_i)\}_{i \geq 1}$ *be a sequence of independent and identically distributed non-negative random vectors (as* (R, D)*) with joint density* f*. We assume that* Ψ *and* $\{(R_i, D_i)\}_{i \geq 1}$ *are independent and that* $E\nu_3(B(0,R) \times [0,D] \oplus \check{K}) < +\infty$ *for any compact subset* K *of* $\mathbb{R}^2 \times \mathbb{R}_+$*. The* cylindrical temporal Boolean model *is the random set defined as*

$$\Phi = \bigcup_{i \geq 1} B(0, R_i) + x_i \times [t_i, t_i + D_i], \tag{3.1}$$

where $B(0, R_i) + x_i$ is the translation of $B(0, R_i)$ at x_i. The set $B(0, R_i) + x_i \times [t_i, t_i + D_i]$ is a cylinder in $\mathbb{R}^2 \times \mathbb{R}_+$, the i-th event.

The probability that a compact subset K of $\mathbb{R}^2 \times \mathbb{R}_+$ is contained in the complement of Φ is given by

$$Q(K) = P(K \subset \Phi^c) = \exp\Bigg(-\lambda E\nu_3 \Big[(B(0,R) \times [0,D]) \oplus \check{K} \Big] \Bigg), \tag{3.2}$$

where $\check{K}$ is the symmetric of K with respect to the origin. If we take $K = \{0\}$ then, the volume fraction is given by

$$p = P(0 \in \Phi) = 1 - Q(\{0\}) = 1 - \exp\{-\lambda \pi E(R^2 D)\}. \tag{3.3}$$

This probability p is called the volume fraction of Φ and it corresponds to the mean volume covered by the random set per unit area and unit time.

3.2.1 The duration distribution of isolated events

Let us assume that the joint distribution of (R, D), the random radius and duration of a typical grain, has a density $f(r, d; \theta)$ where θ are the parameters within a parametric space Θ. However, due to the overlapping among shapes, we cannot observe all events. We can only segment the *isolated events*, i.e those connected components with a smooth boundary, which are either not overlapped by any other event or are hit by another one (or other ones) in such a way that the hitting events lie within it. Such as isolated events can be completely observed from the beginning to the end of their lifetimes as a cylindrical shape.

Let us denote by (R^*, D^*) the random vector associated with an arbitrary isolated event and f^* be its corresponding density function. In order to study the bias in the estimation of the durations using a sample composed of isolated events, it is necessary to relate the densities f and f^*. Note that f corresponds to the observable and non-observable events and f^* to the isolated (and therefore observable) events. The case of the static two-dimensional cylindrical Boolean model was studied in [28].

Theorem 2 *Let* Φ *be a cylindrical temporal Boolean model with intensity* λ. *Let* (R, D) *(respectively* (R^*, D^*)*) be the random radius and duration of the typical grain (the typical isolated grain) of* Φ. *If* (R, D) *has density* $f(r, d; \theta)$ *then the density function of* (R^*, D^*) *verifies*

$$f^*(r, d; \lambda, \theta) \propto f(r, d; \theta)p(r, d; \lambda, \theta), \tag{3.4}$$

where

$$p(r, d; \lambda, \theta) = \exp\bigg(-\lambda \int_0^{+\infty} \int_0^{+\infty} C(u, v; r, d) f(u, v; \theta) du dv \bigg). \tag{3.5}$$

The function C is defined as $C(u, v; r, d) = \nu_3\big(B(0, u) \times [0, v] \oplus S_0(r) \times [0, d]\big)$, being $S_0(r)$ a disc de radius r. The function C is equal to

$$C(u, v; r, d) = \begin{cases} \pi(u + r)^2(v + d) & \text{if } u \geq r \text{ or } \ v \geq d, \\ 2\pi(r - u)^2 v + 4\pi ur(v + d), & \text{if } u < r \text{ and } \ v < d. \end{cases} \tag{3.6}$$

The function $p(r, d; \lambda, \theta)$ gives the probability that a cylinder of radius r and duration d remains as an isolated event in the observed pattern. This is equal to the probability that its boundary is not hit by any other grain i.e. it belongs to the complement of Φ.

Proof. If $(R, D) = (r, d)$, the probability $p(r, d; \lambda, \theta)$ that the boundary of the cylinder $S_0(r) \times [0, d]$ is not covered by any other event can be obtained from Eq. (3.2) as

$$p(r, d; \lambda, \theta) = \exp\Big\{ -\lambda E_\theta\big[\nu_3\big(B(0, R) \times [0, D] \oplus S_0(r) \times [0, d]\big)\big]\Big\}. \tag{3.7}$$

When considering hits by a cylinder of radius $u \geq r$ or duration $v \geq d$, it holds that

$$B(0,u) \times [0,v] \oplus S_0(r) \times [0,d] = B(0,u+r) \times [0,v+d].$$

If $u < r$ and $v < d$, we have

$$\begin{aligned} &B(0,u) \times [0,v] \oplus S_0(r) \times [0,d] = \\ &= \Big[\big(B(0,u+r) \backslash B(0,u-r)\big) \times [0,v+d]\Big] \bigcup \Big[B(0,u+r) \times ([0,v+d] \cup [d-v,d])\Big] \end{aligned} \tag{3.8}$$

and the proof is completed.

□

Although we have assumed cylindrical shapes for the grains, the respective expressions for other kind of random shapes can be likewise derived.

The remaining sample fraction, defined as the ratio of isolated events with respect to all the grains, is given by

$$\delta(\lambda, \theta) = \int_0^{+\infty} \int_0^{+\infty} p(u, v; \lambda, \theta) f(u, v; \theta) du dv. \tag{3.9}$$

We have no closed expression for the density $f^*(\cdot, \cdot; \lambda, \theta)$. However, Theorem 2 opens a simple way to generate random samples of f^*. First, we generate $(R, D) = (r, d)$ with density $f(\cdot, \cdot; \lambda, \theta)$. Second, every generated pair (r, d) is accepted with probability $p(r, d; \lambda, \theta)$ or rejected with probability $1 - p(r, d; \lambda, \theta)$. The accepted (r, d) pairs have density function $f^*(\cdot, \cdot; \lambda, \theta)$. By taking into account that

$$p(r, d; \lambda, \theta) = \exp(-\lambda E[C(R, D; r, d)]),$$

we can obtain an estimate of $p(r, d; \lambda, \theta)$ by computing $E[C(R, D, r, d)]$.

Let $(R_1, D_1), \ldots, (R_m, D_m)$ be a random sample with density f, then the Monte Carlo estimator of $E[C(R, D, r, d)]$ is given by

$$\hat{E}[C(R, D; r, d)] = \sum_{i=1}^{m} \frac{C(R_i, D_i; r, d)}{m}.$$

In summary, we will consider the estimator

$$\hat{p}(r, d; \lambda, \theta) = \exp\left(-\lambda \sum_{i=1}^{m} \frac{C(R_i, D_i; r, d)}{m} \right). \tag{3.10}$$

Given a model with parameters λ and θ, we can generate samples of isolated episodes with durations D^* for fixed λ and θ, say $d_1^*, \ldots, d_m^*$, and we can compute the error in the estimation of the distribution function of the duration.

3.2.2 Evaluating the marginal duration distribution of isolated events

In this subsection, the original density of the durations considering all the events f and the density of the isolated episodes f^* are compared. Three cylindrical temporal Boolean models were simulated corresponding to three duration distributions with mean 6: uniform distribution, $U(4, 8)$; Gamma distribution, $Ga(18, 0.3333)$, and normal distribution, $N(6, 1)$. The radii of the grains were uniformly distributed in the interval $[4, 8]$ pixels. The durations and the radii were independently generated.

Fig. 3.4 shows the estimates of f^* for the three models. The solid line corresponds to f, the dashed line represents a remaining sample fraction δ of around 58%, the dotted line of approximately 20% and the dashed-dotted line of around 10%. Note how the densities tend to be concentrated in the left side of the interval. The higher the remaining sample fraction is, the smaller the bias is.

3.2.3 Error in the estimation of the mean duration using isolated events

The relative error of the mean duration, defined as the difference between the estimated and theoretical values with respect to the theoretical value, was calculated as a function of the remaining sample fraction, δ. We used the same models as in the previous subsection. We considered values of the intensity of the germ process ranging in the interval $[0.0001, 0.001]$ with a step of 0.0001 resulting in 10 values of δ ranging from 0 to 0.8. A total number of 40 replicates for each value of δ and model were simulated. Fig. 3.5 displays the 0.05 and 0.95 percentiles (dotted lines) and the median

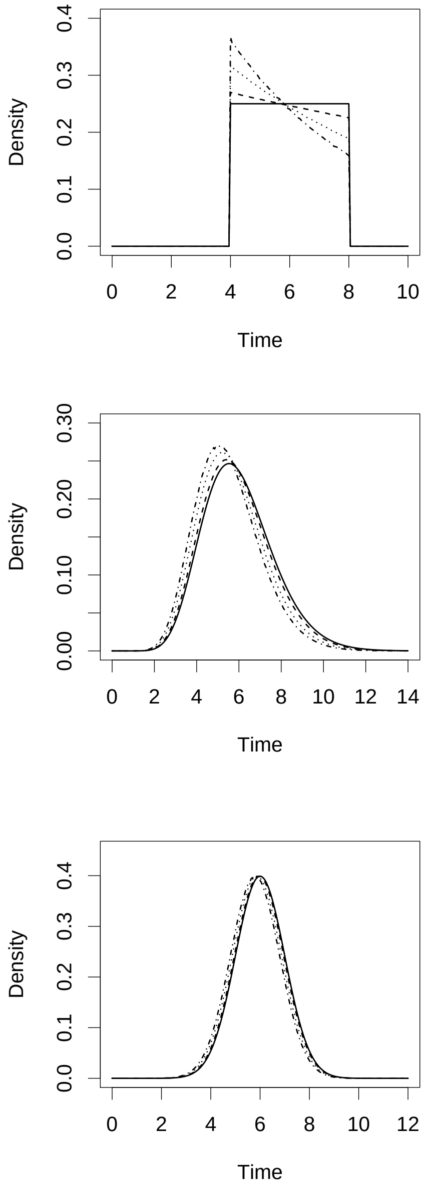


Figure 3.4: Estimated densities f^* of the random duration D^* for different remaining sample fractions. From top to bottom: uniform distribution, $U(4, 8)$; Gamma distribution, $Ga(18, 0.3333)$, and normal distribution, $N(6, 1)$. Durations are in unit time.

(solid line) of the relative error (in percentage) computed from the 40 realizations. When the remaining sample fraction is low i.e. there are many grains overlapped, the relative errors are high. For the uniform distribution the error is lesser than 9%, for the Gamma distribution is approximately 12% and for the normal distribution is around 6%. These values could be unacceptable in many real applications. In Section 3.3, we propose an estimator of the cumulative distribution function of the duration of a cylindrical temporal Boolean model using all the events i.e the observable and non-observable episodes.

3.3 The covariance of a cylindrical temporal Boolean model

In this section we present an estimator of the cumulative distribution function of the duration based on the covariance of a cylindrical temporal Boolean model. We define the *spatial-temporal covariance* of a (spatial-temporal) stationary random set as

$$\mathbb{C}(\mathbf{h}, t) = P(\mathbf{0} \in \Phi(0), \mathbf{h} \in \Phi(t)). \tag{3.11}$$

For a given $\mathbf{h} \in \mathbb{R}^2$ and a given time $t \in \mathbb{R}_+$, this function gives the probability that an arbitrary point at an arbitrary time and that point translated by a vector $\mathbf{h}$ and a time t belong to the random set. The spatial origin $\mathbf{0}$ denotes an arbitrary location and the temporal origin 0 denotes an arbitrary time. Fig. 3.6 provides a geometrical interpretation. If the distribution of Φ is invariant against spatial translations then $\mathbb{C}$ only depends on the modulus h, $h = \|\mathbf{h}\|$.

Proposition 4 *Let Φ be a cylindrical temporal Boolean model Φ with intensity λ and primary grain $\Phi_0 = B(0, R) \times [0, D]$, where the radii and durations are independent. Let f_D be the density of the durations. The spatial-temporal covariance of Φ is*

$$\mathbb{C}(\mathbf{h}, t) = (2p - 1) + (1 - p)^2 \exp\big(\lambda C_B(\mathbf{h}) L(t)\big), \tag{3.12}$$

where $C_B(\mathbf{h}) = E\nu_2\big(B(0, R) \cap B(0, R) + \mathbf{h}\big)$, $L(t) = ED - \int_0^t u F_D(du) - tP(D \geq t)$ and F_D is the cumulative distribution function of the duration.

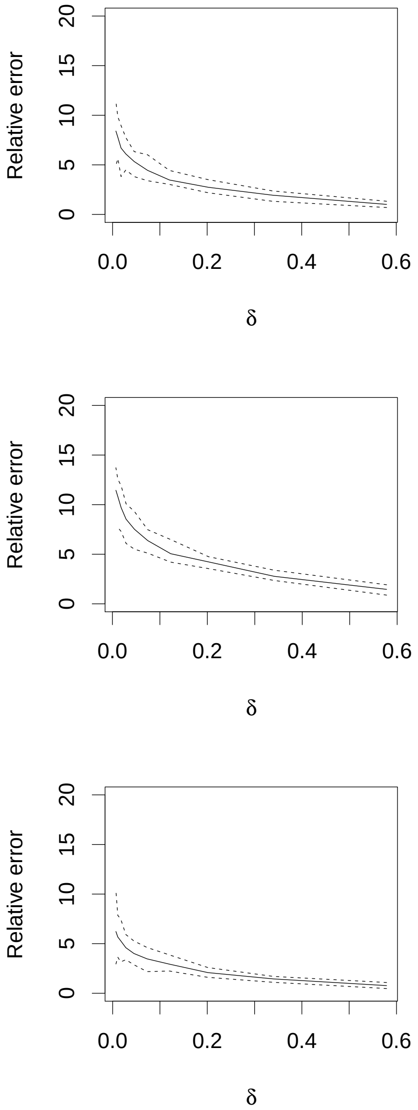


Figure 3.5: Relative errors in the estimation of the mean duration as a function of δ. From top to bottom: uniform distribution, $U(4,8)$; Gamma distribution, $Ga(18,0.3333)$, and normal distribution, $N(6,1)$.

Proof. A cylindrical temporal Boolean model is a particular case of a non-isotropic three-dimensional Boolean model. By taking into account the general expression of the covariance ([51, 72, 79]), we have

$$\mathbb{C}(\mathbf{h},t) = 2p - 1 + (1-p)^2 \exp\{\lambda E\nu_3\big(\Phi_0 \cap \Phi_0 + (\mathbf{h},t)\big)\}. \tag{3.13}$$

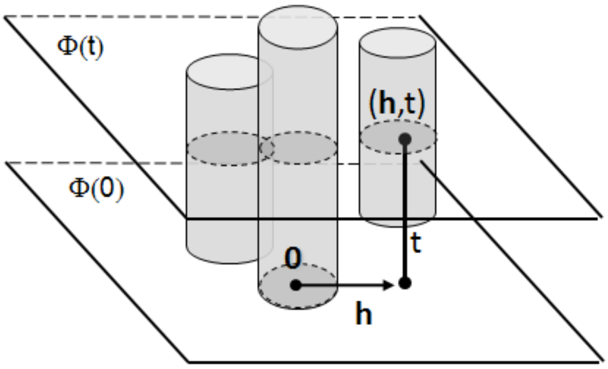


Figure 3.6: The spatial-temporal covariance of a temporal random set.

Let us denote $A_0 = B(0,R)$, then $\Phi_0 \cap \Phi_0 + (\mathbf{h}, t) = (A_0 \cap A_0 + \mathbf{h}) \times ([0,D] \cap [t, t+D])$ if $t < D$ and $\emptyset$ otherwise. It holds that

$$E\nu_3\big(\Phi_0 \cap \Phi_0 + (\mathbf{h},t)\big) = E_{A_0,D}\Big(\nu_2\big(A_0 \cap A_0 + \mathbf{h}\big) \times \nu\big([0,D] \cap [t, D+t]\big)\Big)$$
$$= E_{A_0}\nu_2\big(A_0 \cap A_0 + \mathbf{h}\big) E_D \nu\big([0,D] \cap [t, D+t]\big), \quad (3.14)$$

because A_0 and D are independent. Finally,

$$L(t) = E_D\nu\big([0,D] \cap [t, D+t]\big) =$$
$$\int_t^{+\infty} (u-t) F_D(du) = ED - \int_0^t u F_D(du) - tP(D \geq t). \quad (3.15)$$

$\square$

If we assume that f_D, the derivative of F_D, exists, then from Eq. (3.15) it follows that

$$L'(t) = -tf_D(t) - (1 - F_D(t)) + tf_D(t).$$

Therefore,

$$F_D(t) = 1 + L'(t) \quad (3.16)$$

and

$$f_D(t) = L''(t).$$

We will estimate F_D in the next section using Eq.(3.16) ([21]).

3.3.1 Estimation of the distribution function of the duration

Our data consist of a discrete set of temporal cross-sections corresponding to the observation times, $t \geq 0 : \Phi = \bigcup_{t \geq 0} \Phi(t)$, as shown in Fig. 3.1.

For a temporal Boolean model it holds that $\gamma = \lambda ED$ ([9]), where γ is the mean number of germs per unit area in any frame i.e. the intensity of an arbitrary temporal cross-section. From Eq. (3.13), we have

$$G(h,t) = \log\left(\frac{\mathbb{C}(h,t) - 2p + 1}{(1-p)^2}\right) = \lambda C_{A_0}(h)L(t), \tag{3.17}$$

where $C_{A_0}(h) = E_{A_0}\nu_2\big(A_0 \cap A_0 + h\big)$. We define

$$H(t) = G(0,t) = \lambda a_0 L(t), \tag{3.18}$$

where $a_0 = E\nu_2(A_0)$. Note that $L(0) = ED$ and $L'(0) = -1$. Therefore, $H'(0) = -\lambda a_0$. Then it holds that

$$L(t) = -\frac{H(t)}{H'(0)}, \tag{3.19}$$

and from Eq. (3.16), we calculated the cumulative distribution function of the duration as

$$F_D(t) = 1 - \frac{H'(t)}{H'(0)}. \tag{3.20}$$

3.3.2 Estimators

We replace the (unknown) continuous version of the covariance function by the corresponding discrete version. In order to simplify notation, we assume that the frames observed correspond to times $t = 0, \tau, 2\tau, \ldots, n\tau$ i.e. $n+1$ equally spaced frames. Our sampling information consists of the sequence of sets $\{\Phi(k\tau) \cap W\}_{k=0,\ldots,n}$ where τ is the temporal delay between two consecutive frames and W is the sampling window.

The spatial-temporal covariance is estimated as

$$\hat{\mathbb{C}}(h, r\tau) = \frac{1}{n-r} \sum_{k=0}^{n-r} \frac{\nu_2\Big(\big(\Phi(k\tau) \cap W\big) \cap \big(\Phi((k+r)\tau) \cap W\big) - h \Big)}{\nu_2\Big(W \cap W - h \Big)}. \tag{3.21}$$

The volume fraction p is estimated as the average of the area fractions observed at each frame and it is given by

$$\hat{p} = \frac{1}{n+1} \sum_{k=0}^{n} \frac{\nu_2\big(\Phi(k\tau) \cap W\big)}{\nu_2\big(W\big)}. \tag{3.22}$$

The function $H(t)$ is estimated at the observation points $\{t_1, \ldots, t_n\}$, a discrete set of values, giving $\{\hat{H}(t_1), \ldots, \hat{H}(t_n)\}$. We have used an approach based on functional data analysis in order to obtain a precise estimation of the first derivative from this discrete set. We have chosen a polynomial spline basis $\{\Psi_k(t)\}_{k=1,\ldots,K}$, where each $\Psi_k(t)$ is a piecewise cubic function. We computed the coefficients c_k of the expression $x(t) = \sum_{k=1}^{K} c_k \Psi_k(t)$ by minimizing the least squares criterion

$$SMSSE(x/c) = \sum_{j=1}^{n} \big(x_j - \sum_{k=1}^{K} c_k \Psi_k(t_j)\big)^2.$$

A detailed presentation of this method can be found in [64].

3.4 A simulation study

Section 3.4.1 contains a simulation study to assess the performance of the proposed estimator of the distribution function of the duration from the covariance. In Section 3.4.2 we apply the method to images with random noise.

Six cylindrical temporal Boolean models were generated ([21]), corresponding to three distributions for the duration (uniform, Gamma and normal) and two mean durations (6 and 10 seconds), see Table 3.1. The radii were uniform in the interval $[4, 8]$ pixels. Durations and radii were independent. For each model, three values of the intensity $\lambda = \{0.0001, 0.0002, 0.0003\}$ were simulated that corresponded to low, medium and high volume fractions (see Table 3.2). The volume fractions were derived using the equation $p = 1-\exp\{-\lambda\pi E(R^2 D)\}$. Image sequences were 256×256 pixels, 100 seconds long and sampled at 4 frames per second. Fifty replica for each model and value of λ were generated, thereby analyzing a total of $6 \times 3 \times 50 = 900$ sequences. Fig. 3.7 displays the first frame of the six models simulated for the uniform distribution. In supplementary material, Video1 corresponds to the uniform distribution with $\lambda = 0.0001$, Video2 to the Gamma distribution with $\lambda = 0.0002$ and Video3 to the normal distribution with $\lambda = 0.0003$ with a mean duration of 6 seconds.

Table 3.1: Description of the models simulated.

Duration	Parameters	Model
Uniform	$[4, 8]$	Model 1
	$[8, 12]$	Model 2
Gamma	$[18, 0.3333]$	Model 3
	$[50, 0.2]$	Model 4
Normal	$[6, 1]$	Model 5
	$[10, 1]$	Model 6

Table 3.2: Volume fractions simulated.

	$\lambda = 0.0001$	$\lambda = 0.0002$	$\lambda = 0.0003$
$ED = 6$	0.06	0.13	0.18
$ED = 10$	0.10	0.20	0.29

3.4.1 Estimating the cumulative distribution function

First, we analyzed the absolute error in the estimation of the cumulative distribution function of the duration, F_D, for each model and intensity λ. Fig. 3.8 displays the maximum and the minimum absolute error (solid lines) and the median (dashed line) of $\hat{F}_D$ for 50 replica. The first row corresponds to $ED = 6$ and the second row to

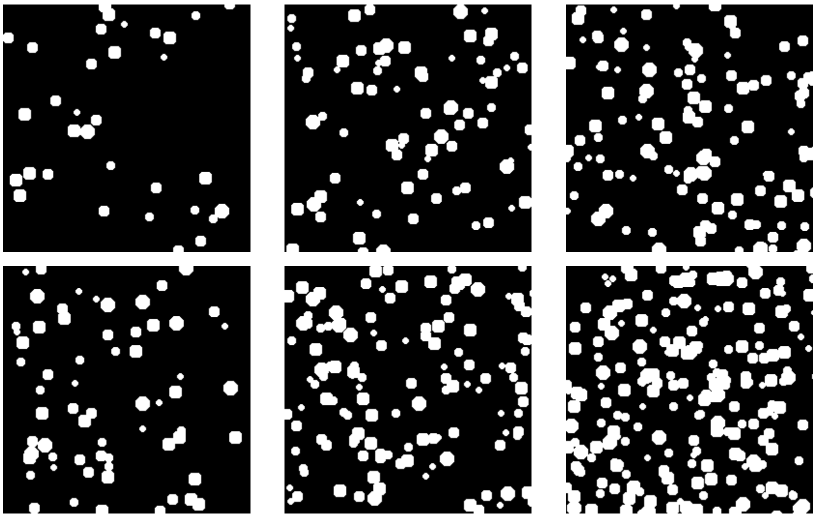


Figure 3.7: The first row corresponds to $ED = 6$ and the second row to $ED = 10$. From left to right, the columns correspond to $\lambda = 0.0001, 0.0002, 0.0003$. Uniform distribution for the durations.

$ED = 10$. Columns correspond to the uniform, Gamma and normal distributions, respectively. The intensity λ is 0.0003. Errors are quite small, lesser than 0.06 except for the uniform distribution $U(8, 12)$. The highest values are observed next to the median of the distribution. Similar results were obtained for $\lambda = 0.0001$ and $\lambda = 0.0002$.

Second, we computed the maximum and the minimum deviations of the 50 realizations from the theoretical distribution function F_D. Fig. 3.9 plots the upper and lower envelopes (solid lines) as a function of the theoretical distribution function F_D. The theoretical value is within the envelopes even for very small values of F_D. The intensity λ is 0.0003. Similar results were obtained for $\lambda = 0.0001$ and $\lambda = 0.0002$.

Finally, we have calculated the relative error of the median of the duration as the difference between the estimated and the theoretical values with respect to the theoretical value. The minimum, median and maximum values are given in Table 3.3. The relative errors were very small, even for high volume fractions. They were lesser than 0.8% -except for the uniform $[4, 8]$, that was 1.33%-. In our opinion, this error would be small enough in many applications. Note that the error in the case $ED = 10$ is smaller than in the case $ED = 6$, because the discretization effect of the acquisition

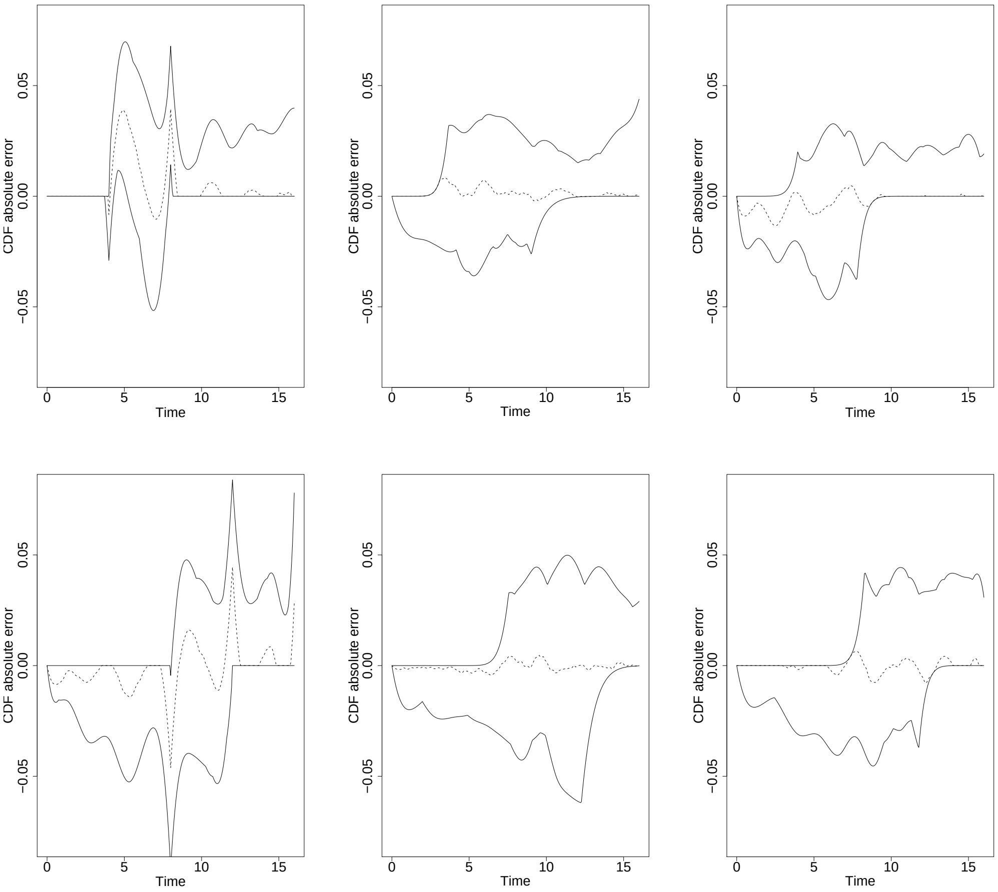


Figure 3.8: The median (dashed line), the maximum and the minimum (solid lines) of the absolute error in the estimation of F_D.

rate is lesser in relative terms.

Table 3.4 shows the relative error of the median of the duration using a sample composed of isolated events. The errors were quite high for the uniform and Gamma distributions, reaching a maximum value of 10.77% for the Gamma distribution. Higher intensities lead to a higher degree of overlapping and therefore greater errors. The minimum and the median of the errors are null for the case of the normal distribution. This is due to the small variance we have used in the simulation of the durations (a value of 1). Almost all isolated grains segmented from the images last a duration equal to the median.

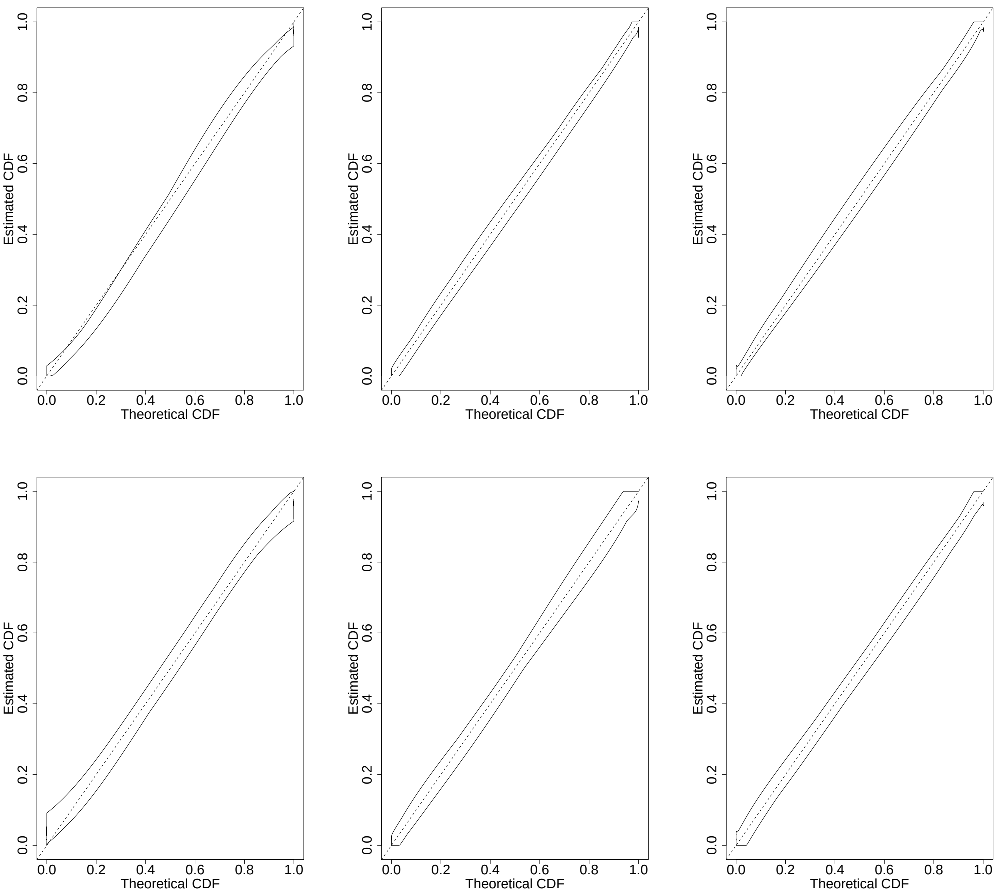


Figure 3.9: The minimum and maximum deviations of the estimated cumulative distribution function F_D from the theoretical function.

The spatial-temporal covariance was estimated using Eq. (3.21) in the interval $[0, 16]$ seconds with a step of one frame. To compute the first derivative of $H(t)$, we used 15 basis functions to transform the raw data into a piece-wise function, which is appropriate for functions re-sampled at 65 values. The R ([63, 65]) package *fda* developed by J.O. Ramsay has been used.

3.4.2 Application to noisy images

In some applications, the image acquisition method introduces random noise resulting in the appearance of corrupted pixels, edge blurring and other spurious effects. In

Table 3.3: Relative errors of the median of the duration (%).

		$\lambda = 0.0001$	$\lambda = 0.0002$	$\lambda = 0.0003$
	Param.	min/median/max	min/median/max	min/median/max
Uniform	$[4, 8]$	0.44/1.33/4.00	0.44/0.44/4.00	0.44/0.44/3.11
	$[8, 12]$	0.27/0.80/2.93	0.27/0.27/1.87	0.27/0.27/1.87
Gamma	$[18, 0.3333]$	0.38/0.52/3.24	0.38/0.52/3.10	0.38/0.52/2.19
	$[50, 0.2]$	0.14/0.67/2.28	0.14/0.67/2.28	0.14/0.54/1.21
Normal	$[6, 1]$	0.44/0.44/2.22	0.44/0.44/2.22	0.44/0.44/2.22
	$[10, 1]$	0.27/0.27/1.33	0.27/0.27/1.33	0.27/0.27/0.80

Table 3.4: Relative errors of the median of the duration using isolated clumps (%).

		$\lambda = 0.0001$	$\lambda = 0.0002$	$\lambda = 0.0003$
	Param.	min/median/max	min/median/max	min/median/max
Uniform	$[4, 8]$	0.00/0.00/4.17	0.00/4.17/8.33	0.00/4.17/8.33
	$[8, 12]$	0.00/1.25/5.00	0.00/2.50/5.00	0.00/2.50/7.50
Gamma	$[18, 0.3333]$	1.97/2.28/6.53	1.97/2.28/6.53	2.27/5.46/10.77
	$[50, 0.2]$	0.58/1.84/1.84	0.67/1.84/6.87	0.58/1.84/6.88
Normal	$[6, 1]$	0.00/0.00/4.17	0.00/0.00/4.17	0.00/3.12/4.17
	$[10, 1]$	0.00/0.00/2.5	0.00/0.00/2.5	0.00/1.87/5.00

this subsection, we analyze the performance of the estimator of the median of the distribution function of the duration in noisy images. We introduced salt-and-pepper noise in the frames independently. This type of noise corrupts a given percentage R of pixels in the image. Three ratios of noise were simulated, 0.5%, 1%, 2% (see Videos Noise1, Noise2 and Noise3 in supplementary material). Table 3.5 shows the relative errors in the estimation of the median of the duration. For each ratio of noise, the minimum, maximum and median values of the errors were calculated. The median of the relative errors was lesser than 2.2% for low and medium noise density. The maximum error was 4% for the highest ratio of noise. We can conclude that the proposed estimator performs well in the presence of this kind of noise.

Table 3.5: Relative errors in the estimation of the median of the duration in noisy images with a percentage of corrupted pixels R (%).

Distr.	λ	Without noise min/med/max	$R = 0.005$ min/med/max	$R = 0.01$ min/med/max	$R = 0.02$ min/med/max
$N(6,1)$	0.0001	0.44/0.44/2.22	0.44/1.33/2.22	0.44/1.78/3.11	2.22/3.11/4.00
	0.0002	0.44/0.44/2.22	0.44/0.89/2.22	0.44/1.78/3.11	2.22/3.11/4.00
	0.0003	0.44/0.44/2.22	0.44/1.33/2.22	1.33/2.22/3.11	2.22/3.56/4.00
$N(10,1)$	0.0001	0.27/0.27/1.33	0.27/0.53/1.33	0.80/1.33/1.87	1.87/2.67/3.47
	0.0002	0.27/0.27/1.33	0.27/0.80/1.33	0.27/1.33/1.87	1.87/2.93/3.47
	0.0003	0.27/0.27/0.80	0.27/0.80/1.33	0.80/1.60/2.40	2.40/2.93/4.00

3.5 Application to cell Biology

Endocytosis is the process by which cells traffic components from the plasma membrane into various intracellular compartments. This process regulates a wide range of physiological functions including cell migration, signaling, nutrient uptake, wound repair and neurotransmission. Clathrin-mediated endocytosis is one of the main endocytic routes.

However, the current knowledge regarding the formation of endocytic protein complexes is relatively limited and largely based on biochemical approaches and genetic analysis. Despite a growing number of studies of endocytosis in living cells (e.g. using TIRFM), a formal method to quantify the correlation between two endocytic proteins in space and time is still lacking.

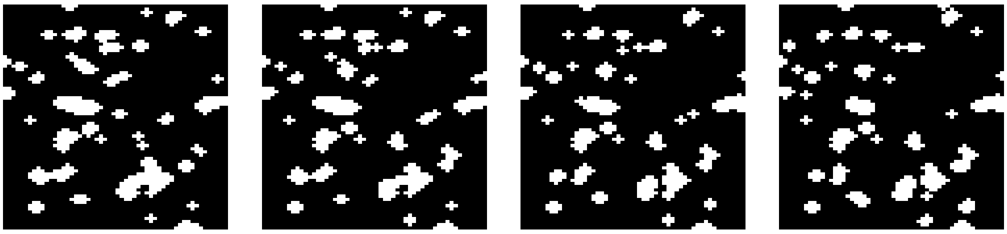


Figure 3.10: Several consecutive snapshots of a fluorescent-tagged protein image sequence. The binary shapes correspond to the segmented areas of fluorescence.

To obtain the duration distribution of the endocytic events is important in the analysis of the influence of different cell treatments on this process. Traditionally, collecting this kind of data has been very time-consuming where much of the work had to be done manually, making it virtually impossible in large image sequences.

Our biological data comprise three sequences of clathrin protein, described in Table 3.6 (see Videos Cell1, Cell2 and Cell3 in supplementary material). RFP stands for Red Fluorescent Protein. Time-lapse images were acquired by TIRFM at 0.25 frames per second. In this technique, a laser beam is sent to the sample at a critical angle. As it encounters the interface between two media with a different refractive index (i.e. water and glass), the beam undergoes total internal reflection. As a consequence, a small excitation wave is generated. The evanescent field is only $100 - 200$ nm thick, and it decays exponentially as it moves away from the coverslip. Therefore, only objects which are within $100 - 200$ nm of the bottom plasma membrane of the cell are illuminated, while the nucleus, inner cytosol and upper plasma membranes are left in the dark. In this way, it is possible to image membrane-associated events with superior signal-to-noise ratio. The setup employed for this study was an objective-based TIRFM (63X magnification) implemented on an inverted IX70 microscope (Olympus) and coupled to a 488-nm laser line (Melles Griot). The laser power was between 80 and 100 mW. The image processing to extract the binary images was based on the application of: a top-hat transform to subtract the background and extract peaks of fluorescence, a template matching with a gaussian kernel to remove eventual noise, followed by a region growing technique in order to delineate each marked object.

Table 3.6: Description of biological image sequences.

Seq.	Protein	# of frames	Size	Video
1	Clathrin-RFP	300	120×67	Cell 1
2	Clathrin-RFP	300	77×60	Cell 2
3	Clathrin-RFP	300	71×54	Cell 3

We have performed the analysis for the three clathrin sequences. The medians of the duration were 84, 74 and 86 seconds, respectively. The cumulative distribution functions of the duration for these three sequences are shown in Fig. 3.11. The estimated median using the sample composed of isolated episodes was 16 seconds for the

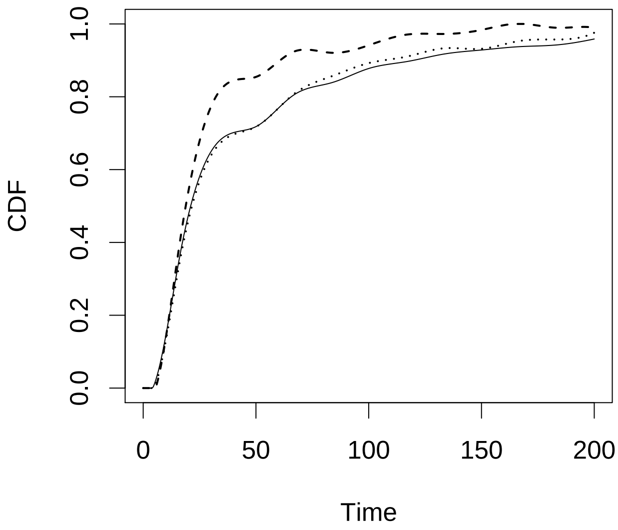


Figure 3.11: Estimated cumulative distribution function of the duration of endocytic events.

three cases. Again, note how the use of a biased sample results in the underestimation of the median of the duration.

3.6 Conclusions

We have shown that the use of a biased sample composed of isolated events can lead to a very high bias in the estimation of the mean of the random duration. The higher the overlapping among grains is the higher the bias is. This can be unacceptable in many real applications. It is necessary to develop procedures based on stochastic models where the spatial temporal overlapping is explicitly assumed. We have shown that the temporal germ-grain models are a good choice and, in particular, the cylindrical temporal Boolean model.

We have proposed an estimator of the cumulative distribution function of the duration of spatially and temporally overlapping events based on the covariance of temporal random closed sets. Results from the simulation study showed that the absolute errors in the estimation of the distribution function of the duration are small. Indeed, the relative error in the estimation of the median of the distribution was lesser than 1% in almost all cases. The estimator also performs quite well in images with a certain degree of random noise such as salt-and-pepper.

Several points remain open for future research. First, the development of new temporal germ-grain models with other germ models. In particular, models with higher aggregation than the Poisson point process which will lead to greater overlapping probability. The estimation will be harder but, possibly, semi-parametric models could be proposed. Importantly, the experimenters should provide information about the mechanisms to be modelled in such a way that this information can be incorporated in the definition of these new semi-parametric models. Second, the estimation of the size distribution for the grains. Last, the method described depends on the assumption of stationarity. Potential extensions will include the use of other models, such as the Cox model.

We have implemented a Matlab toolbox for simulation of temporal Boolean models and estimation. It is available at *http://www.uv.es/tracs/*. Simulated videos and fluorescent-tagged proteins image sequences can also be downloaded, as well as the *R* scripts of the functional data analysis applied in this study. These videos and the source code of the functions are included in the supplementary material.

Chapter 4

The bivariate case: bivariate temporal random sets to model spatial-temporal dependencies

The study of many dynamic processes implies the analysis of inter-dependencies between different types of events produced by coupling mechanisms. For instance, in cell biology, the identification and analysis of protein interactions is essential to understand a given physiological process. In some applications the phenomenon to be modeled can be described by a configuration of independent and randomly located particles with different sizes forming random clumps. Examples can be found in Microscopy, Material Sciences, Biology, Medicine, Chemistry, Geostatistics, Cellular Communications Networks or Image Processing (in particular for texture analysis and classification [33]). Random sets have been often applied to model these irregular patterns [52, 57, 79, 6].

Although the problem of measuring interaction between types of events in bivariate spatial point processes has been reported and analyzed [17, 24, 80, 87], this is not the case for random sets. In point processes, the standard summary methods use bivariate counterparts of the empty space function, the nearest neighbor distance distribution, the Ripley $\mathbb{K}$-function and derived statistics such as the pair correlation function and the J-function. The J-function was extended in [35] to a bivariate spatial process, one being a point process and the other being a random set. However, further extensions which capture dependencies associated with time and spatial properties of overlapping events are needed.

4.1 Introduction

In this chapter, we define the bivariate temporal random set and use it to model pairs of time lapse image sequences. We propose a non-parametric approach to quantify spatial and spatial-temporal interrelations between different types (or categories) of events using the pair-correlation, cross-covariance and the Ripley $\mathbb{K}$ functions. Based on these summary statistics we propose a randomization procedure to test independence between event types by applying random toroidal shifts and Monte Carlo tests. A testing procedure based on a Monte Carlo test and on bootstrap procedures is proposed.

A simulation study assessed the performance of the proposed estimators and showed that these statistics capture the spatial-temporal dependencies accurately. The estimation of the spatial-temporal interval of interactions was also obtained.

The method has a direct application in the study of many dynamic biological processes, which require analysis of the spatial and temporal inter-dependencies between two or more different types of proteins. Being able to analyze cellular protein complexes dynamics under physiological conditions will provide fundamental insights of their biological functions, in many processes such as propagation of signaling cascades and internalization of membrane (also known as endocytosis). Essential to understanding such fundamental processes is to be able to ascertain precisely where and when a given protein is localized in space in time, especially vis-à-vis to secondary and even tertiary proteins (e.g. dynamic colocalization analysis).

Current practice relies on simple visual analysis (e.g. of scatterplots or overlapping RGB merged images) and simple measurements, such as colocalization coefficients. Some commercial software analysis packages (e.g. from Zeiss and Bitplane) can measure colocalization between two fluorescence images, display the colocalization as a two dimensional xy scatterplot and provide some standard quantitative measures such as colocalization coefficients. However, these tools show major drawbacks.

First, they are unable to simultaneously assess spatial-temporal dependencies from image sequences. Second, more general measures of dependencies are needed, i.e. two fluorescent molecules sharing two regions close in space and close in time, but not necessarily sharing the same pixel at a given frame. Third, the current procedures are based on a pixel-by-pixel analysis with no probabilistic modeling. No formal statistical tests to evaluate the colocalization have been applied. Also, no p-values of significance have been given.

Therefore, formal tests and associated statistical methods are needed to mine these complex data. A major aim of the present study is to provide cell biologists with a computational statistical procedure so as to quantify the spatial and temporal interactions of subcellular proteins in living cells.

The method is successfully applied to analyze the interdependencies of several endocytic proteins using image sequences of living cells and validate the procedure as a new way to automatically quantify dependencies between proteins in a formal and robust manner [20].

Images were acquired using Total Internal Reflection Fluorescence Microscopy (TIRFM), a technique that provides extremely high contrast images of events that occur near the cell surface [83]. This technique has enabled simultaneous and dynamic visualizations of multiple proteins in living cells. Fig. 4.1 shows a reconstruction of the area covered by two proteins over 50 frames within a small sub-window of the cell. The green color corresponds to Epsin-GFP and the red one to Clathrin-RFP.

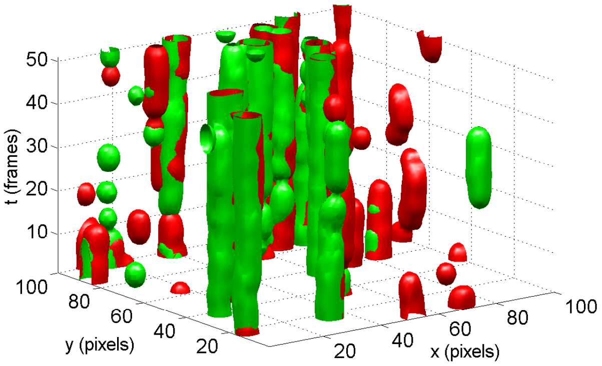


Figure 4.1: A spatial-temporal plot of the fluorescence areas covered by two fluorescently-tagged proteins.

In Section 4.2.1, we define the concept of bivariate temporal random set. A generalization of the Ripley $\mathbb{K}$-function, of the covariance and of the pair correlation functions for bivariate random sets are introduced in Section 4.2.2. Section 4.4 contains the testing procedure based on toroidal shift randomizations and on a Monte Carlo test. A simulation study to assess the performance of the proposed estimators is described in Section 4.5. Section 4.6 is devoted to the biological application. Conclusions are given in Section 4.7.

4.2 Bivariate Temporal Random set

4.2.1 Notation and basic definitions

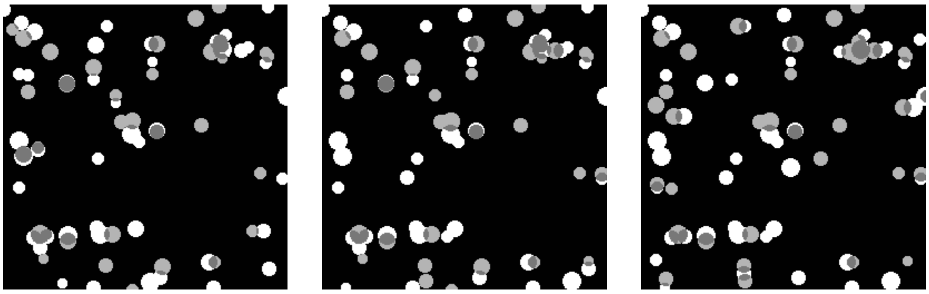
Intuitively, a random closed set Φ is a random mechanism producing random (closed) subsets in space. More formally, let us denote by $\mathbb{F}$ the subsets of $\mathbb{R}^d$ and $\mathbb{F}_K = \{F \in \mathbb{F} : F \cap K \neq \emptyset\}$, i.e. the family of sets hitting the set K, with K being an arbitrary compact subset of $\mathbb{R}^d$. Let σ_f be the σ-algebra generated by $\mathbb{F}_K$. Then a *random set* is the probability space $(\mathbb{F}, \sigma_f, P)$, where P is a probability measure in the measurable space $(\mathbb{F}, \sigma_f)$ [52, 79]. Note that the definition is given in such a way that $\forall \Phi \in \mathbb{F}$, $\{\Phi \cap K \neq \emptyset\}$ is a random event and its probability $P(\Phi \cap K \neq \emptyset)$ can be computed. From now on, we will denote $\Phi \uparrow K = \{\Phi \cap K \neq \emptyset\}$.

We model each binary image of the sequence as a realization of a random set in $\mathbb{R}^2$ and the whole image sequence as a random set in $\mathbb{R}^2 \times \mathbb{R}_+$. Let us denote by $\Phi(t)$ the random set corresponding to the random binary image at the observation time t, with $t \geq 0$. If we stack the observed temporal cross-sections for every time t, we define the *temporal random set* in $\mathbb{R}^2 \times \mathbb{R}_+$ as

$$\Phi = \cup_{t \geq 0} \Phi(t). \tag{4.1}$$

This model can be considered as a particular case of a non-isotropic three dimensional random set.

If Φ_1 and Φ_2 are the temporal random sets associated with the first and second event types respectively, then the random set defined as $\boldsymbol{\Phi} = (\Phi_1, \Phi_2)$ is a *bivariate temporal random set*, where Φ_i is the i-th component. The distribution of $\boldsymbol{\Phi} = (\Phi_1, \Phi_2)$ is characterized by the probability $P(\Phi_1 \uparrow K_1, \Phi_2 \uparrow K_2)$, i.e. the probability that the temporal random set Φ_1 hits K_1 and the temporal random set Φ_2 hits K_2 simultaneously, being K_1 and K_2 compact subsets of $\mathbb{R}^2 \times \mathbb{R}_+$. Fig. 4.2 shows three



(a) (b) (c)

Figure 4.2: A realization of a simulated bivariate temporal random set.

consecutive temporal cross-sections of a realization of a bivariate temporal random set. The white color corresponds to events of type 1, the light grey to events of type 2, and the dark grey to overlapping regions. Video 2 in the supplementary material corresponds to a simulated bivariated temporal random set in which spatial-temporal dependencies were generated.

Fig. 4.3 shows the spatial-temporal reconstruction in $3D$ of this realization of a bivariate temporal random set. We can observe the spatial-temporal overlapping between both proteins.

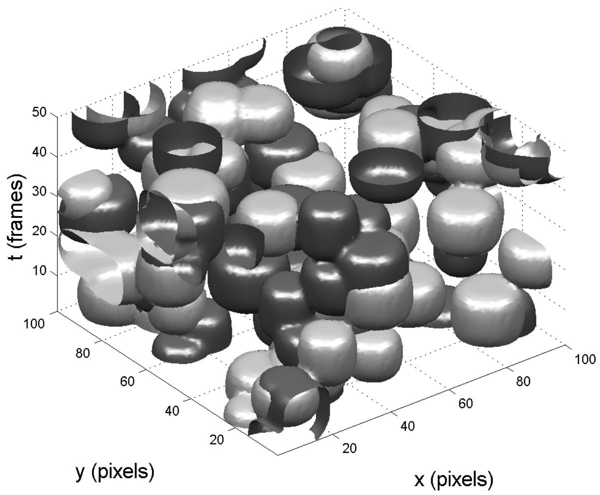


Figure 4.3: A spatial-temporal reconstruction.

Our data consist of a pair of sequences of binary images which have been simultaneously captured, i.e. two discrete sets of equally spaced temporal cross-sections observed through a sampling time interval T and within a fixed sampling window W (a closed and bounded subset of the two dimensional Euclidean space). Let ϕ denote the corresponding realization of Φ and $\phi(t_i)$, the area covered at time t_i . If $t_1, t_2, \ldots, t_n$ are the sampling times, then we will observe $\phi(t_i) \cap W$, where $i = 1, \ldots, n$.

Throughout this study we assume that (Φ_1, Φ_2) is jointly stationary, i.e. its joint distribution is invariant against spatial-temporal translations in $\mathbb{R}^2 \times \mathbb{R}_+$. This implies that Φ_1 and Φ_2 are marginally stationary. This assumption is a natural and pragmatic simplification, which is tenable in many real applications and justifies the use of relatively simple non-parametric summaries. All the summary descriptors of the image sequence we propose rely on this assumption. There is a considerable debate on under which conditions the stationarity assumption can be made (see the introduction in [67]). In spatial processes, departures from stationarity are trends in intensity from top to bottom, trends in any other direction, 'banding' from periodic variation in intensity in any direction or 'patchiness' on one of many scales at least. If a large observation scale is used then a small region of the phenomenon is observed. In this case the stationarity hypothesis could be assumed. When the scale becomes smaller then the spatial heterogeneity is even more evident, which implies that this hypothesis could not be tenable. For a given image data set, the experimenter should (visually) evaluate the images and decide whether the stationarity assumption can be made. If this hypothesis was not tenable then this analysis may not be sufficient. In Section 4.7 we discuss some conditions in which the assumption of stationarity could not be tenable and we comment on different attempts to overcome this problem.

Under stationarity the volume fractions can be considered separately for each component $p_i = P(\mathbf{0} \in \Phi_i(0))$, for $i = 1, 2$, where the spatial origin $\mathbf{0}$ denotes an arbitrary location and the temporal origin 0 denotes an arbitrary time. Note that p_i gives the probability that an arbitrary point at an arbitrary time belongs to the i-th component of the random set and can be interpreted as the mean volume covered by the i-th temporal random set per unit area and unit time.

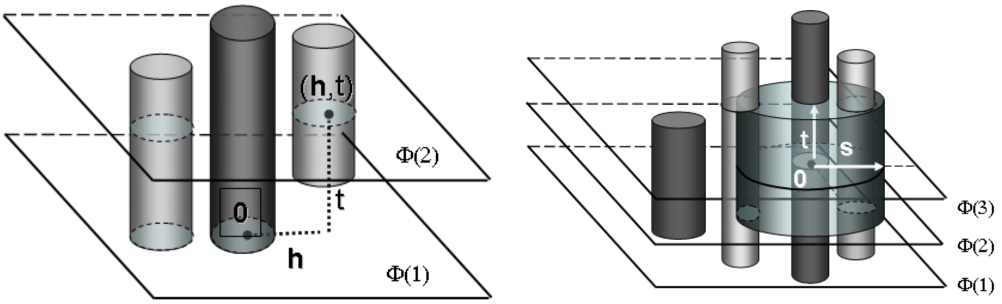
4.2.2 Summary statistics for bivariate temporal random sets

We define a generalized version of the Ripley $\mathbb{K}$-function, the covariance and the pair correlation functions for a bivariate temporal random set, $\mathbf{\Phi} = (\Phi_1, \Phi_2)$. For a given

$\mathbf{h} \in \mathbb{R}^2$ and a given time $t \in \mathbb{R}_+$, we define the *spatial-temporal covariance function* as

$$\mathbb{C}_{ij}(\mathbf{h}, t) = P(\mathbf{0} \in \Phi_i(0), \mathbf{h} \in \Phi_j(t)), \tag{4.2}$$

where $i, j = 1, 2$. This is a natural and direct generalization of the covariance of a random set (details of various theoretical aspects can be found in [57, 72, 79]). The spatial-temporal covariance funtion gives us the probability that an arbitrary point observed at an arbitrary time belongs to the i-th component of the temporal random set and its translation by a vector $\mathbf{h}$ and a time t belongs to the j-th component. This idea is illustrated in Fig. 4.4 (a). The functions $\mathbb{C}_{11}$ and $\mathbb{C}_{22}$ are the covariance functions of the temporal random sets Φ_1 and Φ_2 respectively, and the functions $\mathbb{C}_{12}$ and $\mathbb{C}_{21}$, the cross-covariance functions.



(a) (b)

Figure 4.4: Graphical description of the spatial-temporal cross-covariance (a) and cross $\mathbb{K}$-function (b) for a bivariate temporal random set. The dark grey cylinders represent the i-th temporal random set and the light grey cylinders the j-th temporal random set.

Throughout this work we assume that (Φ_1, Φ_2) are jointly spatial (but not temporal) isotropic, i.e. their joint distribution is invariant against spatial rotations. Under spatial stationarity and isotropy, the covariance function $\mathbb{C}_{ij}$ only depends on the modulus of $\mathbf{h}$, $s = \parallel \mathbf{h} \parallel$, and on t.

The *cross* $\mathbb{K}$-*function* of a bivariate temporal random set can be defined as

$p_j \mathbb{K}_{ij}(s, t) = E[$ *volume covered by the* j*-th component within a cylinder of radius* s *and height* $2t$ *centered at a randomly chosen point of the* i*-th component*$]$,

where p_j is the volume fraction of Φ_j.

More formally, it can be expressed as $p_j \mathbb{K}_{ij}(s,t) = E[\nu_3(\Phi_j \cap B(\mathbf{0},s) \times [-t,t])|\mathbf{0} \in \Phi_i(0)]$, *with* $s,t \geq 0$, where ν_3 stands for the volume and $B(\mathbf{0},s)$ is the disk centered at an arbitrary location of Φ_i with radius s. Fig. 4.4 (b) illustrates this concept. From the covariance function, the cross $\mathbb{K}$-function can be computed as

$$\mathbb{K}_{ij}(s,t) = \int_{-t}^{t} \int_{B(\mathbf{0},s)} \frac{\mathbb{C}_{ij}(\mathbf{u},v)}{p_i p_j} d\mathbf{u} dv, \tag{4.3}$$

with $i,j = 1,2$. From (4.2) and (4.3), it holds that $\mathbb{K}_{ij}(s,t) = \mathbb{K}_{ji}(s,t)$.

Finally, we define the *pair correlation function* of the bivariate temporal random set as

$$g_{ij}(s,t) = \frac{\mathbb{C}_{ij}(s,t)}{p_i p_j}. \tag{4.4}$$

4.3 Estimators

At this point, we propose the estimators for p_i, $\mathbb{C}_{ij}$ and $\mathbb{K}_{ij}$. Essentially, we replace the (unknown) continuous versions of these functions with the corresponding (known) discrete versions. In order to simplify the notation, we assume that the sampling times are $t = 0, \delta, 2\delta, \ldots, n\delta$; i.e. $n+1$ equally spaced frames, where δ is the temporal delay between two consecutive frames. Our sampling information consists of a sequence of sets $\{(\phi_1(k\delta) \cap W, \phi_2(k\delta) \cap W)\}_{k=0,\ldots,n}$.

A natural estimator for p_i is the mean of the area fractions observed at each frame of the sequence

$$\hat{p}_i = \frac{1}{n+1} \sum_{k=0}^{n} \frac{\nu_2[\phi_i(k\delta) \cap W]}{\nu_2[W]}. \tag{4.5}$$

We will estimate the covariance as

$$\hat{\mathbb{C}}_{ij}(\mathbf{h}, v\delta) = \frac{1}{n-v+1} \sum_{k=0}^{n-v} \frac{\nu_2\Big[\big(\phi_i(k\delta) \cap W\big) \cap \big(\phi_j((k+v)\delta) \cap W\big) - \mathbf{h}\Big]}{\nu_2\big[W \cap W - \mathbf{h}\big]}. \tag{4.6}$$

We have adopted the minus sampling method [79] to correct edge effects. If the covariance function does not depend on time, we can estimate it as

$$\hat{\mathbb{C}}_{ij}(\mathbf{h}, 0) = \frac{1}{n+1} \sum_{k=0}^{n} \hat{\mathbb{C}}_{ij}(\mathbf{h}, k\delta). \tag{4.7}$$

Likewise, if there is no spatial dependence, the following estimator can be used

$$\hat{\mathbb{C}}_{ij}(\mathbf{0}, k\delta) = \frac{1}{\nu_2(W)} \int_W \hat{\mathbb{C}}_{ij}(\mathbf{h}, k\delta) d\mathbf{h}. \tag{4.8}$$

Finally, $\mathbb{K}_{ij}(s,t)$ and $g_{ij}(s,t)$ will be estimated by replacing the volume fraction and the covariance function in (4.3) and (4.4) with their corresponding estimators given in (4.5) and (4.6).

4.4 Testing spatial and spatial-temporal dependencies

The functions $\mathbb{C}_{ij}$, g_{ij} and $\mathbb{K}_{ij}$ will be used to describe the joint distribution and to test different null hypotheses. First, we test whether there are spatial dependencies. For instance, spatial dependencies would correspond to a pattern in which both events are associated with an environmental factor in the neighbourhood that makes them happen at similar locations but at any time. Second, if locations and times are jointly analyzed, we will test spatial-temporal interactions. In this case, we would observe that type 1 events occur in the surroundings of type 2 events and both event types are close in time. Spatial-temporal associations are considered to be a consequence of environmental factors.

4.4.1 Independent temporal random sets

First, we give the formal definitions of complete, spatial or temporal independence. Two (temporal) random sets would be completely independent if

$$P(\Phi_1 \uparrow K_1, \Phi_2 \uparrow K_2) = P(\Phi_1 \uparrow K_1) P(\Phi_2 \uparrow K_2),$$

for any compact subsets K_1 and K_2 of $\mathbb{R}^2 \times \mathbb{R}_+$. This is a direct extension of the original definition of independent random sets given in [52].

If Φ_1 and Φ_2 are independent random sets then, for fixed times t_1 and t_2, the random sets $\Phi_1(t_1)$ and $\Phi_2(t_2)$ are independent random sets. Let K_1 and K_2 be two compact subsets of $\mathbb{R}^2$ then

$$\begin{aligned}P(\Phi_1(t_1) \uparrow K_1, \Phi_2(t_2) \uparrow K_2) = \\ P(\Phi_1 \uparrow K_1 \times \{t_1\}, \Phi_2 \uparrow K_2 \times \{t_2\}) = \\ P(\Phi_1 \uparrow K_1 \times \{t_1\})P(\Phi_2 \uparrow K_2 \times \{t_2\}) = \\ P(\Phi_1(t_1) \uparrow K_1)P(\Phi_2(t_2) \uparrow K_2).\end{aligned}$$

The opposite is not true, i.e. if for any times t_1 and t_2 the corresponding random sets $\Phi_1(t_1)$ and $\Phi_2(t_2)$ are independent then it does not follow that the corresponding three dimensional random sets Φ_1 and Φ_2 are independent.

In many real processes, it seems reasonable to consider less restrictive types of independence. From now on, for A, a compact subset of $\mathbb{R}^2$, and B, a compact subset of $\mathbb{R}_+$, we consider the generalized cylinder $A \times B$. We say that Φ_1 and Φ_2 are spatially independent if

$$P(\Phi_1 \uparrow A_1 \times B, \Phi_2 \uparrow A_2 \times B) = P(\Phi_1 \uparrow A_1 \times B)P(\Phi_2 \uparrow A_2 \times B),$$

for any A_1 and A_2 compact subsets of $\mathbb{R}^2$ and any B compact subset of $\mathbb{R}_+$. Likewise, we say that Φ_1 and Φ_2 are temporally independent if

$$P(\Phi_1 \uparrow A \times B_1, \Phi_2 \uparrow A \times B_2) = P(\Phi_1 \uparrow A \times B_1)P(\Phi_2 \uparrow A \times B_2).$$

Under complete independence, the covariance function is

$$\mathbb{C}_{ij}(\mathbf{h}, t) = P(\mathbf{0} \in \Phi_i(0), \mathbf{h} \in \Phi_j(t)) = P(\mathbf{0} \in \Phi_i(0))P(\mathbf{h} \in \Phi_j(t)) = p_i p_j. \tag{4.9}$$

In such a case, from (4.4) it holds that $g_{ij}(s, t) = 1$. Values of $g_{ij}(s, t)$ greater than one represent a positive dependence (attraction) and lesser than one represent a negative dependence (inhibition). Likewise, from (4.3) we have $\mathbb{K}_{ij}(s, t) = 2\pi s^2 t$.

For temporally independent random sets, the covariance function does not depend on time and the cross $\mathbb{K}$-function can be computed as

$$\mathbb{K}_{ij}(s, t) = 4\pi t \int_0^s \frac{r\mathbb{C}_{ij}(r, 0)}{p_i p_j} dr. \tag{4.10}$$

If there is temporal dependence but no spatial association, the covariance function only depends on time and the cross $\mathbb{K}$-function is calculated as

$$\mathbb{K}_{ij}(s,t) = \pi s^2 \int_{-t}^{t} \frac{\mathbb{C}_{ij}(\mathbf{0},v)}{p_i p_j} dv. \tag{4.11}$$

Let us consider the following marginal functions:

$$\mathbb{K}_{ij}^{(1)}(s) = \int_{B(\mathbf{0},s)} \frac{\mathbb{C}_{ij}(\mathbf{u},0)}{p_i p_j} d\mathbf{u}, \tag{4.12}$$

and

$$\mathbb{K}_{ij}^{(2)}(t) = \int_{-t}^{t} \frac{\mathbb{C}_{ij}(\mathbf{0},v)}{p_i p_j} dv, \tag{4.13}$$

then, under spatial independence $\mathbb{K}_{ij}^{(1)}(s) = \pi s^2$ and under temporal independence $\mathbb{K}_{ij}^{(2)}(t) = 2t$.

4.4.2 Testing null hypotheses by using toroidal shift randomizations

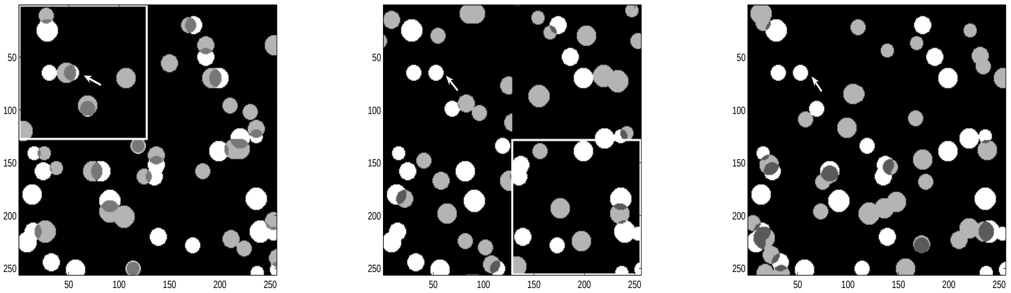
Testing null hypotheses can be achieved by means of toroidal shift methods, which provide a non-parametric way to test complete independence. For convenience, we assume that the sampling window W is a rectangle. For a given $\mathbf{h} \in W$, we denote by $T_{\mathbf{h}}(A)$ the toroidal shift (with respect to W) of any subset A of W. A toroidal shift is a simultaneous, parallel shift of all points in the set by the same randomly chosen shift vector $\mathbf{h}$. The rectangle W is treated as a torus and the set A is shifted within this torus. This method can be extended to the temporal dimension as follows. Let us consider

$$T_{(\mathbf{h},d\delta)}(\phi_i(k\delta) \cap W) = \\ \begin{cases} T_h(\phi_i((k+d)\delta) & \text{if } d = 0,\dots,n-k, \\ T_h(\phi_i((k+d-n)\delta) & \text{if } d = n-k+1,\dots,n, \end{cases} \tag{4.14}$$

where $d = 0,\dots,n-1$, being n the total number of frames. Let us consider the random vector $(\mathbf{H}, D)$ such that $\mathbf{H}$ (distance) and D (lag) are independent; $\mathbf{H}$ is uniform in

W, $\mathbf{H} \sim Unif(W)$; and D is uniform in $\{0, \ldots, n-1\}$. From the original observed image sequence $\{(\phi_1(k\delta) \cap W, \phi_2(k\delta) \cap W)\}_{k=1,\ldots,n}$, we generate $\mathbf{H} = \mathbf{h}$ and $D = d$ obtaining a randomly modified sequence $\{(\phi_1(k\delta) \cap W, T_{(\mathbf{h},d\delta)}(\phi_2(k\delta) \cap W)\}_{k=1,\ldots,n}$ in which the second component has been spatially and temporally shifted over the first one.

In Fig. 4.5 an example of toroidal shift randomization is shown. Images are 256×256 pixels in size. Fig. 4.5(a) displays a frame of a bivariate temporal random set. Fig. 4.5(b) displays a randomization with $\mathbf{h} = (128, 128)$ and $d = 0$. Fig. 4.5 (c) displays a randomization with $\mathbf{h} = (0, 0)$ and $d = 20$. We can observe how the toroidal shift randomizations breaks the dependence between components.



(a) (b) (c)

Figure 4.5: An example of toroidal shift randomization.

Let us see a short explanation of the testing procedure using the cross $\mathbb{K}_{ij}$-function. First, the cross $\mathbb{K}$-function is calculated for the original pair of binary image sequences, giving us $\hat{\mathbb{K}}_{ij,0}$. Next, for m randomized pairs in which we keep the original sequence of type 1 events and randomize type 2 events, we obtain $\hat{\mathbb{K}}_{ij,n}$ with $n = 1, \ldots, m$. A graphical analysis to test independence can be performed by displaying $\hat{\mathbb{K}}_{ij,0}$ along with the lower and upper envelopes obtained from the m functions associated with the randomizations. The region delimited by both envelopes quantifies the variability when independence between components is assumed. The function $\hat{\mathbb{K}}_{ij,0}$ should be contained within both envelopes under the null hypothesis of independence between event types. A similar approach is used to test spatial independence. In this case, the random variable D is chosen to degenerate at 0.

4.4.3 Monte Carlo test

Apart from the graphical test given in Section 4.4.2, a p-value corresponding to a Monte Carlo test can be computed. Independence would imply that the sampling distribution of $\hat{\mathbb{K}}_{ij}$ is invariant to random toroidal shifts. We can conduct a Monte Carlo test of independence by computing a suitable test statistic and comparing the value calculated for $\hat{\mathbb{K}}_{ij,0}$ with the values obtained from the randomizations $\hat{\mathbb{K}}_{ij,n}$, with $n = 1, \ldots, m$. This methodology was proposed in [24] and applied there to point processes. Under the hypothesis of independence

$$P(\mathbb{K}_{ij,0}(s,t) > U(s,t)) = P(\mathbb{K}_{ij,0}(s,t) < L(s,t)) = \frac{1}{m+1},$$

where the lower envelope is defined as $L(s,t) = \min_{n=1,\ldots,m} \hat{\mathbb{K}}_{ij,n}(s,t)$ and the upper envelope is $U(s,t) = \max_{n=1,\ldots,m} \hat{\mathbb{K}}_{ij,n}(s,t)$.

An exact test can be performed using the test statistic $d_n = \int_0^{+\infty} (\mathbb{K}_{ij,n}(s,t) - \bar{\mathbb{K}}_{ij,n}(s,t))^2 dsdt$, where $\bar{\mathbb{K}}_{ij,n}(s,t) = \sum_{r=0,r\neq n}^{m} \frac{\mathbb{K}_{ij,r}(s,t)}{m}$. All the rankings of d_0 are equiprobable under the null hypothesis. If $d_{(j)}$ denotes the j-th largest among d_n, with $n = 0, \ldots, m$, then, $P(d_0 = d_{(j)}) = \frac{1}{m+1}$ with $j = 0, \ldots, m$, and rejection of the null hypothesis on the basis that d_0 ranks the k-th largest or higher gives an exact one-sided test with p-value equal to $1 - \frac{k}{m+1}$.

4.4.4 Estimating the spatial-temporal interval of dependencies

We describe now the proposed bootstrap algorithm to estimate the spatial-temporal interval of interactions.

Let us consider the random events $\{\mathbf{0} \in \Phi_1(0)\}$ and $\{\mathbf{h} \in \Phi_2(t)\}$ and let us denote $s = \parallel \mathbf{h} \parallel$. Supposing that we have rejected the null hypothesis of complete independence, then from (4.9) it holds that there is a pair of values (s,t) such that $\mathbb{C}_{ij}(s,t) - p_i p_j \neq 0$ (or the corresponding expression obtained by replacing the cross-covariance with the cross $\mathbb{K}$-function or the pair correlation function) and therefore $\{\mathbf{0} \in \Phi_1(0)\}$ and $\{\mathbf{h} \in \Phi_2(t)\}$ are not independent. Note that it does not imply that this statement is true for any pair (s,t). In many real applications there are only local dependencies in such a way that for large s and t these two events $\{\mathbf{0} \in \Phi_1(0)\}$

and $\{\mathbf{h} \in \Phi_2(t)\}$ will be independent. For a given s it is expected that the events will be independent for t values larger than a fixed point that depends on s.

Given (s_0, t_0), let (h_n, t_n) (with $n = 1, \ldots, m$) be a random sample uniformly distributed in $\big[B(0, s_0 + \delta_1) \setminus B(0, s_0 - \delta_1)\big] \times [t_0 - \delta_2, t_0 + \delta_2]$ where $\delta_1, \delta_2 > 0$. For every translation vector (h_n, t_n), we perform a toroidal shift of ϕ_1 over ϕ_2 and estimate the covariance functions from these new pair of sequences, $\hat{\mathbb{C}}_{ij,n}(h_n, t_n)$. Then, a two-sided Monte Carlo test applied to the difference $\hat{\mathbb{C}}_{ij,n}(h_n, t_n) - \hat{p}_i\hat{p}_j$ provides us with a p-value. By applying this procedure to different values of (s_0, t_0) we have a function $p(s_0, t_0)$, a map of p-values. This procedure serves to study the spatial-temporal interval of interactions with greater significance at low computational cost, since we can perform a large number of toroidal randomizations for each (h_n, t_n). Note that the covariance function is not calculated for the rectangle $[0, s] \times [0, t]$ as in (4.6), but for a reduced set of values (s_0, t_0). This map provides us with a graphical representation of the p-values and tell us at which distances and temporal lags the two components show significant dependencies.

4.5 A simulation study

Before applying the proposed estimators to examine dependencies in real processes we designed a simulation study. Models with three different types of dependencies between components were simulated: spatial interactions, spatial-temporal interactions and complete independence. In particular, we simulated bivariate temporal Boolean models. A temporal Boolean model formalizes the configuration of independent randomly placed particles with random durations. A linked-pairs scheme for the germs [24] was used to generate spatial and spatial-temporal dependencies between the two types of events. The marginal random sets were temporal Boolean models that have identical intensity and share the property that event types came in pairs. The spatial and temporal shifts between germ types used followed a uniform distribution with the parameters given in Table 4.1. Spatial shifts are expressed in pixels and temporal lags are given in frames. A proportion of 0.1 of germs of one component were not associated with any event of the second one. The simulated image sequences were 256 × 256 pixels in size, 100 seconds long and the sampling ratio was 1 frame per

Table 4.1: Shifts applied to germs of type 1 to generate germs of type 2.

Pair	Type of interaction	Spatial Shifts	Temporal Lags
1	Spatial	$U(0,6)$	
2		$U(0,12)$	
3	Spatial-temporal	$U(0,6)$	$U(-2,2)$
4		$U(0,12)$	$U(-6,6)$
5	None		

second. Grains were cylinders with uniform radius in the interval $[4,8]$ pixels and durations uniformly distributed in $[4,8]$ seconds. The spatial-temporal intensity of the germs process was 0.0001. Durations and grain sizes between components were independent. Video 1 in the supplementary material corresponds to spatial dependencies, Video 2 to spatial-temporal interactions and Video 3 to complete independence [20].

Apart from the function $\mathbb{K}_{ij}$ and the pair correlation function g_{ij}, the function $\mathbb{L}_{ij}$ has been used. This function stabilizes the variance [79, 17]. This, in turn, linearizes the plot under independence between components. The function $\mathbb{L}_{ij}^{(1)}$ is defined as

$$\mathbb{L}_{ij}^{(1)}(s) = \sqrt{\frac{\mathbb{K}_{ij}^{(1)}(s)}{\pi}}. \tag{4.15}$$

To study the spatial-temporal dependence, we have used

$$\mathbb{L}_{ij}(s,t) = \sqrt{\frac{\mathbb{K}_{ij}(s)t}{2\pi}}. \tag{4.16}$$

In this way, under spatial independence we have $\mathbb{L}_{ij}^{(1)}(s) - s = 0$, and under spatial-temporal independence, it holds that $\mathbb{L}_{ij}(s,t) - st = 0$. Even though these three descriptors provide similar information, the combination of them leads us to obtain stronger results. The pair correlation function g_{ij} is easy to interpret and can be used to detect dependencies between components at given values (s,t). The function $\mathbb{K}_{ij}$ and the function $\mathbb{L}_{ij}$ are more robust since they are obtained as integrated versions of g_{ij} within a spatial-temporal window. These are the main reasons for their wide use in other fields, such as the statistical analysis of spatial point patterns [24].

For all the models in Table 4.1, the cross-covariance was estimated by using (4.6)

Table 4.2: Monte Carlo p-values using $\mathbb{K}_{12}$-, g_{12}- and $\mathbb{L}_{12}$-functions.

Pair	$\mathbb{K}_{12}(s)$	$\mathbb{L}_{12}(s)$	$g_{12}(s)$	$\mathbb{K}_{12}(s,t)$	$\mathbb{L}_{12}(s,t)$	$g_{12}(s,t)$	Interactions
1	0.05	0.02	0.00	0.72	0.65	0.25	Spatial
2	0.00	0.00	0.00	0.32	0.41	0.30	Spatial
3	0.00	0.00	0.00	0.00	0.00	0.00	Spatial-temporal
4	0.00	0.00	0.00	0.00	0.00	0.00	Spatial-temporal
5	0.57	0.88	0.27	0.72	0.25	0.65	None

at four different angles $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, at distances ranging from 0 to 30 (with a step of 2 pixels) and at times from -28 to 28 (with a step of 4 frames). Functions $\mathbb{K}_{ij}$, g_{ij} and $\mathbb{L}_{ij}$ were calculated from the cross-covariance. A value of $m = 39$ toroidal randomizations were generated in the Monte Carlo test.

Fig. 4.6 plots the empirical function $\mathbb{K}^{(1)}_{12,0}$ (red line) and the upper and lower envelopes (blue lines) of the toroidal shift randomizations for simulation 1. The green line corresponds to the theoretical value under spatial independence of components. Figs. 4.7 (a) and (b) plots the pair correlation function $g_{12,0}(s)$ and the $\mathbb{L}^{(1)}_{12,0}(s)$ function with the upper and lower envelopes of the toroidal shift randomizations for simulation 1. In this model, a positive spatial association between germs was simulated (see Video 1, supplementary material).

The observed functions clearly deviated from the envelopes, suggesting the existence of spatial dependencies. A clear excess at small distances was observed, which is a feature of positive interactions. The empirical functions lie outside the envelopes within the interval of around $[0, 12]$ pixels. Note that in this simulation, the mean radius of the cylinders was 6 pixels and the maximum spatial shift generated was 6 pixels. Therefore, no dependence can be expected at distances over 12 pixels. The first row in Table 4.2 (columns headed $\mathbb{K}_{12}(s)$, $\mathbb{L}_{12}(s)$ and $g_{12}(s)$) gives the p-values provided by the Monte Carlo test. They were very small values and again provided strong evidence for rejecting spatial independence. The p-values of the analysis of spatial-temporal interactions are summarized in columns headed $\mathbb{K}_{12}(s,t)$, $\mathbb{L}_{12}(s,t)$ and $g_{12}(s,t)$. They are larger than 0.05. Therefore, we can conclude there is no spatial-temporal association in this model, which is in complete agreement with the simulated coupling mechanism.

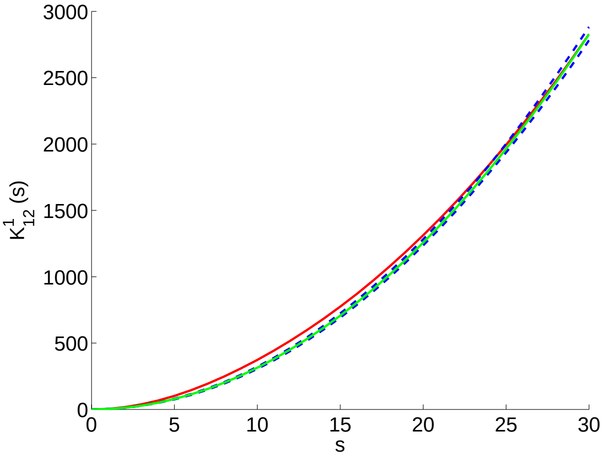
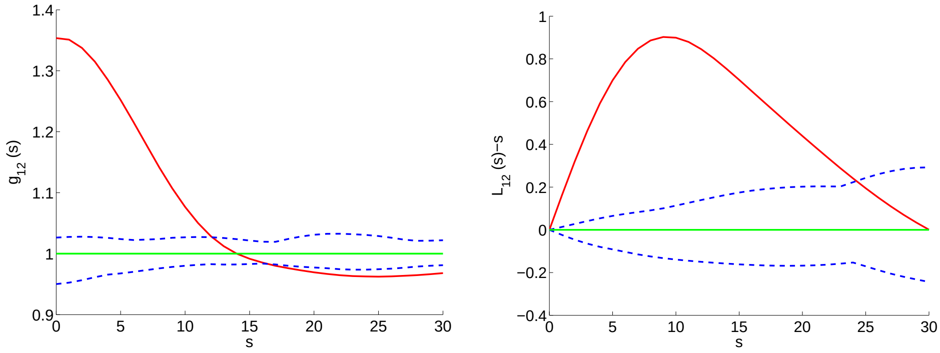
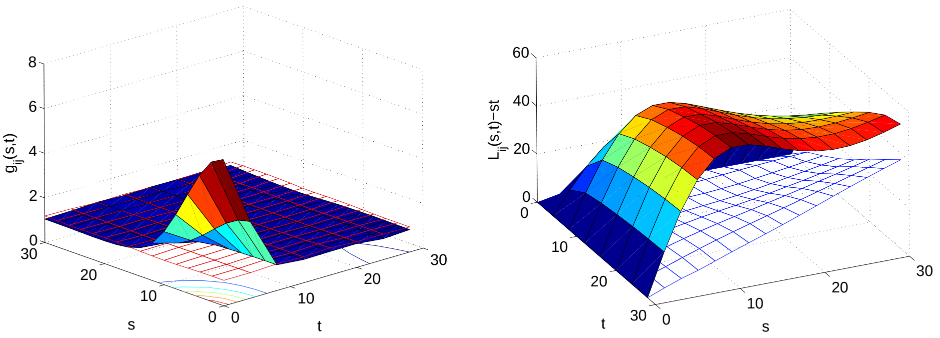


Figure 4.6: Analysis of spatial dependencies for simulation 1. The empirical function $\mathbb{K}^{(1)}_{12,0}$ and its envelopes.



(a) (b)

Figure 4.7: Analysis of spatial dependencies for simulation 1. Plot (a) corresponds to the pair correlation function, $g_{12,0}(s)$ and (b) to $\mathbb{L}^{(1)}_{12,0}(s) - s$.



(a) (b)

Figure 4.8: Analysis of spatial-temporal dependencies for simulation 3.

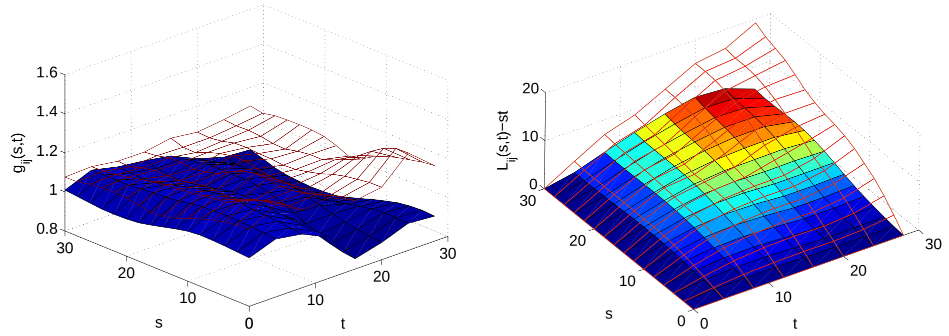
The same comments can be made on simulation 2. In this case, uniformly distributed spatial shifts within the interval $[0, 12]$ were generated between germ types. The observed pair correlation function was outside the envelopes throughout the interval of around $[0, 18]$. The p-values provided by the Monte Carlo test are given in the second row in Table 4.2.

Figs. 4.8 (a) and (b) show the functions $g_{12,0}(s, t)$ and $\mathbb{L}_{12,0}(s, t)$ for simulation 3. Fig. 4.8 (a) displays the empirical pair correlation function (colored mesh) $\hat{g}_{12,0}(s, t)$ along with the upper envelope (grid). Figs. 4.8 (b) corresponds to $\hat{\mathbb{L}}_{12}(s, t) - st$.

Both functions (colored mesh) were clearly larger than the upper envelope (grid) of the randomizations at small spatial-temporal distances, which suggests rejection of the null hypothesis of spatial-temporal independence. The mean radius for the events was 6 pixels, the mean duration was 6 seconds, the maximum spatial shift was 6 pixels and the maximum temporal lag was 2 seconds; hence, there was only interactions up to distances smaller than 12 pixels and 8 seconds. The excess of small spatial-temporal distances is compatible with the simulated underlying interactions model (see Video 2, supplementary material). The p-values given by the Monte Carlo test (third row in Table 4.2) support for this conclusion.

Similar comments can be made on simulation 4. In this case, there were spatial-temporal interactions within the interval of around $[0, 18]$ pixels and $[0, 12]$ seconds. The p-values provided by the Monte Carlo test were again almost null (fourth row in Table 4.2).

Figs. 4.9 (a) and (b) display the empirical pair correlation function $\hat{g}_{12,0}(s, t)$ and the function $\hat{\mathbb{L}}_{12,0}(s, t) - st$ for simulation 5, in which neither spatial nor temporal

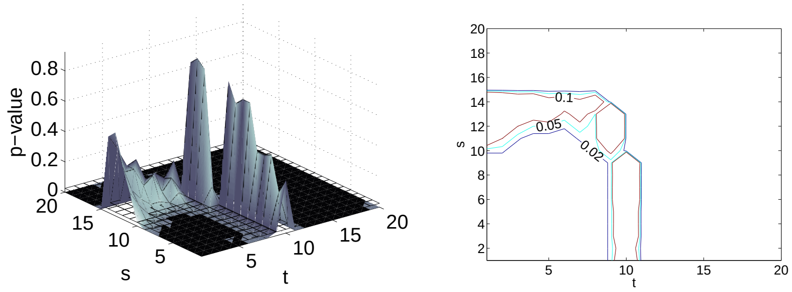


(a) (b)

Figure 4.9: Analysis of spatial-temporal dependencies for simulation 5.

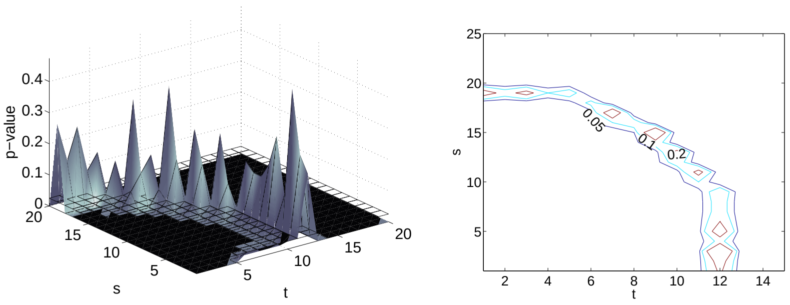
dependence was generated. The observed values (colored mesh) lay close to those of the reference value under independence throughout its range, which suggested that the null hypothesis of independence cannot be rejected. The same conclusion was reached from the Monte Carlo test. The p-values obtained were quite large (fifth row in Table 4.2). These results are in complete agreement with the simulated model.

We also applied the bootstrap algorithm described in Section 4.4.4 to each pair to estimate the spatial-temporal interval of dependencies. Fig. 4.10 plots the map of p-values obtained for Simulation 3. A grid corresponding to a p-value= 0.05 is plotted for visualization purposes. Note that p-values higher than 0.05 are observed for large values of s and t. Spatial-temporal interactions within the rectangle of around $[0, 12] \times [0, 8]$ are observed. Fig. 4.11 plots the map of p-values obtained for Simulation 4. Here, dependencies span within a rectangle of around $[0, 18] \times [0, 12]$.



(a) (b)

Figure 4.10: Estimating the interval of dependencies for simulation 3.



(a) (b)

Figure 4.11: Estimating the interval of dependencies for simulation 4.

We have implemented a software tool in Matlab to: estimate the cross-covariance function, perform testing procedures (based on toroidal shifts and Monte Carlo tests), and simulate bivariate temporal Boolean random models. The implemented toolbox can run under Windows and Linux operating systems. The computing time depends on the length of the sequence, on the image size and on the spatial and temporal intervals of estimation; for the simulated image sequences described, the average cost was ten minutes per pair of image sequences. Our Matlab toolbox (source code and user documentation) for the simulation of bivariate temporal random sets, the estimation of the cross-covariance function and the testing procedures is available at *http* : *//www.uv.es/tracs/index.html* and can be distributed under GNU license.

The simulated videos and the fluorescent-tagged protein image sequences are included in the supplementary material and they can also be downloaded from *http* : *//www.uv.es/tracs/videos/videos.html*.

The videos and the source code of the functions are also included in the supplementary material.

4.6 Application to cell Biology

We applied the proposed method for testing the spatial-temporal dependencies of pairs of proteins involved in clathrin mediated endocytosis. We visualized the endocytic events in real time by imaging cells with green/red fluorescent-tagged proteins with TIRFM. The fluorescent-tagged proteins are seen as fluorescence areas, scattered over the plasma membrane [44, 31]. We studied the spatial-temporal interactions of

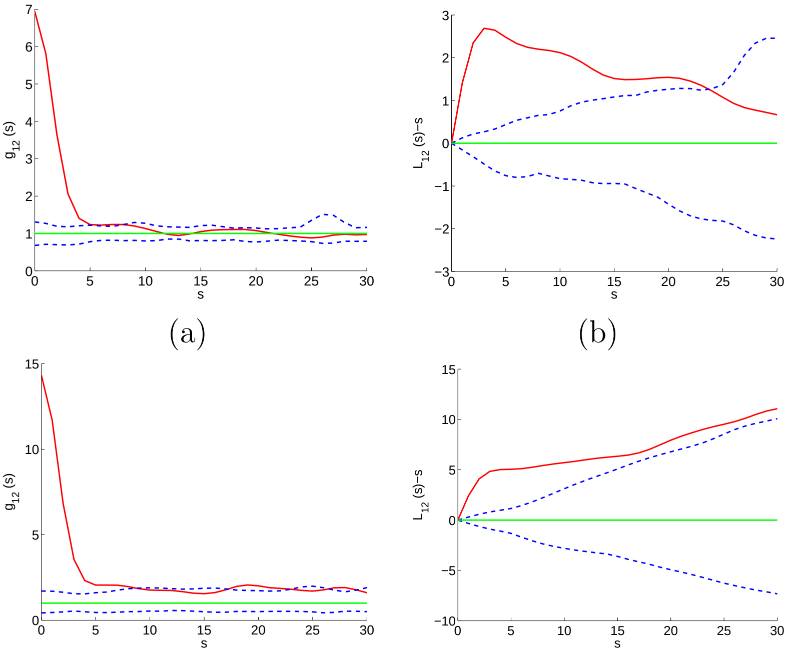
Table 4.3: Pairs of image sequences of fluorescently-labelled proteins.

Pairs	Protein type 1	Protein type 2	# of frames	Size
1	Clathrin-RFP	Hip1R-GFP	151	180×153
2	Clathrin-RFP	Epsin-GFP	203	213×185
3	Clathrin-RFP	Caveolin-GFP	78	135×281
4	Clathrin-RFP	Clatrhin-GFP	143	166×210

Clathrin [11] with other proteins: Hip1R [32, 55], Epsin [15] and Caveolin. Table 4.3 describes the pairs of sequences analyzed. In the supplementary material, Video 4 corresponds to pair 1, Clathrin-RFP is in red and Hip1R-GFP in green. Overlapping areas are in blue. Video 5 corresponds to pair 3, Clathrin-RFP in red and Caveolin-GFP in green.

Pair 1 shows spatial associations between proteins, since the p-values provided by the Monte Carlo test were almost null (see first row in Table 4.4). The same conclusion was drawn from the plots of $g_{12,0}(s)$ and $\mathbb{L}_{12,0}(s)$. Figs. 4.12 (a) and (b) display $g_{12}(s)$ and $\mathbb{L}_{12}(s)$ for Clathrin-RFP and Hip1R-GFP. The empirical functions (red line) take much larger values than the envelopes of the randomizations (blue lines), suggesting spatial dependencies. The same occurs for pair 2, composed by Clathrin-RFP and Epsin-GFP (see Figs. 4.12 (c) and (d) and the second row in Table 4.4). Note that the pair correlation function was expected to reach the reference value of 1 for long distances, since the dependencies between proteins, when they exist, are thought to occur over small ranges. In both cases, the colocalization disappeared at a distance of around $5 - 10$ pixels.

As a negative control, we measured the dependence between Clathrin and Caveolin, a protein that is not implicated in clathrin-mediated endocytosis (pair 3). The observed $g_{12}(s)$ and $\mathbb{L}_{12}(s)$ functions (red line) and their envelopes (blue lines) are displayed in Fig. 4.13 (a) and (b). In this case empirical functions fall inside the envelopes. Therefore we cannot reject the hypothesis of spatial independence. The p-values clearly deviated from 0.05 (see the third row in Table 4.4), suggesting that both proteins are spatially independent. As a positive control, we measured the dependence between two molecules of Clathrin fused with two different fluorochromes (pair 4, Clathrin-RFP and Clathrin-GFP). A high dependence for this homologous

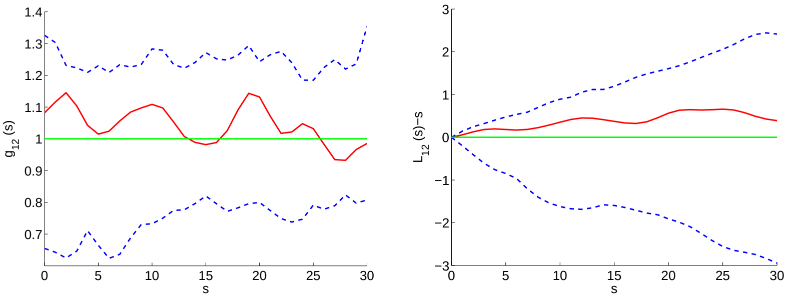


(a) (b)

(c) (d)

Figure 4.12: Plots of $g_{12}(s)$ and $\mathbb{L}_{12}(s)$ for pairs of fluorescent-labelled-proteins. Plots (a) and (b) correspond to pair 1 and plots (c) and (d) to pair 2.

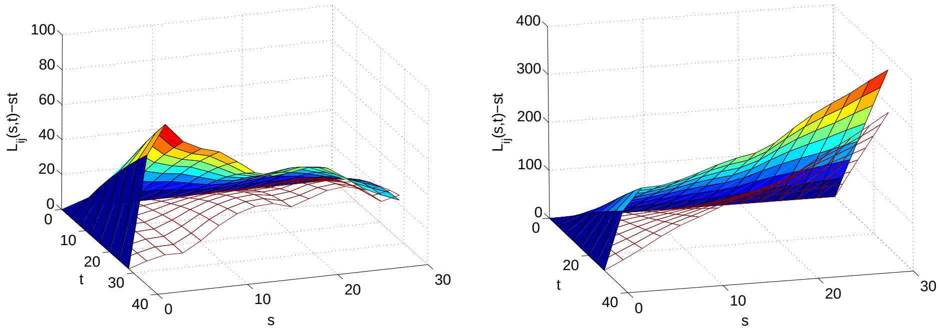
pair of Clathrin was obtained (the p-values are given in the fourth row in Table 4.4).



(a) (b)

Figure 4.13: Plots of $g_{12}(s)$ and $\mathbb{L}_{12}(s)$ for pair 3, Clathrin-RFP and Caveolin-GFP.

For the study of the spatial-temporal associations between proteins, we used $\mathbb{K}_{ij}(s,t)$, $g_{ij}(s,t)$ and $\mathbb{L}_{ij}(s,t)$. Fig. 4.14 summarizes the results obtained for the function $\mathbb{L}_{ij}(s,t)$ for pairs 1 and 2. The observed functions (colored mesh) lay outside the upper envelope of the randomizations (grid), suggesting the existence of spatial-temporal interactions between proteins. Fig. 4.15 displays results for pair 3. In this case, $\mathbb{L}_{ij,0}(s,t)$ is below the upper envelope, suggesting there were no interactions between Clathrin and Caveolin.



(a) (b)

Figure 4.14: Plots of $L_{ij}(s,t) - st$ for pairs of fluorescent-labelled-proteins. (a) Pair 1, Clathrin-RFP and Hip1R-GFP. (b) Pair 2, Clathrin-RFP and Epsin-GFP.

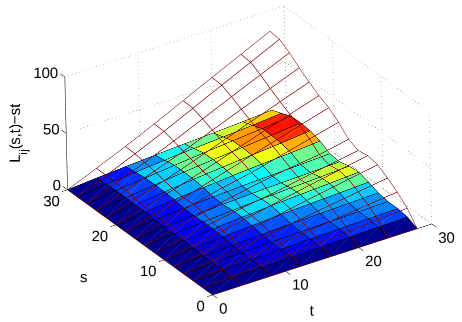


Figure 4.15: Plots of $L_{ij}(s,t) - st$ for pair 3, Clathrin-RFP and Caveolin-GFP.

In Table 4.4, the p-values obtained in the Monte Carlo test for $\mathbb{K}_{ij}(s,t)$, $g_{ij}(s,t)$ and $\mathbb{L}_{ij}(s,t)$ are provided. The three descriptors agree on the existence of spatial-temporal relationships between proteins for pairs 1 and 2, confirming that Hip1R and Epsin are integral components of the clathrin coats [89]. Hip1R and Epsin had been previously observed [32, 55, 91] to colocalize to clathrin-coated pits. However, our methodology now provides a statistical validation to those observations.

Figs. 4.16 (a) and (b) show the map of p-values obtained for pairs 1 and 2. In both pairs, proteins colocalized within a distance of around 3 pixels. The temporal

Table 4.4: p-values of Monte Carlo test applied to biological image sequences.

Pair	Proteins	$\mathbb{K}_{12}(s)$	$\mathbb{L}_{12}(s)$	$g_{12}(s)$	$\mathbb{K}_{12}(s,t)$	$\mathbb{L}_{12}(s,t)$	$g_{12}(s,t)$
1	Clat.-Hip1R	0.00	0.00	0.00	0.00	0.00	0.00
2	Clat.-Epsin	0.00	0.00	0.00	0.00	0.00	0.00
3	Clat.-Cav.	0.59	0.57	0.76	0.95	0.95	0.62
4	Clat.-Clat.	0.02	0.00	0.00	0.02	0.02	0.00

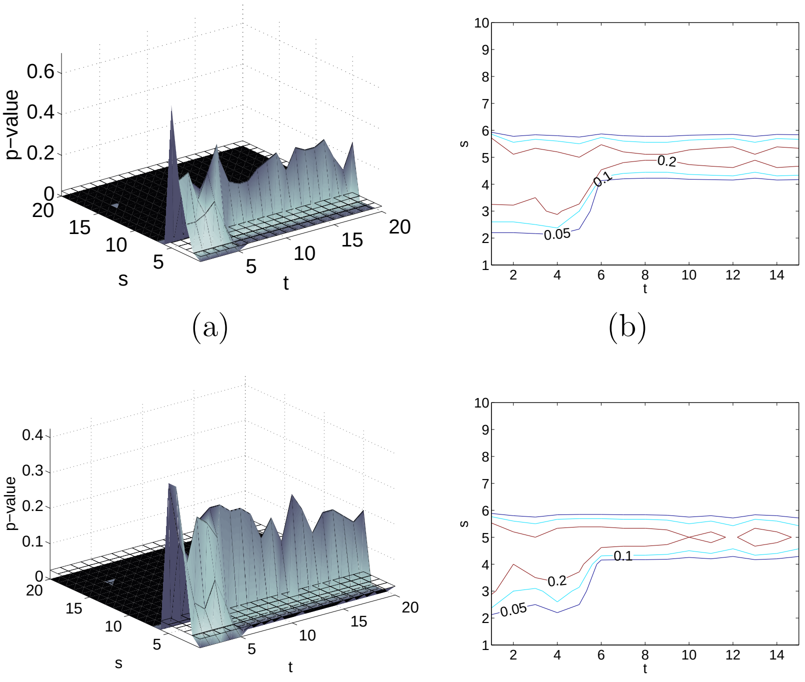
dependencies were as long as the length of the sequence.

4.7 Conclusions

In this chapter, we have shown a new procedure for analyzing either spatial or spatial-temporal interactions between two types of events in a bivariate temporal random set. We have demonstrated its utility both as a summary statistic and as a basis for formal inference. We successfully applied this method to study associations between proteins in quantification of endocytic dynamics.

We can draw the following conclusions. Firstly, the proposed methodology is not only a testing procedure, but it allows us to quantify the degree and the spatial-temporal interval of the interactions. Moreover, no (unnecessary) parametric assumptions are made. The proposed approach can be easily extended to multivariate random sets. From the application on simulated dependencies, we conclude that these summary descriptors accurately capture spatial-temporal interrelations. The results obtained in the analysis of TIRFM image sequences of fluorescent-tagged proteins are in complete agreement with those reported by other techniques. In our opinion, these estimators are sensitive enough to be applied to many other applications.

Secondly, apart from an exploratory analysis, we have proposed a formal testing procedure. This represents an important departure from simple visual inspection of the images, which is the normal practice in cell biology literature, although manual analysis is a very time-consuming task and subject to observer error. In contrast, our proposed methodology is highly reproducible and fully automatic. Such a procedure could be used as a screening tool to study protein interdependencies under different treatments and provides new opportunities for validating cellular models. To our



(a) (b)

(c) (d)

Figure 4.16: Estimating the spatial-temporal interval of dependencies. Plots (a) and (b) correspond to pair 1, Clathrin-RFP and Hip1R-GFP. Plots (c) and (d) to pair 2, Clathrin-RFP and Epsin-GFP.

knowledge, this is the first time these TIRFM image sequences are modeled within a probabilistic context and the hypotheses of independence are tested in a formal way.

In the procedure proposed, we assume stationarity. In many real applications, the bivariate temporal random set analyzed is obtained of a previous segmentation of the original grey-level image sequences. In some cases, the segmented image sequences could not be considered as a realization of a stationary model. How can we circumvent this kind of difficulties? First, if there is a gradient in the background, we could apply segmentation procedures that use the local spatial structure of the images, i.e. then the violation of stationarity becomes a computer vision problem which can often be solved. Second, if there is periodicity in the image, we could restrict the toroidal shifts to those values within the periodicity interval. Third, if there is some non-random structure, we could select an image sub-window and consider a stationary model within this sub-window.

Several aspects call for further investigation. First, the development of new parametric models for the spatial-temporal interactions where biological factors could be included. Second, when stationarity does not hold, simple non-parametric measures of association may not be sufficient. The incorporation of covariates and possible sources of interactions of the events throughout the observation window would lead to non-homogeneous distributions, i.e. non-homogeneous random sets. Parametric or semi-parametric modeling of inhomogeneity is in its infancy. Regarding to such modeling, potential extensions will include the use of other parametric models, such as a linked Cox model [24].

Chapter 5

Automatic detection of exocytic vesicles and statistical analysis of their intracellular distribution

Spatial Statistics has provided several valuable models to analyze any given set of points or processes which can generate point data following some stochastic law. Problems that involve random spatial patterns have been encountered in various Science fields, such as Biology, Ecology, Epidemiology, Geology, etc. In this chapter, we present an application of point processes to the study of cell exocytosis among different cell treatments.

5.1 Introduction

The problem of studying the effect of specific proteins on the spatial distribution of large dense-core vesicles (LDCV or granules) is central to the understanding of regulated exocytosis, as well as for synaptic transmission. Vesicles whose destination is the plasma membrane leave the trans Golgi network in a steady stream. The membrane proteins and the lipids in these vesicles provide new components for the plasma membrane, while the soluble proteins inside the vesicles are secreted to the extracellular space. The fusion of these vesicles with the plasma membrane is the final step of the exocytosis process.

Exocytosis can be affected by inhibition of critical exocytic molecular machinery, resulting in potential perturbation of the spatial distribution of the granules within the cytoplasm and of the morphological properties and number of vesicles.

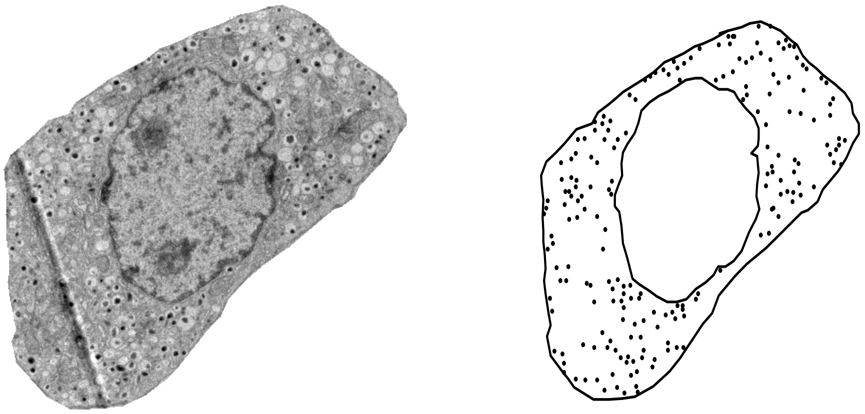
Ultrastructural studies have identified LDCV from transmission electron micrographs of cross-sections of cells (see Fig. 5.1 (a)). Three different types of vesicles are observed: i) stand-alone LDCVs with a dark appearance (granules); ii) dense vesicles with “halos”, i.e. granules with white vacuoles surrounding them; and iii) isolated vacuoles, light round objects. In this study we are concerned with cases i) and ii) corresponding to granules, i.e. small round dark secretory vesicles.

It is usually assumed that white vesicles have some functional relationship with electron-dense adrenal secretory vesicles, since they also change in number, size and appearance, when cells are stimulated. However, we do not consider them in the present work, although this analysis is considered as future work. Regarding other types of electron-lucent vesicles described in [47], we do not analyze them either, as they do not have apparently any relationship with secretory vesicles.

Three different cases were observed in our case: (1) stand-alone or isolated dense granules, with dark appearance; (2) dense vesicles surrounded by irregular white halos or zones, lighter than stand-alone vesicles and (3) isolated vacuoles, that seem to be empty granule membranes. We consider as granules all the dense-core or dark secretory vesicles, independently from the existence or the absence of halos [19].

The granules are formed in the Golgi apparatus and travel across the cytoplasm towards the plasma membrane, where they ultimately fuse [85]. A necessary and critical step prior to vesicle fusing is the close morphological apposition of the vesicle to the plasma membrane (e.g. less than 100 nanometers), a process morphologically described as vesicle ”tethering” or ”docking” [13]. Comparing the distances from the granules to the plasma membrane among cell groups is especially important when studying the factors that influence vesicle docking. This study is founded on a crucial biological interest in understanding the different vesicle pools (defined by electrophysiology kinetic measurements) and their distribution in regulated secretion. Specialized secretory cells include chromaffin cells, mast cells and insulin cells, which are respectively responsible for the release of adrenalin, insulin and histamine.

The spatial distribution of the granules in the Cell Biology literature is usually quantified by means of the distances from the granules to the plasma membrane [41, 43, 46, 47, 61, 62]. Each cell provides a sample of granule distances. It is a common



(a) (b)

Figure 5.1: A section of a wild-type adrenal chromaffin cell and the spatial point pattern associated.

practice to join all the distances obtained from different cells into a unique sample per group [90, 61, 43]. The comparison among groups is performed by comparing these sets under visual inspection combined with simple exploratory data analysis. In our opinion this is not an appropriate analysis, because it has no into account the interdependence between the distances observed in a given cell. Granules may interact among themselves in such a way that their locations are interdependent. Their distances to the plasma membrane will be (possibly) dependent too. Indeed, in compound exocytosis of granules, this has been shown to be the case. Alternatively, the protein fusion machinery may be concentrated in a given site, which could lead to exocytosis in preferred areas of the cell. In fact, our analysis of constitutive exocytosis has shown precisely this fact [45, 70].

Our data consist of a spatial point pattern (the locations of the granules) within a given region (the cytoplasm) for each cell (see Fig. 5.1 (b)). The dots correspond to the granule centers. The external and internal solid lines delimit the plasma membrane and the nucleus membrane, respectively. The annular region corresponds to the cytoplasm.

It is important to note that we have replicated point patterns for each cell group. We model the locations of granules of a given cell as a realization of a spatial point process and we consider the point patterns associated with the different cell groups as replicated point patterns of possibly different spatial point processes. A set of spatial locations is known in the statistical literature as *spatial point pattern*, a realization of a *spatial point process*. Good standard references on point process theory are [17, 24, 80, 79].

In this study, we propose two functional descriptors so as to describe the relative locations of the granules with respect to the plasma membrane: i) the empirical cumulative distribution function of the distances from the granules to the plasma membrane; and ii) the mean number of granules per unit area within a given distance to the plasma membrane. This second descriptor captures the arbitrary morphology of the cytoplasm. The comparison among cell groups is performed by using bootstrap techniques. As a biological application that validates the proposed methodology, we analyze 77 cells of the adrenal glands of wild-type (38) and treated (39) mice in order to study whether the granules in the treated group are closer to the plasma membrane than in the control group (wild-type).

In Section 5.2, the image processing method to segment the granules using electron microscopy images is outlined. The statistical framework to model and analyze the data is detailed in Section 5.3. Section 5.4 is devoted to the results. A simulation study is also included in Section 5.5. Conclusions are summarized in Section 5.6.

5.2 Data and image processing

Images were acquired using Electron Microscopy and they have high spatial resolution. Since experiments imply the analysis of dozens of cells, each of which contains over one hundred granules per cell section, the development of an automatic software tool is needed. The image processing algorithm to automatically detect the granules is described in Section 5.2.1. Granule descriptors based on shape criteria and on the grey-levels distribution are proposed in Section 5.2.2. A Linear Discriminant Analysis is performed in Section 5.2.3 in order to select the relevant descriptors for granule characterization and classification.

5.2.1 Detection of granules

We developed a detection algorithm to collect the locations of granules in the cytoplasm for dozens of cells. To our knowledge, no procedures have been reported in the literature using EM images. Here, we present an image processing within the context of mathematical morphology. In [50], an image processing method based on the Hough Transform is presented in order to segment granules in pancreatic acinar cells using fluorescence from light microscopy images. However, due to the complexity of our images, we consider the Hough Transform would lead to worse results because of its sensitivity to false edges detection.

Step 1: Image pre-processing. The cytoplasm (the region where granules can be observed) was defined for each cell by delineating the plasma membrane and the nucleus membrane. We used the negative of the original image. In this way, granules appeared in the image as bright objects over a darker background (see Fig. 5.2 (b)). Since images often presented a non-uniform illumination, the background was subtracted. The background image (Fig. 5.2 (c)) was obtained by smoothing the negated image with a median filter of size 19×19 pixels, since the diameter of granules in chromaffin cells ranges from 100 to 600 nm [46, 47, 61, 62], equivalent to 6 to 18 pixels respectively in our images.

Step 2: Granule Segmentation. A Top-Hat transformation [72, 76] with a circular structuring element was applied to the negated image (after background subtraction) in order to extract the intensity peaks and to eliminate other organelles in the cytoplasm associated with larger and less bright objects. The radius of the structuring element was 9 pixels (260 nm), equal to the radius of the largest granule to be segmented. The resulting image was composed of small bright regions (see Fig. 5.2 (d)). Afterwards, the Top-Hat image was thresholded with a threshold of 40 (see Fig. 5.2 (e)). This value is a tuning parameter, that depends on the contrast, brightness or resolution of the images.

Step 3: The Watershed. We used a Watershed transformation [72] in order to separate overlapping granules. First, a marker image was obtained from the regional maxima of the distance transformation of the binary image. Second, a Watershed transformation constrained by the marker image was applied to the negation of the distance function (see Figs. 5.2 (f) and (g)).

Step 4: Grey-scale morphological reconstruction. The resulting labelled image

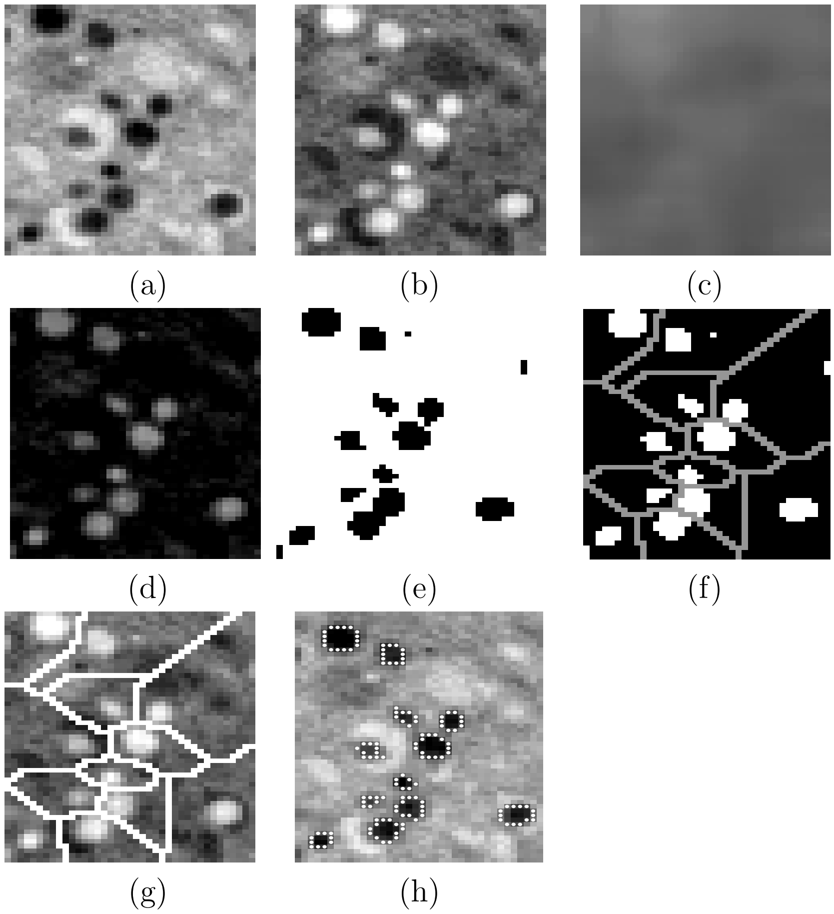


Figure 5.2: Image processing. (a) Original image. (b) Negated image. (c) Background of the negated image. (d) Top-Hat image. (e) Binary image after thresholding. (f) Watershed lines over the binary image. (g) Watershed lines over the negated image. (h) Segmented granules.

from Step 3 provided us with an initial segmentation containing all possible candidate granules. For each object in the candidate granule image (marker), a grey-scale morphological reconstruction of the original image (conditioning image) was applied. The result is a full delineation of each marked object in the original grey-level image (see Fig. 5.2 (h)). White dots represent the segmented granule contours. The reconstruction was restricted to the watershed division lines avoiding objects from encroaching on other very close granules.

The image processing algorithm for granule segmentation using EM images is detailed now. This software was implemented in Matlab and is available on request. The computational cost was less than 10 seconds per image.

Algorithm 1: Image processing algorithm for granule segmentation using EM images

```
begin
    Delineate the cytoplasm in the original image $I_{original}$
    Calculate the negation of $I_{original}$, $I_{neg}$
    Apply a median filter to obtain the background, $I_{bg}$
    Subtract $I_{bg}$ from $I_{neg}$, giving $I_{minus}$
    Apply a Top-Hat transform to $I_{minus}$ to detect intensity peaks, obtaining $I_{Top-Hat}$
    Threshold of $I_{Top-Hat}$, resulting in $I_{bin}$
    Apply a Watershed transform to $I_{bin}$, obtaining $I_{Watershed}$
    Separate overlapping granules by image subtraction
    $I_{markers} = I_{Watershed} - I_{bin}$
    Grey-level reconstruction of the granules by using $I_{markers}$ and $I_{neg}$
end
```

5.2.2 Shape and grey-level descriptors

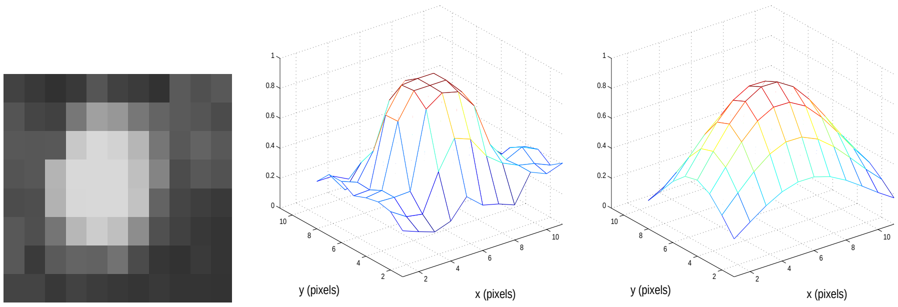
In order to achieve high sensitivity in the detection algorithm, the threshold level applied to the Top-Hat image was set to a low value (40). In this way, the number of false negatives (granules not detected by the algorithm) decreases, whereas the number of spurious objects detected (false positives) increases. In this subsection, several numerical descriptors of the shape and the grey-level distribution of the granules are associated with each candidate granule. We then classify candidates as spurious objects or true granules by means of a Bayes classifier. Let us introduce the descriptors used.

Since true granules appear in the images as symmetric and round objects, we used the following shape descriptors. All these measures range from 0 to 1, being close to 1 if the granule shape is a circle.

- the circularity, defined as $4\pi * area/perimeter^2$;
- the eccentricity, defined as the ratio of the foci distance and the major axis length;

- the ratio of the minor axis and the major axis of the object; and
- the ratio of the diameter of a circle with the same area as the object (the equivalent diameter) and its major axis.

We also described each granule by means of its grey-level distribution. Let x_0 be the center of a candidate granule, and R a 19×19 window centered at x_0. Let x be the two-dimensional coordinates of a pixel of the candidate granule. Let us consider the grey-level observed as a function of x. For each candidate granule, we estimated the granule radius by fitting a two dimensional Gaussian to the raw grey-levels. We followed the procedure described in [14], pp. 504-506. We assumed a symmetric bell for the shape of the granule. Fig. 5.3 (a) shows the raw image of a typical, rounded granule, Fig. 5.3 (b) displays a three-dimensional representation of the grey-level values and Fig. 5.3 (c) plots the Gaussian fit. Let $\hat{\sigma}$ be the estimated marginal standard deviation (radius) of the Gaussian. The area of the candidate granule was estimated as $\pi\hat{\sigma}^2$. For each granule, we calculated the mean energy, the standard deviation and the coefficient of variation of the grey-levels within a circular window centered at the granule with radius $\hat{\sigma}$. Granules usually have a higher mean and standard deviation than other cell organelles. Table 5.1 shows the descriptors proposed to characterize true granules.



(a) (b) (c)

Figure 5.3: Fitting a two-dimensional Gaussian. (a) Raw grey-level image. (b) Three-dimensional plot of the raw image. (c) Gaussian fit.

Table 5.1: Shape and grey-level descriptors.

Binary shape	Grey-level
Circularity	Mean energy
Eccentricity	Standard deviation
$\frac{Minor\ axis}{Major\ axis}$	Coefficient of variation
$\frac{Equivalent\ diameter}{Major\ axis}$	

5.2.3 Linear Discriminant Analysis

A feature vector $\mathbf{x} = (x_1, \ldots, x_k)$ was associated with each granule candidate. We were concerned with a two-class classification problem: granules (true positives) and other spurious objects detected in the cytoplasm (false positives). A Linear Discriminant Analysis (LDA) was performed (by using SPSS) in order to evaluate the descriptors proposed. The performance of the image processing method was assessed by comparing computer results with granule counts carried out by two experts who manually marked each granule in six cell images. The percentages of granules correctly classified for each cell by using all the descriptors were 94.1% in the worst case and 97.3% in the best case. The median observed was 96.2%. We used 1114 candidate granules, of which 1027 were true granules and 87 other spurious objects.

When the granules of all the cells were taken together in a unique analysis, the percentage of correct classification decreased slightly, as expected, due to higher variability in the sample. The percentage of granules correctly classified by using all descriptors was 93.8%. The percentage obtained by using only shape descriptors was 91.7%. Descriptors based on the grey-level distribution led to slightly better discrimination (93.3%).

For each of the three sets of descriptors considered (shape, grey-level, all), the robustness of its classification was measured. Table 5.2 shows a summary of the mean probabilities of correctly classifying true and false positives for each set of descriptors. The three sets of descriptors had a very high probability of classifying true granules into the correct group. Interestingly, descriptors based on the grey-level distribution showed higher performance than the shape descriptors when classifying false positives. The highest values were obtained when combining all the descriptors.

We measured the sensitivity of the detection method, i.e. the ratio of the number

Table 5.2: Mean probabilities of belonging to a given class.

Descriptors	Shape	Grey-level	All
Class 1 (FP)	0.404	0.657	0.707
Class 2 (TP)	0.942	0.947	0.952

of correctly detected granules (true positives) and the total number of granules (true positives plus false negatives). A good reference for the validation of medical image techniques is [77], pages 569 – 604. The sensitivity was 96.2% in the best case and 91.1% in the worst case. The median was 92.6%. False negatives were granules poorly contrasted with the background, too small or with a very irregular shape [19].

5.3 Statistical analysis

The experiment was designed in order to test whether there were differences in the distances from granules to the plasma membrane between two cell groups. Our data consist of the granule locations and the cytoplasm for each cell. We have different finite point patterns defined over different regions, i.e. the cytoplasms of the different cells. In order to compare the different treatments we need to formulate the probabilistic framework.

Let us introduce some notation. For a given cell, let C and Φ be the cytoplasm and the locations of the different granules, respectively. Both sets C and Φ can be considered as random sets where $\Phi \subset C$ (with probability one). The cytoplasm C would be a realization of a random closed set and Φ would be a realization of a finite point process. Now, we will describe the distribution of Φ with respect to the plasma membrane.

We have g groups of cells (in our case $g = 2$) and n_i cells per group, i.e. we have n_i replicates of the finite point process associated with the i-th cell group observed over different (random) regions, the cytoplasms. Let (Φ_{ij}, C_{ij}) be the point pattern and the cytoplasm corresponding to the j-th cell of the i-th group. Let $\Phi(A)$ be the random variable which gives the number of points within A of the point process Φ, where A denotes an arbitrary (Borel) subset of $\mathbb{R}^2$. In particular, $\Phi(C)$ is the total number of points of Φ.

The first hypothesis to be tested is whether the counts are similar for the different groups. Let the random variable $Y = \frac{\Phi(C)}{\nu_2(C)}$ denote the mean number of points of Φ per unit area of cytoplasm in a given cell, where ν_2 stands for the area. Let Y_{ij} be the random variable corresponding to the j-th cell of the i-th group. The random variables $Y_{i1}, \ldots, Y_{in_i}$ are a random sample of (let us say) Y_i. Let $\mu_i = EY_i$ be the mean number of granules per unit area in a typical cell of the i-th group, then the null hypothesis to be tested is $H_0 : \mu_1 = \ldots = \mu_g$. In Section 5.4.1, the application of ANOVA and generalized linear models to our dataset is discussed.

The second hypothesis is concerned with the analysis of differences in the relative location of granules with respect to the plasma membrane among cell groups. An usual practice in biological literature is to group the distances of granules measured in different cells into the same treatment group and compare groups of distances. In our opinion, this procedure is not appropriate, because distances from the same cell are dependent and distances from different cells are independent.

To circumvent this problem, we consider the random variable D defined as the distance from an arbitrary chosen granule to the plasma membrane. If $d_1, \ldots, d_n$ are these distances observed for a cell, these values are a (not random) sample of the random variable D, because the different distances are interdependent. We propose to describe the distances in a given cell by using the empirical distribution function and then to compare the estimated functions using a bootstrap procedure in order to test differences among cell groups. The *empirical cumulative distribution function* (from now on, ECDF) of the distances is given by

$$\hat{F}(t) = \frac{\#\{d_i : d_i \leq t\}}{n}, \tag{5.1}$$

where # stands for the number of, i.e. for each t, $\hat{F}(t)$ is the proportion of distances lesser than or equal to t. If $\hat{F}_{ij}$ denotes the estimated function for the j-th cell of the i-th group and F_i denotes the (unknown) theoretical distribution function of the distance from an arbitrary granule to the plasma membrane for the i-th group then, under the null hypothesis of no difference among groups, it holds that $F_1 = \ldots = F_g$. We will test it using a bootstrap procedure. Results are described in Section 5.4.2.

Although this approach is a natural one, it does not take into account the morphology of the cytoplasm. The different distances observed in a given cell are dependent, among others factors, on the location and size of the nucleus. It seems a valid alternative to consider the mean number of points per unit area of cytoplasm within a given distance from the plasma membrane. Note that a non-stationary point process is assumed and therefore other usual functional pattern descriptors such as the $\mathcal{K}$-function or the empty space function cannot be used [24].

Let $\lambda(x|C)$ be the intensity function of Φ given a cytoplasm C. This intensity function is the one verifying, for a (Borel) subset of $\mathbb{R}^2$, A, the equation:

$$E[\Phi(A)|C] = \int_A \lambda(x|C)dx. \tag{5.2}$$

In particular, $\lambda(x|C) = 0$ if $x \notin C$. Fig. 5.4 shows the intensity functions of two cells of different groups. Fig. 5.4 (a) corresponds to a mutant cell and Fig. 5.4 (b) to a wild-type cell. Both cells have a similar number of granules (105 granules for the mutant cell and 106 granules for the control cell) and a similar number of granules per unit area ($4.62 \cdot 10^{-6}$ granules per unit area of cytoplasm for the mutant cell and $4.60 \cdot 10^{-6}$ granules per unit area of cytoplasm for the control cell). However, the intracellular distributions of their granules are quite different. The mutant cell has granules close to the plasma membrane and the control cell shows granules scattered along the cytoplasm. Color is proportional to the intensity function value, from small values (green) to higher values (yellow, orange and white).

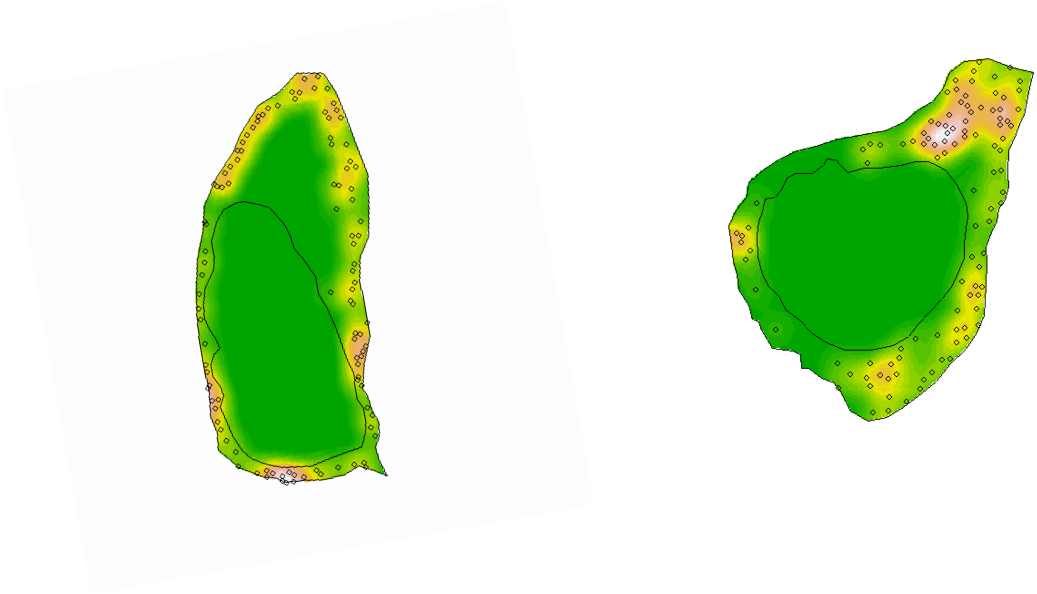
Let $N(x, h)$ denote the number of events of the Cox process within distance h of the point x and $B(x, h)$ the disk of center x and radius h. The intensity function $\lambda(x|C)$ can be estimated as follows (see [24], page 117):

$$\hat{\lambda}(x|C) = \frac{N(x, h)}{|A \cap B(x, h)|}, \tag{5.3}$$

being A the region where the data are observed.

The estimator $\hat{\lambda}(x|C)$ can be seen as a kernel estimator which uses the kernel function $k(u) = (\pi u^2)^{-1}$ if $0 \leq u \leq 1$, and 0 if $u > 1$. The function *kernel2d* included in the package *Splancs* of *R* has been used [63].

Note that a given point x cannot be compared between different images because the location and orientation of the cell within the image is completely arbitrary. We quantify the mean number of points of the point processes within a given distance



(a) (b)

Figure 5.4: Intensity function $\lambda(x)$ of the process for two cells. (a) A mutant cell. (b) A control cell. Color is proportional to the intensity function value, from small values (green) to higher values (yellow, orange and white).

from the plasma membrane. For a given cytoplasm C, let C_ϵ be the set of points of C within distance ϵ from the outer boundary of C, i.e. from the plasma membrane. More formally,

$$C_\epsilon = \{x \in C : d(x, \partial_o C) \leq \epsilon\}, \tag{5.4}$$

where $\partial_o C$ denotes the plasma membrane and $d(x, \partial_o C)$ denotes the Euclidean distance from a point x to the set $\partial_o C$. Let us define the following function

$$h_C(\epsilon) = \frac{1}{\nu_2(C_\epsilon)} \int_{C_\epsilon} \lambda(x|C)dx. \tag{5.5}$$

For a given cytoplasm C and a given positive value ϵ, the function $h_C(\epsilon)$ gives us the mean number of points per unit area within distance ϵ from the plasma membrane. The function h_C is estimated by replacing the integral with the corresponding count $\Phi(C_\epsilon)$ in (5.5). Let $\hat{h}_{ij}$ be the estimated function corresponding to the j-th cell of the i-th group. The different $\hat{h}_{i1}, \ldots, \hat{h}_{in_i}$ are estimators of the unknown h_i. Again, we will test the null hypothesis $H_o : h_1 = \ldots = h_g$ using a bootstrap procedure. The results are detailed in Section 5.4.3. It is important to note that this function depends on C. In some way, C is a co-variable and we do not know how h depends on it. We will assume that this dependence is negligible. We believe that this previous (and necessary) hypothesis is tenable at small values of ϵ.

5.4 Results

5.4.1 Comparing counts of granules

We compared the observed mean numbers of points per unit area of cytoplasm among cell groups. The mean of this random variable was denoted by μ_i for the i-group. We tested if a common value for the different μ_i's can be assumed.

If a normal distribution is assumed, then a usual analysis of variance (a t-test in our case because we have two groups) can be applied. The p-value observed was 0.7795, i.e. no statistically significant difference between the means was observed. Note that the standard analysis based on ANOVA remains as the only choice in Cell Biology and Neurobiology literature in spite of the fact that it is unsuitable in many cases (small counts over small areas, for instance).

The random variable Y is a rate between the count and the area of the cytoplasm. We can consider a random count over a fixed area. The total number of granules in the cytoplasm C, $\Phi(C)$, is a discrete random variable, whereas the area of the cytoplasm $\nu_2(C)$ is not random. We performed a Poisson regression for rates, where a Poisson distribution was assumed for $\Phi(C)$. The mean of the random variable $E\Phi(C)$ is related to the dummy variable indicating the group by using the log link function

$$\log(E\Phi(C)) = \nu_2(C)(\beta_0 + \beta_1 s_1), \tag{5.6}$$

where $s_1 = 1$ in the mutant group and 0 otherwise ([1]).

We tested the null hypothesis $H_0 : \beta_1 = 0$ i.e. there was no difference in the means for the two groups. The observed p-value was 0.6387. Again, there was no evidence against the null hypothesis. Note that we compared nested models where the simpler model had the coefficient β_1 equal to zero. It is well-known that although we do not have a very good global fit, the comparison between nested models is still suitable. In this case we had large deviance residuals, probably indicating overdispersion. The deviance residuals observed ranged from -7.2158 to 8.3725. When a negative binomial distribution was assumed for the random count, the residuals ranged from -0.79253 to 0.84609. The null hypothesis of a null β_1 seems consistent with a p-value observed of 0.781. Note that this new p-value was very close to the one obtained assuming a normal distribution for the rates.

In summary, the null hypothesis of a common mean of granules per unit area of cytoplasm under two different treatments (mutant and control) groups cannot be rejected. Statistical analysis was performed using the function *glm* and *anova.glm* included in the package *MASS* of the *R* language [63].

5.4.2 Comparing distances to the plasma membrane

Each cell provided us with a set of distances. As usual in the literature, we joined all the distances from different cells into a unique set per group and we performed a comparison of these two sets by using ANOVA. The p-value obtained for this analysis was 0.023, i.e. the distances from the granules to the plasma membrane are different for the two groups at a significance level of 0.05. The main drawback of this approach concerns the interdependence between the different distances observed in the same cell. A classic ANOVA assumes random samples taken from each group, in particular that the distances observed for a given cell should be independent among them. That does not hold in our case because granules may interact in such a way that their locations are interdependent and, by definition, are their respective distances to the plasma membrane.

To circumvent this problem, we described the granule distances by using the empirical cumulative distribution function and compared the estimated functions by using a bootstrap procedure. The empirical cumulative distribution functions were estimated at 100-equally spaced points from 28.88 to 2888 nm (corresponding with 1 pixel and 100 pixels). Let $\hat{F}_{ij}(t_k)$ be the empirical cumulative distribution function of the j-th cell, with $j = 1, \ldots, n_i$, in the i-th group, with $i = 1, \ldots, g$, at the point t_k. The functions $\hat{F}_{ij}$'s for different j's are independent estimates of the common unknown theoretical distribution function F_i for the i-th group. The function $F_i(t)$

can be estimated as

$$\hat{F}_i(t) = \sum_{j=1}^{n_i} \frac{\hat{F}_{ij}(t)}{n_i}. \tag{5.7}$$

In order to estimate the sampling variance of $\hat{F}_i(t)$, a bootstrap method is used [30]. We construct a bootstrap sample of F-functions $\{\hat{F}^*_{i1}, \ldots, \hat{F}^*_{in_i}\}$, i.e. the functions $\hat{F}^*_{ij}$ with $j = 1, \ldots, n_i$ are sampled at random with replacement from the set of functions $\{\hat{F}_{i1}, \ldots, \hat{F}_{in_i}\}$. The resampled $\hat{F}^*_i(t)$ is computed as

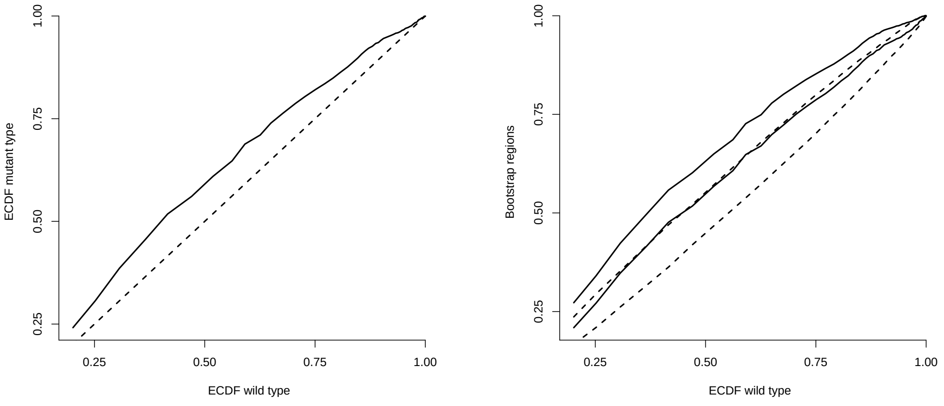
$$\hat{F}^*_i(t) = \sum_{j=1}^{n_i} \frac{\hat{F}^*_{ij}(t)}{n_i}. \tag{5.8}$$

This process is repeated s times (in our analysis $s = 1000$) and the sample variance of the corresponding values of $\hat{F}^*_i(t)$ is the bootstrap approximation to the sampling variance of $\hat{F}_i(t)$.

Fig. 5.5 (a) displays the $\hat{F}_2$ function (mutant group, solid line) against that of the control group (dashed line). The larger the difference of $\hat{F}_2$ and the bisecting line is, the larger the difference between the distribution of the distances of mutant cells and the wild-type cells is. Cells of the mutant group had granules closer to the plasma membrane than control cells, since the $\hat{F}_2$ function was clearly higher than the bisecting line at any point of the empirical cumulative distribution function.

We performed a bootstrap test and obtained the bootstrap regions. Fig. 5.5 (b) shows the bootstrap regions of $\hat{F}_2$ function (solid line) against those of $\hat{F}_1$ (dashed line). For each group, the two plus and minus bootstrap standard error limits for $s = 1000$ are displayed. We can conclude that there are differences between the distribution of the distances to the plasma membrane of the mutant group and wild-type group. Although there is a slight overlapping in the bootstrap regions at very small distances (from 0 to 57.76 nm), they quickly differentiate and do not overlap for distances from 57.76 nm to 375.44 nm, which includes docking distances.

Apart from this graphical test, we needed to formally quantify the differences between groups. A bootstrap approach was again used. A detailed description of this test can be found in [24, 25] (pages 124-127). Let $\hat{F}_{ij}$ be the j-th function of the i-th group. As before, $\hat{F}_{ij}$, with $j = 1, \ldots, n_i$, are independent and identically distributed realizations of the unknown F_i. The null hypothesis to be evaluated is that there is no difference between these functions, $H_0 : F_1(t) = \ldots = F_g(t)$, for all t. Under this



(a) (b)

Figure 5.5: Comparison of ECDFs: (a) $\hat{F}_2$ function (mutant, solid line) against $\hat{F}_1$ (control, dashed line). (b) Bootstrap regions.

null hypothesis, let us denote the common function by $F_0(t)$. This function can be estimated as

$$\hat{F}_0(t) = \sum_{i=1}^{g}\sum_{j=1}^{n_i} \hat{F}_{ij}(t) / \sum_{i=1}^{g} n_i. \tag{5.9}$$

The statistic used is a generalization of the between-group sum of squares used in a typical analysis of variance and it is defined as

$$T = \sum_{i=1}^{g} n_i \int_0^{r_0} (\hat{F}_i(t) - \hat{F}_0(t))^2 dt, \tag{5.10}$$

where r_0 was the maximum value at which the functions were estimated. Since it is not possible to know the null distribution of the statistic T given in (5.10), a bootstrap procedure was used. A bootstrap sample of functions was generated, i.e. a random sample of size $\sum_{i=1}^{g} n_i$ drawn with replacement from the population of $\sum_{i=1}^{g} n_i$ functions $\{\hat{F}_{ij}$ with $i = 1, \ldots, g,$ and $j = 1, \ldots, n_i\}$. We calculated the statistic T for the original sample and the bootstrap sample. In our case, $s - 1$ bootstrap samples were generated and from them the corresponding bootstrap replicates of T. If t_1 is the observed T from the original data and t_k with $t_k = 2, \ldots, s$ the bootstrap replicates of T, under the null hypothesis any permutation of $\{t_1, t_2, \ldots, t_s\}$ has the same probability, i.e. they are interchangeable. The null hypothesis is rejected for large t_1 values. The p-value is $P(T \geq t_1|H_0)$ and the corresponding bootstrap estimate is $\#\{t_i \geq t_1 \text{ with } i = 1, \ldots, s\}/s$, where $\#$ stands for set cardinality.

The procedure was applied with $r_0 = 100$ pixels (which corresponds to 2888 nm) and $s = 1000$. A p-value of 0.029 was obtained. Therefore, we concluded that the distribution of the distances from the granule to the plasma membrane in the mutant sample was significantly different from that of the control specimens [19].

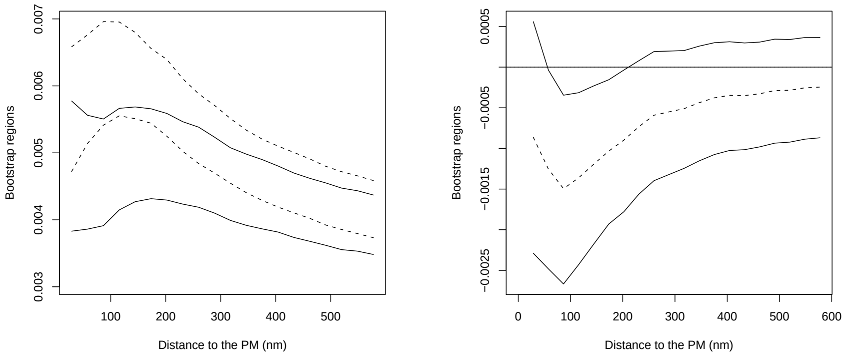
5.4.3 Comparing densities within a given distance to the plasma membrane

A similar procedure was applied to compare the h-functions of the two cell groups. These functions give the mean number of granules per unit area within a given distance from the plasma membrane. Fig. 5.6 (a) displays the bootstrap regions corresponding to h_1 and h_2. Cells of the mutant group (solid line) had a higher density of granules closer to the plasma membrane than control cells (dashed line). Likewise, we estimated the bootstrap region for the difference $h_1 - h_2$. We resampled from each sample and considered the difference between the corresponding $\hat{h}_1^*$ and $\hat{h}_2^*$. The corresponding mean ± two standard deviation were calculated and displayed in Fig. 5.6 (b) (solid line). Note that the line $y = 0$ is not contained in the region from approx. 60 to 200 nm.

The bootstrap test was applied to the h-function with $r_0 = 2888$ nm and $s = 1000$. A p-value of 0.099 was obtained. This p-value is not significative. However, from Fig. 5.6 (b), it can be drawn that there were differences between the two groups. Note that the experiment was designed in order to investigate differences in the number of docking granules at the plasma membrane, i.e. granules whose distances to the plasma membrane is less than 300 nm.

5.5 A simulation study

Let us evaluate the performance of the proposed descriptors by performing a simulation study. Let D_0 be a fixed disk with unit radius centered at the origin, which



(a) (b)

Figure 5.6: Bootstrap regions: (a) bootstrap regions of h_1 (dashed line) and h_2 (solid line); (b) bootstrap regions of the difference $h_1 - h_2$.

will represent a cell section. First, we generated D_1, a circle with radius 0.5 uniformly located within D_0, to simulate the cell nucleus. Second, we generated a non-homogeneous Poisson point process with intensity function

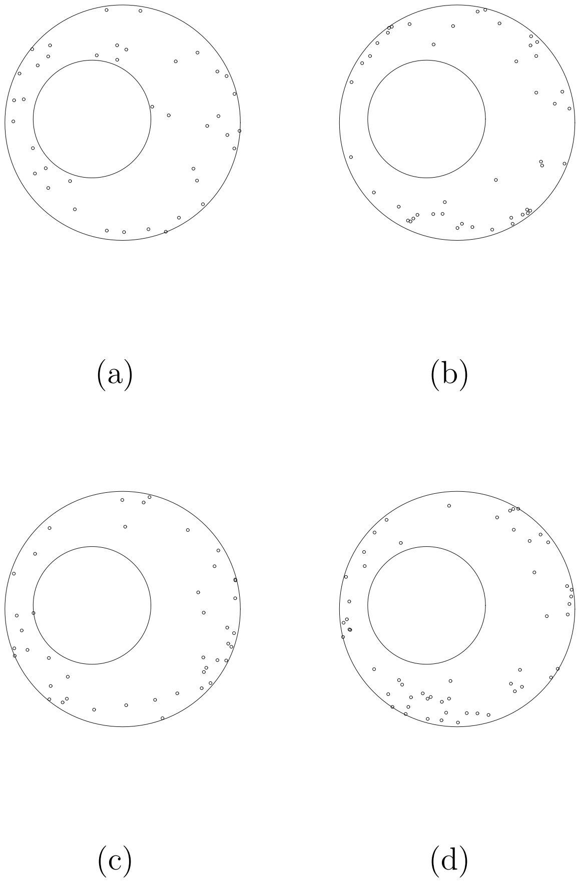
$$\lambda(x|C) = \beta_0 \exp\{-\beta_1 d(x, \partial D_0)\},$$

if $x \in D_0 \backslash D_1$ and 0 otherwise, where ∂D_0 denotes the boundary of the circle D_0. Note that all points generated are contained within $D_0 \setminus D_1$. The greater the parameter β_1 is, the closer the events to the boundary of D_0 are. For a given β_1 value, the corresponding β_0 was chosen in such a way that the mean number of points was constant (equal to 50 points in our case).

Three different sets of fifty replicated point patterns each were simulated. Each set had a different β_1 (and the corresponding β_0). In particular, $\beta_1 = 5$ for the first group, $\beta_1 = 5 + \delta$ for the second group and $\beta_1 = 5 + 2\delta$ for the third group, with values of δ ranging from 0 to 1 with a step of 0.025. Note that greater values of δ correspond to greater differences among the three groups of point patterns. Fig. 5.7 displays different realizations for different δ values.

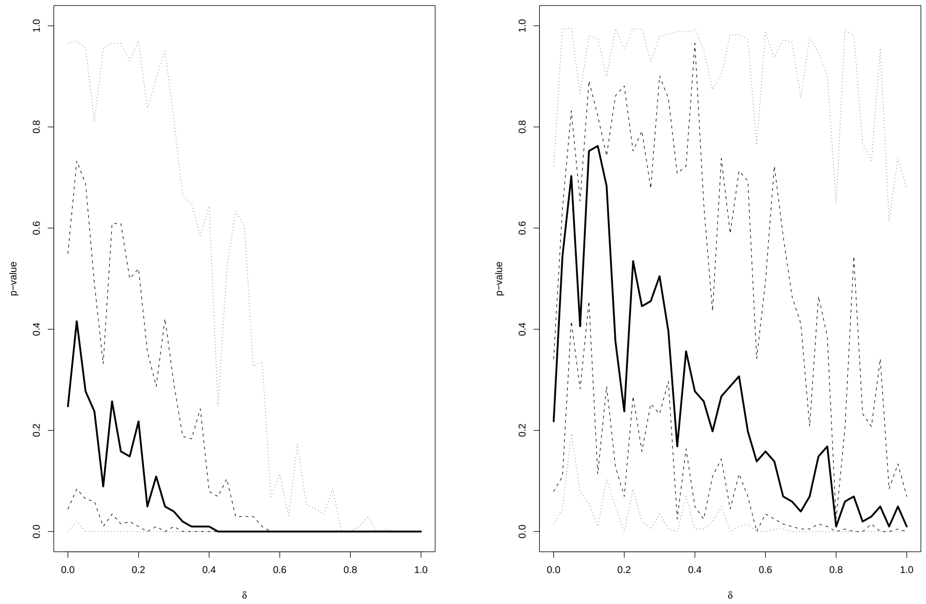
We estimated the empirical cumulative distribution functions for each group and applied the bootstrap procedure. Fig. 5.8 (a) displays a short descriptive of the p-values provided by the test. In particular, for each δ the quantiles $0.05, 0.25, 0.50, 0.75$

and 0.95 are displayed. The observed p-values are almost null for values of δ higher than 0.6. By visual inspection, the point patterns shown in Fig. 5.7 exhibit very slight differences in the relative locations of the events with respect to the boundary. However, these differences are clearly captured by the proposed descriptor and detected by the bootstrap procedure. Fig. 5.8 (b) shows the p-values corresponding to the h-function. The p-values are now almost null for values of δ higher than one. Note the better performance of the test based on the cumulative distribution function.


(a) (b)

(c) (d)

Figure 5.7: Several simulated point patterns. Realizations with (a) $\delta = 0.25$; (b) $\delta = 0.50$; (c) $\delta = 0.75$ and (d) $\delta = 1$.



(a) (b)

Figure 5.8: The p-values provided by the bootstrap test as a function of δ. Plot (a) displays the cumulative distribution function of the distances from the granules to the plasma membrane and plot (b) the respective p-values associated with the h-function. Solid line corresponds to the median, dashed lines represent the lower and upper quartiles and dotted lines correspond to the 0.05 and 0.95 quantiles.

5.6 Conclusions

In this study, an image processing method to segment granules in EM images has been proposed and evaluated. Although this software was used in this proof-of-principle study exclusively for chromaffin cells, the parameters should be easily adaptable to a wide range of cells with granules that undergo regulated secretion. Importantly, the performance of the automatic technique matched that of the manual analysis.

We have proposed several robust statistical methods to compare differences in the intracellular distribution of granules among cell groups with respect to: i) the number of granules per unit area of cytoplasm; and ii) the spatial distribution of the granules with respect to the plasma membrane. The empirical cumulative distribution function and the density of granules within a given distance to the plasma membrane are two good functional descriptors, which allow us to discriminate between cell groups. Moreover, for replicated spatial point patterns, we propose a bootstrap procedure for obtaining a p-value.

The analysis has been applied to two groups of adrenal gland cells from control and mutant mice. For the image set analyzed, there were no significant differences in terms of the number of granules per unit area of cytoplasm. However, there were differences in the distribution of the distance from the granules to the plasma membrane, i.e. mutant cells had granules closer to the plasma membrane than control cells.

Several aspects remain open for future research. First, to identify differences in granule parameters, such as size and grey-level density (related with the catecholamine content). Second, to study the joint distribution of the bivariate point patterns composed of stand-alone granules and granules with white vacuoles surrounding them. We note that the framework chosen is flexible and general enough to incorporate and directly test biological hypotheses in a direct way.

Finally, the proposed methodologies may be adapted to analyze other cellular organelles including endosomes, lysosomes, vacuoles, etc. To our knowledge, this is the first time that these tools are used to study secretion. In summary, we have presented and applied a powerful set of new tools for the detection, analysis and hypothesis testing which will facilitate future research of vesicle exocytosis.

Chapter 6

Conclusions and Future Work

In this thesis we have proposed different statistical methodologies in the context of stochastic spatial-temporal processes. Our major aims have been to characterize dynamic processes producing random shapes with random durations and to analyze the spatial-temporal dependencies between two processes.

We have proposed the *temporal Boolean model*, which formalizes the configuration of independent, randomly placed events with independent durations. Locations of these events are assumed to be the outcome of a realization of a spatial temporal Poisson Point Process. We have proposed some methods for parameter estimation. The methods have been applied to study the biological process of endocytosis.

First, in Chapter 3 we have shown that the use of a biased sample composed of isolated events, when spatial temporal overlapping exists, can lead to a very high bias in the estimation of the mean of the random duration. It was necessary to develop procedures based on stochastic models where the spatial temporal overlapping is explicitly assumed. We have shown that the temporal germ-grain models are a good choice and, in particular, the cylindrical temporal Boolean model.

We have proposed an estimator of the cumulative distribution function of the duration of spatially and temporally overlapping events based on the covariance of temporal random closed sets. Results from the simulation study showed that the absolute errors in the estimation of the distribution function of the duration are small. Indeed, the relative error in the estimation of the median of the distribution was smaller than 1% in almost all cases. The estimator also performs quite well in images with a certain degree of random noise such as salt-and-pepper.

Second, in Chapter 4 we have shown a new procedure for analyzing either spatial or spatial-temporal interactions between two types of events in a bivariate temporal random set. We have demonstrated its utility both as a summary statistic and as a basis for formal inference. We successfully applied this method to study associations between proteins in quantification of endocytic dynamics. The proposed methodology is not only a testing procedure, but it allows us to quantify the degree and the spatial-temporal interval of the interactions. Moreover, no (unnecessary) parametric assumptions are made. The proposed approach can be easily extended to multivariate random sets. From the application on simulated dependencies, we conclude that these summary descriptors accurately capture spatial-temporal interrelations. We have proposed a formal testing procedure. This represents an important departure from simple visual inspection of the images, which is the normal practice in cell biology literature, since manual analysis is a very time-consuming task and subject to observer error. In contrast, our proposed methodology is highly reproducible and fully automatic. The results obtained in the analysis of TIRFM image sequences of fluorescent-tagged proteins are in complete agreement with those reported by other techniques. In our opinion, these estimators are sensitive enough to be applied to many other applications. Such a procedure could be used as a screening tool to study protein interdependencies under different treatments and provides new opportunities for validating cellular models. Since we assumed stationarity in the images, we have proposed different solutions to overcome the violation of stationarity if needed.

Third, in Chapter 5 we have proposed several robust statistical methods to compare differences in the intracellular distribution of granules among cell groups with respect to: i) the number of granules per unit area of cytoplasm; and ii) the spatial distribution of the granules with respect to the plasma membrane. The empirical cumulative distribution function and the density of granules within a given distance to the plasma membrane are two good functional descriptors, which allow us to discriminate between cell groups. Moreover, for replicated spatial point patterns, we propose a bootstrap procedure for obtaining a p-value.

Also, an image processing method to segment granules in EM images has been developed and evaluated. The performance of the automatic technique matched that of the manual analysis. While this software was used in this proof-of-principle study exclusively for chromaffin cells, the parameters should be easily adaptable to a wide range of cells with granules that undergo regulated secretion.

There are several interesting extensions of the methodologies presented, that we propose in the following paragraphs.

Regarding Chapter 3, extensions for future research are the development of new temporal germ-grain models with other germ models. In particular, models with higher aggregation than the Poisson point process which will lead to greater overlapping probability. The estimation will be harder but, possibly, semi-parametric models could be proposed. Also, another aspect to be developed is the estimation of the size distribution for the grains. Last, the method described depends on the assumption of stationarity. Potential extensions will include the use of other models, such as the Cox model.

Regarding Chapter 4, some aspects require further investigation. First, the development of new parametric models for the spatial-temporal interactions where biological factors could be considered. Second, when stationarity does not hold, simple non-parametric measures of association may not be sufficient. The incorporation of covariates and possible sources of interactions of the events throughout the observation window would lead to non-homogeneous distributions, i.e. non-homogeneous random sets. As regards such modeling, potential extensions will include the use of other parametric models, such as a linked Cox model.

Regarding Chapter 5, several aspects remain open for future research. First, to identify differences in granule parameters, such as size and grey-level density (related with the catecholamine content). Second, to study the joint distribution of the bivariate point patterns composed of stand-alone granules and granules with white vacuoles surrounding them. We note that the framework chosen is flexible and general enough to incorporate and directly test biological hypotheses in a direct way. Finally, the proposed methodologies may be adapted to analyze other cellular organelles including endosomes, lysosomes, vacuoles, etc. To our knowledge, this is the first time that these tools are used to study secretion. In summary, we have presented and applied a powerful set of new tools for the detection, analysis and hypothesis testing of secretion which will greatly facilitate future research of vesicle exocytosis.

All the methodologies and methods described in this thesis have been programmed in Matlab and R software. A toolbox for Matlab have been developed to simulate Temporal Boolean Models (univariate and bivariate) and reproduce the proposed method of estimation. It is available in http://www.uv.es/tracs/index.html. More details can be found in Appendix A and in the supplementary material included.

To sum it up, we have defined and presented a methodology to estimate spatial temporal parameters of stochastic models and test procedures that allow the researcher test hypothesis in a formal way. The methodologies and models described herein could bring information of temporal distributions, intensities and size measures of spatial-temporal point processes. The performance of the proposed estimators have been tested in simulation studies. Also, these estimators have been successfully applied on real biological sequences to obtain information of the underlying processes which can not be obtained manually.

Appendix A

Software

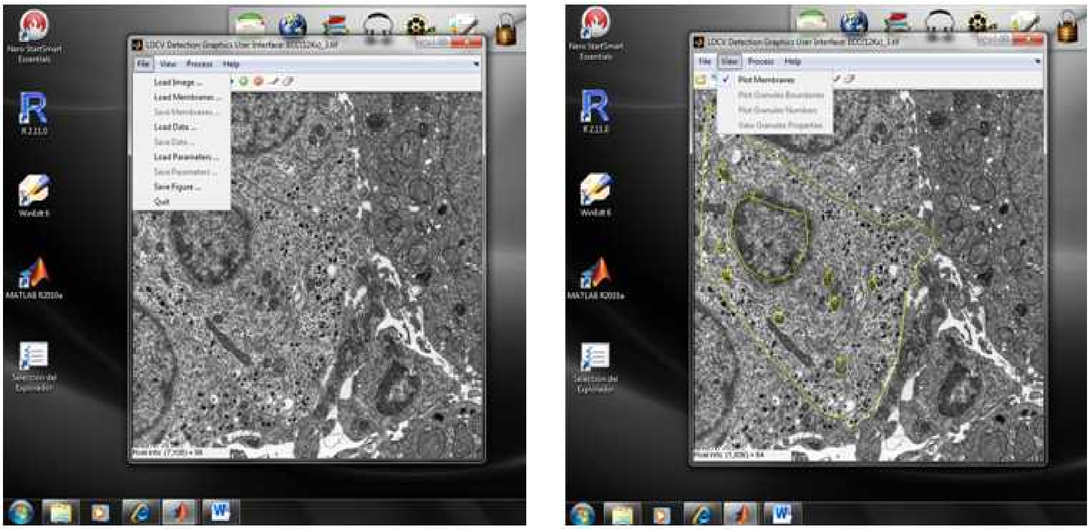
We have used the Matlab environment to perform the image analysis of the sequences. Specifically, we have used two libraries designed for image analysis named *Images* and *Mmorph*.

The statistical analysis was performed by using *R* software [63], an environment specifically designed for data analysis. That software can be freely downloaded from http://cran.r-project.org.

A.1 Image segmentation

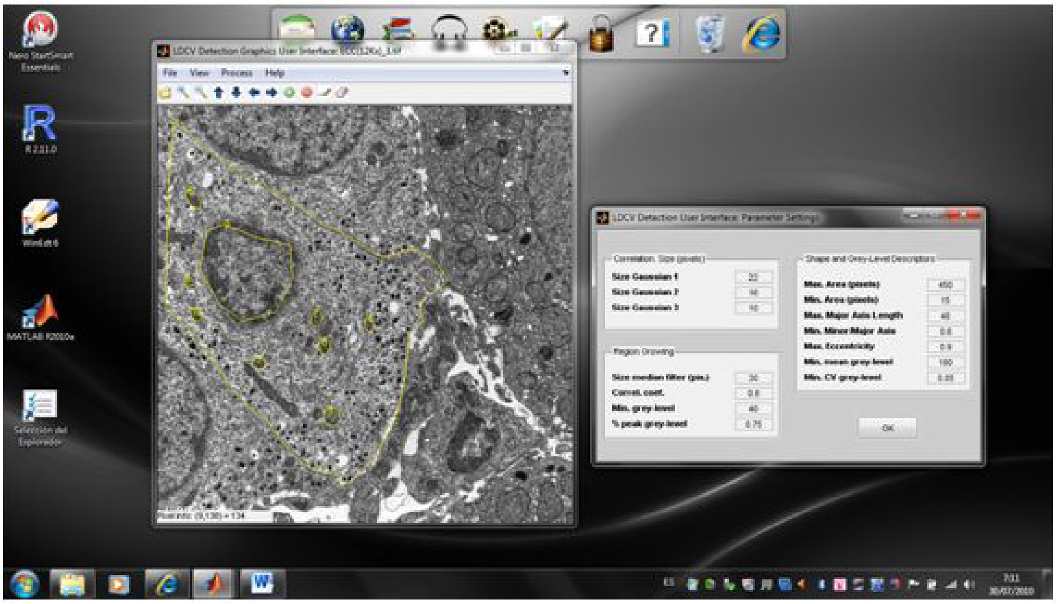
We have implemented a software tool in Matlab that allows the automatic detection of large dense core granules. We think that it is a very useful tool that facilitate the task of the biologist in the study of regulated exocytosis. The implemented toolbox can run under Windows and under the Linux operating system. It is available on request and it is offered under the terms of the GNU public license.

We illustrate the use with an example. Fig. A.1 (a) shows a cell and Fig. A.1 (b) the delineated membranes applied to the cell. Fig. A.2 shows the menu for setting the parameters of the detection algorithm. The obtained results are displayed in Fig. A.3. In Figs. A.4 (a) and (b), two zooms of the cell with the membrane granules delineated in blue are shown.



(a) (b)

Figure A.1: An electron micrograph of cell and the delineated membranes.



Correlation. Size (pixels)

Size Gaussian 1	22
Size Gaussian 2	16
Size Gaussian 3	10

Region Growing

Size median filter (pix.)	30
Correl. coef.	0.8
Min. grey-level	40
% peak grey-level	0.75

Shape and Grey-Level Descriptors

Max. Area (pixels)	450
Min. Area (pixels)	15
Max. Major Axis Length	40
Min. Minor/Major Axis	0.6
Max. Eccentricity	0.9
Min. mean grey-level	100
Min. CV grey-level	0.05

Figure A.2: Parameter setting menu.

The toolbox also allows the elimination of false positives and the manual incorporation of false negatives, by adding the location of a seed. The algorithm automatically calculates the selected data of the granule and its membrane.

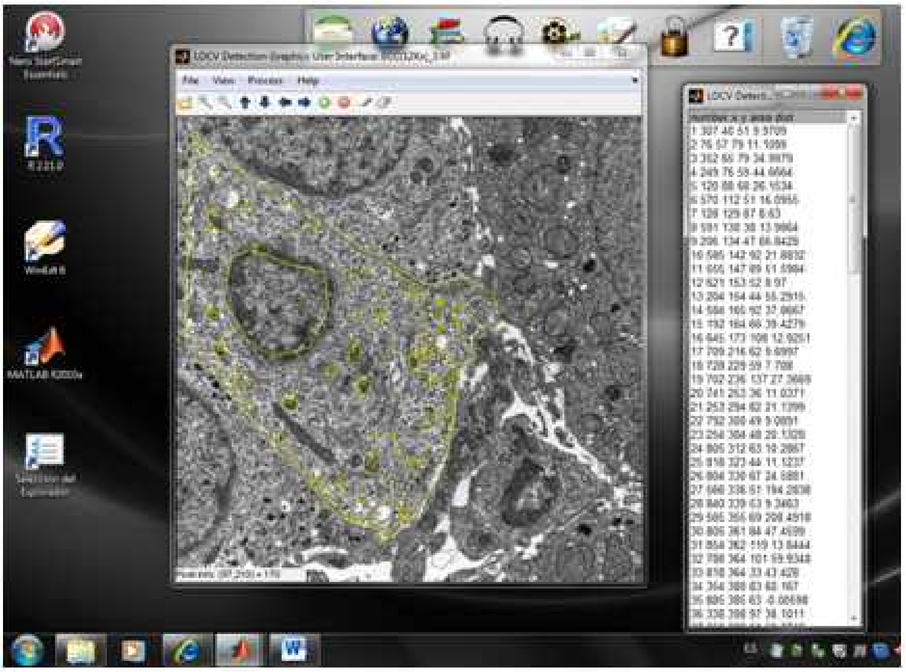
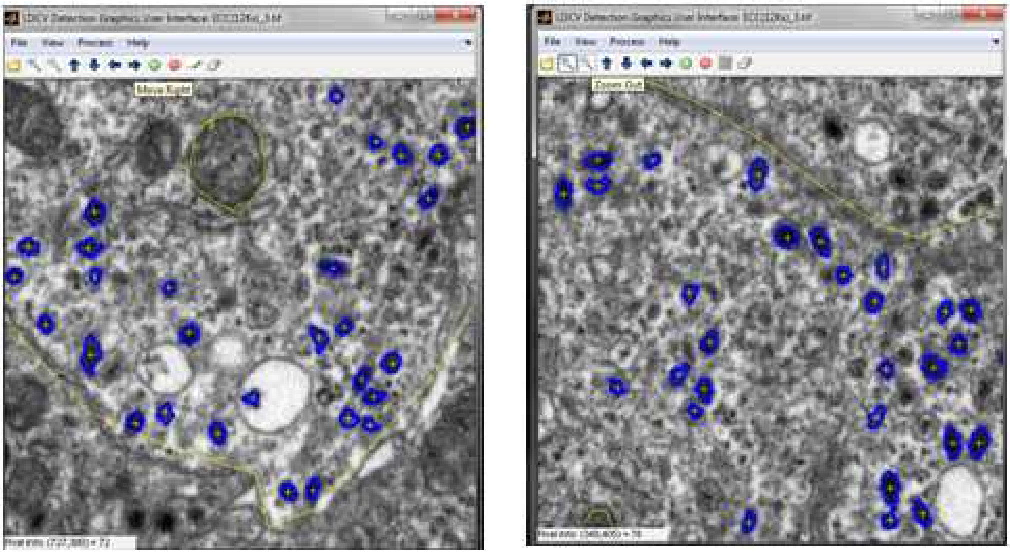


Figure A.3: The results obtained: granule location and data script.

A.2 Statistical analysis of granules

Statistical analysis to compare counts of granules has been performed using the function *glm* from *MASS* library of the software package *R*. An example of use is:

```
nullfit.APO  <- glm (Ngranules ~ offset (log (AreaCytoplasm)),
                family =poisson)
densfit.AGPO <- glm (Ngranules ~ factor(type) + offset (log(AreaCytoplasm)),
                family = poisson)
summary (nullfit.APO)
summary (densfit.AGPO)
anova.glm (nullfit.APO, densfit.AGPO,test="F")
nullfit.ANB  <- glm (Ngranules ~ offset (log (AreaCytoplasm)),
                family =negative.binomial(theta=1,link="log"))
densfit.AGNB <- glm (Ngranules ~ factor (type) +offset (log (AreaCytoplasm)),
                family =negative.binomial(theta=1,link="log"))
summary (nullfit.ANB)
summary (densfit.AGNB)
```



(a) (b)

Figure A.4: Two zooms of the granules with the granule membrane delineated.

The bootstrap procedures and the estimation of the cumulative distribution functions are developed in the software package *R*.

A.3 Functional Data Analysis

A functional datum is a set of discrete measured values $\{(s_j, y_j)\}_{j=1,\dots,n}$. First, it is necessary to convert these values to a function which is computable for any value. We did not use an interpolation process because the discretization might include some observational error. Instead, a smoothing technique was used to transform the raw data $\{(s_j, y_j)\}_{j=1,\dots,n}$ to a function $y(t) = \sum_{k=1}^{K} c_k \Phi_k(t)$, being $\{\Phi_k(t)\}_{k=1,\dots,K}$ a basis functions possessing a certain number of derivatives. A polynomial spline basis was chosen, where each $\Phi_k(t)$ is a piecewise cubic function. The coefficients c_k of the expression $y(t)$ were obtained by minimizing the least squares criterion $SMSSE(y|c) = \sum_{j=1}^{n}(y_j - \sum_{k=1}^{K} c_k \Phi_k(t_j))^2$. A detailed presentation of this method is given in [64].

To perform the functional data analysis, the R package *fda* developed by J.O. Ramsay has been used ([63, 65]). To compute the first derivative of $H(t)$, we used

15 basis functions to transform the raw data into a piece-wise function, which is appropriate for functions re-sampled at 65 values. These values should be modified depending on the kind of data and the sampling frequency.

A.4 Temporal Random Closed Sets (TRACS)

We have implemented a software tool in Matlab devoted to Temporal Random Closed Sets. In the study of bivariate Temporal Random Closed Sets, it allow us to estimate the cross-covariance function, perform testing procedures (based on toroidal shifts and Monte Carlo tests), and simulate bivariate temporal random closed sets. In the Cylindrical TBM work, we used the tool to simulate temporal boolean models, and estimate the spatial-temporal covariance and the cumulative distribution function.

Matlab toolbox (source code and user documentation) for the simulation of temporal random sets, parameter estimation and the testing procedures is available at *http : //www.uv.es/tracs/index.html*.

The simulated videos and the fluorescent-tagged protein image sequences can be downloaded from *http : //www.uv.es/tracs/videos/videos.html*, and they are also included in the supplementary material.

Now, we comment briefly some of the most important functions of the toolbox.

A.4.1 Temporal Boolean Models

Simulation

Regarding simulation of Temporal Boolean Models, the most important functions are:

- Function **tbmsimulation**: This function generates a realization of temporal Boolean model. Possible probability density functions of the grain sizes are: uniform and Gaussian (truncated Gaussian). Possible probability density functions of the grain durations are: uniform, exponential, Gamma and Gaussian (truncated Gaussian).

- Function **tbmgerms**: This function generates the germs, a spatial-temporal Poisson point process.

- Function **tbmgrains**: This function generates the sizes and orientations of the grains. Possible probability density functions of the grain sizes are: uniform and Gaussian (truncated Gaussian).

- Function **tbmdurations**: This function generates the grains durations. Possible probability density functions of the grain durations are: uniform, exponential, Gamma and Gaussian (truncated Gaussian).

Estimation

Regarding parameter estimation of Temporal Boolean Models, the most important functions are:

- Function **tbmcontminT**: Estimation of the parameters of a two-dimensional Boolean model by applying the Minimun Contrast Method using the capacity functional T.

- Function **tbmcontminH**: Estimation of the parameters of a two-dimensional Boolean model by applying the Minimun Contrast Method using the contact distribution function H.

- Function **tbmblockcorr**: Estimation of the mean and confidence interval using the batch-mean method.

- Function **tbmalphaestimation**: Estimation of the α-function of a Temporal Boolean Model (the mean number of germs that died/were born in a time interval).

- Function **tbmtemporalestimation**: Estimation of the temporal parameters of a Temporal Boolean model.

- Function **tracsvolumefraction**: Estimation of the spatial-temporal volume fraction of a realization of a Temporal Random Closed Sets (TRACS).

- Function **tracsisolatedclumps**: This function segments the isolated circular clumps, i.e. clumps whose boundary is not covered by any other clump/grain.

Function tbmcontminT

Estimation of the parameters of a two-dimensional Boolean model by applying the Minimun Contrast Method using the capacity functional T. It provides a four-component vector, being *gamma*: number of germs per unit area; *perimeter*: mean perimeter of the primary grain; *area*: mean area of the primary grain and $volfrac$: the volume fraction.

W is assumed to have no holes. The capacity functional $T_K(t)$ takes values in the interval $[0, 1]$ since it represents the cumulative distribution function and, as a consequence, the empirical values of $log(1 - T_K(t))$ tend to infinite. Hence, only values of $1 - T_K(t)$ in the interval $[0, 0.8]$ are used in the fitting of the second order polynomial derived from the Steiner formula. The use of a square structuring element is recommended because its digital approximation is more accurate than the corresponding approximation of the disk.

Function tbmcontminH

Estimation of the parameters of a two-dimensional Boolean model by applying the Minimun Contrast Method using the contact distribution function H.

This function returns the parameters of the two-dimensional Boolean model: a three-component vector, being *gamma*: number of germs per unit area; *perimeter*: mean perimeter of the primary grain and $volfrac$ the volume fraction.

W is assumed to have no holes. Only values of $1 - H_K(t)$ in the interval $[0, 0.7]$ are used in the fitting of the second order polynomial derived from the Steiner formula.

Function tbmblockcorr

Estimation of the mean and confidence interval using the batch-mean method. This function depends on the `autocorr` function of the Garch Matlab Toolbox.

Function tbmalphaestimation

Estimation of the α-function of a Temporal Boolean Model (the mean number of germs that died/were born in a time interval).

Function tbmtemporalestimation

Estimation of the temporal parameters of a Temporal Boolean model. It returns a structure with the estimation of the λ, $\alpha(s)$, $p(s)$, $\beta(s)$, the probability density function of the durations and the time interval. The value of λ (mean number of germs per unit area and time) is estimated by fitting a first order polynomial at $\alpha(0)$ and calculating its slope using the first four points, see [9].

Function tracsvolumefraction

Estimation of the spatial-temporal volume fraction of a realization of a Temporal Random Closed Sets (TRACS).

Function tracsisolatedclumps

This function segments the isolated circular clumps, i.e. clumps which boundary is not covered by any other clump/grain. The function returns a structure with two fields. DATA.POINTS is a structure with the information of each isolated clump: X,Y is the location of the center, RAD is the radius of the disk, LIFE is a vector with the frames in which the grain is alive and N is the grain duration (in frames). DATA.MOV is a movie in which the isolated grains segmented are represented in light grey.

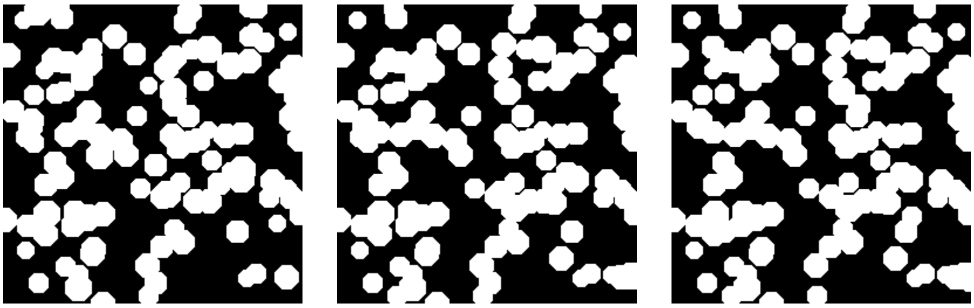
Visualization

We can plot a Temporal Random Closed Set using:

- Function **tracs2dplot**: This functions plots a frame (section) of a Temporal Random Set.
- Function **tracs3dplot**: This functions plots a three-dimensional reconstruction of a Temporal Random Set.

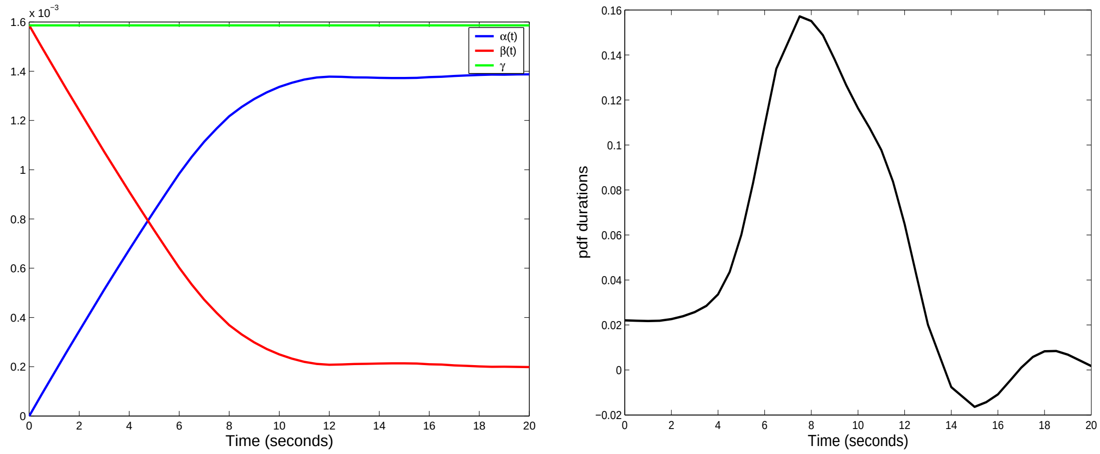
A demo

An example of simulation and estimation of the parameters of a TBM.



(a) (b) (c)

Figure A.5: Several consecutive frames of a simulated TBM with cylindrical grains.



(a) (b)

Figure A.6: Parameter estimation: (a) Estimation of $\alpha(t)$, $\beta(t)$ and γ. (b) Estimated density of the duration.

A.4.2 Bivariate TRACS

Regarding analysis of bivariate TRACS, the most important functions are:

- Function **tbmbivariatesimulation**: Generation of a bivariate temporal Boolean model. Possible probability density functions of the grain sizes are: uniform and Gaussian (truncated Gaussian). Possible probability density functions of the grain durations are: uniform, exponential, Gamma and Gaussian (truncated Gaussian). Possible models of interaction for the germs are: linked-pairs and Poisson cluster (Neymann-Scott process).

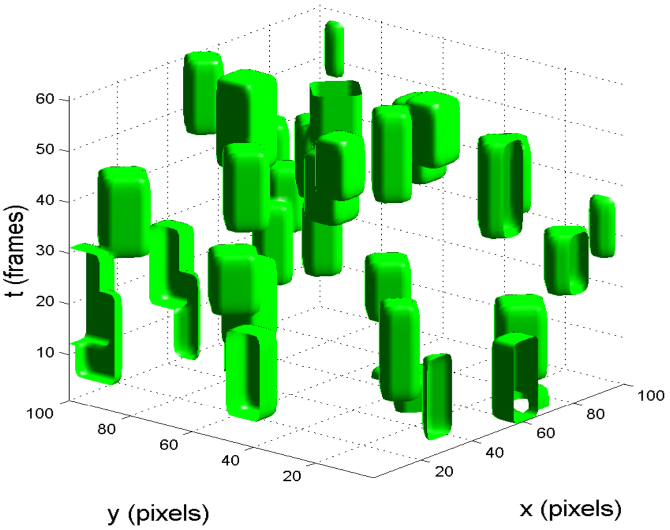


Figure A.7: A reconstruction of a TBM using function tracs3dplot.

- Function **tracsrandomization**: Toroidal shift randomization of a TRACS.

- Function **tracscrossC**: Cross-covariance of a TRACS. This function returns the cross-covariance C within disks of radii S and time interval T (in frames) for movies MOV1, MOV2. C is integrated over the angles $[-\pi : \pi]$.

- Function **tracscrossCh0**: Cross-covariance of a TRACS. This function returns the cross-covariance C given a vector displacement H for two movies, i.e., the value of the cross-covariance at $(H, 0)$. C is integrated over the angles $[-\pi : \pi]$. The estimation method depends on the type of dependencies: spatial or spatial-temporal. We assume there is only spatial dependencies. We assume both sequences have the same length in frames and size in pixels.

- Function **tracscrossCht**: Cross-covariance of a TRACS. This function returns the cross-covariance C given a vector displacement H and time T (frames) for two movies, i.e., the value of the cross-covariance at (H, T). We assume that both sequences have the same length in frames and size in pixels.

- Function **tracscrossK1**: The K1-function within disks of radii given in S.

- Function **tracscrossK**: The K-function within disks of radii given in S (pixels) and time interval given in T (frames).
- Function **tracspvalueMC**: This function provides the p-value of the Monte Carlo test for descriptor F.
- Function **tracspvaluemapMC**: This function calculates and plots the p-values of the Monte Carlo test by applying a bootstrap procedure to the cross-covariance.
- Function **tracsbi3dplot**: Plot a three-dimensional reconstruction of a bivariate TRACS.

Demos

The scripts ScriptSpatial.m and ScriptSpatialTemporal.m show how to calculate the covariance function, the K-function, L-function and the pair correlation functions.

Calculating the p-value map

An example of p-value map is shown in A.8.

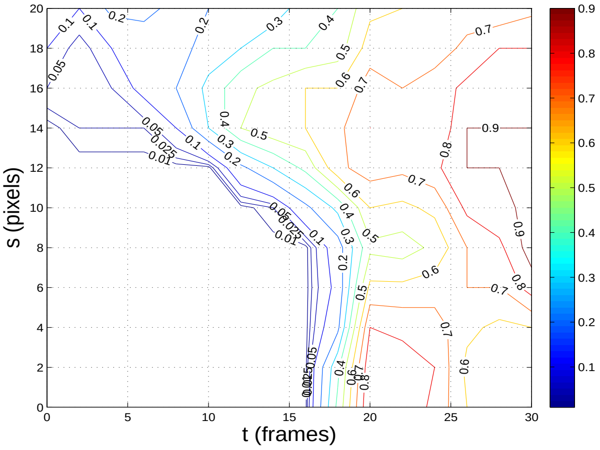


Figure A.8: An example of p-value map.

Appendix B

Supplementary Material

Supplementary videos have been generated in order to visually show the kind of data we are dealing with. In particular, Chapters 3 and 4 which study spatial temporal data refer to these image sequences.

In Chapter 3, the Video1, Video2 and Video3 correspond with simulated models where discs are randomly located in space and time with random radii and random durations. Video1 corresponds to the uniform distribution with $\lambda = 0.0001$, Video2 to the Gamma distribution with $\lambda = 0.0002$ and Video3 to the Gaussian distribution with $\lambda = 0.0003$ with a mean duration of 6 seconds. Videos Noise1, Noise2 and Noise3 correspond to noisy images, with salt-and-pepper noise in the random shapes, independently in each frame, and with ratios of noise 0.5%, 1%, 2% respectively. Videos Cell1, Cell2 and Cell3 corresponds to the biological data and comprise three sequences of clathrin protein.

In Chapter 4, Video 1 in the supplementary material corresponds to a simulated bivariate temporal random set in which spatial dependencies were generated, Video 2 to a one with spatial-temporal dependencies and Video 3 to complete independence. Video 4 corresponds to pair 1, Clathrin-RFP is in red and Hip1R-GFP in green. Video 5 corresponds to pair 3, Clathrin-RFP in red and Caveolin-GFP in green.

The source code of the functions of the Temporal Boolean model toolbox and the automatic granule detection tool is also included in the supplementary material.

Bibliography

[1] A. Agresti. *Categorical Data Analysis.* Wiley, 2nd edition, 2002.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Science, New York and London, fourth edition, 2002.

[3] D. Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Traffic*, 2:764–774, 2001.

[4] G. Ayala. *Inferencia en Modelos Booleanos.* Tesis Doctoral, Universidad de Valencia, 1988.

[5] G. Ayala, J. Ferrandiz, and F. Montes. Two methods of estimation in Boolean models. *Acta Stereologica*, 8(2):629–634, 1989.

[6] G. Ayala, J. Ferrandiz, and F. Montes. Boolean models: maximum likelihood estimation from circular clumps. *Biometrical Journal*, 32:73–78, 1990.

[7] G. Ayala, J. Ferrandiz, and F. Montes. Sparse sampling and maximum likelihood estimation for Boolean models. *Biometrical Journal*, 33(2):237–245, 1991.

[8] G. Ayala, J. Ferrandiz, and F. Montes. On parametric estimation in Boolean models. *Rassegna di Metodi Statistici ed Applicazioni*, 8:1–17, 1993.

[9] G. Ayala, R. Sebastián, M. Díaz, E. Díaz, R. Zoncu, and D. Toomre. Analysis of spatially and temporally overlapping events with application to image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1707–1712, 2006.

[10] T. Brett and L. Traub. Molecular structures of coat and coat-associated proteins: function follows form. *Currrent Opinion in Cell Biology*, 18(4):395–406, 2006.

[11] F. Brodsky, C. Chen, C. Knuehl, M. Towler, , and D. Wakeham. Biological basket weaving: formation and function of clathrin-coated vesicles. *Annual Review of Cell and Developmental Biology*, 17:517–568, 2001.

[12] R. Burgoyne and A. Morgan. Calcium sensors in regulated exocytosis. *Cell Calcium*, 24(5-6):367–76, 1998.

[13] R. Burgoyne and A. Morgan. Secretory granule exocytosis. *Physiological Reviews*, 83:581–632, 2002.

[14] K. Castleman. *Digital Image Processing.* Prentice Hall, 1996.

[15] H. Chen, S. Fre, V. Slepnev, M. Capua, K. Takei, M. Butler, P. Di Fiore, and P. De Camilli. Epsin is an EH-domain-binding protein implicated in clathrin-mediated endocytosis. *Nature*, 394(6695):793–797, 1998.

[16] S. Corner and S. Schmid. Regulated portals of entry into the cell. *Nature*, 4:37–44, 2003.

[17] N. Cressie. *Statistics for Spatial Data. Revised Edition.* John Wiley and Sons, New York, 1993.

[18] D. Dernick, G. Alvarez de Toledo, and M. Lindau. Exocytosis of single chromaffin granules in cell-free inside-out membrane patches. *Nature Cell Biology*, 5:358–362, 2003.

[19] E. Díaz, G. Ayala, M. Díaz, L. Gong, and D. Toomre. Automatic detection of dense-core vesicles in secretory cells and statistical analysis of their intracellular distribution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):2–11, 2010.

[20] E. Díaz, R. Sebastián, G. Ayala, M. Díaz, R. Zoncu, D. Toomre, and S. Gasman. Measuring spatio-temporal dependencies in bivariate temporal random sets with applications to cell biology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1659–1671, 2008.

[21] M. Díaz, G. Ayala, and E. Díaz. Estimating the duration of overlapping events from image sequences using cylindrical temporal boolean models. *Journal of Mathematical Imaging Vision*, 38(2):83–94, 2010.

[22] M. Díaz, G. Ayala, T. León, R. Zoncu, and D. Toomre. Analyzing protein-protein spatial-temporal dependencies from image sequences using fuzzy temporal random sets. *Journal of Computational Biology*, 15:1221 1236, 2008.

[23] P. Diggle. Binary mosaics and the spatial pattern of heather. *Biometrics*, 37:531–539, 1981.

[24] P. Diggle. *Statistical Analysis of Spatial Point Patterns.* Arnold, London, second edition, 2003.

[25] P. Diggle, N. Lange, and F. Benes. Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association*, 86:618–625, 1991.

[26] E. Dougherty and A. Grigoryan. Automatic counting of illuminated spheres in a random Boolean model. In *Image Processing: Algorithms and Systems*, volume 4667, pages 170–180. The International Society for Optical Engineering, 2003.

[27] O. Dousse, P. Mannersalo, and P. Thiran. Latency of wireless sensor networks with uncoordinated power saving mechanisms. *MobiHoc*, 2004.

[28] V. Dupac. Parameter estimation in the poisson field of discs. *Biometrika*, 67:187–190, 1980.

[29] M. Edeling, C. Smith, and D. Owen. Life of a clathrin coat: insights from clathrin and AP structures. *Nature Reviews. Molecular Cell Biology*, 7(1):32–44, 2006.

[30] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall/CRC, 1993.

[31] M. Ehrlich, W. Boll, A. Van Oijen, K. Hariharan, R. Chandran, M. Nibert, and T. Kirchhausen. Endocytosis by random initiation and stabilization of clathrin-coated pits. *Cell*, 118:591–605, 2004.

[32] A. Engqvist-Goldstein, R. Warren, M. Kessels, J. Keen, J. Heuser, and D. Drubin. The actin-binding protein Hip1R associates with clathrin during early stages of endocytosis and promotes clathrin assembly in vitro. *The Journal of Cell Biology*, 154(6):1209–1223, 2001.

[33] I. Epifanio and G. Ayala. A random set view of texture classification. *IEEE Transactions on Image Processing*, 11(8):859 –867, August 2002.

[34] M. Ford, I. Mills, B. Peter, Y. Vallis, G. Praefcke, P. Evans, and H. McMahon. Curvature of clathrin-coated pits driven by epsin. *Nature*, 419(6905):361–366, 2002.

[35] R. Foxall and A. Baddeley. Nonparametric measures of association between a spatial point process and a random set, with geological applications. *Applied Statistics*, 51, part 2:165–182, 2002.

[36] M. Gallego, M. Ibañez, and A. Simó. Non-homogeneous temporal Boolean models to study endocytosis. *To be published*, 2010.

[37] M. Gallego and A. Simó. Random closed set models: estimating and simulating binary images. *Image Analysis & Stereology*, 22:72–91, 2003.

[38] P. García. *Texture analysis using the one dimensional Boolean model*. PhD thesis, University Jaume I, Castellón, Spain, October 1999.

[39] P. García, M. Petrou, and S. Kamata. The use of Boolean model for texture analysis of grey images. *Computer Vision and Image Understanding*, 74(3):227–235, June 1999.

[40] L. Gong, G. Di Paolo, E. Díaz, G. Cestra, M. Díaz, M. Lindau, P. De Camilli, and D. Toomre. Phosphatidylinositol phosphate kinase type Iγ regulates dynamics of large dense-core vesicle fusion. *Proceedings of the National Academy of Sciences*, 102(14):5204–5209, 2005.

[41] L. Gong, I. Hafez, G. Alvarez de Toledo, and M. Lindau. Secretory vesicles membrane area is regulated in tandem with quantal size in chromaffin cells. *The Journal of Neuroscience*, 23(21):7917–7921, 2003.

[42] J. Goutsias, R. Mahler, and H. Nguyen, editors. *Random Sets. Theory and Applications*, volume 97 of *The IMA Volumes in Mathematics and its Applications.* Springer, 1997.

[43] L. Gutierrez, S. Viniegra, J. Rueda, A. Ferrer-Montiel, J. Canaves, and M. Montal. A peptide that mimics the C-terminal sequence of SNAP-25 inhibits secretory vesicle docking in chromaffin cells. *The Journal of Biological Chemistry*, 272:2634–2639, 1997.

[44] M. Kaksonen, C. Toret, and D. Drubin. A modullar design for the clathrin- and actin-mediated endocytosis machinery. *Cell*, 123:305–320, 2005.

[45] P. Keller, D. Toomre, E. Díaz, J. White, and K. Simons. Multicolor imaging of post-golgi sorting and traficking in live cells. *Nature Cell Biology*, 3:140–149, 2001.

[46] L. Koval, E. Yavorskaya, and E. Lukyanetz. Ultrastructural features of medullary chromaffin cell cultures. *Neuroscience*, 96:639–649, 2000.

[47] L. Koval, E. Yavorskaya, and E. Lukyanetz. Electron microscopic evidence for multiple types of secretory vesicles in bovine chromaffin cells. *General and Comparative Endocrinology*, 121:261–277, 2001.

[48] A. Law and W. Kelton. *Simulation Modeling and Analysis.* McGraw Hill, third edition, 2000.

[49] M. Lieshout and E. Zwet. Maximum likelihood estimation for the bombing model. Probability, Networks and Algorithms PNA-R0008, Stichting Mathematisch Centrum, 1090 GB, Amsterdam, August 2000.

[50] B. Loo, B. Parvin, and S. Rothman. Two-and three-dimensional segmentation for measurement of particles in the analysis of microscopic digital images of biological samples. In *Procedings of the SPIE*, volume 2655, pages 209–215. The International Society for Optical Engineering, 1996.

[51] G. Matheron. *Eléments pour une théorie des milieux poreux.* Masson, Paris, 1967.

[52] G. Matheron. *Random sets and Integral Geometry.* Wiley, London, 1975.

[53] C. Merrifield, M. Feldman, L. Wan, and W. Almers. Imaging actin and dynamin recruitment during invagination of single clathrin-coated pits. *Nature Cell Biology*, 4:691–698, 2002.

[54] C. Merrifield, D. Perrais, and D. Zenisek. Coupling between clathrin-coated-pit invagination, cortactin recruitment, and membrane scission observed in live cells. *Cell*, 121(4):593–606, 2005.

[55] M. Metzler, V. Legendre-Guillemin, L. Gan, V. Chopra, A. Kwok, P. McPherson, and M. Hayden. HIP1 functions in clathrin-mediated endocytosis through binding to clathrin and adaptor protein 2. *The Journal of Biological Chemistry*, 276(42):39271–39276, 2001.

[56] I. Molchanov. *Statistics of the Boolean Model for Practitioners and Mathematicians.* John Wiley and Sons, Chichester, 1997.

[57] I. Molchanov. *Theory of Random Sets.* (Probability and its Applications). Springer, 2005.

[58] D. Nott and T. Ryden. Pairwise likelihood methods for inference in image models. *Biometrika*, 86:661–676, 1999.

[59] G. Palade. Intracellular aspects of the process of protein secretion. *Science*, 189:347–1358, 1975.

[60] R. Perera, R. Zoncu, L. Lucast, P. De Camilli, and D. Toomre. Two synaptojanin 1 isoforms are recruited to clathrin-coated pits at different stages. *Proceedings of the National Academy of Sciences*, 103:19332–19337, 2006.

[61] H. Plattner, A. Artalejo, and E. Neher. Ultrastructural organization of bovine chromaffin cell cortex-analysis by cryofixation and morphometry of aspects pertinent to exocytosis. *The Journal of Cell Biology*, 139:1709–1717, 1997.

[62] E. Pothos, E. Mosharov, L. Kuo-Peing, W. Setlik, M. Haburcak, G. Baldini, M. Gershon, H. Tamir, and D. Sulzer. Stimulation-dependent regulation of the pH, volume and quantal size of bovine and rodent secretory vesicles. *Journal of Physiology*, 542.2:453–476, 2002.

[63] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010.

[64] J. Ramsay and B. Silverman. *Functional Data Analysis.* Springer Series in Statistics. Springer, first edition, 1997.

[65] J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2010. R package version 2.2.3.

[66] S. Reader. Using survival analysis to study spatial point patterns in geographical epidemiology. *Social Science & Medicine*, 50:985–1000, 2000.

[67] B. Ripley. *Statistical Inference for Spatial Point Processes.* Cambridge University Press, 1988.

[68] K. Schladitz, S. Peters, D. Reinel-Bitzer, A. Wiegmann, and J. Ohser. Design of acoustic trim based on geometric modelling and flow simulation for non-woven. Technical Report 72, Fraunhofer ITWM, February 2005.

[69] R. Sebastián. *Image Sequence Analysis Based on Spatial Temporal Stochastic Models with Application to Cell Biology.* Tesis doctoral. Universidad de Valencia, 2006.

[70] R. Sebastián, E. Díaz, G. Ayala, M. Díaz, R. Zoncu, and D. Toomre. Studying endocytosis in space and time by means of temporal Boolean models. *IEEE Transactions on Pattern Recognition*, 39:2175 2185, 2006.

[71] R. Sebastián, M. Díaz, G. Ayala, K. Letinic, J. Moncho-Bogani, and D. Toomre. Spatio-temporal analysis of constitutive exocytosis in epithelial cells. *IEEE Transactions on Computational Biology and Bioinformatics*, 3(1):17–32, 2006.

[72] J. Serra. *Image Analysis and Mathematical Morphology.* Academic Press, 1982.

[73] S. Server, H. Damke, and S. Schmid. Garrotes, springs, ratchets, and whips: putting dynamin models to the test. *Traffic*, 1(5):385–392, May 2000.

[74] A. Sim, M. Baldwin, J. Rostas, J. Holst, and R. Ludowyke. The role of serine/threonine protein phosphatases in exocytosis. *Biochemical Journal*, 373:641–659, 2003.

[75] V. Slepnev and P. De Camilli. Accessory factors in clathrin-dependent synaptic vesicle endocytosis. *Nature Reviews Neuroscience*, 1(3):161–172, 2000.

[76] P. Soille. *Morphological Image Analysis. Principles and Applications.* Springer-Verlag, second edition, 2003.

[77] M. Sonka and J. Fitzpatrick. *Medical Image Proccessing and Analysis. Vol. 2.* Spie Press, 2000.

[78] J. Sorensen, G. Nagy, F. Varoqueaux, N. Nehring, M. Wilson, and E. Neher. Differential control of the releasable vesicle pools by SNAP-25 splice variants and SNAP-23. *Cell*, 114:75–86, 2003.

[79] D. Stoyan, W. Kendall, and J. Mecke. *Stochastic Geometry and its Applications.* Wiley, Berlin, second edition, 1995.

[80] D. Stoyan and H. Stoyan. *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics.* Wiley, 1994.

[81] T. Sudhof. Synaptotagmins: why so many? *The Journal of Biological Chemistry*, 277:7629–7632, 2001.

[82] D. Toomre, P. Keller, J. White, J. Olivo, and K. Simons. Dual-color visualization of trans-golgi network to plasma membrane traffic alone microtubules in living cells. *Journal of Cell Science*, 112:21–33, 1999.

[83] D. Toomre and D. Manstein. Lighting up the cell surface with evanescent wave microscopy. *Trends Cell Biology*, 11:298–303, 2001.

[84] D. Toomre, J. Steyer, P. Keller, W. Almers, and K. Simons. Fusion of constitutive membrane traffic with the cell surface observed by evanescent wave microscopy. *The Journal of Cell Biology*, 149:33–40, 2000.

[85] S. Tooze, G. Martens, and W. Huttner. Secretory granule biogenesis: rafting to the snare. *Trends Cell Biology*, 11(3):116–122, 2001.

[86] J. van den Berg, R. Meester, and D. White. Dynamic Boolean models. *Stochastic Processes and their Applications*, 69:247–257, 1997.

[87] M. van Lieshout and A. Baddeley. Indices of dependence between types in multivariate point patterns. *Scandinavian Journal of Statistics*, 26:511–532, 1999.

[88] T. Voets, T. Moser, P. Lund, R. Chow, M. Geppert, T. Sudhof, and E. Neher. Intracellular calcium dependence of large dense-core vesice exocytosis in the absence of synaptotagmin I. *Proceedings of the National Academy of Sciences*, 98:11680–11685, 2001.

[89] P. Yao, I. Bushlin, and R. Petralia. Partially overlapping distribution of epsin1 and HIP1 at the synapse: analysis by immunoelectron microscopy. *The Journal of Comparative Neurology*, 494(2):368–379, 2006.

[90] O. Yizhar, U. Matti, R. Melamed, Y. Hagalili, D. Bruns, J. Rettig, and U. Ashery. Tomosyn inhibits priming of large dense-core vesicles in a calcium-dependent manner. *Proceedings of the National Academy of Sciences*, 101:2578–2583, 2004.

[91] R. Zoncu, R. Perera, R. Sebastian, F. Nakatsu, H. Chen, T. Balla, G. Ayala, D. Toomre, and P. De Camilli. Loss of endocytic clathrin-coated pits upon acute depletion of phosphatidylinositol 4,5-bisphosphate. *Proceedings of the National Academy of Sciences*, 104(10):3793–8, 2007.