La Dra. MARIA ELENA DIA FERNANDEZ, Profesor funcionario y miembro del Departament d'Informàtica de l' Escola Tècnica Superior d'Enginyeria, y el Dr. GUILLERMO AYALA GALLEGO, Profesor funcionario y miembro del Departament d'Estadística i Investigació Operativa de la Facultat de Matemàtiques, ambos de la Universitat de València,

CERTIFICAN: Que la tesis doctoral que presenta Don RAFAEL SEBASTIAN AGUILAR, titulada, "IMAGE SEQUENCE ANALYSIS BASED ON SPATIAL TEMPORAL STOCHASTIC MODELS WITH APPLICATION TO CELL BIOLOGY", PARA OPTAR AL GRADO DE Doctor per la Universitat de València, ha sido realizada bajo su dirección.

Y para que así conste a efectos legales, firman el presente certificado:

Fdo: Maria Elena Diaz Fernandez Fdo: Guillermo Ayala Gallego

Burjassot, Noviembre de 2006

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF VALENCIA

Image Sequence Analysis Based on Spatial Temporal Stochastic Models with Application to Cell Biology

Rafael Sebastian

Advisors
Dr. Elena Diaz
Dr. Guillermo Ayala

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

2006

To my beloved parents and my sister
To my girl Ana

Acknowledgment

Special Agreements to Dr. Maria Elena Diaz from Department of Computer Science of the University of Valencia for her support, encouragement and guidance throughout this PhD, without whose this thesis would have been impossible.

I would like to thank Dr. Guillermo Ayala from Department of Statistics of the University of Valencia who encouraged and engaged me in discussions that resulted in many of the ideas introduced in this thesis.

Finally, I would like to thank to Dr. Derek Toomre and Roberto Zoncu from the Department of Cell Biology of the Yale University since this work would not have been possible without their contributions and their meticulous experiments.

This work has been supported by the Human Frontier Science Organization (HFSP) RGY40/2003, entitled 'Novel Imaging Modelling and Analysis of Exo-Endocytosis Coupling in Neuroendocrine Cells'

Publications Resulted From This Work

Journal Papers

R. Sebastian, M.E. Diaz, G. Ayala, K. Letinic, J. Moncho-Bogani and D. Toomre. *Spatio-temporal analysis of constitutive exocytosis in epithelial cells*, IEEE/ACM Transactions on Computational Biology and Bioinformatics. Vol. 3, Issue 1. January-March 2006.*

G. Ayala, **R. Sebastian**, M.E. Diaz, E. Diaz, 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(10):1707-1712. October 2006* .

R. Sebastian, E. Diaz, G. Ayala, M.E. Diaz, R. Zoncu and D. Toomre. *Studying endocytosis in space and time by means of temporal Boolean models.* Pattern Recognition. 39(11):2175-2185 (November 2006).*

Conference Proceedings

R. Sebastian, E. Diaz, G. Ayala, R. Zoncu, D. Toomre. *Analysis of spatial dependencies of endocytic proteins by using temporal random closed sets.* The 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, New York, USA, 2006.

Posters

R. Zoncu **R. Sebastian,** M.E. Diaz, P. de Camilli, D. Toomre, *Spatial and temporal relationships between ligand-receptor complex formation and clathrin-mediated receptor internalization.* 44th Annual meeting of the American Society for Biology. Washington, DC (USA), 2004

K. Letinic, **R. Sebastian**, G. Ayala, R. Perera, M.E. Diaz, D. Toomre. *Anaylsis of Spatial-Temporal Coordination of Exocytosis in Migrating Astrocytes.* The 58th Annual Meeting of the Japan Society for Cell Biology. Saitama (Japan), 2005.

* Journal Manuscripts reproduced in appendix D

Abstract

In this thesis new modelling methodologies to analyze dynamic processes from image sequences are presented. The problem of analyzing stochastic processes from image sequences in which spatially and temporally overlapping objects could be observed is fundamental to study real phenomenons and it is not completely fulfilled with current approaches.

The original motivation was to develop models to study important stochastic mechanisms observed in cells, the endocytosis and the exocytosis, from image sequences obtained by means of Total Internal Reflection Fluorescence Microscopy, which allows selective illumination of the cellular plasma membrane. In our approach these images are considered the result of a stochastic process that meets the conditions of stationarity and isotropy.

Firstly, we propose the application of a methodology in the context of Point Processes to study the existence of spatial, temporal or spatio-temporal clustered patterns on patterns based on the Ripley $\mathcal{K}$-function to study fusion sites and time in constitutive exocytosis. This kind of methodology has never been used before for the analysis of exocytosis, although it can bring new information to better understanding the underlaying machinery of the cell.

Secondly, we propose an extension of the classical Boolean model to study spatial temporal stochastic processes with special emphasis on the temporal dimension. The *spatio-temporal Boolean model* is defined to model spatial temporal overlapped patterns. Estimators to obtain the mean number of events per unit area and unit time, mean size and density function of durations of the events from image sequences are proposed. Two different approaches to obtain these estimates are presented. Both of them are based on the study of the intensity variation over time, however the first method analyzes these changes by calculating the differences between couples of cross-sections of the

model whereas the second one uses cross-section aggregations and analyzes the intensity increment. The spatio-temporal Boolean model has been applied to the study of the endocytosis since they are observed as fluorescence spots which overlap in space and time for a random period of time.

Results of simulation studies carried out to test the estimators of the spatio-temporal Boolean model showed relative errors ranging $1\% - 5\%$ in both approaches which indicates that the methodologies and models proposed may fulfill the requirements of real applications. The application of spatio-temporal Boolean models on the particular case of the endocytosis allowed us to estimate among others the mean duration of the events, which turned out to be in agreement with manual studies performed in these sequences.

With the models presented in this thesis, it is possible to study real phenomena showing overlapping objects with random shapes and random durations. In the study of exocytosis and endocytosis we have showed that these models could bring valuable estimates in a fast way and at the same time to reduce the bias introduced by the manual selection of samples when studying such processes.

Abstract

En esta tesis se presentan nuevos modelos y metodologías para el análisis de procesos dinámicos a partir de secuencias de imágenes. El análisis procesos estocásticos a partir de secuencias de imágenes, en las cuales se puede producir superposición espacial y temporal de los objetos, es fundamental para el estudio de muchos fenómenos reales, sin embargo las técnicas y modelos existentes no resuelven completamente el problema.

La motivación original de este trabajo fue el estudio de dos mecanismos celulares, la endocitosis y la exocitosis, a partir de secuencias de imágenes obtenidas por medio de la técnica de microscopía Total Internal Reflection Fluorescence Microscopy. Esta técnica permite la iluminación selectiva y de alta precisión de la membrana citoplasmática. Nuestra aproximación al estudio considera las imágenes obtenidas mediante esta técnica como una realización de un proceso estocástico que cumple las condiciones de estacionariedad e isotropía.

En la primera parte del estudio se propone la aplicación de una metodología para el estudio de patrones puntuales agregados espacial, temporal o espacio temporalmente dentro del contexto de Procesos Puntuales. Fundamentalmente se centra en la aplicación de la $\mathcal{K}$-función de Ripley para el estudio de la relación espacial y temporal de la fusiones producidas por la exocitosis constitutiva. Esta metodología nunca ha sido aplicada antes para el estudio combinado de espacio y tiempo en el proceso de exocitosis constitutiva, aunque puede ayudar a conocer mejor el funcionamiento de la célula o a testar nuevas hipótesis.

Posteriormente, se propone una extensión del modelo Booleano con el fin de estudiar procesos estocásticos espacio temporales con especial énfasis en la dimensión temporal. Se define el *modelo Boolean temporal* para el modelado de objetos superpuestos en espacio y tiempo, y seguidamente se definen los estimadores que permiten obtener el

número medio de eventos por unidad de área y tiempo, su tamaño medio, y la función de densidad de las duraciones, a partir de secuencias de imágenes. Se plantean dos aproximaciones diferentes para estimar los parámetros del modelo. Un primer método basado en diferencias, que estudia la variación de la intensidad del modelo a lo largo del tiempo, calculando diferencias entre pares de secciones temporales separadas por distancias que se van incrementando. Un segundo método que se basa en la agregación de secciones temporales del modelo para estudiar el aumento de intensidad del agregado. El modelo Booleano temporal ha sido aplicado al estudio de las áreas de fluorescencia generadas por proteínas que contribuyen a la endocitosis, las cuales se superponen espacial y temporalmente durante períodos aleatorios.

Los resultados obtenidos tras un estudio de simulación diseñado para testar los estimadores mostraron errores relativos de entre un $1\% - 5\%$ para ambos métodos de estimación, lo cual permite su aplicación a estudios reales. La aplicación del modelo para el caso particular de la endocitosis permitió estimar el número de eventos por unidad de área y tiempo, y su duración media, la cual se encontraba próxima a valores obtenidos manualmente para las mismas secuencias.

Mediante los modelos presentados en esta tesis, es posible analizar fenómenos reales en los cuales se observe superposición de objetos con formas y duraciones aleatorias. Su aplicación para el estudio de la endocitosis y la exocitosis permite la obtención de parámetros estimados de una forma rápida al mismo tiempo que reduce el sesgo introducido por la selección manual de muestras del proceso.

Resumen y Conclusiones

En esta tesis se presentan nuevos modelos estadísticos y metodologías para el análisis de procesos dinámicos a partir de secuencias de imágenes, en las cuales se puede producir superposición espacial y temporal de los objetos. Estos modelos son fundamentales para el estudio de muchos fenómenos reales, sin embargo las técnicas y modelos existentes no resuelven completamente el problema.

Las células animales muestran complejos patrones de transporte intracelular, requeridos no sólo para su supervivencia, sino también para el ensamblaje de las complejas estructuras multicelulares. Aunque sí que se han estudiado en profundidad los mecanismos de transporte, incluyendo la formación de las vesículas en el interior de las células y su fusión en la membrana plasmática, apenas se conoce cómo la membrana plasmática mantiene el equilibrio entre los dos fenómenos básicos de endocitosis y exocitosis, procesos que median la entrada/salida de macromoléculas y partículas para la alimentación, regulación y transporte intracelular. Estos procesos son interdependientes, aunque están inexorablemente relacionados entre sí con el fin de asegurar el mantenimiento del tamao y la forma de la célula.

El advenimiento de la *Green Fluorescent Protein* (GFP y sus derivados cian, amarillo y rojo, CFP/YFP/dsRed, respectivamente), junto con la microscopía por onda evanescente, también conocida como *Total Internal Reflection Fluorescence Microscopy* (TIR-FM), han proporcionado una herramienta muy poderosa para examinar el acoplamiento espacio temporal entre los citados procesos biológicos ((5), (69), (70), (73)). Esta técnica proporciona una resolución vertical y un ratio seal/ruido inigualable con otras técnicas de microscopía. Aunque otros grupos han detectado y seguido gránulos o vesículas durante la exocitosis (70), los algoritmos de detección y de seguimiento de vesículas son burdos y están basados

operaciones básicas de diferenciación de imágenes, filtrado y medidas de distancia sobre el centro de masas. Las nuevas técnicas han producido un incremento en el estudio de fenómenos dinámicos a partir de secuencias de imágenes, especialmente en áreas como Biología y Medicina, debido esencialmente a los últimos avances en Microscopía y al desarrollo de técnicas moleculares para el marcaje de proteínas. Propiedades de tamaño, duración, conteo de eventos por unidad de área y tiempo o interacción entre eventos pueden aportar al experimentador información muy valiosa sobre el comportamiento del sistema (en nuestro caso la célula) ante diferentes condiciones (tratamientos). La gran cantidad de información proporcionada por las técnicas actuales hace que la búsqueda de modelos y métodos formales de contraste de hipótesis sea cada vez más necesaria.

En esta tesis se examinan los procesos biológicos básicos de exocitosis y endocitosis dentro del contexto de la estadística espacial y la geometría estocástica. Este marco de trabajo garantiza la formulación y contraste de hipótesis biológicas asociadas a estos fenómenos. Estos métodos han sido raramente aplicados en Biología Celular hasta la fecha, a pesar de su gran utilidad demostrada en otros campos. Aunque existe una gran cantidad de bibliografía sobre procesos puntuales espaciales (11; 14; 52) y sobre conjuntos aleatorios en el espacio Euclideo (35; 36; 62), no lo hay si se considera el tiempo como una dimensión especial, quizás esto es debido a que hasta la fecha existen pocas aplicaciones en donde sea posible disponer de datos tan exhaustivos que permitan describir el comportamiento espacial y temporal de fenómenos dinámicos. En esta tesis se proponen modelos que asuman explícitamente la dimensión temporal y sus correspondientes estimadores, así como métodos de contraste de hipótesis por técnicas de Montecarlo o boostrap. Estas metodologías proporcionarán la objetividad, el análisis exhaustivo de datos y la robustez necesarias en muchos experimentos.

La motivación original de este trabajo fue por tanto el estudio de dos mecanismos celulares, la endocitosis y la exocitosis, a partir de secuencias de imágenes obtenidas mediante la citada técnica de microscopía *Total Internal Reflection Fluorescence Microscopy*. Nuestra aproximación al estudio considera las imágenes obtenidas mediante esta técnica como una realización de un proceso estocástico que cumple las condiciones de estacionariedad e isotropía.

Estudio de la exocitosis constitutiva mediante procesos puntuales espacio temporales

La exocitosis es un fenómeno muy importante en la vida de las células. El proceso de segregación y desplazamiento apical y basolateral de vesículas ha podido visualizarse en células PtK_2 vivas mediante el uso de técnicas de imagen multicolor aplicadas a células con proteínas fluorescentes. En trabajos anteriores se descubrió la existencia de 'zonas calientes' o 'hot-spots' de exocitosis en la membrana plasmática (31), es decir, se observaron zonas de la membrana en las que las vesículas se fusionaban con mayor frecuencia. Sin embargo, los lugares y tiempos de ocurrencia de las exocitosis nunca han sido analizados de forma conjunta hasta ahora. Además, no existen test formales para describir el comportamiento espacial o espacio temporal de las diferentes hipótesis biológicas formuladas, sino que en su lugar se usan descriptores estadísticos sencillos. Una de las preguntas más importantes a considerar es si existe algún tipo de dependencia entre los lugares de la membrana donde se producen las exocitosis y el momento en el que se producen. Con ello se pretende ahondar en el conocimiento actual de los complejos mecanismos que gobiernan la exocitosis y comprender más detalladamente el movimiento de las vesículas hacia la membrana plasmática.

En la primera parte del estudio se propone la aplicación de una metodología para el estudio de patrones puntuales agregados espacial, temporal o espacio temporalmente dentro del contexto de Procesos Puntuales. Fundamentalmente se centra en la aplicación de la $\mathcal{K}$-función de Ripley para el estudio de la relación espacial y temporal de las fusiones producidas por la exocitosis constitutiva. Esta metodología nunca ha sido aplicada al estudio combinado de espacio y tiempo en el proceso de exocitosis constitutiva, aunque puede ayudar a conocer mejor el funcionamiento de la célula o a testar nuevas hipótesis.

Para llevar a cabo el estudio es necesaria la realización de dos tareas independientes:

- Creación de técnicas de análisis de imagen digital para la obtención de posicion y tiempo de exocitosis a partir de secuencias biológicas.
- Análisis estadístico de la información espacio temporal obtenida tras la fase de análisis de imagen.

En la fase de análisis de imagen proponemos un nuevo método de detección de

vesículas basado en criterios de forma del granulo, tamaño y textura, dentro del contexto de la Morfología Matemática. Para el análisis estadístico de los patrones espacio temporales obtenidos utilizamos la siguiente formulación:

1. Asociamos a cada exocitosis i el par (z_i, t_i) donde z_i es su posición y t_i el instante temporal asociado.

2. Utilizamos un extensión de la $\mathcal{K}$-función de Ripley que depende de dos variable, el espacio y el tiempo.

La $\mathcal{K}$-función de Ripley puede ser extendida para tener en cuenta las dimensiones espacial y temporal. Básicamente, la $\mathcal{K}$-función espacio temporal cuenta todos los eventos dentro de un cilindro con radio definido por una distancia espacial h y una longitud temporal t. Se pretende testar como hipotesis nula, sí los eventos producidos por el proceso de exocitosis constitutiva se comportan como un proceso de Poisson espacio temporal, es decir, no existe ningún tipo de efecto clustering. Esta es la hipótesis nula que se plantea desde el punto de vista biológico. En el estudio realizado sobre secuencias biológicas obtenidas mediante TIRFM se utilizaron un total de 3 secuencias (Ver Tabla 1).

Table 1: Descripción de secuencias de imágenes

Céelula	**Marcador**	**Tamaño**	$\nu_2(W)$	T
1	(VSVG3-SP-YFP)	409×334	74276	750
2	(VSVG3-SP-YFP)	407×320	71151	500
3	(VSVG3-SP-YFP)	455×608	117003	300

En el estudio se estudiaron las secuencias desde tres puntos de vista. Los dos primeros estudios dependientes solo de una variable, en los cuales se investigó si los eventos eran resultado de un proceso aleatório en el espacio o aleatorio en el tiempo. En el tercer estudio se estudió si existiía algún tipo de acoplamiento o dependencia espacio temporal entre los eventos. Para ello se utilizaron diferentes versiones de la $\mathcal{K}$-función de Ripley.

Para el estudio de dependencias espaciales o temporales la $\mathcal{K}$-función se define como,

$$\mathcal{K}(s) = \frac{1}{\lambda}\text{E[número de eventos extra hasta una distancia } s \text{ desde un evento arbitrario]}. \tag{1}$$

Esta función se puede estimar a partir de los datos extraidos en la fase de procesado de imagen mediante el siguiente estimador propuesto por Ripley (14; 52),

$$\hat{\mathcal{K}}(s) = \frac{\nu_2(W)}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} I(d_{ij} \leq s), \tag{2}$$

Para el estudio de las dependencias espaciales también se utilizaron los estimadores de *distancia mínima*, $\mathcal{F}$-función y de *vecino más próximo*, $\mathcal{G}$-función. Estas función están detalladas en (14).

En nuestro estudio se propone describir los procesos puntuales espacio temporales producidos por las exocitosis en una determinada célula mediante una extensión a la dimensión temporal de la $\mathcal{K}$-función,

$$\begin{aligned}\mathcal{K}(s,t) = \frac{1}{\lambda}\text{E[\# de eventos extra dentro de una distancia } s \\ \text{y tiempo } t \text{ de un evento arbitrario]}.\end{aligned} \tag{3}$$

El estimador para $\mathcal{K}$-función espacio temporal es,

$$\hat{\mathcal{K}}(s,t) = \frac{\nu_2(W)T}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} I(d_{ij} \leq s) v_{ij} I(u_{ij} \leq t), \tag{4}$$

donde $u_{ij} = \nu_2(t_i - t_j)$ y v_{ij}= 1, si $t_i - u_{ij} > 0$ y $t_{ij} + u_{ij} < T$, sino v_{ij}=2. En este estimador se tuvieron en cuenta los posibles efectos de borde.

Se aplicaron las funciones $\mathcal{F}$,$\mathcal{G}$ y $\mathcal{K}$ en el ámbito espacial, y se llevó a cabo un test de Monte Carlo con 99 simulaciones para contrastar si la hipótesis nula de no dependencia entre las posiciones de los eventos era factible. Los resultados mostraron claramente que los patrones eran el resultado de un proceso agregado, sugiriendo que las exocitosis no se producían de forma aleatoria sobre la membrana plásmatica. Los p-valores obtenidos pueden observarse en la Tabla 2.

Cuando estudiamos la dimensión temporal obtuvimos resultados similares. En este caso realizamos un test de Monte Carlo con 1000 simulaciones, y los

Table 2: Monte Carlo p-valores usando funciones $\mathcal{F}$, $\mathcal{G}$ y $\mathcal{K}$

	Célula 1	Célula 2	Célula 3
F	0.00	0.00	0.00
G	0.00	0.00	0.00
$\mathcal{K}$	0.02	0.02	0.02

Table 3: Monte Carlo p-valores usando función $\mathcal{K}$

Célula 1	Célula 2	Célula 3
0.0002	0.0001	0.0006

resultados de los p-valores fueron muy concluyentes como se observa en la Tabla 3.

Para el estudio de las dependencias espacio temporales se utilizó,

$$\hat{\mathcal{D}}(s,t) = \hat{\mathcal{K}}(s,t) - \hat{\mathcal{K}}_1(s)\hat{\mathcal{K}}_2(t). \tag{5}$$

Esta nueva función tiene en cuenta que si no existe dependencia entre la dimensión espacial y la temporal estas pueden factorizarse.

Los p-valores obtenidos en el estudio de las dependencias espacio temporales están recogidos en el Tabla 4. Como puede observarse el test de Monte Carlo que indica si existen clusters espacio temporales de exocitosis dió resultados negativos.

Table 4: Monte Carlo p-vaores para testar cluster espacio temporal

Cell 1	Cell 2	Cell 3
0.698	0.753	0.783

Los resultados muestran que no existe dependencia entre las posiciones y los tiempos de ocurrencia de las exocitosis. Por el contrario, observamos que se producen hot-spots de exocitosis en el espacio y en el tiempo de forma marginal. Este estudio demostró un comportamiento temporal inesperado y nunca revelado hasta ahora y que desencadena nuevas preguntas sobre el funcionamiento de la exocitosis constitutiva. Esta metodología puede ser extrapolada fácilmente a

otros tipos de células (células cromafín y neuronas).

Análisis y conteo de áreas endocíticas mediante modelos Boleanos

La endocitosis es el proceso mediante el cual las células transportan nutrientes desde la membrana endoplasmática hasta los endosomas. Las vesículas formadas por clathrin son las portadoras más destacadas en el transporte de este tipo de tráfico intracelular. La GTPase dinamin se presenta como un mediador crucial en el proceso de endocitosis de las vesículas formadas por clathrin. Nuevas técnicas de visualización han permitido observar la formación de vesículas envueltas en clathrin en células vivas. Se ha asumido que el mínimo cluster de moléculas de dinamin fluorescente que puede ser detectado corresponde con un 'spot' de vesículas rodeadas de clathrin. Cuando se observa mediante microscopio la membrana plasmática, las áreas de fluorescencia asociadas a diferentes 'spots' endocíticos se solapan formando 'clumps' aleatorios con diferentes formas y tamaños. En este estudio se presenta una metodología que permite obtener el número medio de 'spots' endocíticos en cada imagen, así como su área y perímetro medios, usando para ello una secuencia de imagenes. Para obtener estos datos a partir de una secuencia de imágenes capturadas con microscopio es necesario desarrollar los mecanismos necesarios de análisis de imagen. Uno de los principales retos radica en que los 'spots' endocíticos están total o parcialmente solapados entre ellos, por lo que las técnicas de análisis de imagen solo sirven como información básica de entrada. De esta información se estimarán posteriormente los parámetros relevantes del proceso Boleano subyacente.

La definición a grandes rasgos de un modelo Boleano es, un conjunto de puntos en el plano (gérmenes) producidos mediante una distribución de Poisson estacionaria con intensidad λ, sobre los que se ha colocado una forma aleatoria (granos). Más formalmente: Sea $\Psi = \{x_1, x_2, \ldots\}$ un proceso de Poisson estacionario en $\mathbb{R}^2$ con intensidad λ. Sea $\Xi_1, \Xi_2, \ldots$ una secuencia de conjuntos compactos aleatorios (como Ξ_0), independientemente distribuidos en $\mathbb{R}^2$, que a su vez son independientes del proceso de Poisson Ψ, y que cumplen $E\nu_2(\Xi_0 \oplus \check{K}) < +\infty$, para todo conjunto compacto K. El modelo Boleano es un conjunto aleatorio definido como,

$$\Phi = \cup_{n=1}^{\infty}(\Xi_n + x_n). \tag{6}$$

El conjunto aleatorio Ξ_0 recibe el nombre de grano primario. La distribución Φ se puede caracterizar mediante la función de capacidad. Esta función para un modelo Boleano es,

$$T(K) = 1 - \exp\{-\lambda E\nu_2(\Xi_0 \oplus \check{K})\}. \tag{7}$$

Entre las características del modelo Boleano, la fracción de volumen es de vital importancia, ya que mide la fracción media de volumen ocupada por Ξ en una región por unidad de área,

$$p = E(\nu_2(\Xi \oplus B)), \quad \Xi(B) = 1. \tag{8}$$

Puesto que p no depenede de la elección de la región B cuando Ξ es estacionario, podemos reescribir (para $K = 0$),

$$P(0 \in \Xi) = 1 - exp\{-\lambda E(\nu_2(\Xi_0))\} \tag{9}$$

Partiendo de estas definiciones de G. Matheron (35; 36), vamos a asumir que las imágenes obtenidas mediante microscopía TIRFM en las que se observan endocitosis son el resultado de un modelo Boleano espacial. Por tanto podemos aplicar la formulación de un modelo Boleano para estimar parámetros relevantes. Existen distintos métodos, aquí se realiza una implementación del *método del contraste mínimo* (ver (62)). Los parámetros en los que estamos interesados son, la intensidad del proceso, el área media de cada evento, que en nuestro caso es el área de fluorescencia producida por una endocitosis, y finalmente su perímetro medio.

De las definiciones originales de Matheron (ver (36) pp. 139 y (60)) obtenemos,

$$H_K(t) = 1 - exp\Big\{ - \lambda E[\nu_2(A_0 \oplus tK) - \nu_2(A)]\Big\}, \tag{10}$$

donde K es un disco unidad en el plano y A_0 es el grano primario. La fórmula general de Steiner establece que,

$$E\nu_2(A_0 \oplus \check{K}) = E\nu_2(A_0) + \frac{U(K)EU(A_0)}{2\pi} + \nu_2(K). \tag{11}$$

La función de distribución de contacto $H_K(t)$ se puede calcular directamente de

las imágenes obtenidas del microscopio mediante,

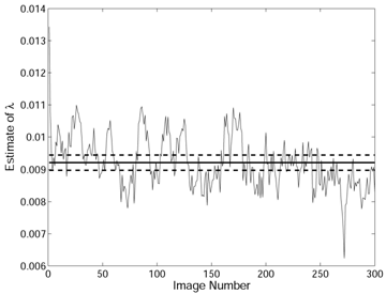
$$H_K(t) = 1 - \frac{1 - T(tK)}{1 - p}, \tag{12}$$

puesto que conocemos el estimador de la función de capacidad $T(K)$ para un modelo Boleano $T(K)$,

$$\hat{T}_W(K) = \frac{\nu_2((\Phi \oplus \check{K}) \cap (W \ominus K))}{\nu_2(W \ominus K)}, \tag{13}$$

donde W es la ventana de muestreo y $A \ominus \check{K} = \{x \in A : x + K \subset A\}$.

Esta metodología se aplicó al análisis de 6 secuencias de imágenes obtenidas mediante microscopio confocal en las que se mostraban endocitosis, gracias al marcado con fluoroporos de la clatrina, una proteína que interviene en el proceso. Puesto que las secuencias cuentas con cientos de frames, los estimadores se aplicaron independientemente para cada uno de ellos. La Fig. 1.2, muestra los resultados obtenido tras la aplicación de estos estimadores sobre una secuencia de 300 imágenes. El elemento estructurante utilizado ha sido un cuadrado, puesto que su aproximación digital es mejor que la de otros elementos como los discos. Las estimaciones obtenidas para cada frame se combinaron utilizando el método del *batch-mean* (33), ya que estaban correlacionados. Finalmente, se obtuvo el valor medio para cada uno de los parámetros de interés en cada secuencia.

(a) 

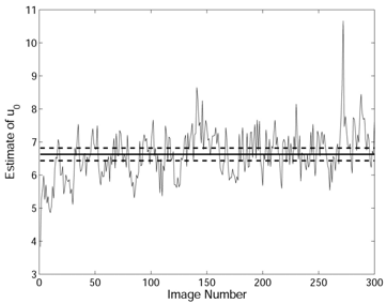
(b) 

Figure 1: **Valores estimados de intensidad λ, y perímetro u_0 para una de las secuencias de clatrina utilizando un cuadrado unidad**. (a) Estimaciones e intervalos de confianza al 95% para $\hat{\lambda}$. (b) Idem para el perímetro medio del grano primario $\hat{u}_0$.

Los resultados obtenidos fueron de una gran precisión con errores muy pequeños en la estimación de los parámetros, como muesta la Tabla 5 donde se resumen

los valores medios y sus respectivos intervalos de confianza.

Table 5: Valores estimados eintervalos de confianza al 95% de 6 secuencias expresando clatrina

Cell	Sequence	$\hat{\lambda}$	$\hat{u}_0$	$\hat{\lambda}\nu_2(W)$
1	1	0.0099 ± 0.0006	9.565 ± 0.28	281.86 ± 17.49
	2	0.0085 ± 0.0006	9.575 ± 0.32	347.20 ± 23.69
2	3	0.0092 ± 0.0002	6.6269 ± 0.19	277.54 ± 7.23
	4	0.0097 ± 0.0003	7.1997 ± 0.24	235.41 ± 8.05
3	5	0.0081 ± 0.0010	9.0042 ± 0.37	170.17 ± 20.88
	6	0.0075 ± 0.0012	8.9756 ± 0.34	220.28 ± 3.54

En las siguientes secciones se propone una extensión del modelo Boleano con el fin de estudiar procesos estocásticos espacio temporales con especial énfasis en la dimensión temporal. Se define el modelo Boleano espacio temporal para el modelado de objetos superpuestos en espacio y tiempo, y seguidamente se definen los estimadores que permiten obtener el número medio de eventos por unidad de área y tiempo, su tamaño medio, y la función de densidad de las duraciones, a partir de secuencias de imágenes. Se plantean dos aproximaciones diferentes para estimar los parámetros del modelo. Un primer método basado en diferencias, que estudia la variación de la intensidad del modelo a lo largo del tiempo, calculando diferencias entre pares de secciones temporales separadas por distancias que se van incrementando. Un segundo método que se basa en la agregación de secciones temporales del modelo para estudiar el aumento de intensidad del agregado.

Análisis de eventos superpuestos espacial y temporalmente y definición de los STBMs

El conteo de eventos superpuestos en una secuencia de imágenes así como la estimación de sus tamaños y duraciones es un problema importante en muchas aplicaciones reales. En este trabajo se propone un procedimiento para llevar a cabo estas tareas basándose en un modelo estocástico. El modelo propuesto recibe el nombre de modelo Boleano espacio temporal, siendo una generalización

del modelo Boleano clásico en el cual las superposiciones entre diferentes formas es explícitamente asumida y modelada.

Definiremos el modelo Boleano espacio temporal como la unión de muestras con formas aleatorias (granos), colocadas de acuerdo a un proceso puntual de Poisson (gérmenes) y con una duración aleatoria asociada (independiente e idénticamente distribuida para los diferentes granos). El modelo propuesto tiene una aplicación directa al estudio del proceso de endocitosis, descrito anteriormente. La ventaja que introduce este nuevo modelo es la capacidad de tener en cuenta la dimensión temporal y ser capaz de obtener la estimación de la duración de los eventos. Obtener estos valores no es posible por métodos de análisis de imagen puesto que los eventos están superpuestos y su número sería intratable en secuencias largas.

A continuación, el modelo Boleano espacio temporal y sus propiedades probabilísticas son formalmente presentados. Proponemos un método para la estimación de parámetros en un intervalo temporal de una secuencia de imágenes basado en diferencias. A continuación, se dan los estimadores para el número medio de gérmenes por unidad de área y tiempo, el tamaño medio de los granos y su distribución de duraciones. Posteriormente, se lleva a cabo un estudio de simulación, en el cual se estudian los errores relativos asociados a los estimadores propuestos, sobre secuencias simuladas cuyos parámetros son conocidos. Finalmente, el método es ilustrado sobre secuencias de imágenes biológicas con el fin de estimar el número medio de eventos endocíticos en la membrana plasmática así como su duración media. Los resultados muestran que los modelos propuestos son efectivos y permiten obtener información de procesos dinámicos en los cuales objetos superpuestos de corta vida son observados.

Definición formal del modelo Boleano espacio temporal. Sea $\Psi = \{(x_i, t_i)\}_{i\geq 1}$ un proceso de Poisson estacionario en $\mathbb{R}^2 \times \mathbb{R}_+$ con intensidad λ. Sea $\{A_i\}_{i\geq 1}$ una secuencia de conjuntos aleatorios compactos independientes e identicamente distribuidos (como A_0) en $\mathbb{R}^2$. Sea $\{d_i\}_{i\geq 1}$ una secuencia de variables positivas independientes e identicamente distribuidas (como D) . Asumimos que Ψ, $\{A_i\}_{i\geq 1}$ y $\{d_i\}_{i\geq 1}$ son independientes y que $E\nu_3(A_0 \times [0, D] \oplus \check{K}) < +\infty$ para cualquier subconjunto K de $\mathbb{R}^3$. El modelo Boleano espacio temporal, STBM, es el conjunto aleatorio definido como,

$$\Phi = \cup_{i\geq 1}(A_i + x_i) \times [t_i, t_i + d_i]. \tag{14}$$

El conjunto $(A_i + x_i) \times [t_i, t_i + d_i]$ es un cilindro en $\mathbb{R}^2 \times \mathbb{R}_+$, el i-ésimo evento. El método de estimación que se presenta a continuación se basa en dos hechos.

- Una sección $2D$ de un modelo Boleano $3D$ es también un modelo Boleano (36; 66).
- Si Φ es un modelo Boleano espacio temporal con intensidad λ y grano primario $A_0 \times [0, D]$, entonces una sección temporal Φ_s es un modelo Boleano $2D$ con grano primario A_0 e intensidad

$$\lambda_s = \lambda ED. \tag{15}$$

En el método por diferencias un modelo Boleano espacio temporal lo consideraremos la unión de los siguientes modelos Boleanos,

$$\Phi_{s_1,s_2} = \cup_{i:t_i \le s_1 \le s_2 \le t_i + d_i} A_i + x_i, \tag{16}$$

$$\Phi_{s_1,s_2^-} = \cup_{i:t_i \le s_1 \le t_i + d_i < s_2} A_i + x_i, \tag{17}$$

$$\Phi_{s_1^-,s_2} = \cup_{i:s_1 < t_i \le s_2 \le t_i + d_i} A_i + x_i. \tag{18}$$

A partir de estos modelos Boleanos independientes definimos una serie de funciones muy importantes. $\beta(s_2 - s_1)$ indica la intensidad de Φ_{s_1,s_2} (número medio de gérmenes vivos en el intervalo temporal $[s_1, s_2]$), y viene dado por,

$$\beta(s_2 - s_1) = \lambda p(s_2 - s_1), \tag{19}$$

donde $p(s) = \int_s^{+\infty} P(D \ge v) dv$. Además, asumiendo estacionariedad, Φ_{s_1,s_2^-} y $\Phi_{s_1^-,s_2}$ tienen la misma intensidad, $\alpha(s_2 - s_1)$ (número medio de gérmenes que han muerto en el intervalo temporal $[s_1, s_2]$) dada por,

$$\alpha(s_2 - s_1) = \lambda ED - \lambda p(s_2 - s_1) = \lambda_s - \beta(s_2 - s_1). \tag{20}$$

El estimador de la función $\alpha(s_2 - s_1)$ puede ser obtenido a partir de las siguientes ecuaciones,

$$P(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2}) = P(0 \in \Phi_{s_2} | 0 \notin \Phi_{s_1}) = 1 - \exp\{-\alpha(s_2 - s_1) a_0\}, \tag{21}$$

$$1 - \exp\{-\hat{\delta}(s_2 - s_1)\} = \frac{1}{2}\left(\frac{\nu_2(\phi_{s_1} \cap \phi_{s_2}^c \cap W)}{\nu_2(\phi_{s_2}^c \cap W)} + \frac{\nu_2(\phi_{s_2} \cap \phi_{s_1}^c \cap W)}{\nu_2(\phi_{s_1}^c \cap W)} \right). \tag{22}$$

donde $\delta(s_2 - s_1) = \alpha(s_2 - s_1)a_0$.
Mediante el método del contraste mínimo podemos obtener estimaciones para $\hat{a}_0$ y $\hat{\lambda}_s$, y posteriormente calcular,

$$\hat{\alpha}(s) = \frac{\hat{\delta}(s)}{\hat{a}_0} \text{ and } \hat{\beta}(s) = \hat{\lambda}_s - \hat{\alpha}(s). \tag{23}$$

Finalmente, podemos obtener la intensidad del proceso de gérmenes espacio temporal y la función de densidad mediante $\alpha'(s) = \lambda(1 - F_D(s))$ y $\alpha''(s) = -\lambda f_D(s)$ donde α' y α'' son la primera y segunda derivadas de α, mientras que F_D and f_D son la función de distribución acumulada y la función de densidad de la variable aleatoria D, respectivamente.

Para testar los estimadores definidos se llevó a cabo un estudio de simulación utilizando 9 combinaciones de parámetros diferentes y generando 15 replicas para cada una de estas. La Tabla 6 muestra los errores relativos para los diferentes parámetros obtenidos tras el estudio de simulación. En todos los casos estos errores se mantienen en un límite muy bajo y por tanto aceptables para un estudio real.

Table 6: Errores relativos con respecto a la distribución de duraciones

Parametro	Distribución	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.
λ	Uniforme	-0.084	-0.003	0.022	0.022	0.047	0.125
	Exponencial	-0.057	0.016	0.053	0.055	0.090	0.153
	Gamma	0.127	-0.010	0.013	0.017	0.044	0.159
λ_s	Uniforme	-0.086	-0.020	0.004	0.000	0.022	0.090
	Exponencial	-0.093	-0.020	0.008	0.009	0.034	0.109
	Gamma	-0.109	-0.026	0.004	0.000	0.022	0.096
a_0	Uniforme	-0.087	-0.049	-0.028	-0.025	-0.005	0.073
	Exponencial	-0.105	-0.057	-0.031	-0.032	-0.010	0.074
	Gamma	-0.093	-0.047	-0.030	-0.027	-0.007	0.066
u_0	Uniforme	-0.055	-0.014	0.005	0.009	0.028	0.096
	Exponencial	-0.078	-0.023	0.002	0.002	0.025	0.113
	Gamma	-0.062	-0.010	0.006	0.007	0.027	0.087

El modelo Boleano temporal fué aplicado al estudio de las áreas de fluorescencia generadas por proteínas que contribuyen a la endocitosis, las cuales se superponen espacial y temporalmente durante períodos aleatorios. En concreto se aplicó sobre 6 secuencias de imágenes. En todos los casos se obtuvieron valores de duraciones que estuvieron en los rangos esperados por los biólogos en base a la bibliografía existente. Además nuevos parámetros nunca obtenidos hasta ahora como la intensidad del proceso del gérmenes fueron satisfactoriamente obtenidos, aunque su valor no ha podido ser contrastado.

Estimación de parámetros en STBMs mediante TBM 2D agregados

Un segundo método ha sido desarrollado para la evaluación y estimación de parámetros de un modelo Boleano espacio temporal, STBM. Este segundo método permite la estimación de los mismos parámetros que el modelo basado en diferencias presentando en el punto anterior, pero en este caso se consideran agregaciones de modelos Boleanos (BM) 2D para obtener los estimadores correspondientes. Esta implementación tiene como ventaja principal frente al método anterior una mayor inmunidad al ruido. El principio en que se basa este método es el hecho de que "La unión de modelos boleanos independientes sigue siendo un modelo Boleano" (Matheron75). Esto nos permite descomponer un modelo Boleano espacio temporal en sus diferentes secciones y estimar sus parámetros mediante agregaciones de estas secciones de una forma específica.

Partimos de un modelo Boleano espacio temporal:

- Construimos nuevas secuencias acumulando sobre la dimensión temporal.
- Acumulamos usando diferentes combinaciones de (k, δ),
  - Número de secciones temporales, k.
  - Distancia temporal entre dichas secciones acumuladas, δ.

Las diferentes secuencias obtenidas por acumulación usando combinaciones de (k, δ) nos permitirán inferir parámetros del modelo Boleano espacio temporal original. El número medio de eventos nacidos antes de una sección temporal determinada y muertos en el intervalo $[s_i, s_{i+1})$ puede obtener mediante,

$$EN_{(-\infty,s_1]}^{[s_i,s_{i+1})} = \theta p(s_i - s_1) - \theta p(s_{i+1} - s_1). \tag{24}$$

En general, el número medio de eventos nacidos en un intervalo $(s_{i-1}, s_i]$ y muertos en otro intervalo posterior $(s_{j-1}, s_j]$ viene dado por,

$$EN_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = \theta p(s_{i-1}, s_i; s_{j-1}, s_j), \tag{25}$$

donde

$$p(s_{i-1}, s_i; s_{j-1}, s_j) = p(s_{j-1} - s_i) - p(s_{j-1} - s_{i-1}) - p(s_j - s_i) + p(s_j - s_{i-1}). \tag{26}$$

En realidad nosotros estamos interesados en estimar parámetros partiendo de un modelo acumulado. Por ejemplo, el número medio de eventos vivos en $\cup_{i=1}^{k} \Psi_{s_i}$ que viene dado por $\sum_{i=1}^{k} \sum_{j=i+1}^{k+1} N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$. La formulación base para realizar estas estimaciones es,

$$\psi(s_1, \ldots, s_k) = E \sum_{i=1}^{k} \sum_{j=i+1}^{k+1} N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = \theta \left[kp(0) - \sum_{i=2}^{k} p(s_i - s_{i-1}) \right]. \tag{27}$$

Los estimadores nos permitirán obtener λ, $\alpha(\delta)$, y ED. Aplicando el método del contraste mínimo podemos obtener para cada combinación de (k, δ) su respectivo $\lambda_s(k, \delta)$ y así obtener,

$$\lambda_s(k, \delta) = \frac{\psi(s_1, s_1 + \delta, \ldots, s_1 + (k-1)\delta)}{\nu_2(W)} = \lambda \left[kp(0) - (k-1)p(\delta) \right]. \tag{28}$$

Esta ecuanción se puede reescribir como,

$$\lambda_s(k, \delta) = (\lambda p(0) - \lambda p(\delta))k + \lambda p(\delta), \tag{29}$$

Es evidente que ahora podríamo obtener las funciones $\alpha(\delta$ y $\beta(\delta)$ a partir de estimaciones para diferentes k, $\alpha(\delta) = \lambda p(0) - \lambda p(\delta)$, y $\beta(\delta) = \lambda p(\delta)$. Finalmente λ y la función de densidad puede obtenerse a partir de las siguiente ecuaciones,

$$\alpha'(0) = -\lambda p'(0) = \lambda P(D \geq 0) = \lambda. \tag{30}$$

y

$$p''(\delta) = f_D(\delta), \tag{31}$$

Un análisis de simulación realizado con 6 tipos de secuencias diferentes permitió medir la precisión de los estimadores. Los errores relativos fueron constantes

y muy bajos, del orden del 5% en los peores casos, que correspondieron a secuencias cuyos eventos tenían una distribución exponencial.

Esta metodología también se aplicó sobre secuencias biológicas para el estudio de endocitosis. En concreto se usaron 6 secuencias en las que se había marcado la proteína clatrina. Los resultados al igual que con el método por diferencias dieron resultados dentro de los rangos temporales esperados, y fueron de la misma magnitud que los obtenido por otros grupos en similares experimentos, aunque en nuestro caso la variabilidad se mostró más baja.

Conclusiones

En este trabajo se han propuesto diferentes métodos estadísticos y computacionales para analizar procesos estocásticos espacio temporales. Estas metodologías han sido aplicadas al estudio de procesos biológicos celulares ampliamente conocidos, la exocitosis y la endocitosis. Estos procesos son intrinsecamente distintos y por tanto generan diferentes tipos de información, lo cual conlleva a diferentes metodologías de análisis dentro del marco de los procesos estocásticos.

Primeramente, se ha propuesto una metodología para el estudio de los patrones puntuales espacio temporales obtenidos tras el análisis de imagen de una secuencia de exocitosis. Esta metodología ha sido probada sobre secuencias reales, en las cuales se han analizado las dependencias, espaciales, temporales y espacio temporales exitosamente, obteniendo información nunca vista hasta ahora. Para ello se ha utilizado una combinación de funciones bien conocidas como son $\mathcal{F}$, $\mathcal{G}$ y $\mathcal{K}$ y test de Montecarlo. Aunque la metodología mostrada aquí se ha aplicado a un caso particular, es suficientemente flexible para ser aplicada a una gran variadedad de casos similares en los que se trate este tipo de información.

Para el problema biológico, se desarrollo una aplicación de procesado de imagen con el fin de obtener las posiciones y tiempos en los que se producían las exocitosis. Este algoritmo permitió obtener de una forma rápida y semi-automática los datos necesarios para el estudio.

En los siguientes capítulos hemos definido y estudiado un caso particular de modelo Boleano, que llamamos *modelo Boleano espacio temporal*, que formaliza

la configuración de eventos independientes con posiciones y duraciones aleatorias. Además este modelo considera la posibilidad de superposición tanto espacial como temporal de estos eventos. Esta aproximación nos ha permitido obtener información de procesos estocásticos con formas y duraciones aleatorias que seguen una determinada distribución.

Se han definido dos métodos diferentes para el estudio de modelos Boleanos espacio temporales. El primero, basado en diferencias entre pares de secciones temporales de un modelo temporal espacio temporal. Con este método podemos observar los cambios en intensidad dependiendo de la distancia temporal entre las diferencias escogida. El segundo método se basa en la agrupación de conjuntos pertenecientes a diferentes secciones temporales de un STBM. De este modo la nueva secuencia agrupada muestra cambios en intensidad que de nuevo revelan información de modelo original. La combinación de diferentes números de conjuntos agregados y de diferentes distancias temporales entre estos conjuntos dan lugar a nuevas secuencias agregadas que permitirán la inferencia de parámetros del modelo. Uno de los inconvenientes de este método es que requiere de mayor tiempo para estimar los parámetros ya que primeramente es necesario crear las nuevas secuencias agregadas y luego analizarlas. Cabe remarcar que el tiempo que este algoritmo necesita es ínfimo comparado con el que requeriría un método manual.

Además, también se han definido los estimadores para la obtención del número medio de eventos por unidad de área y tiempo, el tamaño medio de los eventos, su función de densidad y su duración a partir de una realización de un modelo Boleano espacio temporal. Un exhaustivo estudio de simulación ha sido realizado para estudiar la bondad de los estimadores. En particular, se usaron 135 secuencias simuladas de modelos Boleanos espacio temporales utilizando 9 combinaciones de parámetros con 15 replicas cada una. Los errores relativos obtenidos fueron muy bajos y se situaron en un rango de $0.3\% - 15\%$, siendo su mayoría menores del 5%. A lo largo de los diferentes procesos de estimación con funciones se utilizaron tecnicas de *Functional Data Analysis*, con el fin de mejorar las estimaciones y poder derivar funciones de forma más precisa.

Para ilustrar la aplicación de los modelos Boleanos espacio temporales, estudiamos las propiedades espacio temporales del proceso biológico de endocitosis. Las técnicas de análisis de imagen o las basadas en procesos puntuales son incapaces de obtener información fiable de estas secuencias puesto que los eventos

se solapan en espacio y en tiempo. Sin embargo, los dos métodos basados en STBM fueron aplicados exitosamente a secuencias celulares obtenidas mediante microscopios confocales. Varios parámetros relacionados con la dimensión temporal fueron estimados, tales como i) el número medio de endocitosis por unidad de tiempo y de membrana, ii) la función de densidad de las duraciones de dichos eventos, así como su duración media.
Una aproximación clásica basada en técnicas de análisis de imagen que segmentan componentes conexas hubiese infraestimado tanto el número de endocitosis reales presentes en cualquier instante de tiempo, como su duración media. Por tanto, los modelos STBM así como los métodos de estimación se presentan como una nueva herramienta automática capaz de obtener estimaciones de parámetros relevantes de un proceso. Además este tipo de medidas son muy complicadas de obtener de forma manual, aunque son necesarias para el sondeo de tratamientos en campos como la biología.

Finalmente, es importante mencionar que aunque la exocitosis y la endocitosis son dos procesos estrechamente relacionados, un análisis formal que combine ambos procesos es todavía necesario. Los resultados mostrados aquí indican que la exocitosis constitutiva se produce de forma agrupada sobre la membran plasmática indicando algún tipo de dependencia, mientras que las regiones de fluorescencia producidas por las endocitosis se consideran aleatorias sobre la membrana. Estas asunciones son seguras, ya que no existen evidencias claras de lo contrario.

Existen varias extensiones interesantes a los modelos propuestos, y que permitirían testar una mayor cantidad de hipótesis de mayor complejidad. En general, estas mejoras implican también una mejora en la calidad o en el tipo de datos para el modelo, que no siempre están disponibles debido a restricciones tecnológicas.

En resumen, se ha definido y presentado una metodología para estimar parámetros espacio temporales de modelos estocásticos los cuales nos permiten testar hipotesis complejas. Las metodologías y modelos descritos en este trabajo permiten obtener información sobre distribuciones temporales, intesidades o medidas de tamaño de procesos estocásticos. Todos los estimadores propuestos fueron testados mediante estudios de simulación exhaustivos y aplicados a problemas complejos en el campo de la biología celular.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Problems in Engineering, Computational Science and Biology are using increasingly complex mathematical and computational techniques, creating a bridge between these disciplines that is heavily travelled. Traditional fields, as Biology and Medicine, are encouraging the ways that mathematics could be applied towards new and innovative models which can help to the understanding of different processes.

Here, we study the dynamics of the processes of endocytosis and exocytosis separately from image sequences obtained by means of a microscopy technique know as Total Internal Reflection Fluorescence Microscopy (TIRFM).

The study of these two processes has been done in the context of stochastic geometry. Specifically, we based it on theory related to Point Processes and Random Sets since they present a natural framework for modelling the binary image sequences we deal with. The main goal of this work is to develop a methodology to analyze the spatial pattern of either biological process individually. In order to accomplish this requirement we have i) extended the classical Boolean model to take into account the temporal dimension; ii) proposed the use of estimator for spatio-temporal point processes to model real cellular phenomena.

These new estimators allow us to characterize these spatial temporal cellular processes accurately and in a fast way. So far, this kind of analysis has been performed manually from data obtained by researchers in a one by one basis (19), making it virtually impossible in many situations. In addition, manual selections of this kind of data use to introduce a high bias in the sample. Furthermore, it is important to point out that each one of the experiments carried out to study

the mentioned processes conveys the analysis of hundreds or even thousands of frames which led to an impossible task.

The problem of studying the existence of spatio-temporal clustering is not new in the literature of Spatial Statistics and has been applied in many frameworks that increase every year, specially with the advent of powerful microscopes, geographical information systems, etc. This problem is also very common in Ecology and in Epidemiology, as for instance, finding clusters of infected people around a focus, in space and time, that is, people infected that live in the neighborhood of the focus and that became infected close in time. Sciences such as Agriculture, Astronomy, Meteorology, etc., to name a few, have in spatial temporal Statistic the basis of their research. Spatial Statistics has provided a number of tools in order to analyze any given set of points, models or processes which can generate point data following some stochastic law (11; 14).

Theory on spatial point processes has been developed with extensive amounts of literature devoted to their analysis. However, if the observed process is a hybrid spatial temporal point process, only few practical methods exists. In such cases, practitioners often remove the temporal component and analyze the spatial dependencies. This marginal spatial analysis may lead to misleading conclusions if time is an important factor in the process. In (52) was introduced the analysis in the spatial domain through the widely used $\mathcal{K}$-function. Currently, Ripley's approach to study the dependency structure of point patterns remains the dominant method for analysis.

Computational Biology attempts to create and test models that describe complex cellular processes (19; 43; 46; 47; 64). In addition, problems similar to the biological ones studied here that involve existence of random spatial patterns have been encountered in various other fields, such as Ecology and Geographical Information Systems (14), (11), (51).

One of the first aims of this work is to present a new methodology for the study of the joint spatio-temporal distribution of exocytosis events in the plasma membrane in the context of the theory of spatio-temporal point processes, that may provide new insight into understanding the complex exocytosis mechanisms in membrane trafficking. The use of spatio-temporal point pattern analysis in Biology is strongly new, though some accounts in biological literature as (31) and (46) use spatial domain approximations based on point patterns. We consider the second-order analysis of spatio-temporal point patterns through an extension of Ripley's $\mathcal{K}$-function. From a practical point of view, we apply this methodology to the analysis of constitutive exocytosis patterns in epithelial cells.

Approaches based on random sets have also been considered to study similar processes in other areas of knowledge. In particular, the Boolean model is a well-known model when we study binary images composed of the union of different (independent random) shapes located independently in the 2D Euclidean space. The basic references about this topic are (35; 36; 62). The use of such models allows us to estimate (among other things) the mean number of objects per unit area or the mean area and perimeter of the objects from an image. It has been widely used (26; 40; 66) ever since it was proposed by G. Matheron in the late 1960s , since it can provide a good description for irregular patterns observed in Microscopy, Material Sciences, Biology, Chemistry, Geostatistics or Cellular Communication Networks.

We are concerned with the analysis of overlapping short-lived events from an image sequence. In fact, endocytic spots show a similar pattern and overlap to each other both in space and time and live. In approaching this problem, we have defined the spatio-temporal Boolean model (from now on STBM), a particular case of a non-isotropic 3D Boolean model. This new model formalizes the configuration of independent randomly placed events with independent durations. We will derive some probabilistic properties of the proposed model and consider the estimation of some characteristics such as the mean number and life-time of the germs from a time-lapse image sequence.

1.1 Biological background and motivation

Every cell must eat, and it must communicate with the world around it. In a procaryotic cell, all the eating and communicating takes place across the plasma membrane. The biosynthetic-secretory pathway allows the cell to modify the molecules it produces in a series of steps, store them until needed, and then deliver them to the exterior through a specific cell-surface domain by a process called *exocytosis*. The complementary process, the *endocytosis*, allows the cell to take up molecules and deliver them inside the cell. In this process, the material to be ingested is progressively enclosed by a small portion of the plasma membrane, which first invaginates and then pinches off to form an *endocytic vesicle* containing the ingested substance or particle.

Vesicle recycling through exocytosis and endocytosis is mediated by a coordinated cascade of protein-protein interactions. Here, exocytosis and endocytosis are studied separately so that the coupling between them is understood only indirectly. Other studies have focused on the coupling of these processes by observing

secretory vesicle markers and the endocytotic vesicle marker tagged with green and red fluorescent proteins under an evanescent wave microscope. Nonetheless, this is not the aim of this work.

1.1.1 Exocytosis

Golgi-derived vesicles travel along cytoskeletal elements toward the plasma membrane where they encounter their docking and fusion sites and undergo the process of *exocytosis*. In the so called constitutive secretory pathway, membrane protein and lipids are continuously delivered to the cell surface while some proteins are excreted into the extracellular space (44). Specialized secretory cells also have a regulated exocytosis pathway, where an extracellular signal is required to induce secretion of specific proteins such as neurotransmitters (2), (3) (see Fig. 1.1). In the present study we focus on the constitutive secretory pathway which is important in the delivery of membrane lipids and proteins to the cell surface. These molecules are implicated in numerous cellular processes including signaling, adhesion, migration, and differentiation. This pathway is therefore fundamental for cell function throughout the lifetime of every cell. Not surprisingly, disturbances of exocytosis characterize a variety of human disorders such as diabetes, hypertension and several neurological diseases.

TIRFM has been recently applied (31), showing the existence of preferred sites for exocytosis, but it was a marginal study and no joint information of space and time was considered and analyzed. In posterior papers (58), new environmental factors related with the exocytosis, such as the role of microtubules were taken into account, leading to new hypotheses where some kind of vesicles are transported away from the Golgi via microtubules, so that colocalization of fusion sites is expected. These hypotheses were corroborated in (57) where it was concluded that microtubules are necessary for the domain-specific fusion of Post-Golgi vesicles with the plasma membrane.

Questions that still remain open are,

1. Does the rate of exocytosis vary over time?
2. Is there any mechanism encouraging exocytosis waves?
3. Are exocytosis clustered spatially and temporally?

In the literature there are studies based on either patch clamp capacitance and amperometric techniques or total internal reflection microscopy, but the informa-

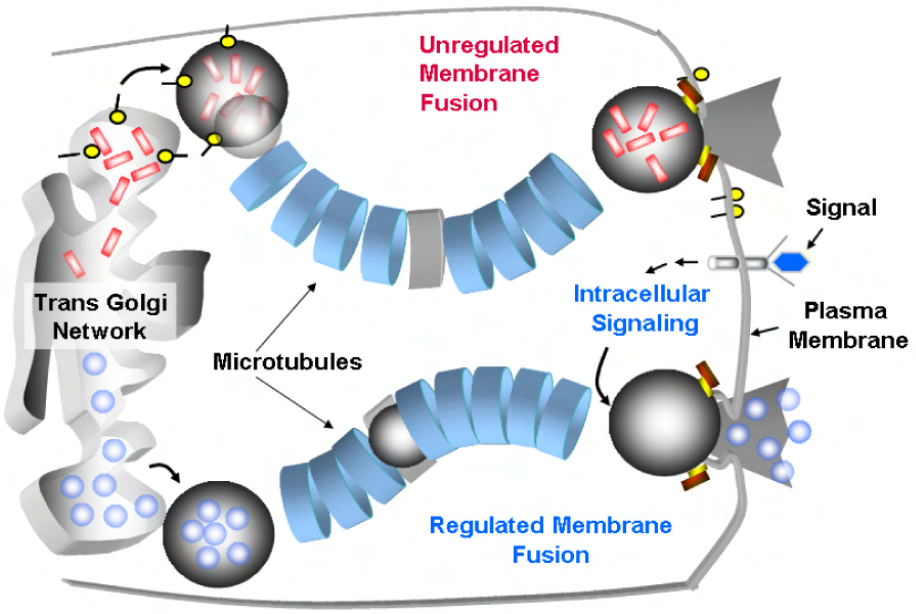


Figure 1.1: **Constitutive and regulated secretory pathways in cells.**

tion of locations and times of occurrence of exocytosis has never been analyzed together (70), (32), (12),(27).

1.1.2 Endocytosis

Endocytosis is the process by which cells carry traffic from the plasma membrane into various intracellular compartments. Among the different endocytic routes, clathrin-mediated endocytosis has been more extensively characterized, and operates in many important cellular processes. The life cycle of a clathrin-coated vesicle, illustrated in Fig. 1.2, involves a sequence of regulated events: a) Cargo loading, where cargo molecules bind to receptors on the plasma membrane; b) Coat assembly, where a molecular lattice of clathrin molecules ensheathes a portion of the plasma membrane containing the cargo-receptor complex; c) Vesicle budding, followed by its pinching-off from the plasma membrane; d) Internalization and coat disassembly; e) Intracellular trafficking of the endocytosed vesicle. Many of these steps have traditionally been inferred from biochemical and cell structural studies (8).

The process of endocytosis has been recently observed in living cells, due to

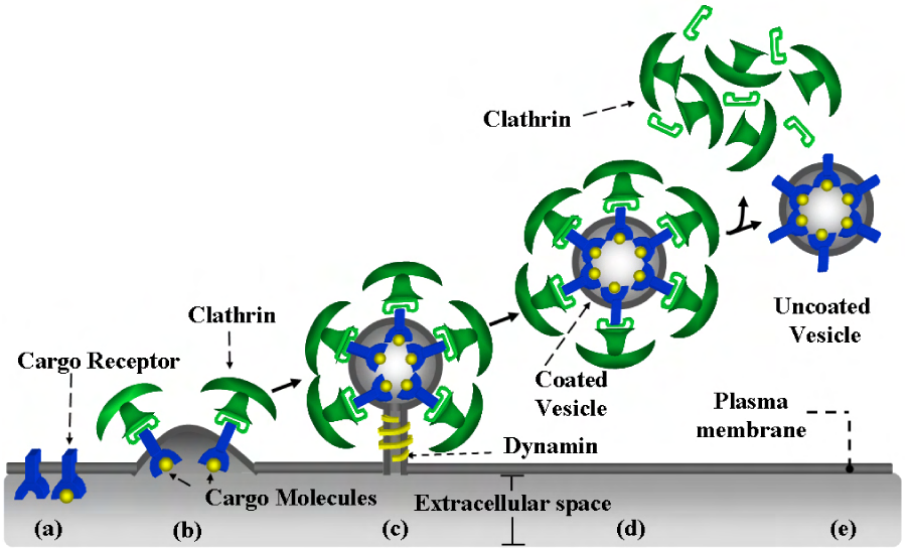
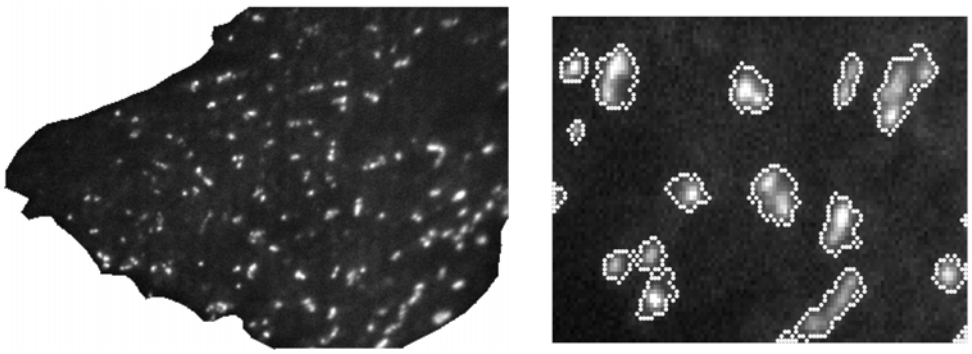


Figure 1.2: **Sequential stages in clathrin-mediated endocytosis**.

the use of fluorescent molecules conjugated to endocytic proteins, as well as the application of specialized microscopy techniques, such as Total Internal Reflection Fluorescence Microscopy (TIRFM).

Multiple proteins mediate this complex process. For instance, the GTPase dynamin has emerged as a crucial mediator of the endocytosis of clathrin-coated vesicles. Dynamin surrounds the neck of clathrin-coated vesicles, and hydrolyzes GTP, resulting in the fission of the vesicle. The precise mechanism of this release is still being debated. 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 (63). It has been reported that a sudden burst of dynamin associated with the coated pit precedes the budding and disassembly of the vesicle (19; 37). An image provided by a microscope is shown in Fig. 1.3 (a). The contour of the cell plasma membrane has been delineated in order to define the region of interest for posterior image processing. Fig. 1.3 (b) displays a zoom of the middle bottom part of the cell. The regions with high fluorescence have been delineated in white color. Each connected component may comprise one or an unknown number of very close and overlapped clathrin-coated vesicles.

In a similar way, clathrin has also been studied in order to characterize the



(a) (b)

Figure 1.3: **Dynamin spots when viewed under TIR-FM**.

endocytosis. 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 (19). An image of a cell expressing clathrin fluorescent protein provided by TIRFM is shown in Fig. 1.4 (a). Fig. 1.4 (b) shows several sub-images of an endocytic event (highlighted with a circular mark) which appears (birth) at time 20 seconds and disappears (death) at time 130 seconds. Fig. 1.4 (c) plots the brightness profile as a function of time of this endocytic spot. Fig. 1.4 (f) displays the segmented endocytic spots after image processing. 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* (as usual in Stochastic Geometry)

From a biological point of view, it is key to characterize endocytic spots by estimating parameters which allow us to know more about the underlying phenomenon of endocytosis, such as the number of endocytic spots per unit area, its mean size (perimeter or area) or its 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 fluorescent areas viewed through the microscope overlap between them forming relative large and irregular regions with high fluorescence, as shown in Fig. 1.3 and 1.4. As the vesicles are smaller

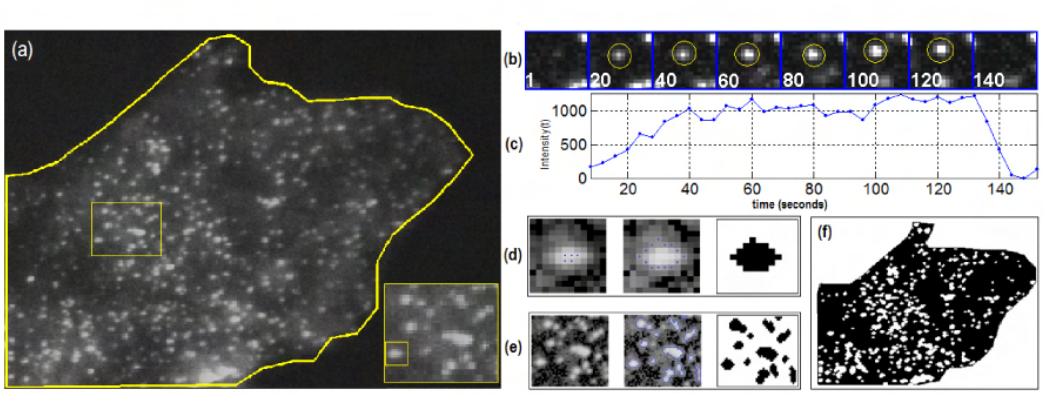


Figure 1.4: **Clathrin spots when viewed under TIRFM**.

than the spatial resolution limit of the microscope, the apparent size of a single vesicle is given by its point-spread function. Where two or more vesicles overlap, we can see objects of irregular morphology in the image, i.e., the clumps (See Fig. 1.3 (b)).

The natural question that arises to the experimenter when analyzing this type of images is: how many endocytic spots are in a given image, especially when they overlap in specific regions? 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. The result is a binary image composed almost solely by these isolated endocytic areas (19). The main drawback to this approach is concerning with the selection procedure. 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. Obviously, 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, or the size of a single endocytic spot (area or perimeter).

1.1.3 Total Internal Reflection Fluorescence Microscopy

Evanescence wave microscopy, also termed total internal reflection fluorescence microscopy (TIRFM), has shed new light on important cellular processes taking place near the plasma membrane. In this way, we can enable the direct observation of membrane fusion vesicles and the movement of single molecules (69). The optical phenomenon of TIR can be observed in everyday life, from fibre optics to sparkling diamonds. The principle based on Snell's law and shown in Fig. 1.5 is as follows: if light travelling in a dense medium strike a less dense medium beyond a certain 'critical angle' θ_c, the light then 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. An important property of the evanescent wave is that intensity falls off exponentially away from the coverslip. Penetration depths lesser than $100nm$ are easily achieved. Electronic Microscopy (EM) has high spatial resolution, but only gives snapshots of the process, by contrast TIR-FM offers a compromise in that good spatial-temporal resolution can be achieved.

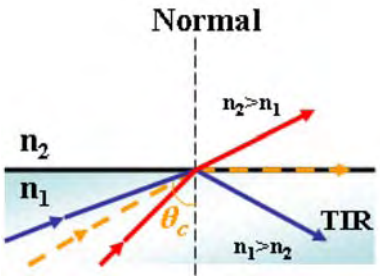


Figure 1.5: **Snell's Law.**

1.2 Stochastic Geometry

Even though the exocytosis and the endocytosis are closely related, their behavior from a statistical point of view is intrinsically different in nature. Exocytosis can be seen as events happening instantaneously in a specific region of the cell, while endocytosis are part of a long process in which events cannot be differentiate

easily from one another, but as a clump of overlapped events. Here we model both types of events using different statistical tools due to either, the source of biological data is physically different and the data that are obtained are also different.

1.2.1 Point Processes to study exocytosis

The concept of random spatial structure is complex, because it refers to both the vertical and horizontal use of space by the elements inside the cell. Each exocytosis is observed at a given location and at a given time. Then, we propose the use of the *Point Processes* theory to analyze the data obtained after image processing of videos of cells showing exocytosis.

A spatial point process is a random mechanism producing locally sets of points in the plane. Let $z_i : i = 1, 2, 3, \ldots$ denote the points of a realization in $\mathbb{R}^2$, a point pattern. We call z_i, the events (fusion events) of the process in $\mathbb{R}^2$. The use of the term 'event' has become standard in spatial point process analysis as a means of distinguishing the location of an observation from any other arbitrary location within the study region. Two important characteristics of a spatial point pattern are stationarity and isotropy, that are directly related with the invariance under spatial shifts and rotation.

Invariance under spatial shifts means that a similar pattern of exocytosis events would have been observed although another view of the cell had been acquired. Invariance under temporal shifts means that the structure of the process has to be unaffected by shifting over the time axis i.e. a similar pattern of exocytosis would have been observed over different periods of time.

The first-order intensity 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

$$\lambda(s) = \lim_{\nu_2(ds)\to 0} \frac{E[N(ds)]}{\nu_2(ds)}, \tag{1.1}$$

where ds is an infinitesimal disk at location s with area $\nu_2(ds)$, E$[\cdot]$ is the expectation operator and $N(ds)$ denotes the number of events in the disk.

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

$$\lambda_2(s_i, s_j) = \lim_{\nu_2(ds_i),\nu_2(ds_j)\to 0} \frac{E[N(ds_i)N(ds_j)]}{\nu_2(ds_i)\nu_2(ds_j)}. \tag{1.2}$$

As we pointed out, a point process is stationary if the process is location invariant. This is equivalent to saying $\lambda(s) = \lambda$ so that the expected number of events at an arbitrary location s is constant for all $s \in A$; and $\lambda_2(s_i, s_j) = \lambda_2(h)$ so that dependence between events depends only on the difference $h = s_i - s_j$, (direction and distance), between s_i and s_j and not on their absolute locations.

Let A be any planar region (the cell membrane), we call $\nu_2(A)$ the area of A, and $N(A)$ the number of events in A. We associate to the i-th event or exocytosis a pair (z_i, t_i) where z_i is the location and t_i the time associated. Much of the theory of spatio-temporal point processes carries over from that of spatial point processes. However, the temporal aspect enables a natural ordering of the points that does not generally exist for spatial processes (59).

A spatio-temporal point process Φ is defined as a a random mechanism producing points in the space at different times.

From now on, the spatio-temporal points (z_i, t_i)'s will be called *events*. These events are thought to be indistinguishable, other than by their times and locations. Our data set is the whole sequence of fusion events $\{(z_i, t_i) : i = 1, \ldots, n\}$ observed for each cell. We will assume that $(z_i, t_i) \in W \times [0, T]$, being W the plasma membrane and $[0, T]$ the time duration of the recorded sequence, hence we are restricted to a specific time interval. In our sequences this time interval corresponds to approximately five minutes. These data are considered as realizations of a spatio-temporal point process, Φ, in such a way that $\Phi(A \times B)$ is the number of points (in our case, number of exocytosis) in the rectangle $A \times B$ with A a Borel subset of $\mathbb{R}^2$, the 2D Euclidean space, and B a Borel subset of $[0, +\infty)$, the time axis.

The probability distribution of Φ is given by

$$\mathcal{P}(\Phi(A \times B) = k), \tag{1.3}$$

for $k = 0, 1, 2, \ldots$, i.e., the probability of finding k points in the planar region A which belong to the set B. The intensity of the point process Φ is a function $\lambda(s, t)$ such that

$$E\Phi(A \times B) = \int_B \int_A \lambda(s, t) ds dt, \tag{1.4}$$

where $E(\cdot)$ denotes the mean of the random variable, in our case, the number of observed exocytosis in $A \times B$. Intuitively, if ds is a small region centered at s

with area $\nu_2(ds)$ and dt is a small time interval centered at t with length $\nu_1(dt)$, then $\lambda(s,t)\nu_2(ds)\nu_1(dt)$ is the probability of an event observed at (s,t).

Φ is called stationary if all properties of the process are invariant under shifts of A and B. More formally, Φ is stationary if

$$\mathcal{P}(\Phi(A \times B) = k) = \mathcal{P}(\Phi((A+h) \times (B+t)) = k), \tag{1.5}$$

where $h \in \mathbb{R}^2$, $t \geq 0$ and $A + h = \{a + h : a \in A\}$.

If Φ is stationary Eq. (1.4) becomes,

$$E\Phi(A \times B) = \lambda\nu_2(A)\nu_1(B), \tag{1.6}$$

where $\nu_2(A)$ is the area of A and $\nu_1(B)$ is the length of B. We have that the function $\lambda(s,t)$ is constant i.e. $\lambda(s,t) = \lambda$, where λ is the mean number of events per unit area and unit time.

Edge effect arises in spatial point pattern analysis when, as is often the case in practice, the region A on which the pattern is observed is part of a larger region on which the underlying process operates. That problem arises in our framework since the microscope only gives us one of the cell sides. The essential difficult is that unobserved events outside A may interact with observe events within A but, precisely because the events in question are not observed, it is difficult to take proper account on this.

There are several ways to avoid the edge effects. We have applied the *adjustment method* described in (14). The adjustment method operates by making an 'on average' adjustment for unobserved events outside A. Hence, if we count the number of exocytosis events observed inside W (plasma membrane), n say, within distance d of a location x, and $a(d)$ denotes the area of intersection between A and a disc of radius d centered on the location x, then a sensible estimate of the actual number of events within distance d of x is $n\pi d^2/a(d)$.

It is natural to consider separately the locations of fusion events and times in such a way that we will have the *marginal* point patterns given by $\Phi_1(A) = \Phi(A \times [0,T])$, i.e., the set of locations corresponding to the whole period of time considered $[0,T]$. Similarly, the second *marginal* point pattern is composed by the times $\{t_i\}_{i=1,\ldots,n}$ without any consideration about the locations, i.e., $\Phi_2(B) = \Phi(W \times B)$. If Φ_1 is (spatially) stationary then $E\Phi(A) = \lambda_1\nu_2(A)$, with λ_1 being the mean number of events per unit area. If Φ_2 is (temporally) stationary then $E\Phi_2(B) = \lambda_2\nu_2(B)$ where λ_2 is the mean number of events per unit time.

In some applications as Geography, it is common the use of regular grids of quadrats on the event distribution in order to obtain the kernel estimation. In-

stead, we could form a count of events per unit area within a moving quadrat or 'window'. We define a window of fixed size and imagine centering this on a number of locations in turn, where these are arranged in a fine grid superimposed over A. We thus obtain estimates of the intensity at each grid point. This produces a more spatially 'smooth' estimate of variation in $\lambda(s)$ than we can obtain by using a fixed grid of quadrats. However, in each of the intensity estimates, no account is taken of the relative location of events within the window and the choice of a suitable window size is not clear. Kernel estimation is a generalization of this idea, where the window is replaced with a moving three-dimensional function (the kernel) which weights events within its sphere of influence according to their distance from the point at which the intensity is being estimated. The method is commonly used in a more general statistical context to obtain smooth estimates of univariate (or multivariate) probability densities from an observed sample of observations.

Fig. 1.6 shows a kernel estimation where color is proportional to the intensity function value, from small values (green) to higher values (yellow, orange and white).

Formally, if s represents a location within A, and s_1,..., s_n are the locations of n observed events, then the intensity $\lambda(s)$ at s can be estimated as

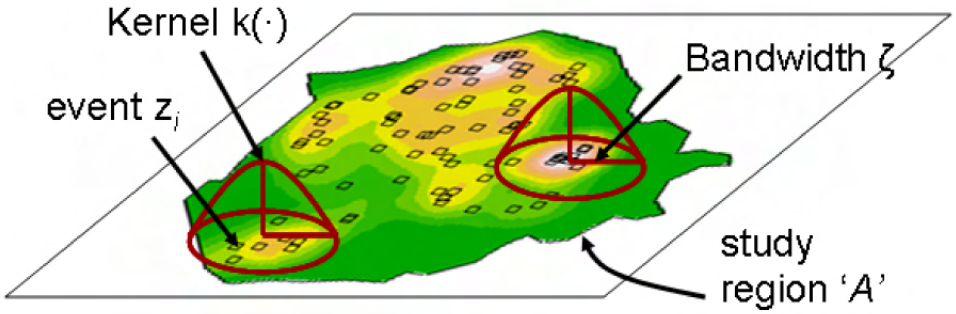
$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right) \tag{1.7}$$

where $k(\cdot)$ represents the kernel weighting function which, for convenience, is expressed in standardized form, i.e., centered at the origin and having a total volume of one under the curve. This is then centered on s and 'stretched' according to the parameter $\tau > 0$, which is referred to as the bandwidth. The value of τ is chosen to provide the required degree of smoothing in the estimate. The kernel estimate $\hat{\lambda}_\tau(s)$ is intended to be sensitive to the choice of bandwidth, τ. As this is increased, there is more smoothing of the spatial variation in intensity; as it is reduced we obtain an increasingly spiky estimate.

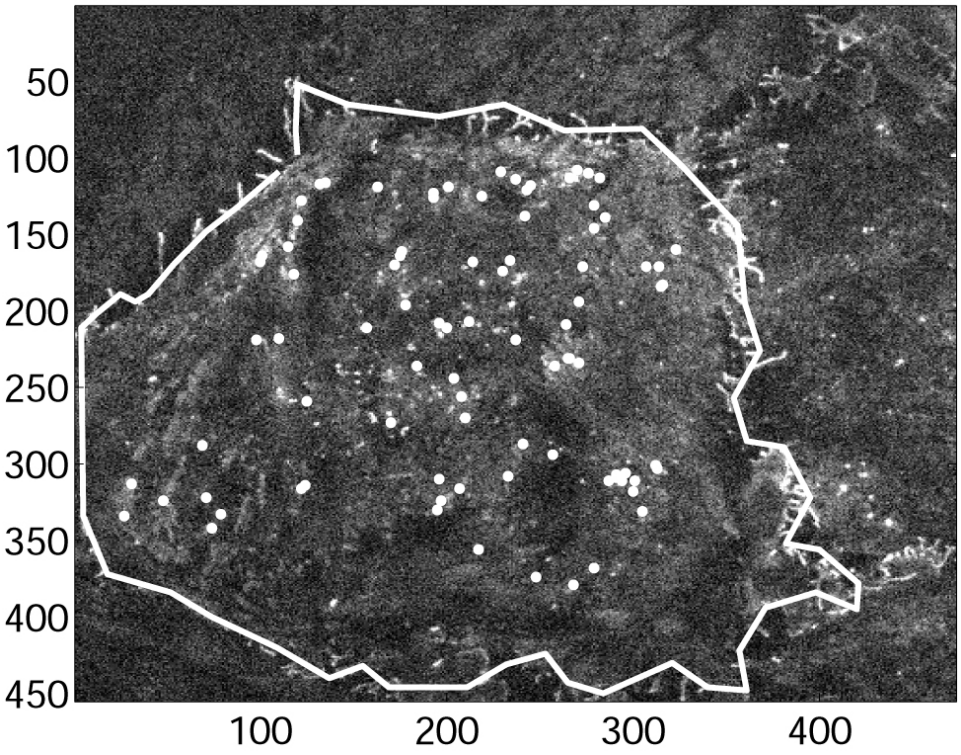
Kernel estimation is considered a good exploratory tool for first-order properties of a point pattern and has been widely used in many context (24). One of the most commonly used kernels is the so-called *quartic kernel* defined as,

$$\hat{\lambda}_\tau(s) = \sum_{d_i \leq \tau} \frac{3}{\pi\tau^2}\left(1 - \frac{{d_i}^2}{\tau^2}\right)^2 \tag{1.8}$$

where d_i is the distance between the point s and the observed event location s_i, and the summation is only over values of d_i which do not exceed τ. The region



(a)



(b)

Figure 1.6: **Kernel estimation of a point pattern. Events correspond to exocytosis observed at any time.**

of influence within which observed events contribute to $\hat{\lambda}_\tau(s)$ is therefore a circle of radius τ centered on s.

A good reference about statistical analysis of spatial point processes is (14) and in the real line (10),(9).

1.2.2 Random Closed Sets to study endocytosis

Until the 1970*s* random sets were only a marginal or exotic part of probability theory. Most papers written mainly by non-mathematicians and widely scattered in the literature, introduced particular models of random sets and some formulae were derived that today are considered as fundamental. This was done without rigorous theoretical base. Nonetheless, the situation changed completely in 1975 when G. Matheron published his fundamental and seminal book (**? ?**). This book has laid the fundamentals of the theory of random closed sets, provided the suitable measure-theoretic machinery and offered the fundamental theorems. Many mathematicians and statisticians have thus been encouraged to study problems of random sets. Various successful applications have led to systematization, simplification, commonness and experience.

First, some basic notation needs to be introduced. If A denotes a Borel subset $\mathbb{R}^2$ then $\nu_2(A)$ and $U(A)$ will denote the area and perimeter respectively of the set A. $A \oplus B = \{a+b : a \in A, b \in B\}$ is the Minkowski addition of the sets A and B. Finally, $\check{A} = \{-a : a \in A\}$ is the symmetric of A with respect to the origin.

Let $\mathcal{F}$ be the class of closed subsets in the Euclidean space $\mathbb{R}$ 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. 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. In fact, the following function characterizes the probability distribution of the random set Φ,

$$T(K) = \mathcal{P}(\Phi \cap K \neq \emptyset) \text{ with } K \text{ any compact subset.} \tag{1.9}$$

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

The Boolean model (also known as the Poisson germ-grain model) is an important and simple type of random closed set. Roughly, a Boolean model could be described as a set of scattered points in the plane, following a stationary Poisson process distribution of intensity λ, where a random shape is placed on every

point. The union of all of these discs is a Boolean model. In this framework, the points of the Poisson process are the *germs* of the model while the discs are the *primary grains*. A particular realization of a Boolean model is shown in Fig. 1.7. The heuristic notion of a clump can be described mathematically as the Boolean model.For the foundation of the theory of random closed sets the reader is referred to (36).

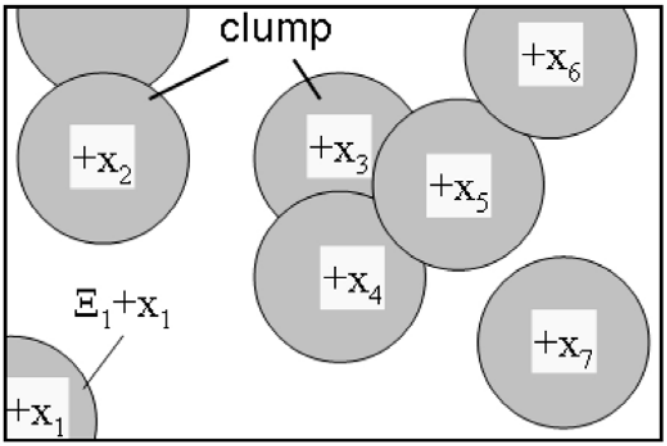


Figure 1.7: **Example of a Boolean model.** Germs appear as x_i while primary grains are Ξ_i.

More formally, let $\Psi = \{x_1, x_2, \ldots\}$ be a stationary Poisson point process in $\mathbb{R}^2$ of intensity λ. Let $\Xi_1, \Xi_2, \ldots$ be a sequence of independent and identically distributed (as Ξ_0) random compact sets in $\mathbb{R}^2$ that are independent from the Poisson process Ψ such that $E\nu_2(\Xi_0 \oplus \check{K}) < +\infty$ for all compact set K. The Boolean model is the random set defined as

$$\Phi = \cup_{n=1}^{\infty}(\Xi_n + x_n). \tag{1.10}$$

The random set Ξ_0 is called the primary grain. As it was previously mentioned, the distribution of Φ is characterized by the capacity functional. This capacity functional for a Boolean model is given by

$$T(K) = 1 - \exp\{-\lambda E\nu_2(\Xi_0 \oplus \check{K})\}. \tag{1.11}$$

Typical possibilities for the primary grains include: discs of random radius, random polygons, segments of random length and orientation. If primary grains

are convex then classical averages of various numerical measures of convex sets play an important part in the theory.

The Boolean model has an important value in the description of given samples and situations in nature. First, a relatively parsimonious description is available for the random set in question, by means of the intensity λ of the germs process and the various characteristics of the grains. Second, it is at least possible that the model assumptions may be suggestive of the process of formation of the structure. We will take advantage of this characteristic to develop further methods in Chapter 4 and Chapter 5. Finally, the formulae to be derived for the Boolean model may be used for estimations of quantities not available for direct measurements.

If we assume that the underlaying point process producing the germs of the Boolean model follows a stationary Poisson process, then the resulting Boolean model is also considered to be stationary. Moreover, if the primary grains have isotropic distribution then the Boolean model is also isotropic in distribution.

Among the basic characteristics of the Boolean model, the volume fraction is of special importance. It measures the mean fraction of volume occupied by Ξ in a region of unit area,

$$p = E(\nu_2(\Xi \oplus B)), \quad \Xi(B) = 1. \tag{1.12}$$

As p does not depend on the choice of the region B when Ξ is stationary, we can rewrite Eq. (1.12),

$$p = P(0 \in \Xi). \tag{1.13}$$

Finally, putting $K = 0$ it is seen that

$$P(0 \in \Xi) = 1 - exp\{-\lambda E(\nu_2(\Xi_0))\} \tag{1.14}$$

1.3 Outline of the thesis

In Chapter 2 a methodology to study different kinds of dependence of a natural stochastic process happening in cells, called exocytosis, is presented. In particular, we test the existence of spatial, temporal and spatial temporal dependence among the events of the process in the context of point processes. Chapter 3 is devoted to the study of the counter process, called endocytosis. Boolean models are used to model the endocytosis observed on the plasma membrane of cells, and several measures are estimated which allows us to better characterize the process. Chapter 4 elaborates on the concepts of the previous chapter to extend the

classical Boolean model to the temporal dimension. This explicit consideration of time allows us to propose new estimators for this spatial temporal stochastic process and to apply them to study the endocytosis process. In Chapter 5 a different method, based on the aggregation of Boolean models, is developed to obtain temporal estimates of a Boolean model in the spatial temporal dimension. This methodology, though different, is based on some of concepts described in Chapter 4. Finally, Chapter 6 gathers the conclusions of the study carried out and presents the future work.

Chapter 2

Spatio-temporal Analysis of Constitutive Exocytosis in Epithelial Cells

2.1 Introduction

Earlier studies suggested that constitutive exocytic fusions sites were randomly distributed in space (56), while other works indicated spatial 'hot spots' for exocytosis (31). As vesicles are transported away from the Golgi via microtubules (68) and have been observed to colocalize near fusion sites, it is possible that cytoskeleton may in part function to specify a domain-specific fusion site (58). Importantly, formal tests were not used to validate biological hypotheses, but rather largely relied on the visual inspection or limited statistical analysis. An in-depth study of the spatio-temporal behavior of exocytosis is lacking and essential for determining if such a correlation exists. Although temporal coupling is well-known from regulated pathway, and in fact defines it, a wide-open question is if similar mechanisms function in constitutive pathways; e.g. with lower constraints. One questions is: Why might this be feasible? Literature states (2) that, a key regulator in regulated fusion is the elevation of cytoplasmic calcium. When its concentration is raised it will trigger the release of many primed vesicles, albeit to different extents in various cell types. However, it is worth to study if spatial and temporal coupling exists in constitutive pathways. Hypo-

thetically, a local elevation of calcium could increase the probability that vesicles in the vicinity are released. Of course, as by definition this constitutive pathway occurs to some extent all the time (e.g. it is not tightly regulated), without the direct imaging of putative spatio-temporal coupling, as has been done here, this biologically relevant issue would be virtually impossible to address.

Electrophysiological studies of exocytosis, while having a high temporal resolution that have revealed multiple vesicle pools, provide practically no spatial information. In contrast, confocal microscopy has been used to examine the kinetics of constitutive secretion (28); although arguably fast processes and single vesicle behaviors can be missed.

TIRFM was used in the present study to obtain spatial and temporal data of exocytosis and test our hypothesis of putative temporal, and/or spatio-temporal coupling (it has been previously demonstrated that spatial coupling exists (31)). To our knowledge no formal tests have been proposed in the biological field to examine these complex data sets (70), (32), (12),(27), (69). The methods we propose and validate here will provide new and better tools for testing biological hypotheses of spatio-temporal interdependencies associated with membrane trafficking and fusion.

In this study, we first considered the number of exocytosis events per unit area and time for the different cells; i.e., the observed spatio-temporal intensities. We then tested to see if a common spatio-temporal intensity could be assumed by comparing rates between the observed counts and the spatio-temporal volume by using generalized linear models (GLM) (1).

Secondly, the null hypothesis of a completely random spatial distribution of events in the plasma membrane was tested. The rejection of this null hypothesis will indicate if a more aggregated or regular distribution should be considered. Specifically, we can address whether the events are randomly distributed or whether there is spatial coordination. While it was previously assumed to be random, recent reports suggest that microtubules may play an important role in positioning exocytic fusion sites (56). Such putative connections can be essential for, as a cell migrates, exocytosis can be spatially repositioned to the leading edge (56). Understanding this process is also essential in diseases such as cancer where one needs to elucidate the cellular mechanisms that alter cell migration during metastasis. Clearly, if the cell already has a means of restricting spatial delivery, this need only be repositioned during migration rather than a new molecular machine being created. In order to examine the spatial arrangement of fusions the $\mathcal{K}$-function, the empty space function $\mathcal{F}$-function and the nearest neighbor function $\mathcal{G}$-function were used (14).

Thirdly, we examined whether the times of occurrence are uniformly and independently distributed (i.e. if they can be considered realizations of a one dimensional temporal Poisson point process) or exhibit some degree of clustering; the latter would imply that there is an underlying mechanism that causes waves of exocytosis. We note that we would expect a constant rate of exocytosis over times of our GFP tagged temperature sensitive marker, vesicular stomatitis virus glycoproteins (VSVG-GFP) (see Appendix A.2 for details), which is released as a wave from the endoplasmic reticulum and eventually exocytosis at the plasma membrane; however, it should have a single 'hump' unless some part of the biological process is rate limiting (as our new data suggests).

Lastly, we examined if there is a dependence between the spatial locations and the times of the fusion events. The null hypothesis tested whether the random mechanisms producing the locations and the times are independent. In this case, the alternative hypothesis was the spatio-temporal clustering, where the fusions which are close in spatial position also occur close in time. This hypothesis was tested by using a temporal extension of the $\mathcal{K}$-function. Monte Carlo tests enabled us to test all the null hypotheses previously mentioned.

The results that we present here offer direct and quantitative evidence that in constitutive exocytosis the spatial and temporal processes are not linked, but that we do see a surprising and novel temporal coupling of this process with exocytosis events distributed in a double 'humped' camel shaped pattern.

In order to test these null hypotheses we need a stochastic model of these complex data. Each exocytosis is observed at a given location and at a given time. Our primary information is the set of spatial locations marked with the occurrence times (or the occurrence times marked with the locations). An example of a spatio-temporal point pattern of fusions observed can be seen in Fig. 2.1. Dots represent the locations of fusions and colors represent the time of occurrence. The colorbar represents the time axis, from 0 sec (dark blue) to end time (dark red). See supplementary material Videos 2.1 and 2.2 (more details in appendix C). A set of temporally referenced points is called *a spatio-temporal point pattern*. *A spatio-temporal point process* is a random mechanism producing a set of points in the plane or in a bounded subset of the plane (as the cell membrane) at random times.

Section 2.2 is devoted to the experimental setup used to obtain the image sequences and the image processing algorithm used to detect fusions. Section 2.3 describes the statistical analysis of the spatio-temporal point patterns observed for three cells. Firstly, counts of fusions per unit area and unit frame are compared among cells in Section 2.3.1; Secondly, the locations (without times) are analyzed

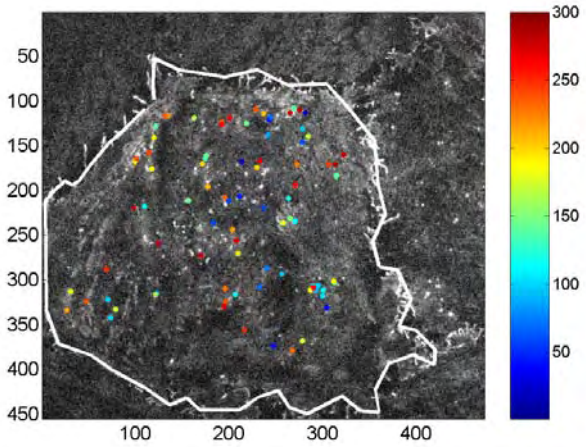


Figure 2.1: **Spatio-temporal point pattern of fusions observed for one cell.**

in Section 2.3.2; Thirdly, the times (without locations) are analyzed in Section 2.3.3 and finally, the spatio-temporal clustering is studied in Section 2.3.4. In Section 2.4 the results obtained are discussed and conclusions are given.

2.2 Material and methods

2.2.1 Image processing

Exocytosis is characterized by a unique signature when it is imaged by TIRFM (69). Firstly, the brightness of a small, spatially localized region in the image increases as the dye spreads symmetrically from the center of the vesicle, the bright area grows and the intensity decreases until complete disappearance. That behavior is illustrated in Fig. 2.2 (a). Exocytosis appears as symmetric and round *bells*, when the grey level image is represented as a function in a 2D space. A fluorescent vesicle is shown before, during and after fusion. As it enters the evanescent field and approaches the lower membrane it becomes exponentially brighter (see 3D intensity profile). Upon membrane fusion, the brightness of a small, spatially localized region increases. Later, as the dye continues to spread, the diameter of the bright area grows, yet its intensity diminishes. This "flash" follows a charac-

teristic spatio-temporal signature. See supplementary material videos 2.3 and 2.4 (more details in appendix C). Fig. 2.2 (b) shows a grey-level intensity profile as a function of time. Three temporal phases can be distinguished: stationary, rise and spread phases. During the *stationary phase*, the total intensity was constant corresponding to docked vesicles. During the *rise phase* the total intensity rises rapidly. Finally, during the *spread phase* the total intensity decreases exponentially, indicating that the cargo is diffusing in the plane of the plasma membrane (for a transmembrane marker).

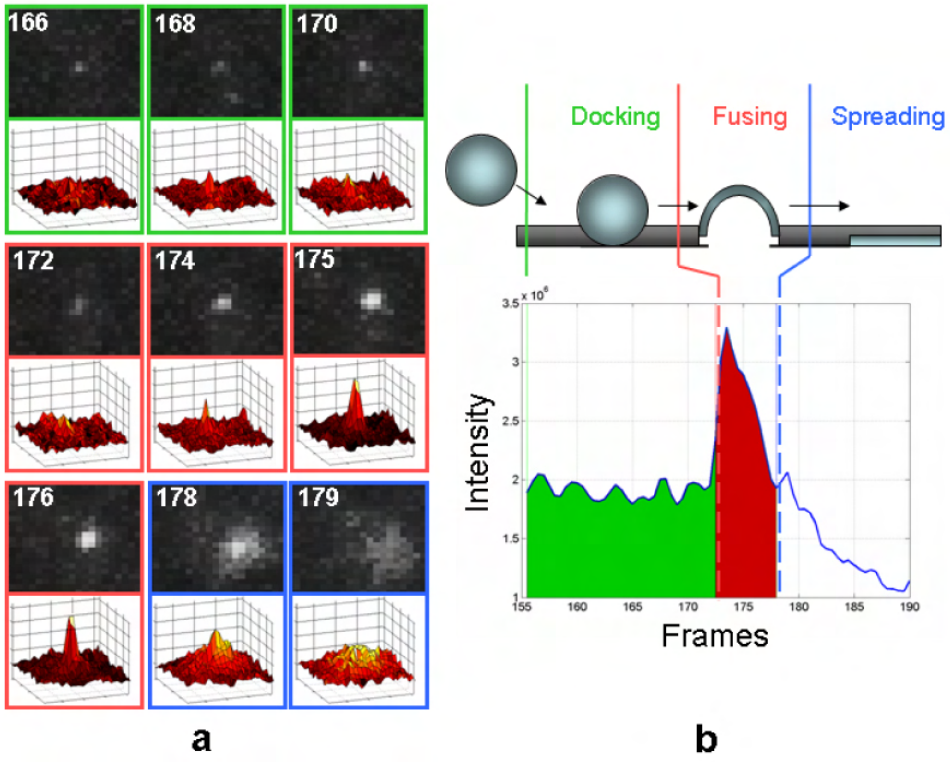


Figure 2.2: **Example of a vesicle fusing with the membrane detected by TIRFM.**

In general, four kinds of objects can be observed in the images: exocytotic events (fusions), lateral moving vesicles, emerging vesicles and static vesicles (see Table 2.1) (56), (31). Fusions can be distinguished from moving, emerging or static vesicles by taking into account that the subtraction of two consecutive images of the sequence is not null, that their grey level intensity spreads sym-

metrically and that their grey level intensity decreases over time. Taking heed of this behavior, the following algorithm was designed.

Table 2.1: Type of objects and its behaviors

Object	Image difference	Symmetry	Intensity decrease
Exocytotic events	not null	yes	yes
moving vesicles	not null	no	yes
emerging vesicles	not null	yes	no
static vesicles	null	yes	no

Step 1: Temporal subtraction. Forward image subtraction of consecutive frames allows the detection of regions whose pixel brightness have changed (25). See Fig. 2.3 (a).

Step 2: Thresholding the subtracted image. After temporal subtraction of consecutive frames, only those points where the changes in intensity are higher than a given threshold T_h are considered as fusion candidates. T_h was set to 15. See Fig. 2.3 (b).

$$S(x,y,n) = \begin{cases} 1 & |I_{t+1}(x,y) - (I_t(x,y)| > Th \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Step 3: Area opening. An area opening transform, $\gamma_{A_{min}}$ (65), is applied to the thresholded image. All the connected components with a surface area (pixel count) less than a certain threshold A_{min} are removed. A_{min} was set to 20. See Fig. 2.3 (c).

$$\gamma_{A_{min}} = \bigvee_i \{\gamma_{B_i} | B_i \text{ is connected and } |B_i| = A_{min}\}. \tag{2.2}$$

The value of A_{min} can change from one image sequence to another, so this a tuning parameter.

Step 4: Top-Hat and Regional maxima. A Top-Hat operator was applied to the original grey level image and regional maxima of the Top-Hat image were calculated to obtain the centers of fusions. A disk of radius of 10 pixels was used as the structuring element. See Fig. 2.3 (d).

Step 5: Image reconstruction. Maxima obtained in Step 4 were used as seeds to reconstruct the area opened image obtained in Step 3. In this way, we could discriminate between fusions and moving vesicles (61), (54). If an object in the opened image had corresponded to a moving vesicle, then its center would have been slightly displaced with respect to its corresponding difference image, and therefore it would not have appeared in the reconstructed image. See Fig. 2.3 (e).

Step 6: Fusion recognition by quantifying intensity decrement and symmetry. So far, we have discarded moving vesicles and static vesicles. In order to reject emerging vesicles, vesicles that suddenly appear in the image, we use the intensity profile. Emerging vesicles exhibit a constant intensity profile through time, whereas fusions undergo an intensity decrement, as shown in Fig. 2.2 (b). The brightness in the neighborhood of the fusion could be fit to a symmetric 2-dimensional Gaussian kernel centered at the point y where the fusion appears by using

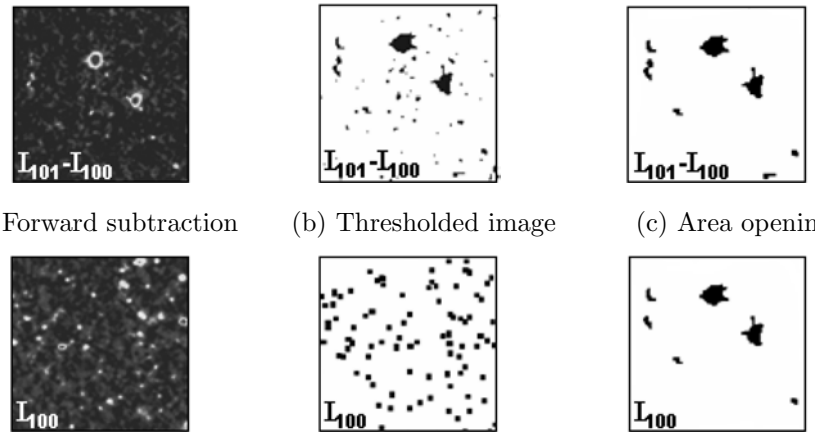
$$I(x) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp{(x-y)^t \Sigma^{-1}(x-y)}. \tag{2.3}$$

The total intensity of the fusion located at y is defined as the integral of the function $I + I_{bg}$ in the circle centered at y with radius 8, where I is defined in Eq. (2.3) and I_{bg} is the mean intensity of the background. The background intensity is calculated as in (73). Note that a circular Gaussian kernel has been used, i.e. $\Sigma = \sigma^2 I_{2\times 2}$ where $I_{2\times 2}$ is the identity matrix.

The output of the algorithm is the location and time of occurrence for each fusion. The performance of the algorithm has been tested by comparing the fusions detected with those observed manually in the three cells analyzed. The method detects on average 86.2 out of 100 true fusions. The specificity was on average 65 out of 100. In this study an automatically detection has been applied which was afterwards manually revised.

2.2.2 Material description

For the image analysis three different video sequences infected with a marker called $VSVG - SP - YFP$ were used. Table 2.2 gives a summary of the image sequences. The size is measured in pixels, being a pixel 180 nm, and time is measured in frames. Each one of the cells was recorded and stored as a different video. Images were acquired at 2 frames per second. Summary on table



(a) Forward subtraction (b) Thresholded image (c) Area opening

(d) Top-Hat image (e) Regional maxima (f) Image reconstruction

Figure 2.3: **Different steps of the detection algorithm.**

2.2 describes the cell number, which is used throughout the study, the kind of fluorescence marker used to image the fusion, the size in pixels of the images, the area of the cell, and finally in the column T, the number of frames for each sequence.

Table 2.2: Description of the image sequences

Cell	**Marker**	**Size**	$\nu_2(W)$	T
Cell 1	(VSVG3-SP-YFP)	409×334	74276	750
Cell 2	(VSVG3-SP-YFP)	407×320	71151	500
Cell 3	(VSVG3-SP-YFP)	455×608	117003	300

2.2.3 Statistical analysis

First, let us describe the data obtained by using the image processing algorithm. We observe, for every cell, different exocytosis (events) occurring in the plasma membrane. In our analysis, each event or exocytosis is defined by its location (where it is observed) and its occurrence time (when it is observed). Let x_i be the location of the i-th fusion and t_i its occurrence time. Let W denote the plasma membrane observed. Each image sequence lasts from an initial time 0

to a final time T (approximately five minutes in our cells). If n exocytosis are observed at the plasma membrane W during the time interval $[0, T]$, our data are then the set $\phi = \{(x_i, t_i)\}_{i=1,\ldots,n}$, a spatio-temporal point pattern. The different spatio-temporal point patterns associated to each cell are analyzed in Section 2.3.4.

If only the locations x_i's are considered then we have a *spatial point pattern* defined over the plasma membrane. If we focus only on the observation times t_i's, we are then concerned with a temporal point pattern. These two point patterns are called the *marginal point patterns* and are analyzed in Sections 2.3.2 and 2.3.3, respectively.

As was mentioned in Section 2.1, the hypotheses to be tested are: first, if only the spatial locations $\{x_i\}_{i=1,\ldots,n}$, are considered, is there some kind of spatial clustering? Second, if only the times $\{t_i\}_{i=1,\ldots,n}$, are analyzed, is there temporal clustering? Finally, is there spatio-temporal clustering? To answer these questions, we need a probabilistic model in order to formulate and test these biological hypotheses formally. In Section 2.3.4 the spatial and temporal information are jointly considered and analyzed.

A spatial point pattern is considered as the outcome or realization of a spatial point process, i.e. a random mechanism producing points in the space. The set of times is a realization of a point process in the one-dimensional Euclidean space. Finally, a spatio-temporal point pattern is a realization of a spatial point process, in which points in the space are marked with random times (14), (9), (24).

Let Φ denote a spatial point process then, for a subset A of the 2D Euclidean space $\mathbb{R}^2$, $N(A)$ will denote the number of events (fusions) of the point process in A. For a spatial point process, the intensity function is defined as

$$\lambda(s) = \lim_{\nu_2(ds)\to 0} \frac{\Phi(ds)}{\nu_2(ds)}, \tag{2.4}$$

where ds is a small region centered at point s with area $\nu_2(ds)$, i.e. $\lambda(s)\nu_2(ds)$ is the probability that one exocytosis would be observed at s. If we assume that the spatial point process is stationary, its probability distribution is invariant against shifts, i.e. a similar pattern of exocytosis events would have been observed if another view of the cell had been acquired, thus the intensity function is constantly equal to λ and can be interpreted as the mean number of fusions per unit area. Note that cells were not excited, therefore we expect the same behavior from the cells every time whenever the cells were alive. In addition, the temporal window was long enough, more than 5 min. In our opinion, it seems that to assume a stationary spatiotemporal point process is reasonable.

A good functional descriptor of spatial arrangement is the $\mathcal{K}$-function defined as

$$\mathcal{K}(s) = \frac{1}{\lambda}\mathrm{E}[\text{number further events within distance } s \text{ from an arbitrary event}]. \tag{2.5}$$

For aggregated patterns, each event is likely to be surrounded by further fusions and, for small values of s, $\mathcal{K}(s)$ will be relatively large. In Fig. 2.4 (a) the fusion locations of a whole image sequence are displayed and the plasma membrane is delimited by a yellow line. Fig. 2.4 (b) illustrates the concept of Empty space distances, which are the set of distances from every point of a regular grid (white dots) to the nearest fusion (black dots). Another kind of distance measurement, the nearest neighbor distance, is shown in Fig. 2.4 (c), and comprises the set of distances from every fusion to the nearest fusion observed. Finally, Fig. 2.4 (d) shows how to calculate the $\mathcal{K}$-function. This function is obtained by locating a disk or radius s over each fusion and counting the number of further fusions (without the center) inside the disk. The count is done for every fusion. If fusions are regularly spaced, each one is likely to be surrounded by empty space and therefore, at small distances, $\mathcal{K}(s)$ will be relatively small.

The $\mathcal{K}$-function can be estimated by means of the estimator proposed by Ripley (14) given by

$$\hat{\mathcal{K}}(s) = \frac{\nu_2(W)}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} I(d_{ij} \leq s), \tag{2.6}$$

where $I(\cdot)$ denotes the indicator function 1 if the comparison is true, and 0 otherwise. A weight w_{ij} is included because of the edge effects. Figs. 2.4 (e) and (f) show an example in which edge effects are observed. Fig. 2.4 (e) shows a hypothetical set of fusions observed on a 3D cell. The same fusions after cell flattening are displayed in Fig. 2.4 (f). Then a border correction adjustment has to be performed since the fusion count inside the disk with no border correction would be two, i.e. two fusions would be missed.

The explanation is that for a given fusion within distance s of the boundary of W (the part of the plasma membrane we observe through the microscope), the observed count of other fusions within distance s necessarily excludes any fusions which may have occurred within distance s but outside the observed W, since we can only image one cell surface side. Naturally, TIRFM only reveal fusion events on the lower side of the cell near the glass and any events that occur on the upper side of the cell will be undetected. Several methods have been proposed to correct this bias. The Ripley estimator has been used, where w_{ij} is the proportion of the

area of the disk with center x_i and radius $d_{ij} = \nu_2(x_i - x_j)$ (Euclidean distance between x_i and x_j) which lies within W.

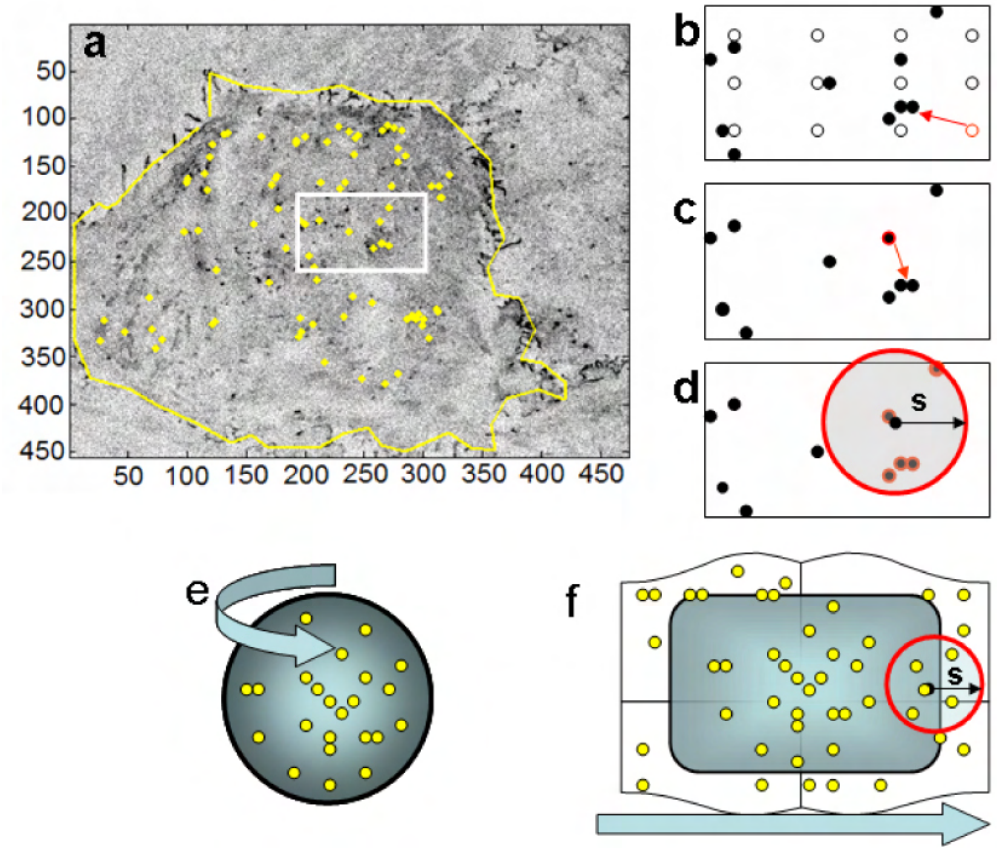


Figure 2.4: **Map of fusions along with explanation of $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{K}$-functions.**

In addition to the $\mathcal{K}$-function previously defined, two other usual functional descriptors of spatial point patterns will be used. They are based on the measure of different kinds of distances between points. First, the *empty space distance* E is the distance from an arbitrary fixed point in the plasma membrane to the nearest fusion. See Fig. 2.4 (b). Second, the *nearest neighbor distance* D is defined as the distance from a fusion (a point of the point pattern) to the closest of the other fusions (points of the pattern), that is, the closest neighbor. See Fig. 2.4 (c). For a given point pattern, a set of distances (nearest neighbor or empty space

distance) is summarized by means of their corresponding cumulative distribution functions. $\mathcal{G}(s) = P(D \leq s)$, is the cumulative distribution function of D and can be estimated by,

$$\hat{\mathcal{G}}(s) = \frac{\#(d_i : d_i \leq s)}{n}, \tag{2.7}$$

where n is the number of fusions, $d_1, \ldots, d_n$ are the nearest neighbor distances observed from each fusion to its closest fusion and # stands for 'number of', i.e., for each distance s, $\hat{\mathcal{G}}(s)$ is the proportion of distances lesser than or equal to s. For the empty space function, $\mathcal{F}(s) = P(E \leq s)$ will be estimated by means of

$$\hat{\mathcal{F}}(s) = \frac{\#(d_i : d_i \leq s)}{m}, \tag{2.8}$$

where $d_1, \ldots, d_m$ are the distances from the m points of a lattice to the nearest fusion. These functions are estimated by using the point pattern observed for each cell.

Let us now consider the spatio-temporal point process, i.e. random locations marked with random times. If Φ denotes the spatio-temporal point process then, as we introduced before in the spatial case, we can define the intensity as

$$\lambda(s,t) = \lim_{\nu_2(ds),dt \to 0} \frac{\Phi(ds \times dt)}{\nu_2(ds)dt}, \tag{2.9}$$

where $\Phi(ds \times dt)$ is the number of events located in $ds \times dt$, a small region centered at s with area $\nu_2(ds)$, and dt is a small time interval centered at t with length dt. Note that $\lambda(s,t)\nu_2(ds)dt$ could be interpreted as the probability of an event observed at (s,t), i.e. at location s and at time t.

Φ is called stationary if all properties of the process are invariant under spatial and temporal shifts. Since we are considering constitutive exocytosis on non-polarized PtK_2 cells and the temporal window is long enough (more than 5 min), the hypothesis of stationary spatio-temporal point process is tenable. If Φ is stationary, then the function $\lambda(s,t)$ is constant, i.e. $\lambda(s,t) = \lambda$, where λ can be interpreted as the mean number of events per unit area and unit time.

We propose to describe the spatio-temporal point pattern of exocytosis observed by using an extension to the temporal dimension of the $\mathcal{K}$-function, which is defined as

$$\begin{aligned}\mathcal{K}(s,t) = \frac{1}{\lambda}\mathrm{E}[&\#\text{ of further events within distance } s \\ &\text{and time } t \text{ of an arbitrary event}].\end{aligned} \tag{2.10}$$

In order to test the null hypothesis of independency between times and locations of fusions, i.e. the existence of an underlying clustering mechanism which may regulate the spatial and temporal behavior of exocytosis, we propose to use function $\mathcal{K}(s,t)$. If the locations and times are independent then it holds,

$$\mathcal{K}(s,t) = \mathcal{K}_1(s)\mathcal{K}_2(t), \tag{2.11}$$

where, $\mathcal{K}_1(s)$ is the *marginal* $\mathcal{K}$-function corresponding to the spatial locations for any event in the whole period of time considered $[0,T]$; and $\mathcal{K}_2(t)$ is the second *marginal* $\mathcal{K}$-function associated to the times $\{t_i\}_{i=1,\dots,n}$ for any event located within W. Let us denote the spatial and temporal intensities by λ_1 and λ_2, respectively. We have the following relations between the spatio-temporal intensity λ and the corresponding spatial and temporal intensities, $\lambda_1 = \lambda T$ and $\lambda_2 = \lambda\nu_2(W)$.

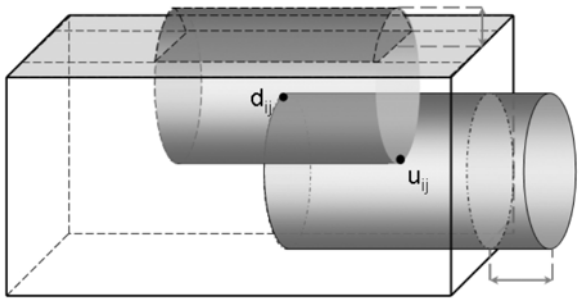


Figure 2.5: **Edge correction for the spatio-temporal** $\mathcal{K}(s,t)$.

In this context the spatial $\mathcal{K}$-function has to be re-interpreted as

$$\begin{aligned}\mathcal{K}_1(s) = \frac{1}{\lambda_1}\mathrm{E}[&\#\text{ of further events within distance } s \\ &\text{and time } t \text{ from an arbitrary event}],\end{aligned} \tag{2.12}$$

and the temporal $\mathcal{K}$-function as

$$\mathcal{K}_2(t) = \frac{1}{\lambda_2}\mathrm{E}[\#\text{ of further events in } W \text{ within time } t \text{ of an arbitrary event}]. \tag{2.13}$$

The estimator of $\mathcal{K}_1(s)$ was already given in Eq. (2.12), so now we give the estimators for the temporal and spatio-temporal $\mathcal{K}$-functions. The temporal $\mathcal{K}$-

function is estimated by

$$\hat{\mathcal{K}}_2(t) = \frac{T}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} v_{ij} I(u_{ij} \leq t), \tag{2.14}$$

where $u_{ij} = \nu_2(t_i - t_j)$ and v_{ij}= 1, if $t_i - u_{ij} > 0$ and $t_{ij} + u_{ij} < T$, otherwise v_{ij}=2. Finally, the estimator for the spatio-temporal $\mathcal{K}$-function is

$$\hat{\mathcal{K}}(s,t) = \frac{\nu_2(W)T}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} I(d_{ij} \leq s) v_{ij} I(u_{ij} \leq t). \tag{2.15}$$

The Ripley estimator was again used to correct the edge effects. The edge-correction weights are defined by the proportion of the surface area of a cylinder centered at d_{ij}, passing through u_{ij}, and inside the study region $W \times T$ (See Fig. 2.5).

2.3 Results

2.3.1 Comparing the intensities of the different cells

The first question to be faced is the fact that we do not have a unique spatio-temporal point pattern to be analyzed. We actually have three different point pattern, each one corresponding to a different cell. We intend to evaluate if a common spatio-temporal intensity can be assumed for the different cells analyzed. Note that an analysis of variance (ANOVA) is not appropriate in this case. Three rates (counts per unit area and unit time) have to be compared. The natural statistical framework for this problem is the generalized linear model. The question to answer is: Can we consider that the observed sets of fusions for each cell are realizations of a given point process with a common spatio-temporal intensity (i.e with the number of events per unit area and time)? We assume this to be a legitimate hypothesis since all cells were treated and analyzed in a similar way in the same set of experiments.

We compared the spatio-temporal intensities estimated for the three cells. The estimated intensity λ_i of the i-th cell is given by

$$\hat{\lambda}_i = \frac{n_i}{\nu_2(W_i)T_i}, \tag{2.16}$$

where n_i is the number of events observed, $\nu_2(W_i)$ denotes the area of the plasma membrane observed and T_i the total observation time. We have a different spatio-temporal sampling window for every cell. Table 2.3 shows the areas of these windows (second column); the observation times in seconds (third column); the numbers of fusions observed, i.e. the counts (fourth column) and finally, the fifth column gives the intensities estimated. Note that the intensities are rates between the counts (the number of fusion events) and the product of the area of the plasma membrane by the total time, i.e. the size of the spatio-temporal sampling window, given in Eq. (2.16).

Table 2.3: Description of the point patterns observed

Cell	$\nu_2(W)$	T	No. events	$\hat{\lambda}$
Cell 1	74276	750	171	0.0000030
Cell 2	71151	500	260	0.0000073
Cell 3	117003	300	89	0.0000025

First, a probability distribution for the response has to be chosen for which a negative binomial distribution was used. Second, a transformation of the mean of the response variable has to be related as a linear combination of the independent variable. In our case, the only independent variable was the experimental factor indicating the cell in which the count was observed.

The mean of the i-th count μ_i is assumed to verify

$$\log\left(\frac{\mu_i}{\nu_2(W_i)T_i}\right) = \alpha + \sum_{j=1}^{2} \beta_j S_j, \tag{2.17}$$

where S_j with $j = 1, 2$ are the dummy variables indicating the corresponding cell. $S_j = 1$ for the j-th cell for $j = 1, 2$. The third cell corresponds to $S_1 = S_2 = 0$. The standard deviation was adjusted in such a way that the residual deviances observed were completely acceptable, i.e. it was a suitable model.

Finally, the null hypothesis of a common intensity for the three cells, that is, $H_0 : \beta_1 = \beta_2 = 0$ was tested. The statistic had a chi-square distribution with 2 degrees of freedom and an observed value of 1.5692 with a corresponding p-value of 0.45. It is clearly non significative. In summary, we cannot reject the null hypothesis of a common intensity for the three cells. We also note that extensive data (more than 1500 seconds) were acquired, providing many fusion events for analysis (more than 500 fusions).

The statistical analysis was done by using the *R* package (49). A short version of the functions used for the algorithm can be found in Appendix B.

2.3.2 Spatial clustering

Let us consider now only the fusion locations observed. All the points observed for each cell were considered without any information about their observation times.

First, it is necessary to evaluate if the observed fusion locations observed were independently and randomly distributed and i.e. if it can be considered as a realization of a Poisson point process. This model formalizes the concept of points located without any kind of interaction between them and with the environment, or on the contrary, whether there is an underlying clustering mechanism which gives rise to preferred sites for exocytosis. The mechanisms for restricted delivery may be explained by the existence of cell structures such as microtubules or heterogeneity in the plasma membrane in relation to the distribution of the protein machinery. A visual inspection of Figs. 2.6 (a), 2.8 (a) and 2.9 (a) suggests that more aggregated patterns (than those provided by a Poisson point process) could be expected. Clusters of exocytosis, 'hot-spots', are observed corresponding to regions with high intensity values (high number of exocytosis per unit area). Color is proportional to the intensity function value, from small values (green) to higher values (yellow, orange and white).

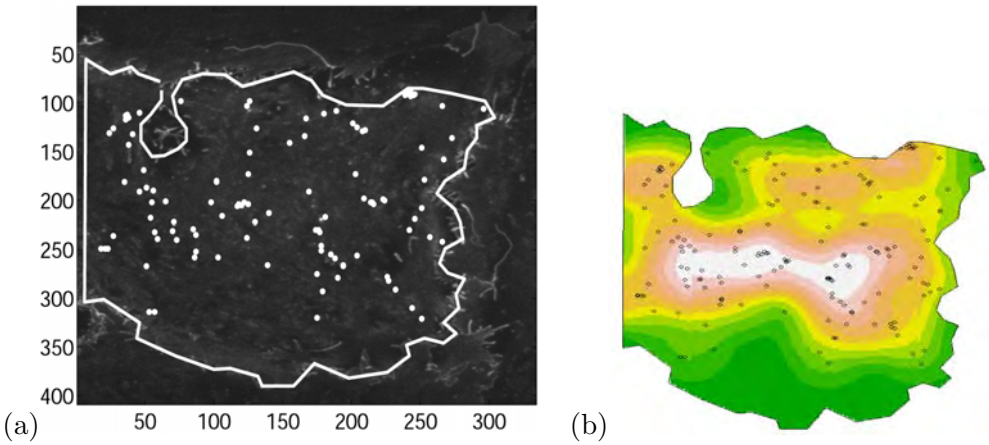


Figure 2.6: **Fusion sites of Cell** 1.

The first step was to evaluate whether the different point patterns can be considered as independent realizations of (possibly with different intensities) Poisson point processes. The intensity function $\lambda(\cdot)$ defined in Eq. (2.4) for the different cells was estimated. Figs. 2.7 (b), 2.8 (b) and 2.9 (b) show the estimated intensity functions for the three cells. It is clear that a clustered spatial point pattern is observed, corresponding with regions with high intensity values.

Using the functions $\mathcal{K}$, $\mathcal{F}$ and $\mathcal{G}$, we test if a Poisson point process is a reasonable model for our three spatial point patterns. In a previous work (31) we stated that fusion events seem to cluster forming 'hot-spots' by means of the $\mathcal{G}$-function which is based on the nearest neighbor distance. Here we use two further functional descriptors based on different distances than $\mathcal{G}$-function. Specifically, based on the empty space distance, $\mathcal{F}$-function, and on the number of neighbors, $\mathcal{K}$-function, in order to compare with previous results.

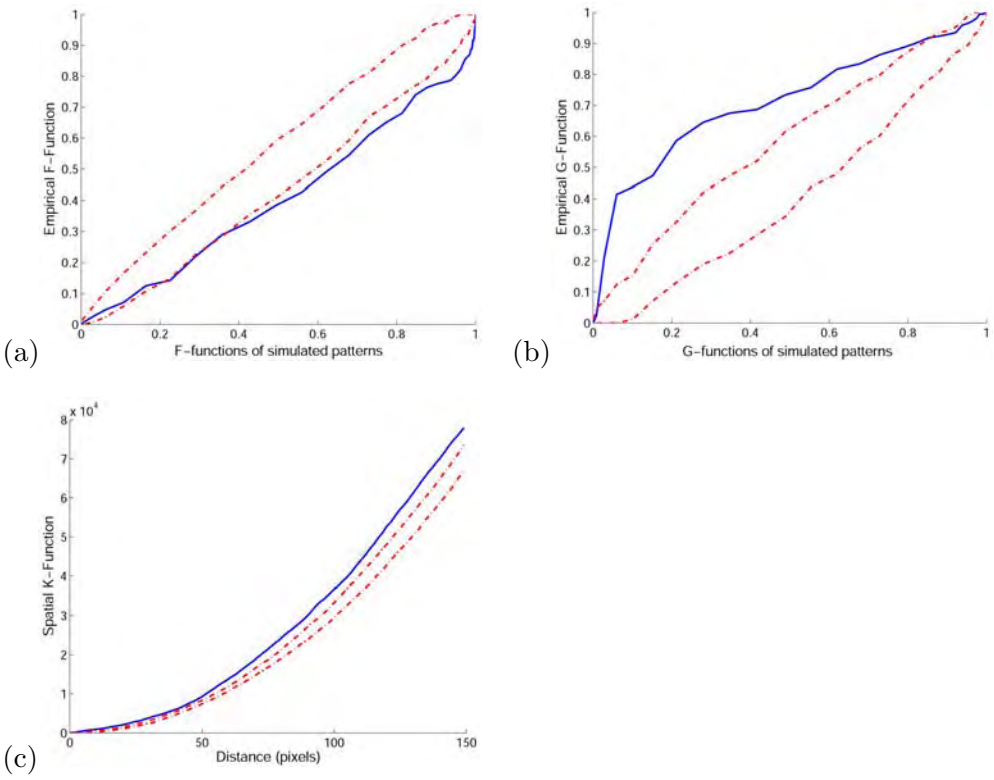


Figure 2.7: **Spatial analysis of Cell 1 by using $\mathcal{F}$, $\mathcal{G}$, $\mathcal{K}$ functional descriptors**.

A Monte Carlo test was used to test the null hypothesis. Let us see a short explanation of the Monte Carlo procedure using the empty space function, $\mathcal{F}$. Analogously, it can be applied to the other two functional descriptors, $\mathcal{G}$ and $\mathcal{K}$.

In this Monte Carlo test, a given spatial point pattern is described by means of the empty space function. Let $\mathcal{F}_0(t)$ be the cumulative distribution function of the empty space function for the point pattern observed. Using the estimated intensity, different Poisson point processes are generated using this intensity over the same window. Let $\mathcal{F}_1(t)$, $\mathcal{F}_2(t)$, ..., $\mathcal{F}_s(t)$ be the distribution functions corresponding to the different generated point patterns (s simulations, in our case $s = 99$). Let L and U be the functions given by $L(t) = \min_{i=1...s}(\mathcal{F}_i(t))$ and $U(t) = \max_{i=1...s}(\mathcal{F}_i(t))$, the lower and upper envelopes of the distribution functions estimated from the simulated patterns. The region delimited by both envelopes quantifies the variability associated with the model considered. If $\mathcal{F}_0$ is contained between the envelopes then the observed pattern corresponds to a Poisson point pattern. Otherwise, the point pattern is not random.

Figs. 2.7 (a), 2.8 (c) and 2.9 (c), plot the observed empty-space functions $\mathcal{F}_0$ (solid line) against the average value of $\mathcal{F}-function$'s of the 99 simulated random patterns. Upper and lower envelopes of the $\mathcal{F}-function$'s calculated from the simulated patterns are also shown (dashed and dotted lines, respectively). If we look at Figs. 2.7 (a) we can clearly distinguish that the observed function deviates from the straight line (bisectrix). The higher this deviation is, the more different the observed pattern from the simulated random patterns is. The range of distances used is from $s_{min} = 0$ to one third of the diameter of the cell, $s_{max} = diameter(W)/3$, approximately 192 pixels. Figs. 2.7 (b), 2.8 (d) and 2.9 (d) show the plots of the observed nearest-neighbor distribution function, $\mathcal{G}_0$ (solid line). $\mathcal{G}_0$ is also out of the envelopes of the simulated random patterns. Finally, Figs. 2.7 (c), 2.8 (e) and 2.9 (e) shows the spatial marginal function $\mathcal{K}(s)$, for the observed pattern (solid line) and the upper and lower envelopes of the simulations. 10 μm are approximately 50 pixels.

Apart from this graphical test, a p-value corresponding to a randomization test can be used as well. The null hypothesis to test is that $\mathcal{F}_0$, the function associated with the observed pattern, is similar to the $\mathcal{F}_i$'s with $i = 1, \ldots, s$, the functions obtained from the simulated patterns. Let us consider

$$d_i = \int_0^{t_0} (\mathcal{F}_i(t) - \bar{\mathcal{F}}_i(t))^2 dt \tag{2.18}$$

where $\bar{\mathcal{F}}_i(t) = \sum^s{}_{j=0,j\neq i} \mathcal{F}_j(t)$ and $i = 0, 1, \ldots, s$. If $\mathcal{F}_0$ cannot be distinguished from the $\mathcal{F}_i$'s with $i = 1, \ldots, s$ then any permutation of $d_0, d_1, \ldots, d_s$ is equiprob-

able. The one-tail p-value would be $1 - \frac{k}{s+1}$ where k is the rank of d_0 (14).

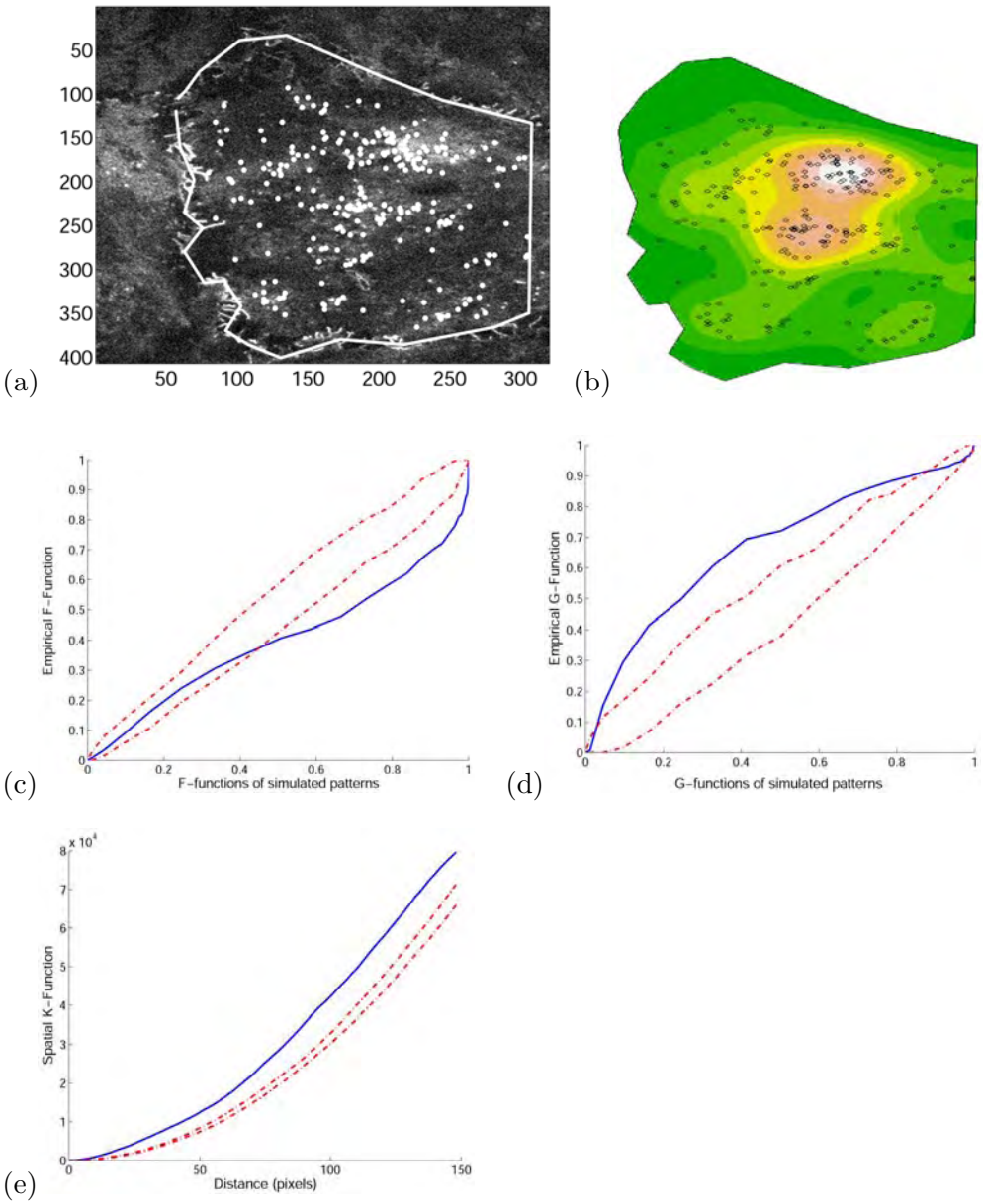


Figure 2.8: **Spatial analysis of Cell 2 by using $\mathcal{F}$, $\mathcal{G}$, $\mathcal{K}$ functional descriptors**.

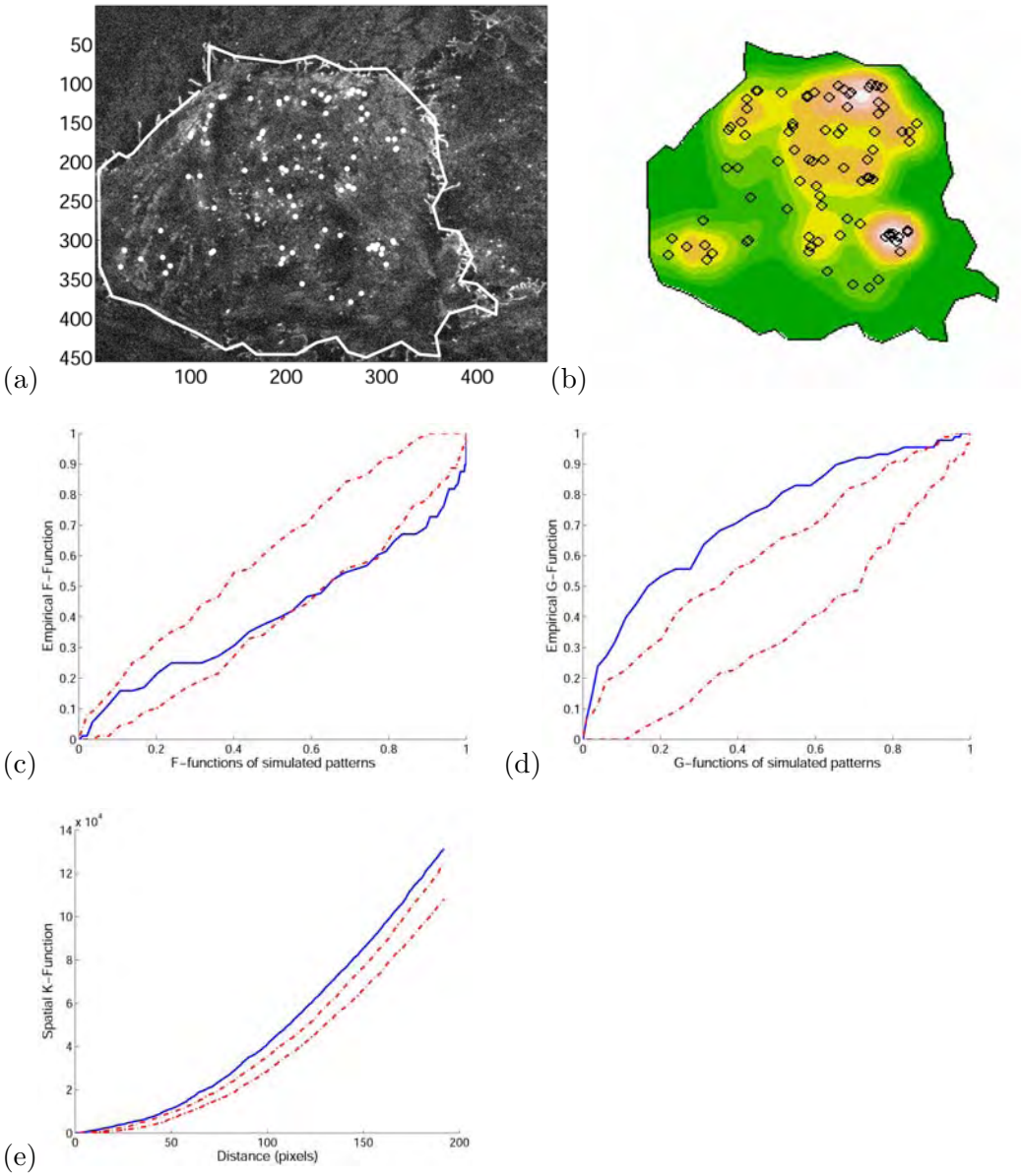


Figure 2.9: **Spatial analysis of Cell 3 by using $\mathcal{F}$, $\mathcal{G}$, $\mathcal{K}$ functional descriptors**.

In Figs. 2.7 (a), 2.8 (c) and 2.9 (c) it is shown that the three observed functions for the three cells analyzed clearly lie below the lower envelopes from the simulated point patterns. If the observed pattern is similar to the simulated patterns, then the observed $\mathcal{F}$-function, $\mathcal{F}_0$, should lie inside the envelopes. A deficiency of empty-space is observed, which is compatible with the underlying clustering mechanism. Remember that the $\mathcal{F}$-function measures the empty spaces in W. The observed $\mathcal{G}$-function along with the lower and upper envelopes against the averaged value of the $\mathcal{G}$-functions from the generated patterns are shown in Figs. 2.7 (d), 2.8 (d) and 2.9 (d). The observed $\mathcal{G}_0$ is clearly out of the envelopes. An excess of small inter-event distances is observed, which is a feature of the clustering mechanism. Lastly, Figs. 2.7 (e), 2.8 (e) and 2.9 (e) display the plots for the $\mathcal{K}$-function. Once more an excess of small distances is observed, suggesting again the proposal of an aggregated model for the observed pattern. Clearly, an aggregated model must replace the Poisson point process.

In Table 2.4 the p-values for the Monte Carlo test for the three cells and functions $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{K}$ are shown. These values again provide strong evidence for rejection of a Poisson point process in favor of an aggregated pattern. A similar result is obtained when analyzing the remaining cells. We can conclude that there are preferred sites for exocytosis.

Table 2.4: Monte Carlo p-values using $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{K}$-functions

	Cell 1	Cell 2	Cell 3
F	0.00	0.00	0.00
G	0.00	0.00	0.00
$\mathcal{K}_1$	0.02	0.02	0.02

2.3.3 Temporal clustering

In this section the marginal temporal point patterns, the times associated with exocytosis observed throughout the whole period of time, are analyzed for the different cells. No spatial information is taken into account now. This point pattern can be considered as a point pattern in the real line. An analysis of this (marginal) information is given in this section. The procedures used are analogous to those detailed in Section 2.3.2. The $\mathcal{K}_2$-function was the functional descriptor used in this case. Note that, with respect to the spatial case, the

estimator has to be modified. Eq. (2.14) shows the estimator of this function for a one-dimensional point pattern.

The null hypothesis to be tested is whether the observed times can be considered as a realization of a Poisson point process in the real line. In order to test the null hypothesis we applied a test consisting of the followings steps. First, we used a Poisson distribution to generate N, the number of events of the simulated pattern. Second, we used a uniform distribution to generate N time values uniformly distributed along the time interval $[0, T]$, where T is the length of the observation time analyzed for each image sequence. Lastly, we evaluated the $\mathcal{K}_2$-function for the simulated pattern. 999 simulations were performed, and the lower and upper envelopes were obtained. Let $\mathcal{K}_0$ be the temporal function for the observed pattern. If the observed pattern corresponds to the observed point pattern then it has to be contained inside the envelopes. Table 2.5 displays the Monte Carlo p-values observed for the cells analyzed.

For all the temporal point processes considered, the null hypothesis is clearly rejected in favor of a more aggregated point process (than the Poisson model). These aggregation effects can also be observed in Fig. 2.10. Fig. 2.10 (a), (c) and (e) show the $\mathcal{K}_2(t)$ estimated and Fig. 2.10 (b) (d), (f) show the estimated density function of the occurrence time for the three cells analyzed. The observed $\mathcal{K}_2$ is above the upper envelope. The density function was calculated for each pattern, clearly revealing the existence of two waves of exocytosis that may indicate rate-limiting step or saturation of the exocytic process. See Fig. 2.10 (b), (d) and (f).

Table 2.5: Monte Carlo p-values using $\mathcal{K}_2$-functions

Cell 1	Cell 2	Cell 3
0.0002	0.0001	0.0006

2.3.4 Spatio-temporal clustering

In this section we are concerned with the study of spatio-temporal clustering, i.e. groups of fusion events occurring close in time and close in space. Note that we are not evaluating spatial clustering or temporal clustering separately. Our objetive now is to test whether events (exocytosis) which are close in space are also relatively close in time, and conversely.

Fig. 2.11 (a) shows the observed spatio-temporal pattern for one of the three

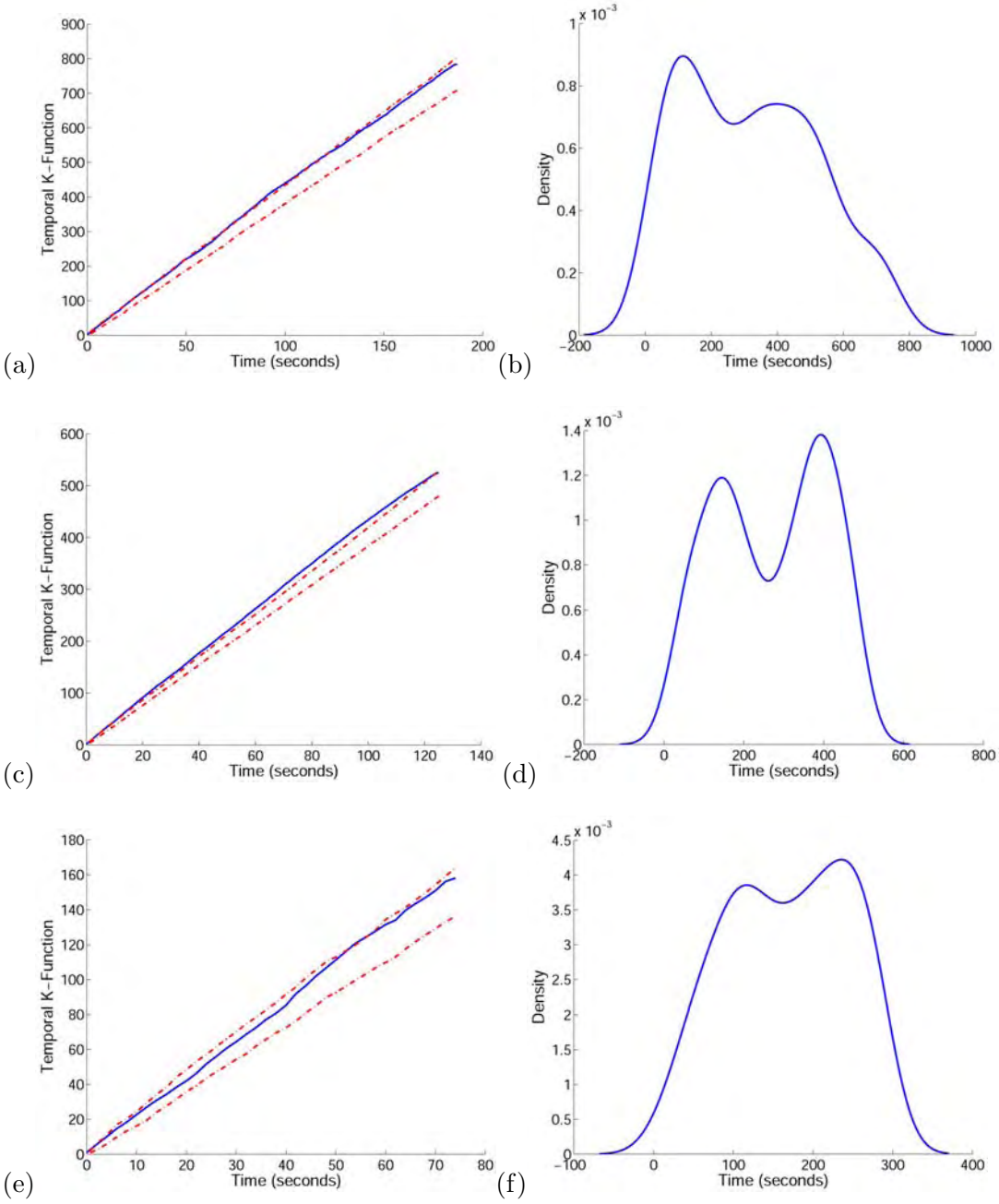


Figure 2.10: **Analysis of temporal clustering**.

cells analyzed. Dots represent the locations of fusions and colors represent the time of occurrence. The colorbar represents the time axis, from 0 seconds (dark blue) to end time (dark red). Hot spots of exocytosis are observed, but they are not connected to "temporal waves" of exocytosis. No spatio-temporal cluster exists under visual inspection. However, formal tests are needed. Given the fusion locations observed, a simulated spatio-temporal pattern in which fusions are spatially and temporally clustered (dots with similar colors are highlighted with a red '>') is shown in Fig. 2.11 (b), where clusters of colors are clearly observed. Fig. 2.11 (c) shows the same number of points randomly distributed in space and time.

The spatio-temporal point patterns observed for the three cells are displayed in Fig. 2.12 in two different ways. In Fig. 2.12 (a), (c) and (e) the colorbar corresponds to the time of occurrence whereas in Fig. 2.12 (b), (d) and (f), the radii of the white circular disks are proportional to the time of occurrence. Large radii correspond to fusions observed later in time. After visual inspection it is hard to say if there is spatio-temporal clustering, i.e. clusters of white disks with similar radii.

If the locations and times are realized independently, then Eq. (2.11) holds. We will use the function

$$\hat{\mathcal{D}}(s,t) = \hat{\mathcal{K}}(s,t) - \hat{\mathcal{K}}_1(s)\hat{\mathcal{K}}_2(t). \tag{2.19}$$

Fig. 2.13 (a), (c) and (e) shows the observed $\hat{\mathcal{D}}$-functions for the three cells. It is not an easy task to evaluate graphically if the values observed are close to zero at small values of s and t, as expected if the locations and times are independent. Nonetheless, we can infer from these plots that there is no spatio-temporal clustering. Axes represent time and space. When locations and times are independent, the function is almost null. $\hat{\mathcal{D}}$-function represents the degree of spatio-temporal association. Thus, for instance, at distance $s = 10$ and time $t = 10$, $\hat{\mathcal{D}}$-function is near zero, which means there is no spatio-temporal cluster within a circle of size 10 and during 10 seconds. When we increase the space and time, the $\hat{\mathcal{D}}$-function necessarily deviates from zero, since clusters of fusions tend to be formed, in the limit when we consider all the cell and the whole sequence all the points would fall inside. Fig. 2.13 (b), (d) and (f) show the results of the randomization test for spatio-temporal clustering, which confirm the first hypothesis of independence.

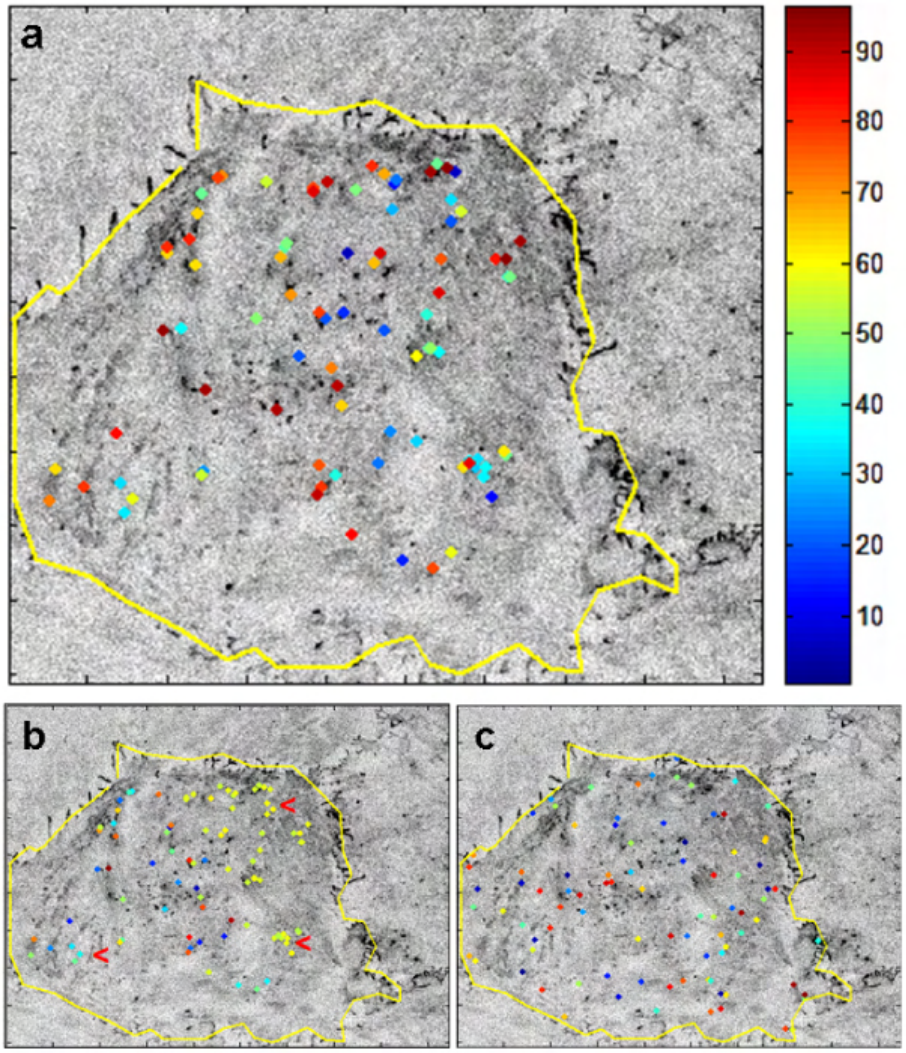


Figure 2.11: **Spatio-temporal point patterns**.

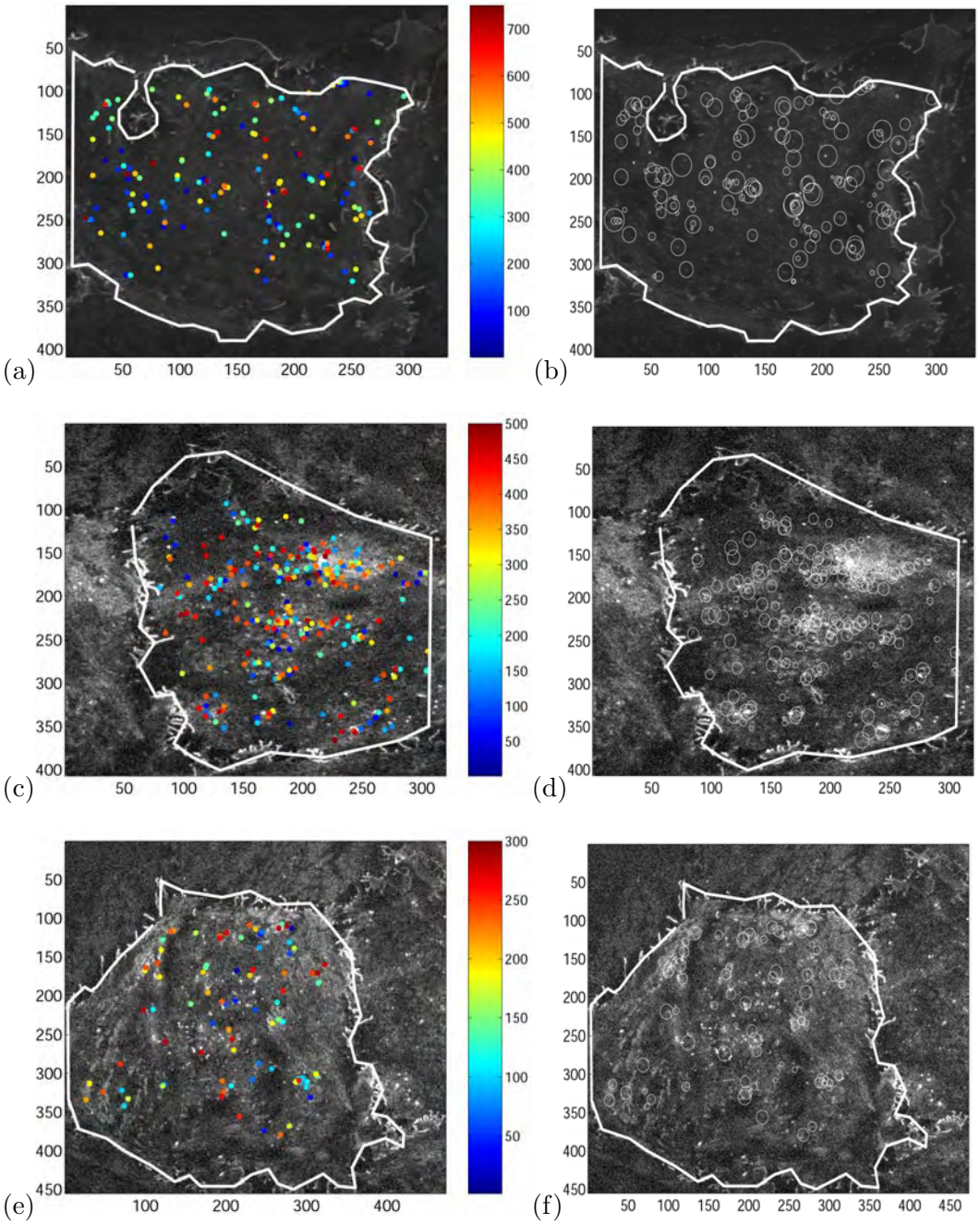


Figure 2.12: **Spatio-temporal point patterns observed for the three cells analyzed.**

We will use the following summary statistic proposed in (15)

$$U = \sum_{k_1=1}^{m_1} \sum_{k_2=1}^{m_2} R(s_{k_1}, t_{k_2}), \tag{2.20}$$

where s_{k_i}, t_{k_j} are the spatio-temporal sampling points; s_{m_1} is the maximum distance used to estimate the spatial function; s_{m_2} is the maximum time-lapse used to estimate the temporal function, and

$$R(s,t) = \frac{\hat{\mathcal{D}}(s,t)}{\sqrt{var(\hat{\mathcal{K}}(s,t))}}. \tag{2.21}$$

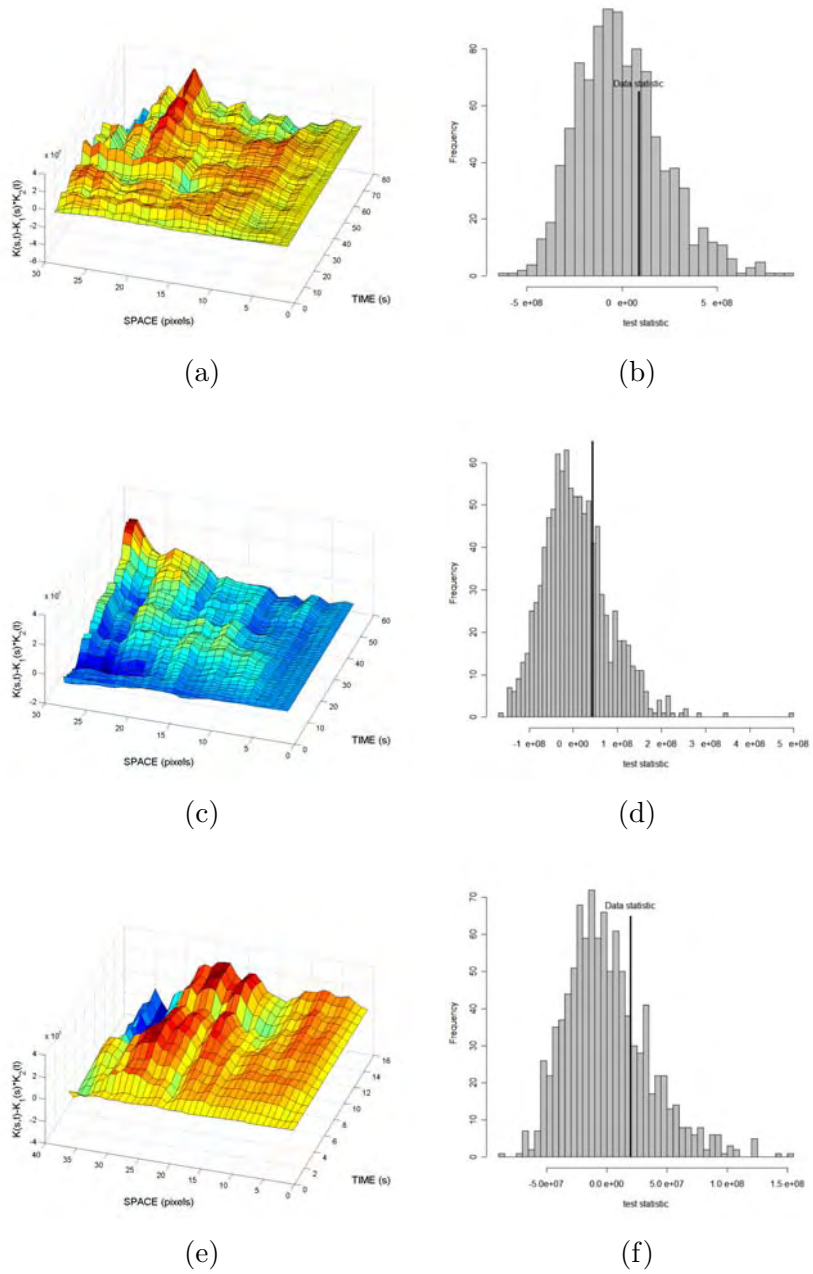
Let u_1 be the observed value of U for the observed point pattern and let $u_2, \ldots, u_m$ be the observed values when random assignment of locations and times are generated. A Monte Carlo test with 99 simulations is applied. Table 2.6 gives the p-values observed for the three cells analyzed. They are really large p-values, greater than 0.5. We can then state that there is no spatio-temporal clustering for any cell. In summary, the locations and times of the different exocytosis can be considered independent.

Table 2.6: Monte Carlo p-values for testing spatio-temporal clustering

Cell 1	Cell 2	Cell 3
0.698	0.753	0.783

2.4 Discussion and Conclusions

A method for studying the spatio-temporal distribution of exocytosis events at the plasma membrane in the context of the theory of spatio-temporal point processes has been presented. Locations and times of occurrence of exocytosis have been jointly considered for the first time and a complete spatio-temporal study of exocytosis observed in three cells has been performed. Results indicate that there is no spatio-temporal coupling of constitutive exocytosis, indicating that every fusion event acts individually and that there is no central coordination by a diffusible factor that influences the final fusion event. It is still, however, conceivable that such factors could act upstream to influence the antepenultimate and penultimate processes, e.g. vesicle docking and priming. These results would



(a) (b)

(c) (d)

(e) (f)

Figure 2.13: $\hat{\mathcal{D}}$**-function and randomization test of spatio-temporal clustering**.

suggest that if a diffusible calcium signal was produced, that it is either too weak or dissipates too rapidly to influence vesicle fusion in time at nearby spatial sites. Importantly, the methodology presented here can also be applied later to processes such as regulated fusion where the spatio-temporal coupling based on global diffusible factors is known to play a central role and should allow one to measure their effects.

A study of locations led us to state that exocytosis tend to form clusters of fusions at specific regions of the plasma membrane. Clearly, an aggregated model must replace the Poisson point process. This study was performed by using different spatial descriptors that allow us to strongly confirm previous results obtained (31). The spatial clustering in the statistical analysis of spatial point processes is considered to be a consequence of either a direct interaction between the points (say, one fusion leads to more fusions in its vicinity) or of environmental factors in the neighborhood of points. Both kinds of factors should be considered when applying our model in a biological context in search of the causes of clustering. For example, from among the environmental factors we could consider cell cytoskeletal architecture underlying the plasma membrane as well as the distribution of various molecules implicated in the regulation of exocytosis.

Surprisingly, contrary to our expectations, temporal release showed a double hump distribution, like a double humped camel's back. But what would be expected? Given that VSVG-GFP is released from the endoplasmic reticulum (ER) as a temporal wave, we did expect some degree of temporal coupling as cargo is kinetically released and egresses out of the ER, trafficks to the Golgi Complex and finally to the plasma membrane. Despite hundreds of studies of identical markers by pulse-chase biochemistry, to our knowledge biochemistry studies have shown no such distribution, but rather the cargo appears on the surface as a single peak (28; 45). But why might this key observation be missed and what is its significance? These data are at odds with others reported in the literature and could reflect differences between ensemble analysis (as is typically done) and the finer temporal-spatial resolution of the processes studied at the single vesicle level. An advantage of biochemistry experiments is that they automatically average the results from thousands (minimum) to millions of cells. Thus, they provide ensemble measurements. We, however, are now doing thousands of measurements (image timepoints) on a single cell. Namely, the analysis is done at the single unit level, or vesicle here. With computational analysis, as here, this can reveal a fundamental process and order of magnitude of resolution better than with ensemble methods such as traditional biochemistry. For instance, we note that each cell is slightly different in the temporal control of exocytosis (or position of the double humps), thus the mere act of averaging many cells can blur the distinction of in-

dividual cellular behaviors. Second, the temporal resolution of such biochemical experiments is typically in the tens of minutes, as each experiment must be done for every timepoint. This is in stark contrast to the millisecond resolution that sensitive imaging can achieve. Even in the events here which are separated by $2-5$ min, such a dual peak would be blurred with a 10 min temporal filter and applying a Nyquist temporal filter would require 1 min sampling intervals and even then could be missed by a blurred average.

The biological significance of these observations is the implication that a pre-fusion step is rate limiting. This would suggest saturation of an earlier trafficking event. For instance, one possibility is that during exocytosis only a given amount of the protein fusion machinery can be used and must be recycled for subsequent rounds. This interesting observation suggests that the precise kinetics of earlier membrane trafficking steps be revisited using single cell time-lapse imaging to see which part of the exocytosis acts as the bottleneck, and hence would be an ideal place for finer regulation. Importantly, only a slight modulation of a constant rate of a critical exo- or endocytic pathway can cause a striking cellular phenotype such as in diseased pathological states. The computational methods that we have established here should provide new opportunities for testing key models of how membrane traffic in cells function and the processes that regulate it. The methodology can be easily extrapolated and applied to the study of other types of cells, such as chromaffin cells and neurons, or to other types of cell processes. The model is flexible and general enough to incorporate and test these biological hypotheses in a direct way. Also more complex hypotheses that take into account heterogeneity in the plasma membrane or internal cell structures such as microtubules could be formulated and tested.

Several aspects remain open. First, formulating a parametric model for the spatial locations where biological factors could be included in the model. A Cox process could be a good general framework since it can deal with the non-homogeneous intensity rate of fusions throughout the plasma membrane. In this new framework the intensity function is a stochastic process, i.e. first, an intensity function is generated and, second a non-homogeneous Poisson process with this intensity function is generated. The Cox process also allows us the incorporation of covariates in the analysis and possible sources of aggregation due to environmental heterogeneity, for instance the existence of microtubules or any local characteristic of the plasma membrane. Another model, a Poisson cluster process, that incorporates an explicit form of spatial clustering could be a good starting point too. Moreover, it is possible to extend those models to see if the amount of cargo depends on the position and time, that is, to study dependencies on what, where and when. This last hypothesis can be formulated considering a

marked spatio-temporal point process. Once more, note that the framework chosen is flexible and general enough to incorporate and test biological hypotheses in a direct way.

Finally, an image processing method that allows us to automatically segment fusions is presented and will greatly facilitate future investigation. While this software was used in this proof-of-principle study exclusively for PtK_2 cells, the parameters should be easily adaptable to analyze secretion in a wide range of cells such as astrocytes, neurons, fibroblasts, etc.

Chapter 3

Counting Endocytic Spots by Means of Boolean Models

3.1 Introduction

In this chapter we are concerned with the estimation of the mean number of endocytic spots per unit area and their mean perimeter observed in a given frame, when overlapping between different endocytic spots exist. We propose to consider the union of the different endocytic areas as a realization of a Boolean model where this overlapping between endocytic spots is explicitly considered and assumed. Estimators based on Boolean models for the mean number of endocytic spots per unit area and their mean perimeter observed in a frame are proposed. It should be noted that, since endocytic vesicles are below the spatial resolution limit of the imaging system, the perimeter of the endocytic spots does not provide a measurement of the real size of the vesicles, but rather is a geometric property of the point-spread function. However, the geometric features of the endocytic spots, which could be associated with vesicles, can be safely used for the purposes of the Boolean model, such as determining the number of endocytic spots per unit area. To our knowledge, it is the first time that these estimators obtained from this type of image sequences are proposed. In following chapters we will generalize this model from $2D$ to $3D$ in order to study temporal and spatial information.

Boolean models have been successfully applied to similar problems in other

disciplines such as, estimation of the area occupied by forest or grass in a given field observed from the air, area of pores in a given material or counting of red blood cells in the viewing field of a microscope (66). Nevertheless, to our knowledge, they have not been applied yet in the live cell image analysis, despite their interest and usefulness.

An endocytic protein, known clathrin has been used in this study to show the application of the model. Clathrin spot bursts appear in the images as small round hyperfluorescent areas with an approximately Gaussian shape on a dark background. These Gaussian shapes overlap each other forming clumps.

In Section 3.2, the image acquisition method, the data and the image processing algorithm are described. In Section 3.3, the methodology based on Boolean models is introduced. In Section 3.4, the estimated values for the sequences of images analyzed are presented. Conclusions are given in Section 3.5.

3.2 Data Collection

3.2.1 Image processing

The image processing method developed is based on the application of the Top-Hat transform to extract peaks of fluorescence, template matching to remove eventual noise and region growing technique that will fully delineate each marked object.

Step 1: Opening Top-Hat. The presence of protein molecules is revealed by brighter pixels surrounded by a relative dark background (17; 62). The *opening* of a grey-level image f with b, denoted $f \circ b$, is defined as

$$f \circ b = \bigvee_x \{b_x + u : b_x + u \leq f\},$$

where b is a gray-scale image, known as the *structuring function*. For the special case when $b(x, y) = 0$ for $(x, y) \in B$, and $-\infty$ otherwise, being B a given subset of $\mathbb{R}^2$, it is referred as a *flat structuring function*. $\bigvee$ stands for maximum and b_x is the translation of the structuring function centered at x. The opening is found by taking the maximum over all morphological translations of the structuring function b that fit beneath the input signal f. It removes narrow peaks from the images, where the meaning of narrowness is relative to the shape and size of the structuring function. It is a smoothing filter that approximates an image from below. Then, the operator $\psi_{opnth}(f) = f - f \circ b$ produces such peaks.

Fig. 3.1 illustrates the use of Top-Hat operator to extract the peaks of hyperfluorescence. Figs. 3.1 (a) and (d) show a subimage of the original grey level image. Opening this image results in Figs. 3.1 (b) and (e). This figures correspond to the local background, with a mean grey level value of 20. A flat structuring function, a disk with radius 10 pixels, was used for these images. The substraction between images in Fig. 3.1 (a) and (b) leads to the detection of peaks (see Figs. 3.1 (c) and (f)). Note that the reference level is approximately zero in Fig. 3.1 (c). The Top-Hat operator implies the subtraction of the local background, which is robust against an inhomogeneous illumination. Grey-level images have been brightened for visualization purposes.

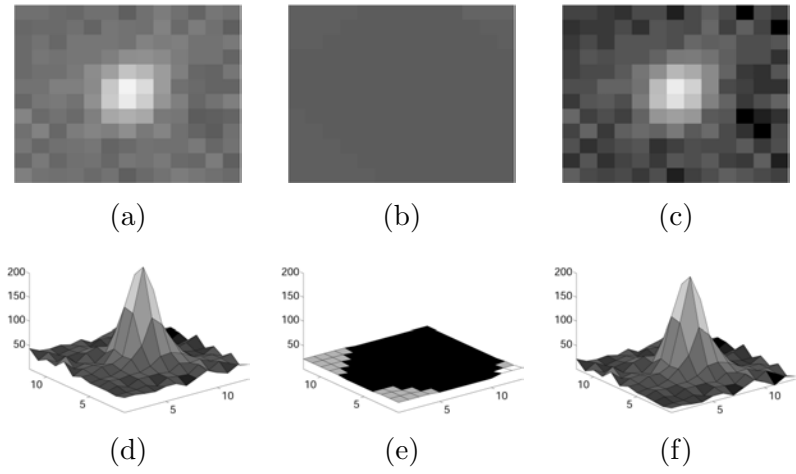
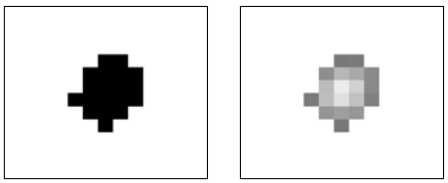


Figure 3.1: **Extracting peaks**. (a) Original image. (b) Opened image. (c) Top-Hat Image. (d), (e) and (f) correspond to the three dimensional plots of the original image, opened image and Top-Hat image, respectively.

Step 2: Thresholding the Top-Hat Image. Owing to noise, the Top-Hat image is thresholded to remove points with low signal. This results in a black-and-white image shown in Fig. 3.2 (a), where white represents the background and black corresponds to the object. Only points with values of the Top-Hat image higher than a given threshold are taken into account for the next step, the template matching. The binary image is used as a mask for the Top-Hat image (see Fig. 3.2 (b)). Points outside the mask correspond to the background or to noise and they are not considered for posterior processing. The value of the threshold depends on the signal-to-noise ratio, and it is a tuning parameter of the method. A value of 50 for 8-bits grey levels images is used. The value was fixed by taking

into account the mean intensity of the background . See Fig. 3.1 (d).



(a) (b)

Figure 3.2: **Thresholding**. (a) Binary image after thresholding the Top-Hat image; (b) Top-Hat image after applying the binary mask.

Step 3: Template matching. Since clathrin spots lead to roughly Gaussian shapes of fluorescence, the next step is to match a template to the image resulted from applying the binary mask to the Top-Hat image. The template is a small image that contains the shape we are trying to find. The template is centered at each image point and the number of points in the template that match those of the image is counted. The procedure is repeated for the entire image and the point which led to the best match, the maximal count, is deemed to be the point where the template lies within the image. A Gaussian template with $\sigma = 1$, shown in Fig. 3.3 (a), is used. Only maxima with higher correlation coefficient than a given threshold (threshold set at 0.7) are used as seeds for the region growing process that will fully delineate each marked object. Seeds are shown as black dots in Fig. 3.3 (b). The size of the Gaussian template depends on the spatial resolution.

Step 4: Region Growing. Points with higher correlation are considered the seeds of the region growing process. We have to judge whether a particular neighboring pixel is to be included in the grown object. If the pixel grey level is greater than a certain absolute threshold, then the pixel is included into the growing object. A value of 30 is used for this signal-to-noise ratio. The Top-Hat image is used instead of the original image in this step. It avoids the growing region to spread in excess due to noise. Fig. 3.3 (b) to (e) show the grown region, delineated with black crosses, for different consecutive iterations.

Fig. 3.4 (a) illustrates the final results of the application of region growing technique on a detail of the original grey-level image. This region is 45×90 pixels in size. Fig. 3.4 (b) is a black-and-white representation of the intermediate step after thresholding the Top-Hat image; whereas Fig. 3.4 (c) is the final step of region growing. Small particles corresponding to noise are removed in the

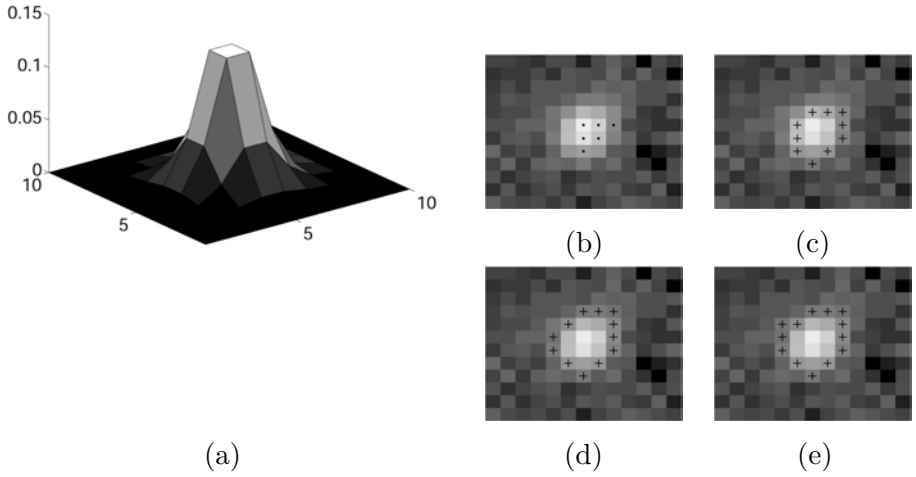


Figure 3.3: **Template matching and region growing**. (a) Gaussian kernel for template matching. (b) Seed (dots). (c), (d) and (e) show consecutive iterations of region growing process.

template matching step. The region growing technique delineates the final shapes. Figs. 3.5 (a) and (b) display the image obtained after Top-Hat thresholding and the final segmentation after region growing for the entire original grey-level image, respectively.

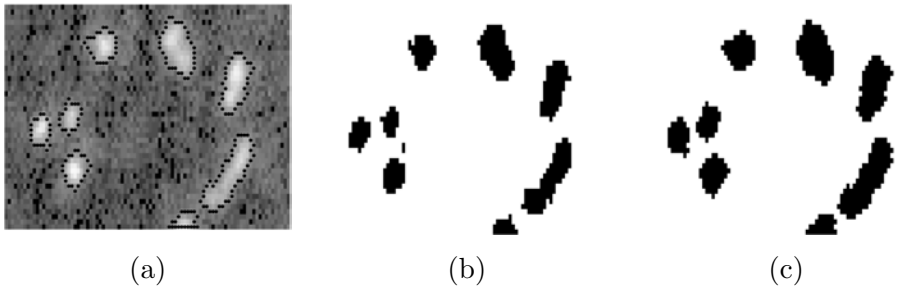
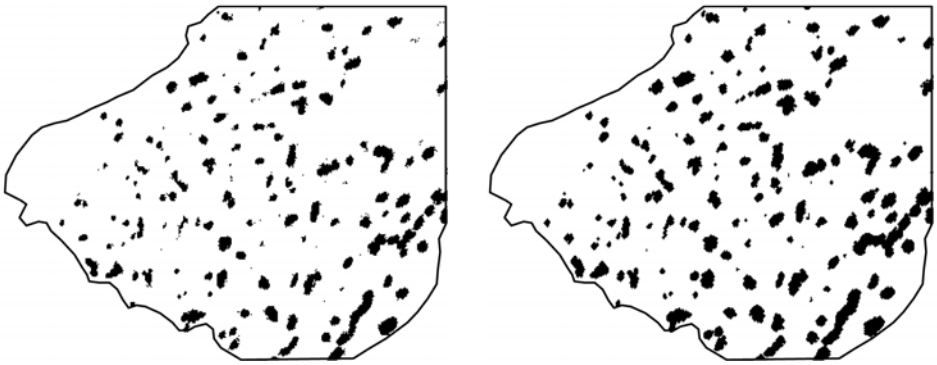


Figure 3.4: **Intermediate steps of image processing**. (a) Original image and grown regions. (b) After thresholding the Top-Hat image. (c) After region growing.



(a) (b)

Figure 3.5: **Segmentation of the original image.** (a) After thresholding the Top-Hat image. (b) Final segmentation after region growing.

3.3 The model

As we mentioned in Section 3.1, experimenters like to answer questions such as: how many endocytic spots are there in a given frame? How large are the endocytic spots? How is the shape of these endocytic spots? Fig. 3.5 (b) shows a binary image where different kinds of randomness have contributed to its formation: the acquisition process, the segmentation process and, possibly, other non-controlled experimental factors. In our approach, this image is considered as a realization of a stochastic model, i.e., an experiment where the results (binary images) cannot be predicted in advance.

The natural mathematical framework that gives some answers to these questions is the Random Closed Set theory. The datum to model is a random binary image, i.e., random subset of the Euclidean space $\mathbb{R}^2$. The research in random set theory is highly active. After the original definition by G. Matheron (35; 36), and the subsequent works of J. Serra (62) and their colleagues the Boolean model was defined. These studies cover at least an 80% of the matter and definitions related to Boolean models.

An applied study of this kind of stochastic model addressed to practitioner is presented in (38), where basically is described as a framework to model (static) 2D or 3D overlapping objects. Ilya Molchanov from the University of Bern maintains a BibTeX bibliography called "Random sets and related topics" within "The Collection of Computer Science Bibliographies" which contains about 3000 items

and it is the fastest way to access to information about the subject. References (11; 26; 39; 66) contain general and clear presentations about these stochastic models. Van den Berg et al. (72) introduce dynamics into the Boolean model, letting the grains move around, and study the properties of this modified model. Dousse et al. (18) model a wireless sensor network by defining a blinking Boolean model, in which grains switch in an uncoordinated way between an 'on' and 'off' mode to save energy, in order to study the latency of the network. In most of the studies that deal with 3D data it is common to analyze each one of the 2D sections separately, even with isotropic data. Dougherty et al. in (16) 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 (55) to estimate the number of straight cylinders in a random system. Boolean models they have been also used in texture analysis and synthesis (4; 20; 22; 23).

3.3.1 The Boolean model

In our approach, the total area segmented covered by fluorescent protein molecules associated to endocytic spots is considered as a realization of a random closed set. Fig. 3.5 (b) displays the area covered in a frame. If different subimages were cropped then they would look similar among them but not equal to each other. This is the intuitive idea underlying to the concept of stationarity. The stationarity hypothesis is a quite natural simplifying hypothesis for the stochastic model, and it seems tenable for the analysis of our images.

The next step is to choose a type of random closed set in order to model our images. Let us consider the underlying phenomenon that generates the images. Different endocytic vesicles produce different endocytic spots of fluorescence located around each other. It seems quite natural to assume that the different endocytic areas are independent realizations of a given random set independently and uniformly located within a given region of the plasma membrane. Independent realizations of a given random set means that the size and shape of a given endocytic spot is not related or conditioned by the other endocytic spots. This assumption is justified by the recent biological data, supporting a random model of endocytic vesicle formation (19). The second basic hypothesis is that the different centers of the endocytic spots are located uniformly in a given region contained in the plasma membrane and the location of each one is independent from the locations of the others. It is a parsimonious stochastic model for our binary images. The mathematical formulation of these intuitive concepts is known

as **Boolean model**. Two basic references about the topic are (38; 66) and should be consulted for a complete presentation.

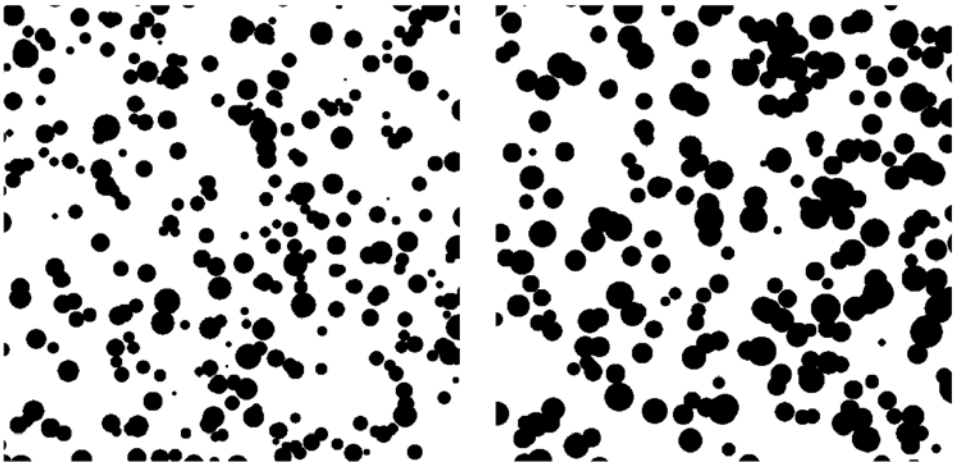
The locations of different endocytic spots, corresponding to the locations of nucleation sites for vesicles, can be considered as a *spatial temporal point pattern*, a probabilistic model producing locally finite sets of points (11; 14; 66; 67). In particular, vesicle locations in each cross-section of the image sequence will be modelled as a realization of a Poisson point process in $\mathbb{R}^2$. In this study the hypothesis of stationarity concerning with the vesicle locations is assumed. Under the stationarity hypothesis, one can define the intensity of the process as the mean number of points per unit area, λ, i.e. the mean number of endocytic spots per unit area in our case. We will assume that the vesicle locations can be considered as realizations of a Poisson point process.

From now on, we will assume that the primary grain A_0 is isotropic, i.e., its distribution is invariant against random rotations. Given an isotropic and stationary Boolean model we are interested in the following three unknown parameters, which allow us to characterize the process: (i) the intensity λ of the germ process (number of germs per unit area); and (ii) the mean perimeter of the primary grain, $u_0 = EU(A_0)$.

In our application the primary grain corresponds to a cluster of clathrin molecules associated to one single endocytic spot. The intensity of the germ process or mean number of points per unit area is the mean number of endocytic spots per unit area of cytosol in our case. In this study, it is not assumed any parametric model for the primary grain as it presents irregular forms, and the size is estimated by the mean perimeter u_0.

In order to illustrate the kind of binary images generated by a Boolean model, two simulated stationary homogeneous Boolean models in a 512×512 rectangular window are shown in Fig. 3.6 (a) and (b). These simulated images have similar aspect to that one obtained after image segmentation, Fig. 3.5 (b). In both simulated images, the primary grain is a ball with normally-distributed random radius, $R \sim N(\mu, \sigma)$. Thus, the model is completely characterized by λ, μ, σ. As it is showed in Fig. 3.6 (a), the first model has an intensity $\lambda = 0.0012$, i.e., around 315 grains in the window, and its primary grains are balls with random radii $R \sim N(8, 3)$. However, most of the observed clumps are constituted by several primary grains and the number of clumps is only 186. Thus, we can estimate correctly neither the number of grains per unit area, nor the mean radius of a grain, unless a Boolean model is assumed. The second model (see Fig. 3.6 (b)) is formed by bigger primary grains ($R \sim N(11, 3)$) with $\lambda = 0.0009$, i.e., 235 grains. Isolated areas in both simulated models can be observed, but

again estimating the size from this biased sample will produce biased results, since smaller grains have a higher probability of belonging to an isolated area. Additionally, higher dispersion in size will introduce higher bias in the estimation of the size and intensity, if only the isolated spots are considered as the sample.



(a) (b)

Figure 3.6: **Simulations of two different Boolean models in a** 512×512 **window**. (a) Balls with random radii $R \sim N(8, 3)$ and $\lambda = 0.0012$. (b) Balls with random radii $R \sim N(11, 3)$ and $\lambda = 0.0009$.

Some previous applications of Boolean models within Image Processing are (13; 22; 23).

3.3.2 Defining the sampling window W

From our point of view, the most difficult hypothesis to be verified concerns the uniform distribution of the endocytic spots along the plasma membrane. This assumption appears in the definition of Poisson point process, the model assumed for the germs. We solve this problem by using only part of the plasma membrane observed. The sampling window finally used is the union of the regions where we have observed fluorescence of protein molecules along the total time interval analyzed. From now on, this bounded sampling window will be denoted as W.

Definition 1 (Sampling Window) *Let us denote* $W_1, W_2, \ldots, W_n$ *the sequence*

of sets in $\mathbb{R}^2$, where W_i corresponds to the union of all white areas of the ith binary image, i.e., the region covered for all the endocytic areas in the corresponding image. At any point $x \in W_i$, we can assume that at least one endocytic spot is detected in the ith-image. The sampling window W is defined as the union of the sets W_i's, i.e., $W = \cup_{i=1}^{N} W_i$.

Fig. 3.7 displays the sampling window for a sequence of 210 seconds. White represents the regions where protein molecules have been observed. It corresponds to an active area of the plasma membrane of 90.57%.

Figure 3.7: **Sampling window**. Obtained by stacking the whole sequence composed by 2100 images.

3.3.3 Parameter estimation. Minimum contrast method.

A robust method to estimate the intensity λ of the process (mean number of endocytic spots per unit area) and the mean perimeter of the endocytic spots u_0 is the minimum contrast method. There is an extensive literature about statistical analysis of Boolean models. This method is detailed by D. Jeulin in (29), who is considered to be the one who bridged the gap between RACS and physical laws. Other fundamental references are (36; 62). Furthermore, for pedagogical details addressed to practitioners (38; 66) are interesting references. See (21) for computational details.

We propose to use this method since it is a simple and robust procedure. The

irregular sampling window makes difficult to apply other alternative procedures such as the method of intensities.

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) then let us define the following function

$$H_K(t) = 1 - \frac{1 - T(tK)}{1 - p}, \tag{3.1}$$

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,

$$H_K(t) = 1 - exp\Big\{ - \lambda E[\nu_2(A_0 \oplus tK) - \nu_2(A)]\Big\}, \tag{3.2}$$

where K is the unit disk on the plane and A_0 the primary grain. (For the original definitions see (36) pp. 139 and (60)).

It is well-known (38; 66) that, if $H^l_K(t) = -\log(1 - H_K(t))$, then

$$H^l_K(t) = \lambda[E\nu_2(A_0 \oplus K) - \nu_2(A_0)]. \tag{3.3}$$

The generalized Steiner Formula establishes that

$$E\nu_2(A_0 \oplus \check{K}) = E\nu_2(A_0) + \frac{U(K)EU(A_0)}{2\pi} + \nu_2(K). \tag{3.4}$$

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

$$E\nu_2(A_0 \oplus B(0, t)) = E\nu_2(A_0) + EU(A_0)t + \pi t^2. \tag{3.5}$$

Thus,

$$\frac{H^l_{B(0,1)}(t)}{t} = \lambda\pi t + \lambda u_0. \tag{3.6}$$

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

$$\frac{H^l_{S(0,1)}(t)}{t} = \lambda t + \frac{2}{\pi}\lambda u_0, \tag{3.7}$$

where u_0 stands for the mean perimeter of the primary grain.

The contact distribution function can be estimated from the estimators of the capacity functional and the area fraction from Eq. 3.1. 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

$$\hat{T}_W(K) = \frac{\nu_2((\Phi \oplus \check{K}) \cap (W \ominus K))}{\nu_2(W \ominus K)}, \tag{3.8}$$

where W is the sampling window and $A \ominus \check{K} = \{x \in A : x + K \subset A\}$. Using the estimator given in (3.8), the left-hand side of (3.6) and (3.7) 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

$$\hat{\lambda} = \frac{c_1}{\pi},\ \hat{u}_0 = \frac{c_0}{\hat{\lambda}}, \tag{3.9}$$

when K is a disk.

It is important to remark on some implementation details. Firstly, the contact distribution function $T_K(t)$ 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 for $1 - T_K(t)$ in the interval $[0, 0.8]$ are used in the linear fitting. Secondly, the unit square structuring element has been used because its digital approximation is more accurate than the corresponding approximation for the disk. Finally, we fit the data in a least-squares sense constraining the resulting polynomial to pass through the three first estimated points.

At this point it is important to test whether the logarithm of the capacity functional is a polynomial of degree the dimension of the structuring element. Fig. 3.8 shows the fitting of function $log(1-T_K(t))$ over a frame showing clathrin protein. Marks in the plot correspond to the values obtained after applying Eq. 3.8 with a square structuring element, and solid line corresponds to the next polynomial of degree 2, $0.0075x^2+0.0711x+0.0242$. All the frames tested showed almost identical fittings and showed the same behavior, so we can conclude that the process could be the result of a Boolean Model.

3.4 Results

We have analyzed 6 image sequences expressing ClatrhinGFP in order to illustrate the methodology. Images were acquired at 0.25 frames per second for the clathrin. All the results are in pixels (1 pixel = 200 nm).

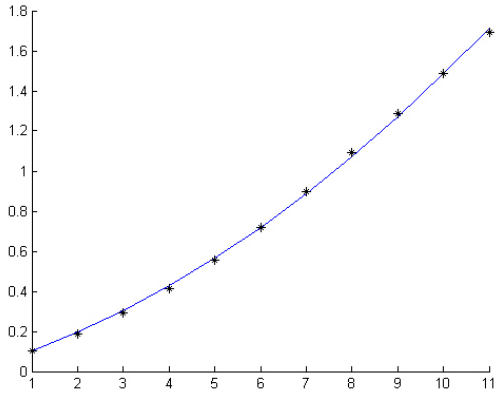


Figure 3.8: **Logarithm of capacity functional and fitting**. Logarithm of the capacity functional and a polynomial fit for these values.

Given a sequence, the intensity $\hat{\lambda}$ and the mean perimeter $\hat{u}_0$ for each frame were estimated from (3.7). Let us denote these estimates as $\{(\hat{\lambda}_i, \hat{u}_{0,i}) : i = 1, \ldots, n\}$, where n is the number of frames. A first basic assumption is that these parameters do not vary significantly along the image sequence. More formally, we assumed that the formation of endocytic vesicles is a stationary phenomenon, i.e. if a different temporal window had been chosen, a similar behavior would have been observed, at least within the periods considered for each sequence.

Fig. 3.9 (a) displays the $\hat{\lambda}$'s estimated values for the third clathrin sequence analyzed using a unit square as structuring element. Fig. 3.9 (b) displays the estimates of the mean perimeter. The horizontal solid line corresponds to the total mean for this sequence. Similar plots were observed for the other sequences.

Other characteristics of the primary grain can be estimated only if a parametric model is assumed. For instance, if a random disk with random radius R is assumed, then we estimate $u_0 = 2\pi ER$. Hence, the mean radius ER and the mean area $a_0 = \pi ER^2$ from the mean perimeter could be estimated if a one-parameter probability distribution is assumed for the radius. Other methods to estimate the mean area of the primary grain can be found in (66).

From a given sequence $\{(\hat{\lambda}_i, \hat{u}_{0,i}) : i = 1, \ldots, n\}$, point estimates and confidence intervals were calculated. The batch-mean method was used (33). A short explanation follows. First, the sequence is partitioned into a set of equally sized

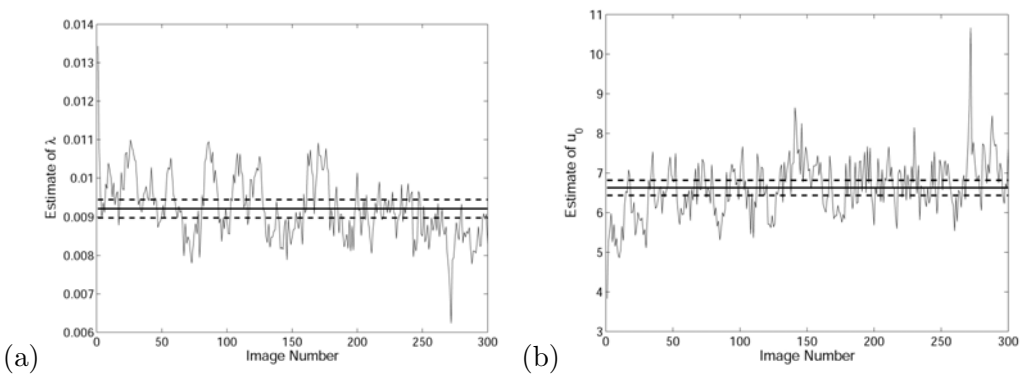


Figure 3.9: **Estimated values for λ and u_0 for the third clathrin sequence using a unit square**. (a) Point estimates and confidence limits at 95% for $\hat{\lambda}$. (b) Idem for the mean perimeter of the primary grain $\hat{u}_0$.

and non-overlapping batchs, from 1 to s, from $s+1$ to $2s$ and so on. The mean of the estimates within a batch is then calculated as

$$\bar{\lambda}_j = \sum_{i=(j-1)s+1}^{js} \hat{\lambda}_i, \tag{3.10}$$

and

$$\bar{u}_{0,j} = \sum_{i=(j-1)s+1}^{js} \hat{u}_{0,i}, \tag{3.11}$$

with $j = 1, \ldots, J$. The batch size s is chosen in such a way that the values $\bar{\lambda}_j$ defined in Eq. (3.10) are approximately uncorrelated. The same comment applies to Eq. (3.11). These batch means are approximately normally distributed as a consequence of the central limit theorem for sequences of dependent variables. Finally, the confidence interval for the mean of a normal distribution is used. This confidence interval is given by $\bar{\lambda} \pm t_{J-1,1-\alpha/2}\frac{S}{\sqrt{n}}$ where $\bar{\lambda} = \sum_{j=1}^{J} \bar{\lambda}_j/J = \sum_{i=1}^{n} \hat{\lambda}_i/n$ and $t_{J-1,1-\alpha/2}$ is the $1-\alpha/2$ quantile of a t-distribution with $J-1$ degrees of freedom. The mean perimeter u_0 is estimated similarly from the original point estimates $\bar{u}_{0,j}$'s. On analogy of signal filtering, it could be said that this method removes the high frequencies observed in the estimates.

We applied the described estimators to six biological image sequences expressing clathrin coupled to the Green Fluorescent Protein (GFP). These sequences

were derived from three movies of fibroblast (COS7) cells, acquired at 1 frame every 4 seconds. The estimates for the sequences expressing clathrin can be found in Table 3.1. Columns headed $\hat{\lambda}$ and $\hat{u}_0$ give the point estimates and the 95% confidence interval using the different image sequences for the intensity and the mean perimeter, respectively. It is important to remark the high accuracy provided by the batch-mean method. We estimated the mean intensity $\hat{\lambda}$ with a 95% confidence interval. The intervals are very short. This comment applies for all the sequences analyzed. A small decreasing in the mean intensity and a small increasing in the mean perimeter is also observed in the estimates, probably as a consequence of photo-bleaching effect. The application of the batch-mean method implies a minimum number of frames. Sequences of 300 leads to about 15 batches of 20 images, which allowed us to use the method. The column headed $\hat{\lambda}\nu_2(W)$ is the estimate of the mean number of endocytic spots for each sequence. The fraction of spots not observed due to the overlapping effect, and that are provided by the model, ranges from 74% to 127% of the number of clumps.

Table 3.1: Estimates and 95% confidence intervals for the six sequences expressing clathrin analyzed

Cell	Sequence	$\hat{\lambda}$	$\hat{u}_0$	$\hat{\lambda}\nu_2(W)$
1	1	0.0099 ± 0.0006	9.565 ± 0.28	281.86 ± 17.49
	2	0.0085 ± 0.0006	9.575 ± 0.32	347.20 ± 23.69
2	3	0.0092 ± 0.0002	6.6269 ± 0.19	277.54 ± 7.23
	4	0.0097 ± 0.0003	7.1997 ± 0.24	235.41 ± 8.05
3	5	0.0081 ± 0.0010	9.0042 ± 0.37	170.17 ± 20.88
	6	0.0075 ± 0.0012	8.9756 ± 0.34	220.28 ± 3.54

An interesting remark is the discretization effect. In a digital image and with primary grains about 4 pixels in size, the perimeter and the area of so small primary grains are the same. Thus, we give the estimation of u_0 as an indicator of the size of the primary grain.

Finally, an interesting exercise in order to evaluate the estimates provided could be to perform a simple (coarse) analysis using the characteristics measured from the connected components. For instance, we know that each clump contains at least the fluorescence associated to an endocytic spot. The perimeter of the connected components could be obtained and the number of clumps counted. A rough estimator of the intensity of endocytic spots would be the number of

clumps divided by the area of the sampling window, i.e., to estimate this intensity using the estimated intensity of clumps. It is clear that the intensity would be underestimated and the mean perimeter overestimated by using this simple approach. In other words, the real number of endocytic spots in each image would be underestimated, since it is assumed that a clump is a single endocytic area instead of an overlapping of several ones. Besides, the mean size of an endocytic area would be overestimated, as some of these clumps could contain several of them. In order to improve the estimation of the mean perimeter u_o, some clumps could be ruled out. In particular, only those clumps composed by just one endocytic spots could be selected. The mean perimeter would be then estimated as the corresponding mean. However, this selection procedure would be again biased since smaller endocytic spots are selected with a higher probability. Only an approach based on Boolean models would lead to unbiased results.

3.5 Conclusions

We have proposed a statistical methodology based on Boolean models to study endocytosis in living cells. Our methodology enabled us to estimate the mean number of endocytic spots per unit area in a robust way. Although the number of endocytic spots can be studied by using electrical techniques such as capacitance measurements, no spatial information is obtained. To our knowledge, it is the first time that these estimates from images are used in this context.

Current studies tend to ignore the overlapping between endocytic areas and calculate the mean area by taking into account only the smallest clumps, i.e., those clumps that may correspond to a single endocytic spot. Nonetheless, using this simple approach the sample would be biased, and the size would be underestimated. By contrast, robust estimators for the mean number of spots per unit area in a given frame and its mean apparent size (perimeter of its point-spread function) are proposed here. Additionally, the estimated values obtained from different frames are jointly used to provide global estimators for both parameters using the batch-mean method. The proposed method provides the experimenter with a robust and formal methodology for the quantitative study of the cell endocytic behavior in response to different levels of stimulation, treatments, etc. Moreover, this type of analysis may provide mechanistical information on the endocytic process. For instance, the role of other proteins such as GTPase dynamin in the endocytosis of clathrin-coated vesicles may be better elucidated.

An interesting extension of the model is the analysis of fluorescence in grey-

level images. In this case, the parameter to be estimated would be the energy (fluorescence) associated with a single endocytic spot. This approach will allow us to estimate if the fluorescence content of endocytic spot is quantized, and to determine the quantal of fluorescence associated to a single endocytic spot.

It has been recently proposed that the minimum cluster of fluorescent clathrin molecules that can be detected in an image might correspond to one single endocytic vesicle (19). Under this hypothesis, the number of clathrin coated vesicles would correspond with the number of endocytic spots.

It is important to note that this framework is flexible and general enough, as well as intuitive and direct, to incorporate biological hypotheses. In fact, although we have assumed that the different endocytic spots are independent realizations of a given random set and they are independently and uniformly located within a given region of the plasma membrane, this framework is the natural one to incorporate a non-homogeneous distribution of germs if required.

Our model may be applied to the study of other biological and biophysical processes. One example is the analysis of the distribution and dynamic behavior of single molecules, such as cellular proteins conjugated to fluorophores, as they interact with biological surfaces (i.e. the plasma membrane) and non-biological ones (catalyst surfaces).

Finally, we should point out that the presented model does not take into account the temporal dimension of the endocytosis phenomenon. However, a further expansion which fill this gap is developed in Chapters 4 and 5. It is called Spatio-temporal Boolean model and will allow us to estimate the mean duration of the endocytic spots ("endocytic delay"), as well as its temporal intensity λ (mean number of endocytic spots per unit time and unit area). These parameters will allow us to better characterize the complex process of vesicle endocytosis. Both extensions of the work can be approached in a natural way within the context of Boolean models.

Chapter 4

Analysis of Spatially and Temporally Overlapping Events

4.1 Introduction

The problems of counting overlapping objects in an image and estimating their shape-size features are common in many real applications. In the previous chapter we show how Boolean models explicitly consider and assume this overlapping. Moreover, we describe a methodology to analyze some of the components of the process of endocytosis.

In spite of the considerable number and variety of applications of Boolean models, the problem of analyzing spatially and temporally overlapping events in image sequences remains largely elusive. More sophisticated models 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 in time. Recent advances in microscopy allow us to capture image sequences of dynamic processes with very high spatial and temporal resolution, such as biological processes. Furthermore, the growth in the volume of data requires specialized methods for effective data analysis, in particular for extracting useful information and for the interpretation of results.

In this chapter, we aim at studying the dynamics of the clathrin-GFP protein which is involved in the endocytosis process. In general, we are concerned with the analysis of overlapping short-lived events from image sequences. Traditionally, obtaining this kind of data has been a very time-consuming process where much of the work had to be done manually, making it virtually impossible in large image sequences. We are interested not only in the study of the static parameters obtained image by image, but in the dynamic parameters of the process as the mean number of endocytic spots per unit area and time as well as their duration distribution. This would permit to obtain an estimation of the duration distribution of Clathrin-GFP from image sequences and should help lead the way to new automated high content endocytic screens. In order to obtain these new spatial temporal estimates, we have extended the classical concept of Boolean model, defining what we call *the spatio-temporal Boolean model*, a particular case of a non-isotropic $3D$ Boolean model. Roughly, a spatio-temporal Boolean model (from now on STBM) is a Poisson point process (producing the locations of germs) coupled with an independent random shape process (the grains) and an independent time duration process for grains (the durations). In the following sections we will examine some of the probabilistic properties of the proposed model and will study the estimation of different characteristics of the model from a time-lapse image sequence. The methodology and the estimators proposed can be safely applied to the study of other proteins or similar spatial temporal patterns.

In Section 4.2 the STBM is introduced, some probabilistic properties and a statistical analysis are given. A wide simulation study to show the relative errors of the proposed estimators is given in Section 4.3. In Section 4.4 an application of the model to Cell Biology is detailed. Finally, conclusions are given in Section 4.5.

4.2 Spatio-Temporal Boolean models

In this section we consider a particular Boolean model defined in $\mathbb{R}^3$. The primary grain Ξ_0 is the product of a 2D random set A_0 and an independent random interval $[0, d_0]$ i.e. $\Xi_0 = A_0 \times [0, d_0]$. The i-th germ is $y_i = (x_i, t_i)$ where x_i is the spatial location and t_i is the birth time of the i-th event.

4.2.1 Boolean model

As we showed in the previous chapter, a basic problem in the statistical analysis of Boolean models is the estimation of: (i) the intensity λ of the germ process; (ii) the mean area of the primary grain, a_0; and (iii) the mean perimeter, $u_0 = EU(A_0)$, where $U(A)$ denotes the perimeter. 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 (38; 62). 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 (34; 42).

In this section, Boolean models are extended to incorporate the temporal dimension. Graphically, a spatio-temporal Boolean model consists of a set of events with location x_i, occurrence time t_i and duration d_i, as shown in Fig. 4.1. In a given cross-section (frame) s_i, each event is observed as a random area located around the position of the germ as in a 'static' Boolean model (see Fig. 4.3 (a)). Cylinders correspond to the different events in the sequence. It is important to note that there is overlapping in space and time and that the degree of overlapping in a given frame is not only dependent upon the sizes of the grains but also upon their durations. See videos 4.1, 4.2 and 4.3 in the supplementary material (more details in appendix C).

Our primary information is the superposition of the different events in space and time associated to the black areas in the image sequence. Fig. 4.2 displays several consecutive frames of a realization of a spatio-temporal Boolean model. Grains appear in the image, stay for a while and then disappear. For example, the grain numbered 1 is visible from frame 2 to frame 8. The grain numbered 2 lasts from frame 4 to 7. These frames belong to an image sequence which was generated using uniform disks with radii in the interval $[8, 12]$ pixels and an exponential duration distribution with a mean of 6 seconds. The intensity of grains is 0.00008. Images are 256×256 in pixels and were sampled at 2 frames per second.

We will assume that the locations x_i's, the occurrence times t_i's and the durations d_i's are independent. This is a basic hypothesis that seems natural for many applications and convenient because it simplifies the stochastic model.

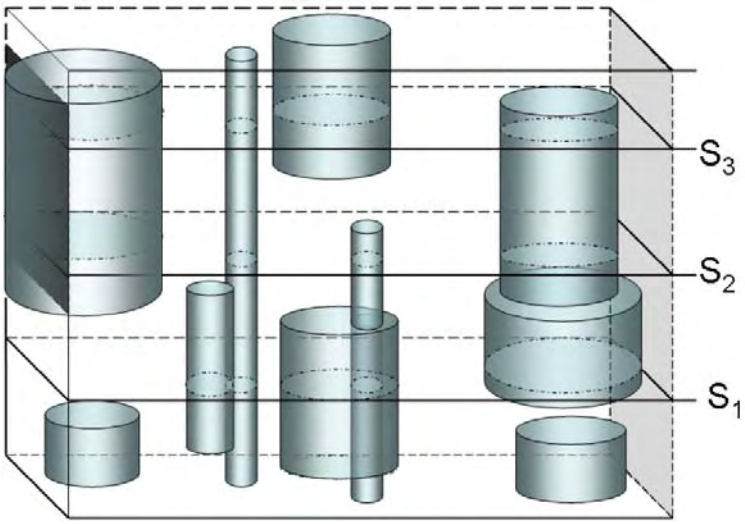


Figure 4.1: **An example of a spatio-temporal Boolean model.** Cylinders represent the grains.

Definition 2 (Spatio-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* $\{A_i\}_{i\geq 1}$ *be a sequence of independent and identically distributed (as* A_0*) random compact sets in* $\mathbb{R}^2$. *Let* $\{d_i\}_{i\geq 1}$ *be a sequence of independent and identically distributed (as* D*) positive random variables. We assume that* Ψ, $\{A_i\}_{i\geq 1}$ *and* $\{d_i\}_{i\geq 1}$ *are independent and that* $E\nu_3(A_0 \times [0, D] \oplus \check{K}) < +\infty$ *for any compact subset* K *of* $\mathbb{R}^3$. *The* spatio-temporal Boolean model *is the random set defined as*

$$\Phi = \cup_{i\geq 1}(A_i + x_i) \times [t_i, t_i + d_i]. \tag{4.1}$$

The set $(A_i + x_i) \times [t_i, t_i + d_i]$ is a cylinder in $\mathbb{R}^2 \times \mathbb{R}_+$, the i-th event. Fig. 4.1 displays a realization of a spatio-temporal Boolean model where A_0 is a random disc. Nonetheless, the primary grain could adopt arbitrary shapes. The temporal cross-sections are (static) 2D Boolean models. Fig. 4.3 (a) displays some consecutive temporal cross-sections. We have spatially and temporally overlapping among the different events.

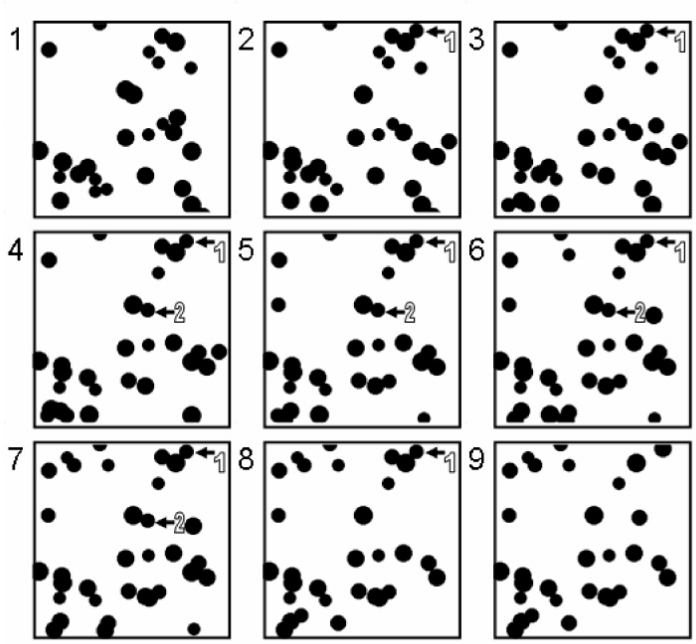


Figure 4.2: **Several shots of a spatio-temporal Boolean model.**

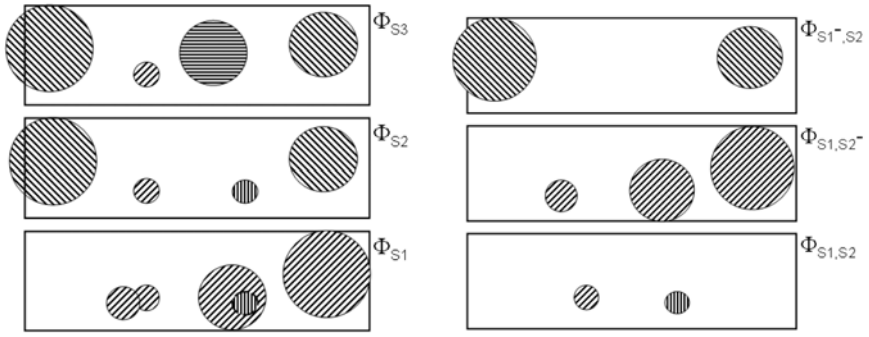


Figure 4.3: **Extraction of cross-sections.** (a) Three consecutive temporal cross-sections derived from Fig. 1. (b) The three Boolean models derived.

4.2.2 Statistical analysis

Our data consist of a discrete set of temporal cross-sections of the model corresponding to the observation times $s_1 \leq \ldots \leq s_k$. Let us denote the temporal cross-section at time s as $\Phi_s = \Phi \cap (\mathbb{R}^2 \times \{s\})$.

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. The projection of Φ_s over $\mathbb{R}^2$ (that with an abuse of notation will be denoted Φ_s) is given by

$$\Phi_s = \cup_{\{i:t_i \leq s \leq t_i+d_i\}}(A_i + x_i), \tag{4.2}$$

i.e. Φ_s is the binary image observed at frame s with the superposition or union of the different black areas associated to those events started before s but remaining at that time s.

The point process Ψ is a (spatially and temporally) stationary Poisson point process in $\mathbb{R}^2 \times \mathbb{R}_+$. Let λ be the intensity of this process, i.e. the mean number of points per unit area and time. Let us first consider the temporal point process composed of the times t_i's whose x_i is located within W. Note that this point process is a random thinning of the original point process. It is well-known that a random thinning of a Poisson point process is a Poisson point process (66). This fact will be a basic argument in the following proofs. The temporal point process is a (temporally) stationary Poisson point process defined in $\mathbb{R}_+$ with temporal intensity given by

$$\lambda_0 = \lambda \nu_2(W). \tag{4.3}$$

Analogously, the point process defined as the locations x_i's such that the corresponding times are in the interval $[0,T]$ is a (spatially) stationary Poisson point process in $\mathbb{R}^2$ with spatial intensity

$$\lambda_1 = \lambda T. \tag{4.4}$$

The set Φ_s, the cross-section at time s, is composed of those x_i's such that their associated time t_i and duration d_i verify that $t_i \leq s \leq t_i + d_i$. Note that t_i and d_i are independent of each other and are independent of their location x_i, therefore we again have a random thinning of the original point process and the following results hold.

If Φ is a spatio-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 and intensity

$$\lambda_s = \lambda E D. \tag{4.5}$$

This result can be found in (35; 62).

A 2D cross-section of a 3D Boolean model is also a Boolean model (36; 66). Since Φ has as primary grain $A_0 \times [0, D]$, it is obvious that Φ_s has as primary grain A_0. The random sets Φ and Φ_s have the same area fraction, therefore $1 - \exp\{-\lambda E\nu_2(A_0)ED\} = 1 - \exp\{-\lambda_s E\nu_2(A_0)\}$ and Eq. (4.5) follows.

Let us consider the estimation of the distribution of the random duration for the events, D. For $s_1 \leq s_2$, we will consider the following three random sets:

$$\Phi_{s_1,s_2} = \cup_{i:t_i \leq s_1 \leq s_2 \leq t_i+d_i} A_i + x_i, \tag{4.6}$$

$$\Phi_{s_1,s_2^-} = \cup_{i:t_i \leq s_1 \leq t_i+d_i < s_2} A_i + x_i, \tag{4.7}$$

$$\Phi_{s_1^-,s_2} = \cup_{i:s_1 < t_i \leq s_2 \leq t_i+d_i} A_i + x_i. \tag{4.8}$$

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. 4.3 (b)). This restriction 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 (16; 55).

Fig. 4.4 illustrates of these sets. Indeed, Fig. 4.4 shows three different cross sections corresponding to two frames plus an intermediate frame. This intermediate frame has been included in order to better explain the meaning of the formulae, but it would not available since it do not correspond to a sampling time. The frames correspond to three cross sections of a spatio-temporal Boolean model, where there are some events which last more than the three sections (actually 2 frames) and therefore are part of boolean model Φ_{s_1,s_2}. An example of an event taking part in this boolean model is shown in Fig. 4.4 (white dot), where an event starts at time t_{i-1}, before s_1, and dies after an unspecified time (after frame s_2). On the other hand, there are events that have appeared before frame s_1 but died before s_2. Those events define the Boolean model denoted by $\Phi_{s_1,-s_2}$. In Fig. 4.4 (black dot) the event that appears at t_i and dies at $t_i + d_i$ belongs to $\Phi_{s_1,-s_2}$. Finally, there is one last Boolean model denoted by Φ_{-s_1,s_2} (dashed dot) which contains all those events that appear between frames s_1 and s_2 and die whenever over the remaining time (or simply never dies). In Fig. 4.4 there is an event that appear at time t_{i+1} and dies at some time after frame s_2, therefore it belongs to the Boolean model Φ_{-s_1,s_2}.

Theorem 1 *Let Φ be a spatio-temporal Boolean model. The random sets Φ_{s_1,s_2}, Φ_{s_1,s_2^-} and $\Phi_{s_1^-,s_2}$ are then independent 2D Boolean models. Let us denote $\beta(s_2 -$*

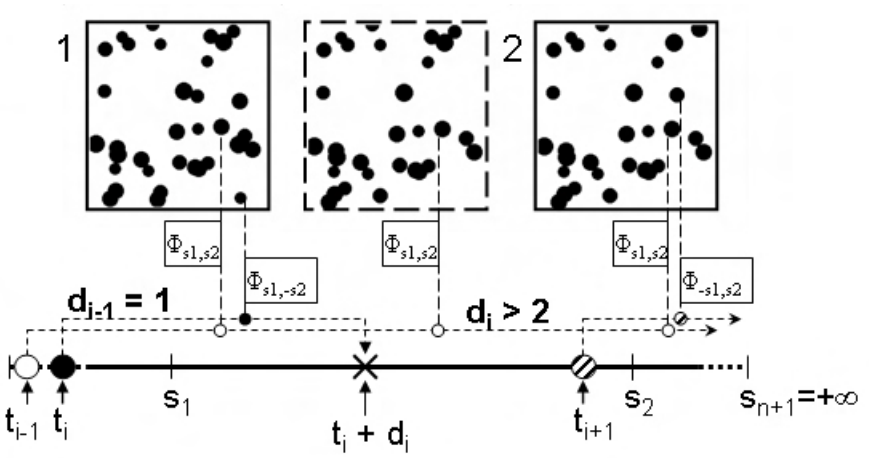


Figure 4.4: **Definition of the different boolean models from a spatio-temporal Boolean model and two cross sections**.

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*

$$\beta(s_2 - s_1) = \lambda p(s_2 - s_1), \tag{4.9}$$

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 a time interval* $[s_1, s_2]$*) given by*

$$\alpha(s_2 - s_1) = \lambda ED - \lambda p(s_2 - s_1) = \lambda_s - \beta(s_2 - s_1). \tag{4.10}$$

Proof 1 *Let* $\xi = \{t_n\}_{n\geq 1}$ *be a Poisson point process in* $\mathbb{R}$ *with intensity* θ*, let* $\{d_n\}_{n\geq 1}$ *be a sequence of independent and identically distributed (as* D*) non-negative random variables. Let* $s_0 = -\infty < s_1 < s_2 < \ldots < s_k < s_{k+1} = +\infty$ *be fixed points in* $\mathbb{R}$ *and* $N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = \#\{n : s_{i-1} < t_n \leq s_i, s_{j-1} < t_n + d_n \leq s_j\}$*, i.e. the number of points of* ξ *which were born in the time interval* $(s_{i-1}, s_i]$ *and died in the time interval* $(s_{j-1}, s_j]$ *where* $j > i$*. Let us consider the point process* $\{t_n \in \xi : t_n \leq s_1 < s_i < t_n + d_n < s_{i+1}\}$*, a thinning of* ξ*, and therefore an inhomogeneous Poisson point process with intensity function* $r(t) = \theta P(s_i \leq$

$t + D < s_{i+1}$) if $t \leq s_1$ and 0 if $t > s_1$. The total number of points of the thinned process, $N^{(s_i,s_{i+1})}_{(-\infty,s_1]}$, is a Poisson random variable with mean

$$
\begin{aligned}
&EN^{(s_i,s_{i+1})}_{(-\infty,s_1]} = \theta \int_{-\infty}^{+\infty} r(v)dv = \\
&\theta \int_{-\infty}^{s_1} P(s_i \leq D + v < s_{i+1})dv = \theta \int_{0}^{+\infty} P(s_i - s_1 \leq D - v < s_{i+1} - s_1)dv = \\
&\theta \int_{s_i - s_1}^{+\infty} P(D \geq v)dv - \theta \int_{s_{i+1} - s_1}^{+\infty} P(D \geq v)dv = \theta(p(s_i - s_1) - p(s_{i+1} - s_1)).
\end{aligned}
\tag{4.11}
$$

Let $\Psi = \{(x_i, t_i)\}_{i \geq 1}$ be a stationary Poisson point process in $\mathbb{R}^2 \times \mathbb{R}_+$ of intensity λ, and $\{d_i\}_{i \geq 1}$ a sequence of independent and identically distributed random variables. If B is a Borel subset of $\mathbb{R}^2$ then $\Psi_B = \{x_i : (x_i, t_i) \in \Psi, (t_i, d_i) \in B\}$ is a random thinning of Ψ and, as is well-known, the point process Ψ_B is a Poisson point process. If we take $B = \{(u, v) : u \leq s_1 < s_2 \leq u + v\}$, then Ψ_B is the germ process associated to Φ_{s_1,s_2}. The intensity of this process, $\beta(s_2 - s_1)$, follows from Eq. (4.11) by taking $s_i = s_2$ and $s_{i+1} = +\infty$. As the different grains A_i are independent and identically distributed random closed sets and the corresponding germ process is a Poisson point process, the proof is completed.

We can similarly prove that Φ_{s_1,s_2^-} is a Boolean model by considering now $B = \{(u, v) : u \leq s_1 \leq u + v < s_2\}$. The addition of the intensities of Φ_{s_1,s_2} and Φ_{s_1,s_2^-} is equal to the intensity of Φ_{s_1} (equal to the intensity of Φ_{s_2}), i.e. equal to $\lambda_s = \lambda ED$. The same proof applies to $\Phi_{s_1^-,s_2}$.

It holds that $\lambda_s = \alpha(s) + \beta(s)$, 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 we can observe

$$\begin{aligned}\Phi_{s_1} &= \Phi_{s_1,s_2} \cup \Phi_{s_1,s_2^-},\\ \Phi_{s_2} &= \Phi_{s_1,s_2} \cup \Phi_{s_1^-,s_2},\\ \Phi_{s_1} \setminus \Phi_{s_2} &= \Phi^c_{s_1,s_2} \cap \Phi_{s_1,s_2^-} \cap \Phi^c_{s_1^-,s_2},\\ \Phi_{s_2} \setminus \Phi_{s_1} &= \Phi^c_{s_1,s_2} \cap \Phi_{s_1^-,s_2} \cap \Phi^c_{s_1,s_2^-},\end{aligned}$$

where $A \setminus B = A \cap B^c$ denotes the set subtraction.

Proposition 1 *Let Φ be a spatio-temporal Boolean Model and Φ_{s_1} and Φ_{s_2} two temporal cross sections with $s_1 < s_2$. The probability that a point is covered only in one of the cross section, i.e, s_1 but not s_2 or s_2 but not s_1, is given by the following expression,*

$$P(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2}) = P(0 \in \Phi_{s_2} | 0 \notin \Phi_{s_1}) = 1 - \exp\{-\alpha(s_2 - s_1)a_0\}. \quad (4.12)$$

Proof 2 *Φ_s is a Boolean model with intensity λ_s and primary grain A_0. By taking into account Theorem 1 it follows that*

$$\begin{aligned}P(0 \in \Phi_{s_1} \setminus \Phi_{s_2}) &= P(0 \in \Phi^c_{s_1,s_2}) P(0 \in \Phi_{s_1,s_2^-}) P(0 \in \Phi^c_{s_1^-,s_2}) =\\ &\exp\{-\beta(s_2 - s_1)a_0\}(1 - \exp\{-\alpha(s_2 - s_1)a_0\}) \exp\{-\alpha(s_2 - s_1)a_0\} =\\ &\exp\{-\lambda_s a_0\}(1 - \exp\{-\alpha(s_2 - s_1)a_0\}), \quad (4.13)\end{aligned}$$

and

$$P(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2}) = \frac{P(0 \in \Phi_{s_1} \setminus \Phi_{s_2})}{P(0 \notin \Phi_{s_2})} = 1 - \exp\{-\alpha(s_2 - s_1)a_0\}. \quad (4.14)$$

Let ϕ_{s_i} be the binary image observed at time s_i, then $P(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2})$ will be estimated as

$$\hat{P}(0 \in \Phi_{s_1} | 0 \notin \Phi_{s_2}) = \frac{\hat{P}(0 \in \Phi_{s_1} \setminus \Phi_{s_2})}{\hat{P}(0 \notin \Phi_{s_2})} = \frac{\nu_2(\phi_{s_1} \cap \phi^c_{s_2} \cap W)}{\nu_2(\phi^c_{s_2} \cap W)}, \quad (4.15)$$

and by using Eq. (4.12) and Eq. (4.15), we will estimate $\delta(s_2 - s_1) = \alpha(s_2 - s_1)a_0$ by solving the equation

$$1 - \exp\{-\hat{\delta}(s_2 - s_1)\} = \frac{1}{2}\left(\frac{\nu_2(\phi_{s_1} \cap \phi^c_{s_2} \cap W)}{\nu_2(\phi^c_{s_2} \cap W)} + \frac{\nu_2(\phi_{s_2} \cap \phi^c_{s_1} \cap W)}{\nu_2(\phi^c_{s_1} \cap W)}\right). \quad (4.16)$$

The mean area (a_0) and the mean perimeter (u_0) of A_0 and λ_s were estimated from each image in the sequence by using the minimum contrast method (38). For a compact set K we can estimate the capacity functional $T(K)$ using the minus-sampling estimator from a given realization (i.e. from each binary image of the sequence) by

$$\hat{T}_W(K) = \frac{\nu_2((\Phi \oplus \breve{K}) \cap (W \ominus K))}{\nu_2(W \ominus K)}, \tag{4.17}$$

where W is the sampling window and $A \ominus \breve{K} = \{x \in A : x + K \subset A\}$. We estimate $T(B(0,t))$ for different t values and fit a quadratic polynomial to $-\log(1-\hat{T}(B(0,t)))$. If $c_0 + c_1 t + c_2 t^2$ is the fitted function then the parameters can be estimated as

$$\hat{\lambda} = \frac{c_2}{\pi},\ \hat{u}_0 = \frac{c_1}{\hat{\lambda}} \ \text{ and } \ \hat{a}_0 = \frac{c_0}{\hat{\lambda}}. \tag{4.18}$$

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

$$\hat{\alpha}(s) = \frac{\hat{\delta}(s)}{\hat{a}_0} \text{ and } \hat{\beta}(s) = \hat{\lambda_s} - \hat{\alpha}(s). \tag{4.19}$$

From Eq. (4.10), 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 $\alpha(s)$ is estimated at the observation points $\{s_1, \ldots, s_n\}$. We will estimate its first and second derivatives at any point, 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. We did not use an interpolation process because we assumed that the discrete values may include some observational error. Instead, we used a smoothing technique 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. We chose a polynomial spline basis where each $\Phi_k(t)$ is a piecewise cubic function. We obtained the coefficients c_k of the expression $y(t)$ by minimizing the least squares criterion

$$SMSSE(y/c) = \sum_{j=1}^{n} (y_j - \sum_{k=1}^{K} c_k \Phi_k(\delta_j))^2.$$

A detailed presentation of this method is given in (50). 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

$$\hat{\lambda} = \hat{\alpha}'(0). \tag{4.20}$$

The probability density of D, f_D, was estimated as

$$\hat{f}_D(s) = -\frac{1}{\hat{\lambda}}\hat{\alpha}''(s). \tag{4.21}$$

4.3 A simulation study

In this section we evaluate the precision of the proposed estimators. The parameters to be estimated are: λ, the spatial-temporal intensity of germs; a_0 and u_0, the mean area and perimeter of the primary grain; ED, the mean duration; λ_s, the intensity of the Boolean model Φ_s; the functions α and β and the density function f_D of the durations.

In order to design the simulation study it is convenient to take into account the spatial temporal volume fraction p. This volume fraction p depends on the spatial temporal intensity λ, the expected duration ED, and the mean area of the primary grain a_0 and is given by

$$p = 1 - \exp\{-\lambda a_0 ED\}. \tag{4.22}$$

Nine different simulations, combinations of the volume fractions $p = \{0.10, 0.15, 0.20\}$ and the mean durations $ED = \{6, 9, 12\}$ seconds were generated, fifteen replica for each (see supplementary material and appendix C). That was done for three different duration distributions (uniform, exponential and Gamma), thereby analyzing a total of $9 \times 15 \times 3 = 405$ simulated image sequences with fixed $\lambda = 0.00006$. Mean areas a_0's were derived from Eq. (4.22) for each combination of ED, p and λ. Radii were assumed to follow a uniform distribution in the interval $[ER - 2, ER + 2]$. Table 4.1 summarizes all parameter combination used to generate the image sequences generated in the simulation study. Image sequences were 512×512 pixels in size, 150 seconds long and sampled at 2 frames per second. We analyzed more than 120000 frames.

Table 4.1: Parameters for the simulations

Simulation	Vol. Frac.	a_0	r_0	ED
#1	0.10	318.34	U(8,12)	6
#2	0.10	205.25	U(6,10)	9
#3	0.10	158.13	U(5,9)	12
#4	0.15	456.58	U(10,14)	6
#5	0.15	318.34	U(8,12)	9
#6	0.15	205.25	U(6,10)	12
#7	0.20	619.94	U(12,16)	6
#8	0.20	384.32	U(9,13)	9
#9	0.20	318.34	U(6,10)	12

(a) (b) (c)

Figure 4.5: **Three frames corresponding to three different STBM with low, medium and high volume fractions.**

Fig. 4.5 (a), (b) and (c), correspond to three images generated with low, medium and high volume fractions, with radii drawn from a uniform distribution with a mean of 10, 12 and 14 and a uniform duration with a mean of 6 seconds.

With regard to the random durations D, if $D \sim U(a,b)$ then $ED = \frac{a+b}{2}$ and $ED^2 = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}$; if $D \sim Exp(\theta)$ then $ED = \frac{1}{\theta}$ and $ED^2 = \frac{2}{\theta^2}$; and if $D \sim Ga(\theta_1, \theta_2)$, i.e. it is Gamma distributed, then $ED = \frac{\theta_1}{\theta_2}$ and $ED^2 = \frac{\theta_1+\theta_1^2}{\theta_2^2}$. In particular, for our simulation study with ED = 6, 9 and 12, the variances of the duration distribution for the uniform cases were $37.3, 82.3, 145.3$, for the exponential were $72, 162, 288$, and for the Gamma $42, 90, 156$, respectively.

In order to assess the errors of the proposed estimators, we used the relative error. If ξ is the parameter to estimate and $\hat{\xi}$ the corresponding estimator, the relative error is defined as

$$RE(\hat{\xi}) = \frac{\hat{\xi} - \xi}{\xi}.$$

Tables 4.2 and 4.3 show a summary of the observed relative errors for λ, λ_s, a_0 and u_0 with respect to the duration distribution and the volume fractions for the fifteen replica generated. The mean values were very small for all the parameters. In particular, the maximum bias in λ was larger when compared with the other parameters studied. We guess that this may be due to the estimation method which is based on the first derivative at the origin of $\alpha(s)$. The observed bias was more severe for higher volume fractions, though these errors would be good enough in many real-life applications.

Table 4.2: Relative errors with respect to the duration distribution

Parameter	Distribution	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
λ	Uniform	-0.084	-0.003	0.022	0.022	0.047	0.125
	Exponential	-0.057	0.016	0.053	0.055	0.090	0.153
	Gamma	0.127	-0.010	0.013	0.017	0.044	0.159
λ_s	Uniform	-0.086	-0.020	0.004	0.000	0.022	0.090
	Exponential	-0.093	-0.020	0.008	0.009	0.034	0.109
	Gamma	-0.109	-0.026	0.004	0.000	0.022	0.096
a_0	Uniform	-0.087	-0.049	-0.028	-0.025	-0.005	0.073
	Exponential	-0.105	-0.057	-0.031	-0.032	-0.010	0.074
	Gamma	-0.093	-0.047	-0.030	-0.027	-0.007	0.066
u_0	Uniform	-0.055	-0.014	0.005	0.009	0.028	0.096
	Exponential	-0.078	-0.023	0.002	0.002	0.025	0.113
	Gamma	-0.062	-0.010	0.006	0.007	0.027	0.087

We estimated the confidence intervals for λ_s, a_0, u_0 and $\alpha(s)$ by using the batch-mean method since we had correlated data (33). The block size obtained was around 7 images each. For the minimum contrast method, we used a square structuring element which gave us better results because of its digital approximation (38; 62). These confidence intervals were very small as shown in Table 4.4,

Table 4.3: Relative errors with respect to the volume fraction

Parameter	Vol. Frac.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
λ	0.10	-0.056	0.004	0.028	0.033	0.059	0.151
	0.15	0.083	-0.005	0.030	0.030	0.059	0.158
	0.20	0.127	0.001	0.033	0.031	0.067	0.144
λ_s	0.10	-0.086	-0.019	0.003	0.003	0.024	0.109
	0.15	-0.094	-0.019	0.005	0.004	0.025	0.100
	0.20	-0.109	-0.027	0.008	0.003	0.031	0.096
a_0	0.10	-0.092	-0.055	-0.032	-0.034	-0.016	0.057
	0.15	-0.103	-0.052	-0.032	-0.029	-0.006	0.040
	0.20	-0.105	-0.046	-0.024	-0.021	0.001	0.074
u_0	0.10	-0.062	-0.012	0.002	0.003	0.019	0.093
	0.15	-0.075	-0.020	-0.000	0.004	0.027	0.071
	0.20	-0.078	-0.010	0.010	0.011	0.035	0.113

where a summary of the observed ratios between the half-width of the confidence intervals and the mean of the estimates is displayed. We observed very small ratios due to the short confidence intervals. Mean ratios were around 5% for the three distributions, those of the exponential being the highest ones, rising to values of 12% due to its higher variances compared with the uniform and Gamma distributions.

In the $\alpha(s)$ estimation each replica provided us with an estimated function and its corresponding confidence interval. The theoretical expression of $\alpha(s)$ in Eq. (4.10) changes according to the $p(s)$ expression which depends on the temporal distribution used. In our study we applied 20 cubic bases to transform raw data estimated into a piece-wise function, which was appropriate for functions like $\alpha(s)$ re-sampled at 40 values.

Let us see the exact expression of the function $p(s)$ for the particular distributions of the random variable D used later in the simulation study.

If the random duration D has an exponential distribution with parameter θ (i.e. its density function is $f(t) = \theta \exp\{-\theta t\}$ for $t \geq 0$ and zero otherwise) then $p(s) = \exp\{-\theta s\}/\theta$.

Table 4.4: Summaries of the ratios between the half-width of the confidence Intervals and the mean.

	Unif (min/mean/max)	Exp (min/mean/max)	Gamma (min/mean/max)
λ_s	0.024/0.051/0.089	0.026/0.055/0.124	0.024/0.051/0.103
a_0	0.020/0.042/0.069	0.022/0.046/0.127	0.019/0.043/0.091
u_0	0.021/0.041/0.068	0.021/0.045/0.129	0.020/0.042/0.084

If D is uniform in the interval $[a, b]$, then we have $p(s) = \frac{a+b}{2} - s$ if $0 \leq s \leq a$; $p(s) = \frac{(b-s)^2}{2(b-a)}$ if $a \leq s \leq b$ and zero if $s \geq b$.

Finally, a Gamma distribution with parameters θ_1 and θ_2 has a density function given by

$$f(t) = \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} t^{\theta_1 - 1} e^{-\theta_2 t} \text{ if } t > 0, \tag{4.23}$$

and 0 otherwise. It can be proved that

$$p(s) = \frac{\Gamma(\theta_1 + 1)}{\theta_2 \Gamma(\theta_1)} (1 - G(\theta_2 s, \theta_1 + 1)) - s(1 - G(\theta_2 s, \theta_1)), \tag{4.24}$$

where $\Gamma(\theta)$ is de Gamma function defined as,

$$\Gamma(\theta) = \int_0^{+\infty} t^{\theta - 1} \exp\{-t\} dt,$$

and $G(s, \theta)$ is the incomplete Gamma function defined as,

$$G(s, \theta) = \frac{1}{\Gamma(\theta)} \int_0^s t^{\theta - 1} \exp\{-t\} dt.$$

Because,

$$\begin{aligned}
p(s) &= \int_s^{+\infty}\int_t^{+\infty} \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1-1}e^{-\theta_2 u}dudt \\
&= \int_0^{+\infty}\int_0^{+\infty} 1_{[s,+\infty)}(t)1_{[t,+\infty)}(u)\frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1-1}e^{-\theta_2 u}dudt \\
&= \int_0^{+\infty}\int_0^{+\infty} 1_{[0,u]}(t)1_{[s,+\infty)}(u)\frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1-1}e^{-\theta_2 u}dudt \\
&= \int_0^{+\infty} (u-s)1_{[s,+\infty)}(u)\frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1-1}e^{-\theta_2 u}dudt \\
&= \int_s^{+\infty} \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1}e^{-\theta_2 u}dudt - s\int_s^{+\infty} \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1-1}e^{-\theta_2 u}dudt \\
&= \frac{\Gamma(\theta_1+1)}{\theta_2\Gamma(\theta_1)}(1-G(\theta_2 s,\theta_1+1)) - s(1-G(\theta_2 s,\theta_1)).
\end{aligned}$$

The last equality follows from the equation

$$G(s,\theta) = 1 - \frac{1}{\Gamma(\theta)}\int_s^{+\infty} u^{\theta-1}e^{-u}dudt, \tag{4.25}$$

$$\int_s^{+\infty} \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1}e^{-\theta_2 u}du = \frac{\Gamma(\theta_1+1)}{\theta_2\Gamma(\theta_1)}(1-G(\theta_2 s,\theta_1+1)), \tag{4.26}$$

and

$$\int_s^{+\infty} \frac{\theta_2^{\theta_1}}{\Gamma(\theta_1)} u^{\theta_1}e^{-\theta_2 u}du. \tag{4.27}$$

Figs. 4.6 (a), 4.7 (a) and 4.8 (a) correspond to simulated frames with low area fraction (0.10) following uniform, exponential and Gamma distribution, respectively. Fig. 4.6 (b) displays (solid line) the mean of the centers (the estimates of the α functions) and the corresponding means for the upper and lower extremes of the confidence intervals (dashed lines) for the 15 replica with duration uniformly distributed from 4 to 8. Figs. 4.7 (b) and 4.8 (b) show the same function using different temporal distribution, in particular, exponential and Gamma distributions with mean 6. For each replica we fit a polynomial spline function over the raw data and then we obtained the mean and interval functions, by averaging the function coefficients instead of the estimated values. Figs. 4.6 (c), 4.7 (c) and 4.8 (c) correspond to the first derivative of α. Figs. 4.6 (d), 4.7 (d) and 4.8 (d) display $\hat{f}_D$ by applying Eq. (4.21).

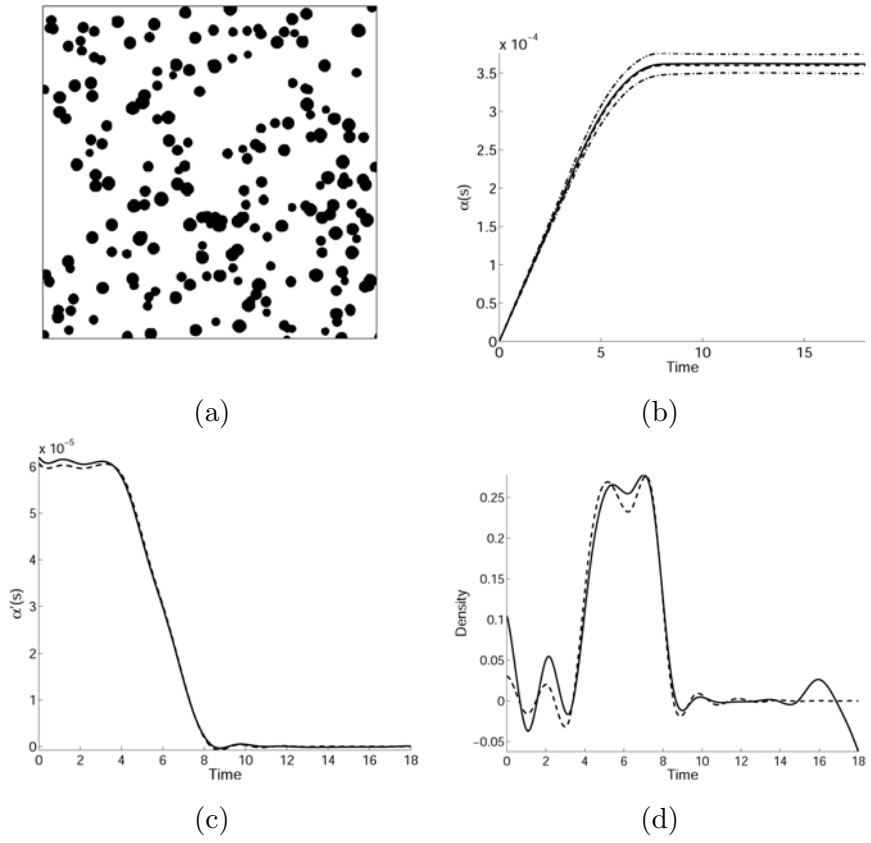


Figure 4.6: **Analysis of the** 15 **replica for uniform duration.** (a) A frame of a simulated video; (b) α-function; (c) α'-function; (d) probability density of durations, $\hat{f}_D$.

Since the duration distribution is obtained from $\alpha(s)$, it is interesting to analyze its relative error with respect to the duration distribution and the different volume fractions. The columns in Fig. 4.9 correspond to the duration distribution (uniform, exponential and Gamma) and the rows to the volume fractions $p = 0.10, 0.15, 0.20$. Each plot displays the minimum and maximum (dot-dashed lines), upper and lower quartile (dashed line) and the mean (solid lines) relative errors. The 3×15 image sequences generated for each combination of volume fractions and duration distributions were pooled to obtain each plot. The mean relative errors were really small and did not depend on p, though they did on the distribution. The worst case corresponded to the exponential and it is likely to

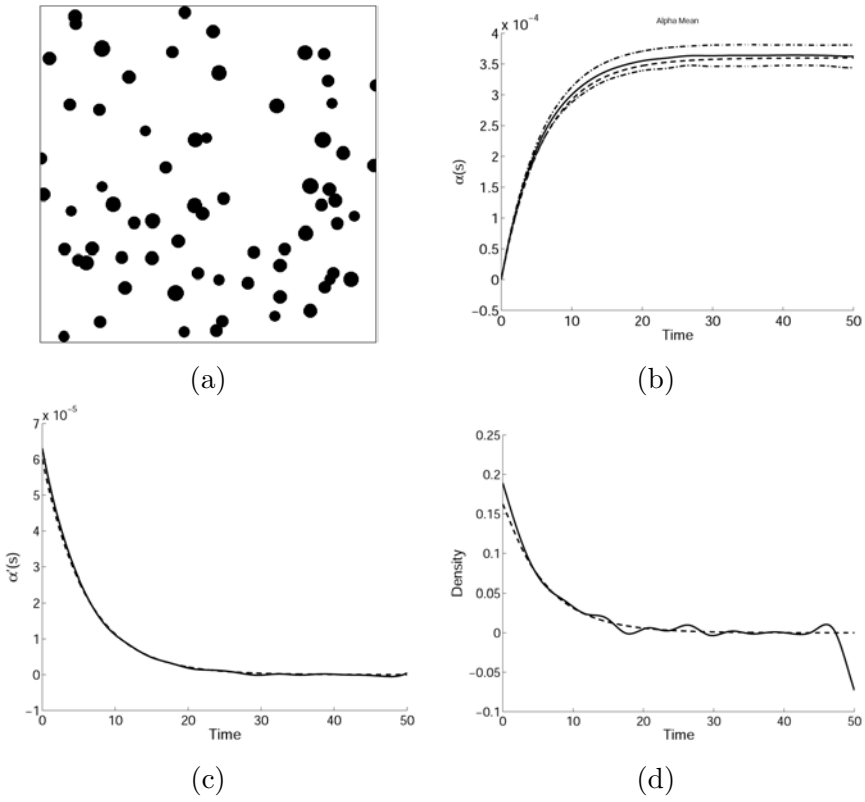
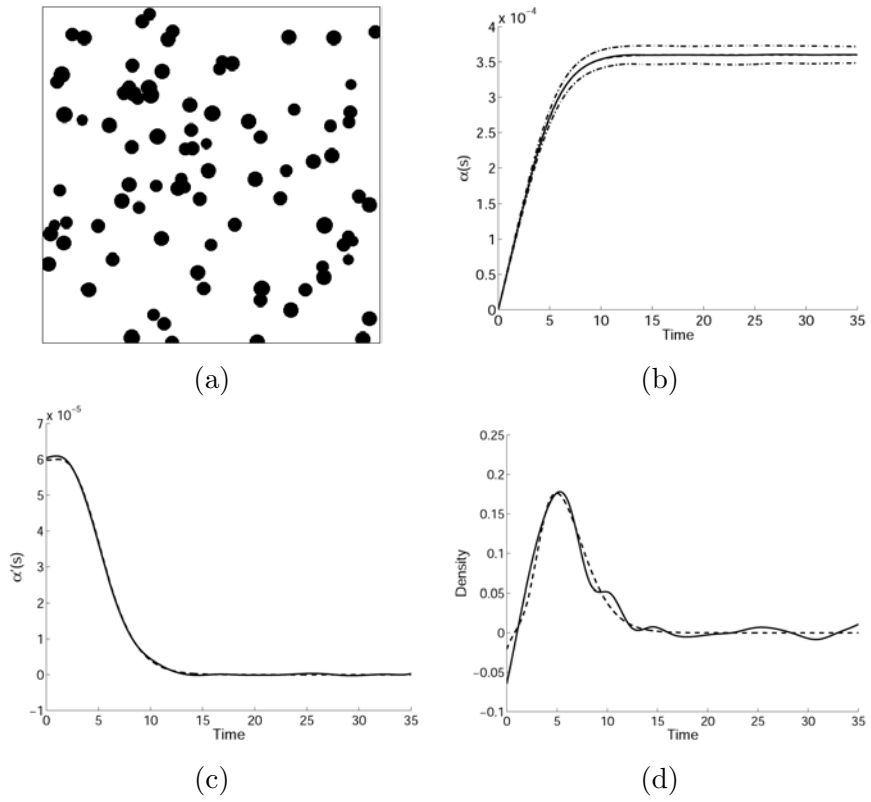


Figure 4.7: **Analysis of the** 15 **replica for exponential duration.** (a) A frame of a simulated video; (b) α-function; (c) α'-function; (d) probability density of durations, $\hat{f}_D$.

be due to the higher variance of the simulated durations.

Finally, Table 4.5 shows the errors for the mean of D (estimated as $\hat{p}(0)$) with respect to the duration distribution. In general, mean relative errors stayed small and were comparable with those in the other parameters.

A library of functions for the generation and estimation of STBM has been written in MATLAB (a trademark of The MathWorks Inc, USA) and is available on request. The toolbox FDA for MATLAB developed by J.O. Ramsay was used to perform the functional data analysis.



(a) (b) (c) (d)

Figure 4.8: **Analysis of the** 15 **replica for gamma duration.** (a) A frame of a simulated video; (b) α-function; (c) α'-function; (d) probability density of durations, $\hat{f}_D$.

4.4 An application to Cell Biology

STBMs were applied to the study of the biological process of endocytosis. Endocytosis happens in discrete events, in which cargo-loaded vesicles detach from the plasma membrane and are trafficked inside the cell. We employed STBMs to estimate the number of discrete endocytic events per unit area and time and their duration distribution. Clathrin-mediated endocytosis is one of the main endocytic routes. Endocytosis has recently been imaged in real time, by conjugating clathrin and other endocytic accessory proteins to fluorescent molecules and imaging cells with specialized microscopy techniques such as Total Internal Reflection Fluorescence Microscopy (TIRFM) (69). TIRFM allows selective illumination of

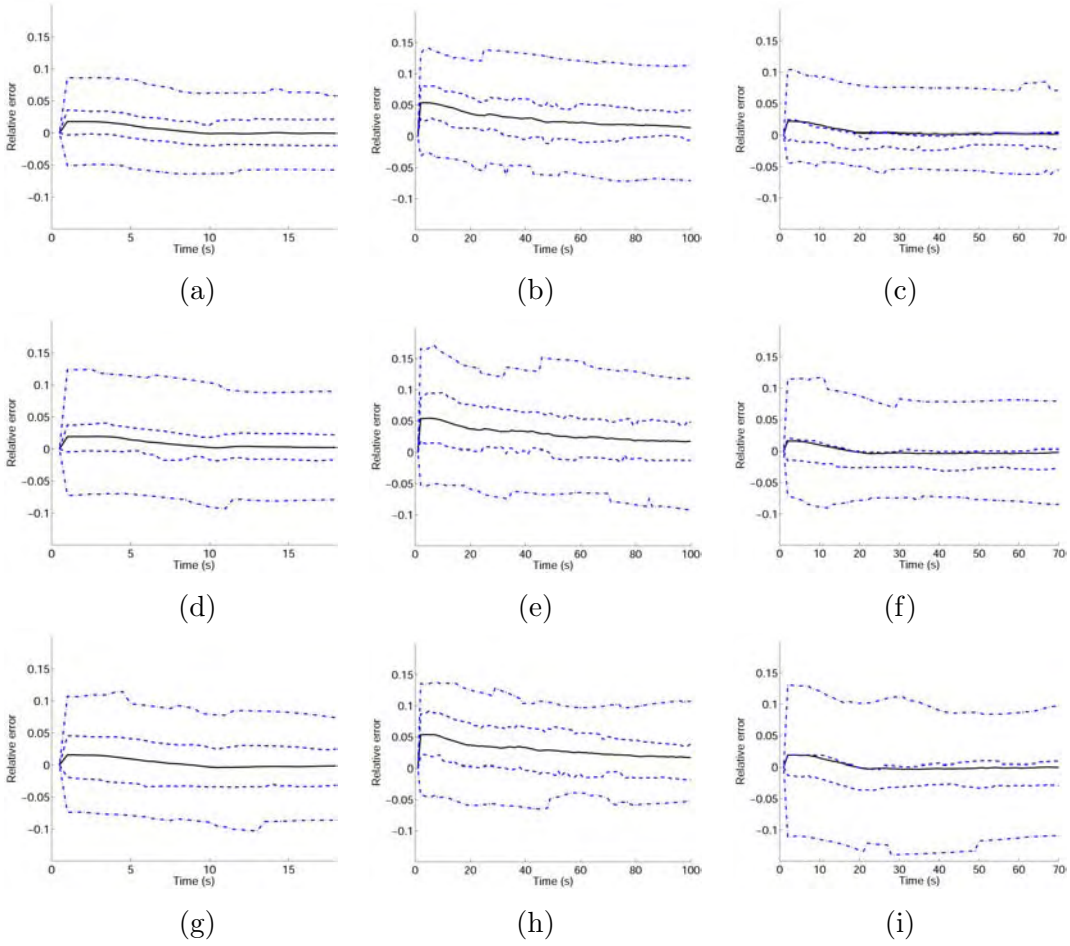


Figure 4.9: **Relative error of α-function and confidence region** with respect to the duration distribution (columns) and volume fraction (rows).

the cellular plasma membrane, thereby providing superior signal-to-noise. Under TIRFM, the assembly of fluorescently-labeled clathrin at a site of ongoing endocytosis results in the appearance and steady growth of a diffraction-limited spot. The time elapsed between the appearance and the disappearance of a fluorescent clathrin spot is defined as *the duration of a discrete endocytic event*. The areas of fluorescence associated to the different endocytic spots overlap, forming random clumps of different sizes, shapes and durations. The duration of discrete endocytic events, and their distribution in space and time, are influenced by many

Table 4.5: Relative errors for ED. Values Min/Mean/Max

ED	Uniform	Exponential	Gamma
6	$-0.031/0.002/0.042$	$-0.117/0.124/0.217$	$-0.044/0.047/0.170$
9	$-0.042/-0.013/0.014$	$-0.021/0.068/0.163$	$-0.158/-0.029/0.031$
12	$-0.057/-0.020/-0.003$	$-0.130/-0.061/0.009$	$-0.105/-0.035/0.003$

biological factors.

It is common practice in 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 of a single endocytic vesicle (19; 71). These techniques based on segmentation and labelling of the connected components lead to an underestimation of the number of vesicles and the duration of endocytosis (17).

We applied our estimators to six image sequences in which clathrin-GFP proteins had been previously segmented. The acquisition rate was 1 frame every four seconds. Table 4.6 describes these image sequences obtained from three cells.

Table 4.6: Description of the six sequences expressing clathrin-GFP

Cell	Seq.	Frames	$\nu_2(W)$
1	1	408	171×165
	2	408	222×184
2	3	306	147×205
	4	306	85×287
3	5	432	141×149
	6	432	132×224

The estimates and their confidence intervals obtained were obtained for the six sequences. Table 4.7 summarizes these results. Data obtained manually (30; 71) showed that the average life time of clathrin events was 73.8 ± 31.5 seconds. Fig. 4.10 shows four binary consecutive frames of clathrin images and the duration distribution obtained.

The estimated α functions and probability density of durations, f_d for each

Table 4.7: Estimated obtained for the six sequences expressing clathrin-GFP

Cell	Seq.	λ	ED (sec.)
1	1	0.000118	84.57
	2	0.000099	85.39
2	3	0.000325	70.81
	4	0.000117	83.07
3	5	0.000110	73.46
	6	0.000108	69.80

(a) (b) (c) (d)

Figure 4.10: **Snapshots of a cell expressing clathrin.**(a)-(d) are the frames of a clathrin video taken at times 1, 40, 80 and 120 s, from left to right.

cell is shown in Figs. 4.11, 4.12 and 4.13.

4.5 Conclusions

We have proposed a new approach for the analysis of objects that overlap spatially and temporally in image sequences. Here, from the original binary image sequence, we have stacked them into a 3D binary image. We have modelled this 3D object as a realization of a spatio-temporal Boolean model, a particular case of a non-isotropic 3D Boolean model. It brings significant improvements for the analysis of this kind of data, enabling us to estimate the mean number of grains per unit area and time, as well as the density function of the random duration, among other things. The proposed procedure is both robust and easy to implement. Functional data analysis techniques have permitted more accurate estimated functions, as well as in deriving and averaging them.

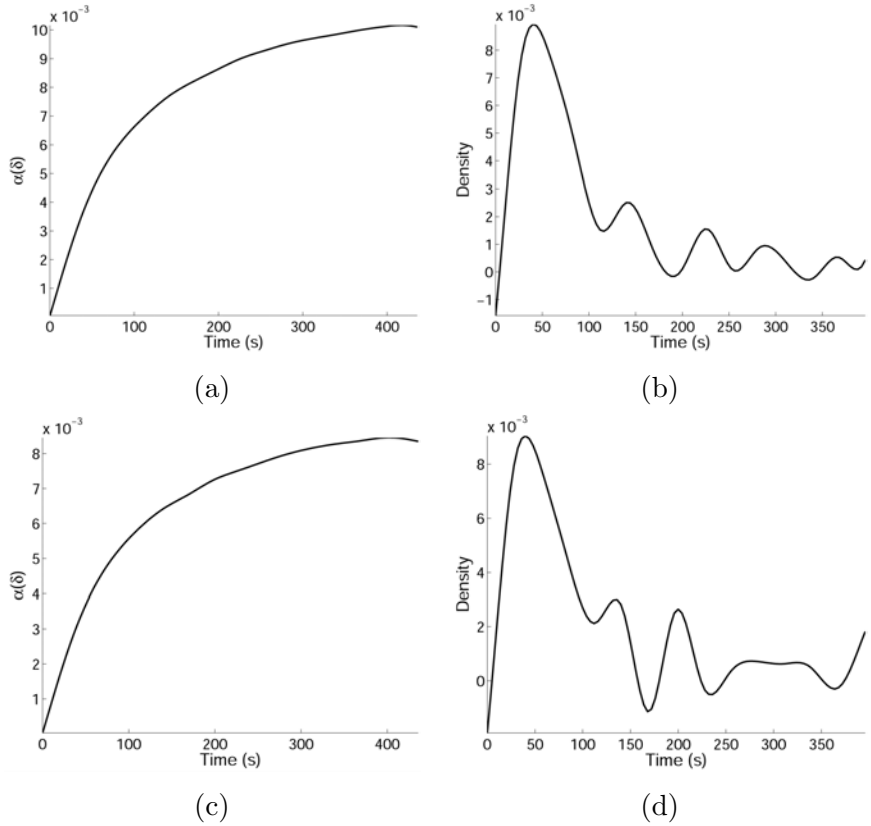


Figure 4.11: **Estimated functions of cell 1.**(a) and (c) Estimated α function; (b) and(d) probability density function of durations, $\hat{f}_d$. (a) and (b) correspond to sequence 1, and (c) and (d) to sequence 2.

A simulation study was performed to show the errors of the proposed estimators. More than 400 image sequences (120000 frames) were simulated with different duration distributions, volume fractions and grain size distributions. The relative error was calculated for all the parameters simulated. Results from the simulation study showed that the bias obtained was small enough to apply the proposed estimators on real-life applications.

We think this errors could be due to the spatial and temporal discretization of the image sequence. On the one hand, images were 512×512 pixels in size, with grain radii ranging from 2 to 14 pixels. Grains with such radii under that resolution are not really disks. On the other hand, the sampling rate was 2 frames

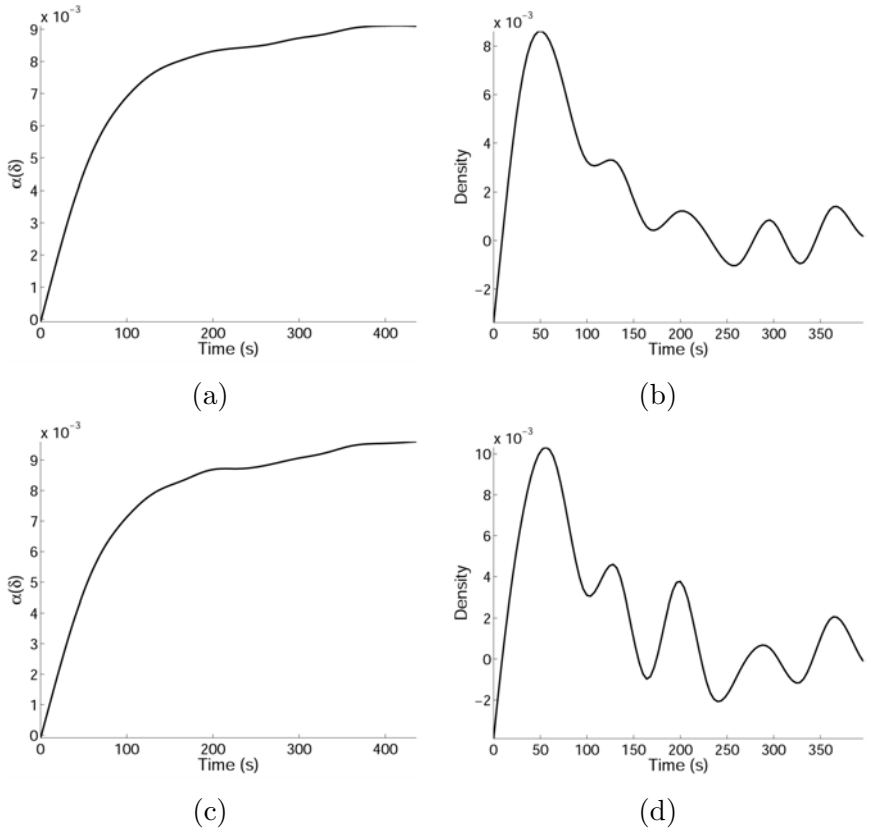


Figure 4.12: **Estimated functions of cell 2.**(a) Estimated α function; (b) probability density function of durations, $\hat{f}_d$. (a) and (b) correspond to sequence 3, and (c) and (d) to sequence 4.

per second, and the mean durations simulated ranged from 6 to 12 seconds. A study of the relative error behavior with respect to the image size and acquisition rate is still to be done. Lower values could be expected if higher acquisition rates and higher spatial resolution were used.

Our study implies dealing with functions estimated at a discrete set of values. An approach based on functional data analysis was used in order to obtain a precise estimation of the second derivative.

The use of STBMs can provide a powerful tool to analyze the behavior of the clathrin-dependent endocytic machinery across a range of different conditions in

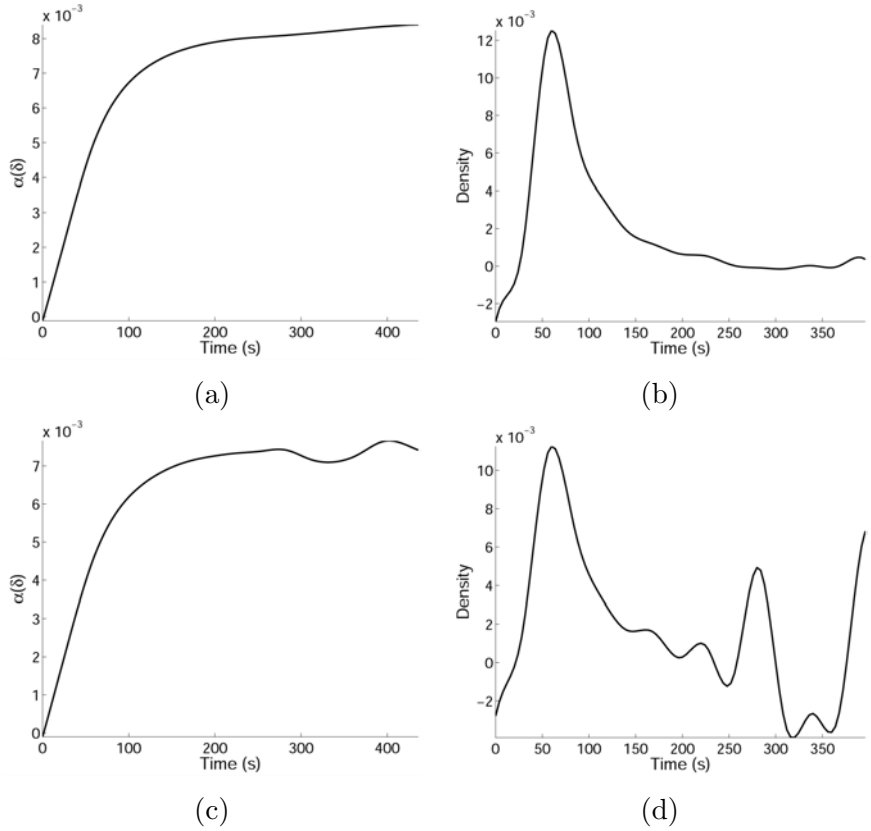


Figure 4.13: **Estimated functions of cell 3.** (a) Estimated α function; (b) probability density function of durations, $\hat{f}_d$. (a) and (b) correspond to sequence 5, and (c) and (d) to sequence 6.

a fast, accurate and automated way. In particular, the application of STBM to real image sequences of cells expressing clathrin protein allowed us to get accurate information never obtained before, such as the probability density function of durations of clathrin. STBM could be used as a robust screening tool to study the dynamics of in-vivo cells expressing certain proteins under different treatments. An important factor to take into account before applying the spatio-temporal Boolean model is the acquisition rate. If the rate is not fast enough a considerable part of the events may be born and dead between two consecutive frames, and so the estimation of the parameters would be biased.

Potential extensions to the proposed model include to assume a non-homogeneous

distribution of germs or occurrence times, since these hypotheses could not be tenable in some real applications. Following previous research concerning Boolean models (41; 48), it could be reasonable to estimate the covariance function of the STBM and, from it to estimate (assuming a parametric model for the primary grain) the distribution of the primary grain. These will be considered in future research.

We have implemented a software tool in MATLAB for simulating and estimating of spatio-temporal Boolean models. This tool is freely available on request. For more detail see appendix B.

Chapter 5

Estimation of STBMs by 2D Boolean Model Aggregation

5.1 Introduction

As we discussed in the previous chapter, the Boolean model is a static stochastic model, no temporal component is incorporated into its definition. A natural extension of this model has been proposed by adding the temporal dimension and it has been used to study some biological mechanisms in cell plasma membranes. Boolean models are extended in order to capture both temporal and spatial properties while formalizing a configuration of independent randomly placed particles with random durations in time. Hence, this model allows us to study jointly the spatial and temporal behavior of overlapping short-lived events from image sequences.

The estimation procedure that we propose in this chapter is based on the analysis of the image sequences generated after iteratively aggregating several frames of the original sequence into one frame. Such an addition of frames brings a new spatio-temporal Boolean model with different parameters, due to the properties of temporal Boolean models. In fact, in the aggregated model the grains keep the same size distribution whereas the spatial temporal intensity increases. The study of such an increment allows us to obtain the distribution of the durations. In previous chapter, the estimation of the parameters was based on the joint analysis of only two frames. Now, we base our estimators on the union of

more than two temporal cross sections of the original spatio-temporal Boolean model. Each temporal cross section is a Boolean model and, as is well-known, the union of Boolean models is again a Boolean model (see Section 5.2). This is our approach here.

Let Φ be a temporal Boolean model and $\Phi_{s_1}, \ldots, \Phi_{s_k}$ temporal cross sections where $s_1 < s_2 < \ldots < s_k$ then $\Phi_{s_1} \cup \ldots \Phi_{s_k}$ is a Boolean model too. Perhaps it is simpler to illustrate the result with $k = 2$. Then

$$\Phi_{s_1} \cup \Phi_{s_2} = \Phi_{s_1,s_2} \cup \Phi_{s_1,s_2^-} \cup \Phi_{s_1^-,s_2}$$

where the three random sets in the right hand side of the equation are independent random sets. Note that s^- indicates that the primary grain considered does not belong to the temporal cross section at s. For $k = 3$, we have

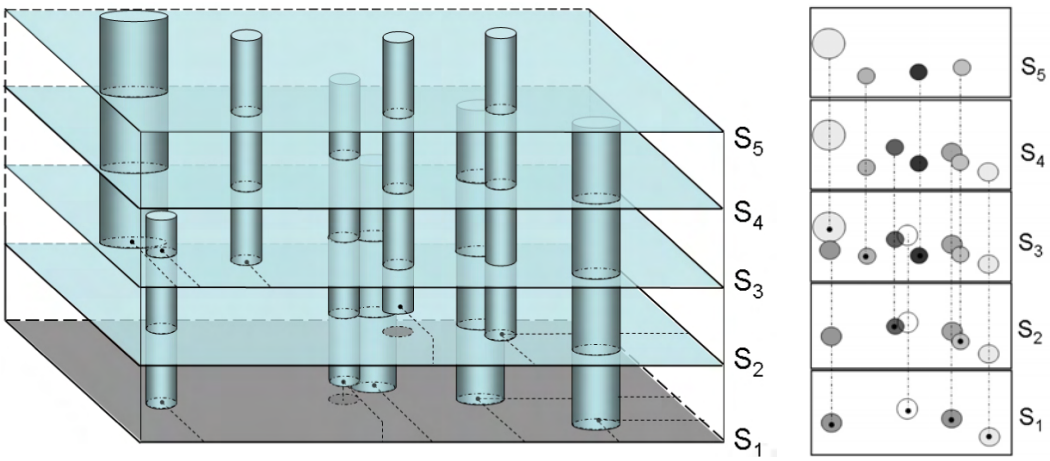
$$\Phi_{s_1} \cup \Phi_{s_2} \cup \Phi_{s_3} = \Phi_{s_1,s_2,s_3} \cup \Phi_{s_1,s_2,s_3^-} \cup \Phi_{s_1,s_2^-} \cup \Phi_{s_1^-,s_2,s_3} \cup \Phi_{s_1^-,s_2,s_3^-} \cup \Phi_{s_2^-,s_3}.$$

All the random sets at the right hand side of the former equation are independent Boolean models. Similarly it can be formulated a similar equation for the general case.

This new methodology to estimate the parameters of a STBM is more robust against noise because the estimates are obtained from group of frames instead of pairs of them. In the previous chapter, the estimation procedure was based on differentiation of couples of frames over time in order to observe changes in intensity. Now we propose to study this intensity change not by using couples of frames but several frames to avoid noise problems.

An example of a spatio-temporal Boolean model is shown in Fig. 5.1(a). The vertical axis corresponds to the time whereas the horizontal plane is the image plane. The union of these cylinders is the realization of the spatio-temporal Boolean model. Fig. 5.1(b) shows some temporal cross sections. Germs are represented as black dots whereas grains are the disks shifted to the germs. 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. Another important remark is that due to the time discretization, a given grain could be born between two consecutive temporal cross sections and it would not be observed (if it was still alive) until the next cross section (frame).

The outline of this chapter is as follows. In Section Section 5.2 considers the statistical analysis of the spatio-temporal Boolean model by aggregation. A simulation study designed to evaluate the estimators proposed is given in Section 5.3. Finally, some conclusions and further developments are summarized in Section 5.4.



(a) (b)

Figure 5.1: **A realization of a spatio-temporal Boolean model with a circular primary grain**. (a) Cylinders correspond to events in the sequence. (b) Five consecutive cross-sections.

5.2 Statistical analysis based on frame aggregation

Here we based our analysis on the fact that the union of independent Boolean models is a new Boolean model. As it was shown in Chapter 4, Section 4.2.2, our spatio-temporal Boolean model can be considered as the union of the different and independent spatio-temporal Boolean models (see 4.6).

Let Φ be a temporal Boolean model and $\Phi_{s_1}, \ldots, \Phi_{s_k}$ temporal cross sections where $s_1 < s_2 < \ldots < s_k$ then $\Phi_{s_1} \cup \ldots \Phi_{s_k}$ is also a Boolean model. An illustration with $k = 2$ follows,

$$\Phi_{s_1} \cup \Phi_{s_2} = \Phi_{s_1,s_2} \cup \Phi_{s_1,s_2^-} \cup \Phi_{s_1^-,s_2}$$

where the three random sets in the right hand side of the equation are independent random sets. Note that s^- indicates that the primary grain considered does not belong to the temporal cross section at s. For $k = 3$, we have

$$\Phi_{s_1} \cup \Phi_{s_2} \cup \Phi_{s_3} = \Phi_{s_1,s_2,s_3} \cup \Phi_{s_1,s_2,s_3^-} \cup \Phi_{s_1,s_2^-} \cup \Phi_{s_1^-,s_2,s_3} \cup \Phi_{s_1^-,s_2,s_3^-} \cup \Phi_{s_2^-,s_3}.$$

All the random sets on the right hand side of the equation are independent Boolean models. Similarly it can be formulated for the general case.

Again, we will consider the estimation of the spatial temporal intensity, the area and perimeter of the primary grain, and the distribution of the random durations.

Our approach is based on the accumulation of different frames. A given temporal cross section of a STBM is a (static) Boolean model. If we accumulate different frames of a STBM, the resulting random set is again a STBM with higher intensity. In this section we will prove it and the corresponding intensity will be determined. Given a sequence of temporal cross sections of a given STBM, we will construct different accumulated versions by choosing different lags and different numbers of frames to be accumulated. For each accumulated video the corresponding intensity will be estimated by using the well-known minimum contrast method (38; 62). These estimated intensities allow us to provide a non-parametric estimator of both the probability density of the random durations of grains and the spatial temporal intensity, λ, of the original spatio-temporal Boolean model. These are the key points of this chapter. Subsection 5.2.1 gives the basic probabilistic results needed later. Basically, we will use different results concerning Poisson point process (11; 14; 66; 67). Subsection 5.2.2 contains a detailed presentation of the estimation procedure.

5.2.1 Previous results

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 the grain will be alive during the time interval $[t_n, t_n + d_n]$. Let $s_0 = -\infty < s_1 < s_2 < \ldots < s_k < s_{k+1} = +\infty$ be a set of fixed (previously specified) points in $\mathbb{R}$, the sampling times. Let $N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$ denote the number of points t_n 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

$$\Psi_{(-\infty,s_1]}^{[s_i,s_{i+1})} = \{t_n \in \Psi : t_n \leq s_1 < s_i \leq t_n + d_n < s_{i+1}\}. \tag{5.1}$$

In particular, we define

$$\Psi_s = \{t_n : t_n \leq s < t_n + d_n\} \tag{5.2}$$

i.e. the events alive at time s. We have the following results:

Proposition 2 *The point process* $\Psi_{(-\infty,s_1]}^{[s_i,s_{i+1})}$ *is an inhomogeneous Poisson point process with intensity function given by*

$$r(t_n) = \theta P(s_i \leq t_n + d_n < s_{i+1}) \; if \; t_n \leq s_1 \; and \; 0 \; if \; t_n > s_1. \tag{5.3}$$

Note that we are applying a thinning to the original Poisson point process. The probability of a given point being removed depends on its temporal location and duration. It is well-known that the resulting point process is a non-homogeneous Poisson point process (11; 66). The total number of points in $\Psi_{(-\infty,s_1]}^{[s_i,s_{i+1})}$ i.e. $N_{(-\infty,s_1]}^{(s_i,s_{i+1})}$ is a Poisson variable with mean given by the following proposition.

Proposition 3 *The mean number of events which were born before* s_1 *and died between* s_i *and* s_{i+1} *is equal to*

$$EN_{(-\infty,s_1]}^{[s_i,s_{i+1})} = \theta\bigg(\int_{s_i-s_1}^{+\infty} P(D \geq t)dt - \int_{s_{i+1}-s_1}^{+\infty} P(D \geq t)dt \bigg). \tag{5.4}$$

Proof 3 (Proposition 3)

$$\begin{aligned} EN_{(-\infty,s_1]}^{[s_i,s_{i+1})} &= \theta \int_{-\infty}^{+\infty} r(t)dt = \\ \theta \int_{-\infty}^{s_1} P(s_i \leq D + t < s_{i+1})dt &= \theta \int_{0}^{+\infty} P(s_i - s_1 \leq D - t < s_{i+1} - s_1)dt = \\ \theta \int_{0}^{+\infty} P(D \geq t + (s_i - s_1))dt &- \theta \int_{0}^{+\infty} P(D \geq t + (s_{i+1} - s_1))dt = \\ \theta \int_{s_i-s_1}^{+\infty} P(D \geq t)dt &- \theta \int_{s_{i+1}-s_1}^{+\infty} P(D \geq t)dt. \end{aligned} \tag{5.5}$$

From now on, we will denote

$$p(s) = \int_{s}^{+\infty} P(D \geq t)dt. \tag{5.6}$$

Note that $p(0) = \int_0^{+\infty} P(D \geq t)dt = ED$ as it is well-known. Then,

$$EN_{(-\infty,s_1]}^{[s_i,s_{i+1})} = \theta p(s_i - s_1) - \theta p(s_{i+1} - s_1). \tag{5.7}$$

In general, we have the following corollary.

Corollary 1 *The mean number of points born before s and dead after s is equal to*

$$EN_{(-\infty,s]}^{(s,+\infty)} = \theta \int_0^{+\infty} P(D \geq t)dt = \theta p(0) = \theta ED. \tag{5.8}$$

Proposition 4 *Let $s_0 = -\infty < s_1 < s_2 < \ldots < s_k < s_{k+1} = +\infty$ be the observation times and $N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$ be the number of points t_n which were born in the time interval $(s_{i-1}, s_i]$ and died in the time interval $(s_{j-1}, s_j]$ where $j > i$. Then*

$$EN_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = \theta p(s_{i-1}, s_i; s_{j-1}, s_j), \tag{5.9}$$

where

$$p(s_{i-1}, s_i; s_{j-1}, s_j) = p(s_{j-1} - s_i) - p(s_{j-1} - s_{i-1}) - p(s_j - s_i) + p(s_j - s_{i-1}). \tag{5.10}$$

Proof 4 (Proposition 4) *We have*

$$N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} | N_{(s_{i-1},s_i]} = n \sim B(n, \tilde{p}), \tag{5.11}$$

where

$$\tilde{p} = P(s_{j-1} \leq T + D < s_j) \tag{5.12}$$

being T a uniform random variable in the interval $[s_{i-1}, s_i)$ and T and D being independent. We have

$$\begin{aligned} P(s_{j-1} \leq T + D < s_j) &= \int_{s_{i-1}}^{s_i} P(s_{j-1} - t \leq D < s_j - t) \frac{1}{s_i - s_{i-1}} dt = \\ &\frac{1}{s_i - s_{i-1}} \left(\int_{s_{i-1}}^{s_i} P(D \geq s_{j-1} - t)dt - \int_{s_{i-1}}^{s_i} P(D \geq s_j - t)dt \right) = \\ &\frac{1}{s_i - s_{i-1}} \left(\int_{s_{j-1}-s_i}^{s_{j-1}-s_{i-1}} P(D \geq v)dv - \int_{s_j - s_i}^{s_j - s_{i-1}} P(D \geq v)dv \right) = \\ &\frac{1}{s_i - s_{i-1}} \left(p(s_{j-1} - s_i) - p(s_{j-1} - s_{i-1}) - p(s_j - s_i) + p(s_j - s_{i-1}) \right) = \theta p(s_{i-1}, s_i; s_{j-1}, s_j). \end{aligned} \tag{5.13}$$

Finally,

$$\begin{aligned} EN_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = E\left[E\left[N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} | N_{(s_{i-1},s_i]} \right] \right] &= E\left[N_{(s_{i-1},s_i]} \right] \frac{p(s_{i-1}, s_i; s_{j-1}, s_j)}{s_i - s_{i-1}} \\ \psi(s_i - s_{i-1}) \frac{p(s_{i-1}, s_i; s_{j-1}, s_j)}{s_i - s_{i-1}} &= \theta p(s_{i-1}, s_i; s_{j-1}, s_j). \end{aligned} \tag{5.14}$$

We are interested in the mean number of points of the point process $\cup_{i=1}^{k}\Psi_{s_i}$ i.e. the union of all points alive in some s_i's. Note that the (random) number of points in $\cup_{i=1}^{k}\Psi_{s_i}$ is equal to $\sum_{i=1}^{k}\sum_{j=i+1}^{k+1} N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}$. We will denote

$$\psi(s_1, s_2, \ldots, s_k) = E\left[\sum_{i=1}^{k}\sum_{j=i+1}^{k+1} N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}\right] = \sum_{i=1}^{k}\sum_{j=i+1}^{k+1} E\left[N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]}\right]. \quad (5.15)$$

We have the following result.

Proposition 5

$$\psi(s_1, s_2, \ldots, s_k) = \theta\left[kp(0) - \sum_{i=2}^{k} p(s_i - s_{i-1})\right]. \quad (5.16)$$

Proof 5 (Proposition 5) *Let us see a graphical proof of the result. See Fig. 5.2. At s_1 we have a mean number of points given by $\theta p(0)$. In the interval $[s_1, s_2)$ the mean number of deaths is given by $p(s_1) - p(s_2)$. The signs + and − in the bottom-left cell refer to $+p(s_2)$ and $p(s_1)$. A similar comment applies to the left cells in Fig. 5.2. The interpretation of the rest of the signs follows. In the node corresponding to s_i and s_j each sign corresponds to the term $p(s_i) - p(s_j)$. The different terms cancel each other out except for a few terms.*

$$\psi(s_1, \ldots, s_k) = E\sum_{i=1}^{k}\sum_{j=i+1}^{k+1} N_{(s_{i-1},s_i]}^{(s_{j-1},s_j]} = \theta\left[kp(0) - \sum_{i=2}^{k} p(s_i - s_{i-1})\right]. \quad (5.17)$$

In particular, if the times s_i are equally spaced, as usual, with $s_i - s_{i-1} = \delta$, then Eq. (5.16) becomes

$$\psi(s_1, s_1 + \delta, \ldots, s_1 + (k-1)\delta) = \theta\left[kp(0) - (k-1)p(\delta)\right]. \quad (5.18)$$

5.2.2 Estimation

Our data consist of the sequence of binary images corresponding to $\{\Phi_{s_i} \cap W\}_{i=1}^{n}$ where Φ is the underlying spatio-temporal Boolean model, W is the sampling

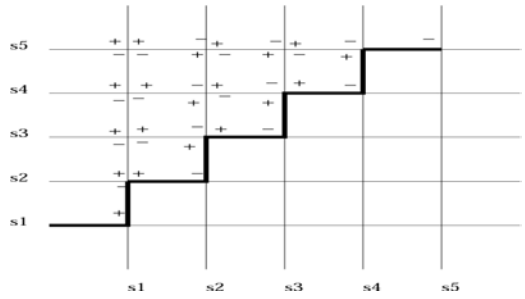


Figure 5.2: Signs of the $p(s_i - s_j)$.

window (the region of $\mathbb{R}^2$ where we can observe the random set) and $s_1 < \ldots < s_n$ are the sampling (or observation) times. Note that Φ denotes the stochastic model. From now on, we will denote the realization of Φ as ϕ. The sequence of binary images really observed will correspond to $\{\phi_{s_i} \cap W\}_{i=1}^n$.

If Ψ denotes now the homogeneous Poisson point process associated with the spatio-temporal Boolean model Φ, then the point process composed of the times t_i's such that the corresponding location belongs to W, $\Psi_2 = \{t_i : (x_i, t_i) \in \Psi, x_i \in W\}$, is a homogeneous Poisson point process in the real line with intensity $\theta = \lambda\nu_2(W)$.

Given the original sequence $\{\Phi_{s_i} \cap W\}_{i=1,\ldots,n}$ with $s_i - s_{i-1} = \delta$ we will consider the accumulated sequence defined as,

$$\tilde{\Phi}_{s_i} = \cup_{j=i}^{i+k} \Phi_{s_j} \tag{5.19}$$

with $i = 1, \ldots, n-k+1$ and k being a fixed integer. If Φ is a spatio-temporal Boolean model, then the different $\tilde{\Phi}_{s_i}$ are (static) Boolean models. The proof is obvious by using the arguments given in (6).

If we assume that $s_i - s_{i-1} = \delta$, we will denote the intensity of $\tilde{\Phi}_{s_i}$ as $\lambda_s(k, \delta)$. Note that this intensity only depends on the lag δ and k (the number of frames that we are accumulating). In particular, if $\delta = 0$ and $k = 1$, we have $\lambda_s(1, 0) = \lambda\nu_2(W) = \theta$. For each frame in the aggregated sequence, we can estimate the intensity by using the minimum contrast method. Afterwards, we apply the batch-mean method (33) over the estimates obtained from each frame in the aggregated sequence in order to calculate the spatial intensity of the whole image sequence, $\hat{\lambda_s}(k, \delta)$. We repeat this procedure for different values of k and δ.

Eq. (5.18) becomes

$$\psi(s_1, s_1+\delta, \ldots, s_1+(k-1)\delta) = \lambda\nu_2(W)\Big[kp(0) - (k-1)p(\delta)\Big]. \tag{5.20}$$

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

$$\lambda_s(k,\delta) = \frac{\psi(s_1, s_1+\delta, \ldots, s_1+(k-1)\delta)}{\nu_2(W)} = \lambda\Big[kp(0) - (k-1)p(\delta)\Big]. \tag{5.21}$$

Let us rewrite Eq. (5.21) as

$$\lambda_s(k,\delta) = (\lambda p(0) - \lambda p(\delta))k + \lambda p(\delta), \tag{5.22}$$

If we consider the right hand side of this equation as a function of the number of frames k, then the coefficient associated with k and the constant have simple interpretations. Let

$$\alpha(\delta) = \lambda p(0) - \lambda p(\delta), \tag{5.23}$$

and

$$\beta(\delta) = \lambda p(\delta). \tag{5.24}$$

It was proved in Eq. 4.10 that $\alpha(\delta)$ is the mean number of new grains 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 are alive between two frames separated by δ frames. Then Eq. (5.22) can be reformulated as reformulated as

$$\lambda_s(k,\delta) = \alpha(\delta)k + \beta(\delta). \tag{5.25}$$

The left hand side of the previous equation is estimated for different δ and k values. Now, we fixed the δ value and consider the estimates $\hat{\lambda_s}(k,\delta)$ for this fixed δ. A weighted least square fit is then performed where the predictor is k and the dependent variable is $\hat{\lambda_s}(k,\delta)$. We repeat that fit for different $\delta \in \{\delta_1, \ldots, \delta_n\}$. This gives rise to the estimated functions $\hat{\alpha}(\delta_i)$ and $\hat{\beta}(\delta_i)$ with $i = 1, \ldots, n$.

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 $\hat{\lambda}$ and the density function of the random duration, f_D. In particular, as we showed in Chapter 4 we have

$$\alpha'(0) = -\lambda p'(0) = \lambda P(D \geq 0) = \lambda, \tag{5.26}$$

where $\alpha'(0)$ denotes the derivative at the origin of the function $\alpha(\delta)$. Furthermore, the function $p(\delta)$ can be obtained as

$$p(\delta) = \frac{\lambda_s(1,0) - \alpha(\delta)}{\lambda}, \tag{5.27}$$

because $\lambda_s(1,0) = \lambda ED$. Moreover, it holds that

$$p'(\delta) = -P(D \geq \delta) = -(1 - P(D \leq \delta)) = F_D(\delta) - 1,$$

and

$$p''(\delta) = f_D(\delta), \tag{5.28}$$

where $p''(\delta)$ is the second derivative of $p(\delta)$. From the previous equations it is clear that a good estimation of the function α will give us good estimators of λ and f_D.

From here on, we adopt an approach based on functional data analysis. We fitted a functional object to the data $\{(\delta_j, \hat{\alpha}(\delta_j))\}_{j=1}^{n}$ and, by using Eq. (5.26), we estimated the spatial temporal intensity λ. Let $\hat{\lambda}$ be the estimated value. Then, by replacing the parameters λ and $\lambda_s(1,0)$ in Eq. (5.27) with the corresponding estimates and the function α with the previously fitted functional object, we had the corresponding functional object for the function $p(\delta)$. By taking into account Eqs. (5.27) and Eq. (5.28) we obtained the non-parametric estimator of f_D. Finally, the mean duration ED was estimated as the value of p at zero.

5.3 Results

In this section we describe the results obtained after performing a study on simulated spatio-temporal Boolean models. These experiments allow us to find out the precision of the proposed estimators and to obtain the error levels expected before running a real application.

In order to obtain more realistic results we performed the study over different spatio-temporal Boolean models. This brought us the possibility of studying several scenarios which could be comparable to some real applications. The parameters of interest to be estimated from the binary image sequences were: the spatial temporal intensity of the germ process, λ, the mean number of germs per unit area in a given frame, λ_s; the mean duration of the events, ED; and finally a functional parameter f_D which is the density of durations. The mean area and perimeter of the primary grain, a_0 and u_0, were obtained as well. Though they do not refer to purely temporal information, we present a summary description of their relative errors.

5.3.1 Simulation Study

Note that the proposed procedure is based on the accumulation of different dependent Boolean models and the estimation of their corresponding parameters by using the minimum contrast method. The behavior of this method depends heavily on the area fraction of the Boolean model i.e. the mean area covered by the Boolean model per unit area. Therefore, we had to take into account the final area fraction after the accumulation of the frames.

Let p_v be the volume fraction i.e. the mean volume covered by the spatio-temporal Boolean model per unit volume (unit area per unit time). By taking into account the properties of the Boolean model, we have $p_v = 1 - \exp\{-\lambda a_0 ED\}$. We set λ to 0.000035 and the volume fraction p_v to 0.11. That allowed us to study the estimator performances as a function of the probability distribution chosen for the duration and the mean of the corresponding distribution. Three different probability distributions were used for the random durations (uniform, exponential and Gamma) and two different values for their means (6 and 8 seconds). The primary grain was circular with uniform distribution for the radius. In particular, a uniform radius in $[11, 15]$ was used when the mean duration chosen was equal to six seconds and a uniform radius in $[9, 13]$ when the mean duration was set to 8 seconds. Note that the mean area of the primary grain a_0 is equal to $a_0 = \pi E R^2$. Six videos of 600 frames each were generated for the different experimental setups previously described. Table 5.1 summarizes the parameter combination used to generate all the videos. The column headed *Distribution* refers to the duration distribution and the column headed R_0 contains the distribution of the random radius of the primary grain. Note that a uniform distribution is specified giving the maximum and minimum of the support. If D follows an exponential distribution, $D \sim Expo(a)$, the parameter a is the mean. When a Gamma distribution is assumed, $D \sim Gamma(a, b)$ then $ED = a/b$. We simulated image sequences of 512×512 pixels in size (each image) and 150 seconds long. Videos were sampled at 4 frames per second, so we had 600 frames for each sequence. Fig. 5.3 displays four consecutive subimages of the two simulated spatio-temporal Boolean models where the random duration follows a uniform distribution.

In order to illustrate the aggregation effect, Fig. 5.4 shows both original frames (see Video 5.1 in supplementary material) and accumulated ones for different values of k and δ. These frames correspond to simulation #1 described in Table 5.1. The first row corresponds to three frames of the original sequence; the remaining rows are frames generated by unions. Specifically, the second row corresponds to $k = 2$ and $\delta = 3$ (see Video 5.2 in supplementary material). Note that the number of grains here is higher compared with the original frames. The

Table 5.1: STBMs used in the study.

	Distribution	R_0
#1	Unif(4,8)	Unif(11,15)
#2	Expo(6)	Unif(11,15)
#3	Gamma(6,1)	Unif(11,15)
#4	Unif(6,10)	Unif(9,13)
#5	Expo(8)	Unif(9,13)
#6	Gamma(8,1)	Unif(9,13)

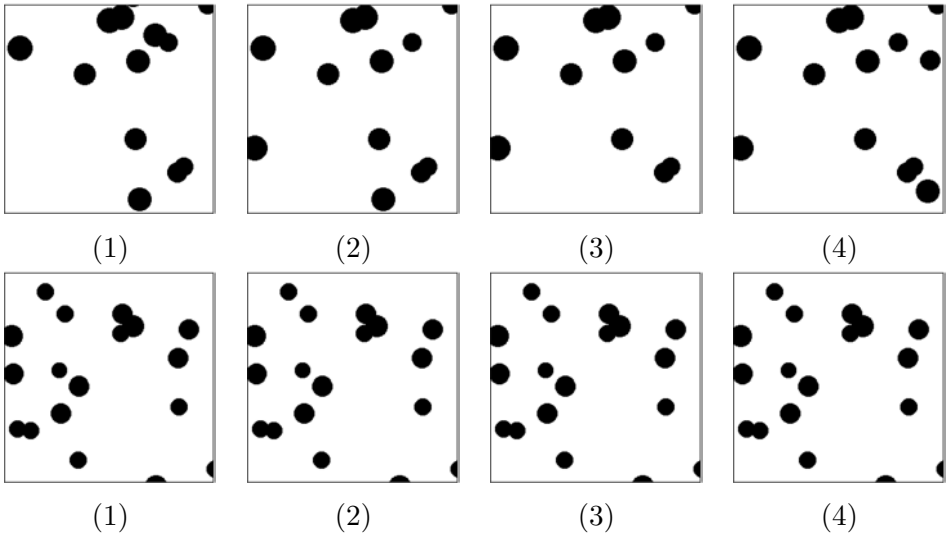


Figure 5.3: **Four consecutive frames of two image sequences with different parameters**. The first row corresponds to simulation #1 with $ED = 6$ seconds. The second row corresponds to simulation #4 in which $ED = 8$ seconds. Here, subimages of 256×256 pixels in size are shown in order to highlight the differences between grain radii.

addition effect depends not only on the number of frames accumulated but also on the δ chosen. This fact can be observed when comparing rows two ($k = 2$, $\delta = 3$) (see Video 5.2 in supplementary material) and three ($k = 2$, $\delta = 6$) (see Video 5.3 in supplementary material). This behavior is due to the fact that for higher values of δ the dependency between the generated Boolean models is

less and more new grains appear. In summary, a greater area fraction can be expected. For really large δ values, the generated Boolean models are then completely independent and the area fractions is a constant value. Finally, the fourth row corresponds to a higher area fraction, with $k = 5$ and $\delta = 9$ (see Video 5.4 in supplementary material). Here, the increment in intensity and overlapping is clearly observable.

Let us see in-depth the results obtained for one of the simulations, simulation #1. We saw the estimators used for the functions α and β in previous sections. In Fig. 5.6 (a) these estimates are plotted. Note that function $\alpha(\delta)$ should be equal to λ_s for large δ values because $p(\delta)$ goes to zero when δ increases. The reverse comment applies to β. Moreover, the sum of $\alpha(\delta) + \beta(\delta)$ should be equal to λ_s. The estimated sum is plotted along with its theoretical value. A slight overestimation of λ_s is appreciable with respect to the theoretical value. This is due to an overestimation of β that did not reach the expected null value. We guess this error is mainly due to the spatial and temporal discretization of the image sequence. On the one hand, since images were 512×512 pixels in size, with grain radii ranging from 9 to 15 pixels, the grains were not really disks. On the other hand, the sampling rate was 4 frames per second, and the mean durations simulated were as short as 6 and 8 seconds. In this example, with $D \sim Unif(4, 8)$ the function $p(\delta) = 0$ for $\delta \geq 8$ and $\alpha(\delta)$ should be equal to λ_s for δ values greater than 8 seconds. Fig. 5.6 (b) shows $p(\delta)$ for a time interval of $[0, 11]$ seconds.

We used an approach based on functional data analysis in order to obtain a precise estimation of the second derivative, the duration density f_D. Estimations of $p(\delta)$ at different δ's were used to construct a continuous function which represented the data more smoothly and allowed us to perform a continuous derivation over time. Because the basis functions are not themselves smooth everywhere, they will not give rise to smooth basis expansions either. From among several possibilities we chose a polynomial spline basis $\{\Phi_k(t)\}_{k=1,\ldots,K}$, where each $\Phi_k(t)$ is a piecewise cubic function, since splines are a good basis for the representation of smooth functions.

A total of 20 different cubic b-spline functions were used to fit the functions on the 44 discrete estimations for the uniform distribution case, and 100 for the exponential and Gamma distributions. It was necessary to impose some constraints on the roughness. In particular, a roughness penalty factor of 0.10 was used for the uniform simulations, 2 for the exponential ones and 0.50 for the Gamma ones. The criterion used to fit a smooth curve then had two terms. The first assessed the fidelity of the curve to the observed data. The second term, the penalty term, measured the extent to which the fitting function was smooth. The

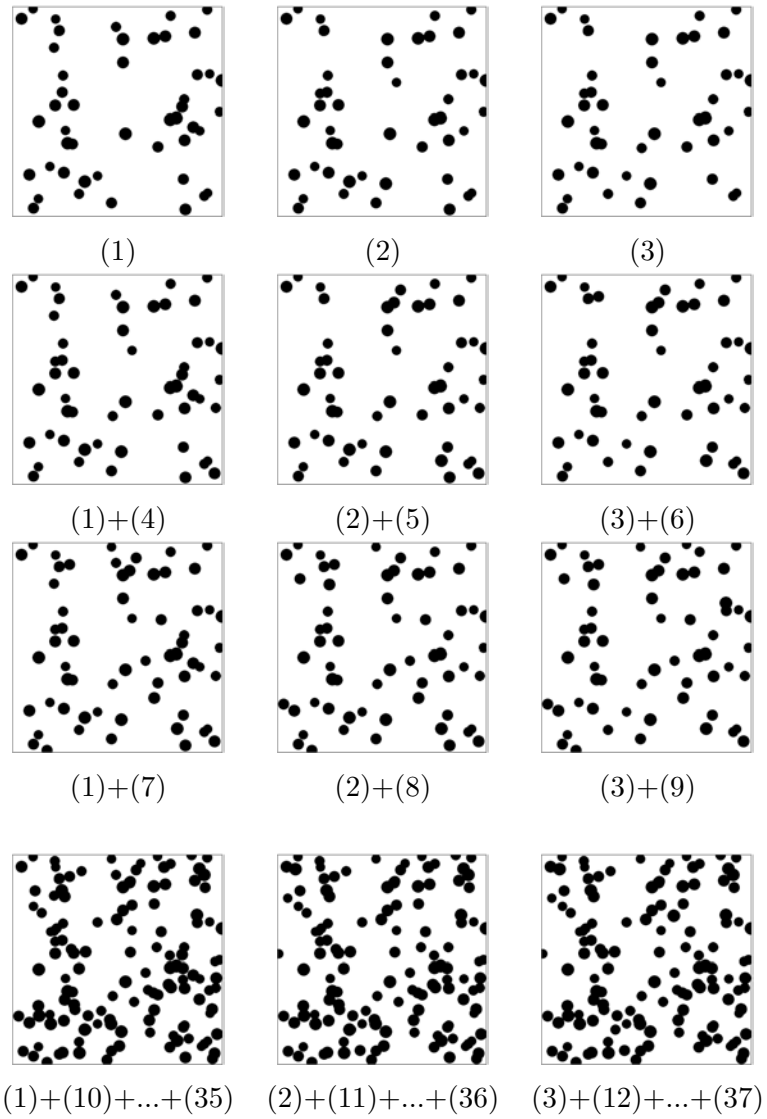


Figure 5.4: **Aggregation effect.** The first row shows three frames of original sequence. The second row shows a sequence made up by aggregating frames with $k = 2$ and $\delta = 3$. The third row corresponds to $k = 2$ and $\delta = 6$. The last row shows a higher aggregation effect with $k = 5$ and $\delta = 9$. For this case the area fraction reached a value of 26.2%.

smoothing was performed over the second derivative because it was the function we were searching for. We realized that we could not have a perfect data fit and a smoothness function at the same time, so the smoothing parameter had to be selected on a per simulation basis. After smoothing and fitting, we applied Eq. (5.28) and obtained the second derivative of $p(\delta)$, that is, f_D. The estimated duration density is shown in Fig. 5.6 (c) along with the corresponding theoretical function.

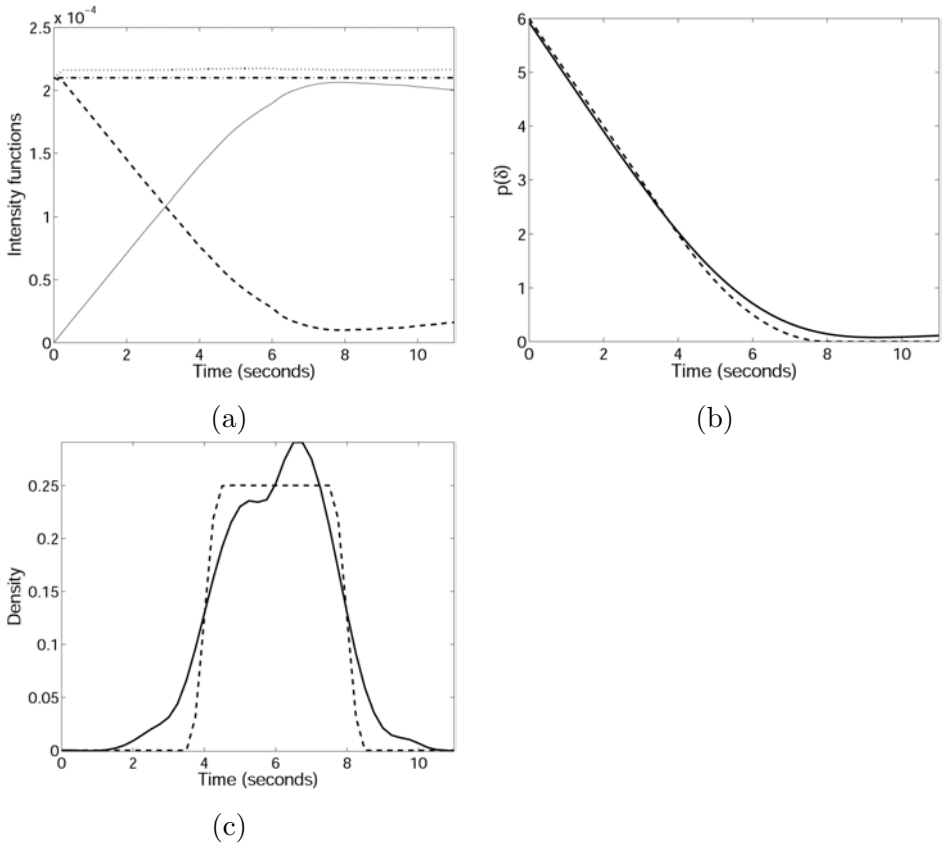


Figure 5.5: **Estimated temporal functions for uniform distribution**. (a) $\alpha(\delta)$ (solid line), $\beta(\delta)$ (dashed line) and the theoretical value of λ_s (dashed-dotted line) along with the estimated one (dotted line). (b) The estimated $p(\delta)$ (solid line) and theoretical one (dashed line). (c) Estimated density function of durations (solid line) and theoretical one (dashed line).

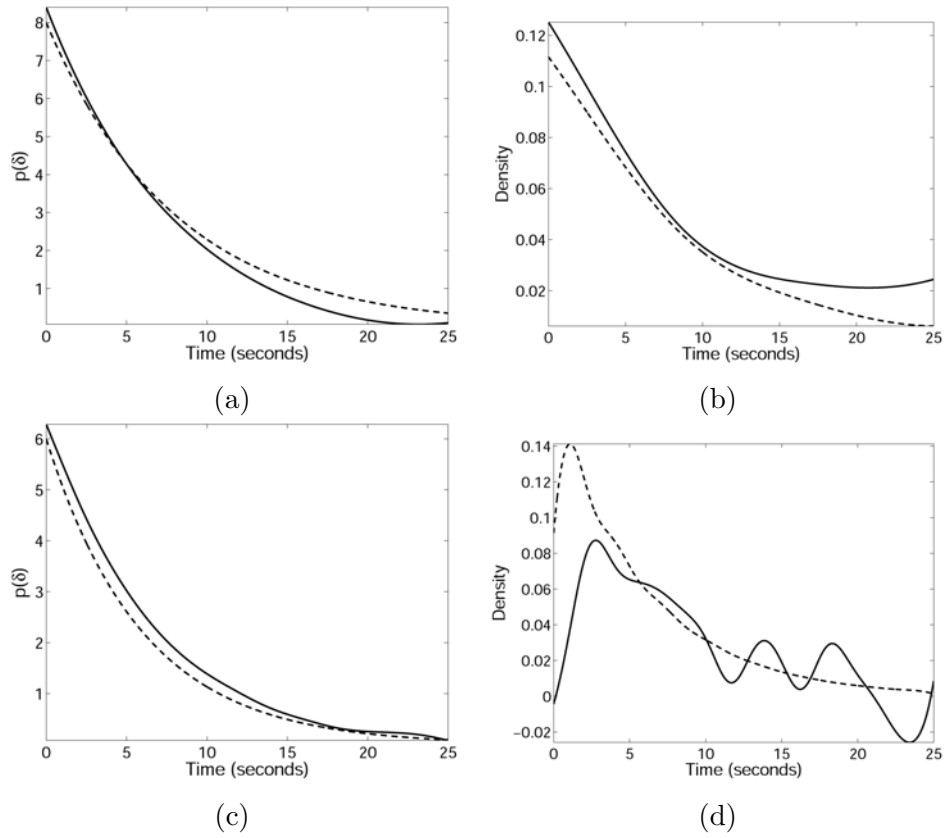


Figure 5.6: **Estimated temporal functions for exponential and gamma distributions**. (a) and (c) show the estimated $p(\delta)$ (solid line) and theoretical one (dashed line) for simulations with Exponential and Gamma distribution, respectively. (b) and (d) show the estimated density function of durations (solid line) and theoretical one (dashed line) for simulations with Exponential and Gamma distribution, respectively.

5.3.1.1 Relative errors

In order to evaluate the errors in the estimators proposed, we analyzed the results of the simulation study by using again the relative error. We would like to remark that we found out that when a heavy aggregation of frames was performed (k larger than 4), $\lambda_s(k, \delta)$ was underestimated. This effect may be due to the minimum contrast method. This method behaves poorly for high area fractions.

Table 5.2: Relative errors for area of the primary grain, a_0

k	ED	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2	6	-0.0612	-0.0594	-0.0554	-0.0488	-0.0383	-0.0243
3	6	-0.0607	-0.0577	-0.0420	-0.0397	-0.0273	-0.0088
4	6	-0.0604	-0.0489	-0.0372	-0.0335	-0.0154	0.0036
5	6	-0.0604	-0.0470	-0.0241	-0.0231	-0.0099	0.0329
6	6	-0.0612	-0.0441	-0.0120	-0.0153	0.0115	0.0427
7	6	-0.0604	-0.0416	-0.0099	-0.0126	0.0149	0.0262
8	6	-0.0604	-0.0387	-0.0085	-0.0106	0.0112	0.0372
2	8	-0.0569	-0.0530	-0.0510	-0.0471	-0.0422	-0.0280
3	8	-0.0668	-0.0602	-0.0564	-0.0532	-0.0495	-0.0289
4	8	-0.0685	-0.0638	-0.0582	-0.0549	-0.0482	-0.0298
5	8	-0.0704	-0.0644	-0.0574	-0.0520	-0.0388	-0.0207
6	8	-0.0703	-0.0641	-0.0563	-0.0461	-0.0230	-0.0090
7	8	-0.0704	-0.0640	-0.0482	-0.0387	-0.0084	0.0135
8	8	-0.0703	-0.0616	-0.0387	-0.0287	0.0120	0.0372

This underestimation causes an underestimation of $\alpha(\delta)$. Therefore, for larger values of δ, the numerator of Eq. (5.27) was never null since $\alpha(\delta)$ never reaches λ_s and so the $p(\delta)$-function was overestimated.

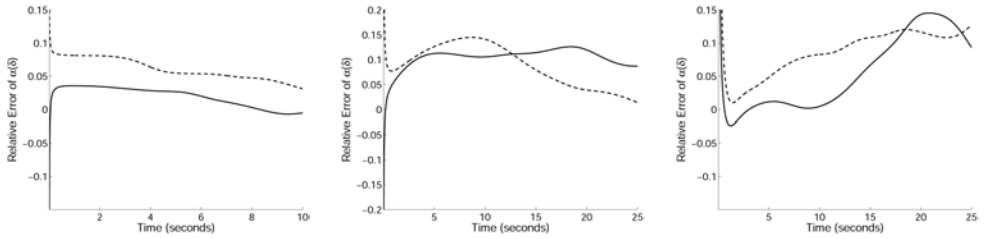
Firstly, the characteristics of the primary grain are supposed to be constant even though an aggregation of frames is performed. In order to see if this underestimation effect was also present in other parameters of the Boolean model such as the area and perimeter of the primary grain, we studied their relative errors. Tables 5.2 and 5.3 show a summary description of the relative errors for the uniform distribution. The different columns correspond to the minimum, first quartile, median, mean, third quartile and maximum. It seems that higher k-values have associated smaller relative errors in the estimation of the mean area a_0. However, there is no clear trend in the estimation of the mean perimeter u_0. Generally, we think that the errors are very small.

Secondly, since the probability density of the duration is estimated from $\alpha(\delta)$, it seems interesting to analyse its relative error by taking into account the distribution used. Fig. 5.7 shows the relative errors of α as a function of δ. For the three cases these errors are quite small. The worst case corresponds to the

Table 5.3: Relative errors for perimeter of primary grain, u_0

k	ED	Min.	$1st$ Qu.	Median	Mean	$3rd$ Qu.	Max.
2	6	-0.0250	-0.0224	-0.0183	-0.0124	-0.0037	0.0109
3	6	-0.0250	-0.0207	-0.0066	-0.0036	0.0084	0.0285
4	6	-0.0250	-0.0134	0.0003	0.0019	0.0194	0.0362
5	6	-0.0235	-0.0116	0.0096	0.0104	0.0258	0.0627
6	6	-0.0244	-0.0094	0.0193	0.0157	0.0384	0.0692
7	6	-0.0250	-0.0077	0.0212	0.0162	0.0378	0.0487
8	6	-0.0233	-0.0085	0.0215	0.0169	0.0354	0.0559
2	8	-0.0154	-0.0116	-0.0094	-0.0058	-0.0026	0.0127
3	8	-0.0253	-0.0187	-0.0155	-0.0125	-0.0087	0.0120
4	8	-0.0278	-0.0229	-0.0172	-0.0154	-0.0105	0.0113
5	8	-0.0313	-0.0261	-0.0177	-0.0141	-0.0047	0.0127
6	8	-0.0313	-0.0257	-0.0162	-0.0093	0.0105	0.0229
7	8	-0.0310	-0.0244	-0.0104	-0.0027	0.0223	0.0434
8	8	-0.0310	-0.0225	-0.0006	0.0070	0.0407	0.0709

exponential distribution, which we think could be due to the higher variance of durations.



(a) (b) (c)

Figure 5.7: **Relative errors for** $\alpha(\delta)$. (a) using a uniform distribution, (b) an exponential distribution, and (c) a Gamma distribution. Dashed lines corresponds to $ED = 6$ and solid lines to $ED = 8$.

Finally, Table 5.4 shows a summary of the observed relative errors for the spatial temporal intensity λ and the mean duration ED. The rows in that table

Table 5.4: Relative errors for function α, spatial temporal intensity, λ and ED

Parameter	ED	Uniform	Exponential	Gamma
α	6	0.0021	0.0009	0.0011
	8	0.0019	0.0012	0.0008
λ	6	0.0170	0.0046	0.0491
	8	0.0029	0.0040	0.0326
ED	6	0.0027	0.0687	0.0531
	8	0.0071	0.0346	0.0625

corresponding to α display the integral of the relative error along the time interval used.

5.3.2 Biological application

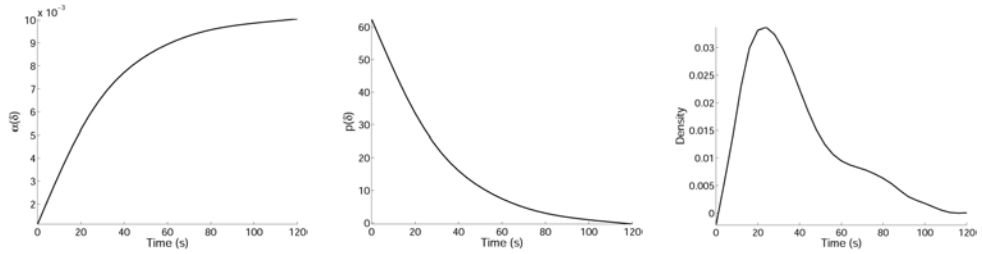
We applied our estimators to the six biological image sequences described in the previous chapter, Table 4.6. These sequences were derived from three movies of fibroblast (COS7) cells, acquired at 1 frame every 4 seconds. The cells expressed clathrin coupled to the Green Fluorescent Protein (GFP). The estimates obtained are summarized in Table 5.5. (see Video 5.5 in supplementary material)

Table 5.5: Estimates for the three cells analyzed

Cell	Seq.	$\hat{\lambda}_s$ (μm^{-2})	$\hat{\lambda}_s \lvert W \rvert$	$\hat{\lambda}$ ($\mu m^{-2}\ sec^{-1}$)	ED (sec)
1	1	0.3625 ± 0.0225	281.86 ± 17.49	0.0052	70.27
	2	0.3085 ± 0.0210	347.20 ± 23.69	0.0361	85.42
2	3	0.3342 ± 0.0087	277.54 ± 7.23	0.0058	57.48
	4	0.3502 ± 0.0120	235.41 ± 8.05	0.0054	65.37
3	5	0.2939 ± 0.0359	170.17 ± 20.88	0.0039	74.95
	6	0.2724 ± 0.0044	220.28 ± 3.54	0.0032	83.70

The column headed $\hat{\lambda}_s$ in Table 5.5 gives the point estimates with a 95% confidence interval for the intensity and the mean intensity in any cross-section, for the different image sequences. In our study, the application of the batch-

mean method implies a minimum number of frames. Sequences of 300 lead to 15 batches of 20 images, which allowed us to use the method. The column headed $\hat{\lambda}_s|W|$ gives the estimate of the mean number of endocytic spots per frame for each sequence. The columns headed $\hat{\lambda}$ and ED give the estimates obtained on the 3 different cells that were analyzed. The column headed $\hat{\lambda}$ provides the mean number of events per unit micrometer and second for each cell. Similar values are observed for all the cells. We compared our results with those obtained manually for a small subset of clathrin spots belonging to the same group of cells. Manual data analysis showed that average endocytosis durations where between $75 - 90$ seconds and matches the results obtained from our analysis (see Table 5.5). Results for one of the sequences analyzed are shown in detail in Fig. 5.8 (a), (b) and (c).



(a) (b) (c)

Figure 5.8: Estimated functions for the sequence 1 from cell 2. (a) $\alpha(\delta)$. (b) The estimated $p(\delta)$. (c) Estimated density function of durations.

5.4 Conclusions

As it was shown in the simulation study, the relative errors of the estimated parameters are reasonably low. The behavior of this method depends heavily on the area fraction of the Boolean model. We found that when a heavy aggregation of frames was performed (k larger than 4), $\hat{\lambda}_s(k, \delta)$ was underestimated, which may be due to the minimum contrast method. This method behaves poorly for high area fractions, the underestimation causing in turn an underestimation of $\alpha(\delta)$. Nevertheless, as was shown in the simulation study, the relative errors of the estimated parameters ranged from 0.0008 to 0.0625, that is, errors of around 5%, which could be considered small enough in many real applications. Lower values could be expected if higher acquisition rates and higher spatial resolutions

were used. Another important comment is that due to the time discretization, a given grain could be born between two consecutive temporal cross-sections and it would not be observed until the next cross-section (frame).

We think this new method can aid the study of complicated processes such as those observed in molecular biology. Specifically, the spatio-temporal Boolean model based on aggregation might be successful applied in the study of protein expression on videos obtained using Total Internal Reflection Fluorescent Microscopy, as was previously shown in previous chapter.

Due to the data source, we would like to remark the need for the functional data analysis. A simple approach for obtaining functional data would be to interpolate the raw values obtained for $p(\delta)$, but because of the skewness introduced this would give worse estimators in the derived functions $p'(\delta)$ and $p''(\delta)$. It is highly advisable to turn the discrete estimated samples into a functional data object in order to obtain smooth derivatives. Here, we show how a function was fitted over the data so that the derivation could be performed on the continuous real line, and results were more accurate. Moreover, the smooth step allowed us to get clean curves and avoid the noise introduced by slight changes in the slope of $p(\delta)$ function.

We should point out that the methodology of aggregation described has certain limitations. The limitations are due to the poor behavior shown by the minimum contrast method when estimating the spatial intensity with large area fractions (around 50%) and small grain sizes, since the overlapping among events is then very large.

It is also important to bring up a matter related to the lag range. The longer the lag is (with respect to the durations), the higher the relative errors are. On the one hand, the longer the lag is the larger the number of events being born and dying in such interval is and we could then expect a higher variability in the intensity estimated. On the other hand, when aggregating several cross sections, the longer the lag is the higher the area fraction is, since the probability of the events in one frame surviving until the next added frame would be really small. In this way, each aggregated frame will increase the intensity in the accumulated image and so the area fraction.

Finally, the analysis of the spatial temporal coverage in our model remains to be done. For instance, this analysis would be interesting for possible applications in the field of sensor networks in which there is overlapping among sensors and the sensor life time needs to be considered.

Chapter 6

Conclusions and Future Work

In this work we have proposed different statistical methodologies to deal with stochastic spatial temporal processes. Either methodologies have been applied to study well-known important Biological processes working in cells, the exocytosis and the endocytosis.

Both processes are different in nature and generate different kind of information, so we have approached them using different statistical modelling in the framework of stochastic processes. From a biological point of view, it is known that this processes should be coupled to a certain extent, though direct visualization and subsequent modelling of the phenomenon is still missing.

Firstly, a methodology to study exocytic spatio-temporal patterns, in which hypothesis of stationarity and isotropy are assumed, is suggested and used for the study of a real process in Chapter 2. We showed a methodology to analyze whether a given pattern of fusion events follows a clustered distribution or on the contrary presents a random arrangement of its events. In particular, we make use of several distance functions to study the spatial, temporal and spatio-temporal dependence between the points of a pattern, $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{K}$-functions. The procedure to test whether a given realization shows a clustered pattern involves the use of Monte Carlo simulations. It consists in the application of one of the mentioned functional descriptors on a number of random simulations, where events have been generated by a Poison process, and its comparison with the observed pattern. A p-value to measure the significance of the outcome is given.

We have applied these functions to test whether the patterns of exocytosis observed are randomly placed over the plasma membrane or on the contrary

there are 'hotspots' where exocytosis are more likely to happen. The analysis has also been performed to study the existence of temporal clustering, i.e, exocytosis happen in bursts or in a steady stream, and spatio-temporal clustering, i.e the exocytic events are produced in specific regions of the cell at given periods of time. Here, the methodology has been shown with illustrative purposes, although it is flexible enough to be applied to a wide range of real processes in which spatial, temporal or spatio-temporal behaviors need to be formally tested.

For the application showed in this thesis, we have developed an image processing algorithm in order to extract the locations and times of occurrence of exocytosis events on the plasma membrane. The algorithm allowed us to obtain semi-automatically and in a fast way the data needed to analyzed the behavior of the process.

Secondly, in Chapters 4 and 5 we have defined and studied a particular case of Boolean model, called the *spatio-temporal Boolean model*, which formalizes the configuration of independent, randomly placed events with independent durations. Moreover, spatially and temporally overlapping among different events is considered and assumed. This approach enables us to obtain information of stochastic processes producing random shapes with random durations following a given distribution. Locations of these events are assumed to be the outcome of a realization of a spatial temporal Poisson Point Process. In this model no particular parametric model for the shape or duration has to be assumed.

Two different approaches are described to study spatio-temporal Boolean models. The first one, presented in Chapter 4, based its estimates on the difference between pairs of temporal cross-sections of the model, which are temporally away. In this way, we can study how this difference in intensity changes depending on the temporal lag. The second approach, presented in Chapter 5 is based on the aggregation of sets of cross-sections of the model separated by a given lag into one cross-sections in order to study the increment of intensity on each frame. Different aggregation and lag combinations give rise to different intensities which permit us to study the model. Specific estimators has been defined to infer the parameters of the original spatio-temporal Boolean model from its respective aggregated sequences. The aggregation method requires to build up as many image sequence as parameter combinations (aggegation and lag) used which, in turn, have to be analyzed to obtain the estimates. Therefore, the aggregation method requires more time to obtain the estimates of the model. We do not consider this fact as an important drawback, since the same process performed manually is highly time-consuming when compared with either method. Nonetheless, we observed a limitation of the aggregation method when the aggregated image se-

quences displayed high volume fractions. Under this circumstances the estimation method, the minimum contrast method, underestimate the real intensity of the aggregated process, so the final estimations are also biased. This effect depends heavily on the original model to be aggregated. In our sequences this effect could be observed after aggregating 5 frames or more.

Estimators for the mean number of events per unit area and time, the mean size of the events and their density function of duration from a realization of a spatio-temporal Boolean model are defined. A wide simulation study has been done in order to test the proposed estimators. In particular, 135 image sequences of spatio-temporal Boolean models were created using 9 different combinations of parameters replicated 15 times each. The relative errors obtained after their analysis range $0.3\% - 15\%$, though most of the errors were smaller than 5%. In this study, techniques based on functional data analysis were applied since some estimators were obtained by derivatives of other functions making them really sensible to the noise introduced during the image analysis step. No significative differences were showed between both estimation methods. We expected a more robust behavior in the aggregation method, since small noise introduced in the images will not affect much this method, whereas it does in the method based on differences.

To illustrate the application of the spatio-temporal Boolean models, we study the spatial temporal properties of a Biological process, the endocytosis. Accurate information about positions and times of occurrence of endocytosis could not be obtained by means of image processing techniques, specifically due to the spatial an temporal overlap of the events, therefore Point Process theory could not be applied. The approach based on spatio-temporal Boolean models was applied successfully to study the kinetics of endocytosis in living cells. This methodology enabled us to estimate in a robust way parameters including: i) the mean number of events per unit area and time and ii) the density function of the durations of endocytic events. Classical techniques based on the labelling of connected components lead to an underestimation of both the number of grains and the mean duration. Spatio-temporal Boolean Models provided a powerful analysis tool. The distribution of duration was shown to be in excellent agreement with the same parameter, measured manually from the same data set. To our knowledge, this is the first automated measurement of the number of endocytic events per unit area and time. This type of measurement is usually difficult to obtain manually from large data sets. In the future this methodology could even be applied to imaged based screening assays - such as drugs that inhibit endocytosis. The proposed method provides the experimenter with a robust and formal methodology for the quantitative study of cell endocytic behavior

in response to different levels of stimulation, treatments, etc. Moreover, this type of analysis may provide mechanistical information on the endocytic process. For instance, the role of GTPase dynamin in the endocytosis of clathrin-coated vesicles may be better elucidated.

Even though exocytosis and endocytosis are thought to be coupled, a formal analysis remains to be done. Results here showed that constitutive exocytosis happens clustered over the plasma membrane whereas endocytic spots are considered to be scattered uniformly over the plasma membrane. These assumption are still safe, since it is not known if an endocytic spot can produce several endocytosis in the same area, thus showing a clearly clustered behavior.

There are further interesting extensions to both methodologies that could bring more information about the analyzed processes. In general, these improvements require more sophisticated models that sometimes could be difficult to apply, specially because more specific data is needed from the process and more complicated hypothesis have to be proposed.

Further extensions to the methodology of spatio-temporal point processes described in Chapter 2 are detailed below:

- One of the plausible extensions to the spatial temporal analysis performed on point processes would be to introduce a new variable to study. That could give rise to the study of marked spatial temporal point processes in which the mark would be either the gray intensity of the fusion or the docking times of the vesicles. Hence, more complex hypothesis could be addressed. As we pointed out, it is the difficulty to obtain the data for applications which can take advantage of it.
- Another variation of the described methodology could be assume a different underlying model for the events, e.g. a Cox process. Depending on the initial hypothesis defined by the researcher, more realistic models could be considered and tested.

Extensions to the spatio-temporal Boolean model described in Chapters 4 and 5 include:

- An interesting direct extension of the spatio-temporal Boolean model is the analysis of fluorescence in grey-level images. This approach will allow us to estimate whether the fluorescence associated to an event (i.e. whether there exists a minimal "unit", which is used as building block for larger

structures), and to determine the quantal of fluorescence associated to a single event. This improvement could be useful for applications like protein fluoresce study, or sensor networks, in which the energy produced for an event would be considered as the radiating energy of the sensor. It is important to note that this framework is flexible and general enough, as well as intuitive and direct enough, to incorporate other hypotheses and applications.

- Further extensions non-directly related with the estimation procedure described here could be considered. One interesting approach to estimate the probability density of durations of a spatio-temporal Boolean model would be the use of an extension of the Covariance so as to take into account the space and time dimension, $C(r, t)$. This method would allow us to obtain estimates for the same parameters of the proposed methodologies. Nevertheless, it is important to take into account that the covariance is obtained performing operations pixel by pixel, so its results are really sensitive the any kind of noise.

- In the specific biological application described herein either models assume that the events are homogenously scattered over the plasma membrane. Although this hypothesis seems to be accepted for most cells expressing clathrin, it can not be applied to the study of exocytosis since we show that this process is spatially and temporally clustered in hotspots. Then a methodology to model non-homogeneous point processes in space and time needs to be developed. In the same way, an extension to the non-homogenous Boolean model in space and time remains to be done. This would be useful to model other proteins taking part in the endocytic process which are know to be clustered, and so could not be modelled with the current spatio-temporal Boolean model.

All the methodologies and methods described in this thesis have been programmed in Matlab and R software. An interface for the study and segmentation of exocytosis has been developed completely in Matlab in order to extract the information needed for the statistical analysis. A toolbox for Matlab have been designed to simulate and estimate parameters from a spatio-temporal Boolean models. Either methods described in this document have been developed and are available under request.

To sum it up, we have defined and presented a methodology to estimate spatial temporal parameters of stochastic models which enable the research to

test hypothesis of a given model formally. The methodologies and models described herein could bring information of temporal distributions, intensities and size measures of stochastic processes observed in the spatio-temporal space. The proposed estimators have been tested by simulation studies and applied on real biological sequences to obtain information of the underlaying process which can not be obtained manually.

Appendix A

Experimental Procedures

A.1 Live Microscopy

Time-lapse images were acquired by Total Internal Reflection Fluorescent Microscopy (TIRFM). In this technique, a laser beam illuminates the sample beyond a critical angle and generates 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 an Evanescent Field is generated. This field is only $100-200$ nm thick and decays exponentially 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 not illuminated. In this way, it is possible to image membrane-associated events, such as endocytosis and exocytosis, with superb signal-to-noise (5; 69). 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 output was between 80 and 100 mW.

A.2 Cell Culture

A.2.1 Exocytosis

PtK_2 cells were grown in MEM, $100U/ml$ penicillin, $100\mu g/ml$ streptomycin, 2 mM L-glutamine, $1x$ nonessential amino acids (GIBCO BRL), and 10% FCS (complete medium). Cells used for microscopy were grown on 1-mm-thick, 30-mm-diam sapphire slides (Rudolf Brügger AG) in complete medium without phenol red. Infection with Recombinant Adenoviruses VSVG3-SP-YFP is based on previously described temperature-sensitive VSVG3-GFP (68), and differs from it by having a longer spacer between the last amino acid of VSVG and the start of YFP. PtK_2 cells were infected for 1 h at 37 degrees Celsius in $1ml$ of complete medium. After changing the medium, the cells were incubated for $6 - 20$ h at 39.5 degrees Celsius to accumulate the protein in the ER, and then used for microscopy. Sapphire slides with PtK_2 cells were transferred to a closed perfusion chamber (POC), which contained a glass coverslip ($0.17 \times 42mm$), a $25 - \mu m$ Teflon spacer, and complete medium with $20\mu g/ml$ cycloheximide. Samples were imaged $20 - 50$ min after a shift to 32 degrees Celsius, conditions under which VSVG3-SP-YFP can exit the ER.

A.2.2 Endocytosis

Clathrin. COS7 cells were grown in MEM, 100 U/ml penicillin, $100\mu g/ml$ streptomycin, 2 mM L-glutamine, and 10% FCS (complete medium). Cells were transfected with a DNA vector encoding for GFP-Clathrin Light Chain (GFP-LCa) using the Amaxa Nucleofector Kit, at 1ug of DNA/100000 cells. Cells were then plated onto MatTek glass bottom culture dishes (MatTek Corp.) at $20 - 50000$ cells/dish. The next day, complete medium was replaced with imaging buffer, containing $136mM$ NaCl, $2.5mM$ KCl, $2mM$ $CaCl_2$, $1.3mM$ $MgCl_2$, $10mM$ HEPES at pH 7.4 in double-distilled water. Cells were imaged at 37 degC. Under these conditions, cells maintained their viability for up to 8 hours.

Appendix B

Algorithms Details

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

B.1 Statistical Analysis of Point Patterns

This section briefly describes some implementation details for carrying out the statistical analysis of point patterns.

The statistical analysis was performed by using *R* software (49), an environment specifically designed for data analysis. That software can be freely downloaded from http://cran.r-project.org.

Statistical analysis to compare counts of fusions has been performed using the function *glm* from *MASS* library of the software package *R*.

If *counts* contains the counts observed and *volume* the spatio-temporal volumes, then the following *R*-code provides us with the statistical analysis given in the paper.

```
null.fit<- glm(counts~offset(log(volume)),
           family=negative.binomial(theta=1,link="log"))
cells.fit <- glm(counts~cells+offset(log(volume)),
             family=negative.binomial(theta=1,link="log"))
```

```
summary(id.fit) anova.glm(null.fit,cells.fit,test="F")
```

For the spatial analysis we used a package for *R* called *Spatstat* (7). This package allows us the definition of objects representing spatial point patterns through the function called *ppp()*, whose inputs are the positions of the points of the pattern (fusion locations) and observation window (cell membrane border in our case). After building a *ppp* object one can evaluate spatial point patterns with the following functions: *Fest()*, *Gest()* and *Kest()*, which calculate the empty space function, the nearest neighbor function and the $\mathcal{K}$-function, respectively.

The following *R*-code was used to create the point pattern and the window associated. The (x, y)-locations of the fusion events are stored in the variables X and Y. Variable *win* corresponds to the observe window defined by a polygonal line stored in the variable *edge*. If the fusions are marked, variable M stores that value.

```
win  <- owin(poly=list(edge[,1],edge[,2]));
exos <-ppp(X,Y,marks=M,window=win);
```

For the temporal and spatial temporal analysis a package called *Splancs* (53) for *R* was used. Function *kernel2d()* was used to obtain the plot of the spatial estimated intensities of the point pattern. Function *stkhat()* was used to calculate both temporal and spatial temporal $\mathcal{K}$-functions. The Monte Carlo test of spatial temporal was performed by means of function *stmctest()*. Spatstat and Splancs can be freely downloaded from the website `http://cran.r-project.org`.

B.2 Functional Data Analysis

The toolbox FDA for MATLAB developed by J.O. Ramsay was used to perform the functional data analysis.

First of all, the function *create_bspline_basis* was used for creating the functional object. This functional object was used to convert our estimates in a continuous function by means of the function *data2fd*. Next, function *smooth_fd* was applied over the function representing the data so as to smooth it. We use a penalty factor in the second derivative of the function. Finally, function *deriv* was used to obtain first and second derivative of the smoothed curves.

```
bsBasisObj=create_bspline_basis([t(1) t(end)],nbasis,norder,breaks);
```

```
derivsmooth=2;
penaltyFactor=0.5;
pEstfd=data2fd(pEst,t,bsBasisObj);
fdParObj=fdPar(bsBasisObj,derivsmooth ,penaltyFactor);
pEstfd2=smooth_fd(pEstfd,fdParObj);
D2pEst=deriv(pEstfd2,2);
```

The basis of bsplines used comprised $15-20$ basis of order 3 in all the studies. This values should be modified depending on the kind of data.

Appendix C

Description of Supplementary Material

Supplementary videos have been generated in order to visually show the kind of data we are dealing with. In particular, Chapters 2, 4 and 5 which study spatial temporal data refer to these image sequences. All the videos are in QuickTime format 'mov'.

Table C.1 summarizes some details of these image sequences. First Column corresponds to the exact name of the video included in the supplementary material. Second and third columns display the chapter in which the videos are referenced and the protein that motivated this particular study. Finally, last column gives the length of the image sequence in minutes and seconds.

A description in detail of each image sequence follows.

***V*ideo** 2.1**,** A cell which was transfected to permit the exocytosis visualization. Exocytosis were automatically detected. A white color square is shown in the video at the position where exocytosis have been observed. The membranes of the cell have been manually delineated in white color for better visualization.

***V*ideo** 2.2**,** A cell which was transfected to permit the exocytosis visualization. Exocytosis positions are marked with a color dot. The colors of the dots represent the time when the exocytosis was observed, and so similar colors denote exocytosis which have happened close in time. This permit to test

Table C.1: Description of supplementary material

Name	Chapter	Studying	Length
Video 2.1	2	Exocytosis	$4'10''$
Video 2.2	2	Exocytosis	$1'40''$
Video 2.3	2	Exocytosis	$3''$
Video 2.4	2	Exocytosis	$3''$
Video 4.1	4	Endocytosis	$1'40''$
Video 4.2	4	Endocytosis	$1'40''$
Video 4.3	4	Endocytosis	$1'40''$
Video 5.1	5	Endocytosis	$25''$
Video 5.2	5	Endocytosis	$24''$
Video 5.3	5	Endocytosis	$23''$
Video 5.4	5	Endocytosis	$16''$
Video 5.5	5	Endocytosis	$60''$

possible spatial temporal clusters by visual inspection.

***V*ideo** 2.3**,** Three dimensional profile of an exocytosis over time, where the z-axis represents the gray intensity of the fusion and the x-y plane the membrane surface of the cell.

***V*ideo** 2.4**,** Two dimensional profile of an exocytosis over in real size.

***V*ideo** $4.1 - 4.3$**,** are simulated image sequence of temporal Boolean models generated for the simulation study. Videos were constructed using different parameters to observe how these differences affect the estimation procedure. Table C.2 summarizes the combination of parameters used. The distribution function for the durations was uniform for all the examples.

***V*ideo** 5.1**,** corresponds to a realization of a TBM in which the grains are disks with uniform radii in the interval $[11, 15]$ pixels, the durations are uniformly distributed (with a mean of 6 seconds) and the spatial temporal intensity (mean number of germs per unit area and time) is 0.000035. Images are 512×512 pixels in size and the sampling rate is 4 frames per second.

Table C.2: Combination of parameters used as an example of TBMs

Name	Vol. fraction p	Radii	Duration
Video 4.1	0.10	U(8,12)	U(4,8)
Video 4.2	0.15	U(8,12)	U(7,11)
Video 4.3	0.20	U(6,10)	U(10,14)

Video $5.2 - 5.5$**,** are accumulated videos generated by aggregation of frames of Video 5.1 for different values of k and δ. Table C.3 summarizes the combination of parameters used in the aggregation process.

Table C.3: Combination of parameters used to generate aggregated TBMs

Name	k	Lag
Video 5.2	1	0
Video 5.3	2	3
Video 5.4	2	6
Video 5.5	5	9

Bibliography

[1] A. Agresti. *Cualitative Data Analysis*. Wiley, 2nd edition, 2002. 2.1

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Publishing, New York and London, third edition, 1994. 1.1.1, 2.1

[3] M. W. Allersma, L. Wang, D. Axelrod, and R.W. Holz. Visualization of regulated exocytosis with a granule-membrane probe using total internal reflection microscopy. *Molecular Biology of the Cell*, **15**(10):4658–4668, 2004. 1.1.1

[4] A. Aubert and D. Jeulin. Estimation of the influence of second- and third-order moments on random sets reconstructions. *Pattern Recognition*, **33**(6):1083–1104, June 2000. 3.3

[5] D. Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Traffic*, **2**:764–774, 2001. (document), A.1

[6] G. Ayala, R. Sebastián, M.E. Díaz, E. Díaz, and D. Toomre. Analysis of overlapping short-lived events. *Submitted*, 2005. 5.2.2

[7] Adrian Baddeley, Rolf Turner, with contributions by Marie-Colette van Lieshout, and 17 others. *spatstat: Spatial Point Pattern analysis, model-fitting and simulation*, 2005. R package version 1.6-1. B.1

[8] S.D. Corner and S.L. Schmid. Regulated portals of entry into the cell. *Nature*, **4**:37–44, 2003. 1.1.2

[9] D.R. Cox and V. Isham. *Point Processes*. Monographs on Applied Probability and Statistics. Chapman and Hall, London, 1980. 1.2.1, 2.2.3

[10] D.R. Cox and P.A.W. Lewis. *The Statistical Analysis of Series of Events*. Chapman and Hall, London, 1966. 1.2.1

[11] Noel A.C. Cressie. *Statistics for Spatial Data. Revised Edition.* John Wiley and Sons, New York, 1993. (document), 1, 3.3, 3.3.1, 5.2, 5.2.1

[12] 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. 1.1.1, 2.1

[13] P.J. Diggle. Binary mosaics and the spatial pattern of heather. *Biometrics*, **37**:531–539, 1981. 3.3.1

[14] P.J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Arnold, London, second edition, 2003. (document), 1, 1.2.1, 1.2.1, 2.1, 2.2.3, 2.2.3, 2.3.2, 3.3.1, 5.2

[15] P.J. Diggle, A.G. Chetwynd, R. Haggkvist, and S.E. Morris. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research*, **4**:124–136, 1995. 2.3.4

[16] E. Dougherty and A. Grigoryan. Automatic counting of illuminated spheres in a random Boolean model. In *Image Processing: Algorithms and Systems*, **4667**, pages 170–180. The International Society for Optical Engineering, 2003. 3.3, 4.2.2

[17] E.R. Dougherty and R. Lotufo. *Hands-on Morphological Image Processing*. SPIE Press, Washington, first edition, 2003. 3.2.1, 4.4

[18] O. Dousse, P. Mannersalo, and P. Thiran. Latency of wireless sensor networks with uncoordinated power saving mechanisms. *MobiHoc*, 2004. 3.3

[19] M. Ehrlich, W. Boll, A. Van Oijen, K. Hariharan, R. Chandran, M.L. Nibert, and Kirchhausen T. Endocytosis by random initiation and stabilization of clathrin-coated pits. *Cell*, **118**:591–605, 2004. 1, 1.1.2, 1.1.2, 1.1.2, 3.3.1, 3.5, 4.4

[20] I. Epifanio and G. Ayala. A random set view of texture classification. *IEEE Transactions on Image Processing*, **11**(8):859 –867, August 2002. 3.3

[21] M.A. Gallego and A. Simó. Random closed set models: estimating and simulating binary images. *Image Analysis & Stereology*, **22**:72–91, 2003. 3.3.3

[22] P. García. *Texture analysis using the one dimensional Boolean model*. PhD thesis, University Jaume I, Castellón, Spain, October 1999. 3.3, 3.3.1

[23] 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. 3.3, 3.3.1

[24] A.C. Gatrell, T.C. Bailey, P.J. Diggle, and B.S. Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions - Institute of British Geographers*, **21**:256274, 1996. 1.2.1, 2.2.3

[25] R. C. Gonzalez and R. Woods. *Digital image processing.* Addison-Wesley Publishing Company, New York, 1992. 2.2.1

[26] J. Goutsias, R.P.S. Mahler, and H.T. Nguyen, editors. *Random Sets. Theory and Applications*, **97** of *The IMA Volumes in Mathematics and its Applications*. Springer, 1997. 1, 3.3

[27] J.P. Henry, F. Darchen, and S. Cribier. Physical techniques for the study of exocytosis in isolated cells. *Biochimie*, **80**:371–377, 1998. 1.1.1, 2.1

[28] K. Hirschberg, C.M. Miller, Ellenberg J., Presley J.F., Siggia E.D., Phair R.D., and Lippincott-Schwartz J. Kinetic analysis of secretory protein traffic and characterization of golgi to plasma membrane transport intermediates in living cells. *Journal of Cell Biology*, **143**:1485–1503, 1998. 2.1, 2.4

[29] D. Jeulin. *Advances in Theory and Applications of Random Sets*. World Scientific, Singapore, 1997. 3.3.3

[30] M. Kaksonen, C.P. Toret, and D.G. Drubin. A modullar design for the clathrin- and actin-mediated endocytosis machinery. *Cell*, **123**:305–320, 2005. 4.4

[31] P. Keller, D. Toomre, E. Diaz, J. White, and K. Simons. Multicolor imaging of post-golgi sorting and trafficking in live cells. *Nature Cell Biology*, **3**:140–149, 2001. (document), 1, 1.1.1, 2.1, 2.2.1, 2.3.2, 2.4

[32] G. Kreitzer. Three dimensional analysis of post-golgi carrier exocytosis in epithelial cells. *Nature Cell Biology*, **5**:126–139, 2003. 1.1.1, 2.1

[33] A.M. Law and W.D. Kelton. *Simulation Modeling and Analysis*. McGraw Hill, third edition, 2000. (document), 3.4, 4.2.2, 4.3, 5.2.2

[34] M.N.M. Lieshout and E.W. Zwet. Maximum likelihood estimation for the bombing model. Probability, Networks and Algorithms PNA-R0008, Stichting Mathematisch Centrum, 1090 GB, Amsterdam, August 2000. 4.2.1

[35] G. Matheron. *Eléments pour une théorie des milieux poreux*. Masson, Paris, 1967. (document), 1, 3.3, 4.2.2

[36] G. Matheron. *Random sets and Integral Geometry*. Wiley, London, 1975. (document), 1, 1.2.2, 3.3, 3.3.3, 3.3.3, 4.2.2

[37] C.J. Merrifield, M.E. Feldman, L. Wan, and W. Almers. Imaging action and dynamin recruitment during invagination of single clathrin-coated pits. *Nature Cell Biology*, **4**:691–698, 2002. 1.1.2

[38] I. Molchanov. *Statistics of the Boolean Model for Practitioners and Mathematicians*. John Wiley and Sons, Chichester, 1997. 3.3, 3.3.1, 3.3.3, 3.3.3, 4.2.1, 4.2.2, 4.3, 5.2

[39] I. Molchanov. *Stochastic Geometry Likelihood and Computation*, **80** of *Monographs on Statistics and Applied Probability*, chapter 7, pages 285–331. Chapman & Hall /CRC, 1999. 3.3

[40] I. Molchanov. *Theory of Random Sets (Probability and its Applications)*. Springer, 2005. 1

[41] I.S. Molchanov. Statistics of the boolean model: from the estimation of means to the estimation of distributions. *Adv. Appl. Prob.*, **27**:63–86, 1995. 4.5

[42] D.J. Nott and T. Ryden. Pairwise likelihood methods for inference in image models. *Biometrika*, **86**:661–676, 1999. 4.2.1

[43] D.J. O'Brien, J.B. Kaneene, A. Getis, J.W. Lloyd, M.R. Rip, and R.W. Leader. Spatial and temporal distribution of selected canine cancers in michigan. *Preventive Veterinary Medicine*, **42**:1–15, 1999. 1

[44] G.E. Palade. Intracellular aspects of the process of protein secretion. *Science*, **189**:347–1358, 1975. 1.1.1

[45] S. Pfeiffer, S.D. Fuller, , and K. Simons. Intracellular sorting and basolateral appearance of the g protein of vesicular stomatitis virus in madin-darby canine kidney cells. *Journal of Cell Biology*, **101**:470–476, 1975. 2.4

[46] A.A. Philimonenko and P. Janaček, J. Hozk. Statistical evaluation of colocalization patterns in immunogold labeling experiments. *Journal of Structural Biology*, **132**:201–210, 2000. 1

[47] I.A. Prior, C. Muncke, R.G. Parton, and J.F. Hancock. Direct visualization of ras proteins in spatially distinct cell surface microdomains. *Journal of Cell Biology*, **160**:165–170, April 2004. 1

[48] J.L. Quenec'h, M. Coster, J.L. Chermant, and D. Jeulin. Study of the liquid-phase sintering process by probabilistic models: application to the coarsening of wc-co cermets. *Journal of Microscopy*, **168**:3–14, 1992. 4.5

[49] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. 2.3.1, B.1

[50] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, first edition, 1997. 4.2.2

[51] S. Reader. Using survival analysis to study spatial point patterns in geographical epidemiology. *Social Science & Medicine*, **50**:985–1000, 2000. 1

[52] B.D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, **13**:255–266, 1976. (document), 1

[53] Barry Rowlingson, Peter Diggle, adapted, packaged for R by Roger Bivand, pcp functions by Giovanni Petris, and goodness of fit by Stephen Eglen. *splancs: Spatial and Space-Time Point Pattern Analysis*, 2004. R package version 2.01-15. B.1

[54] P. Salembier and J. Serra. Flat zones filtering, connected operators, and filters by reconstruction. *IEEE Transactions On Image Processing*, **4**(8):1153–1160, August 1995. 2.2.1

[55] K Schladitz, S. Peters, D Reinel-Bitzer, A. Wiegmann, and J. Ohser. Design of acoustic trim based based on geometric modelling and flow simulation for nin-woven. Technical Report 72, Fraunhofer ITWM, Febrary 2005. 3.3, 4.2.2

[56] J. Schmoranzer, M. Goulian, D. Axelrod, and S.M. Simon. Imaging constitutive exocytosis with total internal reflection fluorescence microscopy. *The Journal of Cell Biology*, **149**:23–31, April 2000. 2.1, 2.2.1

[57] J. Schmoranzer, G. Kreitzer, and S.M. Simon. Migrating fibroblasts perform polarized, microtubule-dependent exoctosis towards the leading edge. *Journal of Cell Science*, **116**:4513–4519, November 2003. 1.1.1

[58] J. Schmoranzer and S.M. Simon. Role of microtubules in fusion of post-golgi vesicles to the plasma membrane. *Molecular Biology of the Cell*, **14**:1558–1569, April 2003. 1.1.1, 2.1

[59] F.P. Schoenberg, D.R. Brillinger, and P. Guttorp. *The encyclopedia of environmetrics*. Jhon Wiley, Chichester, second edition, 2002. 1.2.1

[60] J. Serra. The boolean model and random sets. *Computer Graphics and Image Processing*, **12**:99–126, 1980. (document), 3.3.3

[61] J. Serra. *Image Analysis and Mathematical Morphology. Volume 1.* Academic Press, San Diego and London, fourth edition, 1993. 2.2.1

[62] J.P. Serra. *Image Analysis and Mathematical Morphology*, **1**. Academic Press, 1982. (document), 1, 3.2.1, 3.3, 3.3.3, 4.2.1, 4.2.2, 4.3, 5.2

[63] S. Server, H. Damke, and S.L. Schmid. Garrotes, springs, ratchets, and whips: putting dynamin models to the test. *Traffic*, **1**(5):385–392, May 2000. 1.1.2

[64] B.M. Slepchencko, J.C. Schaff, J.H Carson, and L.M. Loew. Computational cell biology: Spatiotemporal simulation of cellular events. *Annu. Rev. Biophys. Biomol. Struct.*, **31**:423–41, 2002. 1

[65] P. Soille. *Morphological Image Analysis. Principles and applications*. Springer-Verlag, Germany, Berlin, 1999. 2.2.1

[66] D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic Geometry and its Applications*. Wiley, Berlin, second edition, 1995. (document), 1, 3.1, 3.3, 3.3.1, 3.3.3, 3.3.3, 3.4, 4.2.2, 4.2.2, 5.2, 5.2.1

[67] D. Stoyan and H. Stoyan. *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics*. Wiley, 1994. 3.3.1, 5.2

[68] D. Toomre, P. Keller, J. White, J.C. Olivo, and K. Simons. Dual-color visualization of trans-golgi network to plasma membrane traffic alone microtubules in living cells. *J. Cell Sci.*, **112**:21–33, 1999. 2.1, A.2.1

[69] D. Toomre and D.J. Manstein. Lighting up the cell surface with evanescent wave microscopy. *Trends Cell Biol.*, **11**:298–303, 2001. (document), 1.1.3, 2.1, 2.2.1, 4.4, A.1

[70] D. Toomre, J.A. Steyer, P. Keller, W. Almers, and K. Simons. Fusion of constitutive membrane traffic with the cell surface observed by

evanescent wave microscopy. *J. Cell Biol*, **149**:33–40, 2000. (document), 1.1.1, 2.1

[71] J.Y. Toshima, J. Toshima, Kaksonen M., Martin A.C., King D.S., and Drubin D.G. Spatial dynamics of receptor-mediated endocytic trafficking in budding yeast revealed by using fluorescent alpha-factor derivatives. *PNAS*, **103**:5793–5798, 2006. 4.4, 4.4

[72] J. van den Berg, R. Meester, and D.G. White. Dynamic boolean models. *Stochastic Processes and their Applications*, **69**:247–257, 1997. 3.3

[73] D. Zenisek, J.A. Steyer, M.E. Feldman, and W. Almers. A membrane marker leaves synaptic vesicles in milliseconds after exocytosis in retinal bipolar cells. *Neuron*, **35**:1085–1097, 2002. (document), 2.2.1