Chapter 2
Spatial Point Process Analysis of Promyelocytic Leukemia Nuclear Bodies

Philip P. Umande and David A. Stephens

Abstract There has been widespread interest in the nuclear body (NB) Promyelocytic leukemia (PML) because of its link to several human disorders, including Promyelocytic leukemia and AIDS. The notion of PML NB interaction with its surrounding and other NBs such as RNA Polymerase II (RNA Pol II) is of great importance as it can improve our understanding of the function of PML. In this paper, spatial point process methods are used to conduct multivariate analysis to assess the relationship between the spatial locations of PML NBs relative to RNA Pol II. We also propose a model for PML NB locations. By fitting a model to the PML NBs we are able to gain insight into how PML NBs are distributed across the nucleus in relation to themselves and the nuclear boundary.

Keywords Spatial Point Pattern • PML • RNA Pol II • Marked Point Process • Inhomogeneous Poisson Process • K-function

2.1 Introduction

The notion of Promyelocytic leukemia (PML) nuclear body (NB) interaction with its surrounding is one of great importance. Lanctot et al. (2007) have reported that gene expression is mediated by interaction between chromatin and protein complexes. Dellaire and Bazett-Jones (2004) have proposed that PML NBs are dynamic sensors of cellular stress, that associates with regions of DNA damage. Borden

P.P. Umande (✉)
Department of Mathematics, Imperial College, 180 Queens Gate, London SW7 2AZ, UK
e-mail: philip.umande00@imperial.ac.uk

D.A. Stephens
Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal QC, H3A 2K6, Canada

N.M. Adams and P.S. Freemont (eds.), *Advances in Nuclear Architecture*,
DOI 10.1007/978-90-481-9899-3_2,

(2002) has suggested that PML NBs tend to be near certain nuclear compartments such as Cajal/coiled bodies, cleavage bodies, and splicing speckles .

Knowledge of the PML NBs in relation to other structures may provide clues to PML NB functions. Furthermore, its relative location may give insight into what specific targets it regulates (Borden 2002). Following this, most reported strategies for assessing PML NB functions in essence are designed to answer questions relating to which nuclear structures the bodies are near to, what other macromolecules localize with the body, and the effects of disrupting the locations of the body (Borden 2002). Wang et al. (2004) analysed the correlation between the minimum locus-PML distances against their transcriptional activity to show that PML associate with transcriptionally active genomic regions.

The Imperial College Centre for Structural Biology has been able to provide images obtained via confocal microscopy. Such images provide the spatial locations (three-dimensional coordinate space) of PML and centroids of other nuclear bodies (RNA Polymerase II etc.) (see Fig. 2.1), and also the nuclear boundary (see Fig. 2.2). These data will be used for the quantitative analysis presented in this paper. Note that we are considering replicated data, that is, data for several cell nuclei representing multiple independently and identically distributed realisations of some spatial stochastic process. Replication can add complications to statistical inference (see Section 2.4 on modelling PML NB data), but is important as it provides further credibility to the outcome of the statistical analysis.

We shall carry out multivariate (or equivalently, *marked*) point pattern analysis so that we can provide some statistical evidence for biological ideas regarding

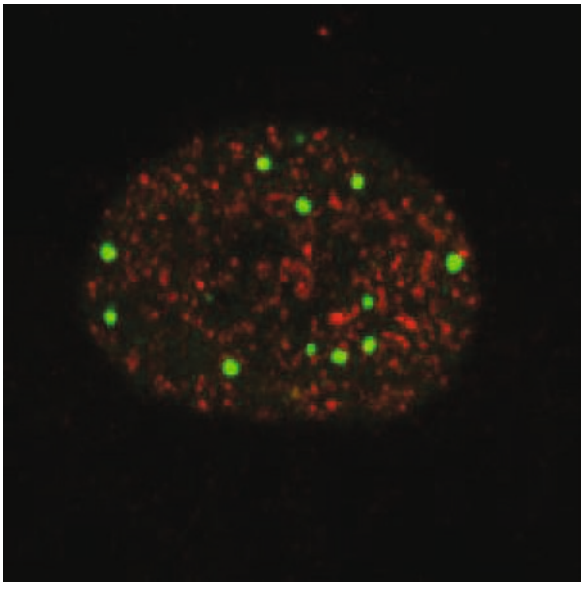

Fig. 2.1 Microscopy image of cell nucleus with PML bodies (*green*) and RNA-Polymerase II (*red*)

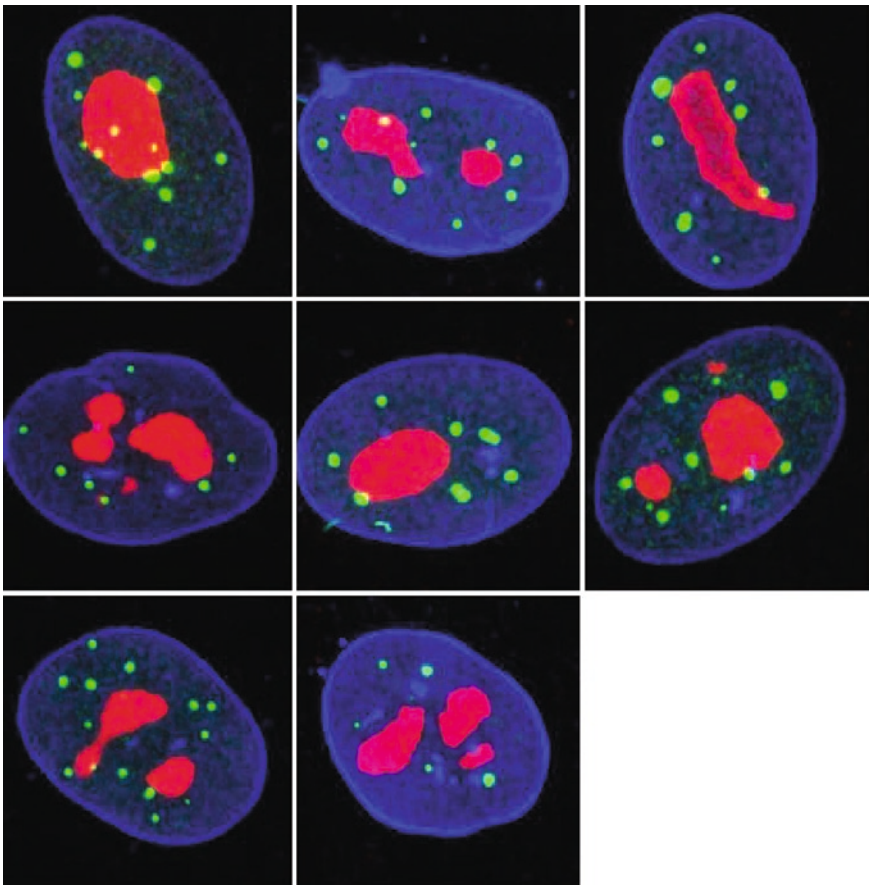

Fig. 2.2 PPSD2 data: Microscopy image of cell nucleus with PML bodies (*green*) and nucleoli (*red*), with nuclear lamina (*blue*)

PML NBs association with other nuclear bodies. Specifically, Borden (2002) stated that general transcription factors such as RNA Polymerase II do not colocalize with PML bodies. This is consistent with experimental results obtained by Xie and Pombo (2006) which supported the view that although PML NBs are present in transcriptionally active areas, they are not generally sites of polymerase II assembly. We shall attempt to use spatial point pattern analysis to confirm or otherwise these findings. We also explore and discuss how one might go about modelling the relationship between PML NB size with its positioning in the nuclear interior.

This paper is structured as follows. In the next section we will provide a background to point process theory and its application for multiple event types. We shall then go onto discuss how multivariate spatial analysis can be used for assessing the

relationship between the spatial location of PML NBs relative to RNA Polymerase II. In the final section we discuss a possible approach for modelling PML NB locations before ending with concluding remarks.

2.2 Spatial Point Processes: Theory, Models and Statistics

The earliest discussions on spatial point processes date back to the early 1950s when used by Skellam (1952) and Thompson (1955) in statistical ecology. The work of Matérn in 1960, later re-published in 1986, has been viewed as pioneering (Cox and Isham 1980; Kemp 1988). More recently, Møller and Waagepetersen (2004) have provided more detailed accounts of modern spatial point process theory, statistics and models; see also Stoyan (2006).

We begin by introducing notation, definitions and some important concepts. These initial definitions and concepts are as those given by Cressie (1991), Karr (1991), Stoyan and Stoyan (1994), Stoyan et al. (1995), and Stoyan (2006). We begin by considering a collection of points observed within the nucleus, resulting in a data set having two or three coordinates if we utilize microscopic slices or the entire confocal image respectively. Throughout we denote locations by $\boldsymbol{x}=(x,y)$ or $\boldsymbol{x}=(x,y,z)$ and use subscripts $i=1,\ldots,n$ to denote the locations of n observations in the data set. We shall denote the 2-dimensional disk (3-dimensional ball) of radius r centered at $\boldsymbol{x}$ by $\mathcal{B}^d(\boldsymbol{x},r)$, although the superscript d will sometimes be omitted.

2.3 Spatial Point Process: Definitions and Calculations

We assume that the points – PMLs, genomic loci, other nuclear bodies stained in the experiment – observed in the images follow a *spatial point process*, that is, the points occur according to a random mechanism that we can characterize in terms of the distribution of the numbers of points that occur in disjoint spatial regions. For a spatial point process, denoted X, we write

$$X(B)=\sum_{\boldsymbol{x}_i\in X}\mathbb{I}_B(\boldsymbol{x}_i),$$

where $\mathbb{I}_B(.)$ denotes the indicator function for set B, to indicate the count of the number of points of X observed in the region B. For set S, the notation $X(S)=n$ means that S contains n points of X. A *spatial point pattern* – a realisation of a spatial point process – is defined through the locations of points (Cressie 1991). We will at times make the assumption that the spatial point process X is *stationary* or *isotropic*; X is *stationary* (or equivalently, *homogeneous*) if it has the property that $X_{\boldsymbol{x}'}=\{\boldsymbol{x}'+\boldsymbol{x}:\boldsymbol{x}\in X\}$ has the same distribution for all $\boldsymbol{x}'\in R^d$. Also, the point process X is said to be *isotropic* if it is invariant under rotation. Stationarity and isotropy

can be very important assumptions in spatial point process analysis. Practitioners will at times (often implicitly) assume that stationarity holds (without carrying out formal tests for stationarity), or be content that it holds approximately so that certain point process techniques can be readily adopted (see Baddeley et al. 1992; Glasbey and Roberts 1997 as examples).

The *intensity measure*, Λ, of X is a point process characteristic analogous to the mean of real-valued random variables, that is defined as

$$\Lambda(B) = \mathbb{E}[X(B)] = \int \boldsymbol{x}(B) P(d\boldsymbol{x}) = \mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \mathbb{I}_B(\boldsymbol{x})\right]$$

that is, the expected number of points lying in the region B. In the *homogeneous* case it suffices to consider an *intensity*, λ since then $\Lambda(B) = \lambda \nu(B)$, where $\nu(B)$ is the area (or volume) of B. In general, we define the *kth moment measure* $\mu^{(k)}$ by

$$\mu^{(k)}\left(B_1 \times \cdots \times B_k\right) = \mathbb{E}\left[\prod_{i=1}^{k} X(B_i)\right]$$

for sets $B_1, \ldots, B_k$, and $\times$ denoting the Cartesian product. Stoyan and Stoyan (1994) provide a more detailed account on higher order moments, including geometrical interpretations.

2.4 Binomial and Poisson Point Processes

2.4.1 The Binomial Point Process

The points $\boldsymbol{x}_1 \ldots \boldsymbol{x}_n$ form a *Binomial point process*, $X_{\mathrm{Bin}(W,n)}$ in the set W if they are independently and uniformly distributed inside W, with

$$\mathbb{P}(\boldsymbol{x}_1 \in B_1, \ldots, \boldsymbol{x}_n \in B_n) = \mathbb{P}(\boldsymbol{x} \in B_1) \cdot \cdots \cdot \mathbb{P}(\boldsymbol{x} \in B_n) = \frac{\nu(B_1) \cdot \cdots \cdot \nu(B_n)}{\nu(W)^n}$$

for $B_1, \ldots, B_n$ subsets of W. The intensity, λ, of this process is given by

$$\lambda = \frac{np(B)}{\nu(B)} \quad \text{where} \quad p(A) = \frac{\nu(A)}{\nu(W)}, \quad A \subseteq W.$$

The simulation of virtually all spatial point processes in W requires simulating a binomial point process (n points uniformly inside W). Simulating n events uniformly inside W, a unit square or cube is straightforward; for a unit cube, one simply superimposes n independent uniform random points, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ where $\boldsymbol{u}_i = (u_{i1}, u_{i2}, u_{i3})$, and $u_{ij} \sim Uniform(0, 1)$. Once simulated inside the unit cube, we can then apply scaling and translation in order to obtain a simulation inside any fixed cuboid. Coordinate

transformation can be used for simulating uniformly inside a sphere. That is, if $\boldsymbol{u}=(u_1,u_2,u_3)$ is a uniform point in the unit cube, then $\boldsymbol{x}=(x_1,x_2,x_3)$ where

$$x_1 = R\sin(\theta)\cos(\phi) \qquad x_2 = R\sin(\theta)\sin(\phi) \qquad x_3 = R\cos(\theta)$$

and $R=u_1^{1/3}$, $\theta=\arccos(1-2u_2)$, and $\varphi=2\pi u_3$ is uniform inside the unit sphere. Once again, the relevant scaling can be applied for the case where simulation inside an ellipsoid is required. For less straightforward sets and irregularly shaped regions, W_0, rejection sampling (see for example Ripley 1987) can be used. This involves, for example, simulating uniformly inside $W\supset W_0$ and retaining the points that lie in W_0. Simulation is repeated until the desired number of points are obtained.

Data sets PPDS2 and PPDS3 provide the locations of points that constitute empty space inside the nucleus. We exploit this information as a means of simulating u points uniformly inside the nucleus. Specifically, for the u points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_u$ that are classified as empty space, n such points are chosen at random (without replacement). Random selection is made by equipping each of the $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_u$ with a unique integer $k\in\{1,\ldots,u\}$. The point $\boldsymbol{x}$ is selected if the randomly chosen $y\in\{1,2,\ldots,u\}$ is its assigned integer. Occasionally, we attempt to simulate uniformly inside the nuclear interior by adopting methods for simulating uniformly inside a convex hull (Fishman 1996).

2.4.2 The Homogeneous Poisson Point Process

A default standard model for point patterns is the *homogeneous Poisson process*. The homogeneous (stationary) Poisson process X_P, is defined by the following postulates:

(i) For some constant $\lambda>0$, and set B, $X_\mathrm{P}(B)$ follows a Poisson distribution with mean $\lambda\nu(B)$. The parameter λ is the intensity. For the three-dimensional case, this can be interpreted as the number of events per unit volume.
(ii) Given $X_\mathrm{P}(B)=n$, the n events in B form an independent sample from the uniform distribution on B.
(iii) If $B_1,\ldots,B_s$ are disjoint sets then $X_\mathrm{P}(B_1),\ldots,X_\mathrm{P}(B_s)$ are independent Poisson random variables with mean $\lambda\nu(B_1),\ldots,\lambda\nu(B_s)$. Therefore

$$\mathbb{P}(X_\mathrm{P}(B_1)=n_1,\ldots,X_\mathrm{P}(B_s)=n_s) = \frac{\nu(B_1)^{n_1}\cdots\nu(B_s)^{n_s}}{n_1!\cdots n_s!}\lambda^{\sum_{i=1}^{s}n_i}\exp\left(-\sum_{i=1}^{s}\lambda\nu(B_i)\right)$$

These postulates are simultaneously the definition of *complete spatial randomness* (CSR). The *void* probabilities of X_P are given by $\mathbb{P}(X(B)=0)=\exp\left(-\lambda\nu(B)\right)$. The first order moment Λ follows from property (i) and is given by

$$\Lambda(B)=\mathbb{E}[X(B)]=\lambda\nu(B)$$

The homogeneous Poisson spatial point process can be simulated inside W directly from the first two postulates. First we simulate $n \sim \text{Poisson}(\lambda\nu(B))$, then simulate n points uniformly inside W (that is, simulate $\tilde{\boldsymbol{x}} \sim X_{\text{Bin}(W,n)}$).

2.4.3 Inhomogeneous Poisson Point Process

One of the most simple alternatives to the homogeneous Poisson point process is the *inhomogeneous Poisson point process*. The *inhomogeneous Poisson point process* is obtained by replacing the intensity λ by a spatially varying density $\lambda(\boldsymbol{x})$. Let Λ be a diffuse Radon measure on $\mathbb{R}^d$. An *inhomogeneous Poisson point process* is a point process possessing the following two properties:

(i) The number of events in a bounded B has a Poisson distribution with mean $\Lambda(B)$

$$\mathbb{P}(X(B) = n) = \frac{(\Lambda(B))^n}{n!}\exp\big(-\Lambda(B)\big), \quad \text{for} \quad n \in \{0,1,2,\ldots\}$$

(ii) The number of points in k disjoint sets form k independent random variables.

The function $\Lambda(B)$ can be written as

$$\Lambda(B) = \int_B \lambda(\boldsymbol{x})d\boldsymbol{x}$$

The function $\lambda(.)$ is the *intensity function* of the inhomogeneous point process. An inhomogeneous Poisson spatial point process in a set W can be simulated by using the rejection or *random thinning* algorithm of Lewis and Shedler (1976). The algorithm for an inhomogeneous Poisson process with intensity function $\lambda(\boldsymbol{x})$ is as follows: Simulate a homogeneous spatial Poisson point process of intensity λ_{max}, where λ_{max} is the largest intensity value over W. Then independently delete each point $\boldsymbol{x}_i$ with probability

$$1 - \frac{\lambda(x_i)}{\lambda_{\text{max}}} \tag{2.1}$$

The retained points form a realisation of events from an inhomogeneous Poisson process with intensity function $\lambda(\boldsymbol{x})$.

2.4.4 Estimating Intensity

Estimation of many point process functions rely on the estimation of the intensity of a stationary point process. Given a sampling region W, a natural unbiased estimator

for the intensity of a stationary point process is $\hat{\lambda} = X(W)/\nu(W)$. The intensity function $\lambda(.)$ of an inhomogeneous Poisson process can also be estimated using parametric or non-parametric methods such as kernel-based estimation (see, for example, section 8.5.1 of Cressie 1991).

2.4.5 Edge-Effects

Estimating point process functions of interest in some bounded region W, the *sampling* or *observation window*, is not trouble-free. Problems generally encountered are those arising from *edge effects*, that is, estimation problems created by not being able to observe data outside the edges of the observation region. These problems are usually encountered when the region W on which the point pattern is observed is a subset of a larger region on which the process is defined. Therefore estimation of summary statistics is biased by having censored events which may be interacting with events in the observation window. The methods of dealing with edge effects can be split into three categories, the simplest being the use of border methods (Diggle 2003), using estimators that explicitly account for edge effects, and wrapping W into a torus by identifying opposite edges. However, the toroidal wrapping technique does not generally apply to the confocal microscopy data.

2.5 Testing for Spatial Point Processes

2.5.1 The Empty Space Function F

Let X be a stationary and isotropic point process. That is, all probability distributions associated with X are invariant under rotation and translation (Baddeley et al. 1992). The *empty space function* of X, denoted F, or $F(r)$ from now onwards, is the probability distribution of the distance from an arbitrary point to the nearest event. That is, for $r \geq 0$

$$F(r) = \mathbb{P}(D(0,X) \leq r) = \mathbb{P}(X(\mathcal{B}^d(0,r)) > 0)$$

where $D(\boldsymbol{x}, A) = \inf\left\{\|\boldsymbol{x} - \boldsymbol{x}'\| : \boldsymbol{x}' \in A\right\}$ is the shortest (Euclidean) distance from $\boldsymbol{x}$ to A. For homogeneous Poisson process with intensity λ is given by

$$F(r) = 1 - \exp\left(-\lambda\nu(\mathcal{B}^d(0,r))\right) = \begin{cases} 1 - \exp\left(-\lambda\pi r^2\right) & d = 2 \\ 1 - \exp\left(-4\lambda\pi r^3/3\right) & d = 3 \end{cases} \tag{2.2}$$

Using (2.2) and by estimating the empty space function of a point pattern we can assess whether there is regularity or aggregation (clustering) in a point pattern. Estimated values of $F(r)$ greater than that given by (2.2) suggests that there is regularity, while lower values suggest aggregation (Baddeley et al. 1992).

Baddeley et al. (1992) state that the empty space function is typically estimated by taking a fine grid in the sampling region W and computing the distance from each grid point to the nearest event. This technique results in edge effects as we are unable to search for points outside W. The only approach currently in use is the *border method* (Baddeley et al. 1992). When adopting this technique, only events that are at least a distance r from the boundary of W are considered.

2.5.2 The Nearest Neighbour Distribution Function G

The G function is the distribution of the distance from a *typical event* of the process to the nearest other point of the process. For stationary point process X, the $G(r)$ function associated with X is given by

$$G(r) \quad = \quad \mathbb{P}(D(0, X \setminus \{0\}) \le r \,|\, 0 \in X) = \mathbb{P}(X(\mathcal{B}^d(0,r)) > 1 \,|\, 0 \in X) \qquad r \ge 0$$

where $X \setminus \{0\}$ is the process excluding a point at zero. By stationarity the point 0 can be replaced by any arbitrary point $\boldsymbol{x}$. An alternative definition of the $G(r)$ function using the Campbell-Mecke theorem (see section 4.4 of Stoyan et al. 1995) is

$$G(r) = \frac{\mathbb{E}\left[\sum_{\boldsymbol{x} \in X \cap W} \mathbb{I}_{(0,r]}(D(\boldsymbol{x}, X \setminus \{\boldsymbol{x}\}))\right]}{\mathbb{E}[X(W)]}$$

For a homogeneous Poisson process with intensity λ the $G(r)$ function is given by

$$G(r) = 1 - \exp\left(-\lambda v(\mathcal{B}^d(0,r))\right) = F(r) = \begin{cases} 1 - \exp\left(-\lambda \pi r^2\right) & d = 2 \\ 1 - \exp\left(-4\lambda \pi r^3 / 3\right) & d = 3 \end{cases}$$

A border-corrected estimate for the G function is given by

$$\hat{G}(r) = \frac{\sum_{\boldsymbol{x} \in X} \mathbb{I}_{(0,r]}(r(\boldsymbol{x})) \mathbb{I}_{W_{\ominus r(\boldsymbol{x})}}(\boldsymbol{x})}{\sum_{\boldsymbol{x} \in X} \mathbb{I}_{W_{\ominus r(\boldsymbol{x})}}(\boldsymbol{x})}$$

where $r(\boldsymbol{x}) = D(\boldsymbol{x}, X \setminus \{\boldsymbol{x}\})$, and $W_{\ominus r}$ is an *erosion* of W, that is, $W_{\ominus r} = \{\boldsymbol{x} \in W: B^d(\boldsymbol{x}, r) \subset W\}$.

2.5.3 The Pair Correlation Function g

The *pair correlation function*, $g(r)$ is the frequency of event pairs within distance r. The *pair correlation function* is widely used in spatial statistics and particularly in astronomy and astrophysics, for example (Kerscher 1998). Provided that the second

order product density exists, then in the stationary and isotropic case we can write the correlation function $\rho^{(2)}(\boldsymbol{x},\boldsymbol{x}') \equiv \rho^{(2)}(r)$ for $r = \| \boldsymbol{x} - \boldsymbol{x}' \|$. The *pair correlation function* is defined for a stationary point process with intensity λ by

$$g(r) = \frac{\rho^{(2)}(r)}{\lambda^2} \tag{2.3}$$

For a Poisson process, we have $g(r) = 1$. Furthermore, $g(r) > 1$ indicates clustering while $g(r) < 1$ is a sign of regularity. The pair correlation function can be estimated using estimator $\hat{\rho}^{(2)}(r)$ for the second order product density.

2.5.4 The *K* Function

The *K* function appears at present to be the most popular second order characteristic used in point process analysis. For a stationary point process with intensity λ, $\lambda K(r)$ is the mean number of events that are within distance r of the typical event,

$$K(r) = \frac{\mathbb{E}[X(\mathcal{B}^d(0,r))]}{\lambda}. \tag{2.4}$$

The Campbell-Mecke theorem yields the alternative definition

$$K(r) = \frac{\mathbb{E}\left[\sum_{\boldsymbol{x} \in X \cap B} X(\mathcal{B}^d(\boldsymbol{x},r) \setminus \{\boldsymbol{x}\})\right]}{\mathbb{E}[X(B)]} \tag{2.5}$$

for arbitrary B with $0 < \nu(B) < \infty$, where $X(\mathcal{B}^d(\boldsymbol{x},r) \setminus \{\boldsymbol{x}\})$ is the count of the number of points in the ball radius r centered at $\boldsymbol{x}$, excluding $\boldsymbol{x}$. For a homogeneous Poisson process with intensity λ, $K(r)$ is given by

$$K(r) = \nu(B^d(0,r)) = \begin{cases} \pi r^2 & d = 2 \\ 4\pi r^3 / 3 & d = 3 \end{cases}$$

A border-corrected estimate of $K(r)$ for region W is

$$\hat{K}(r) = \frac{\nu(W_{\ominus r})}{X(W_{\ominus r})^2} \sum_{\boldsymbol{x} \in W_{\ominus r}} \sum_{\boldsymbol{x}' \in W} \mathbb{I}_{(0,r]}\left(\|\boldsymbol{x} - \boldsymbol{x}'\|\right).$$

2.5.5 Relationships Between Spatial Point Process Functions

The *K* and *g* functions are closely related, as *K* can be expressed in terms of *g* by the equation

$$K(r) = c(d) \int_0^r u^{d-1} g(u) du \tag{2.6}$$

for some specified constant $c(d)$. Some other characteristics have been defined as combinations and variants of those discussed. Of particular importance is the J-function, suggested by Van Lieshout and Baddeley (1999). For a stationary point process the $J(r)$ function is defined as

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

for $F(r) < 1$. The $J(r)$ function is $J(r) = 1$ for a homogeneous Poisson process. However, $J(r) = 1$ does not imply that the point process is a homogeneous Poisson. $J(r)$ can be estimated by using

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

In general, for $r > 0$, $J(r) < 1$ indicates clustering and $J(r) > 1$ is a sign of regularity.

An alternative to the K-function is the L-function, defined as

$$L(r) = \left(\frac{K(r)}{\nu(@^d(0,1))} \right)^{1/d}.$$

which can be estimated using the estimate $\hat{K}(r)$. For a homogeneous Poisson process, $L(r) = r$, so that $L(r) - r = 0$.

The pair correlation and K-functions can be defined for the non-stationary case, see Møller and Waagepetersen (2004). The anisotropic versions of these functions are defined by Stoyan and Stoyan (1994). Baddeley et al. (2000) propose definitions for the non-stationary versions of the F and G function in their concluding discussions on the analysis of inhomogeneous point patterns.

2.6 More Complicated Point Process Models

Earlier we discussed the simplest point process, the homogeneous Poisson process. We can divide the most commonly used and more complicated point process models into three categories, inhomogeneous Poisson models, models for point patterns which exhibit clustering, and models for point patterns which are regular. The exception to this classification are Cox processes, an important class of models that can be used to model both clustering and regularity (see chapter 5 of Møller and Waagepetersen (2004)).

Preliminary analysis on PPDS1 provided some possible evidence for clustering and hence our discussions here are favoured towards models for clustered data. The cluster models we discuss briefly include the Matérn cluster process (see for example Cressie (1991)). This model has, for example, been used for modelling tree roots data (Fleischer et al. 2006). We also consider the Gauss–Poisson process

(Stoyan et al. 1995). Point process models tend to be generalisations of other point process models; the homogeneous Poisson process is a special case of the inhomogeneous one, which can be generalised to a Cox process.

Models can also be formed by the three fundamental operations discussed in section 5.1 of Stoyan et al. (1995). These operations include superposition, thinning and clustering. In a clustering operation the events of a point process are replaced by clusters of points, X_0. The clusters ($X_0 s$) themselves are spatial point processes. It is common practice to refer to the events as "parents" and the events of the clusters as "daughters". The two cluster processes we discuss here are members of a group of processes called *Neyman–Scott* processes. Neyman–Scott processes result from homogeneous independent clustering applied to a stationary Poisson process. Some Neyman–Scott process such as the Matérn cluster process are also Cox processes.

2.6.1 Gauss–Poisson Process

A *Gauss–Poisson* process (Newman 1970) is an example of a Poisson cluster process (Stoyan et al. 1995). The parent points have a homogeneous Poisson distribution with intensity λ and the number of daughters of each parent is one, two or three with probability q_0,q_1, and q_2 respectively. If the parent has one daughter then the daughter is placed at the parent location. If the parent has two daughters then one is placed at the parent and the other is placed randomly at distance s from the first daughter. The resulting pattern only includes daughter points (and hence the parent points are deleted). Some further results for Gauss–Poisson processes can be found in Milne and Westcott (1972).

2.6.2 Matérn Cluster Process

Matérn's cluster process consists of parents that come from a homogeneous Poisson point process with intensity λ_p. Each parent has m daughters which are uniformly distributed inside $\mathcal{B}^d(0,R)$ (with the parent point being regarded as the origin). The parameter m comes from a Poisson distribution with intensity λ_m. Implicit expressions for the K and g function for a Matérn cluster process can be found in Stoyan et al. (1995).

The Matérn cluster process and Gauss–Poisson process can be simulated in the compact window W directly via the model definitions. Although care should be taken with regards to edge effects. A simple way to account for edge effects is to simulate the parent points inside the dilated window $W_{\ominus \mathcal{B}^d(0,R)}$ where R is such that for the $P(X_0 \supset \mathcal{B}^d(0,R))$ is very small or zero (Stoyan et al. 1995). Brix and Kendall (2002) discuss the simulation of cluster point processes without edge effects.

2.6.3 Markov Point Processes

Markov or Gibbs point processes have been intensively used in spatial statistics since 1970 (Stoyan and Stoyan 1994). Although they are models for various types of point patterns, they are usually recognised for their ability to provide a more flexible framework for modeling spatial point patterns that exhibit inhibition (compared to a homogeneous Poisson distribution) (Cressie 1991). Markov point processes were first defined by Ripley and Kelly (1977). As redefined by Cressie (1991), a spatial point process on bounded set $V \subset \mathbb{R}^d$ is said to be *Markov of range* ρ if it is a spatial point process that has conditional intensity at $\boldsymbol{x} \in V$ given the realisation of the process in $A \setminus \boldsymbol{x}$ that depends only only on the events in $\mathcal{B}^d(\boldsymbol{x}, \rho) \setminus \boldsymbol{x}$. Each Markov process is characterised by a likelihood ratio $f(.)$ with respect to a unit intensity Poisson process. Furthermore, $f(.)$ is usually defined up to a normalising constant that cannot be evaluated in closed form Diggle (2003). A popular example of a Markov point process is the Strauss process (Strauss 1975). In this case, for a configuration of $n < \infty$ points, we have

$$f(x) = \alpha \beta^n \gamma^{\varphi(R)}, \quad \beta > 0, 0 \le \gamma \le 1, R > 0$$

Where φ(*R*) is the number of distinct pair of events within distance *r*. The *Papangelou conditional intensity* defined by

$$\lambda^*(\boldsymbol{x}, \boldsymbol{x}') = \frac{f(\boldsymbol{x} \cap \boldsymbol{x}')}{f(\boldsymbol{x})}, \quad \boldsymbol{x}' \in V \setminus \boldsymbol{x}$$

where we take $a/0 = 0$ for $a \ge 0$ (Kallenberg 1984) is a fundamental characteristic (Møller and Waagepetersen 2001). If f is hereditary (that is $f(x) > 0 \Rightarrow f(y) > 0$ for $y \subset x$), then there is a one-to-one correspondence between f and λ^*. Distribution characteristics (such as the summary statistics introduced earlier) for Markov models are difficult to calculate (Stoyan and Stoyan 1994). Further theory on Markov point process can be found in Stoyan et al. (1995) while a good exposition on simulating Markov point processes can be found in Møller and Waagepetersen (2001) and Møller and Waagepetersen (2004).

2.6.4 Cox Processes

A Cox process is a natural approach for generalising the definition of Poisson point process (Møller and Waagepetersen 2004). A Cox Process on $V \subset \mathbb{R}^d$ is often referred to as a 'doubly stochastic' Poisson point process as the intensity measure is replaced by a random locally finite measure Z_Λ. More formally, we say that a point process X is a Cox process driven by Z_Λ if $X | Z_\Lambda = \Lambda$ is an inhomogeneous Poisson process with mean measure Λ. Due to their generality and associated

manageable closed form calculations, Cox processes tend to find important applications as stochastic models (Stoyan et al. 1995). Examples of Cox processes include the Matérn cluster process. A particular useful class of Cox processes is the class of *Log Gaussian Cox Processes*. A detailed account of Cox processes can be found in Møller and Waagepetersen (2002).

2.7 Marked Spatial Point Processes

A rigorous definition of a point process can be found in Karr (1991). A *marked spatial point process* is a mathematical model for random or irregularly placed points lying in some two- or three-dimensional region, for which each point realization has an associated *mark*, a random variable representing the magnitude or type of some feature that can be measured at that spatial location. A *multivariate spatial point pattern* is a special case of a marked spatial point pattern, where there is a finite number of marks, each representing an event-type (Cressie 1991). A bivariate spatial point process may be used to model the locations of two different types of subnuclear bodies in the nucleus, while a marked spatial point process may be used (as done in this paper) to model the size of one type of subnuclear body in the nucleus.

Spatial pattern analysis (whether marked or unmarked) often begins with tests to determine whether objects are uniformly placed within a specified region. For PML data this is equivalent to testing whether the PML NBs are randomly placed within the cell nucleus. Several techniques have been adopted for assessing the spatial distribution of nuclear bodies. A popular and relatively fast approach for assessing whether nuclear bodies exhibit spatial positioning preference is known as erosion or "nuclear peeling" (Shiels et al. 2007); this entails some form of radial analysis, in which the nucleus is subdivided into concentric rings or shells from the periphery to the centre. Other techniques for investigating subnuclear body spatial preference have included those adopted by Bolzer et al. (2005). They used the mean of inter-body distances and Kolmogorov–Smirnov tests to assess the spatial distribution of subnuclear bodies.

Tests for uniformity are known as tests for CSR (see Section 2.4.2). Such tests will often provide useful insight into any spatial features (such as clustering) and they often entail computing and interpreting (possibly several) distance-based summary statistics. Estimation of these statistics is generally complicated by edge effects. This issue arises for the PML data because the cell nucleus is assumed to cover a finite bounded region and thus estimation of the statistic (which are potentially defined for unbounded regions) can be biased. CSR tests performed on PML data PPDS1, using estimates of the $F(r)$-function (which, informally, is the probability that a PML NB is within distance r of an arbitrary chosen other PML NB), provided some evidence to reject the null hypothesis of CSR (see Umande 2008). Thus there is evidence that PML are not uniformly placed inside the nucleus.

2.8 Bivariate Spatial Point Process Analysis of PML NBs and RNA Polymerase II

One of the most popular summary statistics used for CSR tests in the univariate case is the $K(r)$-function (Diggle 2003). The bivariate version of the $K(r)$-function, $K_{ij}(r)$ of a stationary (invariant under translation) marked point process was first introduced by Hanisch and Stoyan (1979). Heuristically, letting λ_k denote the intensity of events of type X_k, $K_{ij}(r)$ is the expected number of events of type j that are within distance r of an event of type i. Informally, this means, if X_i denotes the location of PML NBs and X_j RNA Polymerase II then λ_i is the average number of PML NBs per unit volume of the cell nucleus and $K_{ij}(r)$ is the average number of PML NBs that are within a distance r of the RNA Polymerase II.

For n_1 the number of type 1 events, and n_2 the number of type 2 events, the $K_{ij}(r)$-function can be estimated inside the bounded window $W \subset \mathbb{R}^d$ (i.e. the interior of the nucleus) using the estimator

$$\widehat{K}_{ij}(r) = \frac{\sum_{\boldsymbol{x} \in X_1} \sum_{\boldsymbol{x}' \in X_2} \omega(\boldsymbol{x}, \boldsymbol{x}')^{-1} \mathbb{I}_{\left(\|\boldsymbol{x}-\boldsymbol{x}'\| \le r\right)}}{\nu(W)\hat{\lambda}_1 \hat{\lambda}_2} \tag{2.7}$$

as suggested by Hanisch and Stoyan (1979). Here $\nu(W)$ is the volume of the nuclear interior. The estimate for the intensity parameter, $\hat{\lambda}_i$ is given by $\hat{\lambda}_i = n_i / \nu(W)$. The function ω is an edge-correction factor such as the proportion of the surface area of the three-dimensional ball centred at $\boldsymbol{x}$, passing through $\boldsymbol{x}'$. For relative ease of calculation and efficiency, we prefer the edge-correction $\omega(\boldsymbol{x}, \boldsymbol{x}') = \nu(W \cap b(\boldsymbol{x}, \| \boldsymbol{x} - \boldsymbol{x}' \|))$. To adopt this preferred form of edge-correction, we can utilise the quadrature approximation

$$\nu\left(W \cap b\left(\boldsymbol{x}, \left\|\boldsymbol{x}-\boldsymbol{x}'\right\|\right)\right) \approx \frac{\nu(W)}{U} \sum_{i=1}^{U} \mathbb{I}_{\left(\boldsymbol{x}_i \in W \cap b\left(\boldsymbol{x}, \left\|\boldsymbol{x}-\boldsymbol{x}'\right\|\right)\right)} \tag{2.8}$$

We estimated the $K_{12}(r)$-function using (2.7) and the preferred form of edge-correction (2.8) for cell 4 of PPDS3. The data used to produced Fig. 2.3 suggest that, heuristically, for this particular cell, we would expect to observe fewer RNA Polymerase II bodies within a distance of 0.5 units of the typical PML, compared to if the RNA Polymerase II exhibited CSR. From the biological literature, it appears that the inhibition (and perhaps more generally, spatial relationship between RNA Polymerase II and PML NBs) is driven by the biological function of PML NBs with different cell nuclei. Also, it may be interesting to compare these results with the findings of Xie and Pombo (2006) who reported that PML bodies contain no detectable RNA polymerase II, but are often surrounded by them at a distance greater than 25 nm. From the detail provided in the Appendix, we estimate that 0.5 units is approximately 41 nm.

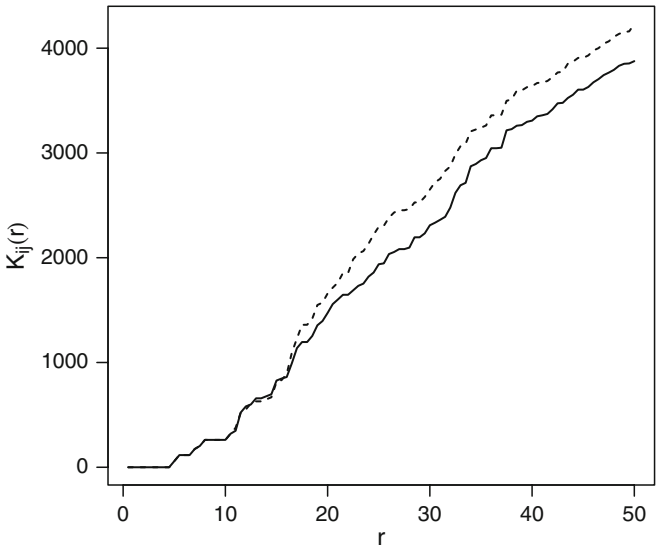

Fig. 2.3 The $K_{12}(r)$-function plot for the PML NB (type 1) and RNA Polymerase II (type 2) in cell nucleus 4 of PPDS3. The solid curve is K_{12} and the dashed curve is K_{21}

Following investigation on edge-correction provided by Umande (2008), we are cautious towards interpreting the results provided in Fig. 2.3 for large values of r. We therefore also consider the outcome of simulation studies. Specifically, a plot of $K_{12}(r)$-functions with simulation envelopes, as shown in Fig. 2.8, can provide further insight. The $K_{12}(r)$-function simulation envelopes for cell nucleus k of PPDS3 was obtained by simulating 100 independent realisations of a homogeneous Poisson process, $\tilde{x}_{1k},\ldots,\tilde{x}_{100k}$, inside the convex hull, representing the nuclear boundary of cell k. Each simulated realisation, $\tilde{x}_{jk}$, of the homogeneous Poisson process had a conditional number of events, n_k, where n_k is the number of RNA Polymerase II found inside cell nucleus k. For each $\tilde{x}_{jk}$ we estimate the $K_{12}(r)$-function without applying an edge-correction, where 1 is an event of type PML and 2 is an event belonging to $\tilde{x}_{jk}$. As we are conducting a like-for-like comparison, in the sense that a biased estimate of the $K(r)$-function is being compared to a another biased estimate of the same function, there is no need for an edge-correction. We hence obtain 100 estimates of the $K_{12}(r)$-function, $\{K_{12_{1k}}(r),\ldots,K_{12_{100k}}(r)\}$ for each cell k.

The upper and lower simulation envelopes for each cell nucleus are respectively $\inf\{K_{12_{1k}}(r),\ldots,K_{12_{100k}}(r)\}$ and $\sup\{K_{12_{1k}}(r),\ldots,K_{12_{100k}}(r)\}$. The results presented in Fig. 2.8 suggest that, apart from cell 4, generally, for a wide range of r, there are fewer RNA Polymerase II bodies within distance r of the typical PML body compared to RNA Polymerase II bodies randomly scattered inside the nucleus. The results for cell 4 are consistent with those obtained for small r (relative to the nuclear magnitude), for the other cells in PPDS3. However, on the contrary to the other cells, for larger r, the PML in cell nucleus 4 typically tend to have a much greater number of RNA Polymerase II bodies within distance r, compared to RNA

Polymerase II randomly placed inside the nucleus. Simulation studies also confirmed that there are enough nuclear bodies in cell 4 of PPDS3 for one to make legitimate observations for features of the spatial point pattern at distances of 0.5 units. This is owed to the high number of RNA Polymerase II bodies (there are 72 RNA Polymerase II bodies in cell 4 of PPDS3).

2.9 A Marked Inhomogeneous Poisson Process Model for PML

Obtaining a model for the spatial distribution of PML NBs is important. Below, we outline how one could fit an inhomogeneous Poisson process to the type of PML data used throughout this paper. By successfully fitting an inhomogeneous Poisson process to replicated PML NB data, suggests a form of a spatial preference for PML within the cell nucleus. Also, very importantly, the formulation of an appropriate model can have potential applications in spotting certain illnesses by comparing the distribution of PML NBs from cells that have been taken from the subject being diagnosed, with the distribution of PML NBs as suggested by the model. At present, this is of course rather ambitious, given the technological limitations.

We have hinted above at a possible candidate model to describe the spatial locations of PML NBs. On rejecting the null hypothesis of CSR, a model that is commonly considered is the inhomogeneous Poisson process. The inhomogeneous Poisson process model is essentially the model that would generate CSR data but with a spatially varying parameter $\lambda(.)$. In terms of the PML NB data, under this model, the number of PML NBs per unit volume is assumed to vary throughout the nuclear interior. The inhomogeneous Poisson process intensity function $\lambda(\boldsymbol{x})$ determines how the PML NBs are distributed throughout the nuclear interior; determining $\lambda(\boldsymbol{x})$ is the core modelling challenge.

Practitioners might consider biological literature when attempting to specify $\lambda(\boldsymbol{x})$. For example, McManus et al. (2006) have reported that chromosomes and regions of chromosomes segregate differently within the nucleus depending on whether or not they are rich in potentially transcribed genes. The individual interphase chromosome territories segregate their gene rich R-bands into the interior of the nucleoplasm, whereas their gene poor G-bands are gathered against the periphery of the nucleus and against the nucleolar surface (see for example Shopland et al. (2003)). Euchromatin sequences are further organised such that they maintain a spatial relationship with the predominant nucleoplasmic nonchromatin structure, the splicing factor compartments (McManus et al. 2006). Smaller nonchromatin structures such as PML associate with specific regions of the genome. In summary, this means there is biological reason to suggest that the spatial location of PML NBs is related to the nuclear boundary. Umande (2008) has used simulation studies and a variant of the empty space function to determine a possible relationship between the placement of PML NBs and the nuclear boundary.

A candidate model that stems from these ideas is one defined through the following postulates:

- MP1 The event (PML NB) locations are a realisation of a homogeneous Poisson process with intensity λ inside bounded $W \subset \mathbb{R}^3$
- MP2 Each event x is retained with probability

$$p(\boldsymbol{x}) = 1 - \exp\left(-\kappa \left\| \boldsymbol{x} - \partial W \right\|\right) \kappa \in \mathbb{R}^+$$

  Otherwise independently thinned (removed) with probability $1-p(\boldsymbol{x})$ where ∂W denotes the boundary of W.

A model defined through postulates MP1-MP2 is an inhomogeneous Poisson process with intensity function $\lambda p(x)$ (see Umande (2008) for a mathematical proof). Furthermore, note that under this model, as $\left|\left| \boldsymbol{x} - \partial W \right|\right| \to 0$, $p(\boldsymbol{x}) \to 0$ which means that PML NBs are less likely to be observed close to the boundary.

We can fit Model 1 to PPDS2 as follows. We first note that for a single replicate, the likelihood, l, of the data $\mathcal{D}$, is given by

$$\ell = p(\mathcal{D}) p(M \mid \mathcal{D}).$$

For k iid replicates the likelihood l_{Rep} is given by

$$\ell_{\text{Rep}} = \prod_{j=1}^{k} p(\mathcal{D}_j) p(\mathcal{M} \mid \mathcal{D}_j)$$

and the log-likelihood is given by

$$\ell = \sum_{j=1}^{k} \log(p(\mathcal{D}_j)) + \log(p(\mathcal{M} \mid \mathcal{D}_j))$$

We can therefore fit the marks separately to the model. However, note that before modelling replicated data that one is uncertain follow the same statistical distribution, it is advisable to begin by testing whether or not the data is "similar" (i.e. whether the data truly does come from the same statistical distribution). Diggle (2003) and Webster et al. (2006) provide details on tests that can be used for testing spatial point pattern similarity. The procedures are not straightforward when applied to data analysed here; bootstrapping techniques and a non-stationary version of the $K(r)$-function are used.

We will now provide an exposition on how we can mark the PML NBs and gain initial insight into the mark distribution by analysing an appropriate spatial point process characteristic and can thus use the marks analysis to completely specify a marked point process model for the PML NB spatial locations. As mentioned above, the extension of other popular characteristics to the multivariate case is generally not difficult (for discrete marks). For the general marked case, the empty space function, $F(r)$ of the marked spatial point process $X^{[m]}$ is the cumulative distribution function of the distance from a randomly selected origin to the nearest event in $X^{[m]}$. That is

$$F(r) = P\left(X^{[m]} \cap (b(0,r) \times \mathbb{M}) \neq \emptyset\right)$$

Also, let B be a subset of $\mathbb{M}$ with $Z_{X^{[m]}}(B) > 0$. We define the nearest neighbour function for events with marks in B by

$$G_B(r) = P^!_{X^{[m]},0}\left(X^{[m]} \cap (b(0,r) \times \mathbb{M}) \neq \emptyset\right)$$

for $r \geq 0$. Here $P^!$ denotes a probability with respect to the Palm distribution. Van Lieshout (2004) introduced a J-function for marked spatial point patterns. The J-function with respect to mark set B, J_B is given by

$$J_B(t) = \frac{1 - G_B(t)}{1 - F(t)}$$

for all $t \geq 0$ and $F(t) < 1$. For an independently marked Poisson process, $G_B(t) = F(t)$ for all t and so $J_B \equiv 1$. Values greater than 1 are a sign of inhibition, while values less than 1 are a sign of clustering.

Van Lieshout (2004) proved, for X a stationary point process on $\mathbb{R}^d$ with intensity $0 < \lambda < \infty$, that if X is *randomly labelled* with mark distribution Z on mark space $\mathbb{M}$ and if $X^{[m]}$ is the marked point process obtained, then for all $r \geq 0$ with $F(r) < 1$, the J-function with respect to a mark set $B \subset \mathbb{M}$ with $Z_{X^{[m]}}(B) > 0$ is given by

$$J_B(r) = J_X(r)$$

where $J_X(r)$ is the J-function of X and where the marked spatial point process $X^{[m]}$ is said to have the *random labelling* property if the marks of the events are conditionally iid given the event locations.

For each PML NB in PPDS2, we calculated an approximate PML body length from the image data used to produce PPDS2. This was done by measuring the maximum distance between any two points, of the points that have been classified as being a part of that PML NB in the image processing stage. That is, in the data provided by the Imperial College London centre for structural biology, each PML NB j is described as a set of points $\{x_{1j}, \ldots . x_{uj}\}$ (see Appendix). The length of PML NB j was calculated as $\inf\left\{\left\|x_{ij} - x_{sj}\right\| : i, s = 1, \ldots, u\right\}$. We use these lengths to assign marks to the PML NBs in PPDS2.

The J-function plots for the cell nuclei of PPDS2 is shown in Fig. 2.4. Figures 2.4 and 2.5 suggests that, since the marked and unmarked $J(r)$-functions are not too dissimilar, we would generally not necessarily expect to observe the PML NBs placed in the nuclear interior, in such a way that depends on their relative sizes (in terms of length). Note also that we found that the proportions of the PML body length to nuclear length, denoted by z_π was consistent with the theoretical proportions provided in the biological literature. All of the PML NBs in PPDS2 were pooled and we calculated the linear correlation between z_π and the proportion of PML NB distance to the boundary to nuclear length. We obtained a correlation of 0.09, suggesting that the two are not strongly linearly correlated. The lack of correlation between the length of the PML and distance to boundary, provides some evidence for random labelling with respect to PML size. This is consistent with the results obtained using the marked J-function. Hence, these results would not support for example, a view that larger PML NBs are found closer to the nuclear periphery or more internally.

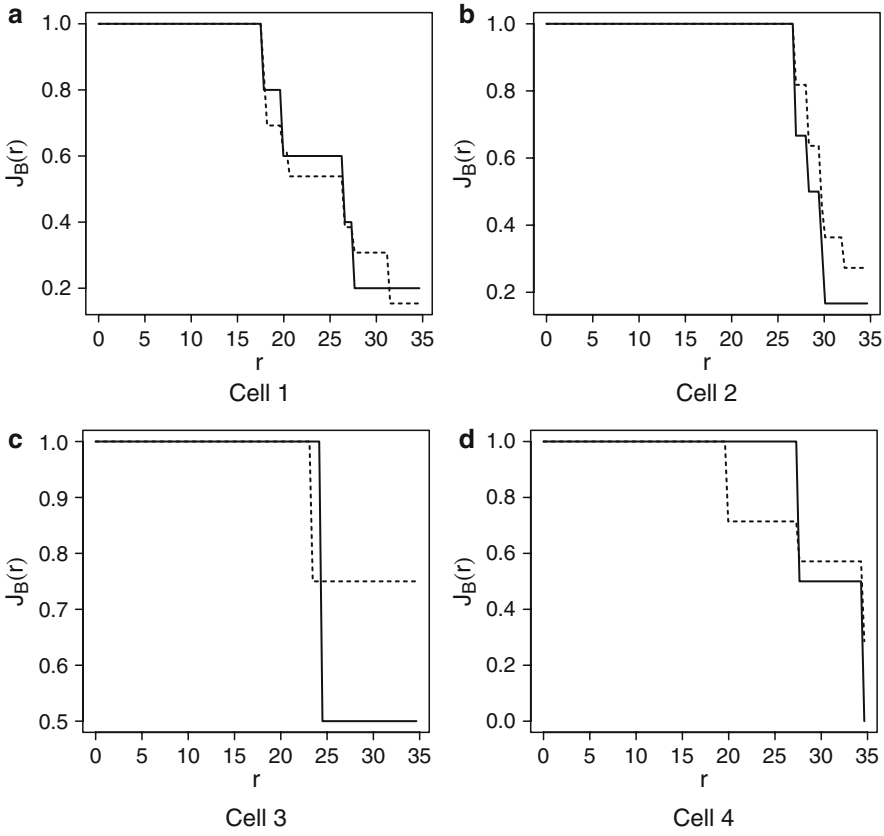

Fig. 2.4 Marked (*solid curve*) and unmarked (*dashed line*) $J(r)$-function for PPDS2 cells 1–4. The mark set $B = [0.01,0.04]$

The final step in the model fitting discussions endeavours to identify each PML NB uniquely by assigning a mark to the PML NB. Each PML NB is now assigned a length. Consider the additional model postulate:

- MP3 Each PML NB $\boldsymbol{x}$ is randomly assigned a proportional length $z_\pi \sim Z$. That is, Z is a random variable that assigns to each PML NB, the mark

$$z_\pi = \frac{\text{PML NB length}}{\text{nuclear length}}.$$

Formal tests on the data (see Fig. 2.6) suggest that a normal distribution is a plausible model for the marks distribution Z.

A Kolmogorov–Smirnov test for the null hypothesis that the marks follow a normal distribution with mean 0.045 and standard deviation 0.019 provided a p-value of 0.92. Caution is required when choosing the mark space since physical restrictions mean that, realistically, the mark space (that the z_π belong to) is $A \subset (0,1)$

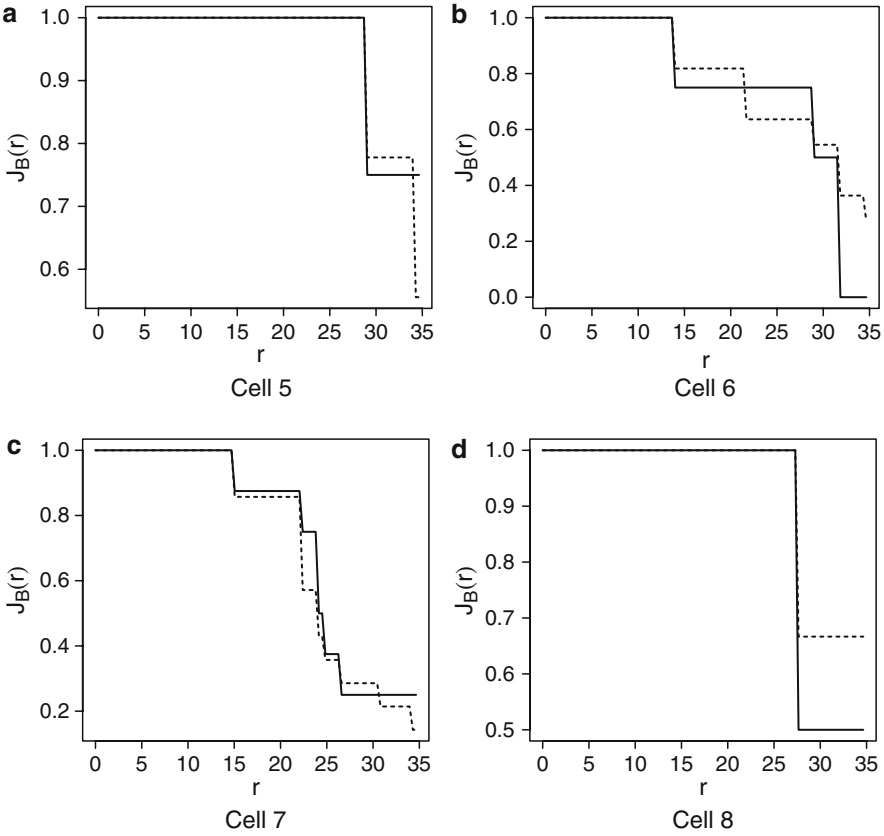

Fig. 2.5 Marked (*solid curve*) and unmarked (*dashed line*) $J(r)$-function for PPDS2 cells 5-8. The mark set $B = [0.01, 0.04]$

(and not for example $\mathbb{R}^+$ as implied by a normal distribution). This is because the PML NBs cannot be longer than the nucleus or have zero length. Hence, it is more appropriate to adopt a truncated normal distribution for Z. More precisely, Z has a normal distribution and lies within the interval (0,1). We estimated the mean and variance of the truncated (0,1) normal distribution, for the PML NB marks, to being (respectively) 0.045 and 0.019 (see for example Barr and Sherrill (1999) for detail on the parameter estimation). The diagnostic plots presented in Fig. 2.7 suggest that the truncated normal model that has been put forward for the PML NB marks distribution is a plausible one.

By using this model for the marks distribution as Z in MP3, and by letting MP3 be an additional final postulate of the Model defined by MP1-MP2, we obtain a marked spatial point process model for the spatial distribution of PML NBs. We may also wish to assess how well the inhomogeneous Poisson process model fits the data. Umande (2008) has carried out such tests on data similar to that used in

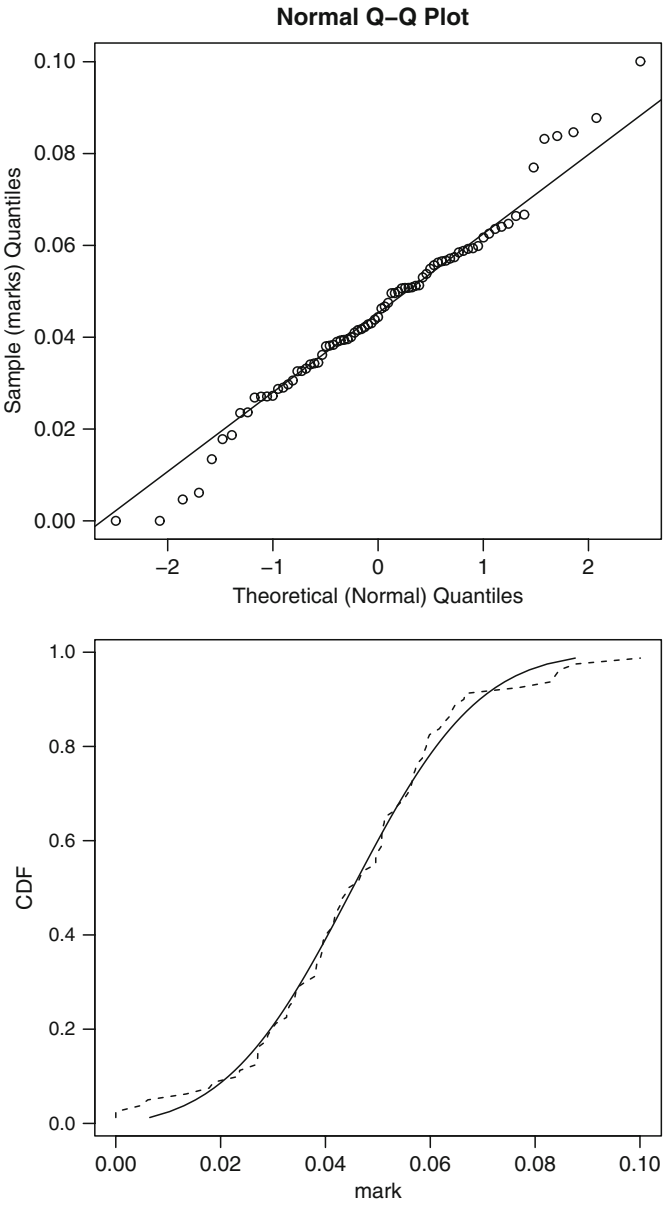

Fig. 2.6 Q–Q plot of the PML NB marks (*top*). The points are approximately linear, suggesting possible normality. The graph on the bottom shows the empirical CDF of the PML NB marks (*dashed curve*) with the CDF of a normal distribution with mean 0.045 and standard deviation 0.019 (*black line*)

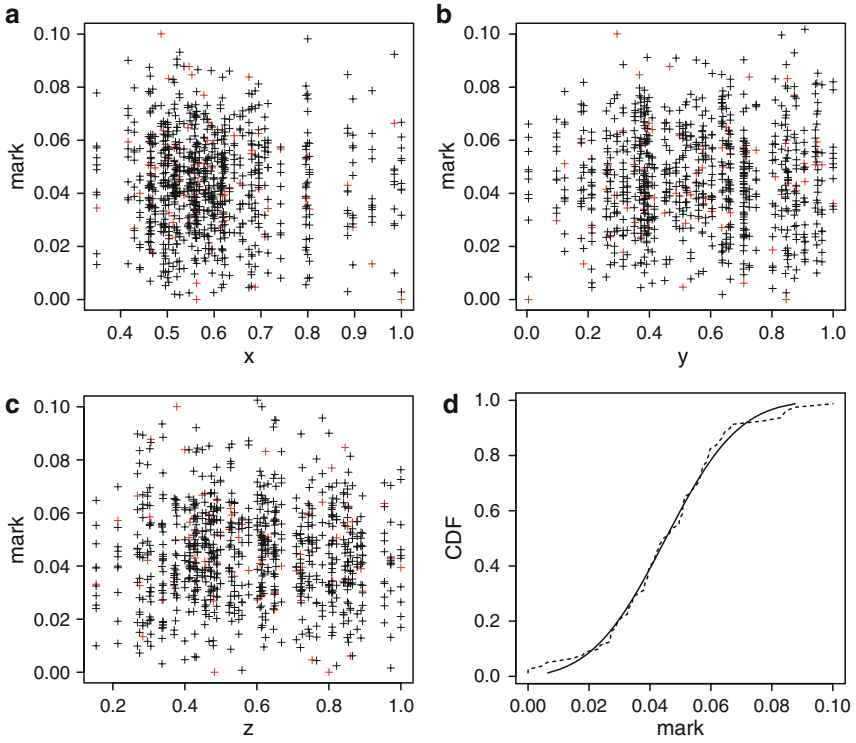

Fig. 2.7 Panel (**d**) shows the empirical CDF of the PML NB marks (*dashed curve*) with CDF of a truncated normal distribution with mean 0.045 and standard deviation 0.019 (*black line*). The scatter plots are of PML NB centroid (x, y, z) coordinates (divided by nuclear length) against simulated realisations of a truncated (0,1) normal distribution with mean 0.045 and standard deviation 0.019 (the model marks distribution). The *red crosses* represent the data and the *black crosses* are for the simulated marks

this paper and found a model defined through MP1-MP2 as a credible model for PML NB locations.

2.10 Conclusion

Tools from spatial point pattern analysis can be invaluable in the investigation of the configuration of nuclear bodies, in particular, the way that PML bodies are distributed across the nucleus in relation to themselves and to other nuclear bodies. By computing inter-object distances and the corresponding K and J functions, simulation-based statistical tests of hypotheses can be formulated and implemented, and these tests allow the validity of important biological models to be assessed.

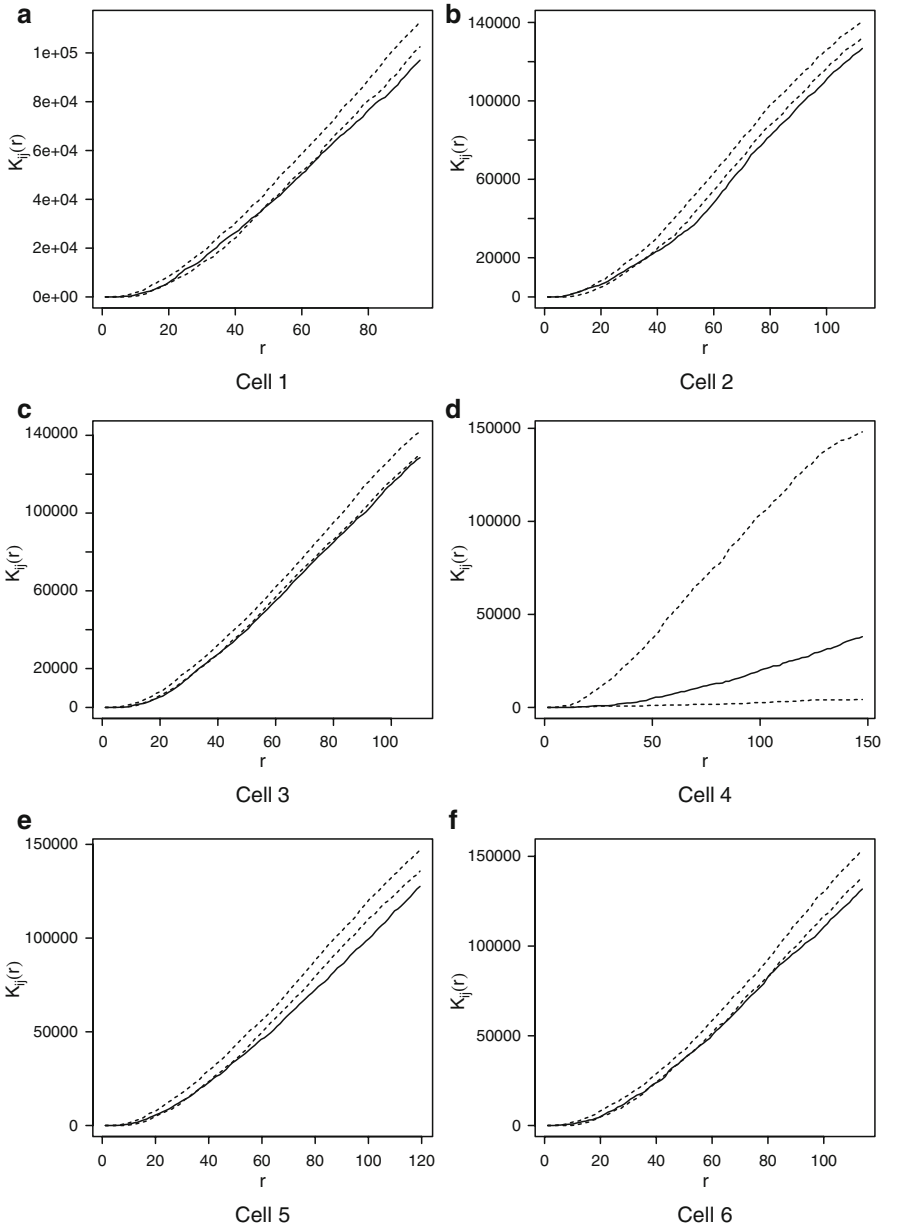

Fig. 2.8 The $K_{12}(r)$-function plot for the PML NB (type 1) and RNA Polymerase II (type 2) in cell nuclei 1–6

Acknowledgements The authors would like to thank Dr. Niall Adams and Professor Paul Freemont for their substantial involvement in this work, and Dr. Elizabeth Batty and Dr Carol Shiels for their experimental work. David Stephens is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant.

Appendix

All of the datsets refered to in this paper (PPDS1, PPDS2 and PPDS3) were provided by the Imperial College London Centre for Structural Biology. The cells used are MRC-5 cell nuclei. Furthermore, the cells used for PPDS2 are all in the G_0 phase of the cell cycle.

- **Distance Units for the Data** Throughout this paper we refer to "units" for reporting measured distances in the cell nuclei. *r* units corresponds to *r* image pixels. There are 12 pixels in 1 micrometer (μm). Hence 1 unit$\approx 0.083\,\mu$m or 83.3 nanometers (nm) .
- **PPDS1** The dataset PPDS1 consists of five cells. We are provided with point coordinates that are the centroids of PML NBs inside the cell nuclei. The sampling region is obtained by calculating the smallest ellipsoid that contains all of the PML points.
- **PPDS2 and PPDS3** Once the confocal image is produced, PPDS2 is obtained by analysing the image data, to form a datset consisting of points that are labeled PML, nucleoli, nuclear boundary, or empty space inside the nucleus. We then run this (large) dataset trough a computer program that forms a point pattern by converting the PML NBs into PML points, by calculating their centroids. PPDS3 is produced in a similar way to PPDS2 but contains an additional labelling to indicate the locations of RNA Polymerase II. Further details, including the number of interior points, *U*, of PPDS2 and PPDS3 are shown in Tables 2.1 and 2.2.

Table 2.1 PPDS2 details

Cell	PML Count	U
1	12	49,942
2	11	50,189
3	9	49,975
4	7	49,969
5	9	49,900
6	11	49,974
7	14	50,091
8	6	50,078

Table 2.2 PPDS3 details

Cell	PML Count	RNA Pol II Count	U
1	6	349	50,216
2	8	269	50,253
3	13	296	50,266
4	12	72	50,266
5	10	226	50,250
6	14	125	49,982

References

Baddeley AJ, Moyeed RA, Howard CV, Boyde A (1992) Analysis of a three-dimensional point pattern with replication. Appl Statist 42(4):641–668

Baddeley AJ, Kerscher M, Schladitz K, Scott BT (2000) Estimating the J function without edge correction. Stat Neerl 54:315–328

Barr D, Sherrill E (1999) Mean and variance of truncated normal distributions. Am Statist 53(4):357–361

Bolzer A, Kreth G, Solovei I, Koehie D, Fauth C, Muller S, Eils R, Cremer C, Speicher MR, Cremer T (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLOS Biol 3(5):826–842

Borden KLB (2002) Pondering the promyelocytic leukemia protein (PML) puzzle: possible functions for PML nuclear bodies. Mol Cell Biol 22(15):5259–5269

Brix A, Kendall WS (2002) Simulation of cluster point processes without edge effects. Adv Appl Probabil 34(2):267–280

Cox DR, Isham V (1980) Point processes. Chapman and Hall, New York

Cressie N (1991) Statistics for spatial data. Wiley, New York

Dellaire G, Bazett-Jones DP (2004) PML nuclear bodies: dynamic sensors of dna damage and cellular stress. BioEssays 26(9):963–977

Diggle PJ (2003) Statistical analysis of spatial point patterns. Arnold, London

Fishman GS (1996) Monte Carlo. Springer, New York

Fleischer F, Beil M, Kazda M, Schmidt V (2006) Analysis of spatial point patterns in microscopic and macroscopic biological image data. In: Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (eds) Case studies in spatial point process modeling, Lecture Notes in Statistics. Springer, Berlin, pp 232–259

Glasbey CA, Roberts IM (1997) Statistical analysis of the distribution of gold particles over antigen sites after immunogold labelling. J Microsc 186(3):258–262

Hanisch KH, Stoyan D (1979) Formulas for the second-order analysis of marked point processes. Statist J Theoret Appl Statist 10(4):555–560

Kallenberg O (1984) An informal guide to the theory of conditioning in point processes. Int Statist Rev 52(2):151–164

Karr AF (1991) Point processes and their statistical inference. Marcel Dekker, New York

Kemp CD (1988) Review: [untitled]. Statistician 37(1):84–85

Kerscher M (1998) Regularity in the distribution of superclusters? Astron Astrophys 336:29–34

Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nature 8:104–115

Lewis PAW, Shedler GS (1976) Simulation of nonhomogeneous Poisson processes with log linear rate function. Biometrika 63(3):501–505

McManus KJ, Stephens DA, Adams NM, Islam SA, Freemont PS, Hendzel MJ (2006) The transcriptional regulator CBP has defined spatial association within interphase nuclei. PLOS Computat Biol 2(10):1271–1283

Milne RK, Westcott M (1972) Further results for Gauss-Poisson processes. Adv Appl Probabil 4(1):151–176

Møller M, Waagepetersen RP (2001) Simulation based inference for spatial point processes, URL citeseer.ist.psu.edu/469797.html

Møller J, Waagepetersen RP (2002) Statistical inference for Cox processes. In: Lawson AB, Denison DGT (eds) Spatial cluster modelling. Chapman and Hall, Boca Raton, FL, p 37

Møller J, Waagepetersen RP (2004) Statistical inference and simulation for spatial point processes. Chapman & Hall, Boca Raton

Newman DS (1970) A new family of point processes which are characterised by their second moment properties. J Appl Probabil 7(2):338–358

Ripley BD (1987) Stochastic simulation. Wiley, New York

Ripley BD, Kelly FP (1977) Markov point processes. Lond Math Soc 15:188–192

Shiels C, Adams NM, Islam SA, Stephens DA, Freemont PS (2007) Quantitative analysis of cell nucleus organisation. Computat Biol 3:1161–1168

Shopland LS, Johnson CV, Byron M, McNeil J, Lawrence JB (2003) Clustering of multiple specific genes and gene-rich r-bands around sc-35 domains: evidence for local euchromatic neighbourhoods. J Cell Biol 162:981–990

Skellam JG (1952) Studies in statistical ecology: I. Spatial pattern. Biometrika 39(3/4):346–362

Stoyan D (2006) Fundamentals of point process statistics. In: Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (eds) Case studies in spatial point pattern modelling, no. 185 in Lecture Notes in Statistics. Springer, New York, p 3

Stoyan D, Stoyan H (1994) Fractals, random shapes and point fields. John, Chichester

Stoyan D, Kendall W, Mecke J (1995) Stochastic geometry and its applications, 2nd edn. Wiley, Chichester

Strauss DJ (1975) A model for clustering. Biometrika 62(2):467–475

Thompson HR (1955) Spatial point processes, with application to ecology. Biometrika 42(1/2):102–115

Umande P (2008) Spatial point pattern analysis with application to confocal microscopy data. Ph.D. thesis, Imperial College London

Van Lieshout MNM (2004) A J-function for marked point patterns. Inst Statist Math 511–532

Van Lieshout MNM, Baddeley AJ (1999) Indeces of dependence between types in multivariate point patterns. Scand J Statist 26:511–532

Wang J, Shiels C, Sasieni P, Wu PJ, Islam SA, Freemont PS, Sheer D (2004) Promyelocytic leukemia nuclear bodies associate with transcriptionally active genomic regions. J Cell Biol 164(4):515–526

Webster S, Diggle PJ, Clough HE, Green RB, French NP (2006) Strain-typing transmissible spongiform encephalopathies using replicated spatial data. In: Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (eds) Case studies in spatial point pattern modelling, no. 185 in Lecture Notes in Statistics.Springer, New York, p 197

Xie SQ, Pombo A (2006) Distribution of different phosphorylated forms of rna polymerase ii in relation to cajal and PML bodies in human cells: an ultrastructural study. Histochemist Cell Biol 125:21–31