Stoch Environ Res Risk Assess (2017) 31:1455–1469
DOI 10.1007/s00477-016-1264-0

ORIGINAL PAPER

Measuring spatial inhomogeneity at different spatial scales using hybrids of Gibbs point process models

Adina Iftimi[1] · **Francisco Montes**[1] · **Jorge Mateu**[2] · **Carlos Ayyad**[2]

Published online: 17 May 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Infectious diseases give rise to complex spatial patterns exhibiting aggregation at different scales. Baddeley (J Stat Softw 55:1–43, 2013) proposed a technique for constructing new Gibbs models for spatial point patterns, combining existing models available in the literature. We use their proposal to model the spatial point pattern of varicella, a highly contagious airborne disease, in Valencia, Spain. We employed descriptive analysis to get a glimpse of the basic properties of the point pattern. Covariate information such as the density of population (children under 14 years old) living in the study region, the distance to the nearest school, and the composition of families (expressed as the average number of persons per family) is used to describe the intensity of the process. We used SatScan to identify main clusters of schools, and to feed the model with this further information. Our analysis shows the relation between varicella cases and school locations, and highlights aggregation in the data at different spatial scales.

Keywords Varicella epidemiology · Spatial statistics · Multi-scale hybrid Gibbs processes

✉ Adina Iftimi
iftimi@uv.es

[1] Department of Statistics and Operations Research, University of Valencia, C/ Doctor Moliner, 50, 46100 Burjassot-Valencia, Spain

[2] University Jaume I, Castellon, Spain

1 Introduction

Formal analysis of a spatial point pattern requires the use of multiple statistical techniques. First and second order summary functions are practical and useful tools to effectively describe and analyse the spatial structure. Diggle (2014, Chap. 7) illustrates model-fitting using summary descriptions for several datasets.

One important class of models for spatial point pattern analysis is the class of (finite) Gibbs point processes. Many theoretical applications of these models can be found in the literature (Geyer 1999; Baddeley and Turner 2000; Møller and Waagepetersen 2004; Diggle 2014). Moreover, Renshaw (2009), Comas et al. (2009) and Comas and Mateu (2011) present different applications in forestry. Funwi-Gabga and Mateu (2012) use the area-interaction model to analyse the behaviour of the *Gorilla horilla diehli* in the Kagwene Sanctuary in Cameroon. Uria et al. (2013) use area-interaction and shot-noise Cox processes to model the distribution of three *Carex remota* cohorts in the north of Spain. Model-fitting, prediction, and simulation of Gibbs models are implemented in the spatstat package (Baddeley and Turner 2005) of R (R Core Team 2014). The most common use of Gibbs models is for *single* spatial scale interaction, which might not be realistic in real-life situations. In practical scenarios, both human interaction and disease spreading exhibit spatial dependence at *multiple* scales. As stated before, point process models are generally used when only one type of spatial interaction (including only one scale of interaction) governs the structure of the point pattern. When there are indications that the spatial structure varies with ranges of distance, a global model is not suitable for describing the complex pattern of such interactions. Baddeley et al. (2013) propose local behaviour models called *hybrid* models. They analyse human

social interaction, studying the spatial locations of people sitting on the grass in a park on a sunny afternoon. They demonstrate that this pattern clearly shows interaction at different scales. Another example of multi-scale use of Gibbs models is Picard et al. (2009). They propose a marked area-interaction multi-scale model and apply the model to three examples from forestry that present different types of structure at different scales: a pine pattern, a bivariate kimboto pattern, and a marked pattern in Gabon, where the marks are represented by the tree diameter. Motivated by this local scale behaviour, we propose an epidemiological application of hybrid models for a highly infectious disease, *varicella*.

In 1992 a multidisciplinary expert committee studied the threat of infectious diseases and related the emergence and re-emergence of these diseases to several factors, such as human demography, technological progress, economic development and land use, international trade, microbial adaptation, and failure in implementing public health measures (Lederberg et al. 1992). Later on, in relation to life-threatening infectious diseases such as smallpox or HIV, Brachman (2003) discusses the fact that modelling has become an important tool in how resources are distributed for purposes of control and prevention. High spreading risk makes infectious diseases difficult to contain and inhibit, in particular airborne diseases, which are spread via air by coughing, sneezing, or talking. Airborne diseases of concern to emergency responders include *meningitis*, *varicella* (also known as *chicken pox*), *tuberculosis* or *influenza*. Airborne transmission depends on various endemic variables. The efficacy of airborne disease transmission is influenced by environmental factors (climate and geographical location) and/or socioeconomic and living conditions. All these factors influence airborne diseases and give rise to complex multi-scale point patterns. In this paper we analyse the spatial point pattern of varicella, an acute infectious disease caused by *varicella zoster virus (VZV)*.

The spatial point pattern of varicella reflects the complex structure inherited by an infectious disease. This complexity makes it difficult to understand the real causes behind this behaviour. Researchers in epidemiology usually make informal statements about disease spread and its relation to covariate information. We make use of stochastic processes in space, in particular of inhomogeneous spatial point processes, to provide a formal statistical procedure to disentangle the local interactions of such complex structures and highlight the role of available covariate information.

VZV can be associated with different sociodemographic variables. Alp et al. (2005) show that the educational level influences the prevalence of antivaricella antibodies, this prevalence being lower for children that are not attending school, followed by those who have at least attended elementary school. Cooper Robbinsa et al. (2011) provide a review of different studies on school-based vaccination. The effect school holidays have on the transmission of VZV is studied in Jackson et al. (2014). The authors show that reductions in contact between children during the summer break lead to a lower transmission of VZV. Due to the nature of the disease, it is well known that a relation exists between school locations and disease spread. However, to the best of our knowledge, this has not been confirmed by formal procedures. We use techniques of spatial point processes to analyse, check, and underline this relation.

We propose the use of hybrids of Gibbs processes to model interactions at different scales and to provide a statistical model that appropriately fits the data. An important issue discussed in this paper is the significance of school locations and how they contribute to explaining the spatial distribution of the disease.

The plan of this paper is as follows. Section 2 presents tools for descriptive analysis, including complete spatial randomness and descriptive analysis with covariates. The dataset on varicella cases is presented and analysed in Sect. 3. The paper ends with some final conclusions and a discussion on different aspects of the analysis.

2 Methods

2.1 Descriptive analysis

The intensity of a point process is the average density of points per unit area. It can either be constant in each point of the process (homogeneous process), or it can vary from one location to another (inhomogeneous process). In most applications the intensity of a point process is not constant. For a location u we denote by $\lambda(u)$ the intensity function.

2.1.1 Complete spatial randomness and inhomogeneity

Testing for *complete spatial randomness* (henceforth, CSR) is equivalent to examining if a random point pattern is a uniform Poisson point process with constant intensity λ. The basic properties to follow are: (i) the number of points in any region A has a Poisson distribution with mean $\lambda \times |A|$, where $|\cdot|$ is the area of the region; (ii) given n points in a region A, the locations of these point are independent and uniformly distributed; (iii) the counts of two disjoint regions A and B are independent.

CSR is tested to rule out the hypothesis of a completely random point pattern (the 'null model'). If this hypothesis is not discarded then the original point pattern has nothing

'readable' because the points are completely unpredictable and have no type of dependence.

There are various methods of testing for CSR. A classical test is the χ^2 test based on quadrat counts (Cressie and Read 1984); a more powerful one is the Kolmogorov–Smirnov test (Berman 1986), in which the observed and the expected distribution of the values of some function T are compared.

In this paper we test for CSR, using *summary statistics* of the original point pattern. We then compare them with the envelopes obtained by simulating point patterns under the null CSR hypothesis.

For a stationary point process X, *the nearest neighbour function, the G function,* is the cumulative distribution function of the distance $dist(x, X \setminus x)$ from a typical point $x \in X$. For a nonstationary point process X, *the inhomogeneous G function* (Van Lieshout 2010), is the intensity-reweighted equivalent of the nearest-neighbour distance distribution function G for homogeneous point processes. An estimator for the G function using generating functionals is given in Equation 5 in Van Lieshout (2010).

For a nonstationary point process X, *the inhomogeneous K-function*, a generalisation of the Ripley's K-function (Ripley 1976), represents the expected value, given that x is a point of X, of the sum of all terms $1/\lambda(y)$ over all points y in the process separated from x by a distance less than r. An estimator for the inhomogeneous K-function (Baddeley et al. 2000) is

$$\widehat{K}(r) = \frac{1}{|B|} \sum_{x \in X \cap B} \sum_{y \in X \cap B \setminus \{x\}} \frac{1\{y \in b(x,r)\}}{\lambda(x)\lambda(y)},$$

for any Borel set B, with $|B| > 0$, where $\mathbb{1}$ denotes the indicator function, $b(x, r)$ is the disk of centre x and radius r, and λ is the intensity of the process. Note that the estimation may be distorted due to the *edge-effect problem*, where, for a point x close to the boundary, the disk $b(x, r)$ may extend beyond the spatial region W. One way to solve this is to use the border method, which takes the erosion of the spatial window W into account. More details and other methods are available in Chiu et al. (2013). The K-function works for pairwise interactions, while the G function handles interactions of all orders.

Rejecting the CSR hypothesis may indicate an inhomogeneous Poisson process, a cluster process, or even a mixed process. The next natural step is to test for *inhomogeneity*. An inhomogeneous Poisson process is obtained when the constant intensity λ is substituted with a spatially varying function, $\lambda(u)$, called the *intensity function*. For a visual inspection of the spatial point pattern we compute a nonparametric intensity estimation, using kernel smoothing (Diggle 1985). To test for inhomogeneity, we compare summary statistics obtained for the empirical point pattern, with envelopes evaluated under the hypothesis of inhomogeneity.

2.1.2 Descriptive analysis with covariates

From a practical point of view, it is of interest to relate the spatial structure with covariate information. Thus we often need to determine if and how a point pattern spatial distribution is associated with different types of covariates. Many times we want to determine whether the intensity of a point process is higher in areas where a certain feature of the population prevails.

Denote by $Z(u)$ the set of covariates for a spatial location u, and consider the intensity as a function of these covariates, $\lambda(u) = \rho\{Z(u)\}$. A more general formulation could be $\lambda(u) = \rho(u, Z(u))$, that is, the inhomogeneity is partially explained by the covariates. The function ρ describes the dependence between the intensity and the values of the covariates.

A nonparametric estimate for ρ is proposed by Baddeley et al. (2012), providing a smoothing estimate of the intensity as a function of (continuous) spatial covariates. This method estimates $\rho(z)$ by the ratio between the (rescaled) density estimate obtained by smoothing the values of the covariates Z at the data points and the density estimate of the reference distribution of Z. This makes it possible to associate the intensity of the process with each covariate separately.

The advantage of parametric modelling using inhomogeneous Poisson processes is that it provides an easy method for analysing spatial point patterns via an intensity function depending on two or more covariates. This can be done by comparing summary statistics computed for simulations of the corresponding inhomogeneous Poisson processes with the intensity function depending on covariates, with the functions estimated from the initial point pattern.

2.2 Models for point process data

2.2.1 Hybrid models

A spatial point pattern $\mathbf{x}$, a realisation of a point process X, is an unordered set $\mathbf{x} = \{x_1, \ldots, x_n\}, n \geq 0, x_i \in W$ of points x_i in a spatial 'window' $W \subset \mathbb{R}^d, d \geq 1$ (Baddeley et al. 2013). Descriptive analysis is required to learn about the basic properties of a spatial point pattern. With a view to finding further characteristics of the spatial structure and association between points, several models for point patterns can be considered.

An important class of models is the Gibbs models, with many applications available in the literature (Geyer 1999;

Daley and Vere-Jones 2003; Baddeley et al. 2006; Renshaw 2009; Comas and Mateu 2011; Funwi-Gabga and Mateu 2012; Diggle 2014). These models are specified in terms of their *probability density*. The probability density function for a Poisson process with intensity 1 is $f(\mathbf{x}) = 1$. The uniform Poisson process with constant intensity $\lambda > 0$ has probability density $f(\mathbf{x}) = \alpha\lambda^{n(\mathbf{x})}$, where $n(\mathbf{x})$ is the number of points in the configuration $\mathbf{x}$ and α is the normalising constant, $\alpha = e^{(1-\lambda)|W|}$. We denote by h the *unnormalised* probability density.

Defining new functional forms for h is not trivial. Baddeley et al. (2013) propose techniques for defining new forms for h combining probability densities of known Gibbs models. In order to define a new class of models, they specify a different unnormalised probability density by multiplying the unnormalised probability densities $h_1, \ldots, h_n$, $n \geq 2$ of n models. A *hybrid* density $h(\mathbf{x}) = h_1(\mathbf{x}) \cdots h_n(\mathbf{x})$, or equivalently $\log h(\mathbf{x}) = \log h_1(\mathbf{x}) + \cdots + \log h_n(\mathbf{x})$, is obtained. Note that the likelihood of a hybrid is a product of component likelihoods, and the Papangelou conditional intensity of a hybrid is the product of the conditional intensities of the components. However, these product forms do not imply any kind of stochastic independence, since the usual factorisation lemma does not apply.

The unnormalised hybrid density has to satisfy certain properties. The functional form of h has to be *integrable* (its integral is finite); has to be *locally stable* (if there is a finite constant B such that $h(x \cup \{u\}) \leq Bh(x)$, for all $x \in \mathbf{x}$ and $u \in W$); has to be *Ruelle stable* (if there are finite constants A and M such that $h(x) \leq AM^{n(x)}$), meaning that h is dominated by an unnormalised Poisson density; and h has to be *hereditary*, or has to have *hereditary positivity* (if, for any configuration $\mathbf{x}$, $h(\mathbf{x}) > 0$ implies $h(\mathbf{y}) > 0$ for all sub-configurations $\mathbf{y} \subset \mathbf{x}$). For more details, see Ruelle (1969) and Møller and Waagepetersen (2004).

Gibbs models are generally applied to point processes with inhibitory patterns. Exceptions that fit aggregated models and are locally stable are the *Widom-Rowlinson penetrable sphere model* (Widom and Rowlinson 1970), the *area-interaction processes* (Baddeley and Lieshout 1995), or the *Geyer saturation process* (Geyer 1999). In this section we present and discuss the details of the *inhomogeneous Baddeley-Geyer hybrid model* (Baddeley et al. 2013).

We first define the unnormalised probability density for the stationary Geyer saturation process with parameters β, γ, r and s, given by

$$h_G(\mathbf{x}) = \beta^{n(\mathbf{x})} \prod_{i=1}^{n(\mathbf{x})} \gamma^{\min(s, t(x_i, \mathbf{x}\backslash x_i, r))} \tag{1}$$

where β controls the intensity of the process, $n(\mathbf{x})$ is the number of points in the pattern and γ is the interaction parameter. Furthermore, $t(x_i, \mathbf{x}\backslash x_i, r)$ is the number of neighbours of x_i in $\mathbf{x}$ within a radio r, that is, the number of points x_j with $j \neq i$ such that $||x_i - x_j|| \leq r$. The parameter $s > 0$ is a saturation threshold which ensures that each term in the product is bounded by γ^s, so that the density is integrable and Ruelle stable for all values of $\gamma > 0$. The process is clustered if $\gamma > 1$.

Considering the hybrid of several Geyer densities of form (1), the unnormalised density of an inhomogeneous Baddeley-Geyer hybrid model is obtained by

$$h(\mathbf{x}) = \beta^{n(\mathbf{x})} \prod_{i=1}^{n(\mathbf{x})} \prod_{j=1}^{m} \gamma_j^{\min(s_j, t(x_i, \mathbf{x}\backslash x_i; r_j))}$$

where $r_1, r_2, \ldots, r_m$ are interaction ranges, $s_1, s_2, \ldots, s_m$ are saturation parameters, and $\gamma_1, \gamma_2, \ldots, \gamma_m$ are the interaction parameters.

We can introduce the effect of covariates by considering a local covariate effect $\beta(x_i) = \rho\{x_i, Z(x_i)\}$, where $Z(x_i)$ is a set of covariates for a spatial location $x_i \in \mathbf{x}$. The resulting probability density is

$$h(\mathbf{x}) = \prod_{i=1}^{n(\mathbf{x})} \beta(x_i) \prod_{j=1}^{m} \gamma_j^{\min(s_j, t(x_i, \mathbf{x}\backslash x_i, r_j))}. \tag{2}$$

This defines inhomogeneity in an elegant and straightforward way. We use this possibility throughout this paper.

Estimating irregular parameters $r_1, \ldots, r_m$ and $s_1, \ldots, s_m$ in the inhomogeneous Baddeley-Geyer hybrid model raises an important issue on which very little statistical theory is available. Baddeley and Turner (2000) propose using profile pseudolikelihood. In this paper we use their technique to estimate the irregular parameters for the hybrid model.

2.2.2 Diagnostics

The next step is to check if the model fits well and if each assumption of the model is appropriate. For a Poisson model, homogeneous or inhomogeneous, a χ^2 goodness-of-fit test based on quadrat counts (Cressie and Read 1984) or a Berman test (Berman 1986) can be applied. For hybrid models, no theory is available to support these tests. As an alternative, goodness-of-fit for hybrid models relies on the summary statistics functions G and K. Baddeley et al. (2011) propose new tools for model validation. They suggest using the residual G and K functions to compare and decide the best fitted model. First, both the nonparametric estimate of the G function and the one based on the corresponding model are computed. Then the residual G function is obtained as the difference between the two measures. Likewise, we obtain the residual K-function. The residual G and K functions should be approximately zero if the model fits well. These functions provide a suitable diagnostic for the goodness-of-fit of a point process model.

Diagnostic plots based on residuals are another tool to measure the goodness-of-fit of a model and also to identify outliers in the data. These plots display the residuals from the fitted model. This diagnostic is followed up by Q–Q plots based on residuals from the model.

3 Application

3.1 Varicella epidemiology

In this section we describe some characteristics of VZV behaviour. The clinical course of varicella is generally mild in children. Adults may suffer from more severe symptoms and also have a higher risk of complications. Children infected with human immunodeficiency virus also may have severe, prolonged illness (Centers for Disease Control and Prevention 2012). After recovery from varicella patients usually have lifetime immunity. Herpes zoster occurs when latent VZV re-activates and causes recurrent disease.

Varicella occurs worldwide and is a highly contagious human disease, no animal or insect source is known to exist. It is highly communicable and endemic in all countries worldwide. In temperate climates, at least 90 % of the population develops varicella by the age of fifteen and 95 % by the time they reach adulthood. Varicella is characterised by fever and a generalised vesicular rash, consisting of 200 to 500 lesions (European Centre for Disease Prevention and Control 2014). The rash progresses rapidly from macules to papules to vesicular lesions before crusting. The most common way of transmission of VZV is person to person from infected rashes. Transmission also occurs by respiratory contact with airborne droplets or direct contact of aerosols from vesicular fluid of skin lesions of acute varicella.

The period of transmission is from one to two days before the onset of the rash to when the lesions are crusted, usually 4–5 days after the appearance of the rash. The incubation period goes from 10 to 21 days, commonly 14 to 16 days.

Most people with varicella make full recoveries, only 2–6 % of varicella cases develop complications (European Centre for Disease Prevention and Control 2014). The diagnosis of varicella is primarily clinical. Laboratory tests are requested for more complicated cases or for epidemiological purposes. Recent studies have shown that varicella vaccination has influenced the incidence of the disease, decreasing the number, size, and duration of the outbreaks.

3.2 Descriptive analysis for varicella data

The city of Valencia is the third largest city in Spain. Figure 1 shows the location of the province of Valencia within Spain (left of the small subplot), the location of the city of Valencia within the Province of Valencia (right of the small subplot), and the city of Valencia itself (large subplot). The varicella incidence in the city of Valencia (study area) has decreased from 2542 cases registered in 2008 down to a number of 921 cases in 2013. In view of the results in Iftimi et al. (2015a, (b), it seems reasonable to work with children under the age of 14 without taking into account the sex distribution of the cases. An important issue this dataset raises is the presence of multiple points, which are caused due to inherent discrete assignment in observations. The point pattern is constructed by considering the address of residence for each case, in particular

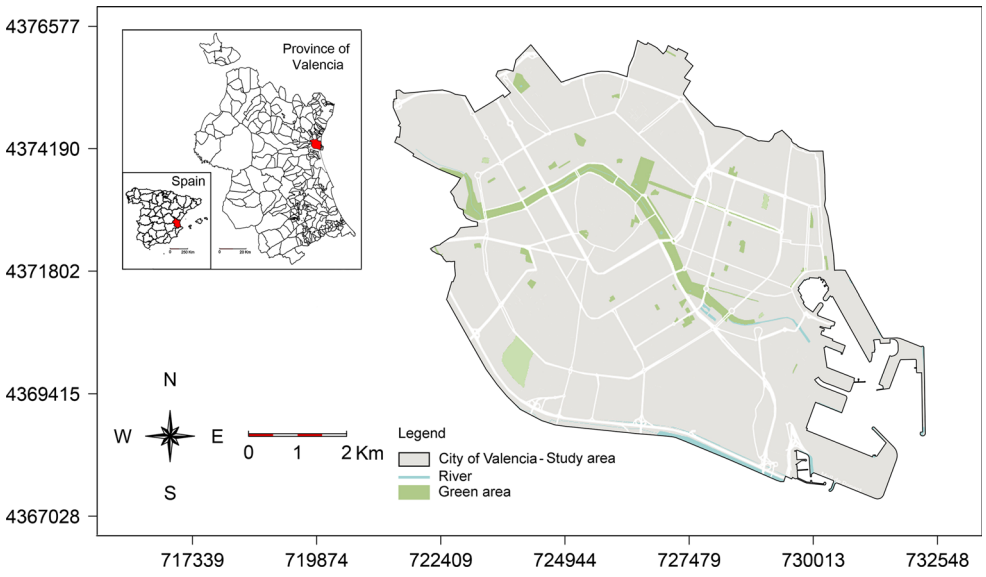


Fig. 1 Location of the city of Valencia (study area). Universal Transverse Mercator (UTM) coordinate system (distance in meters)

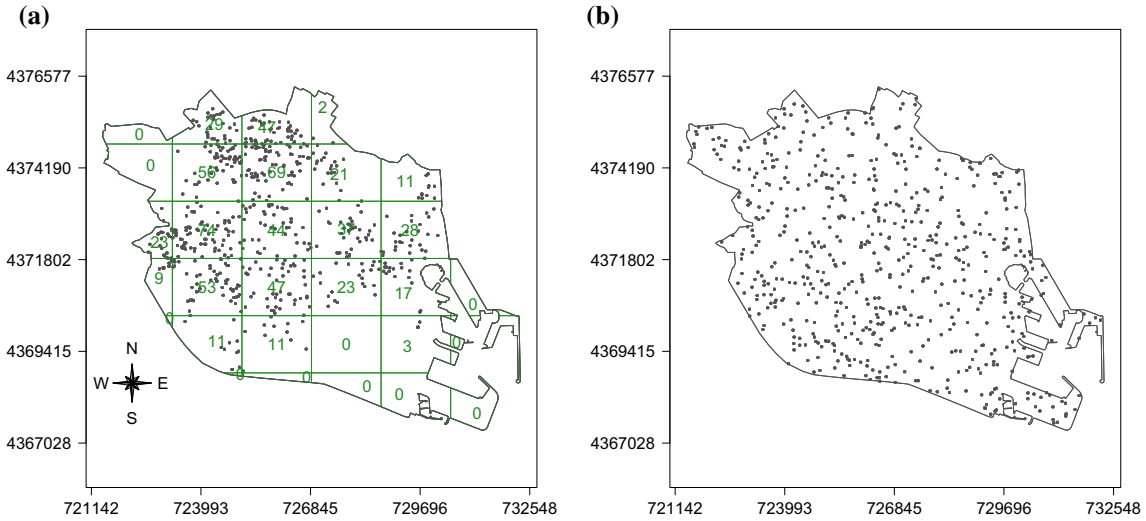


Fig. 2 **a** Spatial point pattern of varicella cases registered during 2013 in Valencia (Spain), together with quadrat plot for the point pattern; **b** Point pattern of independent uniform random points. UTM coordinate system (distance in meters)

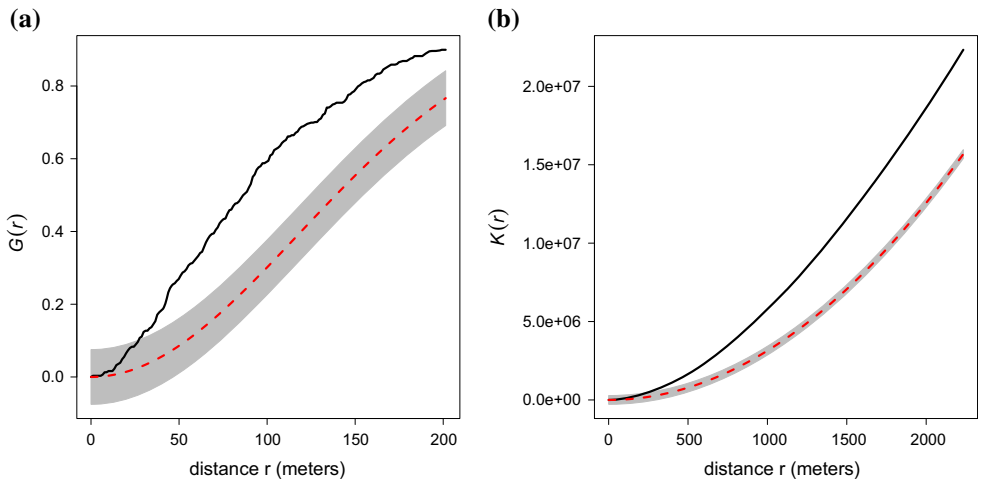


Fig. 3 **a** Empirical *G* function (*solid line*) with pointwise maximum and minimum envelopes obtained under CSR null hypothesis (*grey area*); **b** *K*-function (*solid line*) with pointwise maximum and minimum envelopes obtained under CSR null hypothesis (*grey area*)

the building where the case resides. Therefore, one point (building) can have multiple cases assigned. In this paper we study the ground process, meaning all the locations where the cases were observed, without taking into account the number of cases observed at that specific location. This paper studies the ground process of varicella cases registered in Valencia in 2013, for children under the age of 14.

Figure 2a shows the spatial point pattern of the varicella cases together with the quadrat counting of the points. The pattern of varicella is clearly not randomly scattered, with areas with higher number of cases than the average. Figure 2b shows the same amount of spatial points as the varicella cases uniformly distributed in the region. This figure is a visual confirmation that the varicella pattern is not completely spatially random.

3.2.1 CSR and inhomogeneity

Figure 3a shows the empirical *G* function for the point pattern (solid line) and the pointwise maximum and minimum envelopes obtained from 99 simulations of a process under the CSR null hypothesis. Likewise, Figure 3b shows the estimated *K*-function for the point pattern (solid line) and the pointwise maximum and minimum envelopes obtained by simulating a Poisson process 99 times. The hypothesis of complete spatial randomness is clearly rejected.

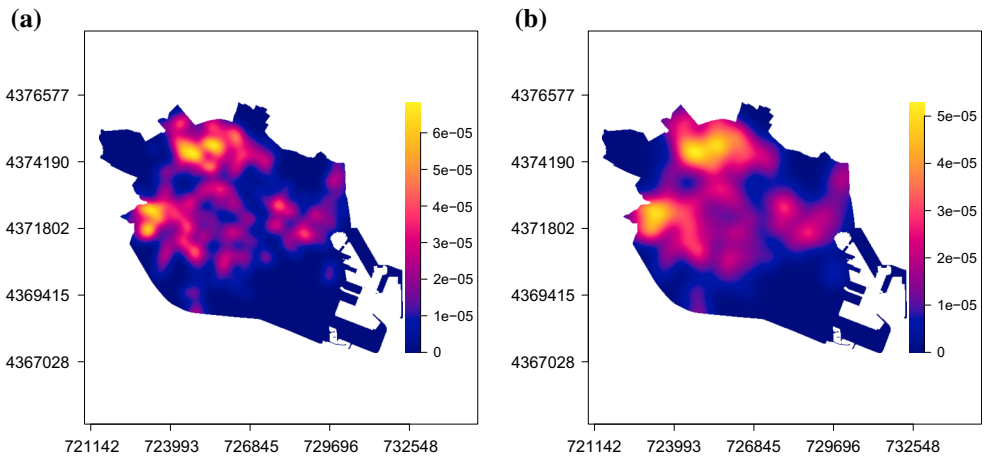


Fig. 4 Kernel smoothing estimate of the varicella point pattern intensity: **a** bandwidth = 153 (Diggle 1985); **b** bandwidth = 300. UTM coordinate system (distance in meters)

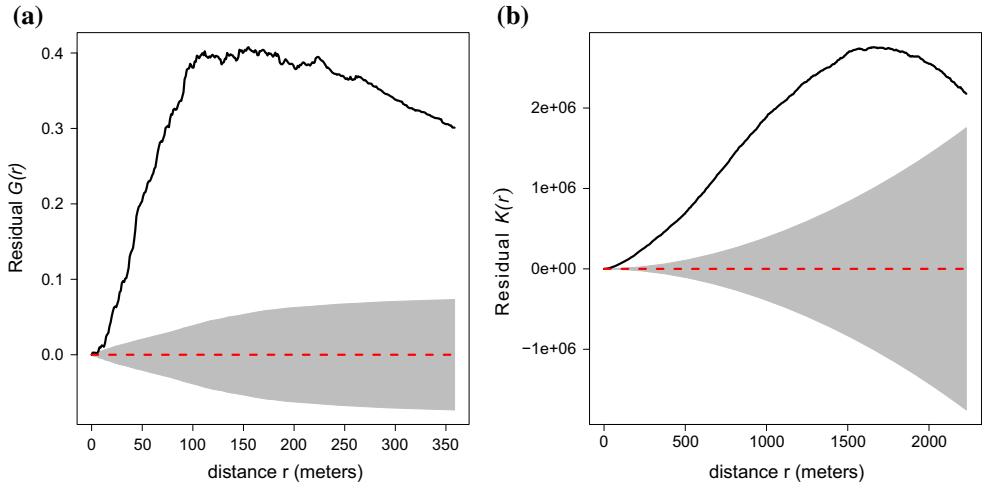


Fig. 5 **a** Residual *G* function (*solid line*); **b** Residual *K*-function (*solid line*); and the corresponding envelopes obtained under the inhomogeneity assumption previously described (*grey area*)

A rejection of the CSR hypothesis is an indicator that our process can be an inhomogeneous Poisson process, a cluster process, or even a mixed process.

The intensity of the spatial point pattern has been estimated nonparametrically using a Gaussian kernel density estimator. This method is sensitive to bandwidth selection. Figure 4a shows the kernel estimate for the bandwidth obtained using a cross-validation method, as proposed by Diggle (1985). Figure 4b shows the intensity estimate for a bandwidth equal to 300. The latter provides a smoother estimate, whereas the former gives a more uneven estimate.

In Fig. 4b we note several areas where varicella incidence is high: one in the north of Valencia, corresponding to the *Rascanya* district, and a second one in the west in the *l'Olivereta* district. The population living in these areas generally tends to have a lower income level in comparison with the central areas of the city.

Figure 5a shows the residual *G* function, defined in Sect. 2.2.2, with the corresponding estimate obtained under the hypothesis of inhomogeneity (the log-intensity depends on a linear combination of the coordinates of the point locations). Likewise, Fig. 5b shows the residual *K* function together with the corresponding envelopes. The irregular shape of the *G* function indicates some type of aggregation in the data.

3.2.2 Introducing covariates into the study

The next step in our analysis is to assess the interaction between the varicella pattern and external factors. Almost

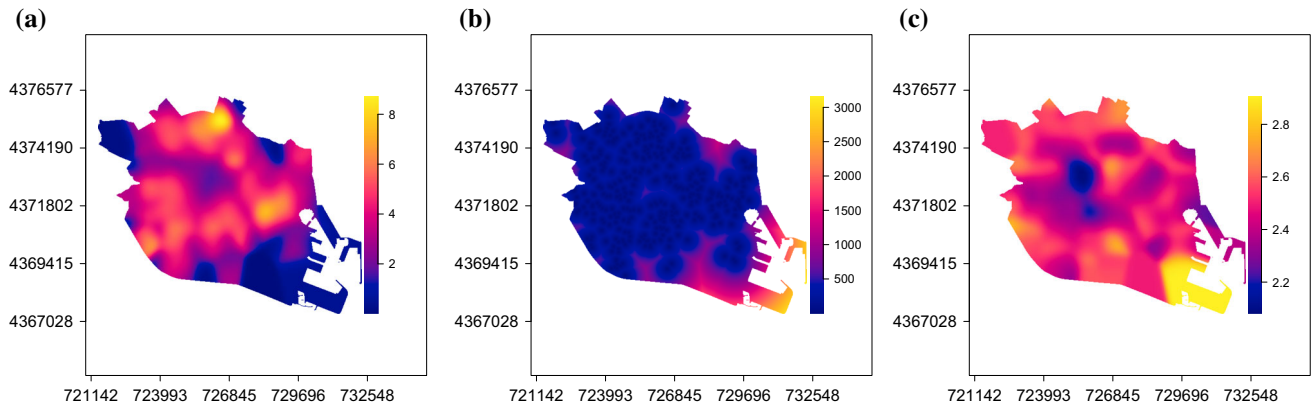
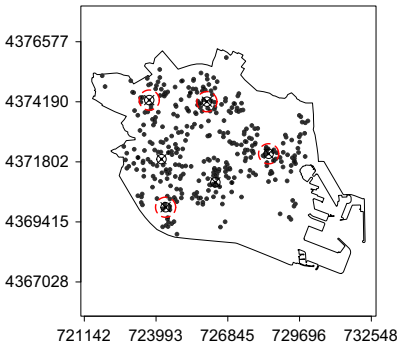


Fig. 6 **a** Kernel estimate of population density under 14 years, 2013; **b** Kernel estimate of the distance function to the nearest school in Valencia; **c** Kernel estimate of the average number of persons per family register. UTM coordinate system (distance in meters)

Fig. 7 Coordinates and corresponding p values for the six main school clusters detected using SatScan software (*left*). Pattern of all schools in Valencia (*black dots*), together with the spatial locations of the main clusters (*dots encircled* and labelled with an X) (*right*). UTM coordinate system (distance in meters)

ID	Northing (m)	Easting (m)	p-value
1	4374194.64	726012.70	2.22E-06
2	4374257.90	723723.00	4.16E-06
3	4372123.33	728467.68	1.30E-02
4	4371909.12	724210.03	2.50E-02
5	4370990.51	726351.50	4.59E-01
6	4369998.62	724372.88	4.59E-01



all types of diseases, and varicella in particular, depend on the population. Whether it is a contagious disease or not, people living in a determined area are the most important factor in initiating and transmitting a disease. Regions with a high population, tend to register more cases, and scarcely populated areas tend to have lower disease incidence. In order to make this correction, we consider the density of population under 14 years as a covariate. Figure 6a shows a kernel estimate of this population density in the city of Valencia for the year 2013. We can see that the spatial distribution of the population shows some sort of a pattern. The centre and the outskirts of the city are areas where the density of population is low, whereas the neighbourhoods surrounding the centre of the city are the most preferred to live in.

The location of the schools in Valencia can play a significant role in the analysis of the disease. Figure 7 shows the pattern of all daycare centres, preschools and schools in Valencia. One of the main objectives of this paper is to find, if present, a relation between the spatial point pattern of varicella and the pattern of schools in Valencia. Varicella is primarily a children's disease, and almost all children either attend daycare centres, preschool or elementary school. Thus, the spatial point pattern in Fig. 7 is considered a major factor in the clustering of varicella cases. We expect schools to behave as *hot spots* for varicella.

We also consider the distance to the nearest school as a covariate. The distance function of a set of locations (in our case the schools), $\mathbf{s} = \{s_1, \ldots, s_m\}, m > 0, s_i \in W$, where W is the study region, is a mathematical function such that, for any spatial location u the function value is the shortest distance from u to $\mathbf{s}$.

$$dist(u) = \min\{eucl.dist(u, s_i)\}, \ i = 1, \ldots, m, \ u \in W. \tag{3}$$

Figure 6b shows a graphical representation of the distance function defined by (3). Areas where the distance function has low values match with neighbourhoods with an elevated number of schools. Similarly, areas with high values of the distance function are regions with fewer or no schools.

Estimating the school distance function can explain the relationship between the disease and the pattern of schools. Alternatively, the schools can be considered *focal points*. The point pattern of all schools in Fig. 7 shows areas where

points tend to cluster. We identify the main clusters, using SatScan software (Kulldorff 2010). A summary of the main identified clusters is presented in the left-hand-side table of Fig. 7. On the right-hand side of Fig. 7 we see the spatial location of all schools as black dots, together with the main clusters detected by the software, encircled and labelled with an X. These six points can be interpreted as focal points (sources) around which varicella cases tend to group. For each source we calculate the distance function in Eq. (3).

Another factor taken into account is the composition of the families. Figure 6c shows a kernel estimate of the average number of persons per family register, meaning the average number of persons a family consists of.

As stated before, the population density in Fig. 6a shows the centre and the peripheral areas of the city having less population than the areas surrounding the central neighbours of Valencia. In the north of Valencia there is also an area where the population density is the highest. Figure 6c shows a central area where the number of persons per family is low. This intensity increases as we move toward the peripheral neighbours of the city.

As explained in Sect. 2.1.2 the intensity can be considered as a function of the covariates, $\lambda(u) = \rho\{Z(u)\}$. Figure 8 shows the nonparametric estimate of $\lambda(u)$ for each individual covariate together with pointwise two-standard-deviation confidence limits (grey shading) (Baddeley et al. 2012). A clear relationship between the intensity of the process and the three covariates can be observed. For areas where the population density is low, the intensity of varicella cases is also low. Areas with high-density population are more likely to have a higher relative risk of varicella. For the distance function defined in Eq. (3), Fig. 8b shows a linear relation. The larger the distance from schools, the lower the incidence of varicella observe. Figure 8c shows high intensity of the relative risk of varicella in areas where small families live. As the number of family members increases, the risk of varicella decreases. This unusual and unexpected behaviour will be further discussed in the last section of the paper.

The nonparametric estimates shown in Fig. 8 could indicate that varicella risk is higher in neighbourhoods where families have a medium-low cultural and socioeconomic status. We fit an inhomogeneous Poisson process where the log-intensity is a linear function of the three covariates shown in Fig. 6, and use the diagnostics described in Sect. 2.2.2 to assess the fitted model. The coordinates of the spatial locations, x and y, were not considered in this model, as they do not improve the overall fitting.

Table 1 shows the estimated parameters for the inhomogeneous Poisson process. All three covariates are significant and, as expected, population density has a positive effect on the overall risk. The sign for the second covariate shows an inverse relation between disease risk and the distance function. The parameter of the last covariate also shows an inverse association between varicella risk and the family register, as shown in Fig. 8c. This would mean that

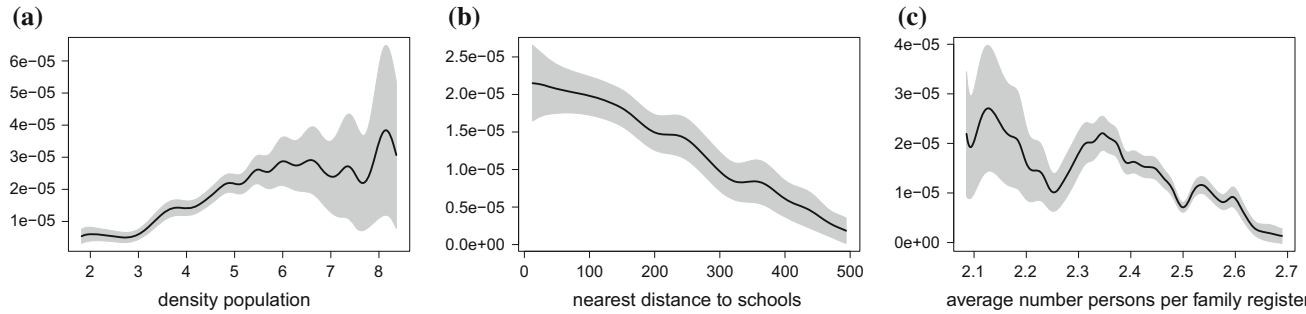


Fig. 8 Non-parametric smoothed estimation of the intensity as a function of spatial covariates: **a** population density under 14 years, 2013; **b** nearest distance to school in Valencia; **c** average number of persons per family register

Table 1 Estimated parameters for the inhomogeneous Poisson model.

	Estimate	S.E.	CI95.lo	CI95.hi	Ztest	Zval
(Intercept)	−4.11E+00	9.25E−01	−5.92E+00	−2.29E+00	***	−4.44E+00
Population density	4.05E−01	3.61E−02	3.34E−01	4.76E−01	***	1.12E+01
Distance function	−2.55E−03	3.68E−04	−3.27E−03	−1.83E−03	***	−6.94E+00
Family register	−3.40E+00	4.04E−01	−4.19E+00	−2.61E+00	***	−8.41E+00

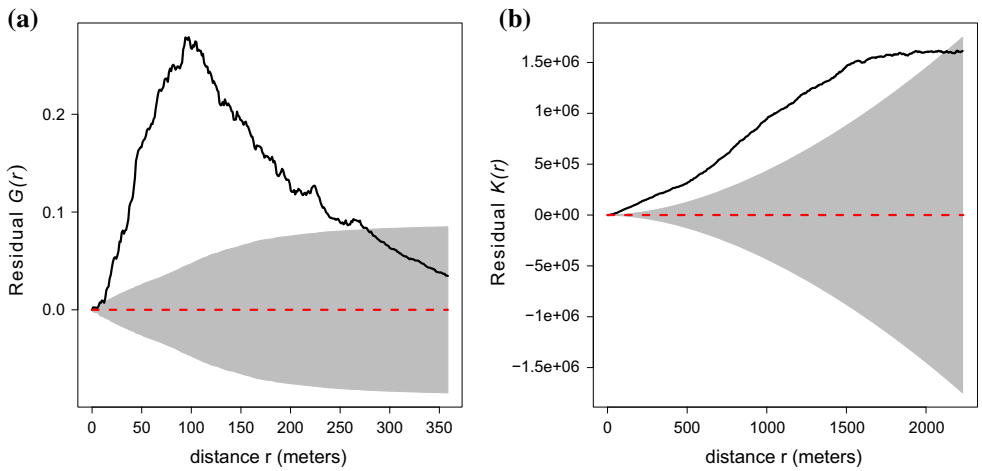


Fig. 9 (a) Residual G function for the inhomogeneous Poisson process; (b) Residual K-function for the inhomogeneous Poisson process

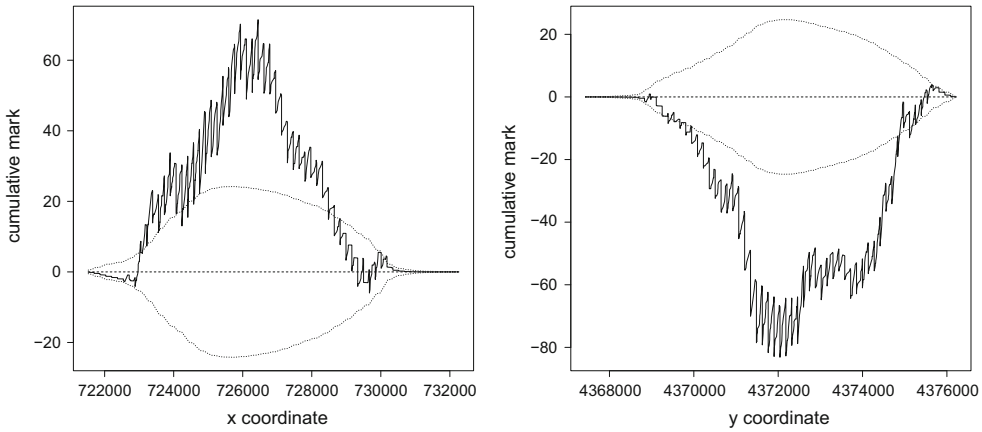


Fig. 10 Cumulative raw residuals for the x and y axis

for regions where large families live, varicella incidence is low. This seems to contrast with the opinion of epidemiologists, and will be further discussed in Sect. 4.

Figure 9 shows the residual G and K functions for the inhomogeneous Poisson process with intensity depending on the covariates. The sharpness and the peak of the residual G function in Fig. 9a evidences 'unexplained' interaction at different spatial scales, between 75 and 125 meters. The residual K-function shows that the model does not fit well.

Figure 10 shows the cumulative raw errors for the inhomogeneous Poisson model. The size of the errors is considerable, and they have a particular shape. A qqplot test confirmed that the errors are not normal.

3.3 Inhomogeneous Baddeley–Geyer hybrid model

To explain the multi-scale spatial interaction, we considered a hybrid Baddeley-Geyer process depending on the three covariates described in Sect. 3.2.2. We then added all possible linear combinations of focal points to this initial model and compared the results, using model diagnostics. The best model for our data includes all three covariates and four of the six focal points described in Sect. 3.2.2. These four focal points are circled in Fig. 7.

We selected the irregular parameters, using maximum profile pseudolikelihood. We searched over a wide range of parameters $1 \leq r \leq 400$ and $s = 1, \ldots, 20$ and identified three significant spatial interaction components. The optimum combination is obtained for distances $r_0 = 47$, $r_1 = 104$, $r_2 = 286$ meters and saturation parameters $s_0 = 1$, $s_1 = 3$, $s_2 = 2$.

To determine the accuracy of the fitted model, we used three complementary methods. First, we looked at the residual G and K functions. Second, we used diagnostic plots to study the size of the errors. Finally, we simulated a point pattern from the fitted model and compared it with the empirical initial pattern.

The interaction parameters for the hybrid model were $\gamma_1 = 1.305$, $\gamma_2 = 1.273$, and $\gamma_3 = 1.643$, which confirm

Table 2 The estimated parameters for the Baddeley-Geyer hybrid model.

	Estimate	S.E.	CI95.lo	CI95.hi	Ztest	Zval
(Intercept)	−7.55E+00	1.28E+00	−1.01E+01	−5.04E+00	***	−5.89E+00
Population density	3.21E−01	4.71E−02	2.28E−01	4.13E−01	***	6.81E+00
Distance function	−1.08E-03	5.05E−04	−2.07E−03	−9.20E−05	*	−2.14E+00
Family register	−2.54E+00	5.85E−01	−3.69E+00	−1.39E+00	***	−4.34E+00
Focal point 1	−8.12E−05	7.28E−05	−2.24E−04	6.14E−05		−1.12E+00
Focal point 2	−4.82E−05	9.54E−05	−2.35E−04	1.39E−04		−5.05E−01
Focal point 3	7.84E−05	9.34E−05	−1.05E−04	2.61E−04		8.40E−01
Focal point 4	5.48E−05	5.74E−05	−5.78E−05	1.67E−04		9.54E−01
β_1	2.66E−01	1.01E−01	6.78E−02	4.65E−01	**	2.63E+00
β_2	2.42E−01	4.16E−02	1.60E−01	3.24E−01	***	5.81E+00
β_3	4.97E−01	2.03E−01	9.89E−02	8.94E−01	*	2.45E+00
γ_1	1.305					
γ_2	1.273					
γ_3	1.164					

that the point pattern shows spatial aggregation at different spatial scales. Table 2 shows the estimates of the parameters for the fitted model. The positive coefficient of the population density shows a positive effect of population density on the incidence of varicella. The negative coefficient of the family register remains an unusual concern.

To assess our fit, we used the residual G and K functions. Figure 11 shows a comparison between these two functions for the inhomogeneous Poisson model and the hybrid model. We see a definite improvement for the latter model. The values of the two functions oscillate around zero and are inside the envelopes, which indicates the model fits well. Using the same scales as in Fig. 10, Fig. 12 shows a diagnostic for the residuals of the fitted model. The range of the raw residuals in Fig. 12 has substantially decreased compared to the residuals of the inhomogeneous Poisson process. The qqplot in Fig. 12 shows that the errors are approximately normal.

Figure 13 shows a comparison between the initial point pattern of varicella and a simulation of the fitted hybrid model. We can see that the simulated pattern presents areas with more points than the initial point pattern, but overall, it does correspond quite well to the data pattern.

4 Conclusions and discussion

We have presented an applied statistical analysis to find specific characteristics and relations between varicella incidence and the school point pattern. We conclude that there is a significant relation between the location of schools and possible varicella outbreaks. The final fitted

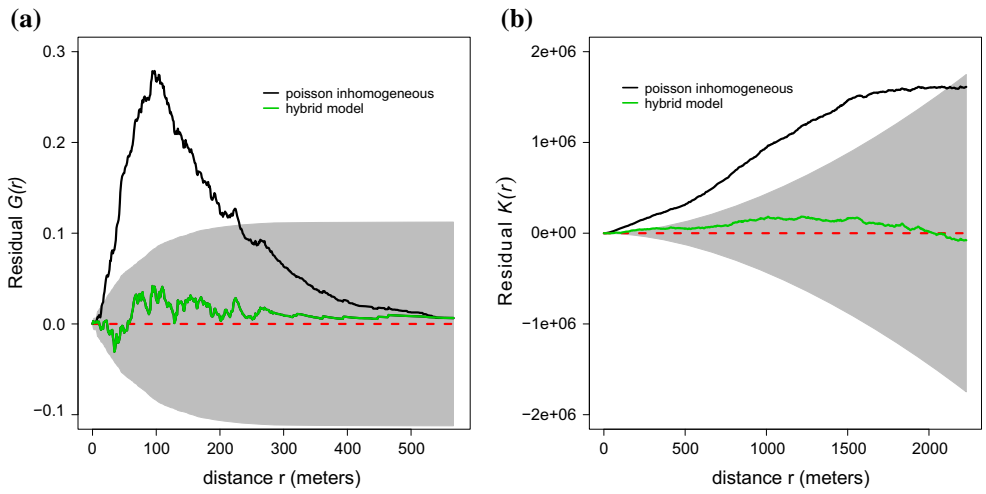


Fig. 11 **a** The residual G function for the hybrid model; **b** The residual K-function for the hybrid model

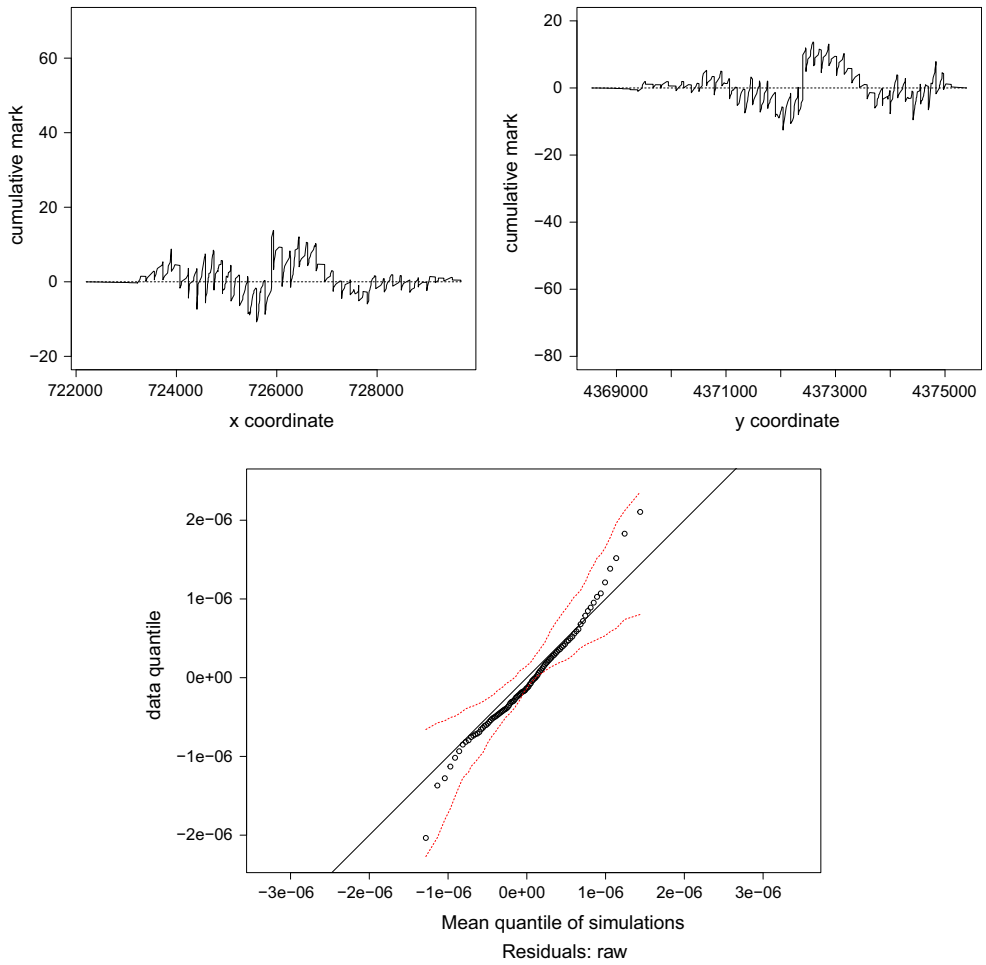


Fig. 12 Cumulative raw residuals for the x and y axis for the hybrid model (*upper row*); qqplot for the raw residuals for the hybrid model (*lower row*)

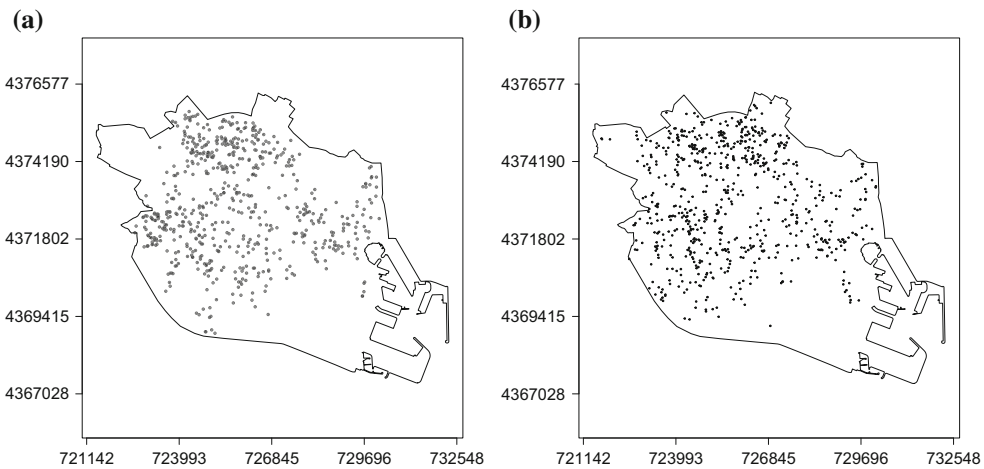


Fig. 13 **a** Spatial distribution of varicella point pattern in Valencia, Spain, cases registered during 2013; **b** Simulation from the hybrid model. UTM coordinate system (distance in meters)

model explains the behaviour of the disease and identifies aggregation at different scales: approximately 50, 100, and 290 meters. Equation (2) shows that every spatial scale r_j is bounded to its corresponding saturation parameter s_j, thus making the interpretation of these parameters more difficult.

As stated in Sect. 3.2, we constructed the point pattern by considering the addresses of residence for each case. As a result, we obtained points with multiple cases which would lead us to a marked point process. We have studied the ground process, meaning the locations where the cases were observed, without the marks, as a first step toward discovering statistical properties of the pattern. The next natural step is to explore the properties of the marked point process. This is one of the main focuses of our future work.

The final fitted model in Sect. 3.3 includes, besides the distance to the nearest school, the focal points. We wanted to analyse the model without taking into consideration any focal point. The reason we did this is, partly, because we already used school information when we introduced the nearest-school distance function. We thought it was somehow redundant to use this information again, by considering the focal points. Figure 14 shows the representation of the errors, using the same scale as in Figs. 10

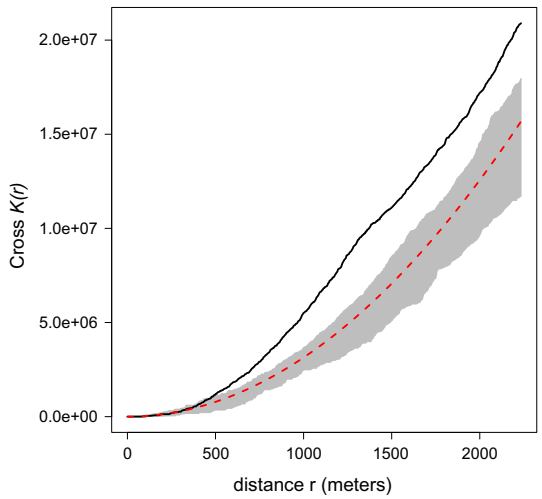


Fig. 15 The cross *K*-function between the varicella and schools patterns

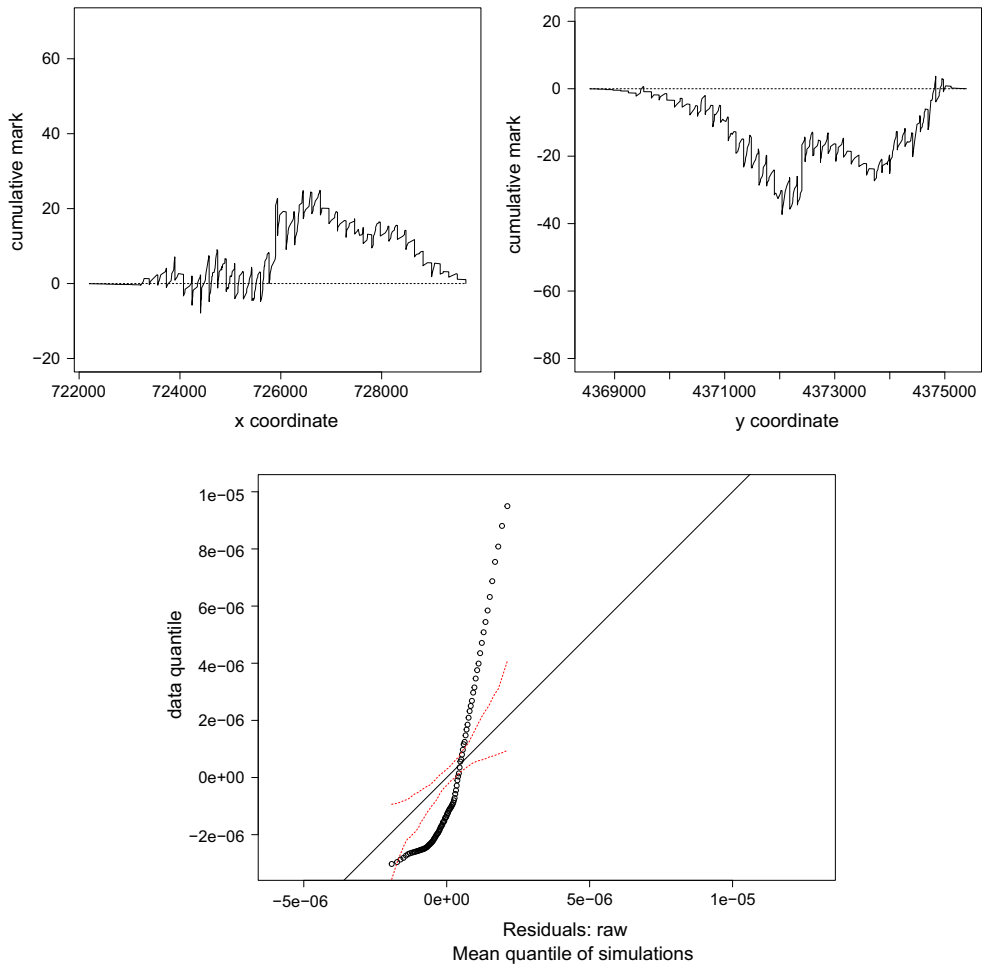


Fig. 14 Cumulative raw residuals for the *x* and *y* axis; qqplot for the raw residuals. Hybrid model without the focal points

and 12 and a qqplot for the hybrid model without the focal points. If we take a careful look, the shape of the errors changes, with certain areas of the study region exhibiting a clearly pronounced shape and high values. The qqplot shows that the errors are not normal. All this indicates that the inclusion of focal points helps to correct the unusual behaviour of the resulting residuals. Another way to verify this is to compute the cross K-function between the pattern of varicella cases and the pattern obtained considering the six focal points. Figure 15 shows the cross K-function estimate (solid line) and the envelopes simulated assuming a stationary (spatially homogeneous) random spatial point process (grey area). As this figure illustrates, the cases are not independently distributed with respect to the focal points.

Our fitted model explains one interesting feature that has been missed by other more epidemiological approaches and therefore adds value to our research. The intensity of incidence of varicella is low at peripheral areas even though families with a large number of children live there. Let us compare Figs. 4b and 6c, the varicella point pattern estimate and the 'average number of persons per family register' estimate. On one hand, we observe that for centrally situated areas (areas with an older population) the average number of persons per family is low. As expected, the intensity of the process in these areas is also low. On the other hand, as stated before, the peripheral areas of the city correspond to low varicella incidence and large families. This seems odd, but a deeper and careful analysis indicates that these peripheral areas of the city are less populated, so less movement of people is expected, and the risk of varicella being contagious is reduced. In addition, children living in large families are quickly infected and become immune faster than those living in families with a reduced number of members.

We have been able to identify such interactions between the incidence of varicella cases and some available covariates. However, other covariates that have not been taken into account here could also play a role. For example, some economic and demographic variables could influence the risk of incidence. The next step will undoubtedly involve considering more covariate information, as well as evolution in time.

Acknowledgments We thank Francisco González of Surveillance Service and Epidemiological Control, General Division of Epidemiology and Health Surveillance – Department of Public Health, Generalitat Valenciana for providing the varicella data. We also thank Ana Míguez from Preventive Medicine and Public Health, University Hospital Dr. Peset, Valencia, for her very useful comments on epidemiology-related issues. Adina Iftimi's research is funded by the Ministry of Education, Culture, and Sports Grant FPU12/04531. The work of Francisco Montes was partially supported by Grants MTM2013-45381-P and MTM2013-43917-P from the Ministry of Economy and Competitiveness. The work of Jorge Mateu was partially supported by Grants MTM2013-43917-P and P1-1B2012-52 (Bancaja project) from the Ministry of Economy and Competitiveness.

References

Alp H, Altnkaynak S, Ertekin V, Kcaslan B, Giiraksin A (2005) Seroepidemiology of varicella-zoster virus infection in a cosmopolitan city (erzurum) in the eastern turkey. Health Policy 72:119–124

Baddeley A, Chang Y, Song Y, Turner R (2012) Nonparametric estimation of the dependence of a point process on spatial covariates. Stat Interface 5(2):221–236

Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (2006) Case studies in spatial point pattern modelling. Springer-Verlag, New York

Baddeley A, Lieshout M (1995) Area-interaction point processes. Anne Inst Stat Math 47(4):601–619

Baddeley A, Moller J, Waagepetersen R (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. Stat Neerl 54(3):329–350

Baddeley A, Rubak E, Møller J (2011) Score, pseudo-score and residual diagnostics for spatial point process models. Stat Sci 26(4):613–646

Baddeley A, Turner R (2000) Practical maximum pseudolikelihood for spatial point patterns. Aust N Z J Stat 42(3):613–646

Baddeley A, Turner R (2005) spatstat: an R package for analyzing spatial point patterns. J Stat Softw 12(6):1–42

Baddeley A, Turner R, Mateu J, Bevan A (2013) Hybrids of gibbs point process models and their implementation. J Stat Softw 55(11):1–43

Berman M (1986) Testing for spatial association between a point process and another stochastic process. Appl Stat 35:54–62

Brachman PS (2003) Infectious diseases—past, present, and future. Int J Epidemiol 32:684–686

Centers for Disease Control and Prevention (2012) Epidemiology and prevention of vaccine-preventable disease. Varicella. The Pink Book: Course Textbook p 12

Chiu SN, Stoyan D, Kendall WS, Mecke J (2013) Stochastic geometry and its applications, 3rd edn. Wiley, New York

Comas C, Mateu J (2011) Statistical inference for gibbs point processes based on field observations. Stoch Env Res Risk Assess 25(2):287–300

Comas C, Palahi M, Pukkala T, Mateu J (2009) Characterising forest spatial structure through inhomogeneous second order characteristics. Stoch Env Res Risk Assess 23(3):387–397

Cooper Robbinsa S, Wardb K, Skinnera SR (2011) School-based vaccination: a systematic review of process evaluations. Vaccine 29:9588–9599

Cressie N, Read T (1984) Multinomial goodness-of-fit tests. J R Stat Soc Ser B 46:440–464

Daley DJ, Vere-Jones D (2003) An introduction to the theory of point processes: volume I: elementary theory and methods. Springer, New York

Diggle P (1985) A kernel method for smoothing point process data. Appl Stat (J R Stat Soc Ser C) 34:138–147

Diggle P (2014) Statistical analysis of spatial and spatio-temporal point patterns. CRC Press, London

European Centre for Disease Prevention and Control (2014) Varicella vaccine in the European Union. ECDC, Stockholm

Funwi-Gabga N, Mateu J (2012) Understanding the nesting spatial behaviour of gorillas in the Kagwene Sanctuary, Cameroon. Stoch Env Res Risk Assess 26:793–811

Geyer C (1999) Likelihood inference for spatial point processes. In: Barndorff-Nielsen OE, Kendall WS, van Lieshout MNM (eds)

Stochastic geometry: likelihood and computation. Chapman and Hall/CRC, Boca Raton

Iftimi A, Martínez-Ruiz F, Míguez Santiyán A, Montes F (2015a) Spatio-temporal cluster detection of chickenpox in Valencia, Spain, 2008–2012. GeoSpatial Health 10(341):54–62

Iftimi A, Montes F, Míguez Santiyán A, Martínez-Ruiz F (2015b) Space-time airborne disease mapping applied to detect specific behaviour of varicella in Valencia, Spain. Spat Spatio-Temporal Epidemiol 14(15):33–44

Jackson C, Mangtani P, Fine P, Vynnycky E (2014) The effects of school holidays on transmission of varicella zoster virus, England and Wales, 19672008. PLoS One 9(6):1–9

Kulldorff M (2010) SatScan User Guide for version 9.0

Lederberg J, Shope RE, Stanley CO (1992) Emerging infections: microbial threats to health in the united states. National Academy Press, Washington DC

Van Lieshout MNM (2010) A J-function for inhomogeneous point processes. Stat Neerl 65:183201

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

Picard N, Bar-Hen A, Mortier F, Chadoeuf J (2009) The multi-scale marked area-interaction point process: a model for the spatial pattern of trees. Scand J Stat 36:23–41

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Renshaw E, Comas C, Mateu J (2009) Analysis of forest thinning strategies through the development of space-time growth-interaction simulation models. Sioch Environ Res Risk Assess 23(3):275–288

Ripley B (1976) The second-order analysis of stationary point processes. J Appl Probab 13:255–266

Ruelle D (1969) Statistical mechanics: rigorous results. Benjamin, Reading

Uria J, Ibanez R, Mateu J (2013) Importance of habitat heterogeneity and biotic processes in the spatial distribution of a riparian herb (carex remota l.): a point process approach. Stoch Env Res Risk Assess 27(1):59–76

Widom B, Rowlinson J (1970) New model for the study of liquid-vapor phase transitions. J Chem Phys 52(4):1670–1684