THE USE OF STATISTICAL POINT PROCESSES IN GEOINFORMATION ANALYSIS

Alfred Stein[a], **Valentyn Tolpekin**[a] **and Olga Spatenkova**[b]

[a]ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands
[b]Helsinki University of Technology, PO Box 1200, FIN-02015

ABSTRACT:

Many objects in space can best be modeled statistically by using point processes. Examples are fires in an urban environment, herds of animals in large areas, earthquakes and forest fires and large speckles on a radar image. Modern developments in point process theory now much better than before allow us to make statistical models to explain the observed patterns. In this paper, we will address the way that point processes can be modeled in space and time. The first application draws from domestic fires at the city level, where we apply a statistical point pattern analysis to derive major causes from related layers of information. The second application considers earthquakes as a marked point process. For earthquakes, large and complex data sets exist including many possibly relevant covariates that may influence their occurrence. The Strauss point process model is explored to analyze earthquake data in Pakistan recorded since 1973, in particular the major earthquake event occurring in 2005. The model, despite some limitations, is rigorous for applying it to such a marked point pattern, representing well the clustering behaviour as determined by a number of environmental factors. Finally, the Strauss point process model is suggested for the use in identifying and explaining the occurrences of speckles in a radar image.

1 INTRODUCTION

Spatial point pattern play an increasingly important role in modern image analysis and geographical information processing. On the one hand, we observe patterns of objects that show a point-like pattern, or at least can be modelled as such, whereas on the other hand several images inherently show point-like patterns. Typical examples of the first category are the locations of settlements in an area, the presence of wildlife herds observable from high resolution remote sensing, whereas in geographical information processing examples include the position of indoor fires in a large city, and the position of earthquakes in space and time. In particular several aspects of object-related noise may exhibit a point-like pattern, and the most common example of such is the presence of speckle on a radar image.

Spatial point pattern analysis is a powerful technique to detect relationships in spatial data distribution. The theory has rapidly grown in recent years and the background is described in an accessible way in (9) and (12), whereas a solid summary is given in (22). Classical examples exist in forestry (2, 13, 14), where either the positions of trees or the positions of gaps in fotrests are mdoeled as a point process. Other examples include studies in epidemiology (12, 17), or wildlife (23). (26) identifies practical difficulties when applying point pattern analysis methods in ecology and provides several relevant gudielines. The analysis methods usually first distinguish between clustering, regularity and randomness, and succeed by providing answers to questions about the scale of clustering and reasons behind the patterns. On the basis of an observed pattern we usually identify a process that generates these. This allows as well an analysis of spatial distributions in time (12). Typical examples discussed below are a Poisson process and a Strauss process, whereas also terms like a clustered or a regular process are used in the literature. It is usually the parameters of such process hat we are interested in, and that we may derive from a collection of observed points, e.g. within a limited window in space and time. Various software tools are now easily available for standard use. In this sense, an increasingly better match may arise between the patterns observable on images and understanding of processes occurring at the earth surface.

The aim of this paper is to briefly introduce the subject and then present some examples of data analysis and recognition. This will include some aspects of the Strauss process model as a specific model for application in spatial analysis.

2 METHODS

2.1 Point patterns

Basic concepts and analysis methods are in e.g. (9, 6, 19). Our interest concerns detection of systematics in the distribution, i.e. regularity or aggregation (clustersing) as deviation from randomness. Complete spatial randomness (CSR) is defined by the following criteria: (i) the number of events in a planar region A of size $|A|$ follows a homogeneous Poisson distribution with mean $\lambda|A|$, where λ is the constant density; (ii) given n events x_i in a region A, the x_i are an independent random sample from the uniform distribution on A (9). In other words, the density of the point pattern does not vary over the bounded region, and there are no interactions among the events.

Density estimation can be based on kernel functions (7) - a bivariate probability density function, which is symmetric around the origin located at a point of estimation. Incidents contribute to density estimation according to their distance from the kernel centre - the closer to the kernel centre, the larger the influence. The range of influence is limited by the kernel bandwidth controlling the smoothness of the result. Density plots with well-chosen bandwidth provide a good summary of the data, whereas a bandwidth that is too large leads to too much smoothing, and a bandwith that is too small over-emphasizes local events, like small variations in the incident pattern. Dependency relationships for local interactions can be described by the nearest neighbour distances defined as the distance from the ith event to the nearest other event in the bounded region of interest. Empirical cumulative probability distribution function $\hat{G}$ for the nearest neighbour distances summarises the incident pattern in an effective way:

$$\hat{G}(w) = \frac{\sum_{w_i \leq w} 1}{n},$$

where w_i is a nearest neighbour distance for the ith event and n is the number of events in the study region. Yet, the observed pattern is usually part of a larger region, where the distribution of events is unknown. Interaction between events lying inside and outside the study region cannot be properly accounted and cause edge effects. A simple but effective adjustment consists in reducing the sample by the buffer defined around the boundary. Events falling inside the buffer are not used for the analysis directly, but unveil the distribution behind the reduced study region.

To ease the interpretation, it is suitable to plot the $\hat{G}-$function against the theoretical curve for CSR, which is (ignoring the edge effects):

$$G(w) = 1 - exp(-\lambda \pi w^2).$$

Importance of the difference between $\hat{G}-$ and $G(w)$ is assessed by using Monte Carlo simulations. For this purpose, empirical cumulative probability distribution functions are generated for nearest neighbour distances for each of 99 realizations of a simulated CSR process with the same density as the original pattern. Its average provides a reference line, maximum and minimum values provide simulation envelopes.

A stochastic mechanism that generates a set of events in the study region is called a spatial point process. To model the dependence of domestic fires on exploratory variables we fit a process, termed a DF process. Its density function reflects the spatial distribution of the different influences. Assuming a stochastic dependence between the points, we use a class of Markov point processes (21), which allows flexible modelling of interpoint interactions. The Strauss process (24, 16) represents an example of Markov point process for pairwise interaction and can be used to simulate a wide range of patterns from simple inhibition to clustering (9, 15). The conditional density of Strauss process is

$$\lambda(u, x) = \beta(u) \cdot \gamma^{t(u,x)},$$

where $\beta(u)$ is the density at location u, $t(u, x)$ is the number of events x that lie within a distance r of u and the inhibition parameter γ controls the strength of interaction between points. For the special case that $\gamma = 1$ the Strauss model reduces to the homogeneous Poisson process with constant density β, the case that $\gamma = 0$ corresponds to a simple inhibition process, whereas for $\gamma > 1$ the model produces a clustered process. The effect of dependence on exploratory variables is expressed with a density being a loglinear function of covariates:

$$log\beta(u) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_n c_n,$$

where the c_i are the explanatory variables and β_i are parameters to be fitted. A linear form is chosen as a first approach in this exploratory study.

2.2 Goodness of fit

Modelling is an iterative procedure, aiming at finding a suitable representation of the data corresponding to observed relationships. The suitability of a model is checked according to several criteria. The Akaike Information Criterion AIC (1) is a versatile measure for model selection. In addition to goodness-of-fit it also considers the number of estimated parameters and the number of observations. A model with the lowest AIC value reflects the best trade-off between bias and variance.

The overall goodness-of-fit for the Strauss models can be assessed based on simulation envelopes of summary functions (9, 18, 3). The $K-$function provides a summary of the spatial pattern over a wide range of scales and is therefore more effective than measures based on the nearest neighbour distances. It is defined as the expected number of other points of the process lying within a distance d of a typical point of the process, divided by the density λ. A suitable estimate of this function given by (20):

$$\hat{K}(d) = \frac{R}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} I_d(d_{ij}),$$

where n is a number of points in the study region with area R, d_{ij} is the distance between ith and jth points, and $I_d(d_{ij})$ is an indicator function, which is 1 if $d_{ij} \leq d$ and 0 otherwise. After adjustment for inhomogeneity this becomes $\hat{K}_I(d, \lambda)$, defined as

$$\hat{K}_I(d, \lambda) = R^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{I_d(d_{ij})}{\lambda(x_i)\lambda(x_j)}.$$

We apply the reduced sample method to adjust the estimate for edge corrections. The $\hat{K}_I-$function is calculated for each of the realizations of a simulated models. In order to test the goodness-of-fit of the model, we consider global envelopes, which represent the largest absolute difference between the simulated and estimated theoretical curves over the entire distance interval. A significance level of 0.05 is achieved after 19 simulations.

Spatstat, an R package designed for analysing spatial point patterns was used for the analysis (4, 5, 3).

3 EXAMPLES AND ILLUSTRATIONS

3.1 Domestic fires

The first example considers domestic fires in Helsinki occurring within a single year. At the city scale, such domestic fires form a spatial pattern. We can derive various summary statistics describing pattern properties. These can represent first order effects describing the number of fires per unit area varying in a study region, or second order effects describing the dependency relationships between fires. A visual inspection of the $\hat{G}$ plot brings the spatial distribution of the pattern to light. An excess of nearest neighbors at short distances indicates clustering in the data, while an excess of long distances neighbors refers to regularity. Buildings and census records form an additional pattern of events. They carry information on types and age of buildings and socio-economical classes (density of population and workplaces, age of households, education, income, unemployment). Second order effects, in particular the $\hat{G}-$function between domestic fires and these patterns of selected influences, provide insight into the relationships in the data. If there are considerably more nearest neighbors at short distances than what would be expected for random distribution, we can assume a correlation between the events. In this way, processes underlying domestic fires are unveiled that indicate the importance of particular exploratory variables.

Modelling the distribution of domestic fires has been used to assess a probability of fire occurrence and analyse the contribution of explanatory variables. A point pattern analysis allows to preserve the level of detail offered by the data itself, in contrast to lattice methods that handle aggregated data. It avoids an ambiguous definition of a lattice scale and therefore enables to draw more accurate conclusions. The methods applied on buildings could be well applied to other cities and may include other phenomena that can be represented as a point pattern or a grid layer, such as crime distribution or house prices.

Conceptualization of studied phenomena as point pattern layers allows to apply well-established statistical methods for spatial point patterns analysis. In addition, spatial statistics offers more than a basis for accepting or rejecting null hypotheses about spatial randomness. The difference from randomness observed from the $\hat{G}-$function indicates a dependence between domestic fires and particular explanatory variables. The $\hat{G}-$function also provides an insight into the aggregation scales for separate variables. Comparison of plots for different variables helps to identify the most important influences. The Strauss model considers all the variables simultaneously and enables to quantify their influence to the distribution of domestic fires through the estimated parameters. The analysis of the distribution in time by splitting the set of events according to different time scales could be enhanced by using periodic splines to directly specify the time domain in the model. Yet, this is beyond the scope of the current manuscript and will be explored in the future.

The point pattern analysis can serve as a basis for generating new hypotheses and complement other data mining methods in the process of knowledge discovery. Still, before drawing reliable conclusions, the obtained results need to be discussed and confirmed with domain experts. Here we can benefit from the visual form of the density and $\hat{G}-$function plots.

The point pattern analysis is hampered by a large number of islands within the study area, which, as being built-up, need to be considered in the study. Rigorous distinction of land and water areas through the observation window and applying an edge correction would give a solution, however, it prolongs the processing time exceedingly. Having in mind that the frequency of incidents on the islands generally decreases with more tedious accessibility, we assume no significant effect of the observation window shape on the final results. To confirm this hypothesis, we performed the $\hat{G}-$function with the reduced sample edge correction method on a restricted area covering only a mainland of Helsinki. As expected, no major differences were observed in the results between the general and restricted observation windows. We therefore proceed the analysis using the simplified observation window.

Fitted models of domestic fires distribution reflect the empirical data. As there always exists a gap between the data and reality, the best model from a mathematical point of view does not need to be the best one in reality. Thus, it is desirable to consider also other criteria and keep the preferred model consistent with a priori knowledge.

Precise interpretation of the fitted parameter values indicates the relations between estimated density and variables involved in the model. However, the parameters can provide an accurate account of the process only in connection with the concrete variables values. The actual degree of influence of particular model variables can be assessed using AIC, for more details see (8). According to Akaike's rule of thumb, two models are significantly different, if the difference of their AIC is more than 2. Thus, model selection based on step-wise variable reduction comparing the AIC values leads to the model representing the most significant variables.

The method is data driven and the reliability of the results depends on the quality of the input datasets. We should therefore consider the data quality carefully and be aware of data quality problems that may occur. In this study we battled with the positional accuracy of the incident dataset. The coordinates of the incident location are inserted via an electronic report filled by the mission commander by clicking a mouse on the corresponding place in the map. This process should ensure the highest possible accuracy. As the commander's main responsibility is in extinguishing the fire, we may put some doubt on the precision of the coordinates of incidents, which may not correspond to incident addresses. Also, temporal variations of explanatory variables are unknown and may influence the results as, for example, population density data are based on permanent addresses. Additional uncertainty emerges with the data processing. We carried out the analysis by splitting set of events into various categories, that may have vague boundaries between them. Although we do not expect major changes in the results, this issue is postponed for a further analysis.

3.2 Earthquakes

In a recent study we investigated earthquake data in the Northern part of Pakistan, an active seismic zone. Data include 1403 earthquakes that occurred in the region between January 1973 and August 2008. The year 2005 is marked by a large seismic activity in the region as compared to the previous years. This is due to a major shock, the Kashmir earthquake of magnitude 7.6, which struck the region on Oct 8, 2005 followed by a range of aftershocks, causing great devastation and misery by killing more than 80000 people and damaging the whole infrastructure of the region. There were 22 earthquakes of magnitude 5.5, out of which 12 occurred the same day as the major earthquake and 15 earthquakes of magnitude 5.5 occurred within 15 days after the Kashmir earthquake. Only 7 other earthquakes of magnitude 5.5 occurred during the past 35 years. The seismicity of the area decreases after the first month after the Kashmir earthquake and the number of events in the preceding months is almost negligible as compared to the first month after the main shock. Locations of all earthquakes within one month after the Kashmir earthquake are in its close vicinity. Only four earthquake locations lie more than 50 km from the aftershocks region. The analysis of the data considering it as a point pattern will be based on this study region and the earthquakes located within it.

The epicentre region lies on the western edge of the Himalayan Arc, which denotes the area of continental convergence between the Indian and Eurasian tectonic plates. The Indian plate moves northwards at a rate of about 40mm/year and subducts below the Eurasian plate. The Kashmir earthquake is associated with fault rupture near the western end of the MBT in Kashmir region of Northern Pakistan. Location of tectonic plates boundaries plays a significant role in determining seismicity of the study area.

To assess the influence of geological faults located in the study region on the earthquakes distribution pattern, the distance of earthquake locations to faults could serve as additional information (covariates) in modelling the point pattern. For that purpose a a geo-referenced Tectonic map of Pakistan. From this map the study area of the earthquakes data was extracted using its bounding coordinates and the faults within the study area were digitized. Aftershocks earthquakes occurred along the plate boundaries with a dense cluster of aftershocks near the point where the two boundaries converge. Thus the location of plate boundaries can possibly serve as an important factor contributing to the distribution pattern of the earthquakes. To evaluate the contribution of plate boundaries location, a pixel image of the shortest distance of each pixel from the pate boundary was obtained. Similarly, to test the effect of active faults in the study area on the earthquake point pattern, distance of each earthquake location was calculated from the nearest fault.

An earthquake hypocenter is the three dimensional point in the earth where the rupture of an earthquake begins. For large earthquakes, the ruptures may extend up to several kilometres, and the hypocenter may be anywhere along the rupture. The epicentre of an earthquake event is the point location on the surface of the globe that represents the projection of the hypocenter onto the surface of the globe.

The explanatory variables, apart from the Cartesian coordinates, consisted of the information about the spatial location of the plate boundaries and geological faults in the study area given as pixel images showing shortest distance to the nearest plate boundary and nearest fault location for each pixel. The application of Strauss point process model proved satisfactory in explaining the spatial trends and capturing the sources of variability introduced by the explanatory variables. The application showed that the locations of plate boundaries and geological faults are significant determinants for the earthquake epicentre locations. When the effects of both these variables were combined along with the magnitudes and geographic locations of the earthquake epicentres, the modelling was significantly improved. The effects of the explanatory variables were quantified by improvement in AIC values. The improvement in the modelling of earthquake location can also be assessed visually by the plots of fitted trends for different types of earthquakes.

3.3 Speckle on remote sensing images

Radar images provide important information about the earth surface that is complementary to optical remote sensing images. In some cases it has advantage over optical images, e.g. in dense cloud cover conditions and observation at night time. Synthetic Aperture Radar (SAR) is a special case of radar system where relatively high spatial resolution is achieved due to coherent processing of many recorded responses.

Speckle is an inherent property of SAR images. It originates from interference of coherent responses coming from many scattering elements within a resolution cell. It results into a large variance of radar image compared to its mean value. Therefore SAR images are difficult to use for automatic classification purposes. In several instances, it is required to reduce the effect of speckle, being this goal of SAR image despeckling.

As an example we may consider the ERS-2 Single Look Complex (SLC) image covering Serowe region in Botswana (see Figure 1. The patterns of the spikes show a point structure that may identify important issues related to land processes. The image presented attached subset of HH image of Botswana. The bright spots in the left part of the image are results from strong reflectors in a city. Hence, also after despeckling, the remaining pattern of extremes shows a pattern that can be readily analyzed and interpreted using a statistical point pattern analysis.

4 DISCUSSION

At this stage, good results are obtained with the combination of spatial point pattern theory and remote sensing, as well as in its combination with geographical information processing. The combination of readily available software tools and the request for an increasingly better data quality may lead to a more regular use of the methodology, thus leading to answering relevant questions. In particular, modern methods on space-time point processes may become beneficial to better understand the development of patterns in space and time.

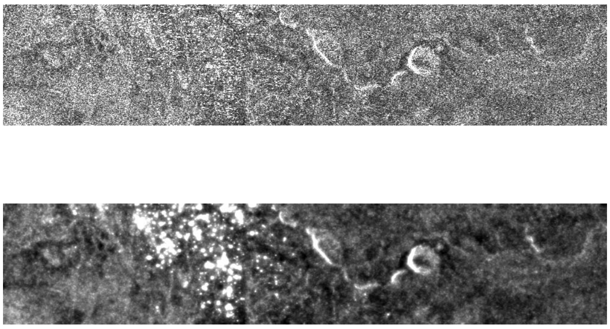

Figure 1: (an ERS-2 Single Look Complex (SLC) image covering Serowe region in Botswana and the same image after despeckling. .

Standard GIS packages do not yet contain easy to use and interpretable spatial statistical software. In the context of geoinformation processing, this is a clear deficiency, as such procedures are useful for a wide set of applications. A better integration of the spatial statistical software and GIS packages is a necessary step forward. An important reason in this respect is that a spatial statistical summary of collected or registered point data may be helpful to further communicate quantitative findings to the user.

A step to further explore concerns the issue of spatial data quality. Spatial data quality is firstly relevant in terms of accuracy of the observations,. In the study described above on earthquakes this plays an important role, as an earthquake occurs at some depth below the Earth crust, whereas its effects are mainly visible at the surface. Moreover, there is never a precise location of such an event, and only an approximate value. The second issue related to point patterns is their attribute, which may be difficult to define in full. The domestic fire example may at several instances relate the question whether any fire that is registered in a house is in fact domestic fire. Buildings may be used for different purposes, and there is usually an issue of not reporting such a fire to fire brigades, or reporting it in a deviate way. Such issues apply to a range of other spatial point patterns in a similar way. An as yet somewhat unexplored domain concerns the use of marked point processes in remote sensing images. When additional information comes available from images, it is not difficult to imagine that such methods can be useful for a range of applications. In particular, we see good opportunities in deforestation studies and in development of urban regions.

Finally, recent progress has been made on spatial processes in modeling of spatial extremes and of modeling point patterns in the space-time domain. A good example of the first type of study is in soil contamination. The second type of analysis is well presented in two recent papers (11), (10).

REFERENCES

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Atkinson, P. M., Foody, G. M., Gething, P. W., Mathur, A. and Kelly, C. K., 2007. Investigating spatial structure in specific tree species in ancient semi-natural woodland using remote sensing and marked point pattern analysis. *Ecography* 30, 88–104.

Baddeley, A., 2008. Analysing spatial point patterns in R. *CSIRO workshop notes [online]*. Available from: http://www.csiro.au/files/files/pn0y.pdf [Accessed 1 April 2008].

Baddeley, A. and Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* 12 (6), 1–42.

Baddeley, A. and Turner, R., 2007. *Spatial Point Pattern analysis* [online]. Available from: http://www.spatstat.org

Bailey, T. C. and Gatrell, A. C., 1995. *Interactive Spatial Data Analysis*. Longman, Harlow.

Bowman, A. W. and Azzalini, A., 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, New York.

Burnham, K. P. and Anderson, D. R., 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, New York.

Diggle, P. J., 2003. *Statistical Analysis of Spatial Point Patterns*. 2nd ed. Arnold, London.

Diggle, P. 2007. Spatio-Temporal Point Processes: Methods and Applications. In Finkenstadt, B., Held, L. and Isham, V. (eds) *Statistical methods for spatio-temporal systems* Chapman and Hall, Boca Raton, pp. 1–45.

Diggle, P. and Gabriel, E. 2009. Spatio-Temporal Point Processes. In Gelfand, A., Diggle, Guttorp, P. amd Fuentes, M. (eds) *Handbook of Spatial Statistics* Chapman and Hall, pp. 449 – 461.

Gatrell, A. C., Bailey, T. C., Diggle, P. J. and Rowlingson, B. S., 1996. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, NS 21(1), 256–274.

Getis, A. and Franklin, J., 1987. Second-order neighborhood analysis of mapped point patterns. *Ecology*, 68(3), 473–477.

Getzin, S., Dean, Ch., He, F., Troyfomow, J. A., Wiegand, K. and Wiegand, T., 2006. Spatial patterns and competition of tree species in a Douglas-fir chronosequence on Vancouver Island. *Ecography*, 29(5), 671–682.

Gregori, P. and Mateu, J., 2002. Spatial point processes: an overview. *In*: J. Mateu, Montes and F. eds. *Spatial Statistics Through Applications*. WIT Press, Southhampton, Boston.

Kelly, F. P. and Ripley, B. D., 1976. A note on Strauss's model for clustering. *Biometrika*, 63, 357–360.

Martínez-Beneito, M. A., Abellán, J. J., López-Quílez, A., Vanaclocha, H., Zurriaga, Ó., Jorques, G. and Fenollar, J., 2006. Source detection in an Outbreak of legionnaire's disease. *In*: A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan eds. *Case Studies in Spatial Point Process Modeling*. Lecture Notes in Statistics 185, Springer, New York.

Mattfeldt, T., Eckel, S., Fleischer, F. and Schmidt, V., 2007. Statistical modelling of the geometry of planar sections of prostatic capillaries on the basis of stationary Strauss hard-core processes. *Journal of Microscopy* 228 (3), 272–281.

O'Sullivan, D. and Unvin, D. J., 2003. *Geographic Information Analysis*. Wiley, Hoboken.

Ripley, B. D., 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13, 255–266.

Ripley, B. D. and Kelly, F. P., 1977. Markov point processes. *Journal of the London Mathematical Society*, 15, 188–192.

Schabenberger, O,. and Pierce, F.J., 2002. Contemporary statistical models for the plant and soil sciences. CRC Press, Boca Raton.

Stein, A. and Georgiadis, N., 2006. Spatial Marked Point Patterns for Herd Dispersion in a Savanna Wildlife Herbivore Community in Kenya. *In*: A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan eds. *Case Studies in Spatial Point Process Modeling*. Lecture Notes in Statistics 185, Springer, New York.

Strauss, D. J., 1975. A model for clustering. *Biometrika*, 62, 467–475.

Walter, C., McBratney, A.B., Viscarra Rossel, R.A. and Markus, J.A. 2005. Spatial point-process statistics: concepts and application to the analysis of lead contamination in urban soil. *Environmetrics* 16 339355

Wiegand, T. and Moloney, K. A., 2004. Rings, circles, and null-models for point pattern analysis in ecology. *OIKOS*, 104(2), 209–229.

9