Discussion of spatial statistics: marks, maps and shapes

Denis Allard

► To cite this version:

Denis Allard. Discussion of spatial statistics: marks, maps and shapes: Discussion on Antonio Pos-solo's presentation. Stu Hunter Conference, Mar 2015, Leuven, Belgium. 7 p. hal-02792549

HAL Id: hal-02792549

<https://hal.inrae.fr/hal-02792549>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Discussion of “Spatial Statistics: Marks, Maps, and Shapes”

Denis Allard

INRA, UR546 Biostatistics and Spatial Processes (BioSP)

Site Agroparc, 84914 Avignon, France

E-mail: allard@avignon.inra.fr

I would like to congratulate Dr. Possolo for his excellent – and very personal – presentation of the subject. I greatly enjoyed his presentation in Leuven, and I had the same pleasure reading his contribution. I will first discuss his categorization into Marks, Maps and Shapes. I will then share some personal – obviously partial – views on three main challenges faced in the field of spatial statistics, that were not addressed in the paper. They relate to spatio-temporal problems, data massiveness and unconventional data.

Introduction

One particularity of spatial statistics is that early methodologies were largely developed in different areas of applications rather than in university statistics departments. Geostatistics was pioneered by Matheron and colleagues in a mining engineering context (Matheron, 1965; see also Chilès and Delfiner, 2012, for an extensive presentation of geostatistics); spatial considerations were implicit in Fisher’s work on randomization and blocking for the analysis of agricultural field trials (Fisher, 1966); forestry was the context of the seminal PhD thesis of Matérn on spatial variations (published in Matérn, 1986).

It was later introduced into mainstream statistical research by two seminal contributions: Besag (1974) proposed models and associated methods for analyzing spatially discrete “lattice” data, while Ripley (1977) introduced a family of functions, known as Ripley’s K and L functions for the analysis of spatial point processes. Since then, spatial statistics has been enriched with important contributions, in particular relating to point processes and object models (Stoyan, Kendall and Mecke, 1987), model based geostatistics (Diggle, Moyeed and Tawn, 1998), hierarchical modeling for spatial data (Banerjee, Carlin and Gelfand, 2004) and spatio-temporal statistics (Cressie and Wikle, 2011).

The historical context briefly sketched above grounds the usual presentation of spatial statistics in three major branches: (1) continuous spatial variations, i.e. random field theory and geostatistics; (2) discrete spatial variations, including lattice processes and areal unit data; (3) spatial point patterns, including marked point processes and object models.

The categorization proposed by Dr. Possolo into Marks, Maps, and Shapes deviates from this classical presentation. The “Marks” category includes lattice models (2) together with point processes (3); “Maps” corresponds evidently to continuous spatial variations (1); and “Shapes” corresponds to

the description, analysis and modeling of objects (3). Lattice gases, presented here as a mark process (each site of the lattice is either white or black) could be, and usually is, presented as a (binary) Markov random field, the most widely-used tool for building models with discrete spatial variations. Obviously, the boundaries between these categories are neither rigid nor impermeable. Data in ecology or environmental sciences are most of the time multivariate and multitype. Complex spatial models with components of every type are therefore necessary for the analysis of such datasets.

Since the early 1990s, spatial statistics, together with its more recent development towards spatio-temporal statistics, has become an ever-growing domain of statistics. The past 25 years has seen an explosion of interest in space and space-time problems, largely fueled by the computer revolution and by the increased availability of geo- and time- referenced data. Figure 1 reports the number of publications returned by a Web of Science query with “Spatial Statistics” as topic, from 1991 to 2014. The number of publications increased by 10% each year, corresponding to a doubling every seven years. This exponential growth illustrates the dramatic growth of research within this field as well as its application to other scientific fields. Spatial statistics is an expanding research domain, and as such it faces new challenges.

I will briefly present three of those challenges, thereby proposing a complementary view to that of Dr. Possolo. They mainly arise from the data acquisition revolution brought about by inexpensive, mobile, connected measurement devices. They are expected to have a profound impact on how data will be modeled and how computations will be carried out in the near future.

Figure 1: Number of publications per year, returned by a Web of Science query with ”Spatial Statistics” as topic.

From spatial to spatio-temporal statistics

At first, it is tempting to consider that spatio-temporal processes are not that much different from purely spatial processes. From a purely mathematical point of view, this is indeed the case: time could be considered as an additional coordinate. From a physical prospective, this view is insufficient. The movie of life goes forward, and usually does not make much sense when played backwards. Time differs intrinsically from space, in that there is a *time arrow*, whilst there is usually no preferred direction in space. In addition, there is usually a very large discrepancy between space and time lags. Two quite different situations are very frequent. In the first situation, corresponding for example to

an air pollution monitoring network, data are dense in time and sparse in space. Multivariate time series seem to offer an adequate framework of analysis. In the second case, data are dense in space and sparse in time. This is for example the case when analyzing remote sensing data at different dates. Here, multivariate spatial fields might seem a good modeling strategy. Unfortunately, these models generally separate time and space, in the sense that spatio-temporal covariance functions will be the product of spatial and temporal covariance functions. Separable models do not allow for spatio-temporal interaction. They correspond to processes that will frequently fail to model physical process adequately, in particular in environmental applications. Statistical models that account for as much physics as possible are thus necessary to address the very spatio-temporal nature of the process under study.

A first approach to account for this issue is to build valid and flexible covariance functions that do not separate time and space. Straightforward constructions include combinations of products and sums of purely spatial and temporal covariances or the use of geometric anisotropy on covariance functions defined in $\mathbb{R}^{d+1}$. In a seminal paper, Gneiting (2002) gave a very general class of non-separable covariance functions built from the spectral representation of covariance functions that has now become the standard class of models for univariate spatio-temporal random fields in geostatistical applications. Let $\phi(r)$, $r \in \mathbb{R}$ be a completely monotone function and let $\psi(r)$, $r \geq 0$, be a positive function with a completely monotone derivative. Then, the function

$$C(\mathbf{h}, u) = \frac{1}{\psi(u^2)^{d/2}} \phi\left(\frac{\|\mathbf{h}\|^2}{\psi(u^2)}\right)$$

is a covariance function of the space-time domain $\mathbb{R}^d \times \mathbb{R}$. More general mixture representations were later obtained in Schlather (2010). Current research aims at building general and flexible classes of covariance functions for multivariate spatio-temporal processes. Scheuerer and Schlather (2012) built covariance functions that honor physical constraints for the modeling of random fields of divergence-free or curl-free random vectors, with an application to the analysis of wind fields over Germany. An open challenge is to characterize the covariance functions that are compatible with particular partial differential equations, such as reaction-diffusion PDEs.

A second approach consists in building mechanistic-statistical models that combine a statistical model for the observation process with a physical model for the dynamics under study (e.g. pest dynamics). Here, spatio-temporal interactions are part of the physical model. An example of such a construction can be found in Roques, Soubeyrand and Rousselet (2011) for the analysis of the northward expansion of the pine processionary moth (PPM). In the Paris Basin, its range has shifted 87 km northward between 1972 and 2004. Since 2005, this expansion has been especially well documented with the establishment of a geo-referenced map of the front winter nests all over France. The statistical model bridges the gap between continuous data (nest densities) and binary data (presence/absence in $2 \text{ km} \times 2 \text{ km}$ cells) and, conversely, provides a way to estimate the parameters of the mechanistic model based on the observed data. The mechanistic model describes the discrete-time evolution of nests as a function of continuous-time adult dispersal and of the environmental effect on PPM fitness. Their combination in a Bayesian setting resulted in maps of both the nest density and the regionalized fitness parameter.

Data massiveness

Spatio-temporal data are becoming available in overwhelming volumes and diverse forms as a result of growing remote-sensing capabilities, ground-based sensor networks, crowdsourcing, citizen science data, as well as medical sensing technologies. The processing of such diverse and massive data poses conceptual, methodological, and technical challenges, which are exacerbated by data diversity.

Maximum likelihood estimation for spatial data usually requires mathematical operations (inverting a matrix, evaluating a determinant) which have computational costs to the order of $O(n^3)$, where n is the number of data at hand. It is thus limited to datasets with $n < 10^4$. Approaches for alleviating the dimensionality problem are required for larger datasets. In the past decade, there have been several approaches proposed to overcome this problem (see Sun, Li and Genton, 2011, for a review), such as tapering the covariance function, using composite likelihood methods, or approximating the random fields by the use of a latent process of reduced dimension. One common feature of all these methods is that they consider an approximate version of the true model by leaving aside some information deemed unimportant. As an alternative route, one may chose to use Gaussian Markov Random Fields (GMRFs) representation, which offer appealing computational properties. This route bridges the gap between the categories discussed above: between continuous and discrete spatial representations on the one hand; between Marks and Shapes on the other hand.

Covariance functions that do not have bounded support tend asymptotically to zero at large distances. Tapering the covariance function consists to set to 0 all small values at large distances while keeping the original covariance at short distances in a way such that the transformed covariance is still definite positive. As a result, covariance matrices are highly sparse, allowing the use of very efficient algorithms for estimation and/or prediction. Under appropriate conditions, tapered maximum likelihood have good properties, such as consistency, asymptotic normality and high relative efficiency. Covariance tapering can also be used for spatial prediction, also known as kriging. It has been shown to be asymptotically accurate and computationally efficient (Furrer, Genton and Nychka, 2009).

Composite likelihoods are products of smaller marginal or conditional likelihoods computed on subsets of data. They are thus fast to compute and do not require storage capacities. Varin (2011) provides an overview of composite likelihood methods. An application to spatio-temporal data is given in Bevilacqua et al. (2012), in which it is shown that pairwise composite likelihood estimation has a relative efficiency close to 1.

Fixed rank kriging captures the main features of the spatial dependence described by the $n \times n$ covariance matrix Σ using data dimension reduction techniques. The matrix Σ is decomposed into the sum of a low rank and a sparse component, $\Sigma = \mathbf{B}\mathbf{F}\mathbf{B}^t + \mathbf{G}$, where $\mathbf{B}$ is an $n \times m$ matrix arising from m basis functions $(B_1(\cdot), \dots, B_m(\cdot))$, with $m \ll n$, $\mathbf{F}$ is an $m \times m$ covariance matrix and $\mathbf{G}$ is a sparse matrix. Then, the Sherman-Morrison-Woodbury inversion formula yields

$$\Sigma^{-1} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{B}(\mathbf{F}^{-1} + \mathbf{B}^t\mathbf{G}^{-1}\mathbf{B})\mathbf{B}^t\mathbf{G}^{-1}.$$

This formula is very efficient since it only requires the inversion of one sparse matrix and one low dimensional matrix. Fixed rank kriging has been applied (Cressie and Johannesson, 2008) with success for the kriging of massive spatio-temporal data, such as total column ozone data over the entire globe.

GMRFs are specified through the set of n conditional distributions of one component given all others. Gaussian properties imply that GMRFs are directly specified by their precision matrix $\mathbf{Q} = \Sigma^{-1}$. When all conditional distributions actually depend on very few neighbors, $\mathbf{Q}$ presents the

appealing property of being very sparse, thus yielding highly efficient algorithms (Rue and Held, 2005). The computational gains from GMRFs have been further boosted by applying Integrated Nested Laplace Approximation on top of predictive process GMRF models (Rue, Martino and Chopin, 2009), thus dramatically reducing the computation burden for the inference of large spatial datasets with a hierarchical structure.

A major breakthrough for modeling very large spatial datasets was achieved in Lindgren, Rue and Lindström (2011) where it is proposed to fit GMRFs with local neighborhood on non regular lattices for Gaussian random fields with Matérn covariance function. Such random fields are shown to be solutions of a particular Stochastic Partial Differential Equation (SPDE). The authors then generalize the construction of the corresponding (SPDE) using triangulation techniques from finite element methods. This avoids the interpolation on regular grid points and allows to refine the resolution where details are required. Lindgren, Rue and Lindström (2011) sets a new avenue for the modeling of massive spatial datasets. It bridges the gap between continuous and discrete spatial variations both elegantly and efficiently. Currently, this approach cannot be used for spatio-temporal data, but important research efforts are devoted to this issue.

Data diversity and unconventional data

Recent datasets in ecological and environmental applications come now in very diverse formats, qualities, types and spatial supports. For example, Figure 2 shows the main landscape features (roads, canals and hedgerows) in an agricultural area near Avignon (France) dominated by orchards. This image is part of a statistical analysis aimed at assessing the role of these features on pest propagation over an agricultural landscape. The full dataset includes data of all types discussed above: countings of pests in insect traps, location and severity of pest attacks in orchards, orchard types and descriptions, etc. Statistical models describing such data are highly structured, involving a spatial hierarchical structure in a Bayesian setting. There are also quite complex spatial dependencies between data of all types seen above, including geometrical objects such as connected segments. The analysis of the northward expansion of the PPM that illustrated mechanistic-statistical models could also illustrate the type of approach to be developed for the analysis of a process with very diverse data.

More recently, novel data classes have appeared, such as networks, trajectories, reported events along routes and streams of data provided by connected devices. With these new types of data, boundaries between categories are blurring at an increasing rate. The current analytical frameworks have not been designed to cope with these data. Today, approaches for the analysis of very diverse datasets are still to be developed. Novel theoretical models and computational tools must be developed to build statistical models and inference methods within this context and to tackle issues such as change of support, impact of uncertainties, bias and preferential sampling.

Acknowledgments

I would like to express my gratitude to the organization committee, (especially Peter Goos, Martina Vandebroek, R. Does and M. Schoonhoven) for giving me the opportunity to discuss these issues at the 2015 Stu Hunter Research Conference that was held in Leuven (Belgium) and in this Special Issue.

Figure 2: Main landscape features in an area near Avignon (France). Blue: canals; Red, roads; Black, hedgerows.

About the author

Denis Allard is Directeur de Recherche, at INRA, France. He obtained a PhD from the Ecole des Mines de Paris (now MinesParisTech) in 1993. Before joining INRA, he served as Assistant Professor at the Dept. of Statistics at University of Washington (Seattle, WA) and has worked as Senior Geostatistician for BP research. He joined INRA in 1996. His research activities are focused on statistical models for spatial and spatial-temporal processes (including point and object processes) and on developing statistical methods for assessing the spatio-temporal interactions implied by these models. Currently, his research includes the development of multivariate spatio-temporal models for describing climate variables in view of building stochastic weather generators. He is a member of the editorial board of *Statistics and Computing*, *Spatial Statistics* and *Mathematical Geosciences*. He chaired the third *Spatial Statistics* conference, which was held in Avignon in June 2015.

References

- Banerjee, S., Carlin, B., Gelfand, A.E. (2004). *Hierarchical Modelling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC Press.
- Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, B*, **36**, 192–225.
- Bevilacqua, M., Gaetan, C., Mateu, J., Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association*, **107**, 268–280.
- Chilès J.-P., Delfiner P. (2012). *Geostatistics: Modeling spatial uncertainty, 2nd Edition*. New-York, NY: Wiley.
- Cressie, N., Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, B*, **70**, 209–226.

- Cressie, N., Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, New-York.
- Diggle, P.J., Moyeed, R.A., Tawn, J.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.
- Fihser, R.A. (1966). *The Design of Experiments, 8th Edition*. London: Oliver and Boyd.
- Furrer, R., Genton, M.G., Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523.
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, **97**, 590–600.
- Lindgren, F., Rue, H., Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). *Journal of the Royal Statistical Society, B*, **73**, 423–498.
- Matérn, B. (1986). *Spatial variations, 2nd Edition*. Berlin: Springer-Verlag.
- Matheron, G. (1965). *Les variables régionalisées et leur estimation. Une application de la théorie des fonctions aléatoires aux Sciences de la Nature*. Paris: Masson.
- Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, B*, **39**, 172–212.
- Roques, L., Soubeyrand, S., Rousselet, J. (2011) A statistical-reaction diffusion approach for analyzing expansion processes. *Journal of Theoretical Biology*, **274**, 43–51.
- Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*, Boca-Raton, FL: Chapman & Hall/CRC.
- Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, B*, **71**, 319–392.
- Schlather, M. (2010). Some covariance models based on normal scale mixtures. *Bernoulli*, **16**, 780–797.
- Scheuerer, M., Schlather, M. (2012). Covariance models for divergence-free and curl-free random vector fields. *Stochastic Models*, **28**, 433–451.
- Stoyan, D., Kendall, W.S., Mecke, J. (1987). *Stochastic Geometry and Its Applications*. Chichester: Wiley.
- Sun, Y., Li, B., Genton, M. (2012). *Geostatistics for Large Datasets*. In E. Porcu et al. (eds.), *Advances and Challenges in Space-time Modelling of Natural Events*, Lecture Note in Statistics 207. Springer.
- Varin, C., Reid, N., Rity, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42.