X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

The ABC of Spatial Statistics

Rosa M. Crujeiras

Departamento de Estatística e Investigación Operativa
Universidade de Santiago de Compostela
(rosa.crujeiras@usc.es)

1. INTRODUCTION

Spatial statistics is concerned with random phenomena whose spatial location contributes directly to the underlying stochastic model. Data with a spatial or geographical reference is encountered in a variety of applied fields such as ecology, meteorology, mining, epidemiology, forestry or satellite imaging, among many others. The spatial feature may be sometimes the driving force of the process, but even when the study of purely spatial effects is not the primary goal, the dependence between observation due to their proximity must be regarded in any descriptive or inferential analysis. Classical spatial statistics generally distinguishes three main situations where a spatial component arises, specifically *geostatistics* (continuously varying spatial processes), *lattice processes* (discrete spatial variation) and *point processes* (random pattern for spatial locations maybe with associated random observations -marks-), owning particular tools and goals, but sharing the common factor of dealing with dependent observations.

In this talk, we will introduce some basic concepts in spatial statistics, presenting the main ingredients for a proper analysis in each of the different situations, discussing possible extensions, weaknesses and brand–new areas of development. Some application examples will be also shown in the presentation.

2. WHAT MAKES SPATIAL STATISTICS *SPECIAL*?

Spatial statistics owns a particular history within the statistical field, since quite a few important developments do not come from the study of mathematics, but from the needs arising in different applications. Some of the first references to a spatial component appear in Fisher's randomized agricultural trials, but the main developments within the geostatistical field were carried out in the 50s by Krige in mining engineering and continued in L'Ecole des Mines de Fontainbleau by Matheron. Almost at the same time, Bertil Matérn introduced a class of correlation models for real–valued, spatially continuously varying stationary processes (the Matérn class), and also contributed to spatial sampling and point process theory. Lattice processes, mainly Gaussian Markov random fields, were approached from a different perspective and presented by Besag (1974) in his seminal paper. Given the unlike origins of

each subdiscipline and obviously their own special features, it is not easy to define a common scheme to assemble geostatistics, Markov random fields and point processes together, at least, from a frequentist perspective. Besides the lack of a joint scenario for spatial data analysis, another substantial feature is the non negligible role of the dependence structure (not being a nuissance parameter), neither in geostatistics, lattice or point processes. A naive way of thinking could lead the practitioner to get inspired by what has been done in time series, where dependence is also a driving force. Both areas deal with dependent discrete observations, but apart from this, the approach to the descriptive or inferential problem is intrinsically different.

In all different situations within spatial statistics, we are interested in explaining the variability of the process at two different scales (as it's own goal or as a means for prediction): the large–scale variability (trend) and the small–scale variability (dependence), which can be approached directly or through a model–based strategy. Keeping this premise in mind, we will now introduce some basic ideas in geostatistics, lattice and point processes, noticing some further extensions.

3. TREND VS. DEPENDENCE IN GEOSTATISTICS

Consider a spatial process $\{Z(\mathbf{s}),\ \mathbf{s} \in D\}$, continuously varying over $D \subset \mathbb{R}^d$ ($d \geq 2$). A simple way of approaching the large–small scale description, is to decompose Z as follows:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d \tag{1}$$

where $m(\cdot)$ is a deterministic function capturing the large–scale variability (spatial trend) and ε is usually a zero–mean stationary Gaussian process. Representation (1) can be viewed as a regression model and therefore, estimation procedures can be borrowed from the regression literature. From this approach, the dependence can be seen as a nuisance parameter which is not of primary interest but which needs to be accommodated appropriately in the model in order to avoid compromising the estimation and interpretation of regression and other parameters. We should notice that, although prediction was in principle the main goal of geostatistical methods, as it will be seen later, estimation helps in the understanding of the stochastic device, and this knowledge may be incorporated in order to improve prediction results.

In a classical geostatistical analysis, one should start with the specification of a parametric model for the trend, or proceed with a nonparametric fit, ignoring the spatial dependence structure to obtaine a provisional estimate of the large scale variation. In this first step, no knowledge of the second–order structure of the process is needed. In a a second step, one would try to explore the dependence structure based on the residuals from the provisonal trend estimator. In the next sections, we recall some ideas on how the trend and dependence structure can be approximated, jointly with a (not exhaustive at all) list of references.

3.1 Regression–inspired estimation

Assuming a spatial unilateral first–order autoregressive moving average model for the errors in (1), Basu and Reinsel (1994) propose an iterative estimation procedure

for the estimation of a parametric linear trend in the spatial setting. In their approach, the error covariance matrix is computed by maximum likelihood or restricted maximum likelihood, based on residuals. Multi–stage *ad hoc* fitting procedures for linear and nonlinear trends were also proposed by Haas (1996) and Neumann and Jacobson (1984), but with no theoretical results. The general case where the trend may be nonlinear and the errors are spatially correlated (with unknown dependence) was considered by Crujeiras and Van Keilegom (2010), introducing a two–step procedure based on least squares estimation both of the trend and the dependence parameter, but that could be extended allowing for maximum likelihood estimators of the covariance structure.

The simplistic formulation in (1), can be imbued in the generalized linear model (GLM) framework, which in the classical setting for independent data, provides a unifying framework for regression models with continuous or discrete response. This formulation has been adapted to account for dependent observations (see Diggle and Ribeiro (2007), Chapter 4, for dicussion on generalized geostatistical linear models). Generalized linear mixed models (GLMM) are obtained by the inclusion of a random effect (or a latent variable) in the linear predictor, which in the simplest case, is assumed to be mutually independent accross observations, allowing for over–dispersion with respect to the corresponding GLM. In order to account for spatial variation, the random effects can be interpreted as a signal from a certain Gaussian field, producing a generalized linear geostatistical model (GLGM), thoroughly investigated by Diggle *et al.* (1998).

3.2 Spatial dependence through variograms.

A very simple model of spatial variation assumes that the process is Gaussian, zero–mean and has stationary increments (intrinsically stationary). This means that the spatial covariance can be characterised through the variogram $2\gamma(\mathbf{s},\mathbf{u}) = \mathrm{Var}(Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{u}))$. If the process is (second–order or weak) stationary, where mean and variance are independent of locations, then the variogram is just a function of the different vector between the locations $2\gamma(\mathbf{s},\mathbf{u}) = 2\gamma(\mathbf{u})$. The variogram is a very well established tool in the analysis of spatial data and Cressie (1993) gives a detailed treatment of variogram properties and estimation. There has been a broad interest in the spatial statistical literature for estimating the variogram, both from parametric and nonparametric perspectives. There exist a large variety of parametric families for variograms, such as the already mentioned Matérn class. Within a parametric family, estimation can be done by least–squares (ordinary, weighted or generalized), maximum likelihood or minimum distance methods.

Least–squares methods are based on a pilot (nonparametric) estimator of the variogram. Given that the variogram is the variance of the difference process, it can be estimated by a sample variance for each spatial vector $\mathbf{u}$ (or defining a tolerance region around it), although a robust transform of the sample variogram in the square–root absolute value scale provides better results, and enables the use of the variogram as an inferential tool. For instance, in the case of independent spatial data, the distributional properties of the sample variogram in the robust scale can be evaluated

and exploited to construct a test for the presence of spatial correlation, as described by Diblasi and Bowman (2001). Similarly, a test of goodness-of-fit assuming known parameters can be constructed, as discussed in Maglione and Diblasi (2004). However, beyond these rather special cases, diagnostic and inferential procedures based on the sample variogram have proved difficult to develop.

Smoothing methods have been also used for variogram estimation, examples including kernel methods or splines. The idea of using smoothers for variogram estimation is based on the interpretation of the (empirical) variogram cloud, that is, square–root differences with respect to distances, as a dispersion plot in an isotropic setting. Smooth variogram estimators can be used as pilots in obtaining least–squares estimators, but one should be aware of inappropriate use of smooth variograms as a means for broadly reflecting the underlying dependence of the data. One of the basic reasons is that the building blocks in the variogram cloud are not independent and they do not have the same variance. Usually, the smooth variogram estimators obtained with a kernel approach, are not valid variograms in the sense that they do not guarantee (or it cannot be proved for a general situation) a conditionally semidefinite negative behaviour (see Cressie (1993) for theoretical details for the variogram).

3.3 Some extensions.

Prediction. Krige's work in geostatistics was primarly motivated by the prediction of the likely yield of a mining operation over a spatial region, given the results of samples of ore extracted from a finite set of locations. Krige's prediction proposal (kriging) assumed that the trend in (1) was constant, and spatial prediction of the field value at a certain location was done considering the best linear unbiased predictor (minimum prediction error variance). The classical procedure has been extended in order to account for more complex trend structures, non–Gaussian responses, multivariate settings, external information, etc.

Prediction in high dimensions. As it was noticed in the Introduction, datasets with a huge amount of information are nowadays being collected, as in satellite image, for instance. Spatial statistics is such situations is challenging, mainly because of the big-n problem (the number of operations in solving a kriging prediction problem is $\mathcal{O}(n^3)$) and due to the fact that observation are taken in large spatial domains, and the spatial process considered therein may exhibit a non–stationary behaviour. Fixed rank kriging, introduced by Cressie and Johannesson (2008) tries to overcome this problem by defining a flexible family of non-stationary covariance functions is defined using a set of basis functions that is fixed in number. Dimensionality reduction seems to be the way to deal with this type of datasets, also from a Bayesian perspective. Wikle (2010) proposes representing the spatial process in terms of a latent Gaussian field with reduced dimension, and take advantage of the hierarchical computational tools.

Space–time geostatistics. From a pure probabilistic perspective, a space–time process can be regarded as a stochastic process in three dimensions, considering time as an extra coordinate. With this view, the dynamical behaviour cannot be appropriately capture, and although random fields theory has been developed in a quite general

context, the study of space–time process still needs a carefully view insight. There has been a large amount of recent works on space–time models, mostly focused on the description of the covariance structure, overcoming the simplistic assumption of separability, where the spatial and temporal dependence can be modelled independently, with no interaction between them. Different class on non–separable spatio–temporal covariances have been proposed by Cressie and Huang (1999), Genton (2007) and Ma(2003, 2008). Despite the great effort to introduce these classes of covariances, there is no a universal recipe on how and when use each of the proposals. That is the reason why some authors have developed different procedures for assessing separability in a spatio–temporal process, such as Fuentes (2005), Mitchel *et al.* (2005) and Crujeiras *et al.* (2010).

Simplifying hypothesis and the SPDE approach. Although separability has been noticed as a simplifying assumption in the analysis of spatio–temporal data, there also exist simplistic scenarios in the pure spatial context, such as those ones considering second–order stationarity or isotropy. Testing procedures for assessing hypothesis in the spatial case are not that abundant. One of the challenges of modern spatial statistics is to develop easy–to–use tools for exploring and assessing these features and construct a unified framework where all these characteristics (anisotropy, non stationarity, lack of separability) could be jointly assemble. Is it possible to obtain such a class of general spatio–temporal processes? And if so, is inference possible? The answer to this question seems to be in the stochastic partial differential equations (SPDEs). Lindgrenn *et al.* (2011) recover this idea from numerical analysis, where Gaussian random fields with Matérn covariance can be obtained as the solution of a fractional SPDE. Far from making inference more complex, the weak solution of such an SPDE provides a Gaussian field with sparse precision (inverse covariance) matrix, that is, a Gaussian Markov Random Field, allowing for approximate inference in a Bayesian framework.

4. SPATIAL POINT PROCESSES

A spatial point process is a stochastic process each of whose realizations consists of a finite or countably infinite set of points in the plane (events) creating a point pattern and that may have associated variables (marks). Classical examples of point patterns include the study of tree patterns in a forest or presence/absence and animal abundance. When building up a model for this kind of processes, one should try to reflect the physical mechanism generating the pattern, albeit in a simple manner. This mechanism may be simple (for instance, a random pattern) or may also reflect properties as inhibition or clustering among events.

A classical way for describing the physical mechanism is through the homogeneous Poisson process, where (a) the number of events in a certain spatial region follows a Poisson distribution with constant intensity, but proportional to the spatial area and (b) the number of events in two disjoint spatial regions are independent random variables. This second property is known as complete spatial randomness, providing a baseline to test against clustering or inhibition among events. Situations where a constant intensity is found in practice are not frequent, so extensions of this type of

processes are needed, being the simplest one the construction of inhomogeneous Poisson processes, that is, Poisson processes with non constant intensity. Nonparametric tests for assessing complete spatial randomness, and identifying clustering or inhibition, have been available for a long time. The classical techniques involve distance to nearest neighbour or nearest event, and are usually callibrated by Monte Carlo procedures. The intensity function in point processes reflects a first order structure. For explaning the second–order variability, the second–order intensity fucntion or the reduced second-moment measure (Ripley's K function) can be used.

A further extension of the Poisson processes assumes that the intensity function is also random, leading to a Cox–Process (Moller and Waagapetersen, 2004, 2007). In this kind of models, the spatial variation is included in a random structure through an underlying random field, namely, the intensity function, assuming independence conditionally on this field. Nonparametric estimation of varying intensity functions is quite problematic from a single realization of the spatial process, although some kernel–based methods have been proposed (see Diggle, 2003). Although usefull for an exploratory analysis, kernel intensity estimators may lose the stationarity property of the Cox process. Log–Gaussian Cox–Processes are a particular case of Cox–Processes, which consider the random intensity function $\lambda(\mathbf{s}) = \exp(Z(\mathbf{s}))$, being Z a Gaussian random field. With this construction, a highly flexible framework is achieved, although inferential procedures have proved difficult to handle. Usually, a Bayesian strategy is adopted, although Monte Carlo Markov Chain (MCMC) methods may result in painfully slow procedures. Nevertheless, with the integrated nested Laplace approximation from Rue *et al.* (2009).

As it has already been noticed, spatial point processes may present associated marks (marked point process) to each event. In this situation, the natural counterpart of independence is that the unmarked and marked point processes are independent. Schlather *et al.* (2004) proposed several summary statistics aimed at investigating departures from the independence hypothesis. In Chapter 20 of *Handbook of Spatial Statistics* (2010), a detailed description of modelling strategies for point processes is presentend.

5. LATTICE PROCESSES FOR DISCRETE SPATIAL VARIATION

In a natural flow of spatial statistical techniques, those related with spatial process with discrete variation are usually introduced after geostatistics and before point processes. In this short paper, we have deliberately leave the description of discrete spatial variation for the last part. The main reason is that, in practice, although our process may be continuously or randomly varying over a spatial region, observations are discrete in space or in space and time. We will briefly introduce what is a Gaussian Markov Random Field (GMRF) and try to explain why it is becoming a powerful tool in spatial analysis.

A GMRF is a Gaussian distributed random vector with some conditional independence properties. Conditional independence can be specified through a sparse precision matrix, which is computationally convenient. In his seminal paper, Besag (1974) specifies a GMRF through the full conditionals, although they cannot be specified completely arbitrarily as they must ensure a proper joint density (see Rue and

Held (2005) for details). These are the well known CAR (conditionally autoregressive) models. A more restrictive approach, although an alternative for CAR models, is simultaneously–autoregression (SAR) modelling (see Cressie, 1993 for details). One of the nice properties of GMRFs is that they can be easily integrated in a MCMC sampling scheme for doing Bayesian inference, although for some cases, the algorithms may run slowly. Applications of GMRF include a large list of fields such as structural time series analysis, longitudinal and survival data, graphical models, semiparametric regression and splines, image analysis and, of course, spatial and spatio–temporal statistics (see the monograph by Banerjee *et al.*, 2004). Specially in this situation, where a huge amount of data must be analyzed, MCMC scenarios may not provide efficient results.

An alternative lies in the idea of the sparse precision matrix of the GMRF and the use of integrated nested Laplace approximations (INLA) for approximate inference (Rue *et al.*, 2009). The same ideas may be extended for the analysis of geostatistical or point–processes, assuming an underlying Gaussian random field with Markov properties (for the intensity in point–processes or for reflecting the small scale variability in geostatistics). Continuously varying Gaussian processes can be linked with their discrete counterpart by the SPDE approach (solution and weak solution, respectively). Nowadays, INLA seems to give an answer to complex problems from a Bayesian perspective, providing approximate inference results.

REFERENCIAS

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, Boca Ratón.

Basu, S. and Reinsel, G.C. (1994) Regression models with spatially correlated errors. *Journal of the American Statistical Association* 89, 88–99.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36, 192-236.

Chiles, J.P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York.

Cressie, N. (1993) *Statistics for Spatial Data*. Wiley, New York.

Cressie, N. and Huang, H.C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94, 1330-1340.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B,* 70, 209-226.

Crujeiras, R.M. and Van Keilegom, I. (2010) Least squares estimation of nonlinear spatial trends. *Computational Statistics and Data Analysis*, 54, 452–465.

Diblasi, A. and Bowman, A.W. (2001) On the use of the variogram in checking independence in spatial data. *Biometrics*, 57, 211–218

Diggle, P. and Ribeiro, P.J. (2007) *Model-based Geostatistics*. Springer, New York.

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

Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998). Model based geostatistics. *Applied Statistics*, 47, 299-350.
Fuentes, M. (2005), Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference*, 136, 447-466.
Gelfand, A.E., Diggle, P., Fuentes, M. and Guttorp, P. Eds. (2010) *Handbook of Modern Spatial Methods*. Chapman & Hall, London.
Genton, M.G. (2007) Separable approximations of spacetime covariance matrices. *Environmetrics*, 18, 681-695.
Haas, T. (1996) Multivariate spatial prediction in the presence of non-linear trend and covariance non-stationarity. *Environmetrics*, 7, 145–165.
Lindgrenn, F., Rue, H. and Lindstrom, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach, *Journal of the Royal Statistical Society, Series B*, 73, 423–478.
Ma, C. (2003) Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference*, 116, 489-501.
Ma, C. (2008) Recent developments on the construction of spatio-temporal covariance models, *Stochastic Environmental Research and Risk Assessment*, 22, 39-47.
Mitchell, M., Genton, M.G. and Gumpertz, M. (2005) Testing for separability of spacetime covariances, *Environmetrics*, 16, 819-831.
Moller, J. and R. P. Waagepetersen (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall, Boca Ratón.
Moller, J. and R. P. Waagepetersen (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34, 643–711.
Neumann, S.P. and Jacobson, E.A. (1984) Analysis of nonintrinsic spatial varibility by residual kriging with application to regional groundwater levels. *Mathematical Geology*, 16, 499–521.
Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall, London.
Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71, 318-392.
Schlather, M.S., Riberiro, P.J. and Diggle, P. (2004). Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society, Series B*, 66, 79-93.
Wikle, C.K. and Van Hooten, M.B. (2010) A general science–based framework for spatio–temporal dynamical models. *Test*, 19, 417–451.