Astronomical Data Analysis Software and Systems IV
ASP Conference Series, Vol. 77, 1995
R. A. Shaw, H. E. Payne, and J. J. E. Hayes, eds.

Spatial Models and Spatial Statistics for Astronomical Data

L. Pásztor

MTA TAKI, H-1022 Budapest Herman Ottó út 15, Hungary

L. V. Tóth

Dept. of Astr., Eötvös Univ., H-1083 Budapest Ludovika tér 2, Hungary

Abstract. A statistical model is a convenient conceptual representation of an observed phenomenon. A statistical model represents the observations in term of random variables which can then be used for description, estimation, interpretation, and prediction based on Probability Theory. Interest in statistical methodology is increasing rapidly in the astronomical community. New questions arising from old and new technologies require new statistical models, and many of the new problems are spatial in nature. Spatial statistics is still a young discipline, the application of which is not yet widespread in the astronomical community.

1. The Formalization of the General Spatial Model

Consider $\{Z(t) : t \in T\}$; where $T \subseteq \Re^d$. Here T is the index set, $Z(.)$ is the spatial process, $\{z(t) : t \in T\}$ is a realization of the process. In the present paper we give a brief overview on the most important spatial statistical models, to illustrate the range of problems that can be addressed and the wide applicability of spatial statistical models in astronomy.

2. Point Processes

A **usual spatial point process** is defined as $Z(t) = 1; \forall t \in T$ (i.e., the index set is the points of $\Re^d$) or $Z(A) = N(A)$ [the number of points within A]; $A \subseteq T$ (i.e., the index set is the units of $\Re^d$), where both $Z(.)$ and T are random. First- and second-order properties of a spatial point process are the intensity function: $\lambda(t) = lim_{|dt| \to 0}\{E[N(dt)]/ \mid dt \mid\}$; and the second-order intensity function: $\lambda_2(t,u) = lim_{|dt|,|du| \to 0}\{E[N(dt)N(du)]/ \mid dt \mid\mid du \mid\}$. Spatial point processes are the mathematical models producing point patterns as their realization. A number of processes are available for modeling the patterns that arise in nature:

Complete Spatial Randomness process (CSR; the white noise of spatial point processes) $\equiv$ homogeneous Poisson process (HPP). The number of points for $\forall A$ has a Poisson distribution with mean $\lambda \mid A \mid$; counts in disjoint sets are independent.

Processes with tendency to produce aggregated patterns. For an inhomogeneous Poisson process (IPP) the number of points for $\forall A$ has a Poisson distribution with mean $\int_A \lambda(t)dt$. Counts in disjoint sets are independent. For a Cox process (CP; doubly stochastic point process) $\Lambda(t)$ is a non-negative valued stochastic process. Conditional on $\Lambda(x) = \lambda(x)$, the events form an IPP with intensity function $\lambda(x)$. For a Poisson cluster process (PCP; Neyman-Scott process) parent events form an IPP. Each parent produces a random number of offspring, realized independently according to a discrete probability distribution. The position of the offspring relative to their parents are independently distributed according to a d-dimensional density function. The final process is composed of the superposition of offspring only. Multi-generation process is the generalization of PCP, where offspring are parents of the next generation.

Processes with tendency to produce regular patterns. For a simple inhibition processes (SIP; hard core processes) no two events may be located within a minimum permissible distance, d, of each other. (Matern models, Matern-Stoyan, Matern-Bartlett, simple sequential inhibition models are examples.) The Markov point process (MPP) is a more flexible framework for modeling inhibition processes. $\lambda(u \mid t_1, t_2, ..., t_i, ..; t_i \in A\backslash\{u\}) = \lambda(u \mid t_1, t_2, ..., t_i, ..; t_i \in b(u,d)\backslash\{u\})$, where $b(u,d)$ is the closed ball of radius d centered at u (Strauss process, Pair-potential Markov point process, Gibbs process).

Multivariate spatial point processes. Defined as $Z(t) = 1, 2, 3, ..., m; \forall t \in T$ (i.e., the index set is the points of $\Re^d$) or $Z(A) = N_i(A)$ [number of i points within A]; $A \subseteq T$ (i.e., the index set is the units of $\Re^d$), where both $Z_i(.)$ and T are random. The m univariate spatial point processes are the components of the multivariate process, which is thus characterized by m intensity functions and $m(m+1)/2$ second-order intensity functions. The terminology reflects the components of the process (e.g., bivariate Cox process).

Examples of applicability in astronomy include: (1) revealing regularity in the spatial distribution of point-like objects, (2) identification of important scales in the spatial distribution of point-like objects, (3) stellar statistics (deriving distributions, testing of predicted distribution functions, identification of clusters and associations of stars, search for wide binaries and multiple systems), and (4) cosmological problems (testing of predicted distribution functions, identification of galaxy clusters, voids, etc.).

3. Theory of Regionalized Variables (Geostatistics)

The spatial index t varies continuously throughout a fixed subset T of a d-dimensional Euclidean space. Term "regionalized" was introduced in order to emphasize the continuous spatial nature of the index set T. The prefix "geo" reflects the fact that the theory's roots are in geographical and geological applications. Random processes are usually characterized by their moment measures. In geostatistics, "semivariogram" plays a crucial role. If $var(Z(t_1) - Z(t_2)) =$

$2\gamma(t_1 - t_2)$ for $\forall t_1, t_2 \in T$; γ is called semivariogram. If $E(Z(t)) = m$ for $\forall t \in T$ and $\gamma(.)$ exist, $Z(.)$ is intrinsically stationary. Semivariogram is conditional negative-definite. If $Z(.)$ is second-order stationary $2\gamma(d) = 2(C(0) - C(d))$. Linear, spherical, and exponential models are simple isotropic (semi)variogram.

The most important application of the (semi)variogram is "kriging," a stochastic spatial interpolation method which depends on the second-order properties of the process. The principal aim of kriging is to provide accurate spatial predictions from observed data. Kriging techniques are all related and refined versions of the weighted moving average originally used by Krige (1951) and based on the simple linear model: $Z(t) = \sum_{i=1}^{n} w_i Z(t_i)$, where $\sum_{i=1}^{n} w_i = 1$. Kriging provides optimum prediction in a sense of minimizing mean-squared prediction error, and also provides the estimation.A useful decomposition is $Z(.) \equiv \mu(.) + W(.) + \eta(.) + \epsilon(.)$, where $\mu(.) \equiv E(Z(.))$ is the large-scale variation, $W(.)$ is the smooth small-scale variation, $\eta(.)$ is the micro-scale variation, $\epsilon(.)$ is the measurement error. These models are widely applied in geosciences.

A number of astronomical applications of the method come to mind: (1) the creation of contour and/or surface maps in the case of incompletely sampled maps in extended radio surveys, (2) testing for completeness in sampling (whether the expected structure is revealed as spiral or filamentary), (3) testing whether resolution is achieved (in the cores of galaxies), (4) the creation of maps with resolution higher than the physical resolution of the observation (interpolations arising from the co-addition of separate sky coverage by *IRAS* or *ISO*), and (5) interpolations to reach a higher virtual resolution for comparisons (e.g., *IRAS* 12 and 100 μm images).

4. Further Cases of the General Model

Spatial model on lattices. The index set T is a countable collection of regularly or irregularly scattered spatial sites and these sites are supplemented with a neighborhood structure. Neighborhood structure is generally modeled either by the connectivity matrix (C is an $n \times n$ matrix, $c_{ij} = 1$ if sites i and j are juxtaposed, $c_{ij} = 0$ if not; n is the number of sites) or by a graph-theoretic formalism (the sites become vertices, which are connected with edges for contiguous objects). Examples for realizations of lattice processes in 2-D are spot maps, mosaics, and digital images. The most important application of lattice models is statistical modeling of spatial images, which is widespread in astronomical image processing (restoration, segmentation, classification, reconstruction, etc.).

Fuzzy sets theory. The elements of T are random sets. The premise of the approach that all the data are imprecise, even after they have been observed.

Multivariate spatial statistics. $Z(t)$ is multidimensional. An example of multivariate spatial statistics is provided by multiband image processing. A generalization of the univariate spatial statistical methods is provided by cokriging, where spatial prediction of a variable is carried out with the aid of another.

Examples of applicability to astronomy include: (1) 2-D classification of objects by their shape on images (e.g., star, galaxy identification on CCD or photographic images), (2) cloud identification from coordinate-velocity "data cubes" (e.g., radio spectroscopic observations), and (3) any advanced image processing technique, like maximum entropy or deconvolution (e.g., maximum correlation method in "HIRES" *IRAS* data processing at IPAC).

Acknowledgments. This research was partially supported by the Hungarian State Research Found (Grant No. OTKA-F 4239). L. Pásztor is grateful to ADASS and the Hungarian State Research Found for the travel grants.

Basic References on the Astronomical Applications of Spatial Statistics

Bahcall, J. N., & Soneira, R. M. 1981, ApJ, 246, 122

Bahcall, J. N., Jones, B. F., & Ratnatunga, K. U. 1986, ApJ, 60,939

Bucciarelli, B., Lattanzi, M. G., & Taff, L. G. 1993, ApJS, 84, 91

Cliff, A. D., & Ord, J. K., 1973, Spatial Autocorrelation (London, Pion)

Cressie, N. A. C. 1991, Statistics for Spatial Data (New York, Wiley)

Diggle, P. J. 1983, Statistical Analysis of Spatial Point Patterns (London, Academic Press)

Getis, A., & Boots, B. 1978, Models of Spatial Processes (Cambridge, Cambridge University Press)

Huang, J. S., & Shieh, W. R. 1990, Pattern Recognition, 23, 147

Journel, A. G., & Huijbregts, Ch. J. 1978, Mining Geostatistics (London, Academic Press)

Matheron, G. 1965, La Theorie des Variables Regionalisées et ses Applications (Paris, Masson)

Molina, R., Olmo, A., Perea, J., & Ripley, B. D. 1992, AJ, 103, 666

Pásztor, L., Tóth, L. V., & Balázs, L. G. 1993, A&A, 268, 108

Pásztor, L. 1993, in Astronomical Data Analysis Software and Systems II, ASP Conf. Ser., Vol. 52, eds. R. J. Hanisch, R. J. V. Brissenden, & J. Barnes (San Francisco, ASP), p. 7

Pásztor, L. 1994, in Astronomical Data Analysis Software and Systems III, ASP Conf. Ser., Vol. 61, eds. D. R. Crabtree, R. J. Hanisch, & J. Barnes (San Francisco, ASP), p. 253pasztorl2

Peebles, P. J. E. 1973, ApJ, 185, 413

Prusti, T., Adorf, H-M., & Meurs, E. J. A. 1991, ESO Sci. Preprint 797

Ripley, B. D. 1981, Spatial Statistics (New York, Wiley)

Ripley, B. D. 1988, Statistical Inference for Spatial Processes (Cambridge, Cambridge University Press)

Upton, G., & Fingleton, B. 1985, Spatial Data Analysis by Example, Vols I–II (New York, Wiley)

Wilson, C. D. 1991, AJ, 101, 1663