

Review of the statistical package S+Spatialstats.

A. Stein and P. Droogers, Wageningen Agricultural, University, P.O. Box 37,
6700 AA Wageningen, The Netherlands

Technical Information

S+ Spatialstats, version 1.0. Mathsoft, Inc. Seattle, Washington, importeur: Statsci Europe, Osney House, Mill Street, Oxford OX2 0JX, United Kingdom. Price £ 995,- Academic and quantum reduction available upon request.

The package runs on

- a. A 80386 or higher processor running Windows 3.1, Windows NT or Windows 95, math coprocessor, VGA or better monitor, 8MB Ram and a minimum of 30 MB free disk space.
- b. A Sun sparc, IBM RS/6000, HP 9000-700, Silicon Graphics IRIS, DECStation or DECAAlpha/OSF, with 12MB of RAM, 60 MB free disk space and an X-windows graphics system.

1. Introduction

Recently the spatial statistical software package *Spatialstats* as an extension to the S-plus package has appeared. S-plus is a statistical package, developed originally in Seattle, in the late nineteen eighties. The package is currently wide-spread, mainly because of its excellent analytical and graphical possibilities. The basis forms the generalized linear model, but it has developed as fully professional statistical package. Strong points include the easy accessible and highly readable graphical output and an object-oriented library. This library can be extended optionally to include user-defined or commercial modules. Recently available modules include a Wavelet package, the DOX package to design experiments and the package *Spatialstats*. Here we discuss the library *Spatialstats*.

Spatial data are currently becoming widely available. The reason is an increasing interest and importance of environmental data, excellent computer facilities to combine and present spatial data and important scientific

development in this field. Following Cressie (1991) a distinction is made into geostatistics, lattice data and spatial point patterns. The library *Spatialstatistics* contains functions and procedures on these three topics, and one on visualisation. As the package focuses on analysis of spatial data, an interface with a geographical information system is crucial. It contains a similar functionality as the other well-known modules of S-Plus. A relatively easy interface as soon as one is familiar with it, and direct access to data and routines.

2. Overview and functionality of the package.

The package starts with exploiting the different graphical routines as they are available in S-plus for both linear and planar data. Linear data, such as data collected along a line in the field, are to a certain extent comparable with time series data, whereas planar data (data collected in a 2-dimensional space) can of course be analysed using ordinary regression, ANOVA, GLM and other standard statistical procedures. Clustering of data, box plots and trend models are accessible and standard data sets are provided to apply them to and to become familiar with the different routines. An extension is included towards the generalized additive model and to splines. Users less familiar with S-plus may find it difficult to become familiar with the data structure and the way most efficiently their own data can be analysed. Sections on this topic are available from the S-plus manual. The *Spatialstats* manual would though have benefitted from inclusion of those parts of the S-plus manual which cover reading and writing of files. The chapter on visualization ends with a brief introduction on spatial point patterns.

The next chapter is on geostatistical data and concentrates on the variogram, kriging and spatial simulations. Geostatistical data are continuous data, tied to their location in space. Examples are the content of a pollutant in an area, the daily temperature in a country or the population density in a region. The variogram is currently a common tool to describe the spatial similarity between data as a function of the distance between locations. Observations close to each other are more similar than observations at a larger distance. This form of spatial dependence can be analysed and described using the variogram. A special form of spatial dependence applies to anisotropy: spatial dependence in one direction can be different from that in another direction - think about vegetation along a highway, with strong similarities parallel to the road, but large differences perpendicular to it. Such an

anisotropy analysis is possible with *Spatialstats*. Model fitting of the variogram is included with five basic models, whereas other models such as the whole effect model can relatively easily be included by the user. Also a weighted non-linear model fitting is available. A nested model is not a standard option, but can probably be analysed with the common statistical routines. The way different models can be estimated and compared is not very user-friendly: they have to be specified in some detail each time an analysis is required. The interpolation technique Kriging applies the variogram to obtain values at unvisited locations. That is in particular useful when a map has to be created. Kriging contains two options: ordinary kriging, i.e. kriging using the variogram, and universal kriging, where the latter interpolates the residuals after trend fitting to which the value of the trend at prediction points is added. Universal kriging therefore allows to deal with non-stationarity in the data as well. Also plots of the kriging standard deviation can be obtained. Apart from interpolation, spatial variability studies currently focus on spatial simulation. The basic idea of geostatistics is the random field theory, and often one wishes to visualize this field. In fact, kriging is a smoother, being connected with splines, whereas sometimes the actual spatial variability is required, for example to study error propagation in agricultural models. The only simulation technique included is the Gaussian random field with exponential errors. As concerns functionality, therefore, S+-spatial statistics has less power than, for example, the geostatistically oriented GS-LIB package (Deutsch and Journel, 1992) or the ISATIS package (Geovariances, 1997). However, it is integrated in a much wider statistical environment.

Chapter 3, on lattice data, contains sections on neighbours, on testing of lattice data for spatial autocorrelation, on spatial regression models and on simulation of lattice data. The procedures are again transparent, can simply be applied and have a clear output and summary. It may be questioned, as with the geostatistical procedures, whether the functionality will suffice for all practical case studies. For an exploratory spatial analysis, though, it is definitely useful.

Chapter 4 is on spatial point patterns, and focuses on measures of spatial randomness, examination of 1st and 2nd order properties and simulation of spatial point patterns. Example of point patterns are the locations of a single tree species in a forest or the location of cafeterias in a large city. Commonly the so-called K-function is used: this function models the probability as a function of the distance r , that from an arbitrary point of the process another point is encountered within a distance r . In *Spatialstats* the K-function is available, but no functions are available to estimate spatial

interaction. Five basic patterns can be simulated: the Poisson and the binomial process, which are both entirely random processes, the Sequential Spatial Inhibition (SSI) process which is pattern with a close similarity to a regular grid, the Strauss process, being a hard-core process, that is a process where around each point an empty space exists with probability 1, and the cluster process which produces a Poisson parent-daughter process. For most practical purposes this will be sufficient. However, recent research on the effects of boundaries of the window on for example estimation of the K-function is not easy to program in a standard way. Also the package would have benefitted from a more easy accessible module to generate standard K-functions with confidence bounds.

The fifth chapter covers the important interface with the Geographic Information System (GIS) Arc/Info. A GIS usually contains the data that have to be analysed and it has the powerful display and visualization options. Many GIS packages are currently available. The *Spatialstats* includes an interface with the GIS package Arc/Info, being the current standard for GIS applications. One of the major drawbacks of many of the GIS packages is that they remain very descriptive: few statistical tools are available in any of these. Therefore, GIS packages have not been the topic of much statistical research until now. For example, Arc/Info allows to estimate means and standard deviations, but further statistical routines are not available. It is not at all clear what occurs in Arc/Info's geostatistical library, and further spatial statistical routines are not yet available. The interface now provided with S+Spatialstats is therefore an excellent and an extremely useful tool. It includes examples of analysing point, polygon and grid coverages. For point coverages attention focuses on geostatistics and nearest neighbour techniques, for polygon coverages on irregular lattice procedures and for grid coverages on geostatistical and regular lattice analyses. Examples include an exploratory data analysis and a geostatistical analysis. This interface really gives power to both the statistical package and modern information systems.

3. Example

We applied the S-Plus Spatialstats program on patterns of pores in the soil. Pores are created by roots and are important to quantify movement of possibly contaminated moisture towards the ground water. During wet times the moisture moves down whereas during a dry period the plant (crop) on the land may benefit from pores by sucking moisture from the ground water. The patterns are usually influenced by different soil types and by different forms of land use. A quantitative description of pores is what is currently aimed at, and spatial statistical procedures can assist us in this respect. In a recent experiment (Stein et al., in press) we analysed the patterns for different combinations of land use and soil types. To visualise the patterns blue coloured moisture was poured over the soil. After a few hours, horizontal cross-sections were made and pictures were taken with a camera at different

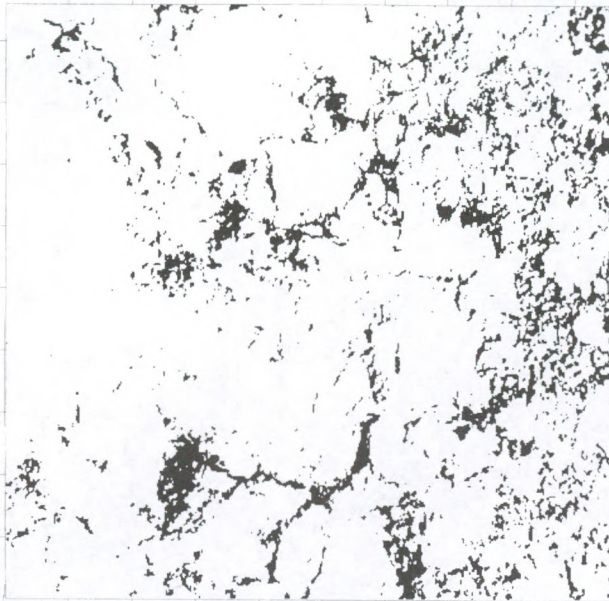


Fig. 1. Methylene-blue coloured pattern of stains in soils at a depth of 0.04 m. Window size is 0.2×0.2 m.

depths. The pictures were digitised and stored in a Geographical Information System (fig. 1).

We distinguished three types of stains on the basis of its morphology parameter $\rho = \text{area/perimeter}^2$: cracks ($\rho < 0.015$), which are longitudinal volumes of air (or moist) caused by cracking of the soil, vughs ($0.015 \leq \rho < 0.04$), which are a type pores of an often irregular shape but with an often unknown origin and ordinary pores ($\rho \geq 0.04$), being caused by roots. On the patterns we applied a spatial analysis by calculating the $L(t)$ function, defined as the square root of Ripley's $K(t)$ function (fig. 2). The $L(t)$ function for this particular example is based upon appr. 700 points on the display. The presentation in S-plus fits the observed point process for the three different types of soil volumes. For the appr. 30 cracks the observed pattern is close to a Poisson process. We notice though a horizontal behaviour of the $K(r)$ function close to the origin, usually indicating a hard core. That can be explained by the positive size of cracks, which were reduced to points for the spatial analysis. Also for vughs (more than 100 occurrences) the hypothesis that they follow a Poisson process is not rejected, although the confidence bounds are close to the $K(r)$ function. For pores, though, the hypothesis is rejected, and a cluster

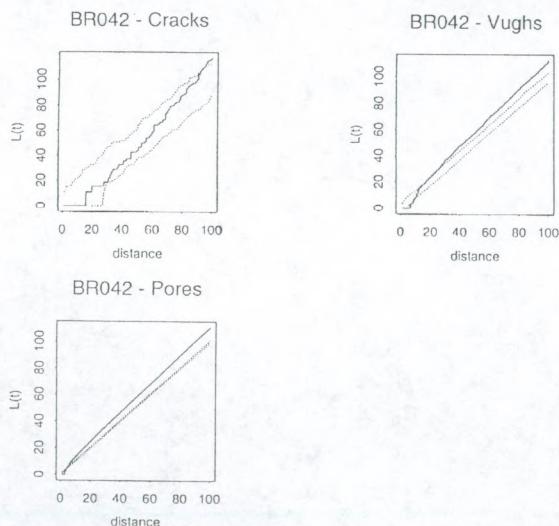


Fig. 2. The $L(t)$ function for cracks, vughs and pores (solid lines). Included are the 95% confidence bounds (dotted lines). Distance is in cm.

process is much more likely to exist: plant roots tend to grow in the vicinity of cracks in the soil.

4. Conclusion

We were in general quite enthusiastic about the Spatialstats module of S-plus. The package is very user friendly. Users familiar with S-Plus will not find any difficulty in using it, documentation is well focused and on-line help is available. In all, we found the package most useful particularly at a relatively standard level. It provides a bridge between modern statistical methods and information systems. The procedures are easily applied, and have a clear, well-ordered output without much redundancy. We surmise though that its functionality will lag behind to more specialised packages rather soon although own routines can be added with Fortran, Pascal and C-interfacing. But that could mainly be a problem for scientists, whereas for many practitioners of spatial statistics and quantitative-oriented GIS users the combination will be extremely useful and serve a broad range of practical purposes.

5. References

- Cressie, N. 1991. *Statistics for Spatial Data*. New York: Wiley.
- Deutsch, C.V., and A.G. Journel. 1992. *GSLIB - Geostatistical software library and user's guide*. Oxford University Press, New York.
- Geostatistics. 1996. *ISATIS, the geostatistical key*. Avon (F): Geostatistics.
- Stein, A., P. Droogers and H. Booltink. *In press*. Point processes and random sets for analysing patterns of methylene blue coloured soil. *Soil Technology*.