EnviroInfo 2004 (Geneva)
Sh@ring – EnviroInfo 2004
 ISBN: 28 29 30 275-3

Advanced analysis and modelling tools for spatial environmental data. Case study: indoor radon data in Switzerland

Mikhail Kanevski [1], Michel Maignan [1] and Georges Piller [2]

Abstract

The present work deals with development and adaptation of advanced geostatistical models and machine learning algorithms (statistical learning theory – Support Vector Machines) for comprehensive analysis and decision-oriented modelling of environmental spatial data. The real case study is based on indoor radon data. The inherent high variability at different spatial scales of noisy indoor radon measurements coupled with the heavy clustering effect of houses locations make this dataset an excellent candidate to assess the feasibility of traditional and advanced models, trend and risk mapping at local and regional scales.

1. Introduction

General methodology of spatial data analysis and modelling includes comprehensive exploratory data analysis, qualitative and quantitative description of monitoring networks, analysis and modelling of spatial anisotropic correlation structures, spatial predictions and simulations using geostatistical models and machine learning algorithms. In the present study this methodology is applied for the real case study of indoor radon data.

The main attention in the present study is paid to the quantification of monitoring network clustering, modelling of spatial correlations, conditional stochastic simulations and risk mapping and application of statistical learning theory for regional classification. Regularised variography helps to improve the visibility of spatial structures on variograms of both raw and transformed data.

Once modelled, the spatial correlation structures of data (variograms) are used to produce traditional maps with ordinary kriging. Due to the spatial variability of data, kriging induce a strong smoothing effect that make this approach only suitable for

[1] Faculty of Geosciences and Environment, University of Lausanne, Dorigny, 1015 Lausanne
Michel.Maignan@img.unil.ch
[2] Swiss Federal Office of Public Health, CH-3003 Bern

data representation. In order to reproduce very high variability and uncertainty of data, advanced geostatistical approaches based on different simulation models were applied. Spatial "Monte Carlo" models generate equally probable simulated maps of concentrations that respect both the spatial structures and distributions of data. By reproducing the variability of data, simulations avoid the well known smoothing effect of kriging and are thus able to produce risk maps of probability to be above a decision level.

Data driven approach based on artificial neural networks, such as general regression neural network or multi-layer perceptron, were also applied to these data in order to reveal trends or spatial patterns. In ther present paper the recent developments based on support vector machines (SVM) are efficiently applied for regional classification of spatial data. It is shown that high flexibility and robustness of SVM are valuable properties for the indoor radon data modelling.

The complete study and results were obtained with Geostat Office, a scientific software package for geostatistical and machine learning analysis of data (Kanevski and Maignan 2004).

2. Methodology

General methodology of analysis and modelling of spatially distributed data consists of several important steps:

- Comprehensive exploratory data analysis, analysis and modelling of distributions, visualisation of data etc.
- Analysis and quantification of monitoring network (distribution of measurement points in space). Estimates of topological, statistical and fractal measures of clustering. Estimation of representative univariate/declsutered distributions
- Moving window statistics. First step into spatial statistics.
- Exploratory variography. Modelling of anisotropic spatial correlations. Description of spatial structures in data. Variography can be used as an exploratory tool to describe the quality of machine learning algorithms.
- Spatial predictions (regression, interpolation), classifications. Mapping of predictions and corresponding uncertainties. Both geostatistical models and machine learning algorithms can be used.
- Modelling of local distribution functions. Modelling uncertainty around unknown values. Risk mapping.
- Conditional stochastic simulations – modelling and sampling of joint distributions. Generation of many equally probable realisations of random function. This is the most advanced modelling tool for the analysis and modelling of spatial uncertainty and variability. Risk mapping.
- Decision oriented mapping using Geographical Information Systems.

For spatial predictions and simulations different approaches can be used, depending on the quality and quantity of data and objectives of the study: family of geostatistical models (variants of kriging), machine learning algorithms (neural networks of different architectures), statistical learning theory (support vector machines, support vector regression). Details of modelling approaches along with case studies can be found in (Kanevski and Maignan 2004).

3. Data Description and Monitoring Network Analysis

For the present study a part of complete data set was extracted. The distribution of measurement points is given in Figure 1. Monitoring network represents complicated structure of distribution of houses in the selected region. The univariate distribution of selected data is positively skewed (coefficient of skewness is about 4) and varies between 1 and 1000 Bq/m^3.

The measuring stations of an environmental monitoring network are usually spatially distributed in an inhomogeneous manner, it means that in some regions there are more measurements, in some less and monitoring network is clustered. There are many circumstances why it happens: geography of the region, risk considerations (more polluted and dangerous regions are sampled more frequently), and others.

The problem of network homogeneity (clustering) is closely connected with global estimations, univariate statistics parameters (e.g. mean and variance), the possibility of monitoring network to detect phenomena under study (fractal or dimensional resolution of the monitoring network), etc. For the global estimations (e.g. mean value), recovery procedure is related to the declustering, i.e. weighting of raw data before statistical analysis.

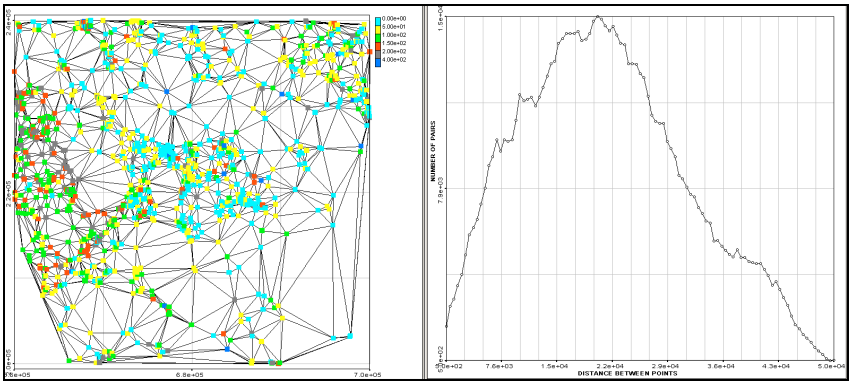

Figure 1.
Data post plot and visualisation of monitoring network (left). Distribution of distances between measurement points (right).

In general, the spatial inhomogeneity of points can be characterized by different geometrical/topological (Figure 1, right), statistical, e.g. Morishita index (Figure 2) and fractal dimension (Figures 2, left).

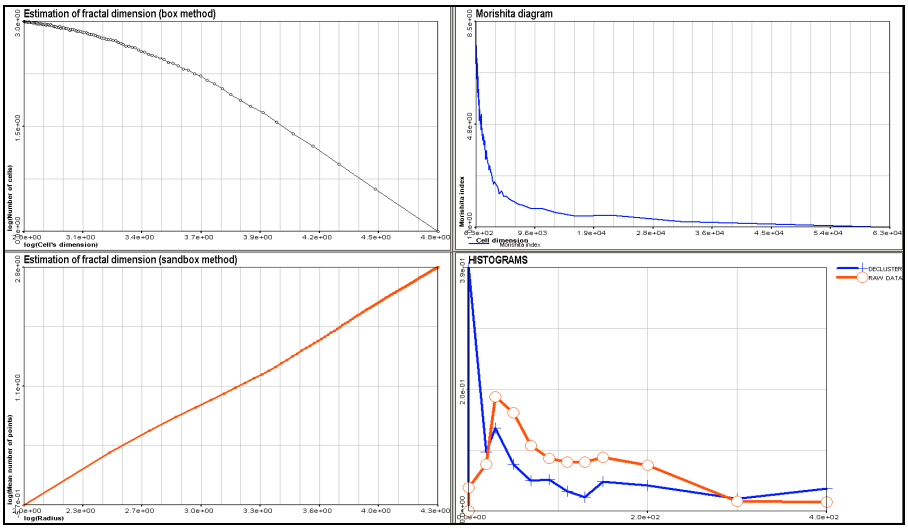

Figure 2.
Calculation of fractal dimensions of monitoring network using box and sandbox counting methods (left). Morishita diagram up-right). Influence of the clustering on histogram estimates – raw data and declustered data (down right).

Fractal analysis of current data demonstrates clustering at two different scales. It can reflect clustering of houses within the commune (local scale about 3000 m) and clustering of communes at regional scale. Spatial declustering procedures should be applied to estimate distribution and global parameters. Several declustering procedures are widely used for spatial data: random declustering, cell declustering, etc. (Kanevski and Maignan 2004). The significance of clustering on the estimation of histogram is given in Figure 2.

4. Geostatistical Modelling

Geostatistics is based on statistical treatment of data: measurements are considered as realizations of a random function. The problem is that usually only one realization of the random function is available (measurements). Therefore, in order to make statistical inference under such conditions some modelling hypotheses should be accepted: ergodicity, intrinsic hypotheses.

These hypotheses mean that different regions of the study are statistically similar (statistical similarity is described by corresponding measures: covariance function, variogram). Under these conditions, different parts of the region can be considered as different realisations of the random function.

4.1 Variography

Variography or modelling of spatial correlation structures is one of the most important step in any geostatistical analysis.

There are several measures describing spatial continuity used to quantify spatial correlations (Goovaerts, 1997; Deutsch and Journel, 1997). The basic idea is to compare similarity/dissimilarity of spatially separated points.

Variogram is the basic tool of the spatial structural analysis so called variography. Variogram (under the intrinsic hypotheses the variogram depends only on separation vector $\mathbf{h}$) can be estimated by

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \left(Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h}) \right)^2$$

where $\boldsymbol{Z}(\mathbf{x})$ – random function, $N(\mathbf{h})$ number of pairs separated by vector $\mathbf{h}$.

Measurement points are discrete points distributed in space. In order to compute correlation measures (covariance function, variogram, etc.) some quantization of lag and angle space should be used. In spatial data analysis correlations can have geometric (different correlations in different directions) or zonal anisotropic behaviours.

A variogram, being a difference of squared function, is sensitive to outliers, i.e. very high values of data. Therefore several robust variogram models have been proposed. An example of experimental variograms for raw radon is presented in Figure 3. In case of smooth and not noisy random function, variogram should start from zero at small lags and increase until correlations disappear after some separation distance. If data are not spatially correlated variogram fluctuates around constant level of a priori variance of data. Erratic behaviour of variogram for indoor radon data is rather typical and reflects very high variability at small scales and measurement errors. Other reasons could be outliers, skewness of distribution, clustering. Some of these effects can be reduced by applying robust spatial correlation measures, or after nonlinear transformation of raw data, or using nonregular variography and regularization technique, etc. In Figure 3 directional variograms were estimated also after nonlinear transformation from raw data to the Nscore values, distributed N(0,1). Such transformation is widely used in simulations. In case of raw data (Figure 3, left) it is impossible to recognize any spatial correlation. After Nscore transformation (Figure 3, right) some anisotropic correlations can be recognized.

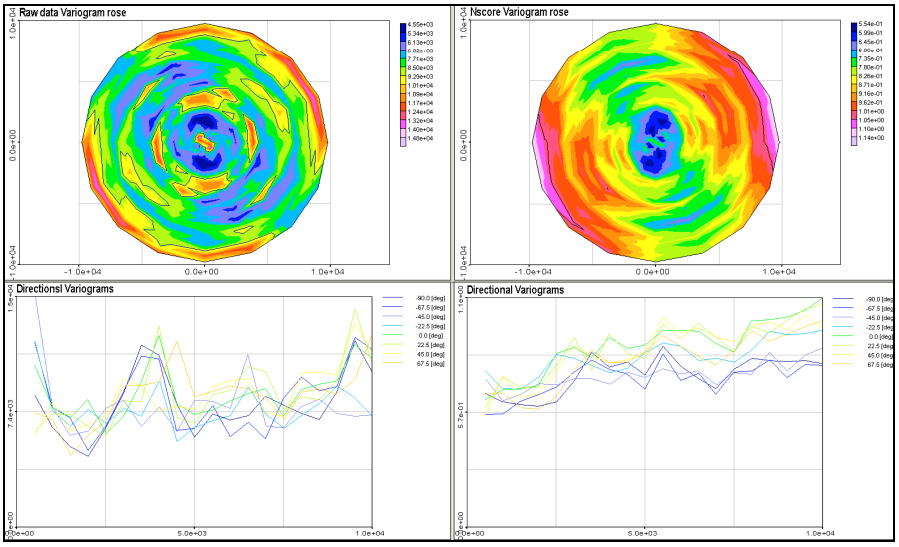

Figure 3
Variogram roses (up) and directional variograms (down) versus lag distance. Variograms of raw data– left; variograms of Nscore values – right.

5. Risk mapping. Simulations

The next step of data analysis methodology deals with spatial predictions. The most widely used geostatistical model for spatial prediction belongs to the family of kriging models: simple kriging, ordinary kriging, universal kriging, etc. Kriging, as a spatial regression model gives rise to the smoothed solutions. Such solutions do not reproduce spatial variability described, e.g. by variogram. Moreover, such solutions, giving "the best" maps in some sense can not be used in risk analysis process, when we are interested in extreme events and not only in averaged (regression) solutions.

Stochastic simulations are an appropriate modelling tools able to reproduce basic statistical properties of the random field including variability. There are several basic simulation models widely used in practice: parametric, based on multigaussian hypotheses; nonparametric, using indicator transforms; simulated annealing.

Traditionally stochastic simulations are trying to reproduce representative/declustered distribution of data, spatial correlations described by variograms and to be conditional, i.e. at the measurement points all realizations equals to measured data (if they are measured without errors). Below only one possible simulation model is considered, conditional sequential Gaussian simulations.

Let us consider a vector-valued random variable $\boldsymbol{Z}=(Z_1, Z_2, \ldots, Z_N)^T$ for which a realisation of the subvector $(\boldsymbol{Z_1, Z_2, \ldots, Z_M})^T$ is known and equal to $(z_1, z_2, \ldots, z_M)^T$ ($0 \le M < N$). The distribution of the vector $\boldsymbol{Z}$ conditional on $Z_i = z_i$ ($I=1,2,\ldots,M$) can be factorised in the form (Chiles and Delfiner 1999)

$$\Pr\{z_{M+1} \le Z_{M+1} < z_{M+1} + dz_{M+1}, \ldots z_{N+1} \le Z_{N+1} < z_{N+1} + dz_{N+1} \,|\, z_1, \ldots z_M\}$$
$$= \Pr\{z_{M+1} \le Z_{M+1} < z_{M+1} + dz_{M+1} \,|\, z_1, \ldots z_M\}$$
$$* \Pr\{z_{M+2} \le Z_{M+2} < z_{M+2} + dz_{M+2} \,|\, z_1, \ldots z_M, z_{M+1}\}$$
$$\bullet$$
$$* \Pr\{z_N \le Z_N < z_N + dz_N \,|\, z_1, \ldots z_M, z_{M+1}, \ldots z_{N-1}\}$$

Using this factorisation random vector $\boldsymbol{Z}$ can be simulated sequentially by randomly selecting $\boldsymbol{Z_i}$ from the conditional distribution $\Pr\{\boldsymbol{Z}_i < z_i \mid z_1, z_2, \ldots z_{i-1}\}$ for $i= M+1, \ldots, N$. and including the outcome z_i in the conditioning data set for the next step. This procedure of decomposition of a joint pdf into the product of conditional pdfs is very general and can be used for spatial random functions as well. Sequential simulation is a theoretically simple and general simulation algorithm, which is conditional by construction.

From the simulation model many equally probable maps of the random function can be generated (Figure 4). Similarity and dissimilarity between these realizations describes spatial uncertainty and variability. Post processing of such realizations gives very complete information for decision making process, including so-called risk maps – maps of probabilities to be above/below decision levels (see Figure 5).

Sequential Gaussian algorithms follows several steps: Nscore transform, checking of multinormality (usually two point statistics only), variography of Nscore values, sequential simulations by randomly visiting nodes of the simulation grid and using kriging to estimate mean and variance of multigaussian function, back transform to original data.

For the present case study sequential stochastic simulation model was developed and 150 realizations were generated. Four selected realizations are presented in Figure 4. An important summary information can be obtained after post processing of simulations: averaged map, variance map, risk maps – maps of probabilities to be above some predefined levels; the most probable maps, etc.

6. Statistical Learning Theory. Support Vector Machines

In recent years there has been an explosive growth in the development of adaptive data driven methods. Geostatistics, in general, is a model-dependent approach based on exploratory analysis and modelling of spatial correlation structures. Efficient data-driven contemporary approach is based on statistical learning theory (Vapnik

1998). In general data-driven learning method is an algorithm that predicts unknown mapping (classification, regression, density estimation) between inputs and outputs from the available data and a priori knowledge if available.

In the present study Support Vector Machines are used as a universal constructive learning procedure based on the statistical learning theory. Recently several research groups have shown excellent performance of SVMs on different problems of classification and regression. Detailed description of the SVM theory can be found in (Vapnik 1998) and application to spatial data in (Kanevski and Maignan 2004).

The problem considered in the present research deals with the classification of spatially distributed data into regions below and above of some predefined levels of contamination (Kanevski and Maignan 2004).

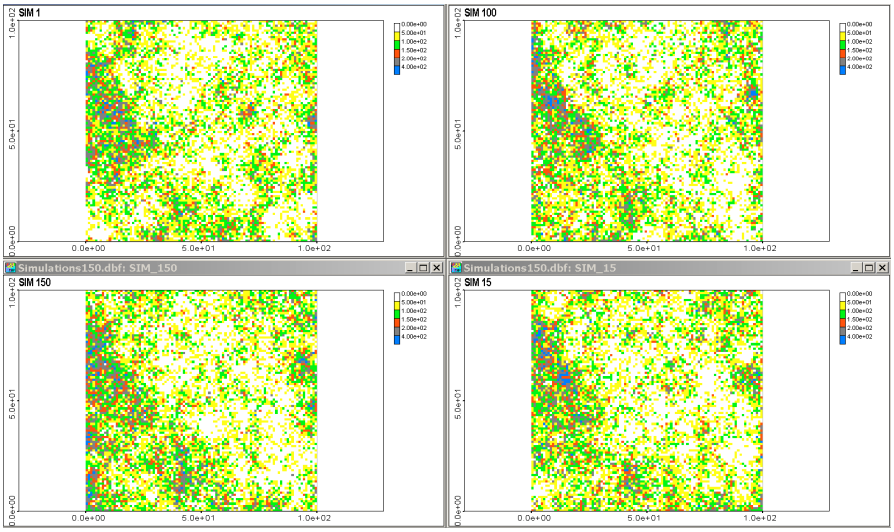

Figure 4.
Selected four realisations generated by conditional stochastic simulations.

The SVM solution of the classification problem using radial basis kernel is given by the following decision boundary discriminating between two classes (+1, -1)

$$F(X) = sign\left[W^T K(X, X_j) + b\right], \text{ where } K(X, X_j) = \exp\left\{-\left\|X - X_j\right\|^2 / 2\sigma^2\right\}$$

The weights W and the number of support vectors (data with nonzero weights) are given by a solution of quadratic optimization problem.

Hyper parameters (kernel bandwidth and regularization parameter) are tuned by training SVM using splitting data into training and testing subsets (training error ~ 0 %, testing error ~ 20%, Figure 6). The number of support vectors is only 20% of raw

data. In fact, sparseness of the solution can be used to optimise monitoring network. The solution with minimum testing error is used for the optimal decision boundary (Figure 7). The quality of the results can be estimate with independent validity set.

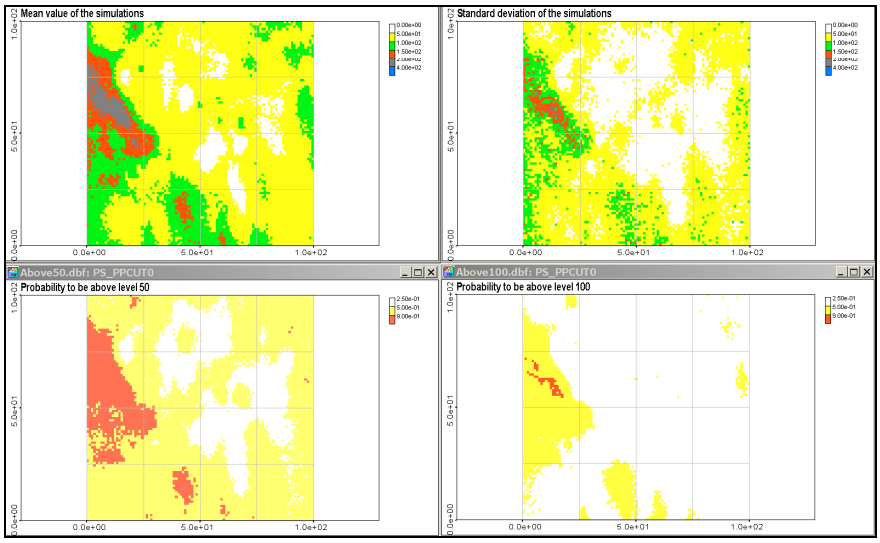

Figure 5.
Post-processing of the simulations: mean value of 150 simulations (top-left), standard deviation of the simulations (top-right), risk/probability maps to be above level 50 (down-left), probability to be above level 100 (down-right).

7. Conclusions

Radon data are highly variable and clustered at different spatial scales. Modern geostatistical approaches and machine learning algorithms are able to recognise spatial patterns in these data. In the present study several advanced modelling tools (simulations, Support Vector Machines) were applied for spatial predictions and classification of indoor radon concentrations. Such approaches are quite general and give much more powerful and useful results, in comparison with traditional mapping. Future developments based on hybrid models using geostatistics and machine learning algorithms seem to be promising, especially when data are nonstationary and multivariate.

References

Chiles J.P. and Delfiner P. (1999). Geostatistics. Modelling Spatial Uncertainty. A Wiley-Interscience Publication. New York.

Deutsch C.V. and Journel A.G. (1997) GSLIB. Geostatistical Software Library and User's Guide. N.Y., Oxford University Press.

Kanevski M. , M. Maignan. (2004). Analysis and Modelling of Spatial and Environmental Data. EPFL Press, Lausanne, Switzerland.

Vapnik V (1998). Statistical Learning Theory. John Wiley & Sons, N.Y.

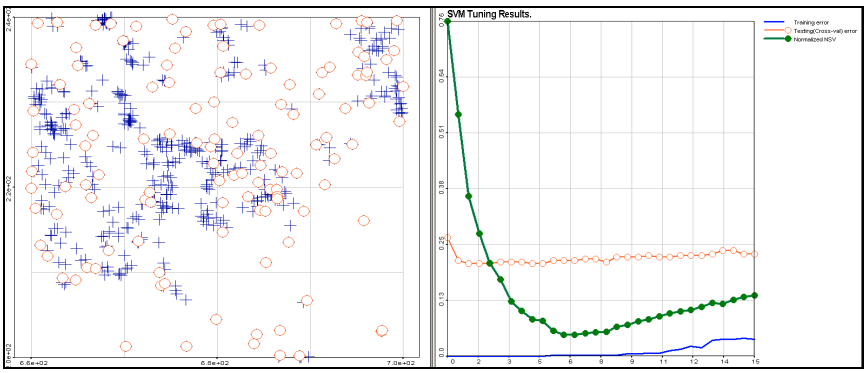

Figure 6.
Training data set and (O) - support vectors (left), error curves versus bandwidth (right).

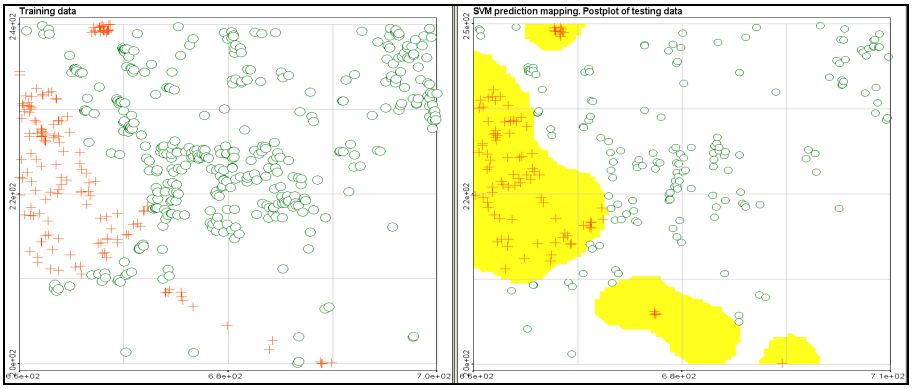

Figure 7.
Visualisation of 2-class problem (left), SVM optimal classification and post plot of testing data (right).