Clustering in Environmental Monitoring Networks: Dimensional Resolutions and Pattern Detection

D. Tuia, C. Kaiser and M. Kanevski

Abstract Monitoring Networks topology and resolutions (spatial and dimensional/fractal) influence the ability of networks to detect spatial phenomena. In the present paper we consider several fundamental questions related to the clustering of monitoring networks and their ability (1) to detect spatial phenomena and (2) to reproduce spatial patterns using geostatistical simulations. Artificial monitoring networks with known level of clustering characterized by their fractal dimension are sampled on a same reference image with known spatial structure. Subsequently, these networks are used to interpolate using Sequential Gaussian Simulation. Resulting images are compared with several methods. Clustering of networks does not harm global detection of spatial structures (i.e., definition of correct variogram model), but influence heavily the uncertainty related to these maps, especially in tasks of detection of areas-at-risk.

1 Introduction

Design of monitoring networks (MN) is an essential task for correct pattern detection and modelling of environmental phenomena. Non-homogeneous spatial distribution (clustering) of measurement points in space can lead to an over- or underestimation of global parameters such as mean or variance and to nonrepresentative probability distribution functions, which are crucial for conditional stochastic simulations (Deutsch and Journel 1997, Kanevski and Maignan 2004).

Monitoring networks design and optimization have been discussed by several authors (see Christakos 1992, Markus et al. 1999, Caeiro et al. 2003). Traditional spatial design techniques have been recently reviewed in a exhaustive way by De Gruijter et al. (2006).

Most of the studies on MN clustering have been dedicated to the consequences of clustering on distributions (without considering spatial aspect of data) while very

D. Tuia et al.
Institute of Geomatics and Analysis of Risk, University of Lausanne, CH-1015 Switzerland
e-mail: devis.tuia@unil.ch

A. Soares et al. (eds.), *geoENV VI – Geostatistics for Environmental Applications*, 497–509
© Springer Science+Business Media B.V. 2008

few of them studied two-point declustering in experimental variogram calculations (Richmond 2002). The present study is a first attempt to characterise the effect of clustering on spatial pattern detection: how clustered networks affect spatial predictions and how the potential losses can be described in terms of spatial patterns. In general, MN can be characterised by spatial and dimensional resolution. Dimensional (fractal) resolution characterises the dimension of the phenomena which can be detected by a particular network: in 2 dimensional space homogeneous networks (no clusters) can detect 2 dimensional phenomena (patterns). Clustered monitoring networks have a dimensional resolution d_f smaller than 2 and are not usually able to detect phenomena having dimension (2-d_f) (Lovejoy et al. 1986). Therefore, a loss of information can occurr, which will cause problems in spatial pattern reconstruction by using interpolations or simulations.

This paper presents synthetic example of simulated spatial patterns sampled with monitoring networks having different level of clustering and different dimensional resolutions.

Section 2 introduces basic notions about dimensional resolution and validity domains that are necessary to characterize the level of clustering of real monitoring networks. Section 3 focuses on the methodology used for the study that is performed in Section 4.

2 Quantitative Description of Network Clustering

There are different measures to quantify MN clustering: topological, statistical and fractal. Each measure characterises different aspects of clustering such as spatial resolution, dimensional resolution or statistical properties of clustering. In general, these measures are connected to each other. In the present study, monitoring networks with dimensional resolutions characterised by fractal dimensions are considered.

2.1 Fractal Dimension of Monitoring Network

Dimensional resolution (ability to detect spatial phenomena) was introduced for characterization of monitoring networks by Lovejoy et al. (1986). By fractals we mean statistically self-similar clustered point objects, whose structure is reproduced throughout the scales and whose dimension is usually not an integer.

In the present paper, fractal dimension d_f is used as a general indicator of clustering, where a dimension lower than 2 can be interpreted as the appearance of clusters at a certain spatial scale. Here, d_f is computed with the box-counting method (Falconer 1990, Peitgen et al. 1992): the area under study is covered by a regular grid of N cells, and the number of cells necessary to cover the whole network, S(L), is computed. Then, the size of boxes L is gradually decreased (accordingly, the number of boxes N is increased). The box-counting operation is repeated m times.

For the fractally distributed measurement points, the number of boxes necessary to cover the network points follows a power-law

$$S(L) \quad \sim \quad L^{-d_f} \tag{1}$$

The fractal dimension of the network d_f can be computed as the slope of the regression line after log-transformation of both sides of Eq. (1).

The equations presented above do not take into account real-life situations where different geographical constrains and finite number of measurement points are important. A recent paper by the same authors (Tuia and Kanevski 2006) has shown that a good way to quantify real monitoring networks is to compare them with a reference network generated within the same domains and having predefined fractality.

2.2 Validity Domains and Fractality of Monitoring Networks

Clustering of networks causes incorrect global estimations of mean and variance of the probability distribution function and erroneous spatial predictions over a regular two dimensional space. In geostatistics, a common practice is to interpolate the variable over the whole two-dimensional surface (often a square) and then to clip the results over the area of interest, e.g., with a GIS. These areas of interest, called validity domains (VD), spatially constrain the predictive space. In most cases, fractal dimension of such regions is less than two. In general, VD are related to geographical, political or economical constraints such as political boundaries or topographic barriers. In such cases, even homogeneous monitoring networks have fractal dimension smaller than two. Therefore, in order to quantify clustering of networks within VD, it was proposed to generate reference networks and to compare them with a real measurement network (Kanevski and Maignan 2004, Tuia and Kanevski 2006). Deviations between these networks (real and reference) were used to quantify the degree of clustering. Interpolation techniques have then been applied only on the area of interest, taking into account the irregular shape of the VD.

In order to analyze the effect of clustering on spatial predictions (reconstruction of spatial patterns), a region characterized by heavy geographical constraints has been chosen: the Swiss canton of Graubünden, which is characterized by a validity domain related to its mountainous landscape and to the organization of its inhabited areas into small settlements. The real monitoring network corresponds to indoor radon data measurements network. According to the methodology developed, three MN have been used for the current study (Fig. 1).

A. **Raw network**: a real MN (RMN), related to samples taken during an indoor Radon sampling campaign. The network is composed of $N = 3258$ unique measurements. The RMN is characterized by a high level of clustering corresponding to the fractal dimension $d_f = 1.38$; Two artificial homogeneous monitoring networks (GR network and Pop network) with the same number of sampling points generated within a validity domain of interest:

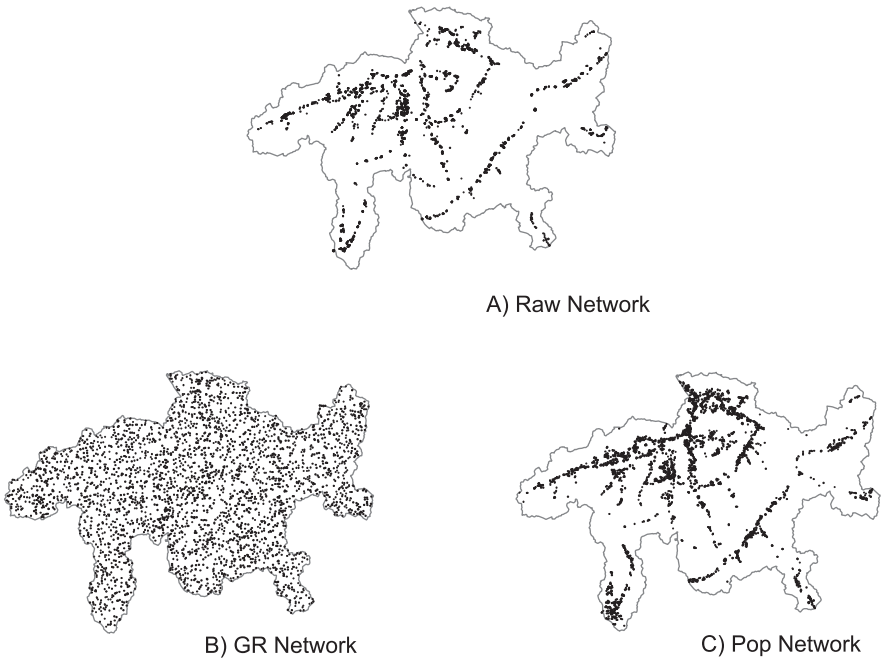

Fig. 1 Monitoring network used for the analysis. (A) RMN; (B) GR network; (C) Pop network

B. **GR network**: 3258 samples homogeneously distributed within the political boundaries of the canton of Graubünden. This VD corresponds to administrative geographical constraints. The dimension d_f of the network is 1.79, where the loss of dimensionality is related to complex boundary effects;
C. **Pop network**: 3258 samples homogeneously distributed within the limits of the populated regions of the canton of Graubünden. This VD corresponds to geographical constraints: the use of populated regions as VD avoids the presence of samples on mountainous regions and provides a distance-related barrier in terms of covariance. The dimension d_f of this network is 1.46.

Thus, the number of observations is constant throughout the networks, and the results should be a function of the network's design, i.e., the level of clustering, as defined by the fractal dimension.

In order to understand and to characterise spatial patterns (and corresponding uncertainties) detected by different networks, a reference model (complete image, CI) was simulated. Then, the CI model was sampled with different MN, which are described above. Finally, conditional simulations were carried out using sampled data and the results were compared with the CI.

3 Simulation of Spatial Patterns

3.1 Sequential Gaussian Simulations of a Reference Pattern

The reference patterns CI have been generated by a nonconditional simulation of a Gaussian random field $Y(u)$ with a given covariance $C_Y(h)$ by using Geostat Office (Kanevski and Maignan 2004):

$$C_Y(h) = \begin{cases} 1 - \frac{3}{2}\frac{|h|}{a} - \frac{1}{2}\frac{|h|^3}{a^3} & h \le a \\ 1 & h > a \end{cases} \quad (2)$$

Only the results on one reference image called SIM1 and generated according to the isotropic spherical variogram with 20 km correlation range are given (Fig. 2).

Once the reference SIM1 image was generated over the coordinates of Graubünden, it was sampled using three monitoring networks described above. These three artificial "measurement campaigns" were used to reconstruct the original pattern with Sequential Gaussian Simulation algorithm and to make the analysis and comparison between the results. The reconstruction of the patterns has been carried out with complete (including variogram analysis and modelling) conditional SGS based on the three sampling campaigns. The use of conditional SGS gives the possibility not only to compare generated patterns but also to quantify uncertainties and the variability between them.

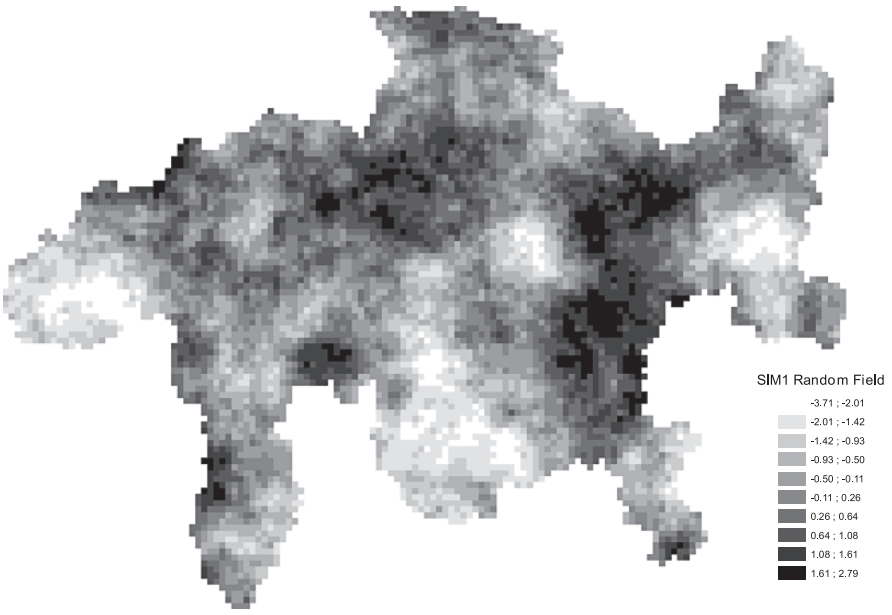

Fig. 2 Artificial phenomenon SIM1 generated by nonconditional simulation over the coordinates of Graubünden

3.2 Tools to Evaluate the Detection of Patterns

Several tools have been applied in order to evaluate the efficiency of pattern reconstruction.

3.2.1 E-type Maps

Conditional SGS provides many realizations of the random field $Z(u)$ governing the phenomenon. The first tool to evaluate the quality of the simulation is the comparison of the maps obtained by computation of the mean value for every simulated node over the M realizations with SIM1. A correct reconstruction of the SIM1 image by the simulations shows the ability of a clustered network to reproduce the underlying spatial structures described, for example, by the γ model.

3.2.2 Probability Isolines

Then, the generation of many realizations of $Z(u)$ allows post-processing analysis, such as the elaboration of maps of probability to exceed a given threshold g:

$$P(X \geq g) = F(g) \tag{3}$$

This procedure allows drawing the isolines corresponding to the same probability to exceed the threshold g. The SIM1 being an artificial reference image, the threshold is defined a priori for the analysis and is not related to a real level of risk.

3.2.3 Spatial Metrics

Finally, the analysis and comparison of the patterns generated by the conditional SGS cannot be made only by simple visual comparison. Several quantitative pattern description metrics coming from landscape ecology (O'Neill et al. 1988, Turner et al. 2001) have been applied on the risk maps discussed above.

3.3 Percentage of Landscape Covered (PLC)

This metric quantifies the percentage of landscape occupied by the patterns. The total area of the pattern is divided by the total area of the landscape (Validity Domain of the political boundaries).

3.4 Land Shape Index (LSI)

The LSI is an indicator of dispersion of a pattern formed by k disconnected patches. It is computed by dividing the total pattern edge length by the edge length of the smallest patch:

$$LSI = \frac{\sum_{i=1}^{k} e_i}{\min e_i} \tag{4}$$

This metric provides a standardized measurement of patches aggregation: the more the LSI increases, the more the patches are disaggregated (McGarigal and Marks 1994).

3.5 Concentration-dependent Fractal Dimension (CDFD, $d_f(Z_{th})$)

CDFD can be estimated with functional box-counting method over simulated nodes exceeding a given threshold Z_{th} for every level of probability tested. The CDFD curve is characterised by the dependence d_{f} (Z_{th}). If the probability level influences the shape of the pattern, then the CDFD curve should decrease with an increase of the level. If the pattern shape is stable, the curve should remain constant or decrease slowly.

4 Discussion

4.1 E-type Maps: Visual Comparison of Results

Fifty stochastic realisations were generated on three sampling networks of the same SIM1 reference phenomenon. In Fig. 3, E-type maps of the realisations are shown.

At a first glance, the networks can reproduce correctly the structure of the phenomenon, i.e., the variogram model. The dependence of the SGS mean results on the clustering of MN is visible by an effect of smoothing of the overall pattern. The GR network gives the best visual result, while the other networks, more clustered, are characterized by smoothed images.

4.2 Probability Isolines: Risk Maps to Draw Pattern Detection

For the SIM1 random field, the value of 1 has been defined as an action threshold for environmental protection (this choice is arbitrary). Then, the SGS simulations allow to draw maps of probability of exceeding that threshold. Figure 4 shows the probability maps related to every set of simulations considered.

The different probability maps studied showed that clustering of the network has a tendency to dilate the regions above a threshold for high uncertainty (i.e., small probabilities). The GR network shows small differences between the regions over the threshold for $P(X \geq 1) = 0.7$ and $P(X \geq 1) = 0.5$, while the Pop, and more clearly the Raw, show increasing differences of patterns between the maps.

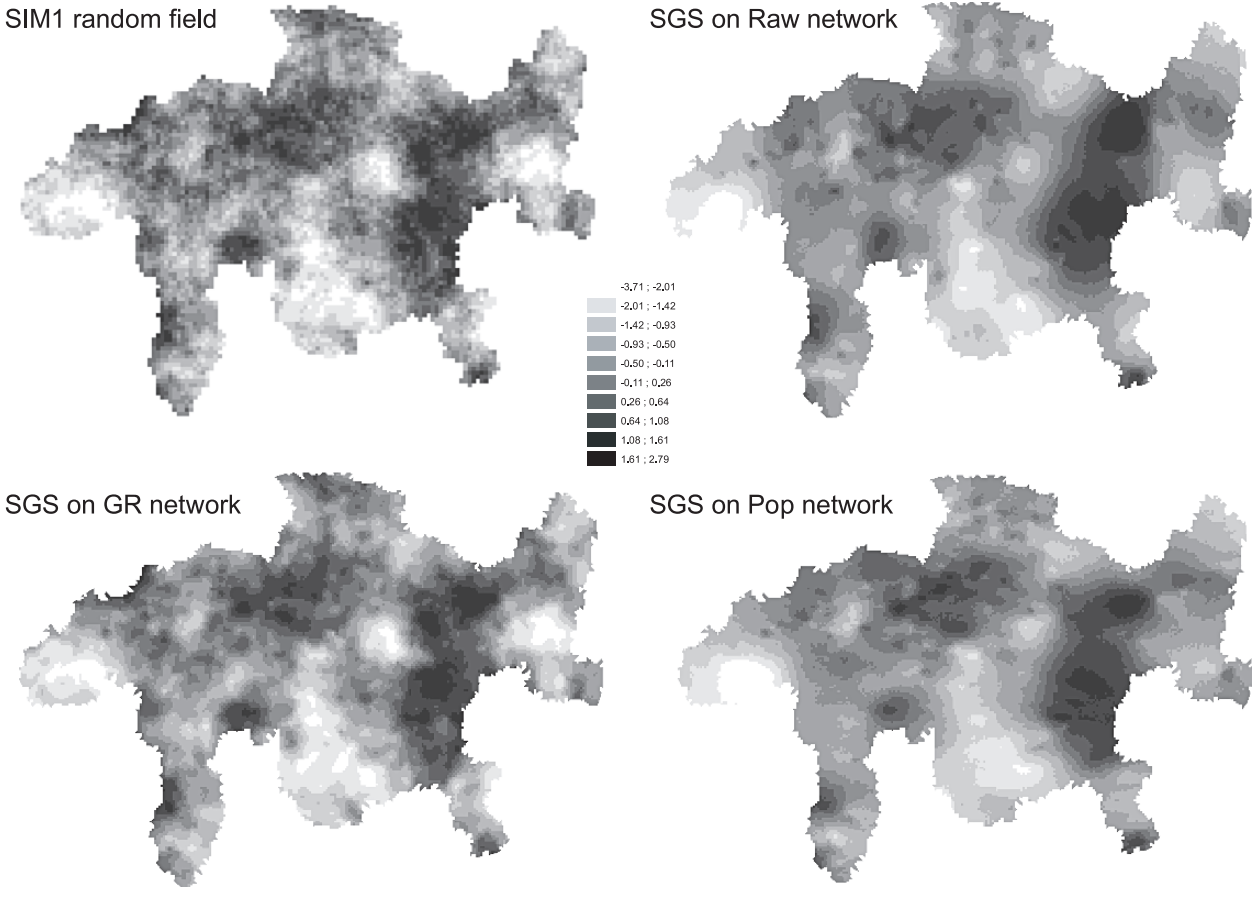

Fig. 3 E-type maps for the simulations obtained on the Graubünden region. Top-right: raw network; bottom-left: GR network; bottom-right: Pop network

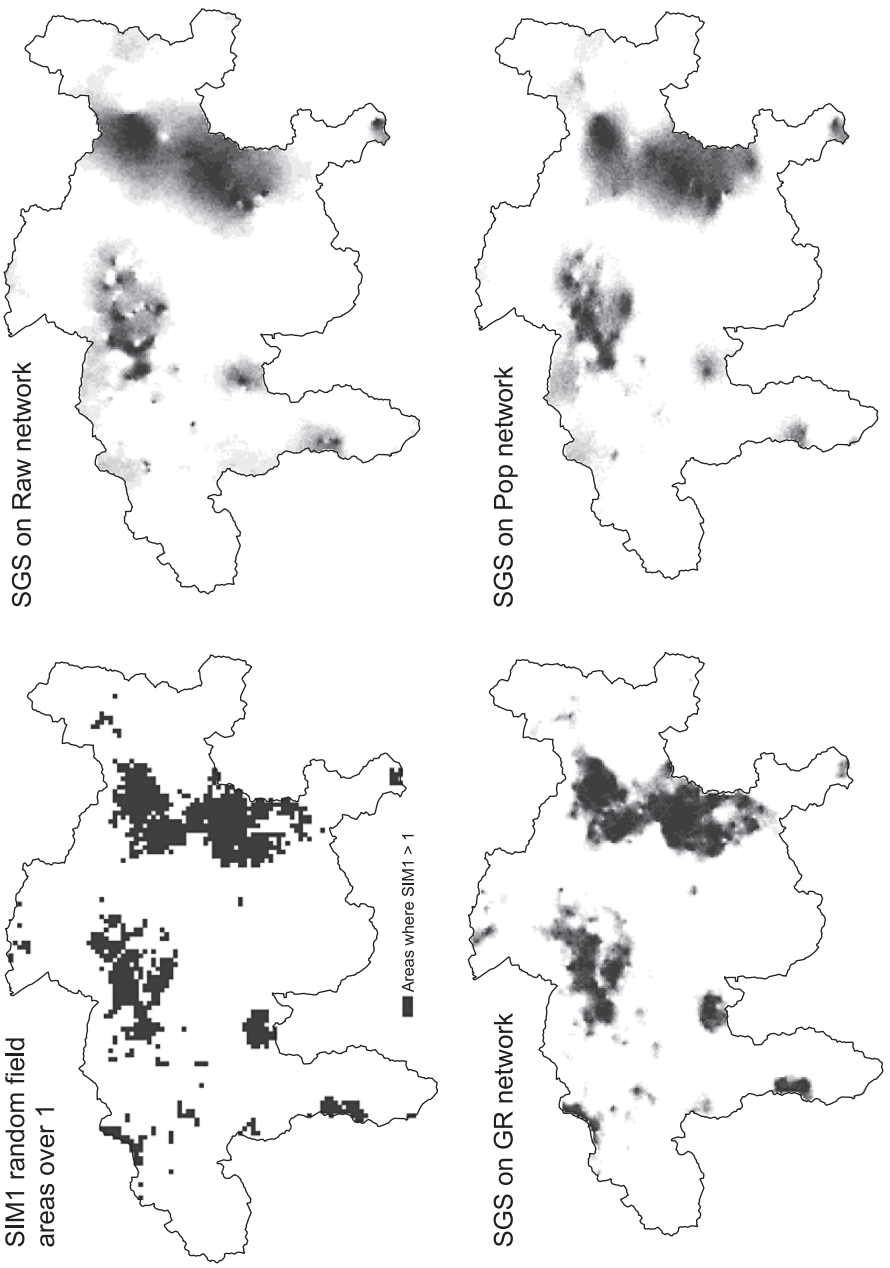

Fig. 4 Probability maps for $P(X \geq 1)$ Bottom-left: GR network; top-right: Raw network; bottom-right: Pop network

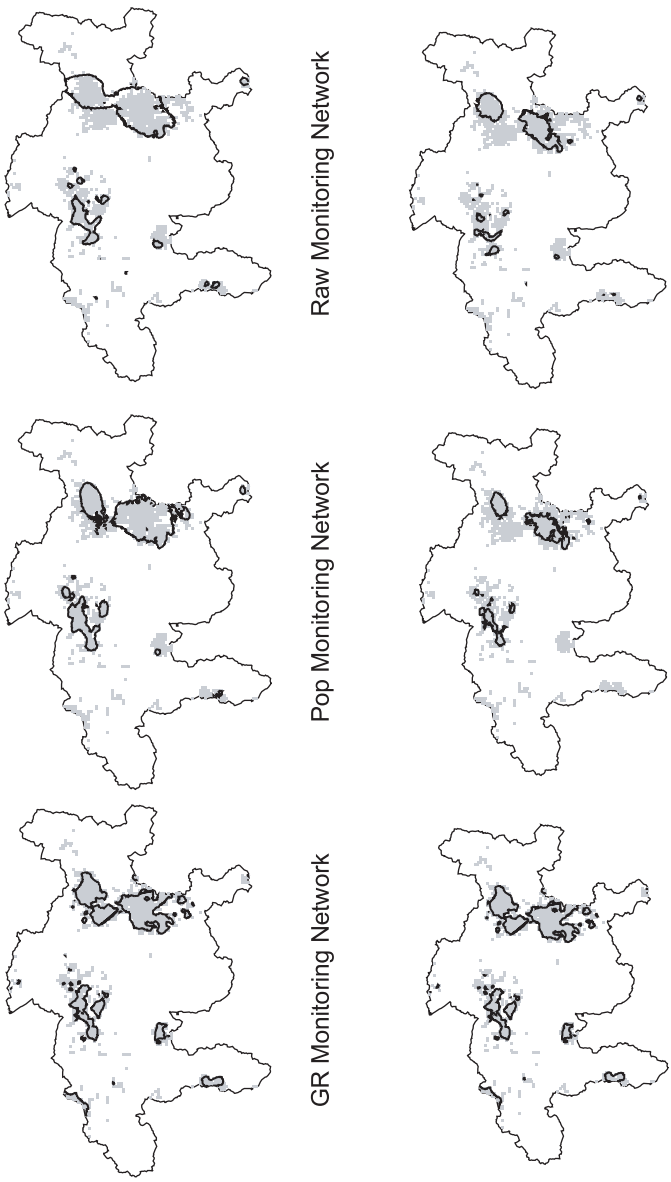

Fig. 5 Probability isolines for $P(X \geq 1) = 0.7$ (top row) and for $P(X \geq 1) = 0.9$ (bottom row). In grey the regions of the random field SIM1 above the threshold. Left: GR network; center: Pop network; right: Raw network

The clustering of networks increases differences between the simulations of $Z(u)$ and destabilizes the shape of the surface at risk for a given probability level.

Comparing only the isolines for probabilities 0.7 and 0.9 (Fig. 5), it can be seen that the clustering of the network leads to a progressive loss of detection of the shape of the area at risk: the GR network allows a correct detection of pattern for small uncertainty levels, while the Pop and Raw networks lose detection on peripheral areas and start having false detections of the phenomenon in areas which are, in reality, below the defined threshold.

4.3 Spatial Metrics

The analysis of the spatial metrics discussed above (Fig. 6) confirmed the observations made following the analysis of probability isolines maps. On one hand, the GR network keeps a higher connectivity level through the probability levels, showed by the slow decrease of PLC (that shows the consistency of small uncertainty levels) and the stability of the CDFD index, which can be explained by the stability of patches keeping their shape and connectivity. On the other hand, clustered networks (Raw and Pop) lead to a faster decrease of PLC, showing a higher uncertainty depending on the probability considered. The CDFD index shows a loss of connectivity of the pattern for small uncertainty (for $F(1) \geq 0.8$) which can even be observed on the map (Fig. 5) by the reduction of the pattern to small patches localized on areas related to high density of samples.

LSI shows the level of aggregation of pattern: the real situation (SIM1) is heavily fragmented, reflecting the complexity of $Z(u)$: the GR network can reproduce partially this disaggregation, which is completely lost with the clustered networks. There are characterized by small values of LSI, i.e., aggregated and smoothed patterns for every probability level.

5 Conclusion

Clustering of monitoring networks has a significant impact on spatial prediction of random fields. Heavily clustered sampling schemes can decrease the quality of definition of areas at risk for environmental and pollution problems. In this study it was shown that even clustered networks can detect correctly the variogram model, but that the realization of the random field cannot provide a correct definition of areas at risk, especially if small uncertainty levels are required.

Patterns created by clustered networks are heavily dependent on the probability level considered. For high levels connectivity is lost, as it is shown by the CDFD analysis. Risk maps can only detect hot spots related to the location of samples.

This study has only used measures of pattern detection based on visual comparison and spatial metrics, which do not analyze patterns in terms of shape or correct reconstruction of the random field. In order to better compare generated patterns,

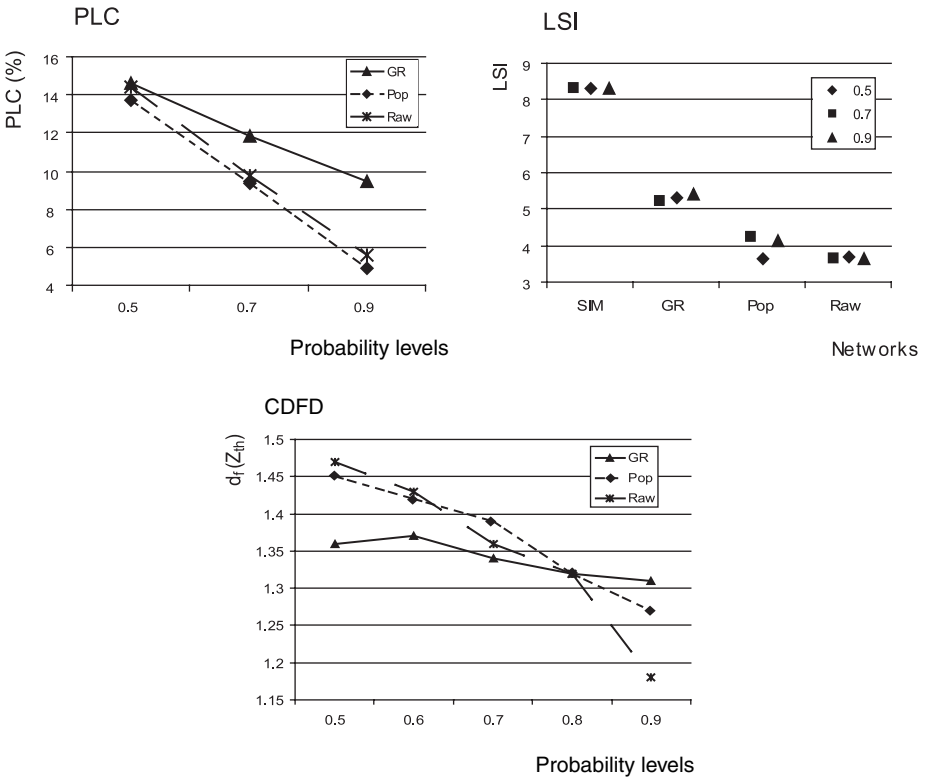

Fig. 6 Spatial metrics. (Top left) Percentage of landscape covered; (Top right) LSI; (bottom) CDFD index

the spatial metrics are being calculated for every independent simulation within our current research. In this way, the point-to-point relationships are preserved and the patterns are coherent for comparison of independent simulations. Approaches based on multiple points statistics and mathematical morphology may be also useful in order to characterize patterns in relation to their shapes and localizations.

Another important issue concerns the fractal dimension used to characterize the clustering of the MN: even if d_f is often used as a global measure of clustering, one must remember that different networks can show the same fractal dimension. The analysis of the robustness of d_f as index of clustering in topologically different situations is central in the process of validation of the index.

Acknowledgments This work has been partially supported by the Swiss National Foundation. Projects "Urbanization Regime and Environmental Impact: Analysis and Modelling of Urban Patterns, Clustering and Metamorphoses" (n.100012–113506) and "GeoKernels": Kernel-Based Methods for Geo- and Environmental Sciences (n.200021–113944).

References

Caeiro, S., Painho, M., Goovaerts, P., Costa, H., Sousa, S. (2003), Spatial sampling design for sediment quality assessment in estuaries. Environmental Modelling & Software, 18:853–859.

Christakos, G. (1992), Random Fields Models in Earth Sciences. San Diego, Academic Press.

De Gruijter J., Brus D., Bierkens M., Knotters M. (2006), Sampling for Natural Resource Monitoring. Berlin Heidelberg, Springer-Verlag.

Deutsch, C., Journel, A. (1997), GSLIB. Geostatistical Software Library and User's Guide. New York, Oxford University Press.

Falconer, K.J. (1990), Fractal Geometry. Mathematical Foundations and Applications. Chichester, John Wiley and Sons.

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

Lovejoy S., Schertzer D., Ladoy P. (1986). Fractal characterisation of inhomogeneous geophysical measuring networks. Nature, 319: 43–44.

Markus, A., Welch, W.J., Sacks, J. (1999), Design and analysis for modeling and predicting spatial contamination. *Mathematical Geology*, 31(1):1–22.

McGarigal, L., Marks, B.J. (1994), FRAGSTATS manual: spatial pattern analysis program for quantifying landscape structure. http://www.umass.edu/landeco/research/fragstats/fragstats.html

O'Neill, R.V., Krummel, J.R., Gardner, R.H., Sugihara, G., Jackson, B., DeAngelis, D.L., Milne, B.T., Turner, M.G., Zygmunt, B., Christensen, S.W., Dale, V.H., Graham, R.L. (1988), Indices of landscape pattern. *Landscape Ecology*, 1:153–162.

Peitgen, H.O., Hartmut, J., Saupe, D. (1992), Chaos and Fractals: New Frontiers of Science. New York, Springer-Verlag.

Richmond A. (2002), Two-point declustering for weighting data pairs in experimental variogram calculations. Computers and Geosciences, 28: 231–241.

Tuia, D., Kanevski, M. (2006), Indoor Radon Data Monitoring Networks: Topology, Fractality and Validity Domains, Congress of the International Association of Mathematical Geology (IAMG), Liège, Belgium.

Turner, M.G., Gardner, R.H., O'Neill, R.V. (Eds) (2001), *Landscape Ecology in Theory and Practice: Pattern and Process*. Springer-Verlag, New York.