Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

Visualization and comparison of spaces of uncertainty using three-dimensional display and multi-dimensional scaling

Pierre Goovaerts
BioMedware, Inc
121 W. Washington St., 4th Floor-T
Ann Arbor, MI 48104
goovaerts@biomedware.com

Abstract

In the analysis of cancer mortality and incidence maps, three types of uncertainty typically arise: (1) the uncertainty about the value of the health outcome over a single geographical unit (local uncertainty), (2) the joint uncertainty about outcome values recorded simultaneously over several geographical units (spatial uncertainty), and (3) the uncertainty about the existence of significant clusters of high disease risk resulting from the propagation of spatial uncertainty through the local cluster analysis (response uncertainty). In each case, the probabilistic way to assess the uncertainty consists of determining the distribution or set of possible outcomes (e.g., local risk value, risk map, or cancer cluster map), which is referred to as the space of uncertainty. The characterization and visualization of the different spaces of uncertainty is still an unresolved issue. This paper presents a distance-based approach to: 1) quantify differences between simulated maps using different distance metrics (e.g. Euclidean versus LISA-based distances), and 2) project the ensemble of maps into a low dimensional Cartesian space using metric multi-dimensional scaling (MDS). This approach is used to visualize the impact of simulation algorithm, distance metric, and number of realizations on the extent of the spaces of uncertainty. Three-dimensional displays of series of quantile maps, probability maps or simulated risk maps are also proposed as an innovative tool to visualize and communicate local and spatial uncertainty to end-users. The discussion is illustrated using county-level mortality rates for cervical cancers recorded in 118 counties of the Western US.
Keywords: geostatistics, stochastic simulation, propagation of uncertainty, variogram, clusters, kriging.

1 Introduction

Cancer mortality and incidence maps are used by public health officials to identify areas of excess and to guide surveillance and control activities. Quality of decision-making thus relies on an accurate quantification of risks from observed rates which can be very unreliable when computed from sparsely populated geographical entities or for diseases with a low frequency of occurrence. By analogy with the terminology used in the geostatistical literature [2,3], three types of uncertainty can be distinguished:

1. **Local uncertainty:** uncertainty prevailing over a single geographical entity at a time; for example uncertainty about cervical cancer mortality rate for any given county in Figure 1.
2. **Spatial uncertainty:** uncertainty about values of health outcomes recorded simultaneously over several geographical units; for example uncertainty about the existence of aggregates of counties with high cancer mortality rates in Figure 1.
3. **Response uncertainty:** uncertainty about results of the application of a transfer function (e.g. cluster detection algorithm) to health outcomes; for example spatial uncertainty translates into a lack of reliability of some of the spatial clusters detected on the map of mortality rates.

In all three cases, the uncertainty can be modeled empirically through the generation of a set of possible outcomes, known as "**space of uncertainty**" [2]. For example, Figure 1 shows the posterior cumulative distribution of cervical cancer standardized mortality risk (SMR) derived for a few counties (*local uncertainty*) using the popular Besag, York and Mollie's (BYM) model.

Spatial uncertainty is modeled through the generation of a set of simulated risk maps (also known as "realizations"), each consistent with the information available, such as a spatial correlation function. For example, Figure 2 (top) shows twenty simulated maps of cervical cancer mortality rates generated by p-field simulation [3,9]. Spatial features consistently observed across all simulations (e.g. lower mortality risk in Utah) are deemed more likely than the ones that are displayed by a few realizations.

Uncertainty about the location of significant clusters of low or high values can be quantified by *propagating the uncertainty* attached to cancer mortality rates through the local cluster analysis (LCA). For example, the twenty simulated risk maps underwent a LCA based on the LISA statistic [1], leading to twenty maps of significant Low-low or High-high clusters (Figure 2, middle graph).

Although the concepts of stochastic simulation and propagation of uncertainty are not new, it appears that little attention has been paid to the definition of the space of uncertainty, and related issues such as the equivalence of spaces of uncertainty generated by different algorithms, and the number of realizations required for sampling this space [10]. To quote Myers [5], "*Underlying this diversity of algorithms was an implicit but never stated assumption that there was some form of equivalence and hence the difference was only computational. . . Neither of these implicit assumptions has really been tested or even considered, most users do not use multiple algorithms and make comparisons nor do they generate multiple finite sets of realizations to compare between the sets*."

Figure 1: Posterior distributions of cervical cancer standardized mortality rates (SMR) modeled using the Besag, York and Mollie's (BYM) model. Note the narrower distribution (steeper CDF) for the heavily populated Santa Barbara County which indicates the smaller uncertainty (greater reliability) of the SMR value.

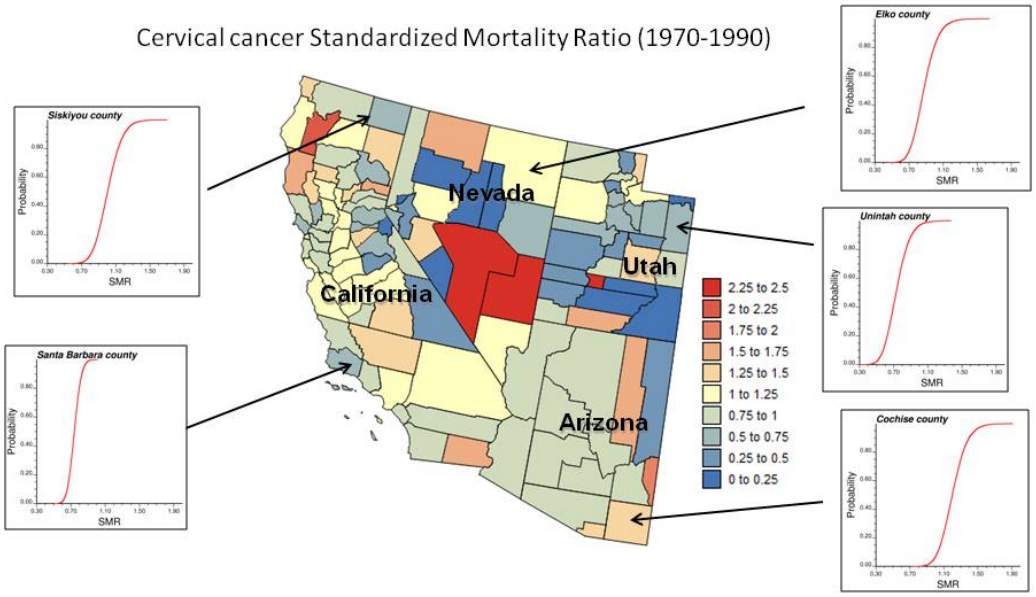
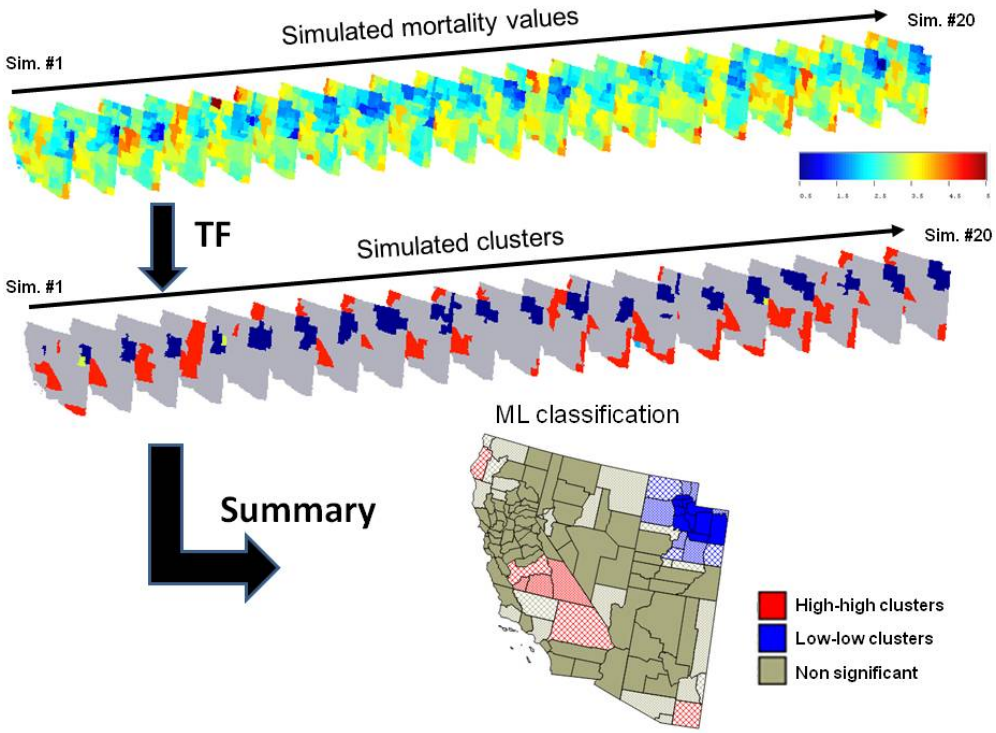


Figure 2: Twenty simulated maps of cervical cancer mortality rates (i.e. SMR multiplied by area-wide rate of 3.08 deaths/100,000 habitants) generated by p-field simulation and the corresponding maps of significant clusters of high or low values detected using a local Moran's I cluster analysis (Transfer Function). Classification results are summarized by mapping for each county the most likely (ML) classification inferred from 20 realizations. The intensity of the shading increases as the classification becomes more certain (i.e. larger likelihood).



Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

From the user's perspective, it is also important to be able to visualize the uncertainty in the local and spatial models of risk values. Plotting cumulative distribution functions as in Figure 1 is unsatisfactory since it quickly generates visual clutter and does not allow an easy visualization of probabilities of exceeding specific thresholds for example. Similarly, the innovative 3D display of simulated maps and corresponding cluster classifications in Figure 2 is not practical for large numbers of realizations.

Using the choropleth map of county-level disease rates in Figure 1, this paper describes preliminary results on: 1) the representation of local and spatial uncertainty by three-dimensional displays, and 2) the mapping and ranking of simulated maps generated by different algorithms using a recently developed distance-based approach [7,8].

2 Visualizing local and spatial uncertainty

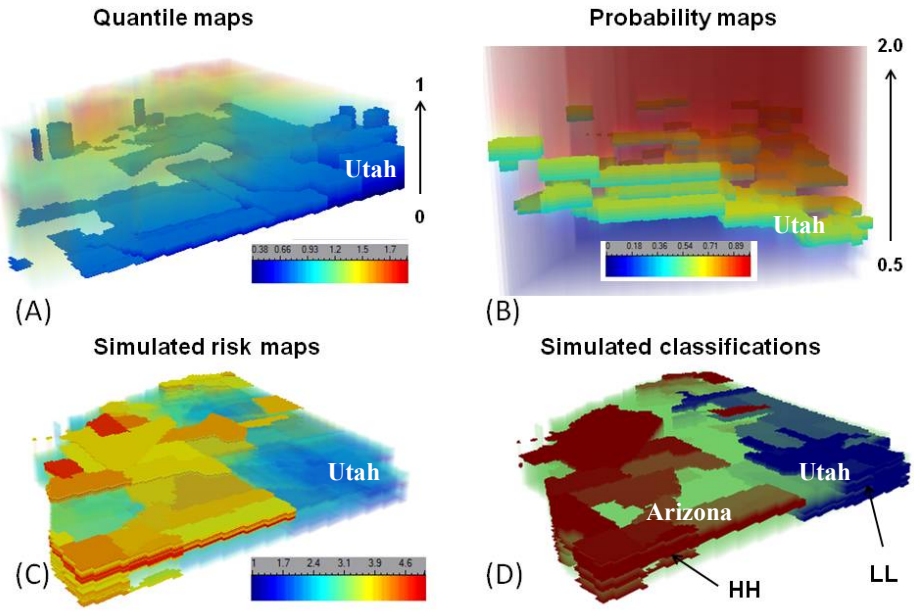
Although stochastic simulation offers a way to generate a large number of potential scenarios, the burden of manually scrolling through dozens of different maps will test the patience of most users and be little informative. The information contained in the set of simulated maps is thus often summarized through a **static display** of probabilities of exceeding particular threshold or some measures of the spread of the posterior distribution. By doing so, one however fails to depict the uncertainty about spatial features and essentially maps the area-specific measures of uncertainty provided by kriging and other rate smoothing methods. Similarly, the information provided by the whole local probability distribution (e.g. Figure 1) is often summarized by computing the probability of exceeding a particular threshold and summary statistics, such as mean or variance, which is unsatisfactory when the distributions are non-parametric.

By analogy with recent work accomplished in the area of 3D visualization of space-time datasets, we explored the use of 3D displays (i.e. vertical stacks of maps) where the vertical dimension can correspond to a quantile, a probability or a realization number. These 3D graphs were all created using SGeMS (Stanford Geostatistical Modeling Software [6]) 3D visualization panel and FORTRAN programs developed to format the data.

Figure 3A shows a vertical stack of 100 maps of mortality risk values corresponding to a cumulative probability increasing from zero to 1. In other words this series of maps correspond to the percentiles of the probability distributions displayed in Figure 1 for a few counties. As the probability increases, the risk value increases. Volume rendering and transparency settings were used to attenuate the visual obstruction of inner cells and highlight only the information relevant to certain users while less important information is not completely masked to avoid losing the overview of the map. In this example, SMR above 0.75 were partially hidden to highlight the large likelihood for the risk to be smaller than 0.75 (i.e. smaller than 75% of the average risk over the study area) in Utah, while larger risks can be observed along the West Coast.

The counterpart of the series of quantile maps is found in Figure 3B that shows a vertical stack of 100 maps of cumulative probabilities corresponding to a risk threshold (SMR) increasing from 0.5 to 2. As the threshold increases, the probability for the risk to be below that threshold rises. In this example, probabilities below 0.4 and above 0.6 were partially hidden to highlight the 0.2-probability interval centred on the median, i.e. the interval bounded by the 40th and 60th percentiles of the probability distributions of Figure 1. The width of the interval reflects the uncertainty about the risk values while the position of the interval along the vertical axis indicates the magnitude of the standardized mortality

Figure 3: Three-dimensional displays of local and spatial uncertainty: (A) stack of standardized risk quantile maps (SMR > 0.75 are partially hidden), (B) stack of probability maps ([0.4,0.6] probability intervals are highlighted by hiding partially all probabilities < 0.4 and > 0.6), (C) stack of incrementally different simulated risk maps (risk values < 3.5 deaths per 100,00 habitants are partially hidden) and (D) the corresponding clusters of high and low values.



Multidisciplinary Research on Geographical Information in Europe and Beyond
Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon, April, 24-27, 2012
ISBN: 978-90-816960-0-5
Editors: Jérôme Gensel, Didier Josselin and Danny Vandenbroucke

risk. For example, intervals are wider and located at the bottom of the threshold axis for counties in Utah (forefront), indicating low risk and large uncertainty because of sparse population. On the contrary, intervals are narrower and located toward the top of the vertical axis for the West Coast (background) where risks and population are larger.

Spatial uncertainty is represented in Figures 3C&D that show vertical stacks of 50 simulated risk maps (or corresponding cluster classifications) where the vertical axis is the realization number. To allow the eye to catch gradual changes successive realizations must be similar enough. Such a similarity can be achieved by ranking the realizations appropriately (e.g. using MDS results described below) or by using a variant of p-field simulation algorithm that generates realizations that are incrementally different [3,9]. The later was used here. The 3D display of realizations combined with volume rendering and transparency settings (Fig. 3C) allows one to distinguish areas that remain stable over all realizations (low uncertainty) from those where large fluctuations occur between realizations (high uncertainty). The stack of classified maps (Fig. 3D) illustrates the location and reliability of clusters of low and high risks. In this particular example, Utah cluster of low risk is highly certain whereas the cluster of high risk in Arizona is less certain since it appears only in a dozen realizations located at the top of the stack.

3 Representing spatial uncertainty using multi-dimensional scaling

Once an ensemble of simulated maps (realizations of a random function) have been created and visualized using three-dimensional displays or other means, a key step is to assess and summarize the variability among this potentially large number of simulations that can be themselves fairly complex. For example, the generation of 1,000 simulated maps of county-level cancer mortality risks over the continental US would translate into more than 3 millions numbers. The distance-based approach, illustrated in Figure 4, proceeds in two steps:

1. a distance matrix *D* that quantifies the dissimilarity between any two simulated maps is created (e.g. the Euclidean distance was computed for all possible pairs of the 50 simulated maps in Figure 4),
2. the ensemble of realizations is projected into a low dimensional space (e.g. 2D at the bottom of Figure 4) using a metric multi-dimensional scaling (MDS) of the distance matrix. Points close in this 2D space correspond to simulated maps that display similar spatial features (e.g. pairs 3-27 and 35-45).

Figure 4: Schematic diagram illustrating the use of multi-dimensional scaling (MDS) to visualize the space of uncertainty modeled by a set of 50 simulated maps of cervical cancer mortality risk. Differences between any two maps *i* and *j* are first quantified using a distance metric $\delta_{i,j}$ (e.g. Euclidean distance here). The matrix *D* of distances is then used to map all simulated maps into a Euclidean two-dimensional space using multidimensional-scaling. Points close in this 2D space correspond to simulated maps that display similar spatial features (e.g. pairs 3-27 and 35-45). More than two dimensions might be required if the Euclidean distance between any two points in the 2D space is poorly correlated with the corresponding dissimilarity in the distance matrix *D*.

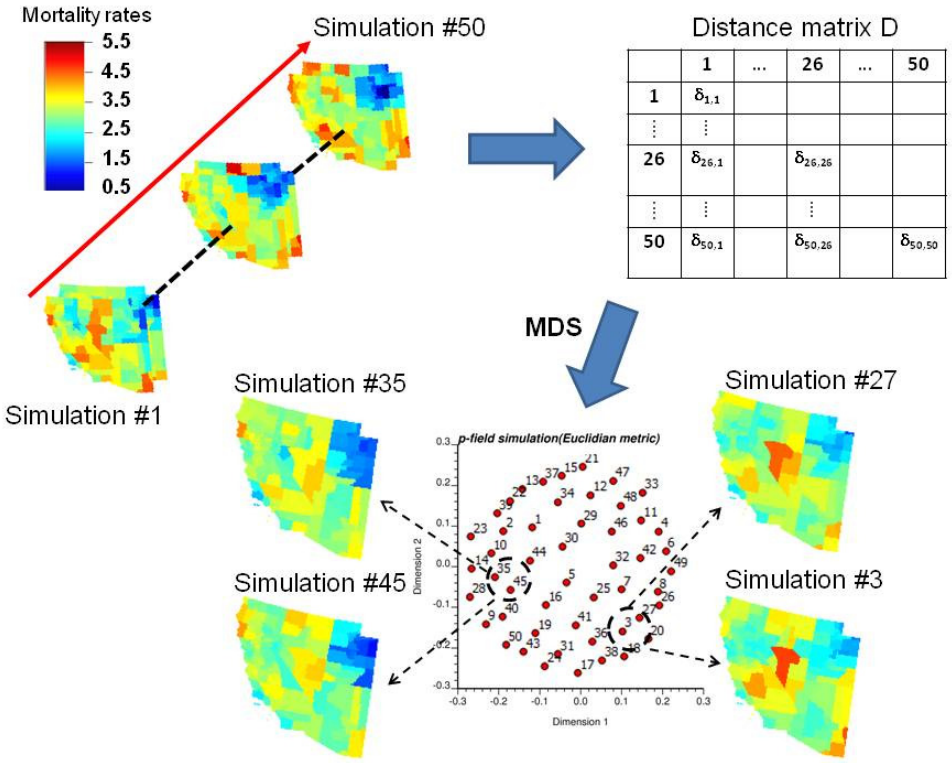


Figure 5: Illustration of the ability of multidimensional-scaling to cluster in a 2D space simulated maps that are similar. The 50 realizations in the top space (A) were generated using a variant of p-field that creates a series of realizations that are incrementally different, while realizations in the bottom space (B) were generated in a random order. Realizations are linked according to the order in which they were created.

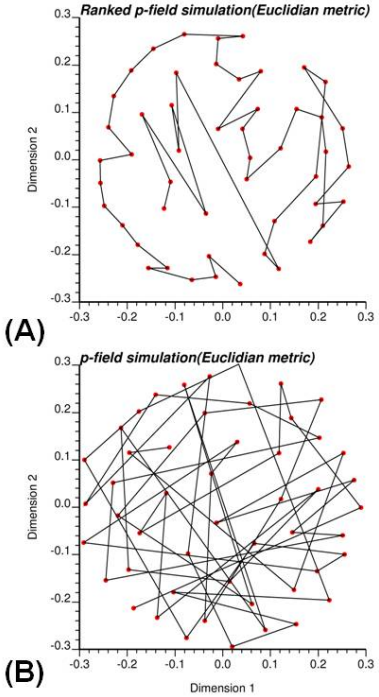
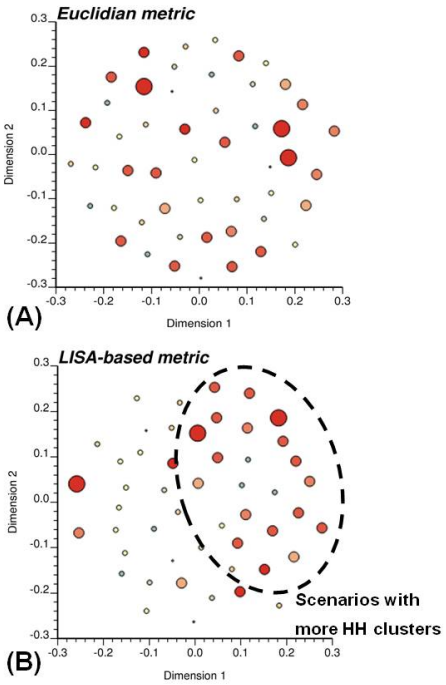


Figure 6: Impact of the choice of a distance metric (Euclidean versus LISA-based distance) on the projection of the set of 50 simulated maps of cervical cancer mortality risk. The size of each dot represents the number of counties classified as high-high (HH) clusters in that simulated map. Using the LISA-based metric (B) leads to aggregates of simulated maps with similar frequency of HH clusters, thereby facilitating the selection of extreme scenarios.



This realization-based representation of uncertainty provides an effective way to visualize the variability among realizations, rank them according to specific features (e.g. spatial connectivity) and select a subset of representative realizations. For example, Figure 5 shows the projection of 50 realizations generated using two types of p-field simulation: 1) Srivastava's technique that generates a series of realizations that are incrementally different through the use of a probability field that is much larger than the area to be simulated [4,9], and 2) general implementation that creates realizations in no specific order. The realizations in Figure 5 are linked in the order they were generated and the distance-based representation clearly illustrates the natural ordering of the realizations generated by the second approach where successive realizations are closer in the 2D space.

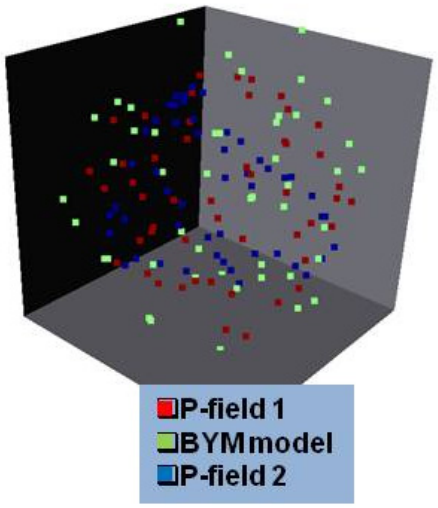
3.1 Distance metric

The cornerstone of the approach is the definition of a distance metric that is tailored to the application at hand. For health studies concerned with the identification of clusters of high risks, a metric based on a Local Indicator of Spatial Autocorrelation (LISA, [1]) seems more appropriate than the general Euclidean distance. The 50 realizations of Figure 4 underwent a MDS analysis using as distance metric the average absolute difference between local Moran's I (LISA-based distance) instead of the average absolute difference between simulated risks (Euclidean-based distance). These realizations are displayed in Figure 6 where the size of each dot (i.e. realization) is proportional to the number of counties that were allocated to a cluster of high risk. Clearly, using the LISA-based metric (Figure 6B) leads to a better discrimination of simulated maps according to the frequency of clusters of high values than the use of the Euclidean-based distance in Figure 6A.

3.2 Algorithmically-defined spaces of uncertainty

The use of stochastic simulation in test of hypothesis relies on the assumption that the space of solutions is sampled fairly exhaustively and uniformly. Some believe that the space of uncertainty must be theoretically defined outside the use of a particular algorithm. Others state that the space of uncertainty can only be defined through the algorithm and consists of all possible realizations that could be generated by the algorithm. This view is particularly suited to the space of uncertainty of

Figure 7: 2D mapping of 50 realizations simulated by three different types of algorithms (BYM model, *p*-field simulation with and without incrementally different realizations). Note how the different spaces of uncertainty overlap, yet are not identical.



responses or output values which cannot be defined analytically because of the complexity (non-linearity) of transfer functions, such as local cluster analysis. In the latter case, the term of algorithmically-defined space of uncertainty is used [2]. Figure 7 indicates that the spaces of uncertainty generated by different algorithms have different size and shape despite a reasonable amount of overlap.

4 Conclusions

Mapping the space of uncertainty through a multi-dimensional scaling (MDS) analysis of distances between simulated maps is very recent and its application has been confined to uncertainty in petroleum reservoir models [7]. It is also a complete generalization of the experimental design technique to reflect the ensemble of sources of uncertainty [8].

The characterization of the different algorithmically-defined spaces of uncertainty has important implications for tests of hypothesis that use these realizations to derive *p*-values. In the future, simulation studies will be conducted to assess their precision (extent of space of uncertainty), accuracy (ability for the space to include the "true" map), and uniform sampling (density of realizations). In addition to the simulation algorithm, the impact of the number of realizations on the properties of the space of uncertainty will be explored, allowing answering important questions such as: Does the extent of the space of uncertainty increase monotonically with the number of realizations? Are some algorithms better suited to generate extreme scenarios using fewer realizations? How many realizations are needed to achieve stable *p*-values in randomization-based tests of hypothesis?

5 Acknowledgments

This research was funded by grants R43CA150496-01 and R44CA132347-02 from the National Cancer Institute. The views stated in this publication are those of the author and do not necessarily represent the official views of the NCI.

References

[1] Anselin L. Local indicators of spatial association - LISA. *Geographical Analysis*, 27:93-115, 1995.

[2] C.V. Deutsch. Algorithmically-defined random function models. In R. Dimitrakopoulos, editor, *Geostatistics for the Next Century*, pages 422-435. Kluwer, Dordrecht, 1994.

[3] P. Goovaerts. Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. *International Journal of Health Geographics*, **5**:7, 2006.

[4] MacEachren A, Brewer CA and LW Pickle. Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A*, 1547-1561, 1998.

[5] D.E. Myers. Choosing and using simulation algorithms. In H. T. Mowrer, R. L. Czaplewski and R.H. Hamre, editors, *Spatial accuracy assessment in natural resources and environmental sciences*, pages 23-29. U.S. Department of Agriculture, Forest Service, Fort Collins.

[6] Remy N, Boucher A and J Wu. *Applied Geostatistics with SGeMS: A User's Guide*. Cambridge University Press, New-York, 2008.

[7] Scheidt C and J Caers. A new method for uncertainty quantification using distances and kernel methods: application to a deepwater turbidity reservoir. *SPE Journal*, 14:680-692, 2008.

[8] Scheidt C and J Caers. Representing spatial uncertainty using distances and kernels. *Mathematical Geosciences,* 41:397-419, 2009

.

[9] R.M. Srivastava. The visualization of spatial uncertainty. In J. M. Yarus and R. L. Chambers, editors, *Stochastic Modeling and Geostatistics: Principles, Methods and Case Studies*. *AAPG Computer Applications in Geology*. pages 339-345. The American Association of Petroleum Geologists, Tulsa, Oklahoma, 1994.

[10] R.M. Srivastava. Matheronian geostatistics: where is it going? In E. Y. Baafi and N. A. Schofield, editors, *Geostatistics Wollongong '96*, pages 55-66. Kluwer Academic Publishers, Dordrecht, 1997.