AUTOMATIC MAPPING ALGORITHMS

for

ROUTINE AND EMERGENCY MONITORING DATA

SPATIAL
INTERPOLATION
COMPARISON
2004

North

Edited by G. Dubois
Radioactivity Environmental Monitoring
Emissions and Health Unit

LEGAL NOTICE

Neither the European Commission nor any person
acting on behalf of the Commission is responsible for
the use which might be made of the following information.

A great deal of additional information on the
European Union is available on the Internet.
It can be accessed through the Europa server
(http://europa.eu.int)

FOREWORD

Grégoire Dubois, September 2005

Radioactivity Environmental Monitoring
Emissions and Health Unit
Institute for Environment and Sustainability
Joint Research Centre
European Commission
Italy

Automating the process of mapping environmental data for early warning and risk management presents a number of challenging and unresolved issues. Variables such as atmospheric pollutant concentrations, radiation levels, rainfall fields, seismic activity, etc., have in common that they are monitored with ground-based networks and that they need to undergo an interpolation process before becoming useful for decision-making or modelling. Because there is nothing such as a universal spatio-temporal mapping algorithm, the tools currently used for spatial data analysis either provide users with very simple mapping functions that allow quick processing of the data, or, what appears to become a new trend, allow them to interact with decision trees that direct them towards more advanced functions. These wizards usually require that the users have some prior knowledge and experience in geostatistics.

Basic algorithms will generate maps in real time, but the results will oversimplify reality, may show very large uncertainties or be even completely unrealistic. Concerning the second approach, maps would most probably show much smaller uncertainties and the outcome of the algorithms would have gone through some quality control. Should the maps be needed in real time, such a strategy would obviously become unworkable.

In case of hazards and emergencies (e.g. pollution peaks, nuclear accidents, flash-floods, earthquakes, etc.), maps need to be generated in real-time with minimum human intervention. Considering the current situation described above, an efficient management of environmental risks is still hindered either by the large uncertainties of the maps produced and/or by the slowness in obtaining more accurate maps.

To explore these issues, the Radioactivity Environmental Monitoring (REM) Group of the Institute for Environment and Sustainability at the Joint Research Centre (JRC) of the European Commission organises Spatial Interpolation Comparison (SIC) exercises. They form part of the JRC's institutional support to Commission Services, DG TREN in Luxembourg. In these exercises, participants are invited to estimate values of a variable observed at N locations with the help of N-n observed measurements. Once the participants have made their estimates, REM discloses the true values observed at all N locations, so that the participants may assess the accuracy of their approach.

The first SIC exercise, held in 1997 (EUR 2003), mainly addressed the state of the art of spatial interpolation at that time to highlight developments in spatial statistics as well as the large impact of human factors on the results obtained. In this exercise, called SIC 2004, participants were invited to fully automate their interpolation algorithms on the basis of some training data sets made available in advance. Moreover, the methods proposed had to describe how to calculate estimates and their associated uncertainties in the shortest period of time and, to address the behaviour of these mapping algorithms in emergency situations, be able to respond to situations in which extreme values are encountered.

Twelve manuscripts in which participants presented their results have been selected by the Editorial Committee for publication in a special issue of the journal *Applied GIS*[1]. This report presents extended abstracts of the twelve contributions as well as some overview papers from invited authors, mainly from the Editorial Committee.

As with the SIC 1997 exercise, readers will not find a unique, "best" method recommended herein, since no generalisation can be made from a single case study applied to a specific problem. They will nevertheless find that results tend to be similar for routine situations, independently of the choice of algorithm. On the other hand, when encountering a few extreme values, the algorithms have yielded a wide range of results, highlighting the many obstacles that still remain before one can rely on mapping systems that will run automatically in emergency situations. This is particularly clear during the early and critical stages of an accident, when measurements are usually sparse and statistical methods unable to cope with the lack of information.

More information on these Spatial Interpolation Comparison exercises and access to the data used may be found on the web site of AI-GEOSTATS (see the "Events" section of www.ai-geostats.org).

References

EUR (2003). *Mapping Radioactivity in the Environment. Spatial Interpolation Comparison 1997*. EUR 20667 EN, EC. 268 pp. Dubois, G., Malczewski, G., and De Cort, M. (eds). Office for Official Publications of the European Communities, Luxembourg.

---

[1] *Applied GIS*, Vol. 1, No. 2: http://publications.epress.monash.edu/toc/ag/2005/1/2

REVIEWERS AND EDITORIAL COMMITTEE

Samy Bengio

Machine Learning Group
IDIAP Research Institute
Switzerland

Dan Cornford

Neural Computing Research Group
Aston University
United Kingdom

Grégoire Dubois

Radioactivity Environmental Monitoring
Institute for Environment and Sustainability
Joint Research Centre
European Commission
Italy

Stefano Galmarini

Radioactivity Environmental Monitoring
Institute for Environment and Sustainability
Joint Research Centre
European Commission
Italy

Pierre Goovaerts

BioMedware Inc.
Ann Arbor, Michigan
USA

Gerard Heuvelink

ALTERRA & Laboratory of Soil Science
and Geology
Wageningen University & Research Centre
The Netherlands

Mikhail Kanevski

Institute of Geomatics and Risk Analysis
University of Lausanne
Switzerland

Jürgen Pilz

Applied Statistics Group
Institute of Mathematics and Statistics
University of Klagenfurt
Austria

TABLE OF CONTENTS

Discussion papers

Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results

G. Dubois and S. Galmarini

Radioactivity Environmental Monitoring, Institute for Environment and Sustainability, Joint Research Centre, European Commission, Via E Fermi, 21020 Ispra (VA), Italy

E-mail: gregoire.dubois@jrc.it

Abstract: The 2004 edition of the Spatial Interpolation Comparison (SIC) exercise dealt with automating spatial interpolation algorithms for environmental monitoring systems designed for routine and emergency Situations. Real-time mapping of critical variables for the environment is still a very rudimentary discipline, often implemented as black boxes, and more advanced mapping methods exist that can provide more robust maps with less uncertainty. In this exercise, participants were invited to test their algorithms on data sets made available by the organizers. This paper describes the design of the exercise in detail and gives an overview of the results obtained by the participants.

Keywords: spatial interpolation, automation, real-time mapping, environmental monitoring, emergencies.

1. INTRODUCTION

Monitoring networks regularly collect and report local observations of a variable that need to be converted into information with spatial continuity, i.e., maps which are frequently essential for decision-making and/or further modelling. Ideally, these maps should be produced automatically in order to allow real-time assessments of the phenomenon monitored and to minimize human intervention in case of emergency. However, the task of automating the spatial interpolation step is everything but straightforward (Dubois and Shibli, 2003). First of all, no uniformly best mapping algorithm exists because each function has its advantages and drawbacks. Second, most functions require a number of parameters to be set, even arbitrarily.

In order to explore the factors to be taken into account when designing environmental monitoring systems, the Radioactivity Environmental Monitoring Group (Institute for Environment and Sustainability, Joint Research Centre, European Commission) has organised Spatial Interpolation Comparison (SIC) exercises that were originally inspired by Englund's work (1990). In these exercises, participants are provided with a set of n measurements at fixed locations and invited to estimate values assumed by the variable at a number of $N\text{-}n$ locations. The real values observed at these $N\text{-}n$ sampling locations are revealed only at the end of the exercise in order to assess the relative performance of the proposed algorithms (Figure 1). While the first SIC exercise (SIC97, see Dubois *et al.*, 2003) allowed participants to interact as much as they wanted with the data and the algorithms, this second exercise (SIC2004) focused on automating the interpolation steps.
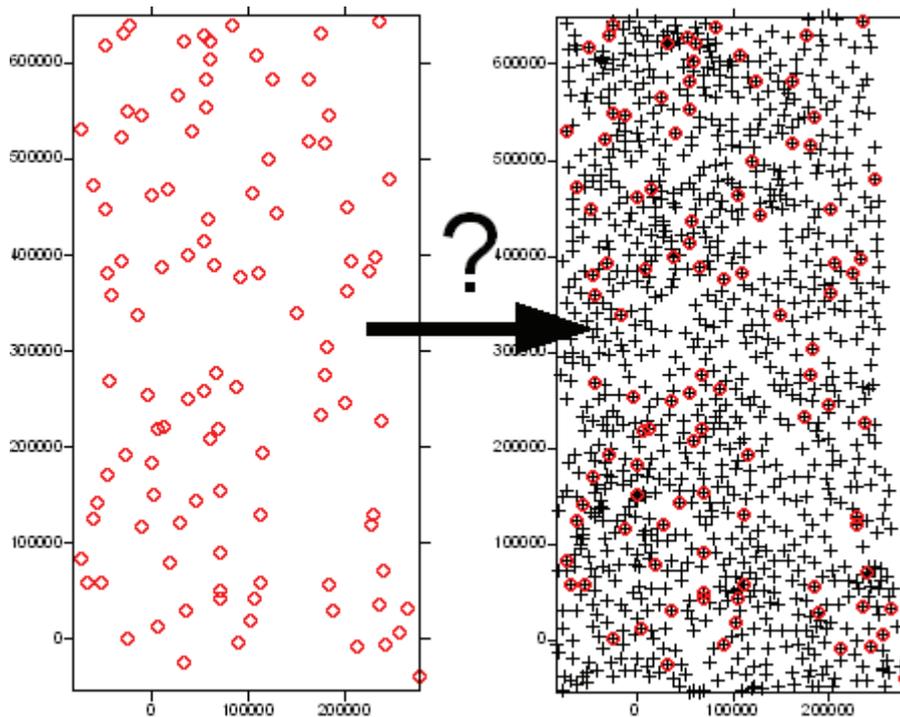
Figure 1. *Description of the concept of SIC exercises: participants are invited to minimize errors when using n observations (left, circles) to estimate values located at N locations (right, crosses). The left map shows the sampling locations of 200 training data used in the frame of SIC2004 and the right map shows 1008 point locations for which estimated values were requested.*

Automatic mapping algorithms are much needed, in particular during emergencies when researchers and decision-makers will need to concentrate on more critical issues than selecting appropriate interpolation parameters. In routine situations, one will understand that researchers and/or authorities can not generate maps in real-time on a 7-day/24-hour basis. Hence, automated algorithms may be very useful in these cases as well.

For these reasons, two scenarios were explored in this exercise. In the first, a routine situation was mimicked, that is the variable monitored behaved without showing strong real-time spatio-temporal fluctuations, as is usually the case, for example, with ozone levels, natural background radioactivity, seismic activity and rainfall fields. The second scenario, in contrast to the previous one, simulated an environmental emergency that is frequently characterized by a sudden and abrupt change in space and time of the reported levels of the monitored variable. This is the case for a number of natural and/or technological disasters such as accidental releases of chemicals or radioactivity, earthquakes, flash-floods, peaks of pollution, etc. To ensure that environmental monitoring systems are able to cope with both routine and emergency situations, real-time mapping algorithms need to be able to adapt to both scenarios without any human intervention. With the request to design a fully automated algorithm, the need to design a single method to cover both cases was the second challenging issue of this exercise.

Practically, in the frame of this exercise, in order to allow participants to design their algorithms properly, information about the investigated variable and about the topology of the monitoring network was provided in advance.

2. DATA DESCRIPTION

The data used in the frame of SIC2004 are measurements of gamma dose rates that have been extracted from the European Radiological Data Exchange Platform (EURDEP, see http://eurdeppub.jrc.it/) database (De Cort and De Vries, 1997). EURDEP is a system developed by the Radioactivity Environmental Monitoring (REM) group to make European radiological monitoring data available to decision-makers. From this database, 10 sets of mean daily values were selected from 2003, by drawing roughly one day at random from each month. A further filtering of these data was applied to select only measurements reported by the German national automatic monitoring network (IMIS) of the Federal Office for Radiation Protection (BfS, http://www.bfs.de/). This selection ensured that the data were homogeneous in terms of measurement technique and that the densest monitoring network in Europe, i.e. the German one, was included. From around 2000 monitoring stations in Germany, 1008 stations (their relative locations are shown in the right part of Figure 1) were selected by drawing a rectangular window. These stations were common to each of the 10 datasets, and all values reported for each day were selected.

Two types of data were provided:

1) coordinates and observations made of the variable X at n fixed locations,

2) and geographical coordinates only for the N-n locations at which values of X should be estimated through the mapping algorithm.

2.1. *Training sets*

From these 1008 monitoring locations, a single sampling scheme of 200 monitoring stations was selected randomly and extracted for each of the 10 datasets, in order to allow participants to train and design their algorithms. These data were made available as text files in which the first column held an integer value used to identify each station uniquely, the second and third columns contained the relative X and Y coordinates (in metres) and the last column gave the gamma dose rate measured in nanoSievert per hour (nSv/h). These 200 sampling locations have a spatial distribution that can be considered as nearly random (Figure 1, left). From the summary statistics in Table 1, one can see that the subsets of 200 points are representative of the whole set of 1008 points. More details regarding these datasets can be found in Dubois & Galmarini (2005).

Set No.	Statistics for the training sets ($n = 200$)					Statistics for the full sets ($N = 1008$)				
	Min.	Mean	Median	Max.	Std. dev.	Min.	Mean	Median	Max.	Std. dev.
1	55.8	97.6	98.0	150.0	19.1	55.0	98.9	99.5	193.0	21.1
2	55.9	97.4	97.9	155.0	19.3	54.9	98.8	99.5	188.0	21.2
3	59.9	98.8	100.0	157.0	18.5	59.9	100.3	101.0	192.0	20.4
4	56.1	93.8	94.8	152.0	16.8	56.1	95.1	95.4	180.0	18.8
5	56.4	92.4	92.0	143.0	16.6	56.1	93.7	94.0	168.0	18.1
6	54.4	89.8	90.4	133.0	15.9	54.4	90.9	91.6	168.0	17.2
7	56.1	91.7	91.7	140.0	16.2	56.1	92.5	92.9	166.0	16.9
8	54.9	92.4	92.5	148.0	16.6	54.9	93.5	94.1	176.0	18.1
9	56.5	96.6	97.0	149.0	18.2	56.5	97.8	98.7	183.0	19.9
10	54.9	95.4	95.7	152.0	17.2	54.9	96.6	97.1	183.0	19.0

Table 1. *Summary statistics for the 10 datasets used to train the algorithm used in SIC2004. Measurement units are nSv/h.*

The remaining 808 stations (that is, 1008-200= 808) were used to define the network topology where estimated values were requested. Only the coordinates of the remaining 808 monitoring stations were provided to the participants; no other information was released about the exhaustive datasets.

2.2. *Dataset of the 1ˢᵗ scenario: routine monitoring*

The first dataset consisted of an 11th dataset of 200 measurements which had coordinates that were identical to the previous 10 datasets used to train the algorithms. Summary statistics for this dataset as well as for the exhaustive dataset are summarized in Table 2. A frequency histogram for the 200 measurements is given in Figure 2.

	Min.	Mean	Median	Max.	Std. dev.	Skew.	Kurtosis
1ˢᵗ input data (*n*= 200)	58.2	96.2	97.6	153.0	17.6	0.1	-0.3
Exhaustive set 1 (*N* = 1008)	57.0	97.7	98.6	180.0	19.6	0.4	0.6

Table 2. *Summary statistics for the first dataset used for the SIC2004 exercise.*
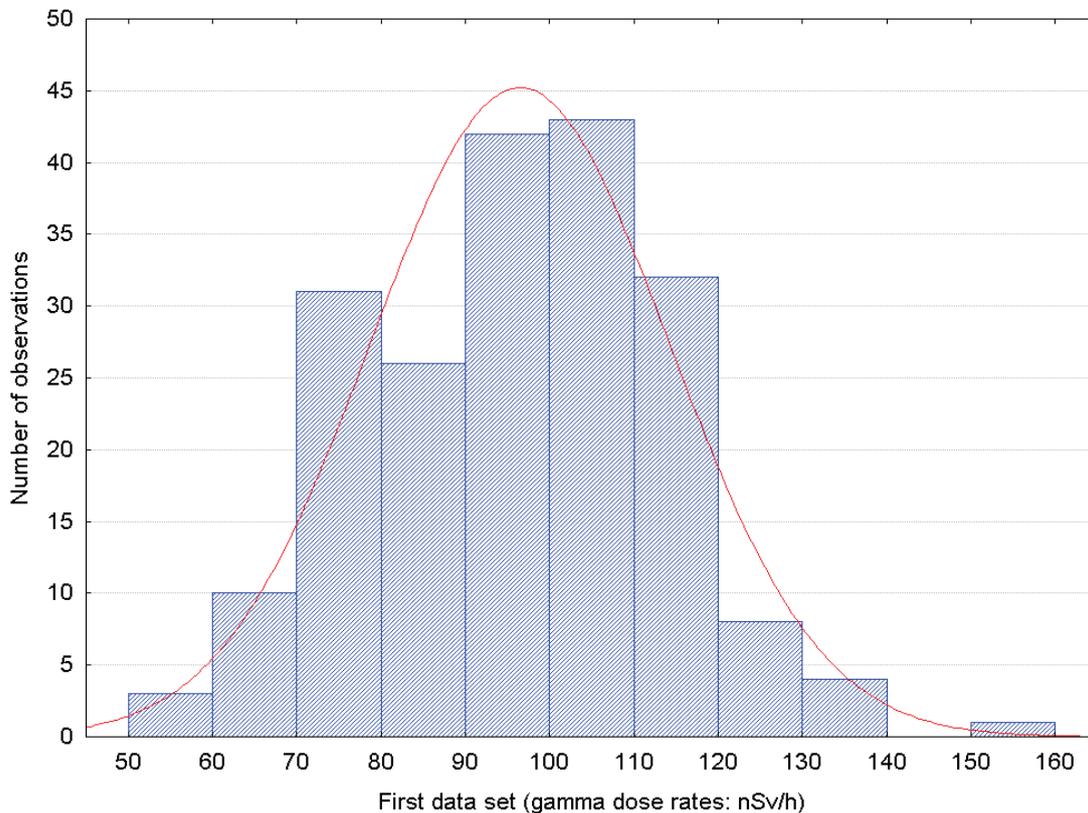
Figure 2. *Frequency histogram of the first data set used for the routine monitoring scenario.*

2.3. *Dataset of the 2ⁿᵈ scenario: monitoring in emergencies*

To explore how the interpolation algorithms would respond to extreme events, a second test dataset, nicknamed "*Joker*", was also submitted to the participants. The "*Joker*" simulated an accidental release of radioactivity into the atmosphere. A small corner located SW of the monitored area was chosen, and a dispersion process was modelled in order to obtain a few values on the order of 10 times more than the overall background levels reported for the first

dataset (see Annex 1 for details on the diffusion model used). As a result, we have the following summary statistics for the second data set (Table 3). The corresponding frequency in Figure 3 shows the strong impact of the simulated release on the statistical distribution of the data.

	Min.	Mean	Median	Max.	Std. dev.	Skew.	Kurtosis
2nd input data ($n = 200$)	58.2	109.0	97.9	1499.0	122.0	10.0	104.4
Exhaustive set 2 ($N = 1008$)	57.0	106.1	98.9	1528.2	92.5	11.3	144.1

Table 3. *Summary statistics for the second ("joker") dataset used for the SIC2004 exercise.*
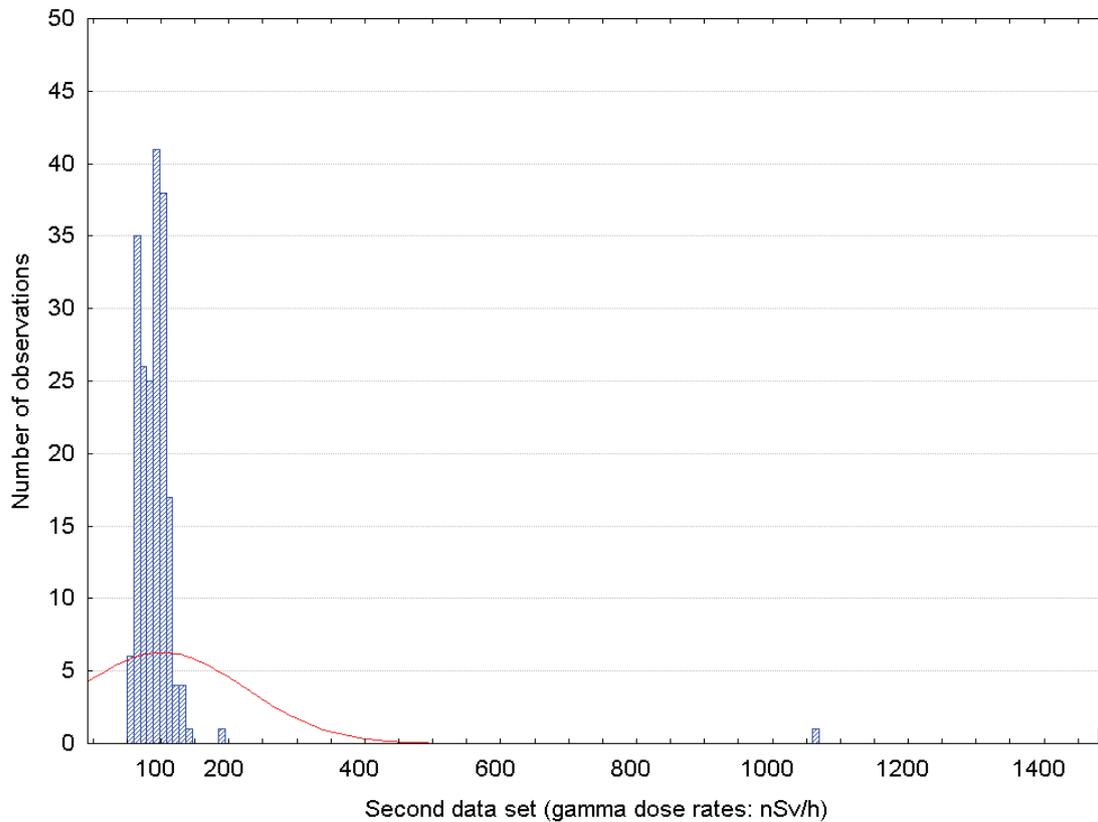
Figure 3. *Frequency histogram for the second dataset (also called "joker") used for the emergency monitoring scenario.*

To further highlight the differences between the two data sets of the exercise, two maps calculated on the basis of the exhaustive sets and are shown in Figure 4. One can clearly see from this figure that the overall background structure of the dose rates remained unaffected and that the simulated release is very local.
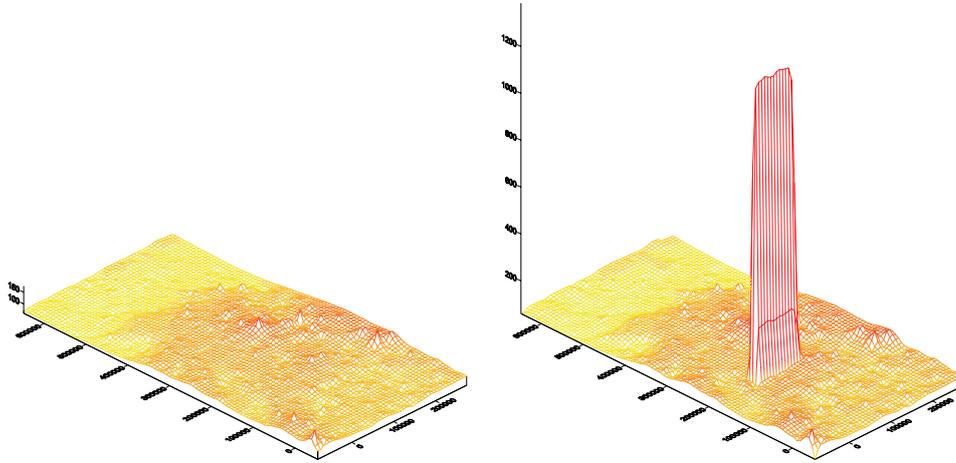
Figure 4. *Isolines (nSv/h) for the 1ˢᵗ data set (left) used in the routine scenario and the 2ⁿᵈ set (right) for the emergency scenario.*

3. THE SPATIAL INTERPOLATION COMPARISON EXERCISE 2004

While we used the exhaustive dataset to plot the maps shown in Figure 4, the participants of the exercise had only 200 measurements to produce similar results. To properly assess the results, a number of evaluation criteria were defined initially, as explained hereafter.

3.1. *Computational objectives of SIC2004*

The framework in which participants had to design their algorithms was defined by the following computational objectives. The mapping algorithms had to generate results:

- in a minimum amount of time (e.g. a few hours);

- without any human intervention, in the sense that only manual downloading and uploading of the data used for the exercise were allowed;

- that were "reasonable", that is the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) had to be kept as low as possible.

To properly verify that the first two criteria were met by the participants, a deadline was imposed by which they had to provide the organizers with all the information about their algorithms and the values assumed by the parameters involved. It is only when the participants had communicated the design of their algorithms that they were given access to a restricted web site that stored the data files used for the exercise. This web site was further used to register the download time of the data and the upload time of the results. When all participants had sent their results, the organizers released the true values to allow the participants to assess the performance of their algorithms by calculating mean errors.

3.2. *Results of SIC2004*

A few participants failed to submit results and withdrew from the exercise. The main reason for withdrawing was that the "joker" data set generated unforeseen numerical instabilities in the algorithms and failed to generate meaningful results. It is also clear that the task of designing algorithms able to predict the unpredictable was very challenging: in fact, the organizers of SIC2004 had a clear idea of the exercise but released nevertheless only a simple statement explaining that the algorithms needed to be able to deal with extreme events.

As a final result of the exercise, a total of 12 manuscripts that reported both the successes and failures of the participants were accepted by the Editorial Committee for publication in a special issue (Volume 1, Number 2) of the journal *Applied GIS*[2]. This report therefore contains only extended abstracts of the accepted papers as well as the core results.

In Table 4 the reader will find the participants' results for either of the two scenarios, that is routine and emergency. The results have been sorted by minimum MAE obtained in the case of the emergency scenario. Other statistics shown in this table are the Mean Error (ME) that allows the readers to assess the bias of the results, the RMSE, as well as Pearson's correlation coefficient *r* between true and estimated values.

First authors	Method Type	MAE (1)	MAE (2)	ME (1)	ME (2)	RMSE (1)	RMSE (2)	*r* (1)	*r* (2)
Timonin	NN	9.40	14.85	-1.25	-0.51	12.59	45.46	0.78	0.84
Fournier	GEOSTATS	9.06	16.22	-1.32	-8.58	12.43	81.44	0.79	0.27
Pozdnoukhov	SVM	9.22	16.25	-0.04	-6.70	12.47	81.00	0.79	0.28
Saveliev	SPLINES	9.60	17.00	3.00	10.40	13.00	82.20	0.77	0.23
Dutta	NN	9.92	17.50	0.20	5.10	13.10	80.60	0.76	0.29
Ingram	GEOSTATS	9.10	18.55	-1.27	-4.64	12.46	54.22	0.79	0.86
Hofierka	SPLINES	9.10	18.62	-1.30	0.41	12.51	73.68	0.79	0.50
Hofierka	SPLINES	9.10	18.62	-1.30	0.41	12.51	73.68	0.79	0.50
Fournier	GEOSTATS	9.22	19.43	-0.89	-0.22	12.51	73.50	0.78	0.48
Fournier	OTHERS	9.29	19.44	-1.12	-0.12	12.56	71.87	0.78	0.53
Savelieva	GEOSTATS	9.11	19.68	-1.39	-2.18	12.49	69.08	0.78	0.56
Palaseanu	GEOSTATS	9.05	19.76	1.40	2.33	12.46	74.54	0.79	0.50
*Fang**	*OTHERS*	*9.67*	*19.91*	*-1.29*	*3.26*	*13.21*	*66.80*	*0.75*	*0.61*
Rigol S.	NN	12.10	20.30	-1.20	-9.40	15.80	84.10	0.67	0.12
Pebesma	GEOSTATS	9.11	20.83	-1.22	0.92	12.44	73.73	0.79	0.50
Pebesma	*OTHERS*	9.94	21.03	-1.35	4.50	13.32	72.12	0.78	0.51
Ingram	GEOSTATS	9.08	21.77	-1.44	0.72	12.47	79.57	0.79	0.35
Lophaven	GEOSTATS	9.70	22.20	1.20	-4.10	13.10	71.20	0.76	0.54
Saveliev	SPLINES	9.30	22.20	1.60	0.60	12.60	76.40	0.78	0.41
Ingram	GEOSTATS	9.47	22.53	-1.15	3.09	12.75	79.16	0.78	0.33
Pebesma	GEOSTATS	9.11	23.26	-1.22	4.00	12.44	76.19	0.79	0.42
Rigol S.	NN	16.00	25.30	-1.70	-11.10	20.80	87.50	0.55	0.02
Hofierka	SPLINES	9.38	26.52	-1.27	4.29	12.68	77.98	0.78	0.38
Dutta	NN	9.62	28.20	0.90	-0.22	12.70	80.10	0.78	0.31
Pebesma	GEOSTATS	9.11	28.45	-1.22	12.01	12.44	81.41	0.79	0.38
Dutta	NN	12.20	28.90	1.50	-1.29	15.90	79.90	0.64	0.33
Rigol S.	NN	21.40	30.50	5.30	3.80	45.80	96.60	0.24	0.20
Ingram	NN	9.72	38.29	-1.54	8.38	13.00	84.24	0.76	0.30
Dutta	NN	9.93	38.50	2.18	17.98	13.30	87.30	0.76	0.27
Ingram	NN	9.48	48.41	-1.22	-3.01	12.73	90.89	0.78	0.38
Pebesma	GEOSTATS	9.11	146.36	-1.22	19.71	12.44	212.10	0.79	-0.27

*Unpublished. See http://www.fanginc.com/rdic/sic1.pdf

Table 4. *Summary table with the results of the SIC2004 exercise for both the routine (1) and the emergency (2) scenarios. MAE= Mean Absolute Error, ME = Mean Error or bias, RMSE= Root Mean Squared Error, r = Pearson's correlation coefficient r between true and estimated ones. GEOSTATS denotes Geostatistical techniques, NN Neural Networks, SVM Support Vector Machine. In each column, the best results have been shaded.*

---

[2] http://publications.epress.monash.edu/toc/ag/1/2

One will notice from the table that a few authors provided more than one result. Most authors have indeed tried out various algorithms after having submitted their results to better assess the drawbacks and advantages of each. All results have been carefully presented and discussed in detail in *Applied GIS*.

Before looking further into these results, we give a few words of caution regarding comparable exercises, and refer the reader to the contributions made by Dan Cornford, Gerald van den Boogaart and Donald Myers, all included in this report. In short, it is clear that no generalisation should be made from a single case study and that results can be compared in many different ways and from many different perspectives.

4. OVERVIEW OF THE RESULTS

4.1. *Methods*

Looking at the methods used, one will realize that the participants chose mainly two types of functions: geostatistical functions and neural networks (NN). This relatively limited number of approaches contrasts with the observation made in the SIC97 exercise (Dubois *et al.*, 2003) in which no automation was required. Participants to this previous edition were keener to try out a larger set of interpolation algorithms (using genetic algorithms, fuzzy logic, etc.) as they had enough time to submit only estimates in which they were confident enough.

4.2. *Results of the routine scenario*

In the first scenario, that is of routine monitoring, almost all of the 31 algorithms submitted gave very similar results, and the correlation between estimated and true values was good in almost all cases. Pearson's correlation coefficient *r* between true values and estimated reached more than 75%; however, some algorithms using neural networks showed poor performance. Looking at the MAE, one will find that Palaseanu obtained the best result overall (MAE = 9.05), while for SVM and NN the best results were, respectively, ranked 12th (Pozdnoukhov, MAE = 9.22) and 17th (Timonin, with MAE = 9.40). Fournier and Furrer generated the lowest RMSE (with respectively RMSE = 12.43 and RMSE = 12.44) with their robust kriging while the best results obtained by Support Vector Machine (SVM) algorithms and NN are ranked 9th (RMSE = 12.47) and 15th (RMSE = 12.59), respectively. Interestingly enough, SVM also generated the lowest bias (ME = -0.04), most results showing larger negative values.

From Figure 5, which may be seen as a rough summary of the results in the routine scenario, one will conclude that the methods generated similar estimation results, except for a few outliers of poor performance stemming from NN.
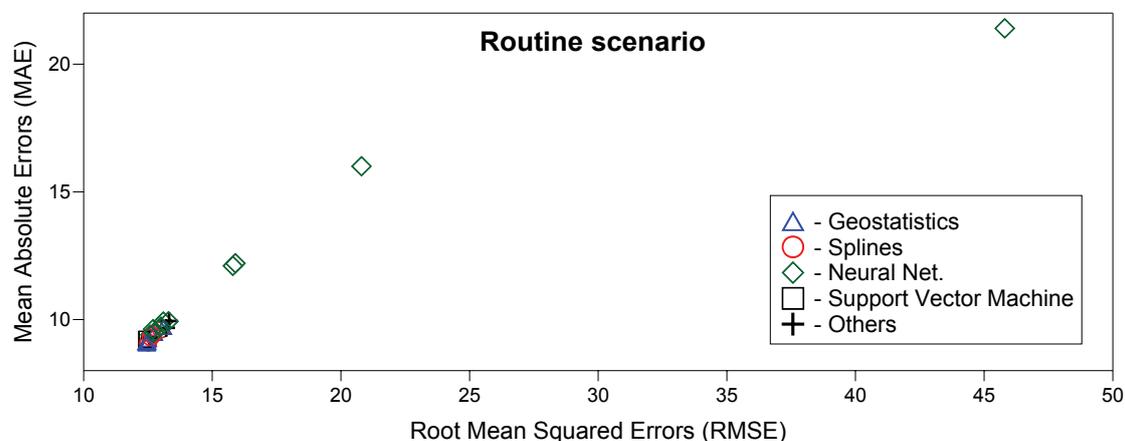
Figure 5. *MAE and RMSE of the methods used in the frame of SIC2004 (routine monitoring)*

4.3. *Results of the emergency scenario*

Clearly, the emergency case destabilized most of the algorithms and much more variability can been found in the results (see Figure 6) as compared to those obtained for routine monitoring. The correlation ranged from 2% to 86% (Ingram) and most values showed poor correlation (below 50%) between estimated and observed values.

In a number of cases, geostatistical techniques failed to generate results because the algorithm set to fit a model of the spatial correlation, a requirement in geostatistics for deriving weights used for the estimation step, did not find any positive definite model (see Armstrong and Jabin, 1981). Consequently, unresolved equations led to no results. Contrary to the first scenario, in which most participants got very similar results, the one discussed here saw clearly a "winner" as Timonin and Savelieva obtained both the lowest MAE (= 14.85) and RMSE (= 45.46). These results were followed by the robust kriging approach by Fournier and Furrer for what concerns the MAE (MAE = 16.22) and Ingram and Csato for the second lowest RMSE (RMSE = 54.22).

4.4. *Overall results*

Geostatistical functions confirmed their reputation for robustness, and, independently of the approach adopted here, they all generated good results in both exercises. Splines generated results that are similar to those obtained by geostatistical techniques. Interestingly enough, although kriging and splines are equivalent in many ways (see Dubrule, 1984; Hutchinson and Gessler, 1994), the critical dependence of geostatistical methods on a model of the covariance function did not appear to hamper these functions in the emergency scenario when compared to splines.
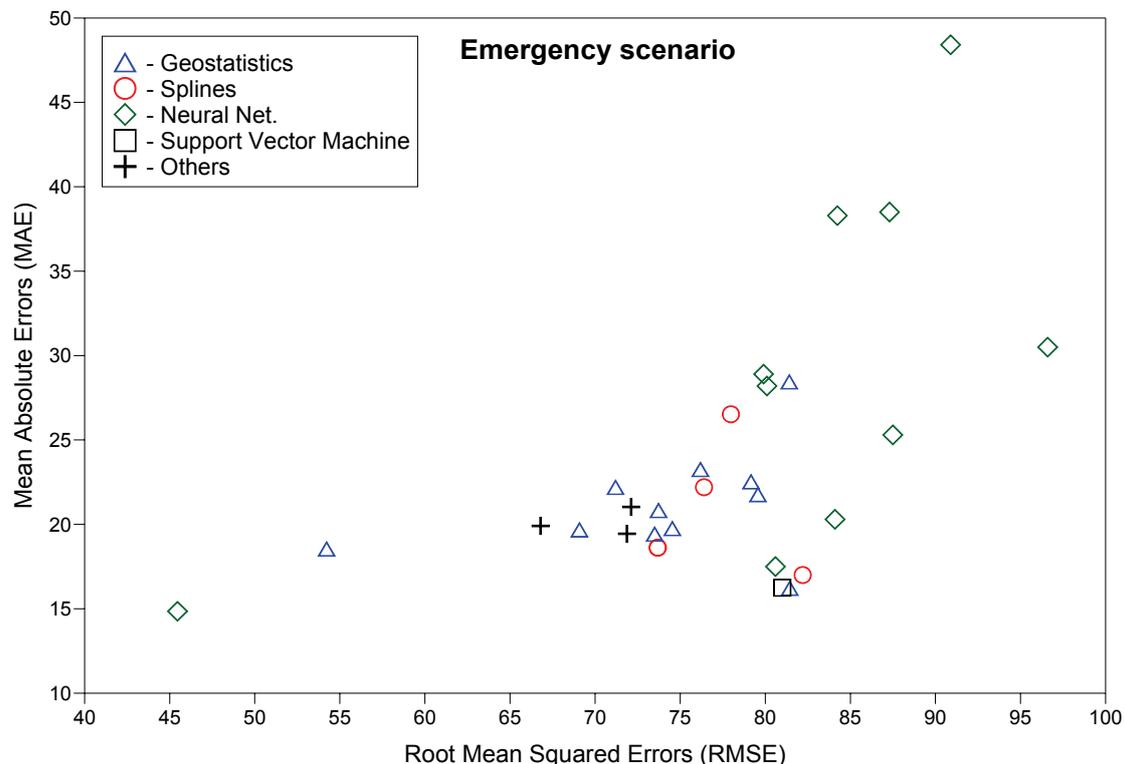
Figure 6. *MAE and RMSE of the methods used in the frame of SIC2004 (emergency monitoring); note the difference in scales from Figure 5. For graphical reasons, one point with MAE > 50 and RMSE > 100 was discarded.*

NN algorithms gave estimation results that range from the best to the worst cases. This highlights the difficulty in training such a type of algorithm, in particular because the number of measurements used here for training was not very large. We refer the reader Nicolas Gilardi and Samy Bengio's paper below for a discussion of the use of NN and machine learning algorithms.

5. CONCLUSIONS

Before coming to any sort of general conclusion, we will reiterate our warning to the readership. The lessons learned in this exercise should be taken with much caution as it is a single case study dealing with a very limited number of issues. The participants were informed about the need to design and tune their automatic mapping algorithms in order to have functions able to deal with extreme events: not more, not less. Extrapolating these results and observations to other problems related to the monitoring of environmental variables that can present extreme events can thus be very hazardous. It is, however, also true that there is a clear need for real-time mapping algorithms for environmental monitoring systems and for tools designed for decision support. In the best cases, monitoring systems have either very dense networks that will allow the use of simple mapping functions or, should the monitored phenomenon be well known, compensate the lack of data by using a modelling approach. There are obviously many other situations in which the monitored phenomenon is very complex and/or the network is very sparse. For these cases, (geo)statistical methods have proven to be very useful to make the maps generated more robust and minimise uncertainty. The boundaries for the need for statistical tools and physical models clearly overlap each other, and the reader might enjoy reading Stefano Galmarini's concluding paper in this report that discusses these aspects in the context of atmospheric modelling. In the frame of the Spatial Interpolation Comparison exercise, only the statistical challenge of generating maps automatically on the basis of data characterized by very different statistical distributions was explored. The outcome of this exercise highlighted a number of issues and problems related to the automation of such algorithms, and further efforts need to be made to understand better the obstacles in setting up such systems.

Real-time mapping of a variable that undergoes only small changes in space and time seems to be achievable without too much difficulty. This may not sound as a big surprise but we will reiterate the fact that the methods applied here are relatively advanced and that all can provide some information about uncertainty, something which traditional deterministic algorithms usually do not do. Hence, in the near future one may expect a new generation of automated environmental monitoring systems that will rely on geostatistics, neural networks and support vector machine functions for generating maps in real time.

For variables that can be critical to our environment when they exceed certain thresholds, the situation is much more complex, because the data will be characterized by statistical distributions, usually skewed, that may be very different from the distribution that is usually observed in routine situations. Not only is the automatic recognition of the statistical distribution a tricky problem, but so is the handling of a few extreme values. Most algorithms that are not exact interpolators would discard a few isolated extreme values as these functions generally tend to smooth out irregularities. This is somehow intrinsic to any statistical technique and one needs efficient logical filtering of the data prior to their statistical treatment. In other words, some expert knowledge will always be needed to help the automatic mapping system to distinguish false measurements from the first outliers that would signalise the early stage of a real hazard. Most statistical methods assume that the underlying distribution of the measurements is normal, and this factor currently prevents monitoring systems from offering better results and justifies more research in non-normal statistics.

Many other obstacles against building fully automatic mapping systems can be found, but, contrary to the situation for outliers and early-warnings, those discussed below may find a number of numerical solutions in the future. The spatial correlation model of the monitored variable can be automatically fitted, but the automatic selection of a shape (linear, exponential, spherical, normal etc.) of the model prior to its fitting still needs to be solved. In the case of strong spatial anisotropies of the monitored phenomenon, something that is typical in the early phase of an atmospheric release of a pollutant (e.g. normal plume), the spatial pattern of the variable will have a strong impact on the spatial correlation models up to a point to reach numerical problems if improperly defined. The automatic detection of such anisotropies is thus also an important issue that is currently in development as shown in this volume by Dionissios Hristopulos. Many other difficulties proper to (geo)statistical methods (e.g. hypothesis of stationary isotropic random fields, skewness of the statistical distributions, mixture of various populations, etc.) remain partly unresolved, but in the future one may expect new developments in these directions. Neural networks all have similar drawbacks and inconveniences, the over-fitting problem probably being the most challenging one as discussed by Nicolas Gilardi and Samy Bengio. The virtual "winners" of SIC2004, Timonin and Savelieva, got the lowest average errors using Generalized Regression Neural Networks. The implementation of their algorithm in a real early-warning system may generate constant false warnings or not detect any change in the data at all, depending on how much effort has been put in training the networks. The choice of the intensity of such training still remains very much based on the experience of the data used and these participants may have had less luck in other circumstances. Here as well, the behaviour of these functions under different constraints needs to be largely explored before they can become effective in emergency systems.

We will conclude by underlining a number of key topics that have not been discussed or developed in the frame of this exercise but that still should be brought to the attention of the reader. The whole issue of decision making, for example, has been left aside, and the management of data collected by environmental monitoring networks has been barely discussed. How can we distinguish false measurements from extreme values reported by early-warning systems? What are possible cost functions one would have to minimise when defining acceptable errors and uncertainty levels? Although limited to the case of radioactivity, the following chapter written by Chris Twenhöfel and his colleagues provides much information about typical and more subtle problems encountered when dealing with measurements reported in real-time by a network monitoring a variable critical to our environment.

One could also have debated the appropriateness of an exercise in which a regression (interpolation) problem had to be solved instead of a classification exercise in which safe and unsafe areas could have been defined instead. This requires different designs of the algorithms and will most probably be the topic of the next Spatial Interpolation Comparison exercise. The contribution of Pierre Goovaerts to this report highlights this point as he applied a regression technique for classification purposes. By using regression techniques to generate maps one may present contours that appear natural to our perception of the environment, whereas using classification techniques would probably be more adapted to decision making as it is of primary importance to delineate areas where a critical value is exceeded. For these reasons, the Indicator Kriging approach is a very interesting one for dealing with both types of problems.

Looking back at SIC97, one will realize that algorithms and experience in the field of spatial statistics have progressed very quickly. Still, not a single participant did use commercial GIS software available on the market, which would indicate either a lack of advanced mapping functions or of transparency in the algorithms implemented.

Acknowledgments

References

[1] De Cort, M. and de Vries, G. (1997). The European Union Radiological Data Exchange Platform (EURDEP): Two Years of International Data Exchange Experience. Radiation Protection Dosimetry, 73: 17-20.

[2] Dubois G. and Galmarini, S. (2005). Introduction to the Spatial Interpolation Comparison (SIC) 2004 exercise and presentation of the data sets. Applied GIS, **1**(2): 09-01 - 09-11.

[3] Dubois, G. and Shibli, S.A.R. (2003). Monitoring of environmental radioactivity: automatic mapping or expert-dependent systems? In: *"Mapping radioactivity in the environment. Spatial Interpolation Comparison 1997"*. Dubois, G., Malczewski, J. and De Cort, M. (Eds.). Joint Research Centre, European Commission. EUR 20667 EN, EC, pp. 253-268.

[4] Dubois, G., Malczewski, J. and De Cort, M. (Eds) (2003). *Mapping Radioactivity in the environment. Spatial Interpolation Comparison 1997.*. Joint Research Centre, European Commission. EUR 20667 EN, EC. 268 pp.

[5] Dubrule, O. (1984). Comparing splines and kriging. *Computers and Geosciences,* **10**: 327–338.

[6] Englund, E.J. (1990). A Variance of Geostatisticians. *Mathematical Geology*, **22**(4):417-455.

[7] Hutchinson, M.F. and Gessler, P.E. (1994). Splines - more than just a smooth interpolator. *Geoderma*, **62**: 45-67.

[8] Armstrong, M. and Jabin, R. (1981). Variogram models must be positive-definite. *Mathematical Geology*, **13**: 455-459.

Operation of the Dutch 3rd Generation National Radioactivity Monitoring Network

C.J.W. Twenhöfel, C. de Hoog van Beynen, A.P.P.A. van Lunenburg, G.J.E. Slagt, R.B. Tax, P.J.M. van Westerlaak and F.J. Aldenkamp

National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, the Netherlands.

E-mail: chris.twenhofel@rivm.nl

Abstract: The Dutch National Radioactivity Monitoring Network (NRM) measures airborne α and β activity concentrations and ambient dose rates in the Netherlands under routine and emergency conditions. The α/β monitors and data acquisition system were recently upgraded. An overview of the present 3rd generation network is given and the operational experience of the network over the first eight months of 2005 is reported. The upgraded system performs well. In the operational period 16 warnings were generated, none were related to nuclear accidents.

Keywords: monitoring networks, nuclear accidents, emergency management.

1. INTRODUCTION

Concerns after the Chernobyl accident has led in many countries to the development of early warning and monitoring systems for radioactive contaminations in the environment. The Dutch National Radioactivity Monitoring Network (NRM) was designed with these objectives in mind. When a radioactive cloud is observed, a warning message is generated which may trigger the national and local emergency response organisations.

In the Netherlands the national (off-site) nuclear emergency management is organised in the Unit Planning and Advice nuclear (EPAn). The National Institute for Public Health and the Environment (RIVM) plays a comprehensive role in the technical information structure of EPAn. During nuclear emergencies RIVM runs the Back Office for Radiological Information (BORI), one of three back-offices of EPAn. Primary task of BORI is to provide timely evaluations of the radiological situation based on measurements and model calculations. For this RIVM operates and maintains a number of advanced technical facilities [1], including the NRM automated network for the continuous surveillance of radioactivity in the environment. In addition to the emergency management task of the network, routine measurements are performed to monitor the exposure of the population to various sources of ionizing radiation (see e.g. ref. [2]).

Recently the α/β monitors and the data acquisition system were upgraded. In this paper the 3rd generation NRM is described and the operation under routine and emergency conditions is shortly evaluated.

2. THE DUTCH NATIONAL RADIOACTIVITY MONITORING NETWORK

2.1 *Objectives*

The primary task of the NRM is to provide early warnings against nuclear accidents. An increase of radiation levels above certain thresholds will issue a warning to the national and local authorities, which after careful validation may activate the national emergency plan for nuclear accidents [3]. The overall objectives of the NRM are summarised by:

1. The provision of early warnings against (major) nuclear accidents.

2. The assessment of the magnitude and geographical extent of a radioactive contamination in the air and on the ground during and after a nuclear accident.

3. The assessment of actual radiation doses to the population and its development in time.

Under regular (non emergency) conditions the network provides information on the natural background radiation levels in the Netherlands [2, 4] and it provides a reference value for other radiation measurement networks, e.g. the fence monitoring systems of the nuclear power plant (NPP) in Borssele (450 MWe) and the radioactive waste storage facility (COVRA) in Vlissingen. During nuclear emergencies the NRM data form the basis for the calculation and validation of actual radiation doses via the direct pathways external radiation and inhalation.

2.2 *Technical Description of the NRM*

The objectives of the NRM are realised by a dense network of 153 ambient dose rate monitors and 14 α/β air sampling monitors. Figure 1 shows a photographic impression of the locations of the equipment in the field.

The ambient dose rate monitors (γ-monitors) contain a Bitt RS03/X proportional counter tube for the detection of external irradiation. The counter is placed at 1 m above the ground in an area free from obstacles (in practice a free radius of 3 times the height of the nearby obstacle is aimed for) to reduce environmental influences on the measurements. The readings are expressed as ambient dose-equivalent rate (ambient dose rate), H*(10) in nSv/h. Measurements are averaged over a 10 minute interval and stored in a relational database at RIVM.

At 14 locations a Berthold BAI 9128 airborne radioactivity monitor measures aerosol bounded gross α and gross β activity in air. The monitor determines an artificial β concentration from the measured gross α and β activity, collected from a, 3-4 m³/h, outside air flow on a moving filter. The use of this filter complicates the response of the monitor for fast varying radioactive air concentrations since a complicated reconstruction is generally required to extract momentary air concentrations from deposited activity on the moving filter. The use of a compensation method for natural activity in the determination of artificial β concentrations results in a superior sensitivity when compared to the γ monitors. The resulting low warning threshold makes them well suited for generating early warnings. At all 14 monitoring locations the α/β measurements are complemented by ambient dose rate measurements via a Bitt RS02/RM10. The design and mounting position of these monitors differ slightly from the 153 Bitt RS03/X monitors: they originate from the 1st generation radioactivity network and will be replaced in the near future.

Air activity concentrations are converted to doses by calculations. Hereto nuclide specific measurements are carried out to provide the required information on nuclide composition. Two nuclide specific monitors were added to the network: an iodine monitor (Herfurth H1399) measures gaseous radioactive iodine concentration in air and a Canberra FHT 59 CP measures nuclide specific gamma activity of airborne particles. These monitors are located at RIVM, Bilthoven and measurements are performed on a continuous basis.

During emergencies the number of nuclide specific measuring sites is increased by the activation of nine so-called 'pilot-flame' institutes, located around the country and with two mobile measuring vans. The emergency measurement program is co-ordinated by RIVM. It provides for a determination of nuclide specific airborne and nuclide specific surface activity concentrations from deposition. The nuclide specific measurements and mobile facilities are not further discussed in this paper.

Figure 1. *Housing of one of the α/β air activity monitors (left). These locations are shared with the equipment of the Air Quality Measuring Network of RIVM. The figure on the right shows the location of an ambient dose rate monitor in Bilthoven.*

2.3 *Geographical coverage*

The density of the geographical distribution of the monitors is a compromise between costs and the effectiveness of which a (significant) ground or air contamination is detected in time and a warning generated. Figure 2 shows the distribution of NRM field locations of ambient dose rate, airborne activity, and the fixed measuring sites for nuclide specific measurements at RIVM and the supporting 'pilot-flame' institutes. This setup resulted in an acceptable overall detection probability for (significant) radioactive releases nearby (warnings generated by dose rate measurements) and those originating from far distances (warnings generated by the more sensitive artificial β measurement). Around Borssele nuclear power plant the density of ambient dose rate monitors is increased in rings at distances of about 5 and 10 km from the power plant. This high local monitoring density is required to increase the probability of detecting a release from the national power plants to an acceptable level [5]. Another increase in the monitoring density is found in the Dutch border region, especially close to the nearby foreign nuclear power plants.

Due to the relatively high initial and maintenance costs associated with the operation of the air sampling monitors the number of airborne activity monitors was limited to 14. This is considered a minimum number to allow the determination of a national overall picture of air contamination. Model calculations showed a detection probability above 95% for severe reactor accidents at locations more than 200 km from the Dutch border, within 1 hour after the radioactive cloud passes the Dutch border [6].

2.4 *IT-Infrastructure and Data Acquisition*

The NRM uses a central system to collect and display the measurement data and to alert the users in case of elevated radiation levels. The system is made up of two servers in a cluster configuration. A back-up system automatically takes over in case of a malfunction in the operational system.

The monitoring stations are equipped with a local processing unit that calculates the averages over 10 minute intervals and send out an alert in case a threshold is exceeded. The public telephone network is used for the collection of data from the 14 α/β stations. The 153 gamma stations communicate via the emergency telephone network (ETN). This network is separated from the public network but uses the same technology. In regular circumstances the data from all stations is collected once every hour. Using 10 modems the data acquisition cycle is complete in 15 minutes. The acquisition frequency can be increased during an emergency to the maximum of once every 15 minutes. Apart from this, data of one or more monitoring stations can be collected by hand at any moment.
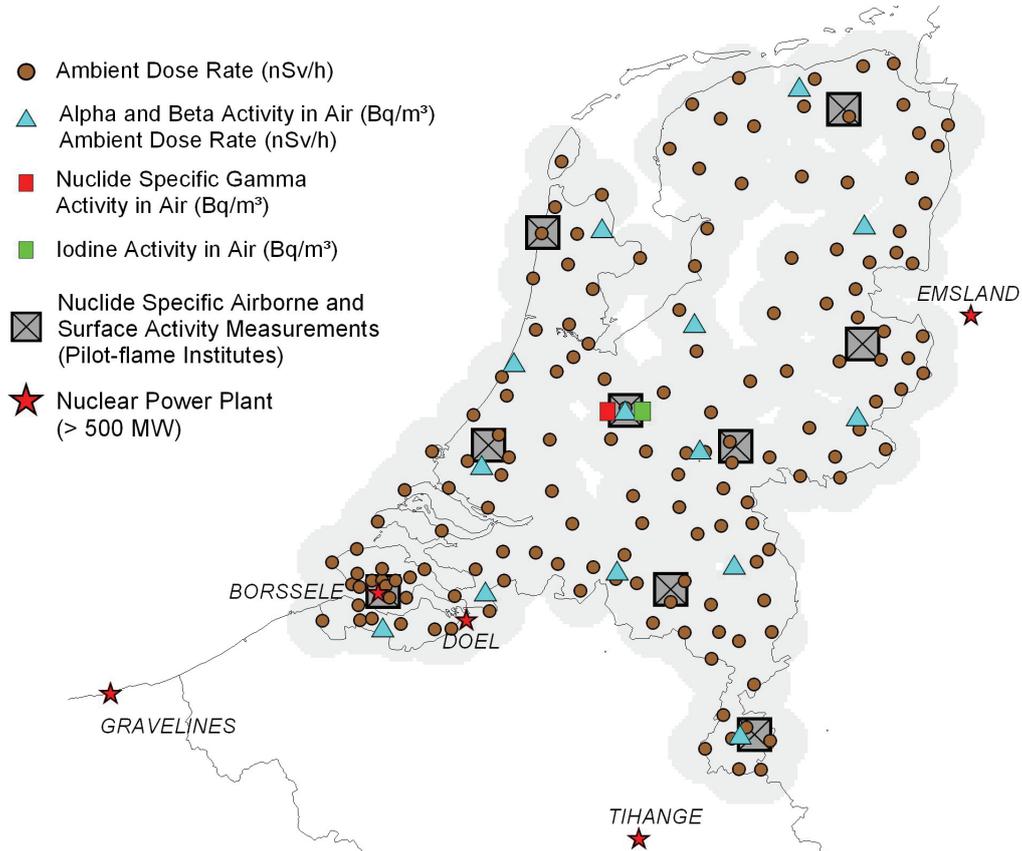
Figure 1. *Field locations of the 153 ambient dose rate monitors (dots) and the 14 locations of the α/β air sampling monitors (triangles). A 15 km proximity area to the nearest γ- monitor is indicated in grey. The rings around Borssele nuclear power plant (lower-left) are clearly visible. The dark grey squares represent the locations of the 'pilot-flame' institutes where additional nuclide specific measurements in air and deposition are performed.*

In the course of 2005 the connection to the 14 α/β monitoring stations will be converted from the public telephone network to a dedicated IP-VPN, the 153 gamma monitoring stations will follow in 2006. The main reason for this conversion is the impending discontinuance of the emergency telephone network in the Netherlands. An additional advantage over the current method of communication is that a direct network connection makes the monitoring station continuously on-line. This means that the central system can communicate directly with the radiation monitors. The use of a local processing unit is no longer necessary as is the need to send alerts from the monitoring stations, since these can now be generated by the central system.

A web server is used to present the data to the users. Users connected to the local area network (LAN) of the RIVM have a direct connection to the server. A dedicated IP-VPN is available for users belonging to the regional emergency response team (e.g. the fire brigades). This IP-VPN is based on Internet technology but uses a separate, private physical network. This network is managed and monitored by the provider and the availability and capacity of all connections guaranteed. Other privileged users can contact the web server through the public telephone network (analogue or ISDN), the ETN or a VPN over Internet.
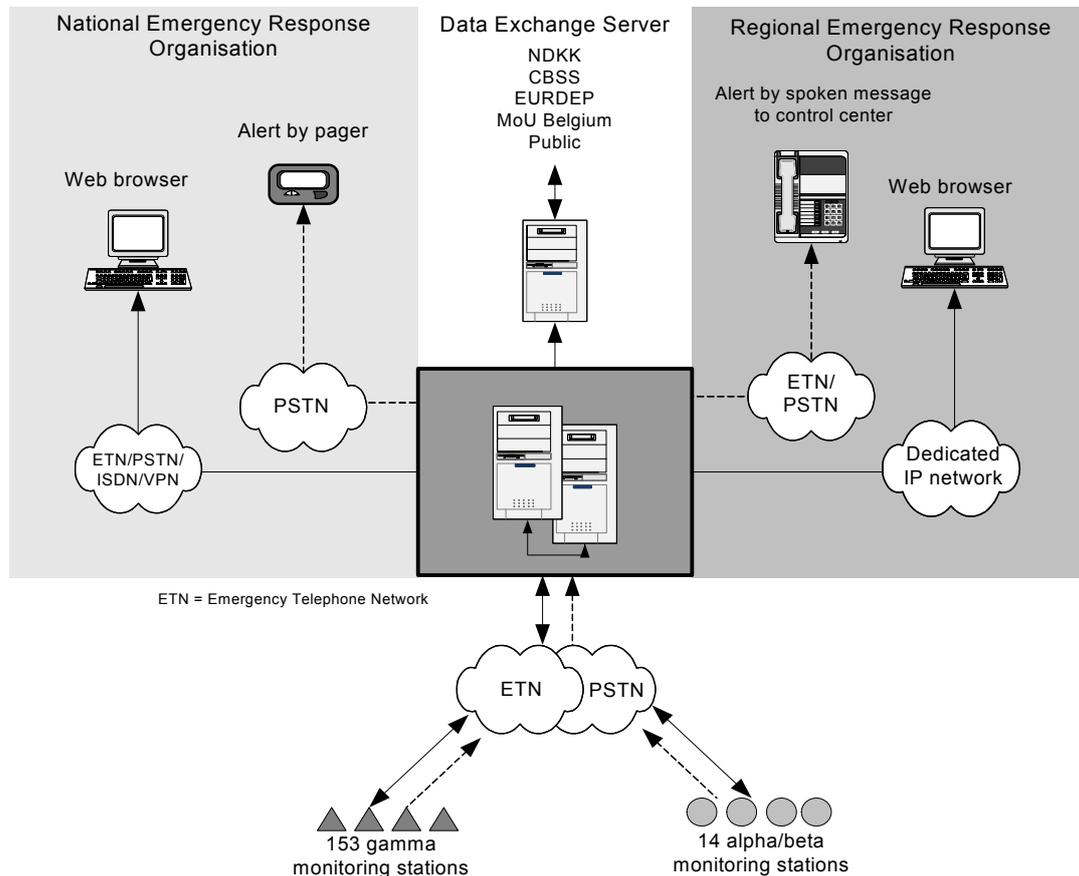
Figure 2. *Infrastructure of the NRM.*

The central server also provides data for external users. By using the "export" facility it is possible to select data based on a time period and/or monitoring stations and export this data in different formats such as CSV, XML and EURDEP. This last format is used to send data to the European Data Exchange Platform, EURDEP [7] hosted by the Joint Research Centre in Ispra, Italy. Other data exchange protocols exist or are being developed; examples are the bilateral agreement between the Netherlands and Germany under the Dutch-German commission for nuclear facilities in the border region, the NDKK [8], the memorandum of understanding between the Netherlands and Belgium and the upcoming agreement with the CBSS (Council of the Baltic Sea States) member states.

An additional but important feature of the system is the possibility to conduct exercises by supplying the system with simulated data and alerts. Only the users that are selected to participate in the exercise have these simulated data presented. All other users see the real data and alerts so the normal operation of the system is unimpaired.

2.5 *Warning procedure*

An important task of the NRM is to provide early warning against severe radiological accidents. All measurements are therefore compared to warning levels stored in de local processing unit. When a level is exceeded the officer(s) on duty is alerted immediately by the central system by means of a call to a pager or a spoken message by telephone. When it is clear that the warning message is not a result of technical deficiencies, a senior officer may further decide what actions are to be taken. In case of a severe radiological accident this will activate the national emergency response organisation.

To make this possible the monitoring stations can send an alert to the central system. This function is required because in normal circumstances the monitoring stations are contacted only once every hour, which would delay the alert for up to one hour. When an alert is received by the central system, an automatic read-out of all monitoring stations is performed to assure that the most recent data is available to the users.

The applied warning levels must be balanced between the capability to detect small contaminations and an acceptable number of false alarms. The limiting factor is the natural fluctuation in the background. For ambient dose rate the long term average in the Netherlands is between 55 – 115 nSv/h [9, 10] with superimposed enhancements typically up to 80 nSv/h [9]. Only under extreme conditions higher elevations are observed. The current maximum recorded elevation by the NRM is 165 nSv/h. It was caused by heavy rainfall in Rijssen on August 11, 2002.
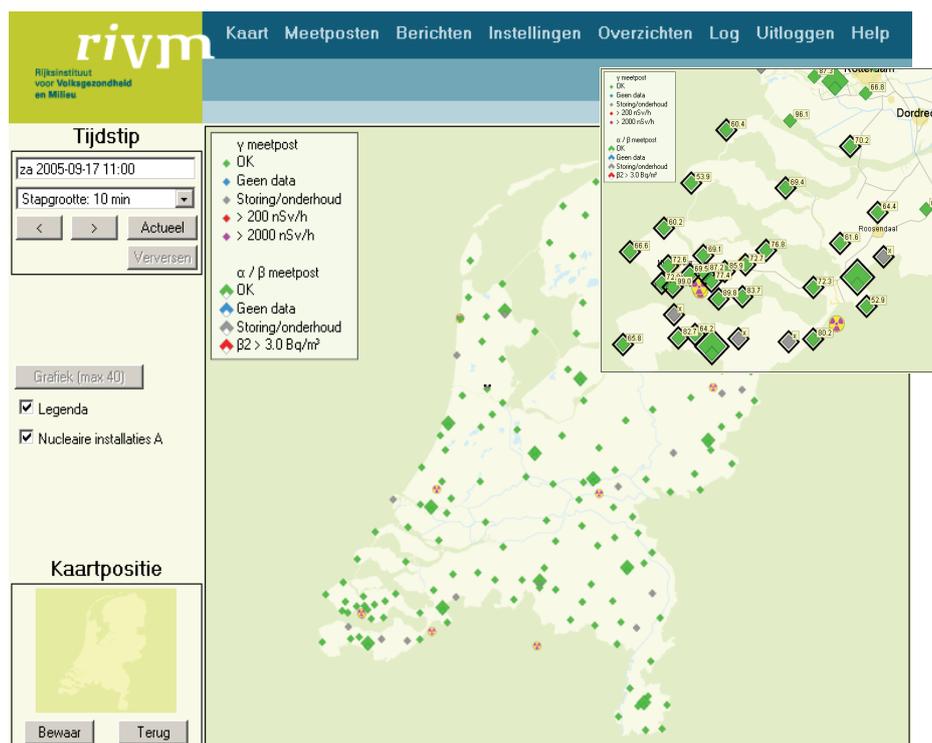
Figure 3. *Map view of the NRM WEB application used for visualisation of the NRM data. The NRM web service runs at the central system at RIVM. The insert shows an enlargement of the map for the region "Zeeland", where the nuclear power plant is located.*

Under normal conditions and a correctly adjusted compensation for the background the air activity monitors allow detection of artificial β activity down to a fraction of 1 Bq/m^3. Based on these findings and the experiences from previous operating periods, warning levels were set at thresholds of 200 nSv/h ambient dose rate and 3 Bq/m^3 artificial β activity

concentration. A warning is generated each time a threshold is exceeded. A 2nd warning level is set at 2000 nSv/h ambient dose rate; this warning will trigger not only RIVM, but also the regional emergency response centre of the corresponding region directly. In the operational period of the 3rd generation NRM this has happened only once, it was caused by an enhancement of ambient dose rate due to a strong source used for gamma radiography. An overview of the current warning levels is given in Table 1.

Quantity	WARNING LEVEL	Received by
Ambient Dose Rate	200 nSv/h	RIVM
Ambient Dose Rate	2000 Sv/h	RIVM + Regional Fire Brigades
Artificial β activity in air	3 Bq/m^3	RIVM

Table 1. *Warning levels of the NRM. A value above 200 nSv/h ambient dose rate or 3 Bq/m^3 artificial β activity will trigger an alarm to RIVM. A level above 2000 nSv/h ambient dose rate will also trigger an alarm to the regional response centre.*

2.6 *The WEB Application*

The national and regional emergency response centres can visualize the data in the NRM by means of the web server function of the central system. This web server provides a map based view of all monitoring stations at a selected date and time. It is possible to zoom in on a specific region and to select one or more stations for which additional data is displayed e.g. a graph of measurements versus time can be generated. In addition it is possible to look at detailed information from a monitoring station. If a user has sufficient authorization it is also possible to start an extra readout of one of more monitoring stations or to generate an "exercise alert".

3. OPERATION

3.1 *Routine Measurements*

Routine measurements are performed every 10 minutes. From these measurements hourly, daily and yearly averages are calculated and stored in the database. Figure 4 gives an example of the yearly averaged ambient dose rate and gross α activity in air over 2004 [10]. The ambient dose rate consists of a cosmic component of about 40 nSv/h [2] and a terrestrial component which originates mainly from γ radiation of radionuclides ^{232}Th and ^{238}U, their decay products and ^{40}K. The averaged terrestrial component varies between 15 and 75 nSv/h [2, 9]. Actual values are dominated by the type of soil on which the monitor is located and the ground cover in the direct vicinity of the measurement location. The data presented here have a systematic error due to an overestimation of the cosmic component and an underestimation of the terrestrial component due to the non-uniform angular response function of the monitors. It assumed that the ambient dose rate, dependent on location, is overestimated between 5 and 10 nSv/h [2].

The natural α activity concentration in air originates dominantly from the decay of radon (^{222}Rn and short-lived progeny). The yearly averaged α activity in the Netherlands is also given in Figure 4. The variation in the natural background originates from the exhalation process of radon from the soil and the dispersion and deposition process of ^{222}Rn and progeny. The determination of an artificial β activity (not shown) is complicated by the fact that the natural β-activity has to be subtracted, which in its turn is derived from the measured gross α-activity. During 2004 the averaged artificial β-activity did not deviate significantly form zero.

Individual 10 minute measurements fluctuate heavily around the here presented average values. Besides fluctuations in the radon exhalation process, fluctuations are also caused by (local) meteorological conditions; especially precipitation appears an important factor for the activity levels from radon and progeny. An example is given in Figure 8.
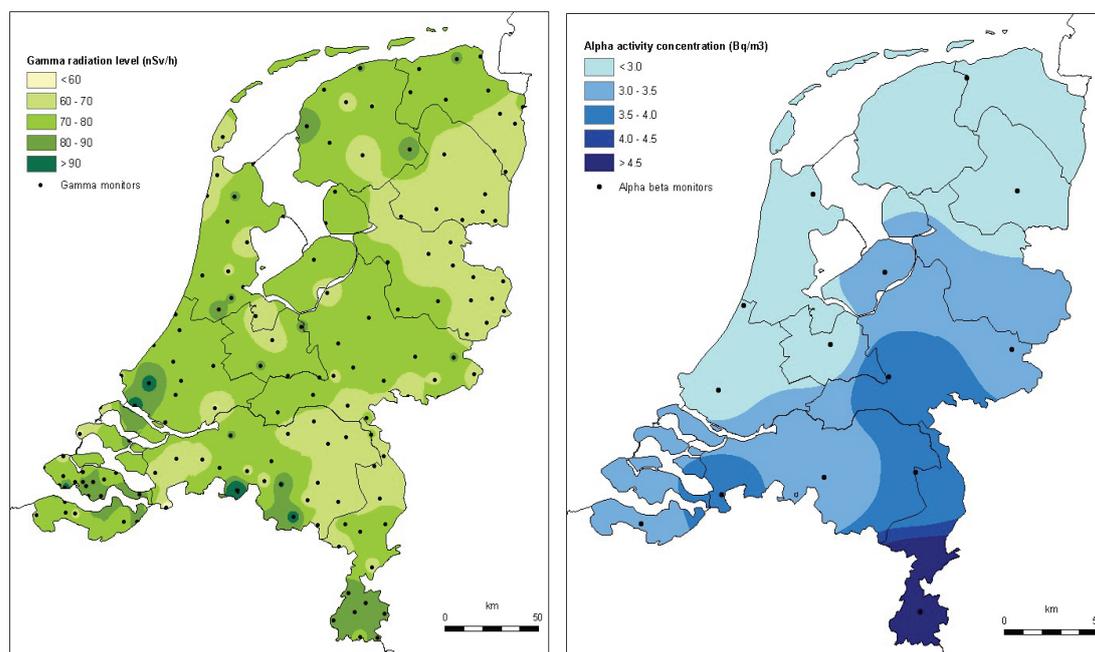
Figure 4. *Yearly averaged ambient dose rates (left) and α activity concentration over 2004 (right) in the Netherlands. An inverse distance weight interpolation algorithm was applied to interpolate between the data points. Preliminary results from ref. [10].*

3.2 *Monitoring in Support of Radiological Emergency Management*

During large scale releases of radioactivity in the environment the NRM has an important role in the monitoring strategy during the response phase of a nuclear accident [11]. Generally two phases are distinguished: an early phase, dominated by the need for radiological information supporting the decision process for early countermeasures and an extensive monitoring phase in which an overall radiological picture develops over a period of several days.

In the early phase the focus of the measurement program is dominantly in the direct exposure pathways; external radiation and inhalation. During passage of the radioactive plume this information is required to support the decisions on protective actions: sheltering, evacuation and iodine prophylaxis. Intervention levels are generally given in terms of effective and organ doses, i.e. they are not measurable quantities. A conversion of the measurement data is performed via model calculations. Additional nuclide specific measurements from the pilot-flame institutes or from the mobile monitoring facilities and the use of dispersion models further improve this conversion and allows spatial interpolations and prognostic overviews of the radiological situation.

In the extensive monitoring phase the measurement data of the NRM, combined with other sources of information, is used in the mapping of contaminated areas in the country. When identified and the magnitude of the contamination is determined, the focus shifts to the ingestion pathway with an emphasis on the monitoring of food- and feedstuffs and drinking water. In this period the NRM maintains its operation as a continuous surveillance network of the environmental contamination.

3.3 *System availability*

Availability of the system is expressed in two parameters describing (a) the fraction of successful connections between the central system and the monitors, i.e. the monitor responded and supplied a return status to the acquisition system and (b) the fraction of the total number of successful measurements obtained in this acquisition cycle per successful connection. These numbers are determined on an hourly basis. The fractions are shown in Figure 6 for a three month period of 2005. This period is considered representative for the operating experience of the upgraded NRM network.
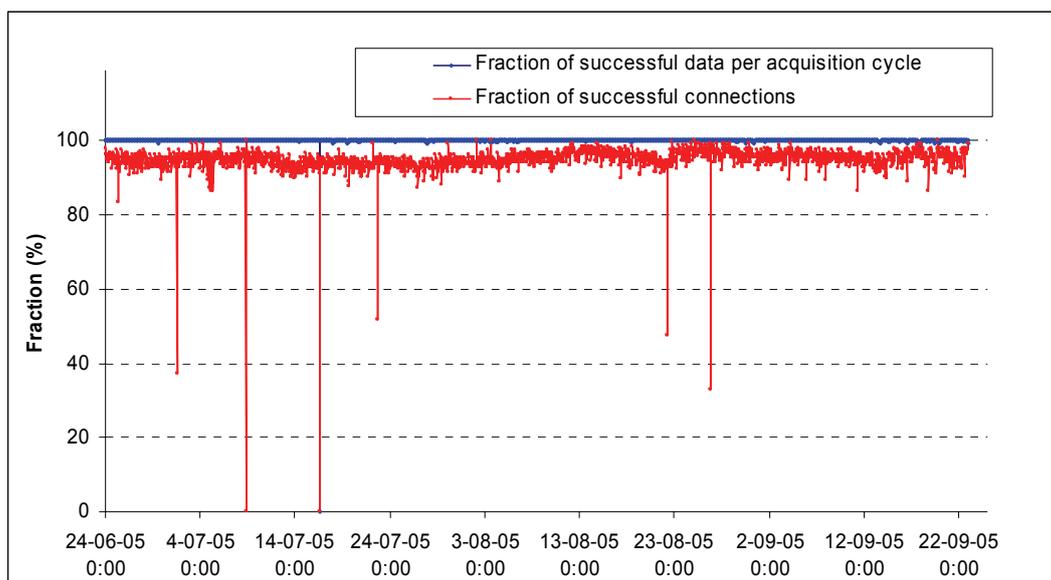
Figure 5. *Availability of the NRM over the most recent 3-month period of operation. Upper curve: the fraction of successful data read in an acquisition cycle, lower curve: the fraction of the successful connections between the central system and the monitors. Evaluation of availability is on an hourly basis. The larger drop-outs in the connection statistics are associated with maintenance activities on the network.*
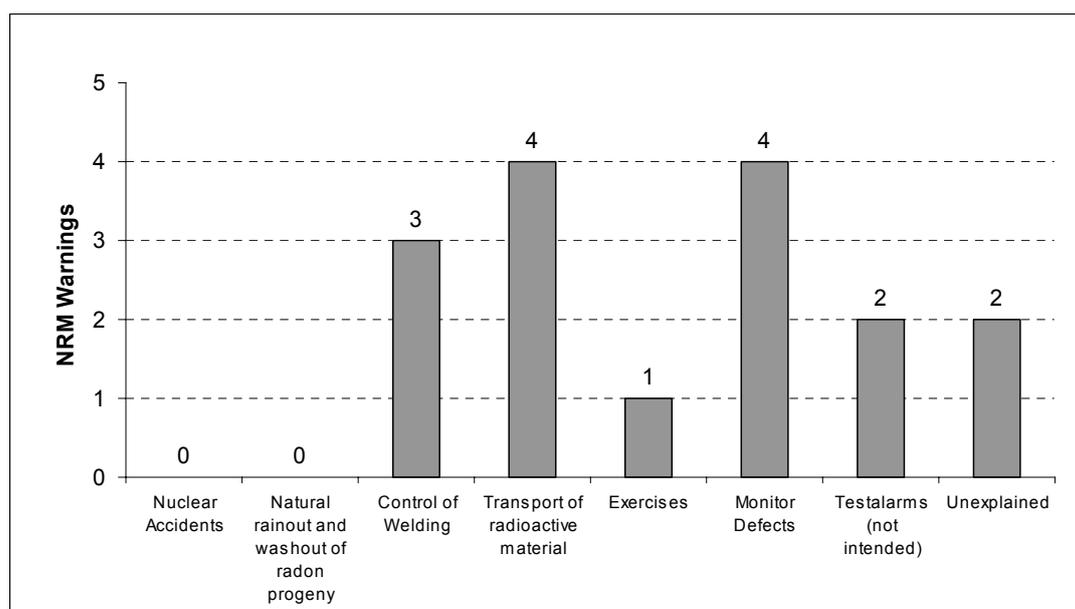
Figure 6. *Overview of warning messages generated by the system. In total 16 warnings were generated over the first 8 months of 2005. None of these events were caused by nuclear accidents.*

The overall availability is easily constructed from the product of the above fractions. Relocations of existing monitoring stations (10–12 in the reported period) are not included in the availability calculation. The availability of the NRM should be above 90% to remain in a normal maintenance mode of operation. Below 90% availability, maintenance and recovery operations are prioritized. Over the reported period the averaged availability was 95% and it was below 90% on 21 hours, but the system was generally recovered already on the next acquisition cycle.

While the number of successful readings is generally close to 100%, the fraction of successful connections varies round 95%. The reason for this is the poor quality of the emergency telephone network (and to a lesser degree also the public network for the 14 α/β monitors) over which the connections have to be established. A future upgrade to the dedicated IP-VPN connections may significantly increase the availability of the network. Until then, the system can, after a successful next hourly acquisition cycle, still read the missing measuring values from the monitor's memory.

3.4 *Elevated Levels and Warnings*

Warning messages from the system are evaluated over the operational period of the 3rd generation network; January – August 2005. In this period nuclear accidents did not occur. However automatic warning messages were generated on 16 occasions. Figure 7 shows an overview of NRM warnings in this period, i.e., measurement values are above the threshold of 200 nSv/h ambient dose rate or 3 Bq/m^3 artificial β activity. Alarms generated by (intended) system tests are not included in the analysis.

Elevated levels of radioactivity may have several causes: (1) nuclear accidents, (2) natural processes, (3) local human or industrial activities, (4) malfunction of equipment and finally (5) they may remain unexplained. Some examples of elevated levels are given in Figure 8. The relation to warning levels is shortly discussed below.

Besides enhancements a decrease of activity levels is also possible, a good example was an observed reduction of ambient dose rate of about 10 nSv/h during several days in the region Friesland in March 2005. The reduction was caused by a thick snow cover of several tenths of centimetres.

The most important natural source of enhanced radioactivity is the washout and rainout of radon progeny [12]. With the current NRM warning level at 200 nSv/h an overshoot of this level has become rare. In the relatively short evaluation period there were no warnings generated by natural causes; although a large number of increased dose rate measurements were recorded they all stayed below the first warning level of 200 nSv/h.

The use of strong γ-sources by e.g. radiographic applications, in the vicinity of an NRM ambient dose rate monitor will also give rise to elevated recordings (see Figure 8c). These events are reported frequently throughout the year and at different monitoring sites in the country. In some occasions this resulted in a warning message from the system, in one occasion the warning level of 2000 nSv/h was exceeded which alarmed the local authorities as well. An on-site inspection revealed radiographic activities in close proximity of a gamma dose rate station. Other human or industrial activities include the processing or unloading of raw materials leading to unintended enrichment of primordial radionuclides in the local environment; for a nice example see e.g. ref. [13]. Another human activity caused alarm messages on several occasions: based on location and timing these elevated levels could be traced back to the transport of radiopharmacons passing a nearby NRM monitoring station (see Figure 8b).

The remaining warning messages are usually false, i.e. caused by malfunctioning equipment or unintended test alarms. The exercise alarm shown in Figure 7 was caused by the Dutch National Exercise in May 25, 2005.
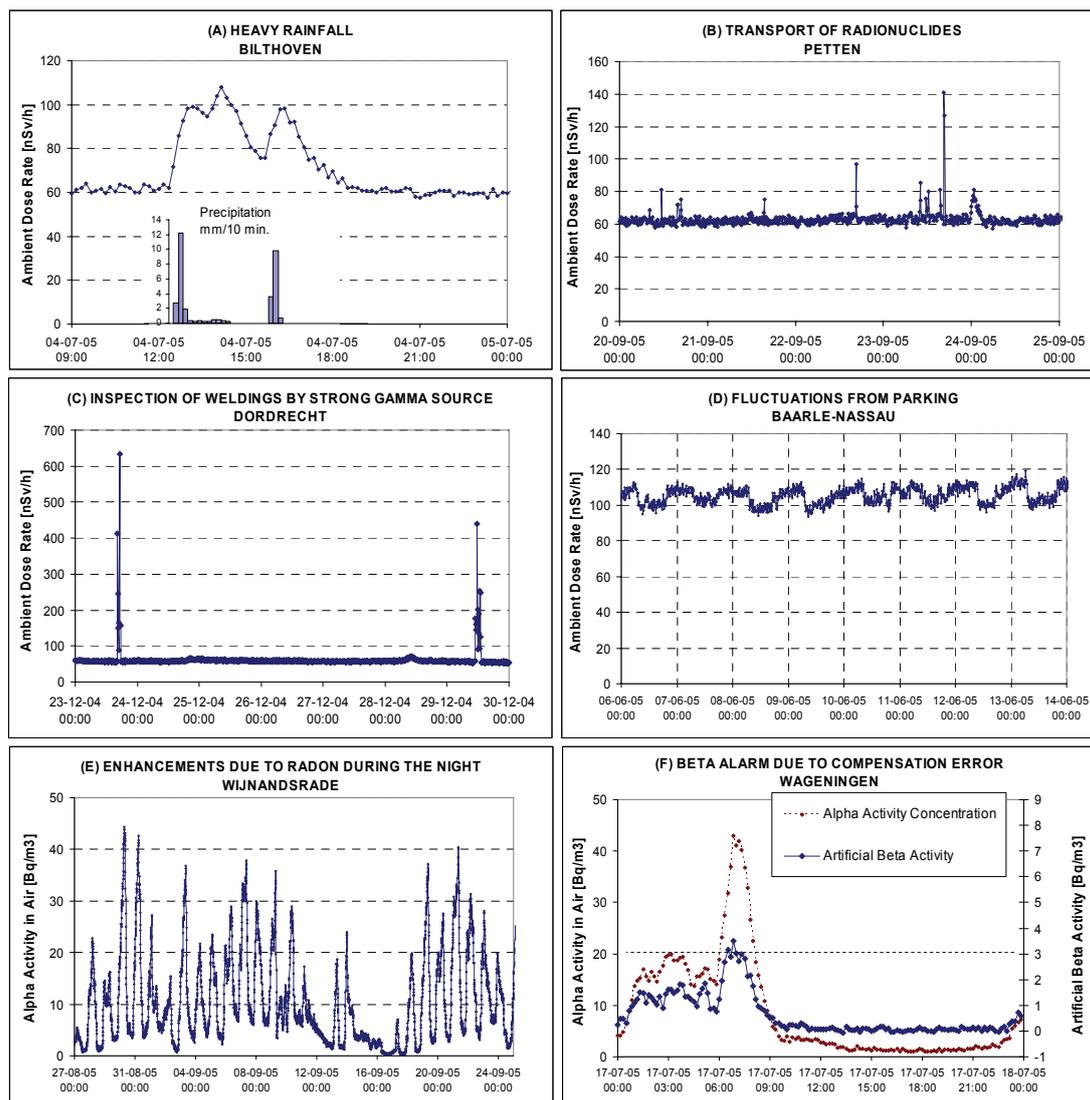
Figure 7. *Examples of NRM measured data from the 3rd generation NRM system. (a) An enhancement of ambient dose rate after heavy rainfall in Bilthoven shows an excellent correlation with the precipitation rate measured by the meteorological institute (KNMI) in De Bilt; precipitation data is provided by KNMI. (b) Transport of radiopharmacons produced at the HFR facility in Petten. On September 24, around midnight a small enhancement in the ambient dose rate from precipitation is also observed. (c) High levels of ambient dose rate during short periods of time: a typical case of nearby radiography; the inspection of welding in the vicinity of an NRM site. (d) The effect of a parking place in Baarle–Nassau. During the day parked cars in the vicinity of the monitor effectively shield part of the external radiation from the ground. (e) Periodically enhanced α activity levels in air. Due to the decrease of the inversion layer during the night the concentration of radon in the lower atmosphere dramatically increases. (f) A compensation problem in the α/β monitor. With fixed compensation ratio fluctuations in the natural α activity is automatically followed by the artificial β activity channel of the monitor. Here warning messages were generated between 07:00 and 08:00 hour, since β activity levels exceed the warning threshold of 3 Bq/m³.*

5. CONCLUSIONS

The performance of the 3rd generation NRM was evaluated over the first eight months of 2005. Compared to the setup used in a previous evaluation [13], the alarm thresholds were increased (1996), new airborne activity monitors were introduced (2003), new data acquisition and database systems were installed (2005) and a new protocol for generating warnings was defined (2005). In the previous setup a warning message was generated only if an enhancement above the warning threshold was accompanied by another enhancement in a nearby monitor. In the current situation any enhancement above the threshold will trigger an alarm. Based on the current findings the number of yearly warnings in 2005 is expected about 30% higher compared the 1990-1993 evaluation [13] and almost 10 times higher then in the previous years, i.e. after the redefinition of the warning protocol.

It is concluded that the network meets it primary goal, i.e., it enables early warnings against nuclear accidents. The availability of the network is served by the planned migration of the communication infrastructure from the public and emergency telephone network to the dedicated IP-VPN. It is expected that the availability numbers will then be well above 95%, and the response time of the network significantly enhanced.

REFERENCES

[1] De Ridder, M., Twenhöfel, C.J.W. (2004). "*The Design and Implementation of a Decision Support and Information Exchange System for Nuclear Emergency Management in the Netherlands*", Information Systems for Crisis Response and Management (ISCRAM), Proceedings ISCRAM2004, May 3-4 2004, Brussels.

[2] Smetsers, R.C.G.M., Blaauboer, R.O. (1996). *"Variations in Outdoor Radiation Levels in the Netherlands"*, ISBN 90-367-0621-1. Thesis, Groningen University.

[3] Tweede Kamer, vergaderjaar 1988-1989, *"Nationaal Plan voor de Kernongevallen-bestrijding"* 21015, nr. 3. Ministry of Housing, Physical Planning and the Environment, VROM 90044/2-89 1174/26. The Hague (1989).

[4] Smetsers, R.C.G.M., Blaauboer, R.O. (1994). *"Time-Resolved Monitoring of Outdoor Radiation Levels in the Netherlands"*, Rad. Prot. Dosimetry Vol. 55, No. 3, pp. 173-181.

[5] Slaper, H., de Haan, B.J., Veling, E.J.M. (1989). "*Dichtheid van meetnetten voor radioactiviteit met betrekking tot signalering van kernongevallen*". RIVM report nr. 248903001.

[6] Slaper H. (1998). Internal report, unpublished.

[7] De Cort, M., de Vries G. (1996). *"The European radiological data exchange platform: Reference manual and European automatic monitoring network"*. EUR Report 16415.

[8] Salfeld, H.Ch., Hable, K., Kok, Y.S., Reinen, H.A.J.M., (2003). *"Bilateral communication and information exchange, an attempt relating to Mandate 8 of the Dutch-German Commission (NDKK)"*. Off-site nuclear emergency management—capabilities and challenges, Salzburg, Austria.

[9] Smetsers,R.C.G.M. and Blaauboer, R.O. (1997). *"Source-Dependent Probability Densities Explaining Frequency Distributions of Ambient Dose Rate in the Netherlands"*. Rad. Prot. Dosimetry, Vol. 69, No. 1, pp. 33-42.

[10] Knetsch, G.J., ed. (2005). *"Monitoring of radiation in the Environment in the Netherlands"*, RIVM report 861020011/2005, to be published.

[11] Van Sonderen, J.F. (1997). "*Monitoring Strategy in Support of Radiological Emergency Management*", Rad. Prod. Dosimetry Vol. 73, Nos 1-4, pp. 115-118.

[12] Smetsers, R.C.G.M. and Blaauboer, R.O. (1997). "*A Dynamic Compensation Method for Natural Ambient Dose Rate based on 6 Years Data from the Dutch National Radioactivity Monitoring Network*". Rad. Prot. Dosimetry, Vol. 69, No. 1, pp. 19-31.

[13] Smetsers, R.C.G.M., van Lunenburg, A.P.P.A. (1994). *"Evaluation of the Dutch Radioactivity Monitoring Network for Nuclear Emergencies over the Period 1990-1993"*. Rad. Prot. Dosimetry Vol. 55, No. 3, pp. 165-172.

Contributions to SIC 2004

The following contributions are the extended abstracts only of the results of the Spatial Interpolation Comparison exercise 2004 published in

Applied GIS, Volume 1, Number 2. (2005)

- Savelieva E. *Ordinary Kriging Abilities for Radioactive Contamination Modelling*

- Pebesma E.J. *Mapping radioactivity from monitoring data: automating the classical geostatistical approach*

- Fournier B., R. Furrer *Automatic Mapping in the Presence of Substitutive Errors: A Robust Kriging Approach*

- Lophaven S., H.B. Nielsen, J. Søndergaard *Automatic Mapping of Monitoring Data*

- Palaseanu-Lovejoy M. *Bayesian automating fitting functions for spatial predictions*

- Ingram, B., L. Csató, D. Evans. *Fast Spatial Interpolation using Sparse Gaussian Processes*

- Hofierka J. *Interpolation of Radioactivity Data Using Regularized Spline with Tension.*

- Saveliev A., A. V. Romanov, S. S. Mukharamova. *Automated mapping using multilevel B-Splines*

- Rigol-Sanchez J. P. *Spatial interpolation of natural radiation levels with prior information using back-propagation artificial neural networks.*

- Timonin V., E. Savelieva. *Spatial Prediction of Radioactivity Using General Regression Neural Network*

- Dutta S., R. Ganguli and B. Samanta *Investigation of two Neural Network Methods in an Automatic Mapping Exercise*

- Pozdnoukhov A. *Support Vector Regression for Automated Robust Spatial Mapping of Natural Radioactivity*

Using Ordinary Kriging to model radioactive contamination data

Elena Savelieva

Nuclear Safety Institute Russian Academy of Sciences, B. Tulskaya 52, 113191, Moscow, Russia

E-mail: esav@ibrae.ac.ru

Extended abstract

This work deals with an application of ordinary kriging (OK) for spatial interpolation of data in a completely automatic ("one-click mapping") manner. The attraction of OK is due to its well-known advantages such as rather fast computation, direct estimation of uncertainty, easiness of an understanding and treatment.

The prior information provided as 10 sets of monitoring observations taken at different days was used to analyse and model the spatial correlation of the phenomenon. Furthermore, the prior information was expected to be consistent within a rather long time range and therefore assumed to reflect the structure of the contamination pattern at any given day. Experimental semivariograms for different days presented similar features: no anisotropy at a short range (<60 km) and some anisotropy in East-West direction at medium range (from 60 to 200 km); the same variogram "sill" (a priori variance); and it is reached at the same range. The semivariogram averaged over these 10 semivariograms again shows the same features. Consequently, the semivariogram model fitted to the averaged semivariogram could be good representation of the spatial structure at any day of observation.

The predictions made using the selected semivariogram model for 10 prior data sets are characterised by the reasonable fluctuations due to the day-to-day variations of the initial data. So it was decided that the method could be used to predict similar patterns at different times. Also, it was considered to check how the model would perform for the data with unknown structure, which could be anomalous and contain outliers, as indicated about the emergency data set in the SIC2004 announcement.

The obtained results are presented in Table 1 (the main statistics for the observed and estimated data) and Table 2 (the mean absolute error (MAE), the bias (or mean error, ME), the root-mean-squared error (RMSE) and Pearson's r coefficient of correlation between the estimated and true values) computed basing on the 808 validation locations. In general, results for the routine data set are better, which is not surprising because of its similarity with the 10 prior spatial distributions. The same parameters for the emergency data set are worse, but they are not meaningless. The general underestimation appears in the results for both data sets.

It is evident that means and medians are reproduced rather well for both data sets. Evidence for the smoothing effect lies in the underestimated standard deviation. The maximum value is underestimated as well due to this smoothing effect. The smoothing is much stronger for the emergency data set, where the actual correlation range is smaller than the one used for estimation. An interesting observation is the underestimation of the minimum value in the second data set. On the contrary, overestimation of the minimum was expected due to the smoothing effect. This minimum, together with 6 other data points estimated to be lower than the minimum of the raw data, is located in the area of the strongest pattern gradient, where it is even questionable to apply local kriging because the stationarity assumption is violated (constant mean can't be expected even close to these locations). It looks as if the estimates at these points compensate the high hot spot values so as to keep the local mean constant.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (routine data set)	57	180	98.02	98.8	20.02
Estimates (routine data set)	66.8	130.35	96.63	98.5	15.19
Observed (emergency data set)	57	1528.2	105.42	98.95	83.71
Estimates (emergency data set)	37.53	651.18	103.23	97.89	53.17

Table 2. *Comparison of the estimated and measured value*s *(nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.11	-1.39	0.78	12.49
Emergency data set	19.68	-2.18	0.56	69.08

Table 2. *Comparison of the errors.*

The approach applied here gave satisfactory results for both routine and emergency data sets. The results on the routine data set are fair, as was expected. The other situation was observed with the emergency data set. Because of the flux (outliers) the experimental semivariogram features very short correlation range – more than two times shorter than the one for the routine prior data. There is no real correlation between the flux values and the parts of the region not yet affected by the flux. Also, since the theoretical limitations of the OK (constant mean) are violated in the area around the flux, estimation in the area around the anomaly becomes biased (strong underestimation). The unaffected parts of the region are predicted well by the OK with the pre-selected parameters, as they display pattern similar to that of the prior data. Still the flux location was detected well, despite the strong smoothing effect.

OK estimates are always accompanied by the estimation variance usually treated as a measure of the uncertainty. In the current work the same uncertainty model was used for both data sets as it depends only on the semivariogram model and on the spatial location of the samples. To check the goodness of this measure was performed a special test. It proved that for the routine data set the uncertainty is described fairly well. The same uncertainty model did not appear to be adequate for the emergency data set, but still it is satisfactory. The application of the irrelevant semivariogram model certainly brings some negative consequences, but still the OK estimates are not meaningless.

Being a rather fast and simple method, OK can easily be incorporated into a decision support system with semi-automatic semivariogram tuning. A site-specific prior spatial correlation model may be prepared in advance. More detailed expert analysis may be performed later, when more detailed information from the contaminated site has been collected.

Automating the classical geostatistical approach for mapping radioactivity from monitoring data

Edzer J. Pebesma

Department of Physical Geography,
Faculty of Geosciences,
Utrecht University,
Netherlands

E-mail: e.pebesma@geo.uu.nl

Extended abstract

In the context of a comparison of spatial prediction algorithms, we applied the classical geostatistical approach to see how well it would automate, and how well it performed in case of an anomaly in the data. In case of the test using the routine data, the method performed well. In the case with the simulated emergency data (anomaly), automatic variogram modelling was hindered seriously, and in terms of RMSE best results were obtained by using the variogram from the test data without the simulated extremes. Although the 10 days of available training data showed a strong temporally persistent spatial pattern, cokriging did not improve predictions.

The training phase involved the following steps: (i) an outlier (observation record 838) was identified and removed from further analysis; (ii) a variogram model was selected from the following families: exponential, Gaussian, spherical, linear, and Matern with smoothing parameters ranging from 0.05 to 10; selection was done on the basis of leave-one-out cross validation (iii) cross variograms were modeled in an attempt to see whether cokriging would improve predictive power; (iv) the neighbourhood size was varied from the 10 nearest observations to all available (i.e., 200) values; (v) anisotropy has been considered.

At the end of the training phase, we ended up with a procedure where we

1. omit record 838;

2. calculate an omnidirectional sample variogram from the raw data, using a cutoff of 500,000 m and an interval (lag) width of 20,000 m;

3. fit a spherical variogram model with nugget to this sample variogram, using weighted least squares (Cressie, 1993), fitting with weights $h^{-1}N_h$, with distance h and number of point pairs N_h;

4. use ordinary kriging for prediction, given the observed data and the fitted variogram model, using a local kriging neighbourhood of 125 observations.

For the routine data the prediction procedure worked as expected. For the simulated emergency data the variogram fitting routine hampered. Best results were obtained when a variogram from the test data (without emergency extremes) was used.

For the routine data set, only a single run was performed. For the simulated emergency data set, in the paper the following scenarios were compared: (i) "singular fit", which uses the variogram resulting from the routine data, (ii) the "linear fit" which uses a fitted linear model. As a standard benchmark, inverse distance weighted interpolation (idw), using distance weighting power 2 and a 125 nearest neighbourhood was added as a scenario for both data sets in the paper. Tables 1 and 2 show summary statistics for observations, estimates and errors (predicted – observed). For the tables below, only the results for the "linear fit", scenario (ii) are shown for the emergency data set.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (routine data set)	57	180	98.02	98.8	20.02
Estimates (routine data set)	69.92	125.14	96.8	99.14	14.47
Observed (emergency data set)	57	1528.2	105.4	99	83.7
Estimates (emergency data set)	73.2	255.2	109.4	103.6	31.8

Table 3. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.11	-1.22	0.788	12.44
Emergency data set	23.26	4.00	0.417	76.19

Table 2. *Comparison of the errors.*

We designed a classical geostatistical "algorithm" for automatic mapping background radioactivity levels. The full paper addresses sensitivity to the following model decisions: (i) outlier detection, (ii) choice of an isotropic variogram model, (iii) choice of variogram cutoff distance and lag width, (iv) verification of directional dependence, (v) verification whether cokriging improved predictions, (vi) optimize kriging neighbourhood. Choices made were based on visual data inspection (i, iii, iv) and leave-one-out cross validation RMSE values (ii, v, vi). The issues were evaluated *in this particular order*; the idea behind it is that prior to the analysis we expected this to be the order of importance.

In principle, we believe that kriging interpolation is a suitable method for automatic interpolation. Automation of the manual steps carried out in this study, notably outlier rejection, variogram model choice and variogram parameter estimation ("model fitting") may not be advisable in cases where anomalies can be expected or should be detected.

The primary automatic mapping algorithm devised in this study failed when confronted with an anomaly (the emergency data set). Two reasons for that are (i) the algorithm had not been subject to strong testing, and (ii) no testing at all had been done with data containing an anomaly. In this light, the current procedure should be seen as a first attempt, containing many aspects that can be improved.

Codes

The procedure was applied by using a scripting environment (the R statistical program, extended by the gstat geostatistical modeling package). Using a scripting environment made it trivial to detect and correct errors made earlier in the process.

Automatic mapping in the presence of substitutive errors: a Robust Kriging approach

Baptiste Fournier[1], Reinhard Furrer[2]

[1] *Chaire de Statistique Appliquée, Ecole Polytechnique Fédérale, Lausanne, Switzerland*

[2] *Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, Colorado*

E-mail: baptiste.fournier@epfl.ch

Extended abstract

The demand for automatic interpolation is great in many scientific domains. For example, in geosciences automatic monitoring networks are becoming more common, more precise and cover bigger areas, providing a steady flow of data. Anomalies in the data occur and the available time to react is usually short. Such anomalies or "outliers" can be erroneous measurements (substitutive outliers) or real values produced by a local change of state of the process of interest. The goal of this article is to develop a new automatic mapping algorithm that is insensitive to erroneous measurements and that produces maps reflecting as close as possible the true process.

For the construction of our model, we suppose that the observations are realizations of an underlying physical process, possessing an additive structure given as follows. The process is composed of, first, a spatial trend or large scale variation and, second, of a second-order stationary Gaussian process exhibiting spatial correlations and substitutive outliers. By this, we mean that we do not observe directly the process of interest, but sometimes (or rather somewhere) a contaminated version thereof. Further, we assume that this contamination is non-contiguous across space.

The best unbiased linear predictor (BLUP), often called kriging predictor in geostatistical science, is sensitive to contamination. In this article, we derive a BLUP for the substitutive error model. In short, it is a weighted mean over all possible contamination configurations of the sites. However, this weighted mean is not computationally feasible and we use an approximation thereof. To approximate the exact predictor, we first use a forward search algorithm to derive the locations of the most probable contaminated sites. Calling this our initial contamination scenario, we look for a neighbourhood of this scenario in the space of all possible scenarios. The idea behind this is that the bulk of possible scenarios are weakly probable and so, starting from what we found being the most probable, we want to define a set of almost all probable scenarios. We approximate the BLUP estimate with a linear likelihood-weighted mean procedure over the subset of scenarios we found. This represents a robust procedure with respect to substitutive errors.

The resulting linear predictor has a Gaussian distribution. Therefore, it is straightforward to determine prediction intervals. As for the exact predictor itself, the exact mean squared prediction error is computationally too expensive to evaluate and we use a similar approximation technique. The approximated prediction intervals are based only on the weights of the selected scenarios, i.e. all the non-negligible weights in the predictor.

In parallel to the analysis of the routine and emergency data, we compare the performance of the robust predictor with five other predictors (weighted least absolute deviation, local mean, local median and Hawkins and Cressie's robust kriging procedure). The presented predictor performs well, especially for small and moderately sized contamination scenarios.

In the robust prediction framework, we also need a robust trend estimation for the large scale variation of the process and a robust variogram estimation and fitting procedure. For the

former, we fit a local regression model and impose a conditionally linear surface. For the latter, we use a robust scale estimator and then fit the empirical variogram using a weighted least square technique.

We apply our estimator to the routine data and the simulated emergency data. We use the ten training data sets as prior information to set several model parameters and initial values for minimization algorithms. Results are shown below in Table 1 and Table 2. The predicted values for the routine data set and the simulated emergency data set are very similar. The minor difference is in part due to differences in the trend estimation. The robust trend and variogram estimation are slightly affected by the "outliers" of the emergency data set. By nature of the robust predictor, the two large values of the simulated emergency data set are classified as outliers and thus the peak is not reproduced, clearly reflected in the error statistics of Table 2.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (routine data set)	57.00	180.00	98.02	98.80	20.02
Estimates (routine data set)	70.74	124.60	96.70	99.54	14.30
Observed (emergency data set)	57.00	1528.00	105.40	98.95	83.71
Estimates (emergency data set)	70.33	125.30	96.83	99.48	14.58

Table 1. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.06	-1.32	0.79	12.43
Emergency data set	16.22	-8.58	0.27	81.44

Table 2. *Comparison of the errors.*

Not only robust prediction procedures but also some classical interpolation methods suffer from the drawback of being smoothing methods when trying to predict anomalies. For illustration, we compare the results of our robust kriging approach with two local predictors. The first is a nearest neighbor kriging technique and the second is a local mean interpolator. The three methods represent approaches with decreasing theoretical and computational complexity, decreasing smoothing characteristics, but increasing outlier sensibility. However, given the fundamentally different nature of the predictors the choice of the prediction method should probably be made on the basis of reasonable model characteristics and not on resulted error statistics.

Coded in the high-level programming language R, it takes a few minutes on a reasonable sized desktop computer to predict a spatial field with several hundred locations using the presented method. For example, the computation times for the routine and emergency data set were roughly 660 and 210 seconds respectively (Linux, 2.6 GHz Xeon processor).

The robust kriging predictor is a fully automatic approach: no human intervention or decision is required for the mapping. However, decision makers are required to classify outliers as measurement errors or as genuine anomalies. The new predictor is not capable – by construction – to distinguish between an outlier and a large "true" value. Current research consists in expanding the substitutive error model to clustered contamination scenario processes and contamination processes with a spatial structure. Such a model should be able to detect contaminated regions and predict fields such as the second data set more accurately.

Automatic mapping of monitoring data

S. Lophaven, H.B. Nielsen, J. Søndergaard

Informatics and Mathematical Modelling, Technical University of Denmark

E-mail: snl@imm.dtu.dk

Extended abstract

This abstract shortly presents an approach, based on universal kriging, for automatic mapping of monitoring data. The mapping approach was implemented by the authors in the Matlab-toolbox DACE, primarily developed for Design and Analysis of Computer Experiments.

The performance of the mapping approach was tested on two datasets, the routine and the simulated emergency datasets. The routine dataset contained daily mean gamma dose rates in Germany reported by means of the national automatic monitoring network (IMIS). In the simulated emergency dataset an accidental release of radioactivity in the environment was simulated in the South-Western corner of the monitored area. For both datasets predictions were computed at 808 locations, based on 200 observations, and the performance was measured by comparing the predictions with the true values.

In order to automate the mapping approach an algorithm, which chooses between combinations of regression model and correlation function, was applied. The development of the algorithm was based on 10 datasets containing daily mean gamma dose rates in Germany at the same 200 locations as in the routine and simulated emergency datasets.

Table 1 shows descriptive statistics for the 200 observations in the two datasets, as well as for the 808 predicted values based on each of the datasets.

N = 808	Min.	Max.	mean	Median	std. dev.
Observed (routine data set)	57.0	180.0	98.0	98.8	20.0
Predictions (routine data set)	66.7	132.5	96.8	99.9	14.4
Observed (emergency data set)	57.0	1528.2	105.4	99.0	83.7
Predictions (emergency data set)	80.0	494.9	109.5	103.4	36.4

Table 1. *Comparison of the predicted and measured values (nSv/h).*

It is seen that the mapping approach tends to smooth the observations, i.e. the range of the predictions is smaller than for the original dataset, which is also indicated by smaller standard deviations. In particular the high concentrations in the second dataset, due to the simulation of an accidental release of radioactivity, are severely underestimated.

In order to assess the efficiency of the automatic mapping approach, the mean absolute error (MAE), the bias (or mean error ME), and the root mean squared error (RMSE) of the predictions at the $n = 808$ locations are reported in Table 2. Furthermore, Pearson's r coefficient of correlation between the predicted and true values is shown. It is seen that the suggested approach performs much better when applied to the routine dataset, thus it is not that well suited for handling extreme values.

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.7	1.2	0.76	13.1
Emergency data set	22.2	-4.1	0.54	71.2

Table 2. *Comparison of the errors.*

In summary the presented approach has a tendency to smooth the actual data values, and it therefore underestimates extreme values, as it is seen when applied to the simulated emergency dataset. On the other hand it was able to identify the simulated radioactivity. Therefore, we believe that even though the prediction of extreme values is unreliable, the mapping approach could be used as a warning system, as it is known from classical process control. Such a system is activated by extreme predictions, and an approximate location of release can be estimated by the mapping approach. To obtain reliable predictions the physical knowledge about the transport processes of radioactivity should somehow be applied.

Codes

The `Matlab` kriging toolbox `DACE` and documentation can be downloaded from http://www.imm.dtu.dk/~hbn/DACE

`Matlab` code for cross-validation, estimation of sample semivariograms etc. can be obtained by contacting the first author.

Bayesian automating fitting functions for spatial predictions

Monica Palaseanu-Lovejoy

Geography, School of Environment & Development, The University of Manchester, Manchester, UK

E-mail: mpal_lovejoy@hotmail.com

Extended abstract

This paper describes a Bayesian model of inference and tests its suitability in automatic mapping using the R statistical package and the geoR routine developed by Ribeiro and Diggle in 2001. The training datasets used to set up the inference equations were the result of only one spatial underlying process. This paper demonstrates that the proposed Bayesian model is robust as long as it models similar spatial processes, but fails to accurately make predictions when the basic spatial process is "contaminated" by a second spatial process with a different spatial correlation range. The second contaminating spatial process is seen as the result of an environmental accident, or point source pollution. None of the training sets provided information on such contamination.

Bayesian inference assumes the mean and the covariance parameters to be random variables, and integrates over the parameters' space to get the predictive distribution of any measurement. The formal solution to any estimation or prediction translates into the inference of the conditional distribution of the unknown parameters based on what has been observed. The parameters' prior distribution sums up our knowledge and beliefs about the underlying spatial process. The parameters' posterior distribution is obtained by updating the prior distribution throughout data measurements. It has been demonstrated that the Bayesian predictive distribution corresponds to the average of classical predictive distributions for known values of the parameters' underlying spatial process, weighted by the parameters' posterior distribution. In this case, prediction variance will account for the uncertainty in the parameters' estimations.

In practice, it is often impractical to explicitly develop a mathematical equation for the Bayesian predictive distribution. Thus, the common practice is to generate a random sample from the Bayesian predictive distribution using Monte Carlo sampling algorithms, and calculate from these simulations any statistical parameters of interest. The estimate for the Bayesian point prediction can be either the mean or the median of the simulations, depending of its distribution.

The "box and whisker" method was employed to check for potential outliers. The datasets had very few close outliers, but none of them had extreme outliers. This suggested that the employment of robust statistical methods to compute parameters such as mean and variance would suffice to fully characterize the prior distributions. None of the prior datasets gave any information about the existence of extreme outliers as outcome of a different spatial process.

The Bayesian model considered all parameters as being unknown. Empirical trials suggested that a transformed Gaussian model, with a circular correlation structure, was the best option for prediction. To make the parameters more "personal" to the different distributions of the training datasets, their respective classical and robust mean and robust variance were incorporated into the parameters variance (σ^2) and range (ϕ) definitions. These definitions were established throughout successive trials and errors in the effort to characterize all 10 available datasets.

The Bayesian model described above was used to predict radioactivity values in 808 locations using the routine and the emergency datasets provided for validation. If the routine dataset behaves in a similar manner with the training datasets, the second dataset has a different

distribution with extreme outliers. The Bayesian model performs quite well for middle range values, but fails to predict values close to the extreme outliers. The comparisons between the 808 estimated values and the observed values for both datasets are given in the following table:

N = 808	Min.	Max.	Mean	median	std. dev.
Observed (routine data set)	57	180	98.018	98.8	20.022
Estimates (routine data set)	66.6	128.41	96.62	98.52	15.26
Observed (emergency data set)	57	1528,2	105.42	98.95	83.71
Estimates (emergency data set)	68.74	419.46	103.09	102.61	24.91

Table 4. *Comparison of the estimated and measured values (nSv/h).*

The Bayesian method predicts reasonably well the mean and the median of the routine and the emergency datasets, but the predictions have a much smaller standard deviation than the measured datasets, especially for the emergency dataset. Also, if for both datasets the minimum values were slightly overestimated, the maximum value is grossly underestimated for the emergency dataset. The comparison of the two datasets' errors and the Pearson's r coefficient of correlation between the estimated and the true values are given in Table 2.

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.05	1.4	0.79	12.46
Emergency data set	19.76	2.33	0.5	74.54

Table 2. *Comparison of the errors.*

For both datasets, a multivariate analysis between the measured and predicted values has been performed. The correlation between the measured and the prediction values increases when robust methods are used, indicating that extreme values are not as correlated as the majority of the values (Table 3). In the case of the emergency dataset, in almost 50% of the cases the predicted values do not follow the distribution of the measured values, while for the first dataset the divergence between measured and predicted values is in only 20% of the cases.

Data sets:	Classical correlation	Robust correlation
Routine data set	0.79	0.82
Emergency data set	0.50	0.56

Table 3. *Multivariate analysis.*

The Bayesian uncertainty interval for a 95% level of confidence is calculated as the difference between the 97.5 percentile simulations value and 2.5 percentile simulations value, respectively. Even if the Bayesian model fits the routine dataset better than the emergency dataset, only 84.4% of the first data measured values are inside the Bayesian confidence interval, while for the second dataset, 98.5 % of the measured values are actually inside the confidence interval.

In conclusion, extreme outliers generated by a contaminating process with different range and variance than the one producing the majority of the data are difficult to predict using a model calibrated exclusively on background data. In order to cope with contamination situations, a two stage procedure is proposed instead that uses two sets of models, one calibrated solely on background data, and one calibrated to predict contaminated data without any assumptions regarding the spatial processes involved, such as multiquadric radial basis functions. If the original dataset does not properly approximate the underlying random spatial process, the Bayesian predictions are potentially a better estimate than the classical estimators are.

Interpolation of radioactivity data using Regularized Spline with Tension

Jaroslav Hofierka

Department of Geography and Regional Development, Faculty of Humanities and Natural Sciences, University of Prešov, Prešov, Slovak Republic

E-mail: hofierka@fhpv.unipo.sk

Extended abstract

In this article, the bivariate Regularized Spline with Tension (RST) implemented in GRASS GIS as `s.surf.rst` command was used to make the required estimates of radioactivity at 808 locations, based on "training" sets of radioactivity measurements at 200 locations throughout Germany. The presented approach consisted of several parts. First, the optimum set of parameters for the RST method and provided data was selected using cross-validation (CV) and visual evaluation of the interpolation results for the training data sets. Then the estimates for 2 test data sets were made using a parameterised interpolator. The obtained interpolation results were verified using evaluation data set and discussed in the last part. The interpolation accuracy was analyzed using the mean absolute error (MAE), the mean error (ME), and the root mean squared error (RMSE) of the predictions at 808 evaluation locations.

The RST method was used to interpolate two data sets representing radioactivity measurements at 200 locations. A cross-validation analysis showed that the size of the training data sets was too low to find optimal parameters using the cross-validation procedure. The resulting surfaces were strongly smoothed and less realistic than expected (found RST parameters: tension = 10, smoothing = 0.6 for the routine data set and tension = 10, smoothing = 1.6 for the emergency data set). Therefore empirical interpolation parameters (tension = 40 and smoothing = 0.1) were used to interpolate the test data. Despite the fact that this empirical selection used in the test did not produced interpolation results with lower overall predictive error (Table 1 and Table 4), it better preserved local fluctuations and anomalies of the phenomenon. The detection of these features is important in radioactivity monitoring and emergency situations. The poor reliability of cross-validation was also confirmed by verification using the evaluation data set (Table 2 and Table 5 vs. Table 3 and Table 6, respectively).

This verification analysis has shown that the RST method cannot be parameterised by CV for low-density data if one of the interpolation goals is to detect anomalies and fluctuations in data. Also, it has been shown that the RST method with empirical parameters produced interpolation results that are more appropriate for detection of anomalies and outliers.

N = 808	Min.	Max.	Mean	median	std. dev.
Observed (routine data set)	57.00	180.00	98.02	98.80	20.01
Estimates (routine data set)	56.96	142.84	96.91	98.58	17.58
Observed (emergency data set)	57.00	1528.20	105.42	98.95	83.65
Estimates (emergency data set)	-144.06	1255.06	105.24	97.98	95.15

Table 1. *Comparison of the measured and estimated values using empirical RST parameters (nSv/h).*

N = 808	Min.	Max.	Mean	median	std. dev.
Observed (routine data set)	57.00	180.00	98.02	98.80	20.01
Estimates (routine data set)	68.34	124.53	96.75	99.15	14.45
Observed (emergency data set)	57.00	1528.20	105.42	98.95	83.65
Estimates (emergency data set)	66.83	257.09	109.71	99.78	39.70

Table 2. *Comparison of the measured and estimated values using parameters found by cross-validation (nSv/h).*

N = 808	Min.	Max.	Mean	median	std. dev.
Observed (routine data set)	57.00	180.00	98.02	98.80	20.01
Estimates (routine data set)	68.23	131.63	96.72	99.30	14.33
Observed (emergency data set)	57.00	1528.20	105.42	98.95	83.65
Estimates (emergency data set)	67.49	681.90	105.82	100.50	56.88

Table 3. *Comparison of the measured and estimated values using parameters found by verification (nSv/h).*

Data sets	MAE	ME	Pearson's *r*	RMSE
Routine data set	10.28	-1.11	0.73	14.00
Emergency data set	28.35	0.18	0.38	100.13

Table 4. *Comparison of the errors for interpolation with empirical parameters.*

Data sets	MAE	ME	Pearson's *r*	RMSE
Routine data set	9.38	-1.27	0.78	12.68
Emergency data set	26.52	4.29	0.38	77.98

Table 5. *Comparison of the errors for interpolation with parameters found by CV.*

Data sets	MAE	ME	Pearson's *r*	RMSE
Routine data set	9.10	-1.30	0.79	12.51
Emergency data set	18.62	0.41	0.50	73.68

Table 6. *Comparison of the errors for interpolation with parameters found by verification.*

The application of CV to the optimization of RST parameters for radioactivity modelling showed the limits of this optimization technique in areas where input data are not sufficiently sampled. Anomalies and local fluctuation require a representative sampling especially in automatic mapping applications. Despite of that, it has been shown that the RST method is flexible enough to produce interpolation results that reflect the behaviour of the modelled phenomenon even for less dense data sets. However, the question of automatic parameterisation is still open and depends heavily on available data and the purpose of spatial interpolation. This study also showed that the minimization of the overall interpolation error can be in contradiction to the detection of anomalies and local fluctuations.

The use of RST in automatic mapping and decision support systems are quite straightforward, as the method is deterministic, controlled by a set of parameters that can be set manually or by optimization technique. Thus the user can use optimized or default parameters without in-depth knowledge of the method. This study showed a great flexibility and robustness of this method.

Automated mapping using multilevel B-Splines

Anatoly A. Saveliev[1], Andrey V. Romanov[2], Svetlana S. Mukharamova[1]

[1]*Faculty of Ecology, Kazan State University, Kazan, Russia*

[2] *R&D Center ScanEx, Moscow, Russia*

E-mail: saa@ksu.ru

Extended abstract

Multilevel B-Splines approximation (MBA) algorithm has been applied to the SIC2004 exercise as a high performance automatic mapping means in emergency situations.

In the context of the phenomenon model, the interpolation task was subdivided into two subtasks, i.e. trend surface estimation and the stochastic component approximation (residuals from the trend surface). The MBA algorithm (Lee, et al, 1997) was chosen for use in both subtasks because it is very fast and provides control over the degree of the surface generalization (smoothness) and over the "area of influence" for each known value and over the (local) anisotropy.

Ten-day average of the prior data at 200 known locations was used to interpolate trend surface with kriging and MBA; both methods have shown similar results, but the MBA outperforms kriging in speed.

In stochastic component interpolation, MBA and kriging have shown also similar results even though the MBA did not assume the use of a-priory information, i.e. variogram.

The table below summarizes the minimum, maximum, mean, median and standard deviation of 808 estimated values in comparison with those obtained for two data sets investigated. To estimate the quality of the model suggested, trend surface model was constructed that did not depend on the new data by MBA interpolating of prior data means. Ten MBA approximations of prior data were also made and averaged as another estimation of the trend surface.

N = 808	Min.	Max.	Mean	median	std. dev.
Observed (routine data set)	57.0	180.0	98.0	98.8	20.0
Estimates (routine data set)	66.2	127.2	96.4	98.3	15.4
Observed (emergency data set)	57.0	1528.2	105.4	99.0	83.7
Estimates (emergency data set)	66.9	341.8	104.8	98.6	38.5
Ten-day models average	54.4	157	94.6	95.3	17.7
Trend surface model	62.2	136.6	95.1	97.3	15.4

Table 1. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	Spearman's r_s	RMSE
MBA estimation					
Routine data set	9.3	1.6	0.78	0.78	12.6
Emergency data set	22.2	0.6	0.41	0.74	76.4
Trend surface model					
Routine data set	9.6	3.0	0.77	0.77	13.0
Emergency data set	17.0	10.4	0.23	0.77	82.2

Table 2. *Comparison of the errors.*

MBA interpolation method proposed in the paper can be used as an automatic mapping means in emergency situations. It shows excellent performance and quite good interpolation quality when applied to routine data set being free of outliers. It is also flexible enough for being adapted automatically to specific data given. The automatic detection of outliers and delineation techniques proposed in this paper make it possible to overcome problems without any losses in performance.

Special computer program for Windows 2000 was developed for SIC2004 exercise and used for all the calculations mentioned. The program developed by authors is based on the MBA, variography, kriging, grid and vector libraries.

As results from the comparison of MBA and kriging in stochastic component interpolation, the MBA outperforms kriging in approximation speed and produces the results of the same quality without preliminary data analysis. Meanwhile, kriging requires exact covariance model to be known at each location. This is impossible in case of non-stationary spatial law.

The comparison of the results obtained with MBA interpolation and trend surface model showed their similarity on the validation part of the routine data set. Trend surface estimation was found to be responsible for main fraction of the result errors. MBA failed to restore this surface with the acceptable quality. In emergency data set, the MBA showed good enough delineation of the high-value area, but not the direction of the plume.

The average computational time for each of two sets was below one second with a standard PC used (1.6 GHz Intel Pentium 4).

The results presented showed average quality of MBA interpolation both for the routine and emergency data sets. Although in the latter case, additional means can improve the situation and automatically detect, delineate and interpolate outliers. The possibility for automated selection of parameters, high speed computation and ability to process huge data sets are considered as a main advantage of the MBA approach proposed.

Fast spatial interpolation using Sparse Gaussian Processes

Ben Ingram[1], Lehel Csató[2], David Evans[1]

[1] *Neural Computing Research Group, Aston University, Aston Triangle, Birmingham.*
B4 7ET. United Kingdom

[2] *Dept. of Empirical Inference, Max Planck Inst. for Biological Cybernetics*
Spemannstrasse 38, 72076 Tuebingen, Germany

E-mail: ingrambr@aston.ac.uk

Extended abstract

A Gaussian process (GP) inference method, equivalent to Kriging, was chosen for this entry. GP inference assumes that the variables of interest have a jointly Gaussian distribution. The inference process is the estimation of the mean and covariance functions from the data and also the prior knowledge about the model. The advantage of GP inference lies primarily in its non-parametric nature and the intrinsic ability to produce probabilistic predictions at any location, seen or unseen. This theoretical flexibility is hindered by the amount of computation needed to obtain the predictions, for which one needs to invert the covariance matrix of the data. The rapid analysis of data in geostatistics is problematic due to this inversion, not only because it is prohibitively time consuming but that also requires a significant amount of memory storage. A number of different solutions were adopted to overcome this problem, such as partitioning or sub-sampling the data. These methods are rather ad-hoc, lack a sound theoretical basis and it is unclear what information is being lost by their application.

In this entry a Bayesian framework has been used to estimate the radioactivity level. These estimates are obtained using a latent global GP. The mean function of the posterior process (after data processing) is the estimate of the radioactivity level and the variance of the GP marginal is the uncertainty associated to this predicted value.

One of the objectives of SIC2004 is to generate results in the shortest amount of time. Instead of the conventional GP method with a single large matrix inversion, a sequential estimation scheme was used. In this sequential scheme a single element of the dataset is considered and then a sequence of intermediate posteriors is built so that the last one is an approximation to the global posterior. The speed improvement is achieved by a subsequent modification to the above sequential inclusion. At each step the possibility of removing some input locations from the representation of the posterior process is calculated. The removal is such that it guarantees that the posterior still includes the maximum amount of "information" about the current data item. This iterative removal leads to a modified sparse approximation to the true posterior which relies merely on a subset of the training inputs, making the operations scale cubically only with the size of this subset.

An important issue in the GP framework is that of model selection. Ideally one should choose a model (class of priors) which reflects the underlying phenomenon. Similarly, knowledge about the data collection process should be encoded in the likelihood function. There was no specific information about the data collection method, thus additive Gaussian noise is assumed with the exact value of the noise variance inferred in the algorithm. In GP inference (and geostatistics) the covariance function of the GP specifies the data model. This function specifies the relationship between neighboring input locations, how much these two observations resemble each other or how much influence a given data point has on its neighbours. We chose to use a covariance function based on existing studies but also to estimate the parameters of this family in an automated way.

According to existing studies, the radioactivity typically displays noisy and complex spatial patterns related to altitude, geology and weather. No additional knowledge relating to these environmental factors or their relationship to the data were given to participants. Since there is no principled way in the machine learning community to determine the "best" covariance function and since this function is usually selected from a small subset of pre-specified functions often based on intuition and experience, we decided that the variogram of the "prior datasets" would be used to help determine the function family that would best model the data.

We chose a mixture of covariance functions which best represents the structure of the sample variogram. It was noted that a squared exponential alone is not an appropriate model to fit this sample variogram. One reason to discard this covariance function is that within the datasets there are insufficient observations at short distances for us to infer the process behaviour at this range. Since the locations where estimates are required for the contest are often at smaller separation intervals, and since it is believed that there is very little variation at this level, it was decided that a squared exponential kernel would be more appropriate to model the local behaviour. However, for the long range variation in the data it would appear to be modelled more accurately by a simple exponential covariance function, hence a mixture of these two functions has been used.

In addition to using the 10 days prior datasets to estimate the variogram model, this data was also used to estimate the parameters for the mixture. In estimating the parameters we used the fact that data within a single day are correlated but we assumed independence for different days. Since there was no additional information about the individual days, it was assumed that each day had an equal weighting.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (routine data set)	57.00	180.00	98.01	98.80	20.02
Estimates (routine data set)	68.82	125.41	96.75	98.96	14.41
Observed (emergency data set)	57.00	1528.20	105.42	98.95	83.71
Estimates (emergency data set)	74.25	634.49	100.78	95.68	39.22

Table 1. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.10	-1.27	0.788	12.46
Emergency data set	18.55	-4.64	0.856	54.22

Table 2. *Comparison of the errors.*

Clearly the estimations for the Emergency dataset are going to have a greater error since the mean function of the GP will be influenced by the extreme values. For the Emergency dataset RMSE is much higher than the MAE, this is because the RMSE is more sensitive to high residuals. The predictions for the routine dataset do not reach the maximum range of the observations particularly well, but the error measures are quite low when compared with other benchmark methods that were used. It can be seen that sparse Gaussian processes can be used to obtain accurate predictions, in addition calculation speed can be increased significantly.

Spatial interpolation of natural radiation levels with prior information using Back-Propagation Artificial Neural Networks

J. P. Rigol-Sanchez

Department of Geology, University of Jaen, Spain

E-mail: jprigol@ujaen.es

Extended abstract

Automatic mapping of environmental variables requires a method able to make use of prior data to adjust spatial model parameters and to generate spatial estimations when presented with new (temporally lagged) data from the same geographical domain. If spatial data follows a temporal sequence time-series forecasting techniques might ideally be incorporated into the procedure. Unlike other methods, Artificial Neural Networks (ANNs) are particularly adept at coping with non-normal and inter-correlated inputs, and managing the incorporation of additional data and expert knowledge about a particular geographical domain within the estimation process.

In this study, we use feed-forward back-propagation networks trained with the Resilient propagation algorithm for automatic mapping of daily measurements of natural radiation. Networks were trained to learn: (a) the relationship between daily measurements and their spatial coordinates, and (b) the relationship between daily measurements made at one site and information from a number of surrounding sites (measurement and distance). Network model development was carried out using available prior information corresponding to the 10-day series of natural background radiation measurements. Inputs to the model included spatial coordinates (eating and northing) of the reference site, and information corresponding to the six closest neighbour observations surrounding each reference site. An additional experiment was performed to test the ability of a network trained using a new simulated routine dataset including outliers to estimate the simulated emergency data.

Model development procedure generated a partially connected trained network with a single hidden layer with 13 units (26-13-1), and with 293 links. Table 1 shows summary statistics for the 808 estimated values and observed values for the routine data and the simulated emergency data. Results for the routine data suggest that the model underestimated observed values. For the simulated emergency data, results suggest that the model clearly failed to estimate extreme values. Table 2 shows the mean absolute error (MAE), the bias (or mean error ME), and the root mean squared error (RMSE) of the predictions at the n = 808 locations for the routine data and the simulated emergency data.

N = 808	Min.	Max.	Mean	median	std. dev.
Observed (first data set)	57.0	180.0	98.0	98.8	20.0
Estimates (first data set)	50.3	154.0	96.4	96.8	23.3
Observed (second data set)	57.0	1528.2	105.4	99.0	83.7
Estimates (second data set)	50.0	154.3	94.4	95.4	24.8

Table 1. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
First data set	16.0	-1.7	0.55	20.8
Second data set	25.3	-11.1	0.02	87.5

Table 2. *Comparison of the errors.*

Results obtained in this study indicate that reasonably accurate estimations of natural background radiation levels could be obtained with a network trained using (background) prior data. On the other hand, estimation of data including extreme values proved to be a difficult task. During the exercise the network model was never presented with outlier values during the training stage. Trained networks produced wrong results when presented with extreme data out of the range of the training data. Nevertheless, output values for the simulated emergency data with outliers included large errors. ANNs are data-driven models which are good at interpolating but not at extrapolating. This means that in general a trained network will produce reliable estimations only in a data range commensurate to the range of the (target) data used for training. Consequently, an ANN developed for estimating natural background radiation levels will generally not be able to correctly estimate extreme values if outliers were not used for training.

Results of the experiment performed using a new training dataset including simulated extreme values indicate that outliers could be better detected if outliers were also considered in model development. However, extreme values were not accurately estimated in any case. In addition, this model produced false outliers both for the first (with no outliers) and the second dataset. These results suggest that a single model for estimation of both background values and outliers might not be appropriate because performing well for one task implied performing bad for the other one. Thus, an approach using two different models for each purpose might be more successful. For instance, a model might be developed for estimating background values and another one for estimating extreme values. These two different models could be used in routine and emergency conditions, respectively. However, this approach poses a new problem since to launch the "emergency mapping algorithm" a signal must be triggered by the routine mapping system.

Model development (architecture selection and training) required several hours of computational time (6 to 7 hours). Once models were determined the estimation phase required an average computational time of 1.5 seconds for each dataset.

In summary, ANNs can be used for automatic mapping of environmental (background) data with moderate success. ANN models for spatial interpolation can incorporate prior information into the estimation process. However, network models developed using prior background radiation data presented large errors when estimating extreme values (extrapolation). The ability to interpolate accurately depends ultimately on the availability of data commensurate with the particular target scale of output. Consequently, network models trained for estimating background values will usually not produce adequate results for extreme values. On the other hand, network models trained to estimate both background and extreme values tend to produce false outliers.

The ANN approach presented here to automatic mapping of environmental data was clearly inappropriate for dealing with outliers. Results obtained suggest that developing two different models for estimating background values and extreme values, respectively, might be a potentially more successful approach. This however introduces new operational problems. This approach would need further investigation.

Spatial prediction of radioactivity using General Regression Neural Network

Vadim Timonin, Elena Savelieva

Nuclear Safety Institute (IBRAE) of RAS, Moscow, Russia

E-mail: vadim@ibrae.ac.ru

Extended abstract

This work describes an application of General Regression Neural Network (GRNN) to spatial predictions of radioactivity. GRNN belongs to a class of neural networks widely used for mapping continuous functions. It is based on a non-parametric (kernel) Parzen-Rosenblatt density estimator. In the current work we used an anisotropic Gaussian kernel with the kernel size (2D value) being the only tuning parameter.

The ordinary GRNN training procedure is a mean-square-error (MSE) minimization procedure, accomplished using a cross-validation (*k-fold*, *leave-one-out*) approach. A target function (MSE) is computed from results for all samples taken out, one by one, and estimated based on the other samples and a given $\sigma=(\sigma_x, \sigma_y)$ value. The values of σ change, taking values from a rectangle marked by points [$(\sigma_{Xlow}, \sigma_{Ylow})$, $(\sigma_{Xlow}, \sigma_{Yhigh})$, $(\sigma_{Xhigh}, \sigma_{Yhigh})$, $(\sigma_{Xhigh}, \sigma_{Ylow})$]. The boundaries were defined basing on the prior data sets. Usually such 2D error function has one smoothed, well-defined minimum. The value of σ giving the minimum error may be adopted as a solution or tuned using a gradient search method. In the current study a conjugate gradient search method was used for final tuning of σ. This algorithm is absolutely automatic, not requiring any optimization parameters and also very fast and effective. Applying a gradient search for final parameter tuning allowed us to catch a value for σ good enough for the emergency data set, which appeared to fall outside the initial interval.

Under the assumption of independent, identically distributed (i.i.d.) random variables over the whole input space (in our case a normal (Gaussian) distribution), a GRNN estimate (Z^*) can be treated as a mean of a random variable. Variation can be considered as the squared of the residuals – $(Z^*-Z)^2$. Thus, uncertainty (variation) is computed from the residuals at the training locations (where the real values are known) by interpolating over the whole space. For this purpose a new special GRNN predicting uncertainty is built. The training procedure is absolutely the same as for variable prediction.

Table 1 presents the basic statistics (minimum, maximum, mean, median and standard deviation) of the 808 estimated and observed values for the two data sets. The mean absolute error (MAE), the bias (or mean error, ME), the root-mean-squared error (RMSE), Pearson's correlation coefficient r between the estimated and true values for the predictions at the $n = 808$ locations are presented in Table 2. Values of MAE and RMSE for the emergency data set are worse, but the correlation coefficient is much better. Thus, the emergency data set is predicted better than the routine one. Negative values of the ME indicate that GRNN, on the average, more tends to reduce (underestimate) high values than increase (overestimate) low ones. It is the case for both data sets.

The most important problem with GRNN predictor is smoothing. The smoothing degree depends on GRNN parameters tuned for the data. The presence of smoothing for the current tasks can be seen in Table 1. Minima are overestimated, maxima are underestimated, and standard deviations for the predicted distributions are much lower than the ones of the real fields. But the means and medians are reproduced very well. It means that GRNN provided a globally unbiased, smoothed estimate.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (routine data set)	57.0	180.0	98.0	98.8	20.0
Estimates (routine data set)	69.9	125.9	96.8	98.8	13.9
Observed (emergency data set)	57.0	1528.2	105.4	99.0	83.7
Estimates (emergency data set)	58.7	1330.5	104.9	100.3	71.5

Table 5. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.40	− 1.25	12.59	0.78
Emergency data set	14.85	− 0.51	45.46	0.84

Table 2. *Comparison of the errors.*

The smoothing characteristic for GRNN lowered the height of the peak in the estimated pattern for emergency data set. But visually it does not look smoothed – a collection of several rather sharp tops. Automatic GRNN allows us to adapt quickly to completely new data and to make very good predictions.

GRNN appeared to be very fast (raw training and prediction calculations took about 30 seconds for each data set on a Pentium III 500 MHz PC) and easy to use tool.

In the current work GRNN was successfully applied to two different data sets – usual routine data and unusual emergency data. Thus the current study led to the main conclusion that GRNN is a promising tool for automatic spatial prediction of radioactivity in both routine and emergency situations. It is very fast, easy to apply and interpret. It provides reasonable results, even for complex, unexpected data. The smoothing effect is a visible drawback. But it can be ignored if we are not too much interested predicting accurately high and low values, but want to detect the presence of a hot spot and indicate that a critical level is exceeded. We only need to keep in mind that maximums would always be underestimated – smoothed.

The direct estimation of the prediction uncertainty has some theoretical problems. Initially the method does not provide a measure of uncertainty; it is calculated under rather strong assumptions. But there are some ways, also discussed in this work, to introduce uncertainty estimation in the process. The other possibility is to move to a probability treatment of uncertainty.

Investigation of two neural network methods in an automatic mapping exercise

Sridhar Dutta[1], Rajive Ganguli[2] and Biswajit Samanta[3]

[1]*Dept. of Mining Engineering, University of Alaska Fairbanks, USA*

[2]*Dept. of Mining Engineering, University of Alaska Fairbanks, USA*

[3]*Dept. of Mining Engineering, Indian Institute of Technology, Kharagpur, India*

E-mail: ftsd1@uaf.edu

Extended abstract

In this paper the performance of two neural network (NN) methods viz. a radial basis function network (RBFN) and a multilayer feed forward network (MFFN) were investigated to predict the radioactivity levels at a given test site. Two different techniques namely the random method and genetic algorithm (GA) were used to divide the data (200 observations) into the training subset and the calibration subset during the model development stage. The training data was used to train the NN models while the calibration data was used to generalize (calibrate) them. In other words, the various models were built using four different approaches. In the first, a RBF model was trained on the randomly divided data and its parameters were obtained in an unsupervised fashion. In the second approach, model was similar to the first except it was trained on the GA divided datasets. For the third, the RBF model was trained on the GA divided datasets, and determination of the model parameters was done using a supervised training pattern. The fourth model was a MFFN employing the Levenberg-Marquardt (LM) training algorithm. The motive behind the use of GA was to ensure statistical similarity in the two datasets. The actual performance of the models was measured on the 800 observations supplied for the prediction purpose. A comparative evaluation of the two networks was done using Root mean square error (RMSE), Pearson's *r*, Mean error (ME) and Mean Absolute error (MAE). For both the datasets, Table 1 and Table 2 show a comparison of the estimated and observed values while Table 3 and Table 4 show a comparison of the corresponding errors in the estimates.

Routine Dataset	N = 808	Min.	Max.	Mean	Median	Std. dev.
Observed Values		57	180	98.0	98.8	20.0
RBF Estimates (Random Division)		67.4	126.6	97.7	99.3	14.3
RBF Estimates (GA division + Unsupervised training)		65.6	123.4	97.2	98.9	14.1
RBF Estimates (GA division + Supervised training)		63.8	108.93	96.5	98.8	10.1
MFFN Estimates		68.2	126.8	100.2	101.2	14.60

Table 1. *Comparison of the estimated and measured values (nSv/h) for Routine Dataset.*

Simulated Emergency Dataset	N = 808	Min.	Max.	Mean	Median	Std. dev.
Observed Values		57	1528.8	105.4	98.95	83.71
RBF Estimates (Random Division)		61.4	129.8	103.3	103.4	17.01
RBF Estimates (GA division + Unsupervised training)		45.9	191.9	106.7	101.62	37.23
RBF Estimates (GA division + Supervised training)		50.1	203.8	105.2	101.63	38.84
MFFN Estimates		72.5	273.9	123.4	103.5	51.4

Table 2. *Comparison of the estimated and measured values (nSv/h) for Simulated Emergency Dataset.*

Routine Dataset	MAE	ME	Pearson's r	RMSE
Original Method (Random)	9.92	0.2	0.76	13.1
RBF (GA division + Unsupervised Training)	9.62	0.90	0.78	12.7
RBF (GA division + Supervised training)	12.2	1.50	0.64	15.9
MFFN	9.93	2.18	0.76	13.3

Table 3. *Comparison of the errors for the Routine Dataset (All Methods)*

Simulated Emergency Dataset	MAE	ME	Pearson's r	RMSE
Original Method (Random)	17.5	5.10	0.29	80.6
RBF (GA division + Unsupervised Training)	28.2	-0.22	0.31	80.1
RBF (GA division + Supervised training)	28.9	-1.29	0.33	79.9
MFFN	38.5	17.98	0.27	87.3

Table 4. *Comparison of the errors for Simulated Emergency Dataset (All Methods)*

From the tables, it is evident that neither of the methods performed well in the simulated emergency dataset; although the performance was reasonably well for the routine dataset. The performance of the various models were probably governed by the similarities (or lack thereof) between the training/calibration subsets and the prediction subsets. Additionally, due to relative sparseness of the training/calibration subsets, the two subsets (in both datasets) did not achieve the degree of similarity as one would desire. The performance of the RBFN with GA data division and supervised training of gradient descent algorithm was marginally better for the simulated emergency dataset compared to the other methods. However, such an improvement might not be statistically significant. On the other hand, for the routine dataset the RBFN with GA data division and unsupervised training generated the best results.

Support Vector Regression for automated robust spatial mapping of natural radioactivity

Alexei Pozdnoukhov

IDIAP Research Institute, Martigny, Switzerland

E-mail: Alexei.Pozdnoukhov@idiap.ch

Extended abstract

The present paper deals with the Support Vector Regression (SVR) model. It is based on Statistical Learning Theory (SLT), which is a general mathematical framework devoted to the problem of estimation of the dependencies from empirical data. SLT gave rise to a family of Machine Learning algorithms, based on Support Vector approach. Concerning spatially distributed data these learning methods predict unknown mapping between inputs (spatial co-ordinates and secondary variables) and outputs (random function) from available data and a priori knowledge. The general property of the algorithms is that they aim at minimizing the empirical error (error on available training data) together with controlling the complexity of the model.

This paper is devoted to the automated application of SVR model for spatial data mapping in the presence of prior data. The complete methodology of SVR application should consist of a number of steps: monitoring network analysis, understanding of clustering, exploratory data analysis, exploratory variography, understanding of spatial continuity; data preparation: splitting of data in training, testing and validation data subsets, SVR training and testing, selection of the optimal SVR hyper-parameters, understanding the quality of the results by using exploratory analysis (statistics, variography) of the residuals, validation of the results, spatial data mapping with optimal SVR model. For the suggested case study, based on SIC2004 competition, this scheme has to be modified. SVR parameter tuning was carried out based on prior data, and on a standard cross-validation procedure as well. The uncertainty of the measurements was estimated and used to refine the SVR model.

The task of the SIC2004 competition, as it was perceived by the authors, was to provide a stable noise-robust automatic tool for mapping the natural radioactivity. SVR is a robust regression approach, which performs particularly well in the presence of noise and outliers in data. Thus, it is an appropriate method for mapping the slowly changing and relatively stable spatial variables, such as natural radioactivity level.

The resulting SVR model is a linear combination of kernel functions. The kernel function is defined for a pair of locations, and can be roughly thought of as a similarity measure between these two locations. The weights in the linear combination are obtained after solving an SVR Quadratic Programming optimisation problem; hence the weights are determined in a unique way. Some of the weights obtain zero values, and the corresponding measurements exert no influence on the model. Other samples with non-zero weights are called Support Vectors.

There are several parameters that control the properties of the model. Gaussian RBF of the bandwidth σ were selected as a kernel function. The optimal value of σ mainly depends on two characteristics of the data: correlation radius and data variability. Then, the upper bound of the weights C. This parameter also defines the trade-off between training error and model complexity. The last ε parameter is the width of the insensitive region of the loss function. This is the parameter that defines the sparseness of the SVR solution - those samples that lie inside the ε-tube possess zero weights.

The parameters were tuned according to the minima of a testing error on the prior datasets. The values of C and ε were determined and fixed. Note that parameter ε was tuned individually for every training sample, proportionally to the estimated standard deviation of the measurements given 10 prior datasets. Then, automatic cross-validation procedure was applied for the given routine and simulated emergency datasets to determine the refined value of σ.

The algorithm was applied for prediction, and the corresponding results are presented in the tables below.

N = 808	Min.	Max.	mean	median	std. dev.
Observed (routine data set)	57.0	180.0	98.0	98.8	20.0
Estimates (routine data set)	68.9	125.2	98.0	100.7	14.1
Observed (emergency data set)	57.0	1528.2	105.4	99.0	83.7
Estimates (emergency data set)	69.0	125.6	98.7	101.5	14.6

Table 1. *Comparison of the estimated and measured values (nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
Routine data set	9.22	-0.037	0.79	12.47
Emergency data set	16.25	-6.70	0.28	81.0

Table 2. *Comparison of the errors.*

The analysis of the results was carried out. The spatial structure of the residuals was investigated. The variogram of the prediction residuals of the routine dataset illustrated almost pure nugget effect. It indicates that all the spatially structured information was extracted from data. This observation suggests a good performance of the model for the routine dataset.

Concerning the results on the simulated emergency dataset, the following remarks have to be made. As it was mentioned before, the model was built to provide robust and noise insensitive tool for predicting the natural radioactivity level. The approach was not aimed at modelling the anomalies. SVR model deals particularly well with the outliers by "neglecting" these unusual values. This is an advantage of the model if the outliers are due to noise. But if the appearance of these anomaly values is due to some different spatial process, the robustness of the classic SVR can be considered as a drawback. However, there is an evident way to assess the presence of such abnormal data points in the dataset by considering the residuals of the training data and the corresponding weights. For the present case study, the training residuals for the locations of modelled emission were abnormally (10 times) large. This is a clear indication that these locations need a special treatment, since they indicate outliers (probably noise) or come from a different phenomenon from the one, which was modelled (artificial emission and not a natural radioactivity).

Generally, Support Vector Regression is a flexible model that allows modelling environmental variables. For the present case study we observed promising performance for predicting the natural radioactivity level. Prior data were used for tuning the parameters of the model. The information on the measurements variance was included into the SVR by the precise choice of the parameters of insensitivity. At the same time, SVR model provides limited capabilities for modelling unusual phenomena due to its robustness. The standard SVR algorithm considers extreme values as outliers; however, it also provides clear way for the detection of these unusual training samples.

Discussion Papers

Spatial Interpolation Comparison exercise
2004

Are comparative studies a waste of time? SIC2004 examined

Dan Cornford

NCRG, Aston University, Birmingham, B4 7ET, UK.

E-mail: d.cornford@aston.ac.uk

Abstract: There are several problems with model comparison studies based on single datasets. First, they really tell you more about the data set than the models used for prediction, secondly they do not generalise well and finally the results are often difficult to interpret and contradictory. SIC2004 suffers from all of these problems, so why did we do it and what have we learnt? This paper addresses the higher level questions about what general properties of the models seem to have been most important in SIC2004 and what we can learn from this. It attempts to unify the diverse models into a common framework and assess their relevance to this automatic mapping task.

Keywords: Model comparison, SIC2004, kernel methods, machine learning, Geostatistics.

1. INTRODUCTION

This paper is aimed at reviewing the entries to SIC2004 in terms of their theoretical basis, and using learning theory to attempt to explain their relative performances. In order to achieve this, the algorithms are grouped on several criteria, this grouping being summarised in one large table (Appendix I). In particular this paper looks at the training algorithm and its justification, the use of the prior information, the structure of the models and the flexibility of the models. The aim is not to review the results as in Dubois and Galmarini (2005), but to understand why the results have been obtained.

The reason for this paper is to explore some of the deeper issues that arise from the type of comparison study that SIC2004 was. Unlike some other comparisons there was an element of novelty to SIC2004. First, prior information was made available to assist the participants in defining their methodology and associated hyper-parameters. Secondly the aim was to define fully automatic methods: the time to carry out the mapping was also important. Finally, a new dataset, the 'joker', was introduced at the last minute to assess the robustness of the automated methods. This 'joker' would penalise participants who had 'biased' their algorithms too strongly to the prior data sets.

This paper will start with an overview of the problem, from the context of modelling, will then move on to consider the basis for the different models introduced by the participants in SIC2004, and finally attempt to draw some useful lessons from these contributions.

2. SIC2004

SIC2004 was devised to test the ability of mapping procedures to provide fast, accurate estimates of radioactive emissions from a large monitoring network. Details of the data and source can be found in Dubois and Galmarini (2005). However, only a small subset of 200 observations out of the total 1008 observations was made available to participants. The 10 days selected were typical of the 'background' levels of radiation observed across Germany. These measurements represent very local observations of radiation, which probably have rather small spatial support, although this is likely to depend on weather conditions, particularly boundary layer stability and thus turbulent fluxes.

A natural approach to such a mapping (prediction) problem is to define the process that generated the data. Since this is a physical system, with some controlling dynamics, this leads to the data assimilation scenario described in Galmarini (2005). Here a dynamical model

for radiation would be created using our knowledge of the physical processes such as the generation and decay of radon gas, the transport of that gas in the lithosphere and atmosphere, the natural decay of mineral sources all around us, and solar radiation. This would clearly be a complex model, however many of the component parts of this model are already well know, and reasonably observed. For example, the atmosphere, at least at large scales, is well represented in Numerical Weather Prediction models (Kalnay, 2003), and these routinely make forecasts of boundary layer winds and stability, which could be used to assist in defining transport in (and out of) the boundary layer.

Of course large parts of this model will be imprecisely known; both the initial conditions (e.g. where the radioactive material that is generating the radiation / radon is, what the state of the atmosphere is?); model error (e.g. what are the parameters for the 'flux efficiency' of radon through a soil model?); and model structure (e.g. what is the real mechanism for the transport of radon in the lithosphere – faults / cracks / diffusion?). A solution to this is to use probabilistic methods: model the distribution over the state and all parameters (all models?) and then work out the predictive distribution using Bayes' theorem. This is the essence of data assimilation; exploit what we know about the processes in a system to optimise the use of the observations we make of that system (Kalnay 2003). The basis for data assimilation is essential Bayes' theorem, so we can view the models as defining priors over space / time processes, and thus use all the useful tools from Bayesian statistics. This is clearly a very large undertaking, since it requires the coupling of models that operate on very different scales (radioactivity is atomic, the atmosphere is practically limited to around 1 km at present). No participants took this route, however I am convinced this is the correct means to provide estimates of the space / time radioactivity field. The 'joker' set shows clearly that we ought to take care when making such models, since they must also include very extreme events, which we know do occur (e.g. Chernobyl).

The situation that presents itself, so far as all participants approached it was that there was a data set, which exhibited space / time structure and that rather than making a comprehensive model for this, they would attempt to approximate the output of the unknown (unknowable?) model using some other prior model. This is really the domain of statistical pattern recognition (Bishop 1995, Nabney 2001), geostatistics (Cressie, 1991) and other statistical prediction methods. Here the model generally represents an attempt to define the structure of the output of the physical processes, through some prior over smoothness or continuity, often imposed through some form of basis, covariance or kernel function. In essence all these methods seek the same goal: to smooth the data optimally. The question is what is optimal? Since we no longer have recourse to physical processes the prior is generally over function classes (explicit in Gaussian process models, implicit in others), and there is now considerable freedom and little theory to assist in making the optimal choice. At this point I note that geostatistics (and spline based methods) have process based interpretations, linked through the choice of the covariance function and the associated differentiability of the sample paths and this was used by Ingram and Csato (2005).

2.1 *Use of prior information*

Cross validation is the most frequently used method of model selection, largely because it works in practice. When choosing an appropriate model type for a given data set, there is no universal model that can exhibit all behaviours imaginable over its parameter range. Thus the first choice to be made is that of model structure. In this exercise a variety of models were investigated; from inverse distance based models, through explicit local averages, to kriging (and variants), splines, feed-forward neural networks, to the semi-parametric 'generalised regression network'. It seems that many (if not all) participants attempted to fit a variety of models to benchmark their preferred approach. This is a very sensible approach and should be rule number one for any potential modeller: make sure you try the simplest models first; these may be all your processes require, or all your data supports.

What is interesting from the results of the mapping of the first test data set, which has the same structure as the 10 prior data sets, is that as measured by Root Mean Squared Error (RMSE) all methods are essentially equally accurate, being in the range 12.4-13.1 nSv/h. Focussing on RMSE seems most natural since almost all participants assumed a sum of squares error function when selecting optimal hyper-parameters. The figures of mean absolute error and mean error (bias) are also very similar. This suggests that all models are essentially doing equally well (or badly). This is not overly surprising given that the sampling density of the observations captures a large scale geologically controlled structure well, but does not capture the local variability adequately. Thus given that all participants used cross validation on the same data for model / hyper-parameter selection, it is not overly surprising that the methods produced broadly comparable results.

The results on the second data set are far more variable, as might be expected. It is rather obvious that those participants who fixed the hyper-parameters on the basis of the prior data sets would not produce very good results on the 'joker' data because this had a very different and localised structure (interestingly this is not totally borne out in the results). In some ways the 'joker' data was a rather mean trick to play, after all what is its role? In this case the aim was automated mapping, so the role was to test the robustness of the algorithms to extreme data. The data was very extreme, totally outside the sort of situations encountered in the training data. Several participants referred to wanting to perform outlier detection on the extreme data: this could make sense as part of an integrated mapping approach, one for the background, on for the extreme values, but over time, as the source dispersed this would become increasingly problematic. Again a better solution is to include such events in your model, although participants could not have been expected to do this.

2.2 *Hyper-parameters*

All modelling involves the estimation of certain parameters. It is often useful to group these parameters into a hierarchy (as is always done in Bayesian modelling which this paper makes extensive reference to). In general there are the first level parameters (the weights in neural networks, or kriging or splines) which are estimated directly from the training data (or indirectly through the covariance model / solution of a linear equation). In the statistical setting these parameters (or weights), denoted w, would typically be set by maximising the likelihood (of the training data, D, given the model, M and w), $p(D|w,M)$. I will explicitly include the model, M in the conditioning. Under an assumption of a Gaussian likelihood, we obtain the sum of squares error function that many seem to have minimised without justification of its origin.

Most models also include some higher level parameters: the regularising terms in neural networks or splines, the parameters of the covariance functions in geostatistical approaches. These I will refer to as hyper-parameters and denote by, α. I assume that structural information on the model is contained in M. Many participants used the prior training data to fix the hyper-parameters in their algorithms. This is probably a good idea in one sense, in that estimation of hyper-parameters is in many cases rather numerically intensive and often difficult to implement using statistically principled approaches. The most popular approach to hyper-parameter estimation was grid based (generalised) Cross Validation (CV). Most participants still (implicitly) sought a maximum likelihood estimate for w and α, maximising $p(D|w,\alpha,M)$. Those that applied spline based methods essential adopted a pragmatic Bayesian maximum a posteriori probability estimate for the spline curvature penalty (Wabha, 1991). Only two participants, Ingram and Csato (2005) and Palaseanu (2005) adopted formal statistical learning methods, both using Bayesian approaches in their solution. These will be contrasted in greater length later.

Many of the participants used heuristic methods to estimate hyper-parameters in their models. Where geostatistical methods were applied, most participants used weighted least squares fitting of the variogram / covariance model to a sample variogram / covariogram (Cressie, 1991). This is a well known method, based on a method of moments estimator and

fitting a non-linear model to this. It works well in practice, and produces very similar performance to maximum likelihood, or Bayesian approaches to hyper-parameter estimation. This illustrates something we all probably feel from experience; a well designed heuristic is often as good as a fully principled model. Of course the heuristic approach is often less computationally expensive than a brute force, principled solution, so is quite attractive in many applications.

3. RESULTS

In order to compare the results easily a summary measure of performance was created, based on the criteria all participants were asked to report. While this summary is completely subjective, it allows us to rank the methods in terms of their overall performance and their performance on test set 1 and 2. The measure is given by:

'skill score' = MAE + abs(ME) + RMSE + 10*(1- Pearsons r)

It is possible to propose a whole range of skill measures, this one is quite natural, and simply weights the mean absolute error, the bias (mean error) and root mean squared error equally, with some scaling applied to the correlation coefficient so that it is in the same order of magnitude as the other components. Note this skill score should be lower for better methods, and higher for worse methods. The skill score is computed for test set 1 and 2 (the 'joker'), and these are added to give an overall skill score.

3.1. *Overall skill score*

Table 1 shows the methods ranked by their overall skill score.

First Author	Classification (Dubois & Galmarini 2005)	Isotropic?	Hyperp. fixed?	Overall skill score
Timonin	GRNN	no	no	**88**
Ingram	Sparse Sequential Gaussian Process	no	no	**104**
Savelieva	Ordinary kriging	no	yes	**121**
Hofierka	Regularized Splines with Tension 1	no	no	**123**
Pebesma	Ordinary kriging	yes	yes	**125**
Palaseanu	Bayesian Kriging	yes	sampled	**127**
Lophaven	Universal Kriging	yes?	yes?	**129**
Pebesma	Inverse Distance Weighted (power 2)	yes	yes	**129**
Saveliev	Multilevel B-Spline A	yes	yes	131
Ingram	Gaussian Process	no	no	134
Pozdnoukhov	Support Vector Machine	yes	yes	135
Ingram	Gaussian Process 2	no	no	137
Fournier	Robust Kriging	no	no	138
Dutta	RBF random division	yes?	yes	140
Hofierka	Regularized Splines with Tension 2	no	no	141
Saveliev	Multilevel B-Spline B	yes	yes	145
Rigol Sanchez	ANN back propagation (add training)	no	yes	155
Hofierka	Regularized Splines with Tension 3	no	yes	163
Ingram	Radial Basis Function	no	yes	165
Ingram	MLP	no	yes	174
Rigol Sanchez	ANN back propagation	no	yes	177
Dutta	LMA Ward Net	no	yes	183

Table 1. *Methods ranked by overall skill score.*

There is significant variability in the skill scores of the methods, with the best method (on this criteria) the generalised regression network of Timonin (2005), achieving a skill score of 88, and the worst method getting 183. The majority of methods achieved scores in the range 120-

140. It is interesting that the best method overall was non-parametric, and that the two strongly theoretical and statistical approaches of Palaseanu (2005) and Ingram and Csato (2005) are 6th and 2nd best respectively. However a number of more engineering approaches are equally successful, with both kriging and spline based methods making the top 6.

First Author	Classification (Dubois & Galmarini 2005)	How is 'training' data used	skill score1
Pozdnoukhov	Support Vector Machine	To fix hyperparameters	24
Pebesma	Ordinary kriging	Set hyperparameters	25
Fournier	Robust Kriging	Structural	25
Ingram	S. Sequential G. Process	Set initial values for hyperparameters	25
Hofierka	R. Splines with Tension 1	Not used?	25
Palaseanu	Bayesian Kriging	To specify the prior distributions	25
Ingram	Gaussian Process	Not used	25
Savelieva	Ordinary kriging	Set hyperparameters	25
Timonin	GRNN	To define range of kernel widths	25
Hofierka	R. Splines with Tension 2	Set initial hyperparameters	26
Ingram	Gaussian Process 2	Not used	26
Ingram	MLP	Structure of network	26
Saveliev	Multilevel B Spline A	Set hyperparameters + G. process	26
Lophaven	Universal Kriging	Structure of model: mean and cov fn	26
Ingram	Radial Basis Function	Structure of network	27
Pebesma	Inv Dist Weight (pow 2)		27
Dutta	RBF random division	Structure of network	27
Saveliev	Multilevel B Spline B	Set hyperparameters	28
Hofierka	R. Splines with Tension 3	Not used?	28
Dutta	LMA Ward Net	Structure of network	**30**
Rigol Sanchez	ANN b-p (add training)	To determine structure	**32**
Rigol Sanchez	ANN back propagation	To determine structure	**43**

Table 2. *Methods ranked by skill score on test set 1.*

It is useful to examine the performance of the algorithms under test set 1 and test set 2 separately, since these present very different challenges. Table 2 presents the summary statistics for test set 1, Table 3 does the same for test set 2, the 'joker'. Table 2 illustrates that most methods perform comparably, with the majority of scores in the range 24-28. Given the sampling variability, the natural variability across time, and other sources of error possible, it seems that for the purposes of providing an automated map, almost any of these algorithms could be chosen to perform the task.

This illustrates that where similar prior data is available it really makes little difference as to which specific algorithm is used to set hyper-parameters, and even perform the prediction so long as all models are capable of representing the structure resolved by the observations (as is the case here). Of course there are other factors to consider, for example the speed of the algorithm. All of the algorithms submitted would run fast or very fast on the data presented, except the Bayesian kriging of Palaseanu (2005) which would require computationally intensive Markov Chain Monte Carlo sampling, and too a much lesser extent the sparse sequential Gaussian process of Ingram and Csato (2005) which would require several hyper-parameter re-estimation loops. However some of the algorithms might not scale well to larger data sets (for example if all the test data and training data were combined for estimation). In this case the non-parametric method of Timonin (2005) would scale badly since all data is used for prediction, while the fully Bayesian kriging of Palaseanu (2005) would become almost impossible to carry out, even off-line. However, most of the

geostatistical approaches adopt a search neighbourhood which cuts down on the complexity of the prediction, and since for many hyper-parameters are fixed this imposes no further penalty. The sparse methods of Pozdnoukhov (2005) and Ingram and Csato (2005) would perform particularly well as the data set size increased, since they are flexible enough to accommodate the additional data with reasonable computational expense.

Another issue not highlighted in these tables is the ability of the methods to make probabilistic predictions. Almost all methods can be used to provide estimates of the prediction error, but some of these estimates are rather ad-hoc, being based on models for the residuals, fitted after the determination of the mean function. This is likely to produce biased estimates. It is not possible to compare the performance on the probabilistic predictions, since the data is not directly available. There are two methods that are likely to provide optimal estimates of the probability distribution at the prediction location: the Bayesian kriging takes into account model uncertainty in hyper-parameters as well as prediction error, and the sparse sequential Gaussian process computes the approximate full posterior process at each prediction location (with fixed hyper-parameters). Other geostatistical approaches also compute an approximate posterior process at each prediction location, but the use of the prediction neighbourhoods means that the level of approximation is not well controlled, unlike in the sparse sequential Gaussian process.

First Author	Classification (Dubois & Galmarini 2005)	Isotropic?	Hyperp. fixed?	skill score2
Timonin	GRNN	no	no	62
Ingram	Sparse Sequential G. Process	no	no	79
Savelieva	Ordinary kriging	no	yes	**95**
Hofierka	Regularized Splines with Tension 1	no	no	**98**
Pebesma	Ordinary kriging	yes	yes	**101**
Palaseanu	Bayesian Kriging	yes	sampled	**102**
Lophaven	Universal Kriging	yes?	yes?	**102**
Pebesma	Inverse Distance Weighted (pow 2)	yes	yes	**103**
Saveliev	Multilevel B Spline A	yes	yes	**105**
Ingram	Gaussian Process	no	no	**109**
Pozdnoukhov	Support Vector Machine	yes	yes	**111**
Ingram	Gaussian Process 2	no	no	**111**
Dutta	RBF random division	yes?	yes	**112**
Fournier	Robust Kriging	no	no	**114**
Hofierka	Regularized Splines with Tension 2	no	no	**115**
Saveliev	Multilevel B Spline B	yes	yes	**117**
Rigol Sanchez	ANN back propagation (add training)	no	yes	**123**
Rigol Sanchez	ANN back propagation	no	yes	**134**
Hofierka	Regularized Splines with Tension 3	no	yes	**135**
Ingram	Radial Basis Function	no	yes	**138**
Ingram	MLP	no	yes	**148**
Dutta	LMA Ward Net	no	yes	**153**

Table 3. *Methods ranked by skill score on test set 2 ('joker').*

Table 3 shows the results ranked by their 'skill score' on the second, 'joker' test set. This test set exhibited behaviour very different from the priors given in training phase. Thus the key issue here was the flexibility of the model chosen. The best results, by some distance were obtained by the generalised regression network. This is very similar in spirit to a radial basis function model, with basis function at each observation. The only free parameter is the width of basis function, which in this case can be defined to be anisotropic along the principal axes

(a diagonal covariance matrix). As long as suitable hyper-parameters are determined this method is very flexible: it imposes a strong smoothing on the data (due to the Gaussian kernels). Timonin (2005) ensured with their combined (semi-exhaustive) grid search, followed by optimisation, approach to hyper-parameter estimation that their algorithm performed well on the 'joker' data. However this parameter estimation will not scale especially well with the size of the data set, and is quite time expensive.

Ingram and Csato (2005) also allowed their model the flexibility to learn its hyper-parameters from the data as part of the automatic procedure, and due to the evidence based nature of the hyper-parameter update in their algorithm they were able to cope with the very large release of radioactivity, although their estimates of the hyper-parameters look somewhat odd in the 3D depiction.

I was quite surprised to see the ordinary Kriging approach of Savelieva (2005) also perform well on this test set, given that they had fixed the hyper-parameters. I would have expected better results if the hyper-parameters had been re-estimated, however it was very difficult to tell exactly how the hyper-parameters had been set in the first place, so maybe their manual choice was lucky? Most methods achieved results of 100-120 on this test data, and this is what dominates the overall ranking in Table 1. It was a little odd that methods with fixed hyper-parameters seemed to do almost as well as methods with re-estimated hyper-parameters, although the best two by some margin both re-estimated. Give the anisotropic nature of the release in the 'joker' data, it is not surprising that in general methods that assumed isotropic behaviour did slightly less well. Almost all methods assumed some degree of stationarity, even Timonin (2005) have a single fixed width for their kernel functions, and at least at some level it is clear that the 'joker' ought not be regarded as data coming from a stationary process. However methods which do not require stationarity such as support vector machines and feed-forward neural networks also do rather poorly on test set 2. This might be explained by rather simplistic hyper-parameter estimation, or the use of fixed hyper-parameters in these methods.

5. CONCLUSIONS

What general conclusions can be reached from SIC2004? I think it shows that if you have a good prior knowledge of what you are trying to predict, and if you use this sensibly, then pretty much any non-linear method can be used to make the prediction (Table 2). The choice of algorithm then comes down to computational speed, cost of implementation, scaling with data size and the ability to make probabilistic predictions (Appendix I). It would be interesting to assess in any future exercise, not just the quality of the prediction of the mean, but also the quality of the probabilistic prediction, since for decision making it is the probability distribution that is critical, not just the mean value.

If on the other hand the data is very different from that you have seen previously, as in the joker set, then all that careful tuning of your algorithm on the training data sets goes out the window! It is interesting that some participants actually modified their prior beliefs when they noticed that a 'joker' data set was being introduced. Overall most models performed relatively poorly on the 'joker' data but this is generally something we would expect. It suggests that automatic mapping might still be something we need to work toward, rather than it being a solved problem. This might be especially true once we address issues such as non-stationarity, as exhibited by the 'joker', which is still a vexed problem for many methods. The most obvious way to account for this, in my view, is by the inclusion of physically based models, with a data assimilation strategy. However to accomplish this we must make further advances in the probabilistic modelling of complicated physical systems.

Acknowledgments

Codes

The spreadsheet with the results presented in Appendix I can be downloaded in Excel format from: http://www.ncrg.aston.ac.uk/~cornfosd/sic2004/

References

1. Bishop C M (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

2. Dubois, G., Galmarini, S. (2005). "*Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results*". This Volume.

3. Kalnay, E (2003). *Atmospheric Modelling, Data Assimilation and Predictability.* Cambridge University Press.

4. Nabney, I. T. (2001). *Netlab: Algorithms for Pattern Recognition.* Springer.

5. Wahba, G (1991). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics.

APPENDIX I

Overall summary table, comparing all methods (over two pages).

First Author	Method	Type	Kernel function	Isotropic?	Hyper-params fixed?	Training method (training set)	Training method (test set)
Timonin	NN	Kernel	Gaussian	no	no	Gradient based CV	Grid and gradient based CV
Ingram	NN	Kernel	Gaussian + exponential	no	no	Sequential Bayesian MAP	Sequential Bayesian MAP
Savelieva	Kriging	Geostats	Spherical	no	yes	Unclear - by eye?	Linear system
Hofierka	Splines	Kernel	Regularised spline with tension	no	no	CV	CV on test data (!)+ linear system
Pebesma	Kriging	Geostats	Spherical	yes	yes	Weighted least squares, grid CV	
Palaseanu	Kriging	Geostats	Circular	yes	sampled	Bayesian posterior, sampling	Bayesian posterior, sampling
Lophaven	Kriging	Geostats	Polynomial trend + 'general exponential'	yes?	yes?	SSE and CV?	?
Pebesma	IDW	Simple	Linear	yes	yes		Linear system
Saveliev	Splines	Kernel	B-spline + linear in GP	yes	yes	Grid based CV	Linear system
Ingram	NN	Geostats	Gaussian + exponential	no	no	Maximum Likelihood	Maximum likelihood
Pozdnoukhov	SVM	Kernel	Gaussian	yes	yes	Grid based CV	Quadratic programming
Ingram	NN	Geostats	Gaussian	no	no	Maximum Likelihood	Maximum likelihood
Fournier	Kriging	Geostats	Loess + exponential	no	no	SSE + WLS	SSE + WLS
Sridhar	RBF	NN	Gaussian	yes?	yes	SSE and CV (GA)	?
Hofierka	Splines	Kernel	Regularised spline with tension	no	no	CV	CV + linear system
Saveliev	Splines	Kernel	B-spline	yes	yes	Grid based CV	Linear system
Rigol Sanchez	MLP	NN	Sigmoidal	no	yes	SSE and CV (6 networks averaged)	?
Hofierka	Splines	Kernel	Regularised spline with tension	no	yes	CV	Linear system
Ingram	RBF	NN	Gaussian	no	yes	Maximum likelihood, with CV	Maximum likelihood, with CV
Ingram	MLP	NN	Sigmoidal	no	yes	Maximum likelihood, with CV	Maximum likelihood, with CV
Rigol Sanchez	MLP	NN	Sigmoidal	no	yes	SSE and CV	?
Sridhar	NN	Ward net	Gaussian + tanh + Gauss. complement	no	yes	SSE and CV (GA)	?

NN, neural network; RBF, radial basis function; IDW, inverse distance weighting; CV, cross validation; SSE, sum of squared errors; WLS, weighted least squares; GA, genetic algorithm.

First Author	Scaling	How is 'training' data used	Overall Speed	Prob. Pred.?	Exact (Dubois and Galmarini)	Skill score1	Skill score2	Overall skill score
Timonin	Moderate	To define possible values of kernel width	Fast?	Partial	GRNN	25.4	62.4	87.9
Ingram	Good	Set initial values for hyperparameters	Fast	Yes	Sparse Sequential Gaussian Process	25.0	78.9	103.8
Savelieva	Good	Set hyperparameters	Fast	Yes	Ordinary kriging	25.2	95.3	120.5
Hofierka	Moderate	Not used	Moderate?	No	Regularized Splines with Tension 1	25.0	97.7	122.7
Pebesma	Good	Set hyperparameters	Fast	Yes	Ordinary kriging	24.9	100.5	125.4
Palaseanu	Poor	Prior distributions for hyperparameters	Slow	Yes	Bayesian Kriging	25.0	101.6	126.6
Lophaven	Good	Structure of the model: mean and cov fn	Fast	Yes	Universal Kriging	26.4	102.1	128.5
Pebesma	Moderate		Fast	No	Inverse Distance Weighted (power 2)	26.9	102.5	129.4
Saveliev	Good	Set hyperparameters + GP	Fast	Partial	Multilevel B Spline A	25.7	105.1	130.8
Ingram	Poor	Not used	Fast	Yes	Gaussian Process	25.1	108.6	133.7
Pozdnoukhov	Good	To fix hyperparmeters	Fast	No	Support Vector Machine	23.8	111.2	135.0
Ingram	Poor	Not used	Fast	Yes	Gaussian Process 2	25.6	111.5	137.1
Fournier	Good	Structural	Fast	Yes	Robust Kriging	24.9	113.5	138.5
Sridhar	Good	Structure of network	Fast	No	RBF random division	27.5	112.4	139.9
Hofierka	Good	Set initial hyperparameters	Fast	No	Regularized Splines with Tension 2	25.5	115.0	140.5
Saveliev	Good	Set hyperparameters	Fast	Partial	Multilevel B Spline B	27.9	117.3	145.2
Rigol Sanchez	Good	To determine structure (GA)	Fast	Partial	ANN back propagation (add training)	32.4	122.6	155.0
Hofierka	Good	Not used	Fast	No	Regularized Splines with Tension 3	28.1	134.9	163.0
Ingram	Good	Structure of network	Fast	No	Radial Basis Function	26.6	138.0	164.6
Ingram	Good	Structure of network	Fast	No	MLP	25.7	148.5	174.2
Rigol Sanchez	Good	To determine structure (GA)	Fast	No	ANN back propagation	43.0	133.7	176.7
Sridhar	Good	Structure of network	Fast	No	LMA Ward Net	29.6	153.1	182.6

Scaling: how well the method extends to large N data sets. Overall speed of the method: fast = O(N), moderate O(N^2), slow = O(N^3). Note some of these are uncertain since it cannot be given with absolute certainty because the papers were not precise enough! GP, Gaussian Process; GRNN = generalised regression neural network; others as before.

The comparison of one click mapping procedures for emergencies

K. G. van den Boogaart

Institut für Mathematik und Informatik, Ernst-Moritz-Arndt Universität Greifswald, Germany

E-mail: boogaart@uni-greifswald.de

Abstract: Mapping in case of an emergency is an important, live saving issue and a big challenge. However comparative studies based on one or two datasets can be misleading. This is explained here on the example of SIC 2004 and of one of its goals, the comparison of one-click-mapping procedures for emergencies. Such comparisons have an inner tendency of preferring ill-suited methods and are often based on criteria unrelated to emergency mapping. A meaningful comparison must be based on multiple datasets, scenarios, robustness issues and emergency decision making oriented criteria.

1. THE GOAL OF EMERGENCY MAPPING

In the Spatial Interpolation Comparison (SIC 2004) one of the declared goals was to check automated mapping procedures to be fit for the case of an emergency. Clearly this is a very important problem: Emergencies typically evolve surprisingly and fast. Emergency forces at once need accurate and reliable spatial information e.g. for the evacuation of affected areas, decisions on allocation of save places and transportation routes, and for directing forces to the right places. Experts in spatial interpolation or even semiskilled persons are not readily available and no time is left for scientific considerations on the mapping procedure to be chosen. Important aspects of the emergency might be unknown prior to the mapping (e.g. the radioactivity released by the detected accident). Thus it is of vivid importance to have reliable, automated, and robust interpolation procedures readily at hand at all related emergency organisations. It might be important to keep the problem open and to improve procedures epsilon by epsilon from a scientific point of view, but it is the interest of all mankind to have something that works in the emergency offices now. As one of its goals SIC took this challenge and attacked this problem of general importance, something that could decide upon life and death of each of us. However, as I will explain here we should beware of generalising its results unreflectedly.

2. MODELLING THE UNEXPECTED

The participants willing to prepare such an automated mapping procedure, were informed about this goal of one-click-mapping in case of an emergency and provided with background information on the prospected dataset, a dataset on radioactivity. After preparation of such an automatic mapping algorithm they were provided with two datasets: A truly observed dataset without any signs of an emergency and one with an artificially introduced model of an extreme elliptical outcrop of radioactivity from a point source, which might be seen as one typical kind of emergency related to a reactor accident in case of no wind and diffusive transport only.

Other patterns of outcrops induced by directed transport through wind, rain, surface water, and ground water or other shapes of sources such as a nuclear war or acts of terrorism (multi point), a leaking truck or pipeline (linear), a returning satellite (sprinkled area), mine (complex known shape), or in house contamination after a laboratory accidents are equally reasonable. A mapping procedure fit for an emergency must be fit for all of these situations.

3. COMPARING AND SELECTING THE BEST METHOD

Afterwards we compare the different methods based upon their performance with this single dataset and are tempted to draw conclusions which method or group of methods performs "best" in case of an emergency. This comparison has been done by Cornford (2005) for SIC 2004 and throughout this paper I will use his findings to illustrate the problems we face her. Like in sports mankind is often helplessly attracted by the winner and tries to identify with him, tries to explain, why he won and how to follow the trace of the hero. But let us take one step back and ask the question what kind of method will win a general comparative study based upon a single dataset? Clearly the answer is: The method most adequate for the given dataset. Typically there is no uniformly optimal method, which is optimal for all kinds of datasets. This is a general finding in statistics (c.f. Witting, 1985; Shao 1998; Huber, 1964). Extreme values might represent the emergency or confusing outliers to be ignored. Trends might be neglectable, polynomial, or based on the topography corresponding to ordinary, IRFk or external drift kriging (Cressie, 1993; Goovaerts 1997). While nice fields are stationary and self similar, emergencies are local and atypical. The drift of the contamination might be isotropic, anisotropic or along special structures (like a valley), pretending anisotropy. Having this in mind it should be clear that there is no black box methodology which somehow magically adapts itself optimally to all these situations. Any adaptive method would pay a price in getting suboptimal in simple situations (e.g. Gaussian random field with mean zero and a known Gaussian covariogram), where clean simple methods such as simple kriging are known to be optimal. So the winner will be the method, which takes the right decision (e.g. concerning outliers), and can adapt to the complexity of the given dataset without over fitting parameters and not adapting to problems not actually present in the example. The methods and their corresponding models can not be ordered on a line from simple to complex, but include different kinds of complexity. So one of the methods fit for the right amount and right kind of complexity will take over, make the run and win the comparison.

What will now happen to the method we need for emergency mapping? This method must be good in all situations, and will thus never be the optimal procedure for any single dataset. It has to be robust and robust methods are never optimal in statistics. It will need to avoid over adaptation to ununderstood structure in the data. There will be some structure and it will never be understood and coded into the method in the moment of the emergency. The method will need to behave well with instationary data, especially in the special regions not looking like the rest of the field, because exactly in these regions the emergency is going on. So if there are enough methods in the comparison any of the procedures fit for the emergency will win no single contest based on a single dataset.
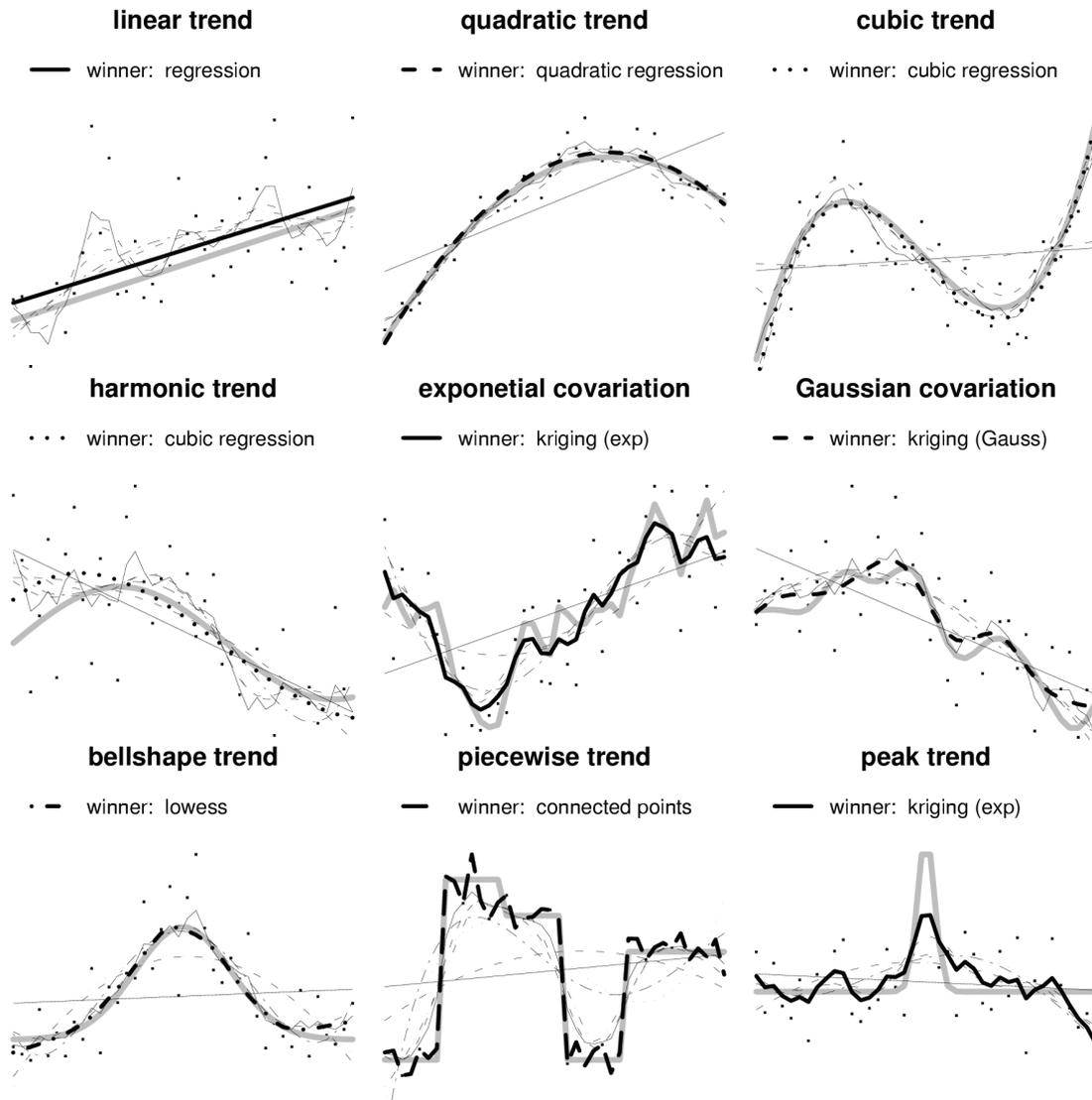
Figure 1: *Nine comparison studies for nonparametric regression. The grey line marks the true values. The dots are the values to be interpolated. The interpolation of the eight compared methods are drawn: 1)linear regression(solid), 2)quadratic regression (dashed), 3) cubic regression (dotted), 4) fourth order polynomial regression (dashed), 5) LOWESS (dash-dot), 6) kriging with an exponential covariogram (solid), 7) kriging with a Gaussian covariogram (dashed), 8) linearly connected observations (dashed). The method with the smallest mean square prediction error for each method is plotted in double line width.*
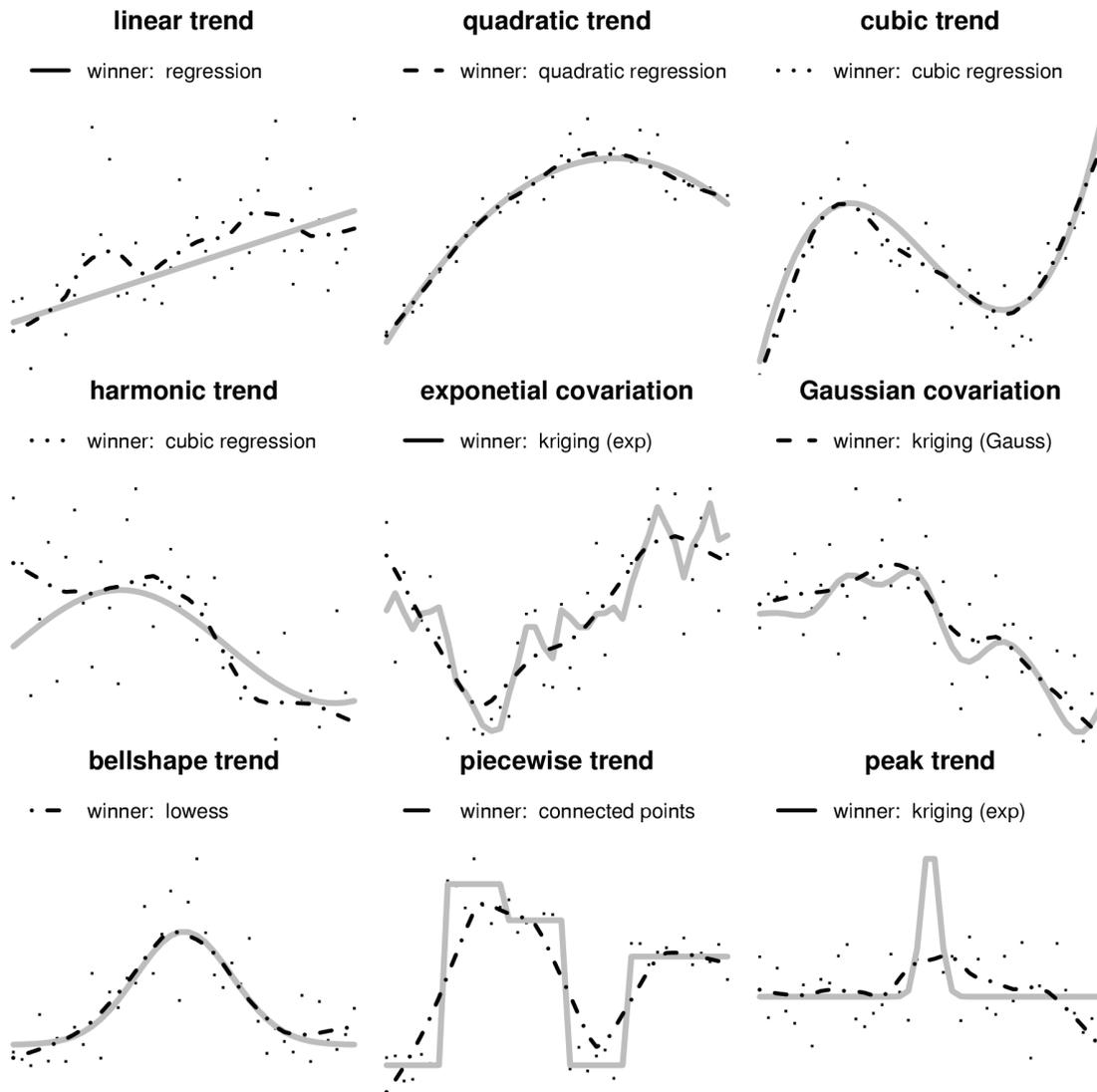
Figure 2: *Same as previous figure. Only the interpolations computed by LOWESS are given.*

4. AN ILLUSTRATIVE EXAMPLE OF WINNING A COMPARISON

To illustrate this behaviour of comparison studies I have prepared 9 comparison studies based on 9 different datasets for a simpler problem: The nonparametric regression problem. For each of the studies I created a deterministic function and provided 40 observations with i.i.d. Gaussian observation errors. Eight different regression and interpolation techniques were applied to each dataset: 1) linear regression (Witting, 1985), 2) quadratic regression, 3) cubic regression, 4) fourth-order polynomial regression, 5) linear interpolation, 6) LOWESS-smoothing (Cleveland, 1979; Cleveland 1981), 7) kriging based on a exponential covariogram and 8) kriging based on a Gaussian covariogram (Cressie, 1993). The interpolation and the comparison were done with the software R (R 2004). You can find the results in figure 1: The headline of each panel describes how the function was generated. The solid grey line shows the true regression function (which might be quite complex) and the black dots the observed data. Each panel further displays the seven interpolations (linear interpolation is not drawn). Above the plot we see a remark which of the techniques is the winner of the comparison in mean squared error. The line of the winning method is plotted in double width.

As we can see no method won more than 2 of the comparison studies. Five of the datasets (linear, quadratic, cubic, exponential covariation, Gaussian covariation) are closely related to the model behind the winning method. The remaining 4 datasets do not fit assumptions of any of the provided methods exactly and produced four different winners. So in conclusion, when Nonparametric Regression Comparison 2004 would have provided two of these datasets (e.g. the first one, which all methods did quite well and one of the others) we more or less get a random selection of which one should be the best method. E.g. this methodology of comparison proposed cubic regression, which is known to be an unstable method, twice.

So beware of comparative studies. They propose the method, which should not go into the agency, which tries to rescue you and your family during the next catastrophe, which will have other difficulties than that of the dataset the SIC 2004 comparison was based on.

In first place this argument seams biased and demagogic. One can easily argue the other way: Following the ideas in Cornford (2005) we could say the method by Timonin earning the best overall score and the best score for the emergency dataset and other methods of similar performance are distinguished by their ability to adapt to different situations away from the fine tuning done on the regular datasets and the emergency calls for built in adaptivity to the situation. However there were other methods in the contest with similar adaptivity properties and performing substantially worse, especially with the emergency dataset. The adaptation introduced by Timonin might be the right one and it might perform well in all situations, but we do not know this from SIC 2004.

5. A CLOSER LOOK TO EMERGENCY ONE CLICK MAPPING

Selecting a good one click mapping procedure for emergencies is mandatory and live saving. How should it be done? Emergencies are various. The procedure must give useful and reliable maps in all cases. There is no need for and no way to the best possible map in emergency one click mapping. Thus the procedures should be compared with multiple structurally different real and artificial datasets. The important aspect is not an optimal performance in average or a minimax type best worse performance but a reasonable and useful performance in all cases. Thus additive overall scores and scores based solely on purely mathematical criteria like mean square are not adequate. The purpose of emergency mapping is that of decision making. Decision makers need reliable information on exceeding of thresholds and can deal better with "*I don't know*" than with wrong information like "*not dangerous*". The area of interest in emergency decision making is the area of the event. An overall performance of it in the rest of the map is of no importance. Thus we should compare the methods based on their performance in the area of event, the special area, where everything is different.

One click mapping always contains prejudices: e.g. how to deal with outliers, which could either constitute the event of interest or destroy the map. When different decisions lead to substantially different maps a single map as an answer will be a dangerous guess. Maybe our emergency officer will realise that the map is wrong, when his map is missing a radioactive plume after the reactor accident, because it was eliminated as outliers, or maybe he accepts the map in wishful thinking. However when he is provided with both maps in case of big differences between the possible outcomes, he most probably will be able to decide which of the maps makes sense and which not. So we should reformulate the problem of one click mapping in a way that we first ask which kind of output can be handled by the decision maker. Soft skills of the method can be very important: A transparent relation of the data to the map might help to detect errors. Outliers and other problems should be listed and mentioned. The machine should never give the impression of infallibility.

6. ILLUSTRATED PITFALLS IN MULTIPLE COMPARISON

Going back to our illustrative figure 1 and having a closer look to the only method in comparison actually made for nonparametric regression: LOWESS. Even only winning only one of the comparisons LOWESS seems not much worse then the best prediction in all cases except in the peak trend case. So from a multiple comparison study using the criterion that it should perform acceptable in most cases we might have selected LOWESS. But now let us remember the special properties of emergencies: Some outliers might represent the hot spot of the emergency. And outlier situation is exactly represented by the peak trend case, where LOWESS would predict an elevated amount, but nothing above the background level. So if it comes to the emergency, LOWESS would fail and tell us that there is no emergency and only slightly elevated values. This is exactly because LOWESS was designed for being robust against outliers. Only one of the methods (kriging with exponential covariogram) managed this situation adequately, and this just because the range of the covariogram was in the right order of magnitude by chance. Now one will argue that kriging is based on estimation of the range from the data. But the central assumption of being able to estimate the covariance is stationarity. However stationarity is not given in the emergency mapping situation, like stationarity is not given in the peak trend situation. So seemingly none of the methods in my comparison is really fit for the purpose of say "nonparametric emergency regression".

7. ADAPTIVITY AND EMERGENCY MAPPING

Following the arguments of Cornford (2005) one result of SIC 2004 was that adaptive methods show a better performance with the emergency dataset. Seemingly adaptivity is an important property of mapping procedure fit for many datasets. On one hand it is seems clear that adaptive methods will lead to better results and thus introducing a single adaptive method into the comparisons it should have won more than two of the comparisons. Now LOWESS is strictly speaking an adaptive method even, when I fixed its hyperparameter of the influential region to 1/3. On the other hand comparing 9 different adaptive methods would lead to similar situation again.

However adaptivity is based on stationarity. Adaptive methods adapt their parameters to interpolate in the whole area in average as good as possible. However the emergency itself is typically represented by an atypical and special part of the area. The emergency is a break of stationarity and hopefully only affecting a small portion of the area. E.g. consider an outlier detection based on the typical local variability of the data and trained 95% normal low variability low value data and on 5% extreme data around an outcrop will declare will declare these 5% of extreme data as outliers and in thus remove the event from the data. A variogram estimation based on the same dataset will estimate a convex combination of variograms with 5% the variogram near the emergency and 95% of the variogram in the unaffected rest of the area. Thus the variogram fits in the area of the event better than the not adapted variogram, but not substantially. Thus adaptive methods do not adapt to the important region of the event, which are in the focus of interest in case of emergency mapping, but to the global map. Clearly, when afterwards comparing the fitness of the methods by comparison of the mean squared error in the whole area the adaptive methods will perform better. But the emergency teams need an accurate map in the area of the event, where adaptivity only has a slight and not necessarily positive effect.

8. A PRELIMINARY SOLUTION TO EMERGENCY MAPPING

So, before SIC 2009 will finally solve the problem of the optimal emergency mapping procedure by a comparison based on multiple datasets and emergency mapping related techniques, summarising all the requirements, I would propose ordinary kriging with a local kriging neighbourhood based on a linear variogram without nugget as a preliminary candidate for the emergency mapping. A kriging procedure without nugget will interpolate the observations (Cressie 1993) giving an easy (linear) relation of interpolation and data, so that anybody can see which observations generate extreme values. The emergency officer can thus

himself decide, whether he believes the data or not. Furthermore the interpolation should be scale invariant in space and value since the scale of the emergency event is a priory unknown and can not be estimated from the data due to missing stationarity (c.f. self similarity (Kent & Mardia, 1994). Emergencies are non stationary and thus the variogram should be consistent with non second order stationary fields to avoid unjustified far range inference. The procedure should use a local kriging neighbourhood to be fast and to avoid far range influences between the affected and the unaffected region. All these properties are summarised in the proposed method. Clearly this type of kriging without any variogram estimation will not win any single competition and even not give a very small expected squared error. But it will transparently outline any emergency region which is actually measured in the data and its kriging error provides a useful non quantitative unbounded measure of local density of the measurements providing a way to find blind spots. And now using SIC2004, we should not forget that according to Cornford (2005) ordinary kriging by Savelieva and Pebesma performed quite well in both provided examples.

9. NON EMERGENCY ONE CLICK MAPPING OF STATIONARY FIELDS

To honour the results of SIC 2004 in the right way we should also be aware of the fact that automated one click mapping of stationary fields is well different from one click mapping in case of an emergency, which was discussed here. One click mapping of stationary fields needs to adapt itself in the right way to the dataset. This is e.g. possible by cross validation due to the stationarity. Clearly still the best robust procedure will not win comparisons, however it will still be one of the well working procedures in all most comparisons, allowing to select candidates from each comparison performed.

10. SUMMARY

Before we decide on a useful automated mapping procedure for emergencies we need more comparisons like SIC 2004 with more different datasets and settings. We need to test the fitness of the methods with multiple scenarios and multiple true datasets. We should not look at overall summed scores, but towards its robustness properties asking for reasonable and useful results in all cases. The performance and utility of the methods should be assessed in terms of decision making and not in terms of mean square errors. Maybe in a novel comparison could not only ask methods to participate but also datasets from different emergencies all over the world.

References

1. Cressie, N.A.C. (1993). *Statistics for Spatial Data (*rev. ed.): J. Wiley & Sons

2. Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829-836.

3. Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, **35**, 54.

4. Cornford, D (2005). Why comparative studies are a waste of time! , *This Volume*

5. Dubois, G. and S. Galmarini (2005). Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results. *This Volume*

6. Goovaerts, P (1997). *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York

7. Huber, P.J. (1964). Robust Estimation of a Location Parameter, in Breakthroughs S. Kotz, N.L. Johnson (Ed) (1991) Breakthroughs in Statistics Volume II, Methododology and Distribution, Perspective in statistics, Springer Series in Statistics, Springer, New York

8. Kent, J.T., K. V. Mardia (1994). The link between kriging and thin-plate splines, In Probability, Statistics and Optimisation: a Tribute to Peter Whittle, Kelly, F.P. editor, Wiley, Chichester, pp. 325-339

9. [R 2004] R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

10. Shao, J. (1998). *Mathematical Statistics*, Springer, New York

11. Witting, H. (1985). *Mathematische Statistik I. Parametrische Verfahren bei festem Stichprobenumpfang*, B.G. Teubner, Stuttgart, 538 pp.

Spatial Interpolation Comparison exercise 2004: a real problem or an academic exercise?

Donald E. Myers

Department of Mathematics, University of Arizona, Tucson, AZ 85721 USA

E-mail: myers@math.arizona.edu

Abstract: SIC2004 was posed as an exercise to compare various "automatic mapping" algorithms. Several important issues arose out of this exercise. One is whether automatic mapping algorithms are sufficient for the problem, i.e., the detection and location of radioactivity release events. A second issue is how automatic the algorithms should be. Most of the participant responses did not use automatic methods but rather required significant intervention by the user, hence as a possible monitoring methods they are less than efficient. Finally most participants "trained" their methods on sets of normal data, it then became clear that the training was inadequate to apply to a data set with a simulated event. It is also found that a number of the different algorithms that purport to be different are in fact equivalent, e.g., radial basis functions and kriging (which differ primarily in the practice). This suggests a need to pose new questions about such an exercise.

Keywords: automatic mapping, interpolation, kriging, radial basis functions, cross-validation, comparative measures

1. INTRODUCTION

The instructions for SIC2004 focused on "automatic mapping" algorithms and in particular to compare the different algorithms both for a data set quite similar to the training sets and the "joker". As noted by Cornford (2005) the differences for former were rather slight and for the latter not as different as might have been expected. The underlying problem in SIC2004 is a real one in space-time, namely to monitor radioactivity and perhaps to sound an alarm in case of an "event". Galmarini (2005) discusses an actual instance when unusual radioactivity levels were detected. Both the results from the various participants and also the various discussions, especially the two noted above, suggest that it would be worthwhile to ask some other questions. In particular is "automatic mapping" the real need and how "automatic" should the algorithms be? It seems clear that none of the reported methodologies were really automatic, they were merely completed in a very short time frame but with more or less involvement of the participant. That is not a criticism of what was done but merely an observation. Another question is, what is being compared and are the right measures of any differences being used? The following is not intended as a disapproval of the competition but suggest that it could not answer some very important questions and also to point out that sometimes science progresses when research does not give the expected answers. Van den Boogaart (2005) raises similar issues but in slightly different ways.

2. WHAT KIND OF COMPARISONS?

There are many articles in the literature "comparing" different spatial interpolation "methods". See Matheron (1981), Creutin and Obled (1982), Franke (1982), Lam (1983), Warrick et al (1988), Grimm and Lynch (1991), Laslett et al (1987), Laslett and McBratney (1990), Laslett (1994), Myers (1994), Zimmerman et al (1999). Most are empirical (as opposed to theoretical) and in many cases they are really comparisons of particular methods by specific individuals or groups on specific data sets. If the competition were in the nature of

a contract negotiation, i.e., to pick the person or group to be given the responsibility of designing and implementing an "automatic mapping" scheme for this particular type of data and this set of monitoring stations, then it certainly provides some guidance as to a reasonable choice. If the objective is to identify possible methodologies and perhaps provide some information on the relative efficacy of each then the competition has succeeded (although the list of possible methodologies may not be complete). If the objective is to focus on methodologies (as opposed to particular implementations and usage by specific groups or individuals) then it does not provide that information, even assuming that the nature of the data sets will remain fixed (the number and locations of the monitoring stations as well as the nature of the data collected). To do that it would be necessary to have one group carry out all the studies with clear standards for the software, for the nature and extent of the possible human interaction with the software during computations, common hardware configurations and perhaps a more elaborate set of measures of success. At the very least it would be necessary to have a description of what was done (and the software used) sufficient to allow others to duplicate all the reported results. Reproducible results are critical and the requirement for them should not be interpreted as an insult or an assault on anyone's integrity. We are all sometimes guilty of having a solution in search of a question, i.e. a favorite methodology or technique which we apply to any problem that comes along.

Having identified various methodologies, it is important to ask about the underlying assumptions and also whether there are real differences in the methodologies or rather they simply arise in different contexts (e.g., statistics vs. numerical analysis) with different terminologies and different common practices. This question will be addressed in part subsequently.

3. MAPPING OR SOMETHING ELSE?

EURDEP is a broad environmental monitoring program which includes almost continuous sensing of radioactivity levels at a very large number of locations in the various European countries. It would seem that the data collected by this network could and would be used for a number of purposes. A most obvious one is simply to detect the malfunctioning of one or more sensors. A second is to detect any events, whether accidental or intentionally caused, that might lead to dangerous radioactivity levels either on a wide spread basis or locally. This would likely involve detecting areas where the levels are significantly higher than others. But it would also be important to detect an increase occurring over a very large area. Thus there would be an interest in both spatial and temporal "outliers" or anomalous readings. While maps might be used for these purposes, it is unlikely to be an automatic process, i.e. such maps would likely have to be interpreted by humans. Would an automatic mapping system have provided a timely alert in the case of Chernobyl?

The presumed purpose of "automatic" methods is that they would produce results very rapidly and at frequent intervals. One might expect that many if not most of these maps would provide little or no real information, i.e., no change over time and no spatial anomalies. If temporal change is to be detected by comparing successive maps, how many must be stored and for what time interval will they be compared? Even if a "best" automatic mapping methodology is identified and implemented, there will still be a need to have automatic interpretation methods and automatic map monitoring methods. This suggests that perhaps the focus should be on methods aimed at detecting spatial and temporal anomalies directly from the data.

This is not say that mapping methods, whether automatic or not, are not useful or important. Given the large number of sensors in Germany (reported as over 2000), mapping techniques might be used to determine whether a smaller number of sensors would provide the same information. This problem has received a considerable amount of attention, particularly in the geostatistics literature. Indirectly this is almost the question that was posed in the study except that the number of sensors in the subsample was already fixed as well as their locations.

4. HOW AUTOMATIC IS AUTOMATIC?

There were seismographs in place and in operation in many places in Southeast Asia on December 26, 2004 when the quake and subsequent tsunami occurred. In particular the instrument at the National Geophysical Research Institute (Hyderabad, India) was functioning and recorded the event. The problem was two fold; December 26 was a Sunday and the event occurred early in the morning (local time) and there was no one in attendance to see the recording until it was too late, secondly the countries most impacted by the tsunami did not participate in the Pacific tsunami warning network and there was no system in place to communicate the warning to those most likely to be affected. In the case of the tsunami rapid and immediate communication was critical. In the event of a radioactivity event, depending on the nature and the extent of the event, the need for rapid communication might be greater or lesser. There is also the potential difference between alerting the public and alerting governments and governmental officials.

It would seem highly desirable for the software to not only detect spatial and/or temporal anomalies but also to communicate that information to the appropriate person or persons. If the algorithm did rely on mapping then the algorithm should also include automatic analysis of the resulting maps. Is automatic mapping a principal part of the overall software package or only a small part? Will the software be running on computers adequately linked to networks both for continuously accessing the data and also for sending out alerts?

Clearly software testing as well as security for the network and the computing would be an ongoing need. It is equally clear that there is the potential for sensor malfunction and if the system is truly automatic, i.e., data is fed directly from the sensors to the software, then there must be some sort of system for cross checking for possible malfunctions to avoid issuing a false alert or failing to issue an alert when an event has occurred.

Pebesma (2005) gives the clearest example of an automatic mapping scheme, i.e., an R script that uses the *gstat* package for R (the script is given in the Appendix of the paper). Unfortunately as the author subsequently notes, the attempt illustrates the need to test the code on a variety of data sets and also illustrates the kinds of numerical problems that might arise. A code such as this would need some system to alert an operator if it were really operating in an automatic environment. None of the other submissions are really as automatic as this and nearly all encountered problems, i.e., because they had been designed to function with data files very similar to the training sets and they did not perform well on the joker data set and the authors did not have at hand ready resolutions for the problems.

Goovaerts (2005) gives another example automating the interpolation using indicator kriging. Like as in Pebesma (2005) there is an automatic algorithm for fitting the variograms, however in this application of indicator kriging (to estimate and fit the local conditional probability distributions it is necessary to fit multiple variograms (for multiple cut-offs). Weighted least squares is an attractive choice, it is akin to the common "fitting by eye" but it also has several dis-advantages. There are no results to show that closely fitting a model to the variogram cloud ensures best behavior of the model in kriging. Secondly, it is necessary to make assumptions about the maximum lag distance to use (and in some instances to make decisions about the distance tolerance and angle windows a priori). To utilize the conditional probability distributions it is usually necessary to modify the partial empirical distributions to ensure that there are no order relation problems but this technique, Deutsch and Journel (1998), is an ad-hoc one with no error bounds. In a non-automatic context the user can inspect the fits and possibly modify them but not in the automatic algorithm. These concerns simply point out the difficulty of a truly automatic algorithm.

5. KRIGING, RADIAL BASIS FUNCTIONS, SPLINES AND KERNEL FUNCTIONS

It has been known for many years that the Radial Basis Function (RBF) interpolator is exactly the same as the Intrinsic Random Function (IRF-k) interpolator although it is derived quite differently and initially it appears different. The RBF interpolator includes as a special case the Thin Plate Spline (TPS). Matheron (1981) gives an explicit construction showing the equivalence of kriging and the TPS. The smoothing spline is a special case of cokriging but in fact to change from TPS to the smoothing spline requires only a minor alteration in the usual kriging equations.

Madych and Nelson (1988) give a very succinct derivation of the both the RBF interpolator and the equations determining the coefficients without using Fourier transforms, The problem is almost exactly that of the study, given a data set estimate/predict the values at non-data locations. Although in the RBF context the problem is to fit a continuous function to the data. To obtain a unique solution, it is assumed given a continuous to strictly conditionally positive definite function (c.p.d.). In the notation of Madych and Nelson, an (ordinary) covariance function would be conditionally positive definite of order zero, the negative of a variogram would be conditionally positive definite of order 1. The negative of the multiquadric function is conditionally positive definite of order 1. The multiquadric function is not quite a variogram, it is not normalized, i.e., its value at zero lag is not zero. However unlike positive definite functions, conditionally positive definite functions of order 1 still have that property if a constant is added, the multiquadric can be normalized. Note that if a function is conditionally positive definite of order k, it is also conditionally positive definite of any higher order.

Madych and Nelson show that for any c.p.d. h, there is a subspace C_h of the continuous functions, each member of which interpolates the data. Moreover there is a semi-norm and the element of C_h with minimum norm is given by

$$s(x) = \mathbf{b}^T \mathbf{h}_x + \mathbf{a}^T \mathbf{P}(x) \qquad (1)$$

where

$$
\begin{bmatrix} \mathbf{H} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \times \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} \qquad (2)
$$

H is a matrix whose elements are of the form $h(x_i\text{-}x_j)$, $\mathbf{h}_x$ a vector with elements of the form $h(x\text{-}x_j)$. In the case of a covariance function or a variogram the coefficient matrix in eq(2) is the same as in the universal kriging equations and eq(1) is the same as eq(1) in Lophaven et al (2005). Lophaven et al refer to this as the universal kriging estimator (although that is not the form usually given in the geostatistics literature and certainly not the form originally given by Matheron). However it is equivalent to it by an elementary application of linear algebra

$$
\begin{bmatrix} \mathbf{H} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \times \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \mathbf{h}_x \\ \mathbf{P}(x) \end{bmatrix} \qquad (3)
$$

$$s(x) = \lambda^T \mathbf{v} \qquad (4)$$

The system in (3) is obtained by constraining (4) to be unbiased and minimizing the error variance. Minimizing the error variance is analogous to minimizing the norm in C_h and hence $s(x)$ in eq(1) is analogous to the expected value over C_h. It is of course well-known that (for a random variable X with finite variance) that the minimal variance predictor of X is the expected value. The interpolating function given by (1) is found in several of the different submissions albeit sometimes with different names and with different derivations of the equations determining the coefficients. In the geostatistics literature (1) would be known "dual kriging". Neither the theorems, nor their proofs, in Madych and Nelson provide any insight into the choice of h. There is a minimal norm interpolating function for each choice of h.

In summary it is nonsensical to argue whether kriging, RBF or spline (and perhaps kernel) interpolation is better. In particular to claim that interpolation with the multiquadric RBF was better in 1997 than ordinary kriging misses the point. There are differences in practice and even in the kinds of results between the use of these equivalent methods. The RBF derivation does not require any notion of stationarity and hence no insight into the order of the polynomial term in (1). From the stochastic model it is easy to see that the polynomial component determines the interpolation far away from the data locations, i.e., it is an unbiased estimator of the mean. In kriging, i.e., when using the interpolator in the form in (4) it is natural to consider emphasizing only the data locations that are close to the location to be estimated, hence the use of a moving search neighborhood. This avoids having to invert very large matrices (by inverting a large number of smaller matrices). In the RBF form there is no obvious way to use a moving search neighborhood but at least two approximating schemes have been developed; Faul and Powell (2000) used a Krylov subspace method whereas Schaback and Wendland (2000) used a "greedy" algorithm. In each case the ultimate RBF interpolator is approximated arbitrarily well using RBF interpolators based on increasing subsets of the data. The RBF form does not naturally lead to the use of cross-validation as does kriging but it can be extended as shown in Myers (1992b).

6. UNCERTAINTY

In one way or another all of the algorithms presented in the various studies use the same basic idea. To wit; the values at two locations close together are more similar than the values at two locations far apart. The key point is how to quantify "similar", continuity is a form of spatial similarity. One important consequence of this basic idea that seems to have been overlooked by all the contributors is that when predicting or estimating the value at a non-data location there is uncertainty associated with the estimation. In a real application the true value will be unknown and hence the true error unknown. The error can only be dealt with in terms of uncertainty. Moreover this uncertainty is not spatially constant. The kriging variance is a relative measure of this uncertainty and clearly incorporates information about the spatial distribution of the data locations and also the proximity of the estimation location relative to the data locations as well as quantifying the spatial correlation. No one seems to have bothered to generate plots of the kriging standard deviation, there is no counterpart for the RBF interpolator (nor the splines). Uncertainty might also be dealt with in the context of a Bayesian formulation. Uncertainty is important in cross-validation.

Some authors, Stein (1999), Diggle et al (2003) refer to the practice of estimating the covariance function/variogram from the data and then using it in the kriging equations as though it were the true model as "plugin" geostatistics. They argue that there is an uncertainty associated with the kriging estimator that is not quantified in the kriging variance. That is true but replacing it with the uncertainty of assuming a multivariate normal distribution in order to use ML or REML estimation or perhaps a Bayesian formulation replaces one form or source of uncertainty with another. A univariate transformation of the data to normality, e.g., using a Box-Cox transformation, does not ensure multivariate normality.

7. CROSS-VALIDATION

A number of contributors used cross-validation either to choose the covariance function/variogram or the drift function or the search neighborhood. Cornford (2005) uses a variant to construct a "skill" measure to compare the results of the various contributors. There are at least seven statistics that quantify how well the interpolator is working. These are

1. Mean Error, ME $=(1/n)\Sigma_{[i=1,...,n]}[v_i - v_i^*]$

2. Mean Absolute Error, MAE $=(1/n)\Sigma_{[i=1,...,n]}|v_i - v_i^*|$

3. Mean Square Error, MSE $=(1/n)\Sigma_{[i=1,...,n]}[v_i - v_i^*]^2$

4. Mean Normalized Square Error, MNSE $=(1/n)\Sigma_{[i=1,...,n]}\{[v_i - v_i^*]/\sigma_i\}^2$

   where σ_i^2 is the kriging variance at location i.

5. Correlation of v_i and v_i^*

6. Correlation of $v_i - v_i^*$ and v_i

In addition one should examine the histogram of the errors and of the normalized errors. Based on Chebyshev's Inequality

$$P[|v - v^*| > k\sigma] < 1/k^2$$

Hence one would expect the proportion of normalized errors exceeding 3.0 (in absolute value) would not be more than 1/9. In general Chebyshev's Inequality underestimates the probability since it requires no distributional assumptions. While intuitively one would expect the ME, MAE and MS to be small, in the case of kriging one can say more. The expected ME is zero and the expected MNSE is 1, this is more definitive than just requiring the MSE to be small. It also takes into account the uncertainties associated with the different estimates. In the case of Simple Kriging, the correlation of the v_i and v_i^* would have a maximum and ideal value of 1 but in the case of ordinary or universal kriging the maximum correlation is less because of the Lagrange multipliers. Likewise in the case of Simple Kriging, the correlation of $v_i - v_i^*$ and v_i ideally should be zero but in the case of ordinary and universal kriging the minimum value is greater than zero. The spatial distribution of errors is important in cross validation (and even in computing the "skill" measure of Cornford (2005)). The different cross validation statistics do not necessarily act in concert with one another, i.e., making a change in the covariance function or variogram may make one "worse" while leaving others unchanged or perhaps slightly improved. Some are more sensitive to changes than others.

One of the possible uses of the cross validation statistics is to identify "unusual" location/value pairs. Pebesma (2005) noted the presence of one such pair and omitted that pair from subsequent analyses. There are different reasons why a location/value pair may appear unusual and if possible one should attempt to first ascertain whether the reason(s) might be unrelated to the interpolation methodology. Obviously if there are a large number of unusual location/value pairs this suggests difficulties with the interpolation methodology. Only after the exploratory analysis is complete should one use the cross validation statistics to compare different covariance function/variogram models and parameters.

8. THE ALGORITHMS AND THEIR ASSUMPTIONS

It was noted by all the participants that the ten training data sets were quite similar, this is further documented in the introduction by Dubois and Galmarini (2005). This allowed the participants to choose the algorithm, set parameters and particular features of the algorithm to "tune" the results. Those training sets represent the situation when nothing untoward is happening. As was noted by many of the participants, their models and methods did not work as well on the Joker data which was intentionally chosen to be somewhat different. Clearly some of the underlying assumptions changed yet few of the algorithms provided for any checking of the assumptions nor provisions for incorporating those changes. In many instances the underlying assumptions are made because they are convenient rather than because they reflected characteristics of the data sets.

A. Stationarity

Anyone who uses kriging (ordinary, universal) knows that there is an underlying assumption of at least intrinsic stationarity. They also know that there are no straightforward tests for this stationarity. There are however characteristics to look for in the data that would suggest the assumption is not satisfied. Simple kriging requires second order stationarity. What is not always recognized is that Indicator kriging requires a stronger form of stationarity, namely at least all bivariate probability distributions are translation invariant. So-called Model based geostatistics, including Bayesian kriging requires not only multivariate normality but also the strong form of stationarity, i.e., translation invariance of all finite joint distributions. Bayesian methods are often touted as more robust, i.e., they incorporate uncertainties associated with fitting the covariance function but they do not incorporate the uncertainties associated with the untestable assumption of a stationary multivariate normal distribution. It might seem that there are no stationarity assumptions associated with the use of splines or Radial Basis Functions. However it has been long known that the thin plate spline is a special case of kriging, i.e., with a particular choice of the generalized covariance hence a weak form of stationarity is implicit. The Radial Basis Function interpolator is in fact really identical to kriging, the coefficient matrix for the equations determining the weights in the RBF interpolator is exactly the same as the coefficient matrix in the system of equations determining the weights in the kriging estimator. It only takes a bit of linear algebra to transform the one estimator into the other. Hence to claim that the RBF is better or worse than kriging is nonsense. Any difference is attributable to a difference in the choice of the generalized covariance, the size of the search neighborhood and other user choices (or perhaps the software implementation).

B. Distributions

Although G. Matheron reportedly was motivated by the form of the conditional expectation in the case of multivariate normality, in most of the usual variations on kriging (simple, ordinary, universal, point vs. block, IRF-k) the equations for the weights can be derived without any assumptions about the underlying distributions. Only an appropriate form of stationarity is necessary, which in turn ensures the existence of the relevant generalized covariance (although in indicator kriging a stronger form of stationarity is necessary). In fact the stationarity is primarily used to justify the process of estimating the generalized covariance.

The algorithms that either used Bayesian methods or Gaussian processes however make explicit use of an assumption of multivariate normality as well as the strong form of stationarity. While an automatic algorithm could include a check for univariate normality and/or a Box-Cox transform, the use of this transform does not ensure multivariate normality. To the extent that the methods did not perform well on the joker data it is not clear whether this may be due to violating the distributional assumptions.

9. USER CHOICES

Most interpolation algorithms including the ones used in this study require the user to make some choices, although in some cases these have already been made by the programmers. These may be dictated by or at least partially determined by the characteristics of the data set including its size. The computational demands for 200 data locations is not the same as for 1008 locations. An example of the importance (and constraints resulting from) user choices appears in a number of the participant contributions, namely the possible complexity and model choices for variograms. I.e., how many components are permitted in the nested variogram model and which types are permitted? Note that one often sees nested models used in published papers where the components are the same type, e.g., spherical, but have different sills and ranges. In general weighted least squares will not generate such models, the minimization technique can not distinguish multiple occurrences of the same type in a variogram model.

10. OTHER ALTERNATIVES

While it has not been claimed that all possible mapping algorithms have been tested in this competition, it is worthwhile to at least consider others. One that perhaps should have received more attention is data mining. There is no single, unique algorithm for data mining but the overall objective is pretty well understood. Namely, to determine if there are interesting "patterns" or useful "information" in a data set, neither of these terms is assumed to have a hard definition or understanding but rather the meaning is assumed to vary from one data set to another and from one context to another. While many applications of data mining are non-spatial, spatial data mining has attracted considerable attention in recent years. In some instances the spatial data mining techniques have incorporated one or more of the mapping algorithms reported in SIC2004. See Hosking et al (1997), Ester et al (1997), Chawla et al (1999) and Lu et al (2003) for just a few examples of papers on spatial data mining.

11. CODES AND SOFTWARE

There are some codes and programs that are regularly reviewed and evaluated by disinterested parties. For example, commercial statistical packages such as SAS, SPSS, S-PLUS, STATA are reviewed and reported on in journals such as The American Statistician. GIS packages such as ARCVIEW, ARC/INFO are commonly reviewed. Free software such as R (and various packages for R) are subjected to extensive checking by many individuals and groups. STATLIB (2005) maintains codes (including source codes) for algorithms published in various statistics journals. Regrettably there is no system nor common practice with respect to software and codes for geostatistics algorithms. The various programs provided by GSLIB were clearly intended to be illustrative rather than end products (according to the authors).

Any effort to compare different algorithms as in SIC2004 must include evaluation of the software implementations. Evaluation must include checking to see that the algorithm is correctly implemented. It is well known that there are various possible implementations for standard computations, e.g., finding the eigenvalues of a matrix, solving a system of linear equations but these are not necessarily equivalent nor equally valid. Numerical Recipes (1988) provides both standard algorithms, comparative information as well as tested, documented codes (FORTRAN and C).

This is not impugn any of the participants but a complete comparison of the results would have to include an evaluation of the software, certainly any claims about computing speed or accuracy have to be considered in the light of the software implementation as well as the hardware configuration.

References

1. Chawla, S., S. Shekhar, W. L. Wu, and U. Ozesmi (1999). Modeling spatial dependencies for mining geospatial data: An introduction. In Harvey Miller and Jiawei Han, editors, Geographic data mining and Knowledge Discovery (GKD)

2. Cornford, D. (2005). Are comparative studies a waste of time? SIC2004 examined! *This Volume.*

3. Cressie, N. (1990). Response to Letter to the Editor. The American Statistician **44**, 256-259

4. Creutin, J.D. and C. Obled (1982), Objective analyses and mapping techniques for rainfall fields: An objective comparison. *Water Resources Res.* **18**, 413–431.

5. Diggle, P.J., P.J. Ribiero, Jr and O.F. Christensen (2000). An introduction to model based geostatistics. In Spatial statistics and computational methods, J. Møller (ed), Springer, NY 43-86

6. Deutsch, C.V. and A.G. Journel, (1998). *GSLIB: Geostatistical software library and user's guide* (2nd edition), Oxford University Press, Oxford, p. 369

7. Dubois, G. and S. Galmarini (2005). Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results. *This Volume.*

8. Dubois, G. and S.A.R. Shibli, (2005). Monitoring radioactivity in the environment: automatic mapping or expert-dependent systems? in G. Dubois (ed) Automatic mapping algorithms for routine and emergency monitoring data. Office for Official Publications of the European Communities, Luxembourg, EUR 20667 EN, EC. pp. 253-268.

9. Ester, M., H.-P. Kriegel and J. Sander (1997). Spatial data mining: a database approach. in Int. Symp. Large Spatial Databases (SSD'97), 47 66, Lecture notes in computer science, Berlin, Germany, July 1997.

10. Faul, A.C. and M.J.D. Powell (2000). Krylov subspace methods for radial basis function interpolation. In Numerical Analysis, Dundee (ed) Chapman and Hall/CRC Res. Notes in Math 420, 115-141.

11. Franke, R. (1982). Scattered data interpolation: tests of some methods. *Mathematics of Computation*, **38**, 181–200.

12. Galmarini, S. (2005). Real-time geostatistics for atmospheric dispersion forecasting and vice versa? *This Volume*

13. Goovaerts, P. (2005). Automatic Interpolation of Network Data using indicator kriging. *This Volume.*

14. Grimm, J.W. and J.A. Lynch (1991). Statistical analysis of errors in estimating wet deposition using five surface estimation algorithms. *Atmospheric Environment* **25a**, 317–327.

15. Hosling, J., E. Pednault and M. Sudan (1997). A statistical perspective on data mining. *Future Generation Computer Systems* **13**, 117-134.

16. Lam, N.S. (1983). Spatial interpolation methods: A review. *American Cartographer* **10**, 129–149.

17. Laslett, G.M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *Jour. Am. Stat. Assoc.* **89**, 391–409.

18. Laslett, G.M. and A.B. McBratney (1990). Further comparison of spatial methods for predicting soil pH. *Soil Science Society of America Journal* **54**, 1553–1558.

19. Laslett, G.M., A.B. McBratney, P.J. Pahl and M.F. Hutchinson (1987). Comparison of several spatial prediction methods for soil pH. *Jour. of Soil Science* **38**, 325–341.

20. Lu, C-T., D. Chen and Y. Kou (2003). Algorithms for spatial outlier detection. in Proc. Of the third int. conf. on data mining (ICDM '03) , 0-7695-1978-4/03 IEEE

21. Madych, W.R., and S.A. Nelson (1988). Multivariate interpolation and conditionally positive definite functions. *Approx. Theory & its Appl.* **4**, 77- 89.

22. Marcotte, D. and M. David M. (1988). Trend Surface Analysis as a Special Case of IRF k Kriging. *Mathematical Geology* **20**, 821-824.

23. Matheron, G. (1973). The intrinsic random functions and their applications. Advances in Applied Probability **5**, 439-468.

24. Matheron, G. (1981). Splines and kriging: Their formal equivalence. In Down-to-Earth-Statistics: Solutions looking for geological problems. D.F. Merriam (ed) Syracuse University Geological Contributions, Syracuse, NY, 77-95.

25. Myers, D.E. (1988). Interpolation with Positive Definite Functions. *Sciences de la Terre*, **28**, 251-265.

26. Myers, D.E. (1992a). Cokriging, Radial Basis Functions and the role of Positive Definiteness. *Computers Math. Applications* **24**, 139-148.

27. Myers, D.E. (1992b). Selection of a radial basis function for data interpolation. in Advances in Computer Methods for Partial Differential Eq. VII, R. Vichnevetsky, D. Knight and G.Richter (eds), IMACS, 553-558.

28. Myers, D.E. (1994). Statistical Methods for Interpolation of Spatial Data. *J. Applied Science and Computations* **1**, 283-318.

29. Pebesma, E. (2005). Mapping radioactivity from monitoring data: automating the classical geostatistical approach. *Applied GIS* **1**(2), doi: 10.2104/ag050011.

30. Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling (1988). *Numerical Recipes in C: The art of scientific computing*. Cambridge University Press, p735

31. The R project for statistical computing (2005). http://www.r project.org/

32. Schaback, R. and H. Wendland (2000). Adaptive greedy techniques for approximate solution of large RBF systems. *Numerical Algorithms* 24, 239-254

33. STATLIB (2005): a system for distributing statistical software, datasets and information. http://lib.stat.cmu.edu/

34. Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging*. Springer, NY p247.

35. van den Boogaart, K.G. (2005). The comparison of one click mapping procedures for emergencies. *This Volume.*

36. Warrick, A.W., R. Zhang, R., M.K. El Haris and D.E. Myers (1988). Direct comparisons between kriging and other interpolators. in Proceedings of the Validation of Flow and Transport Models in the Unsaturated Zone, Ruidoso, NM, 23 26 May, 505-510.

37. Zimmerman, D., C. Pavlik, A. Ruggles and M.F. Armstrong (1999). An Experimental Comparison of Ordinary and Universal Kriging and Inverse Distance Weighting. *Mathematical Geology* **31**, 375-390.

Automatic Interpolation of Network Data using Indicator Kriging

Pierre Goovaerts

BioMedware, Inc. Ann Arbor, MI, USA

E-mail: goovaerts@biomedware.com

Abstract: This paper presents an implementation of indicator kriging that performs automatically the following tasks: selection of thresholds for binary coding of data, computation and modelling of indicator semivariograms, modelling of probability distributions at unmonitored locations, and estimation of unknown attribute values as the mean of these distributions. The approach is very flexible since there is no priori assumption regarding the shape of the histogram, and it provides a non-parametric measure of uncertainty. Application of indicator kriging to the spatial prediction of gamma dose rates shows that it performs well in absence of anomalies. Despite its robustness with respect to outliers, the method fails to predict accurately the extreme values in the "joker" dataset. This little exercise reminds us that geostatistical prediction is not a magic wand and, in absence of auxiliary information, outliers at unmonitored locations will probably go undetected.

Keywords: geostatistics, indicator kriging, least-square fitting, E-type estimate.

1. INTRODUCTION

In the last few years, one has witnessed the rapid development of software for the geostatistical treatment of environmental data and the integration of layers of estimated values in geographical information systems. One of the main challenges when developing such software is the ability to accommodate the needs of both novice and experienced users. Advanced techniques, such as non-parametric geostatistics (Journel, 1983), might not be used, or worst, be incorrectly applied if the software does not offer the option of performing most of the steps automatically. It is thus critical to evaluate how well the algorithm and its default options would perform in various situations. A particular situation, which would benefit greatly from automatic data processing, is the analysis of data collected by automated monitoring networks. Decision must be made quickly, which prohibits any thorough exploratory data analysis and lengthy processing. The analysis can, however, capitalize on training data collected at prior times for fine-tuning of the algorithm.

My interest in the Spatial Interpolation Comparison (SIC) 2004 exercise was exacerbated by my recent appointment as chief scientist for a software research and development firm. Incorporation of geostatistical functionalities into software targeting more the health scientist than the geostatistician presents new challenges, including the need to develop algorithms for the automatic processing of data. In particular, a key step in any geostatistical study is the computation and modelling of experimental semivariograms. The burden for the user increases with the number of semivariograms, which makes the implementation of indicator kriging with multiple thresholds even more challenging (Goovaerts, 1997).

The main objective of this paper is to present an automatic implementation of indicator kriging and to explore its performance for the two datasets of the SIC 2004 exercise (Dubois and Galmarini, 2005). The analysis is conducted using a Fortran program that integrates Gslib routines for semivariogram computation and indicator kriging (Deutsch and Journel, 1998), with a Fortran code for automatic semivariogram modelling (Pardo-Iguzquiza, 1999). The mean of the local probability distributions estimates the dose rate, while the local uncertainty is quantified using the standard deviation of these distributions. The accuracy of

the spatial prediction and uncertainty model is investigated using the magnitude of prediction errors and the accuracy plots, respectively.

2. METHODOLOGY

Consider the problem of estimating the value of an attribute z at an unsampled location $\mathbf{u}$. The information available consists of a set of n observations $z(\mathbf{u}_\alpha)$ which display some degree of spatial correlation. Most interpolation methods compute this estimate as a linear combination of $n(\mathbf{u}_\alpha)$ neighbouring observations:

$$\hat{z}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u}_\alpha)} \lambda_\alpha z(\mathbf{u}_\alpha) \tag{1}$$

In geostatistics, the unmonitored value $z(\mathbf{u})$ is interpreted as a realization of a random variable $Z(\mathbf{u})$ which is fully characterized by the probability distribution $F(\mathbf{u};z) = \mathrm{Prob}\{Z(\mathbf{u}) \leq z\}$. The weights λ_α account for the data configuration (i.e. clustering of observations), the proximity of data to the unmonitored location $\mathbf{u}$, as well as the spatial pattern modelled from the experimental semivariogram:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} \left[z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h}) \right]^2 \tag{2}$$

where $N(\mathbf{h})$ is the number of data pairs within a given class of distance and direction. The weights are the solution of the following system of $(n(\mathbf{u}_\alpha)+1)$ linear equations, known as ordinary kriging (OK) system:

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta \gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta) - \mu = \gamma(\mathbf{u}_\alpha - \mathbf{u}) \qquad \alpha = 1,...,n(\mathbf{u})$$

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta = 1 \tag{3}$$

where μ is a Lagrange multiplier accounting for the constraint on the weights.

2.1 *Indicator kriging*

Extreme values can strongly affect the characterization of spatial patterns, and subsequently the prediction. Several approaches exist to handle strongly positively skewed histograms (Saito and Goovaerts, 2000). One common approach is to first transform the data (e.g. normal score, Box Cox or lognormal transform), perform the analysis in the transformed space, and back-transform the resulting estimates. Such transform, however, does not solve problems created by the presence of numerous censored data (i.e. observations below the detection limit). Moreover, except for the normal score transform (Deutsch and Journel, 1998), it does not guarantee the normality of the transformed histogram, which is required to compute confidence intervals for the estimates.

Another way to attenuate the impact of extreme values is to use more robust statistics and estimators. Indicator kriging (IK) is based on a preliminary coding of each observation $z(\mathbf{u}_\alpha)$ into a vector of indicators, defined for a set of K thresholds z_k discretizing the range of variation of the attribute:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) \leq z_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1,...,K \tag{4}$$

Both predictor (1) and semivariogram estimator (2) are then expressed in terms of the indicator data $i(\mathbf{u}_\alpha; z_k)$ instead of the original attribute values $z(\mathbf{u}_\alpha)$:

$$\hat{i}(\mathbf{u}; z_k) = \sum_{\alpha=1}^{n(\mathbf{u}_\alpha)} \lambda_\alpha(z_k) i(\mathbf{u}_\alpha; z_k) \tag{5}$$

$$\hat{\gamma}_I(\mathbf{h}; z_k) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} \left[i(\mathbf{u}_\alpha; z_k) - i(\mathbf{u}_\alpha + \mathbf{h}; z_k) \right]^2 \qquad (6)$$

For each threshold z_k, the model fitted to the corresponding indicator semivariogram $\hat{\gamma}_I(\mathbf{h}; z_k)$ is used to solve a kriging system of type (3) and obtain the kriging weights for the estimator (5). Although the indicator transform causes a loss of information regarding the exact z-value, the influence of extreme values on semivariogram estimation is attenuated since they are assigned a value of 0 or 1 regardless their magnitude.

Indicator kriging does not provide a direct estimate of the unmonitored attribute value, rather it yields a set of K probability estimates:

$$\begin{aligned} \hat{i}(\mathbf{u}; z_k) &= F_{IK}(\mathbf{u}; z_k \mid (n)) \\ &= \text{Prob}\{Z(\mathbf{u}) \le z_k \mid (n)\} \quad k = 1, ..., K \end{aligned} \qquad (7)$$

These probabilities form a discrete probability distribution, which is referred to as conditional cumulative distribution function (ccdf). Because the K probabilities are estimated individually (i.e. K indicator kriging systems are solved at each location) the following constraints, which are implicit to any probability distribution, might not be satisfied by all sets of K estimates:

$$0 \le F_{IK}(\mathbf{u}; z_k \mid (n)) \le 1 \quad \forall k \qquad (8)$$

$$0 \le F_{IK}(\mathbf{u}; z_{k'} \mid (n)) \le F_{IK}(\mathbf{u}; z_k \mid (n)) \quad \text{if } z_{k'} \le z_k \qquad (9)$$

All probabilities that are not between 0 and 1 are first reset to the closest bound, 0 or 1. Then, condition (9) is ensured by averaging the results of an upward and downward correction of ccdf values (Deutsch and Journel, 1998). Once conditional probabilities were estimated and corrected for potential order relation deviations, the set of K probabilities must be interpolated within each class (z_k, z_{k+1}] and extrapolated beyond the smallest and the largest thresholds to build a continuous model for the conditional cdf. In this paper, the resolution of the discrete ccdf was increased by performing a linear interpolation between tabulated bounds provided by the sample histogram of dose rate data (Deutsch and Journel, 1998).

2.2 *Post-processing of uncertainty models*

Knowledge of the probability distribution at any unmonitored location allows one to: (1) compute several quantities, such as the probability of exceeding any critical threshold for the attribute z or the expected value for this attribute, and (2) generate a set of simulated z-values (i.e. Monte-Carlo simulation). In this paper, the main objective was to estimate the unmonitored value and assess the uncertainty attached to that prediction. Common estimators of the z-value are the mean or the median of the ccdf, while the uncertainty is measured by the spread of the ccdf. Here the mean or E-type estimate was selected to predict gamma dose rates, and the uncertainty was quantified using the standard deviation of the ccdf.

3. IMPLEMENTATION AND TESTING OF THE APPROACH

The different steps of indicator kriging were implemented in a single code that proceeds as follows:

- Sample data are imported, and the K thresholds are computed as equally spaced p-quantiles of the sample histogram.
- The K indicator semivariograms are estimated and a model is fitted using weighted least-square regression.
- Indicator kriging is performed for each threshold, and the set of probabilities is corrected for order relation deviations.
- The discrete ccdf is completed by interpolation/extrapolation, and both E-type estimate and ccdf standard deviation are computed at each location.
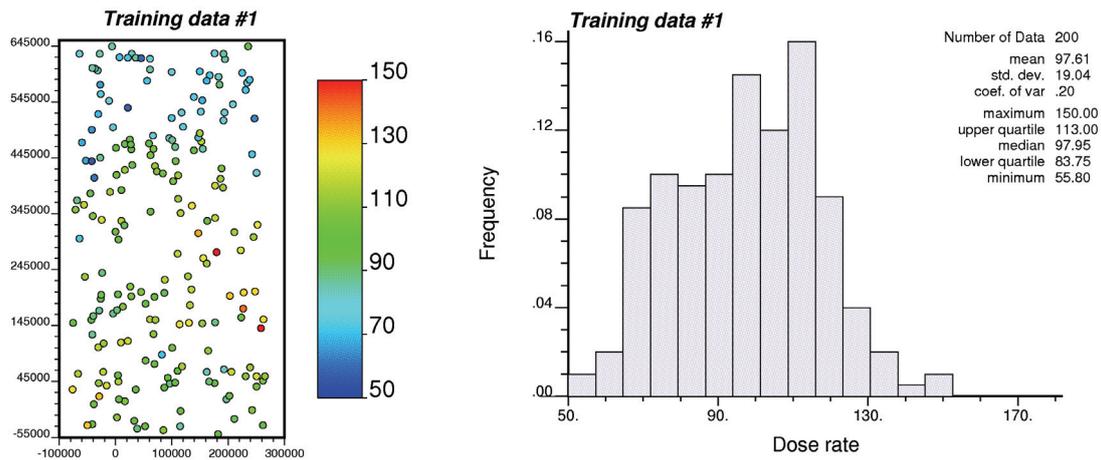
Figure 1. *Location map and histogram for the first training dataset. Measurement units are nSv/h for dose rates and m for distances.*

### 3.1 *Estimation and modelling of indicator semivariograms*

The SIC 2004 participants received a training dataset consisting of 10 days of daily measurements at 200 monitoring locations which were selected randomly from a larger network of 1008 stations. These data are described in details in Dubois and Galmarini (2005). Figure 1 shows the spatial distribution and sample histogram for the first day of measurements. These data display a NS trend with smaller rates measured in the northern part of the study area. All 10 data sets are strongly correlated with each other, which facilitated the choice of a common set of parameters. The analysis of these data was particularly useful to fine-tune the estimation of indicator semivariograms.

Nineteen thresholds were selected in order to reach a balance between loss of information caused by too few thresholds and increasing CPU time as more thresholds are considered. The sample size was large enough to yield non-erratic semivariograms for the extreme 0.05 and 0.95-quantile thresholds, see Figure 2. Analysis of the training data did not reveal any direction-dependent variability within the range of distances used for interpolation, hence only omnidirectional semivariograms were computed. Twenty-five lags of 10 km each were found to provide enough data pairs within each lag class for reliable estimates of indicator semivariograms and enough resolution in terms of spatial lags.

Sample indicator semivariograms were modelled using the weighted least-square algorithm developed by Pardo-Iguzquiza (1999). The weight for a lag **h** was computed as $\sqrt{N(\mathbf{h})}\big/\gamma^2(\mathbf{h})$ to privilege small semivariogram values computed from many data pairs. For each threshold, a nested model including a nugget effect and two basic structures (S=spherical or E=exponential model) was fitted. Three combinations of models (S/S, E/E, and S/E) were tried and the one that minimizes the sum of squares of differences between the experimental and model semivariogram values was selected. Figure 2 shows that for the first thresholds the semivariograms display a very small nugget effect and a short-range structure (less than 50 km), while larger ranges and nugget effects occur for higher thresholds. This reflects the existence of clusters of low dose rates with better spatial connectivity over short distances

3.2 *Cross-validation*

The indicator kriging code was run on each of the 10 training datasets. Each observation was removed at a time from the data set and the ccdf at this location was modelled using the remaining 199 data (cross-validation approach). The mean of the ccdf was compared to the measured value, and the quality of the prediction was assessed using the following three statistics:
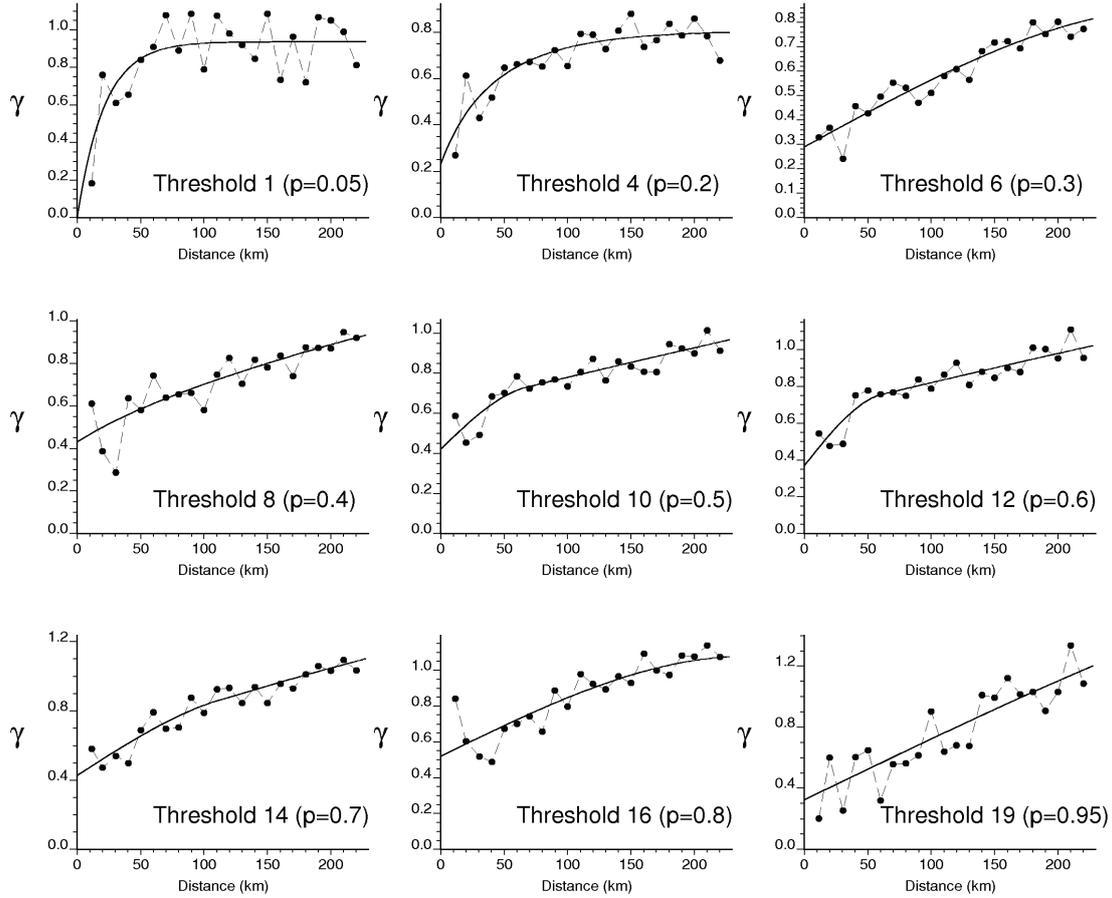
Figure 2. *Omnidirectional indicator semivariograms computed at nine thresholds corresponding to p-quantiles of the histogram for the first training dataset. These semivariograms were rescaled by the variance of the indicator variable. The solid line depicts the isotropic model fitted using weighted least-square regression.*

$$MAE = \frac{1}{n}\sum_{\alpha=1}^{n}\left|\hat{z}(\mathbf{u}_{\alpha}) - z(\mathbf{u}_{\alpha})\right|, \qquad ME = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\hat{z}(\mathbf{u}_{\alpha}) - z(\mathbf{u}_{\alpha})\right),$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{\alpha=1}^{n}(\hat{z}(\mathbf{u}_{\alpha}) - z(\mathbf{u}_{\alpha}))^{2}},$$

The mean absolute error (MAE), the bias (or mean error ME), and the root mean squared error (RMSE) of the predictions were supplemented by the Pearson's *r* coefficient of correlation between measured and estimated dose rates. The relationship between the magnitude of the prediction error and the spread of the ccdf was quantified using the correlation coefficient between absolute errors and ccdf standard deviation (SD).

Cross-validation was used primarily to detect any systematic bias in the predictions, and no attempt was made to minimize the prediction errors by trying different combinations of kriging parameters. The prediction was conducted using the 16 closest observations (no maximum search radius was used). Using more neighbouring data slightly decreases the prediction error, but at the expense of significantly larger CPU time. The resolution of the discrete ccdf was increased by performing a linear interpolation between tabulated bounds provided by the sample histogram of dose rates. The minimum and maximum of the probability distributions were systematically identified with the smallest and largest observations to avoid any arbitrary selection of the range of possible values.

Table 1 shows that the prediction errors are of same magnitude for all training datasets and that the bias never exceeds 0.5%. The correlation between measured and estimated values is reasonably good, and this relationship is illustrated for the first training dataset at the top of Figure 3. This scatterplot shows that the average E-type estimate is close to the sample mean (unbiasedness) while the standard deviation is much smaller than the sample one (smoothing effect). The correlation between absolute prediction errors and ccdf standard deviations (SD) is not as good, see Table 1 (7th column). Figure 3 (right top graph) shows that the standard deviation tends to be smaller in areas where dose rates are uniformly low or high, while high SD values are observed for sampled locations that are isolated and/or surrounded by a large range of dose values. However, locations of higher local variability are not systematically the most difficult to predict, which explains the small magnitude of the correlation coefficient.

At any location $\boldsymbol{u}$ a series of symmetric p-probability intervals (PI) bounded by the $(1-p)/2$ and $(1+p)/2$ quantiles can be computed from the ccdf $F(\mathbf{u};z|(n))$. For example, the 0.5-PI is bounded by the lower and upper quartiles $[F^{-1}(\mathbf{u};0.25|(n)), F^{-1}(\mathbf{u};0.75|(n))]$. A correct modeling of local uncertainty would entail that there is a 0.5 probability that the actual z-value at $\boldsymbol{u}$ falls into that interval or, equivalently, that over the study area 50% of the 0.5-PI include the true value. Cross-validation yields a set of z-measurements and independently derived ccdfs at the n locations $\mathbf{u}_\alpha$, allowing the fraction of true values falling into the symmetric p-PI to be computed as:

$$\overline{\zeta}(p) = \frac{1}{n}\sum_{\alpha=1}^{n} \zeta(\mathbf{u}_\alpha; p) \tag{10}$$

where $\zeta(\mathbf{u}_\alpha; p)$ equals 1 if $z(\boldsymbol{u}_\alpha)$ lies between the $(1-\underline{p})/2$ and $(1+p)/2$ quantiles of the ccdf, and zero otherwise. The scattergram of the empirical, $\overline{\zeta}(p)$, versus expected, p, fractions is called the "accuracy plot"; see example in Figure 3. Deutsch (1997) proposed to assess the closeness of the empirical and theoretical fractions using the following "goodness" statistic:

$$G = 1 - \frac{1}{K}\sum_{k=1}^{K} w(p_k)\,|\,\overline{\zeta}(p_k) - p_k\,| \tag{11}$$

where $\underline{w}(p_k)$ =1 if $\overline{\zeta}(p_k)>p_k$, and 2 otherwise. Twice more importance is given to deviations when $\overline{\zeta}(p_k)< p_k$ (inaccurate case), i.e. the case where the fraction of true values falling into the p-PI is smaller than expected. Goodness is consistently high among training datasets and the accuracy plot indicates the model becomes inaccurate as the interval gets wider.

| Training sets	Mean	MAE	ME	RMSE	Correlation True vs est	Correlation Error vs SD	Goodness
1	97.61	8.96	-0.56	11.74	0.789	0.252	0.951
2	97.42	8.96	-0.65	11.72	0.795	0.291	0.936
3	98.76	8.34	-0.46	10.98	0.803	0.265	0.940
4	93.82	8.14	-0.29	10.81	0.765	0.271	0.949
5	92.41	7.90	-0.07	10.53	0.771	0.280	0.952
6	89.81	7.92	-0.30	10.29	0.761	0.284	0.949
7	91.73	8.16	-0.39	10.81	0.744	0.304	0.942
8	92.44	7.94	-0.48	10.58	0.769	0.297	0.947
9	96.62	8.51	-0.43	11.20	0.789	0.268	0.934
10	95.35	8.43	-0.48	11.21	0.758	0.291	0.939

Table 6. *Prediction statistics obtained by cross-validation for the training datasets (nSv/h).*
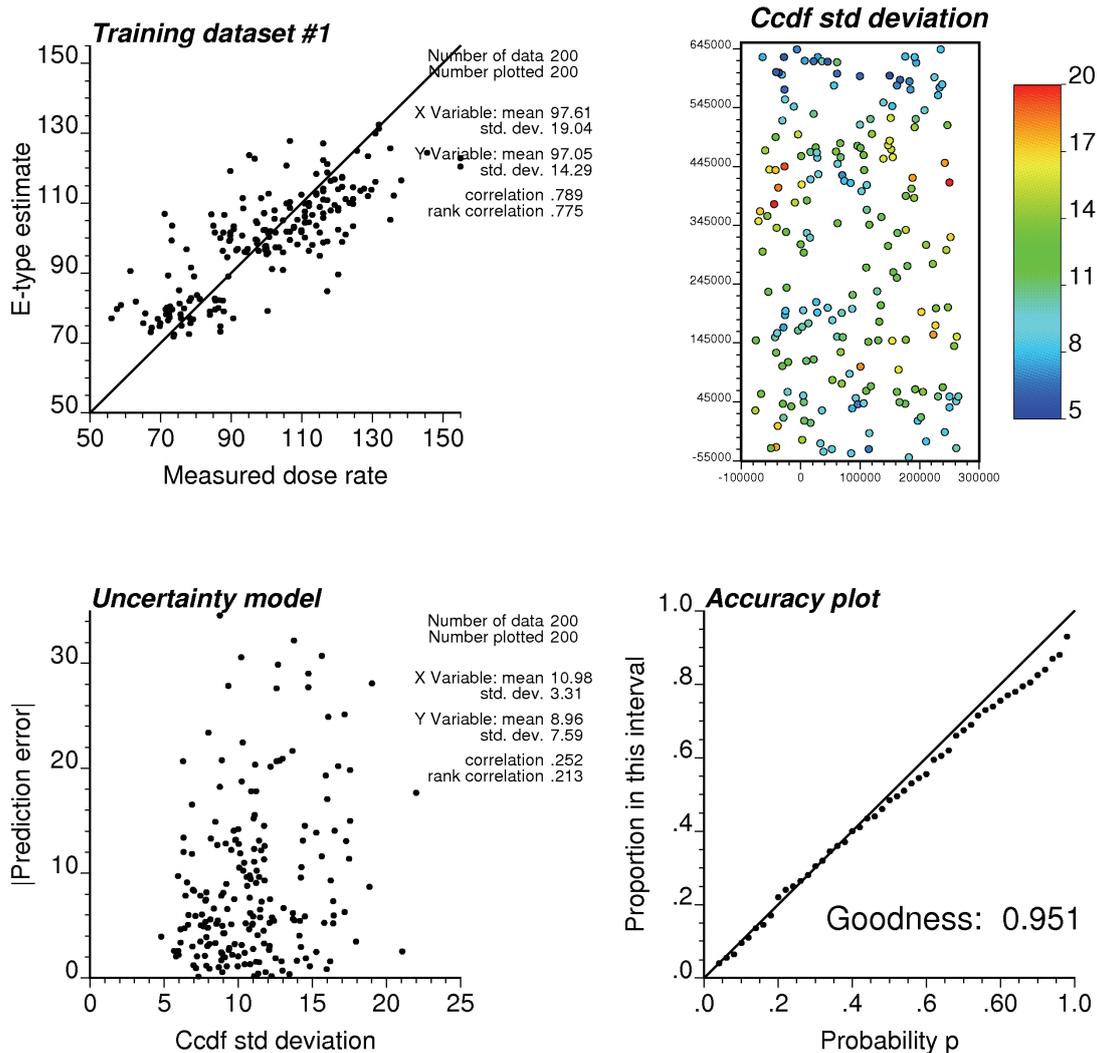
Figure 3. *Cross-validation results for the first training dataset, illustrating the relationship between estimated and measured dose rates, as well as the poor correlation between the magnitude of estimation errors and the standard deviation of the ccdf. The accuracy plot shows the good correspondence between expected and empirical proportions of data falling within probability intervals of increasing size.*

4. RESULTS

The procedure described in Section 3 was applied to two datasets of 200 measurements with the same spatial coordinates as the training datasets. The estimation was conducted at 808 other locations that were not provided to SIC 2004 participants at the time of the exercise; see Figure 4. For each dataset an entire run took 5.62 seconds on a Pentium 3.20 GHz. Table 2 gives the summary statistics for the two datasets which are described in more details in Dubois and Galmarini (2005). The first dataset is very similar to the training data in terms of distribution of values; compare statistics of Table 2 with sample statistics in Figure 1. The second dataset, also known as "joker" dataset, is characterized by the presence of anomalous values or outliers, which are mimicking the results of the emission of an undefined radioactive substance in the South West part of the study area. The maximum for the 2nd dataset is one order of magnitude larger than for the 1st data, which has a profound impact on the sample standard deviation.
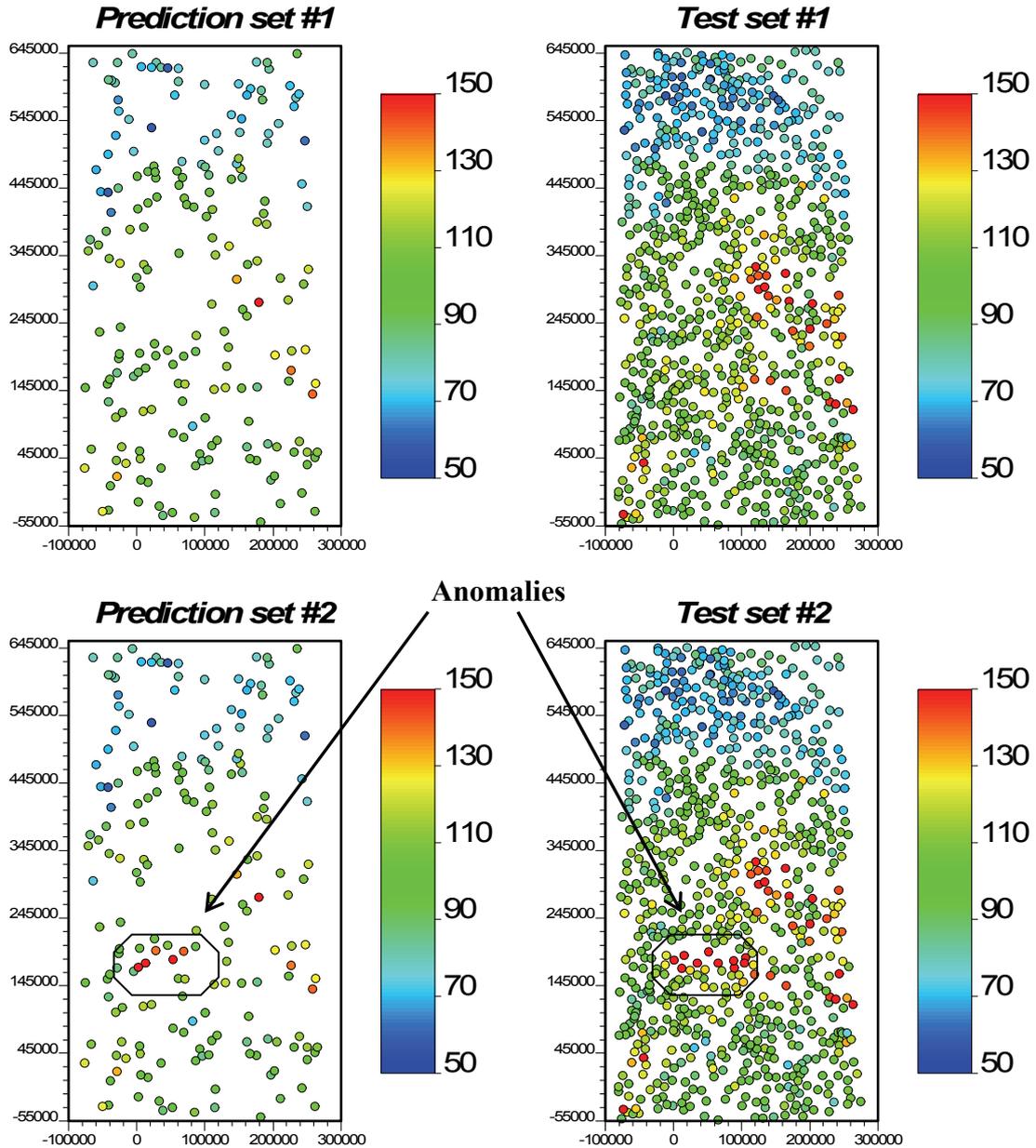
Figure 4. *Location maps of prediction (200 dose rates) and validation/test (808 dose rates) sets for the first and second ("joker") datasets. Measurement units are nSv/h for dose rates. Note the location of anomalies in the second dataset.*

N = 808	Min.	Max.	mean	median	std. dev.
Observed (first data set)	57.0	180.0	98.0	98.8	20.0
Estimates (first data set)	69.9	125.4	96.4	98.7	13.84
Observed (second data set)	57.0	1528.2	105.4	99.0	83.7
Estimates (second data set)	69.9	262.3	109.4	103.0	31.3

Table 2. *Comparison of statistics for the measured and estimated dose rates (nSv/h).*
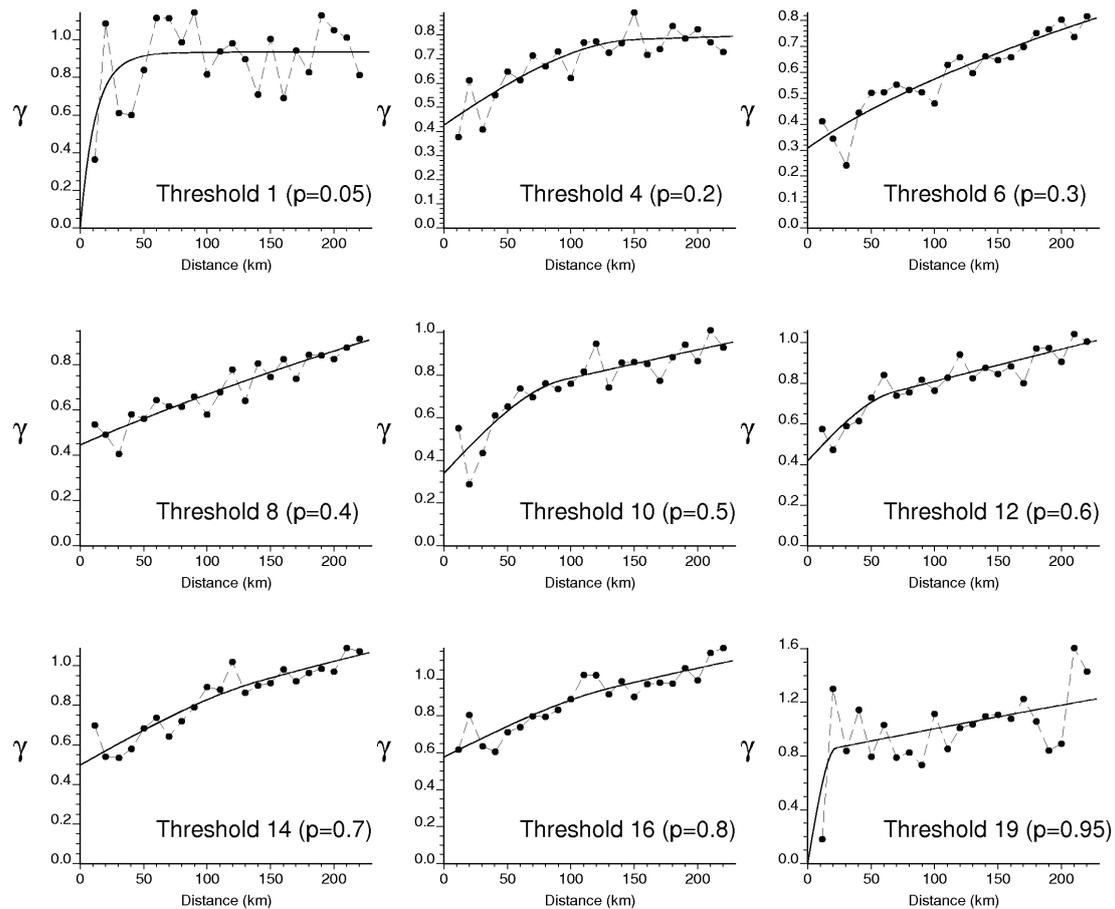
Figure 5. *Omnidirectional indicator semivariograms computed at nine thresholds corresponding to p-quantiles of the histogram for the second ("joker") dataset. These semivariograms were rescaled by the variance of the indicator variable. The solid line depicts the isotropic model fitted using weighted least-square regression.*

4.1. *Prediction errors*

The first step was the computation and automatic modelling of nineteen indicator semivariograms. The semivariograms for the first dataset are very similar to the ones computed for the training dataset and are not displayed here; recall Figure 2. Figure 5 shows the indicator semivariograms for the second dataset. Note that the semivariogram for the 19th threshold has a much shorter range than for the training datasets, which is caused by the anomalous cluster of extremely high values delineated in Figure 4 (left bottom graph). The E-type estimate was estimated at the nodes of a 2 km grid. Figure 6 (top graphs) illustrate the impact of the anomalies on the spatial pattern of the map of dose rates, in particular the higher estimates found in the bottom left corner of the map.

The prediction performance of the indicator approach was assessed by modelling the ccdf at the 808 test locations displayed in Figure 4 (right column). The E-type estimate was derived at each location and plotted against the recorded dose rate; see Figure 6 (bottom graphs). Tables 2 and 3 show that for both datasets the estimator is relatively unbiased and, as expected, smoother than the observed attribute. For the first dataset the correlation between measured and estimated values (r=0.79) is similar to the one found for the training dataset. The performance is much less satisfactory for the second dataset (r=0.40) because the smoothing effect induces a strong underestimation of the seven anomalous values. The rank correlation coefficient, which is more robust with respect to extreme values, is however as good as for the first dataset.
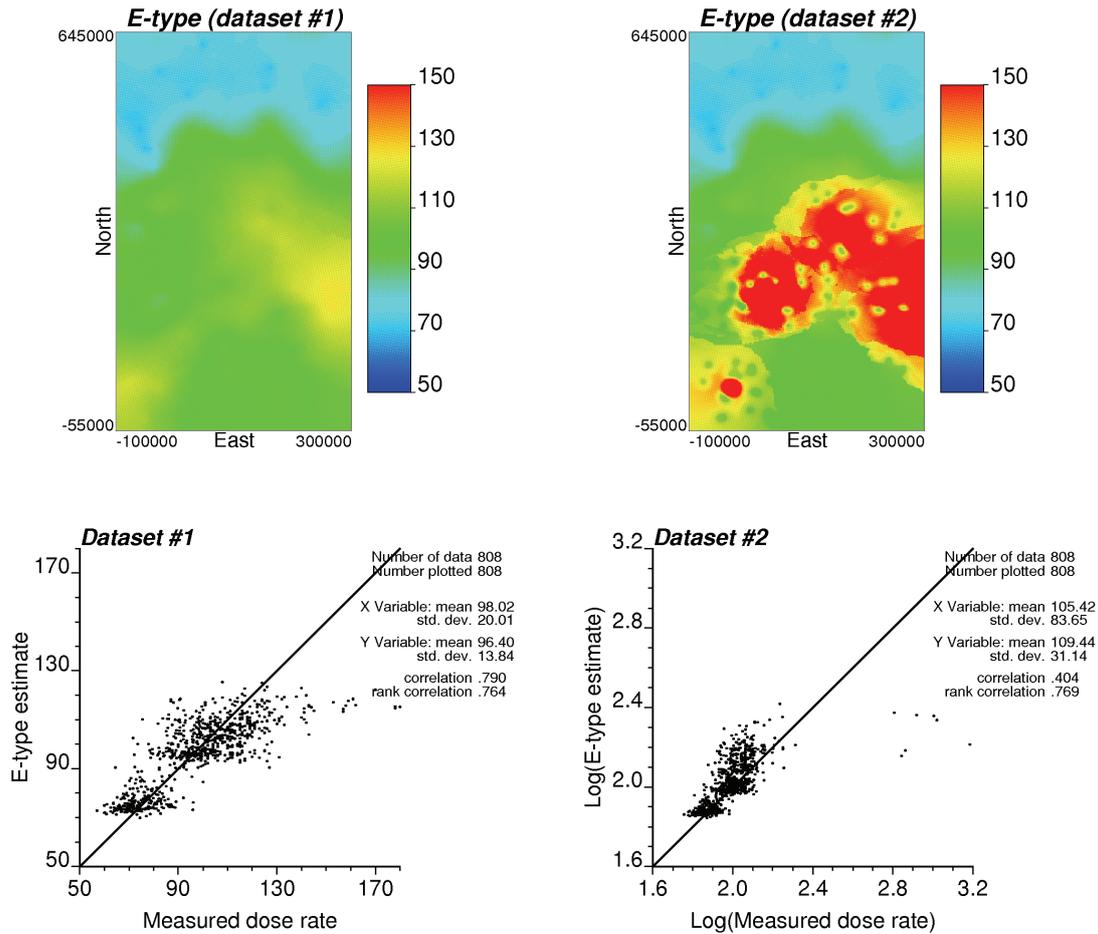
Figure 6. *Maps of E-type estimate for each dataset. Bottom graphs show the scatterplots between estimated and measured dose rates at 808 test locations (units=nSv/h).*

Data sets:	MAE	ME	Pearson's r	RMSE
First data set	9.24	-1.61	0.789	12.54
Second data set	21.85	4.02	0.404	76.68

Table 3. *Prediction statistics for the two validation datasets (nSv/h).*

.

4.2. *Measures of uncertainty*

Indicator kriging provides at each unmonitored location a non-parametric model of uncertainty, which takes the form of a probability distribution for the dose rate. The standard deviation of the ccdf is mapped at the top of Figure 7. Ccdfs tend to be wider in the high-valued areas (proportional effect) as well as where small and high dose rates are intermingled. The two large zones of lower standard deviation, hence smaller uncertainty, correspond to the homogeneous zones of low dose rates in the North and of intermediate rates in the South. There is not a very good correspondence between the spread of the ccdf and the magnitude of prediction errors at the 808 test locations for the first dataset. For the "joker" dataset there is a good correlation between the ranks of prediction errors and ccdf standard deviation. This is somewhat expected since prediction errors are larger for extremely large dose rates which tend to be located in high-valued areas where the proportional effect entails larger variance of the local probability distributions.
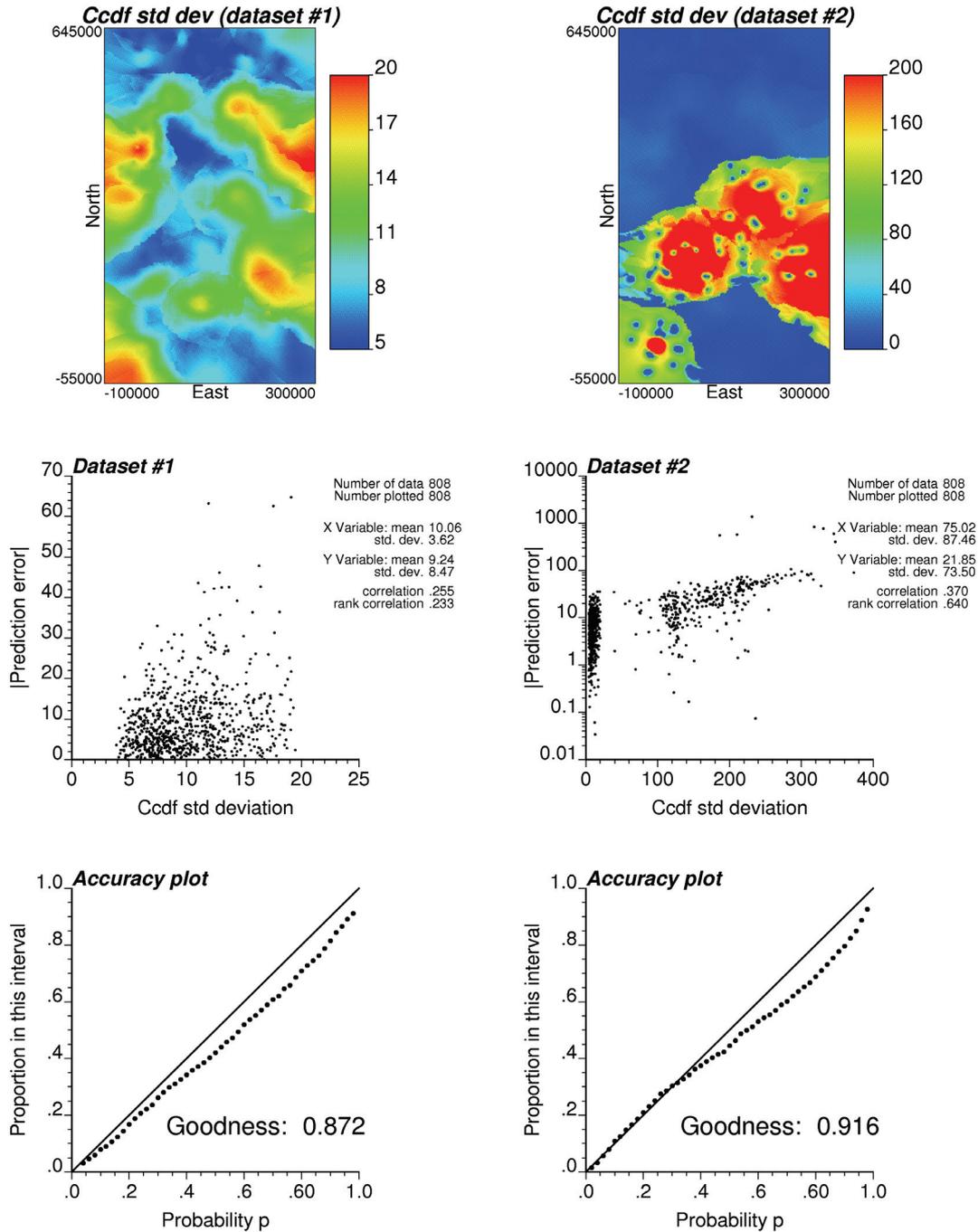
Figure 7. *Maps of the standard deviation of the ccdf for each dataset. Middle graphs show the scatterplots between the magnitude of estimation errors and the standard deviation of the ccdf (units=nSv/h). The accuracy plot shows the good correspondence between expected and empirical proportions of data falling within probability intervals of increasing size.*

The accuracy plots at the bottom of Figure 7 exhibit the same behaviour as the one observed for the training datasets. As the probability interval widens, it includes a proportion of data that is smaller than expected (inaccurate case). Yet, the model provides good results for narrower intervals, in particular for the second dataset.
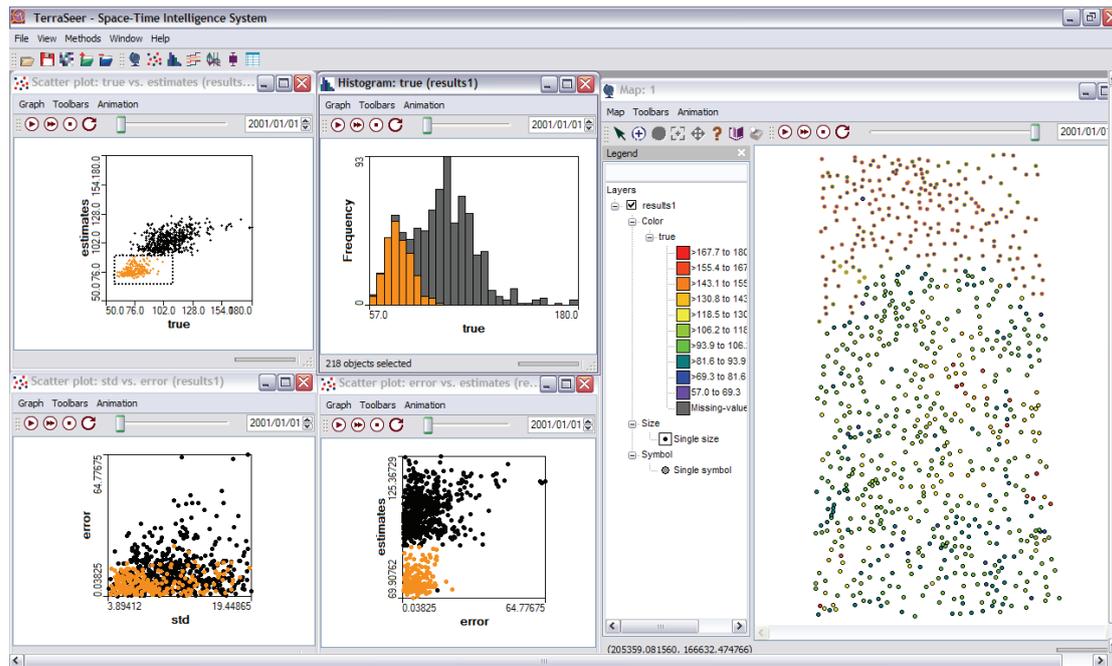
Figure 8. *Location map of the dose rates recorded at the 808 test locations for dataset #1. Using dynamically linked windows, the subset of low estimated rates is identified in all the scattergrams (orange highlights) and it is shown to correspond to the low-valued zone in the Northern part of the study area.*

5. CONCLUSIONS

This study demonstrated that nowadays indicator kriging with multiple thresholds is accessible and computationally tractable thanks to the development of automatic procedures for semivariogram fitting and the growing processing speed. In particular, the user should no longer feel restricted in using a maximum of nine thresholds for indicator coding, as typically done in many analyses. Using more thresholds attenuates the loss of information caused by the coding of continuous attributes into a set of binary indicators, which is a common criticism of the technique. Indicator kriging seems a better alternative to data transforms, since: (1) the transform procedure might be hard to automate (except for normal score transform, there is no guaranty that the transformed distribution will meet the requirements of the analysis) and (2) one-to-one transforms, such as normal score transform, require an arbitrary untie of censored data.

Indicator kriging performs well for the first dataset, although no clear relationship was found between the spread of the local probability distribution and the magnitude of the prediction errors. The "joker" dataset represents a very specific situation where the underlying assumption of stationarity no longer applies. The use of indicator semivariograms allows one to model the spatial connectivity of different classes of dose rates, which is useful for contamination events that typically operate at a much smaller scale than the physical processes responsible for the spatial background over the region. These extreme values have a strong influence on global statistics, which indirectly affects the interpolation/extrapolation of discrete ccdf values using the procedure implemented in this paper (i.e. linear interpolation between tabulated bounds provided by the sample histogram of dose rates). In this paper, this results in the large zone of high predicted rates displayed on the maps of E-type estimate.

The scattergrams of measured versus predicted dose rates at the bottom of Figure 6 clearly indicate the existence of two subsets of values. Using a statistical software with

dynamically linked windows (TerraSeer, 2005), the lower subset is identified as the zone of small dose rates recorded in the Northern part of the study area, see right-hand-side window of the screen shot in Figure 8. This zone is even more apparent when looking at the map of E-type estimates in Figure 6. One could then question the decision of pooling together the two subsets for the geostatistical analysis. An alternative would be to stratify the study area into two different populations, which then would be analyzed and modelled separately. The use of within-stratum histograms for ccdf interpolation would certainly reduce the overestimation of low dose rates and the underestimation of high rates. The stratification of the area goes beyond the split of the sample dataset, since any unmonitored location needs to be *a priori* assigned to a given population. This operation requires further information regarding the physics of the data, and it is certainly worth investigation.

Acknowledgments

References

1. Deutsch, C.V. (1997). Direct assessment of local accuracy and precision. In *"Geostatistics Wollongong '96"*. Baafi, E.Y., Schofield, N.A. (Eds.), Kluwer Academic Publishers, pp. 115-125.

2. Deutsch, C.V., Journel, A.G. (1998). *GSLIB: Geostatistical Software Library and User's Guide: 2nd edn*. Oxford Univ. Press, 369 p.

3. Dubois, G., Galmarini, S. (2005). "*Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results"*. This Volume.

4. Goovaerts, P. (1997)*. Geostatistics for Natural Resources Evaluation*. Oxford University Press, 483 p.

5. Journel, A. G. (1983). Non-parametric estimation of spatial distributions. *Mathematical Geology*, 15(3): 445-468.

6. Pardo-Iguzquiza, E. (1999). VARFIT: a Fortran-77 program for fitting variogram models by weighted least squares. *Computers and Geosciences,* 25: 251-261.

7. Saito, H., Goovaerts, P. (2000). Geostatistical interpolation of positively skewed and censored data in a dioxin contaminated site*. Environmental Science & Technology*, 34(19): 4228-4235.

8. TerraSeer (2005). Space Time Intelligence System. Evaluation version available at http://www.terraseer.com/products/stis/stis_demo.html

Identification of Spatial Anisotropy by means of the Covariance Tensor Identity

Dionissios T. Hristopulos

Department of Mineral Resources Engineering
Technical University of Crete
Chania 73100 GREECE

E-mail: dionisi@mred.tuc.gr

Abstract: This paper focuses on the identification of geometric anisotropy by means of the covariance tensor identity (CTI) method, which is based on established relations between the anisotropic parameters and the covariance tensor of the gradient field. The method is applied to the mean daily values of gamma dose rates measured in Germany, which were used in the Spatial Interpolation Contest (SIC) 2004 (Dubois and Galmarini, 2005). The CTI method avoids multiple variogram calculations in arbitrary directions and gives numerical values for the anisotropy parameters. If the latter are known, the coordinate system can be transformed to render the spatial dependence statistically isotropic, thus allowing the application of non-directional spatial interpolation models. It is also shown that the CTI method enables the detection of incidents leading to a systematic change of the anisotropy, such as the simulated release of radioactivity in the SIC 2004 "joker" data set.

Keywords: automatic mapping, radioactivity monitoring, Spartan random fields.

1. INTRODUCTION

Spatial data often exhibit continuity properties that depend on the direction in space. In the geostatistical framework, estimates of anisotropy are usually based on empirical methods, such as visual inspection of geological maps, estimation of the experimental variogram in the North-South and East-West directions, or in multiple but still arbitrarily defined directions (e.g., Goovaerts, 1997, p.98, p.104; Wackernagel, 2003, p.62-64). Two types of anisotropy are usually considered in geostatistical studies: *Geometric anisotropy* (or range anisotropy) implies that the iso-level variogram contours are ellipses in two dimensions (Wackernagel, 2003; p. 62) and ellipsoids in three dimensions. In the principal axes system, the main axes of the iso-level variogram contours are aligned with the coordinate system. *Zonal anisotropy* implies a different sill value in different directions. The focus of this paper is on geometric anisotropy. Determining the parameters of anisotropy allows a coordinate transformation so that the spatial distribution is statistically isotropic in the new system (i.e., the iso-level variogram contours become circles). The automatic identification of anisotropy parameters, especially if it enables detection of sudden changes in the spatial distribution, is an important aspect of an automated mapping framework.

The *Covariance Tensor Identity* (CTI) method employs a relation between the ensemble average of the *gradient tensor* and the respective derivatives *of the covariance function* (Swerling, 1962), which depend on the geometric anisotropy parameters. The mean gradient tensor is estimated from the spatial (sample) average. The anisotropic parameters are thus obtained by matching directional properties of spatial moments (estimated from the data) with the respective properties of the covariance function. The CTI method was introduced in (Hristopulos, 2002) and tested with synthetic data on square lattices and irregular supports. It was recently applied to lignite quality data from a mining site in the Western Macedonia province of Greece (Hristopulos, 2004; Galetakis and Hristopulos, 2004). To the author's

knowledge, the only other proposal for the systematic estimation of geometric anisotropies is the Bayesian updating method of Ecker and Gelfand (1999).

2. METHODOLOGY

Let us denote the *spatial random field* (SRF) representing gamma dose rate measurements by $X(\mathbf{s}_i)$, $i = 1, \ldots N$, where $\mathbf{s} = (x, y)$, and by $c_X(\mathbf{r})$ its covariance function. The anisotropic parameters in two dimensions are the *rotation angle* θ and the *anisotropic ratio R* (which is equal to the ratio of the semi-axes of the ellipses). The rotation angle determines the orientation of the principal axes with respect to the initial coordinate system, and the anisotropic ratio is given by $R = \xi_x / \xi_y$, where ξ_x, ξ_y are the correlation lengths of $X(\mathbf{s})$ along the two principal axes.

2.1 *The isotropy transformation*

If the anisotropy parameters are known, the coordinate system can be aligned with the principal axes after rotation by θ. Position vectors in the two systems are related by $\mathbf{s}' = \mathbf{U}(\theta)\mathbf{s}$, where $\mathbf{U}(\theta)$ is the 2D rotation matrix. A subsequent *rescaling* of the axes leads to an *isotropic principal coordinate* (IPC) system, in which the iso-level variogram contours become circles. The coordinates in the rotated and rescaled IPC system are given by the equations

$$
\begin{aligned}
\tilde{x} &= \quad x\cos\theta + y\sin\theta \\
\tilde{y} &= \quad R(-x\sin\theta + y\cos\theta)
\end{aligned}
\tag{1}
$$

In the IPC system there is no distinction between directions, and isotropic prediction methods can be used.

2.2 *The Covariance Tensor Identity*

The CTI is strictly valid for *Gaussian, second-order stationary and differentiable* spatial random fields (Hristopulos, 2002). For Gaussian SRF's, differentiability is primarily controlled by the existence of the second-order covariance function derivatives (e.g., Yaglom, 1987; Abrahamsen, 1997). Assuming differentiability raises a number of questions: First, most classical covariance models used in geostatistical applications correspond to non-differentiable SRF's. However, this is not a major obstacle, since differentiable models also exist, such as the family of Matérn covariance functions and the covariance functions of *Spartan Spatial Random Fields* (SSRF's) (Hristopulos, 2003). Second, the data may contain an uncorrelated (nugget) component, which is not differentiable. Both the SIC2004 data set and the mining data mentioned above, contain such a component. As shown below, the CTI method provides "reasonable" estimates even in such cases. If the nugget component is large, the accuracy and precision of the estimated anisotropy parameters is expected to decline. Quantitative estimates of the nugget's impact on the estimates of anisotropy need to be developed. The CTI method is applicable to both normal and lognormal SRF's (e.g., Hristopulos, 2004). Finally, for practical application of the CTI, trends are removed to obtain a zero-mean fluctuation SRF.

If the CTI conditions hold, the zero-lag *covariance tensor* is defined by:

$$
C_{X(ij)} \equiv \lim_{\mathbf{r}\to 0}\left[\partial_{ij}^2 c_X(\mathbf{r})\right], \ \forall i, j = \text{x,y},
\tag{2}
$$

where $\partial_{ij}^2[\cdot]$ denotes the second-order partial derivative. If $\partial_x X(\mathbf{s}) = \partial X(\mathbf{s})/\partial x$ and $\partial_y X(\mathbf{s}) = \partial X(\mathbf{s})/\partial y$ are the partial derivatives of the SRF, the *gradient tensor* is defined by:

$$X_{ij}(\mathbf{s}) \equiv \partial_i X(\mathbf{s}) \partial_j X(\mathbf{s}) , \tag{3}$$

and the *mean gradient tensor* is given by $Q_{ij} \equiv E[X_{ij}(\mathbf{s})]$, where $E[.]$ denotes the *ensemble average*. The *covariance tensor identity* relates the mean gradient tensor to the covariance tensor as follows (Swerling, 1962):

$$Q_{ij} = -C_{X(ij)} . \tag{4}$$

2.3 *The CTI System of Equations*

As shown in (Hristopulos, 2002), the CTI leads to the following system of equations between the elements of the gradient tensor and the anisotropic parameters:

$$\begin{aligned}
Q_{xx} &= \zeta_X \, \xi_x^{-2} \left(\cos^2 \theta + R^2 \sin^2 \theta \right) \\
Q_{yy} &= \zeta_X \, \xi_x^{-2} \left(R^2 \cos^2 \theta + \sin^2 \theta \right) , \\
Q_{xy} &= \zeta_X \, \xi_x^{-2} \left[\sin \theta \cos \theta (1 - R^2) \right]
\end{aligned} \tag{5}$$

where $\zeta_X = -c_X''(h)\big|_{h=0}$ is evaluated in the IPC system with respect to the dimensionless lag $\mathbf{h}$. Since the covariance function has its maximum at $h = 0$, it follows that $\zeta_X > 0$.

Eqs. (5) do not involve an equation for Q_{yx}, because the symmetry of mean gradient tensor Q_{ij} implies that only three independent equations exist. If $\zeta_X \neq 0$ and the principal correlation lengths are finite, Q_{xx} and Q_{yy} are positive. On the other hand, $Q_{xy} = Q_{yx} = 0$ if $R = 1$. If $R \in (0, \infty)$, the rotation angle can be restricted in the interval $\theta \in [-\pi/4, \pi/4]$, since any pair (R, θ), where $\theta \in [\pi/4, \pi/2]$ can be mapped onto a pair (R', θ'), where $\theta' = -(\pi/2 - \theta) \in [-\pi/4, 0]$ and $R' = 1/R$. Based on the third of Eqs. (5), this restriction implies that $Q_{xy} < 0$ is equivalent to $R > 1$.

The parameters ξ_x and ζ_X, which are not related to anisotropy, are eliminated by forming the *scaled gradient tensor moments* Z_{diag} and Z_{off}:

$$Z_{\text{diag}} \equiv \frac{Q_{yy}}{Q_{xx}} = \frac{R^2 \cos^2 \theta + \sin^2 \theta}{\cos^2 \theta + R^2 \sin^2 \theta} , \quad Z_{\text{off}} \equiv \frac{Q_{xy}}{Q_{xx}} = \frac{\sin \theta \cos \theta (1 - R^2)}{\cos^2 \theta + R^2 \sin^2 \theta} . \tag{6}$$

The equations (6) constitute the *CTI system*, which involves the unknown anisotropic parameters and the scaled moments that can be estimated from the data.

2.4 *Estimating the Mean Gradient Tensor*

For irregularly-spaced distributions, the mean gradient tensor is estimated from the sample average of finite-difference gradient approximations. The sample average may introduce *ergodic errors*, while finite differences introduce *discretisation errors*. Let us denote the nearest neighbor of the point $\mathbf{s}_k$ by $\mathbf{s}_k^*$, and the finite difference step the *i*th direction by $\delta s_{k(i)} = s_{k(i)}^* - s_{k(i)}$. Let $\delta s(q)$ denote the qth percentile of the *nearest neighbor distance* distribution; an estimate of the mean gradient tensor is given by

$$\hat{Q}_{ij}(q_1, q_2) = \frac{1}{N_{\text{eff}}(i,j)} \sum_{k=1}^{N_{\text{eff}}(i,j)} \left[\frac{X(\mathbf{s}_k^*) - X(\mathbf{s}_k)}{\delta s_{k(i)}} \right] \left[\frac{X(\mathbf{s}_k^*) - X(\mathbf{s}_k)}{\delta s_{k(j)}} \right]. \tag{7}$$

The estimator $\hat{Q}_{ij}$ includes only points $\mathbf{s}_k$ such that $\delta s(q_1) \le \delta s_{k(i)}, \delta s_{k(j)} \le \delta s(q_2)$. The lower ($q_1$) and upper ($q_2$) percentiles are selected after experimentation during the training stage. The $N_{\text{eff}}(i,j)$ is the number of data points that contribute to $\hat{Q}_{ij}$ and depends on q_1, q_2. The selection of q_1 and q_2 aims to avoid too closely spaced or too distant neighbors, because the former exacerbate errors due to uncorrelated (nugget) fluctuations, while the latter cause errors in derivative approximation and reduce the anisotropy.

2.5 *Solving the CTI Equation System*

The roots of the nonlinear system (6) determine the anisotropic parameters. In (Hristopulos, 2002; 2004) the solution was obtained by minimizing the *objective function* $\Phi(\theta, R) = \left(Z_{\text{diag}} - \hat{Z}_{\text{diag}} \right)^2 + \left(Z_{\text{off}} - \hat{Z}_{\text{off}} \right)^2$, where $Z_{\text{diag}}, Z_{\text{off}}$ are given by Eqs. (6), while $\hat{Z}_{\text{diag}}, \hat{Z}_{\text{off}}$ are the respective estimates obtained from the data. Here we introduce an explicit solution for the anisotropic parameters that will be presented in more detail elsewhere:

$$\tan(2\hat{\theta}) = \frac{2\hat{Z}_{\text{off}}}{1 - \hat{Z}_{\text{diag}}}, \quad \hat{R}^2 = 1 + \frac{1 - \hat{Z}_{\text{diag}}}{\hat{Z}_{\text{diag}} - (1 + \hat{Z}_{\text{diag}})\cos^2 \hat{\theta}}. \tag{8}$$

The solution is valid if $\hat{Z}_{\text{diag}} \ne 1$ and $\hat{Z}_{\text{off}} \ne 0$. If $\hat{Z}_{\text{diag}} = 1$ and $\hat{Z}_{\text{off}} = 0$, then $\hat{R} = 1$ and the rotation angle is indeterminate (isotropic case).

The stability of the solution with respect to *model error* (i.e., if the stationary, differentiable SRF model is not a good approximation of the fluctuations), nugget fluctuations, and discretisation error should be further investigated. The solution (8) is not unconditionally stable for any values of $\hat{Z}_{\text{diag}}$, $\hat{Z}_{\text{off}}$. To see this consider the case $\theta = \pi/4$: Eqs. (6) imply that $Z_{\text{diag}} = 1$, and $Z_{\text{off}} = (1 - R^2)/(1 + R^2)$, i.e., $|Z_{\text{off}}| \le 1$. Next, assume that, due to errors, the estimated values are $\hat{Z}_{\text{diag}} \approx 1$ and $\hat{Z}_{\text{off}} > 1$; then, it follows from Eqs. (8) that $\hat{\theta} \approx \pi/4$ and $\hat{R}^2 \approx -1$. This shows that the solution (8) may become unstable if $\hat{Z}_{\text{diag}}$, $\hat{Z}_{\text{off}}$ involve significant errors, e.g., due to outliers. In such cases, the anisotropic parameters are obtained by minimization of the objective function to obtain a real, albeit inaccurate, number for $\hat{R}$.

2.6 *Use of prior information*

The mean gradient tensor was calculated separately for each of the ten training sets, and the anisotropy parameters were estimated from Eqs. (8). The estimates exhibit a behaviour that lacks systematic time dependence. The prior information on the anisotropic parameters can be used to detect events that cause significant deviations from the established stationary pattern. Rigorous detection algorithms require quantifying the magnitude of various errors in the estimates of the anisotropic parameters. These will be investigated in future work. In this study, the uncertainty of the anisotropy parameters is determined empirically from the fluctuations obtained in the analysis of the training sets.

2.7 *Tuning the algorithms*

The following steps are repeated for each training set (time slice):

1. Exploratory analysis: coordinate normalization, determination of a trend model and residual fluctuations.

2. "Classical" analysis of anisotropy by means of direction-dependent (North-South, East-West) variogram calculations.

3. Estimation of the mean gradient tensor from Eq.(7) for various values of the nearest-neighbour percentiles q_1 and q_2.

4. Estimation of anisotropic parameters from Eqs. (6)-(8). Uncertainty analysis of the estimates $\hat{R}$ and $\hat{\theta}$. Choice of an "effective" range for q_1 and q_2.

5. Coordinate transformation to the "optimal" IPC coordinate system. Verification of isotropy by means of (i) re-evaluation of the anisotropic parameters and (ii) re-calculation of the variograms.

Steps (2), and (5.ii) are not required for implementing the method. Nonetheless, they are used in this study for comparison with the established variogram analysis.

3. RESULTS

The CTI method was applied to the training sets and the two data sets. The training sets show similar behaviour, and permit establishing a 'typical range' for the anisotropy parameters. This range is updated as new observations become available. The anisotropy in the first data set is in agreement with training values. The second (joker) data set exhibits noticeably different behaviour. Even after removal of two extreme values that lead to an unrealistic estimate of the anisotropic ratio, it is obvious that anisotropy is different than in the other sets. In 3.1-3.6 below we elaborate on the procedure outlined in section 2.7.

3.1. *Exploratory Analysis*

The coordinates are normalized by means of the transformations $x' = (x - \overline{x})/s_x$ and $y' = (y - \overline{y})/s_y$, where s_x, s_y are respectively the standard deviations of the x and y coordinates. For the present study $s_y/s_x \approx 2.16$. *The following analysis of anisotropy uses the normalized coordinates*. The data were subsequently detrended (step 1) using the quadratic global polynomial

$$m_X(\mathbf{s}) = a_0 + a_x x + a_y y + a_{xx} x^2 + a_{yy} y^2 + a_{xy} xy, \tag{9}$$

to model the trend. The values of the polynomial coefficients are given in Table 1. Anisotropic dependence of the trend is clearly marked by the difference between the coefficients a_x and a_y as well as a_{xx} and a_{yy}. The coefficients reveal stronger variation in the vertical (North-South) than in the horizontal (East-West) direction. The *fluctuations* are given by the residuals $\chi(\mathbf{s}) = X(\mathbf{s}) - m_X(\mathbf{s})$. In Table 2, the skewness and kurtosis coefficients of the fluctuations are compared with those of the initial data. The former are overall closer to the Gaussian values. The histograms of the residuals are plotted in Figure 1 along with the optimal Gaussian probability density function (p.d.f.) fits. The histograms for the original data are given in (Dubois and Galmarini, 2005). The near-normality of the fluctuations will be used to identify outliers, in particular in the joker set.

Trend coefficients	a_0	a_x	a_y	a_{xx}	a_{xy}	a_{yy}
First training set	106.05	2.97	−9.89	0.25	0.55	−8.72
Second training set	105.73	2.70	−10.21	0.45	0.59	−8.78
Third training set	105.99	2.88	−10.63	0.39	0.51	−7.65
Fourth training set	101.04	2.86	−8.14	0.47	0.23	−7.72
Fifth training set	100.14	2.29	−7.90	0.42	1.15	−8.16
Sixth training set	97.00	2.07	−7.40	0.60	1.39	−7.80
Seventh training set	99.37	2.14	−7.15	0.14	1.10	−7.80
Eighth training set	99.31	3.20	−8.39	0.26	0.59	−7.15
Ninth training set	105.06	3.42	−9.93	−0.10	0.90	−8.37
Tenth training set	103.00	3.21	−7.90	0.07	1.04	−7.73

Table 7. *Polynomial trend coefficients for the gamma dose rates in the ten training sets.*

3.2. *Variogram Calculations*

Experimental semivariograms are calculated in the North-South (vertical) and East-West (horizontal) directions, using 16 distance classes and an angle tolerance of $\pm 0.1\pi$, for lags up to 50% of the maximum distance between data.

Non-Gaussian Deviations	Skewness (initial)	Skewness (residuals)	Kurtosis (initial)	Kurtosis (residuals)
First training set	0.03	−0.09	−0.55	−0.22
Second training set	0.08	−0.04	−0.48	−0.12
Third training set	0.12	−0.13	−0.31	−0.18
Fourth training set	0.20	0.11	−0.05	0.39
Fifth training set	0.20	0.14	−0.20	0.29
Sixth training set	0.11	−0.05	−0.50	0.19
Seventh training set	0.12	−0.02	−0.40	−0.11
Eighth training set	0.15	0.10	−0.09	0.24
Ninth training set	0.05	−0.01	−0.41	0.01
Tenth training set	0.12	0.01	−0.22	0.11

Table 2. *Skewness and kurtosis coefficients of the gamma dose rates in the ten training sets (initial values and detrended residuals).*
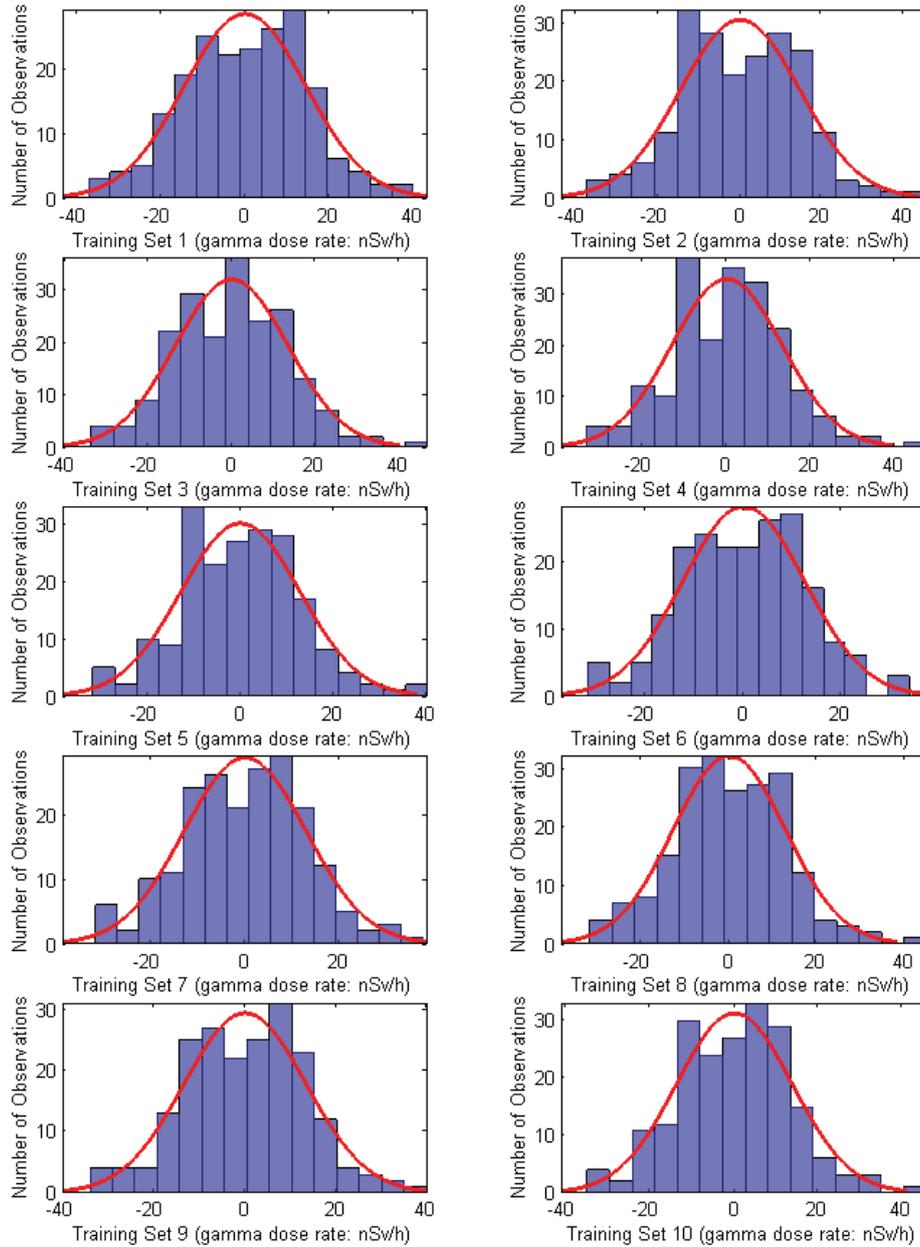
Figure 1. *Histograms and fits with the optimal Gaussian p.d.f. model for the fluctuations of the gamma dose rates in the ten training sets.*

Visual inspection of the semivariograms in Figure 2 implies shorter vertical than horizontal correlation range and a finite nugget component.

3.3. *Estimation of the Mean Gradient Tensor*

The three components of the mean GT were calculated from the data using Eq. (7). In all calculations, q_1=1% was used. A smaller value (e.g., 0.5%) allows random fluctuations to dominate the average and smears the anisotropy, while a larger value (e.g., 5%) reduces considerably the number of points involved in the average of the off-diagonal term $\hat{Q}_{xy}$.

The estimates of the mean GT for q_2 between 20% and 100% are shown in Figure 3. The main features of these plots are: (1) $\hat{Q}_{yy} > \hat{Q}_{xx}$, which indicates a "tendency" for higher correlation range in the x direction (in agreement with the semivariograms). (2) The difference between $\hat{Q}_{xx}$ and $\hat{Q}_{yy}$ is reduced as q_2 increases, while at the same time $\hat{Q}_{xy}$ tends to zero (from negative values), indicating an increasingly isotropic behaviour. The GT dependence is similar in all training sets. This supports the idea that, to a first approximation, the anisotropic pattern is *stationary in time*.
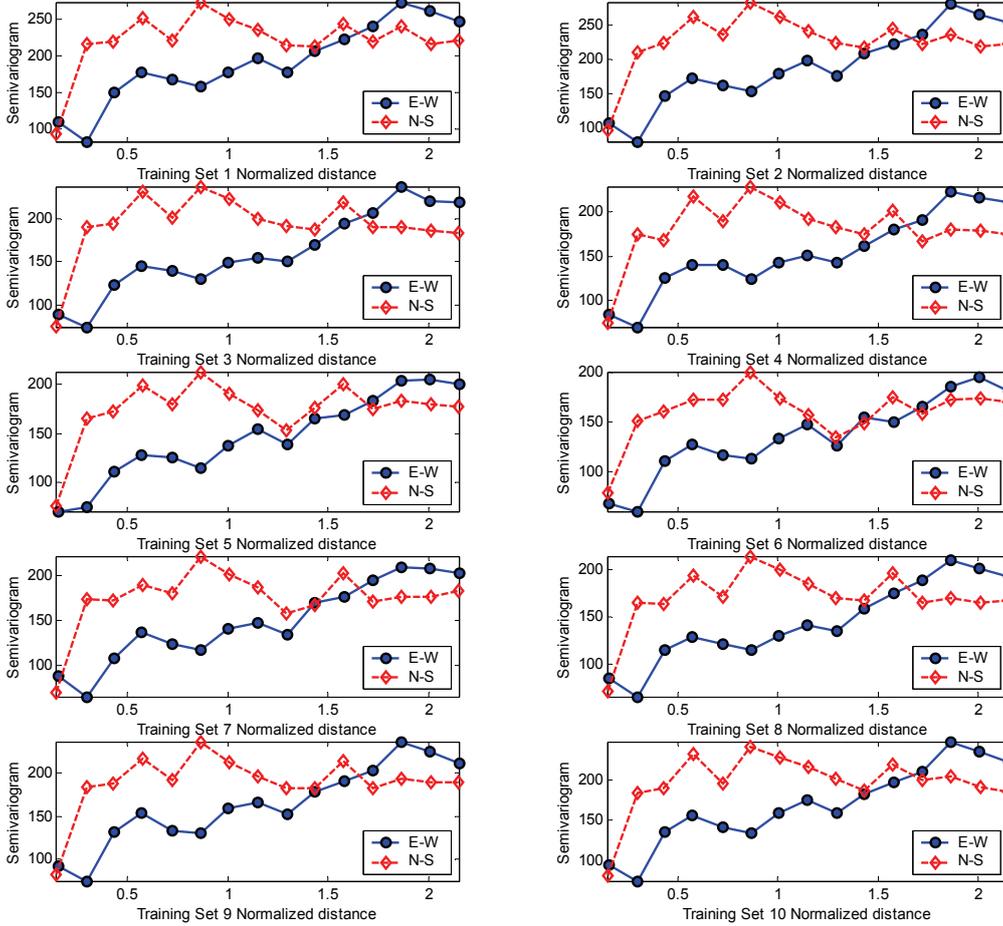
Figure 2. *Semivariograms of gamma dose rate fluctuations for the ten training sets.*

As shown in Figure 4, the number of sites contributing to the sample average increases with q_2; also, $N_{eff}(x,y)$ is smaller than $N_{eff}(x,x)$ and $N_{eff}(y,y)$, since $N_{eff}(x,y)$ involves only sites the neighbours of which obey restrictions in both the x and y directions.

3.4. *Estimation of the Anisotropic Parameters*

The graph of the estimated anisotropic ratio versus q_2 (Figure 5), follows overall a declining trend. In all sets, $\hat{R}$ decreases sharply as q_2 changes from 20% to 25%, due to a significant drop in the absolute value of the off-diagonal term $\hat{Q}_{xy}$. The rotation angle, plotted in Figure 6, follows the same trend as $\hat{R}$, taking values between $30°$ and $-13°$. The larger variation is explained because θ is indeterminate as $\hat{R} \to 1$ for larger q_2.

Uncertainty analysis: The estimate $\hat{Q}_{xy}$ at $q_2 = 20\%$ involves only 15 points, and thus it is not very reliable. If the points $q_2 = 20\%$ are discarded, the anisotropic ratio varies between 1.18 and 2.35 for all training sets and all q_2. Within each training set the variation is between 10% and 20% of the mean value.
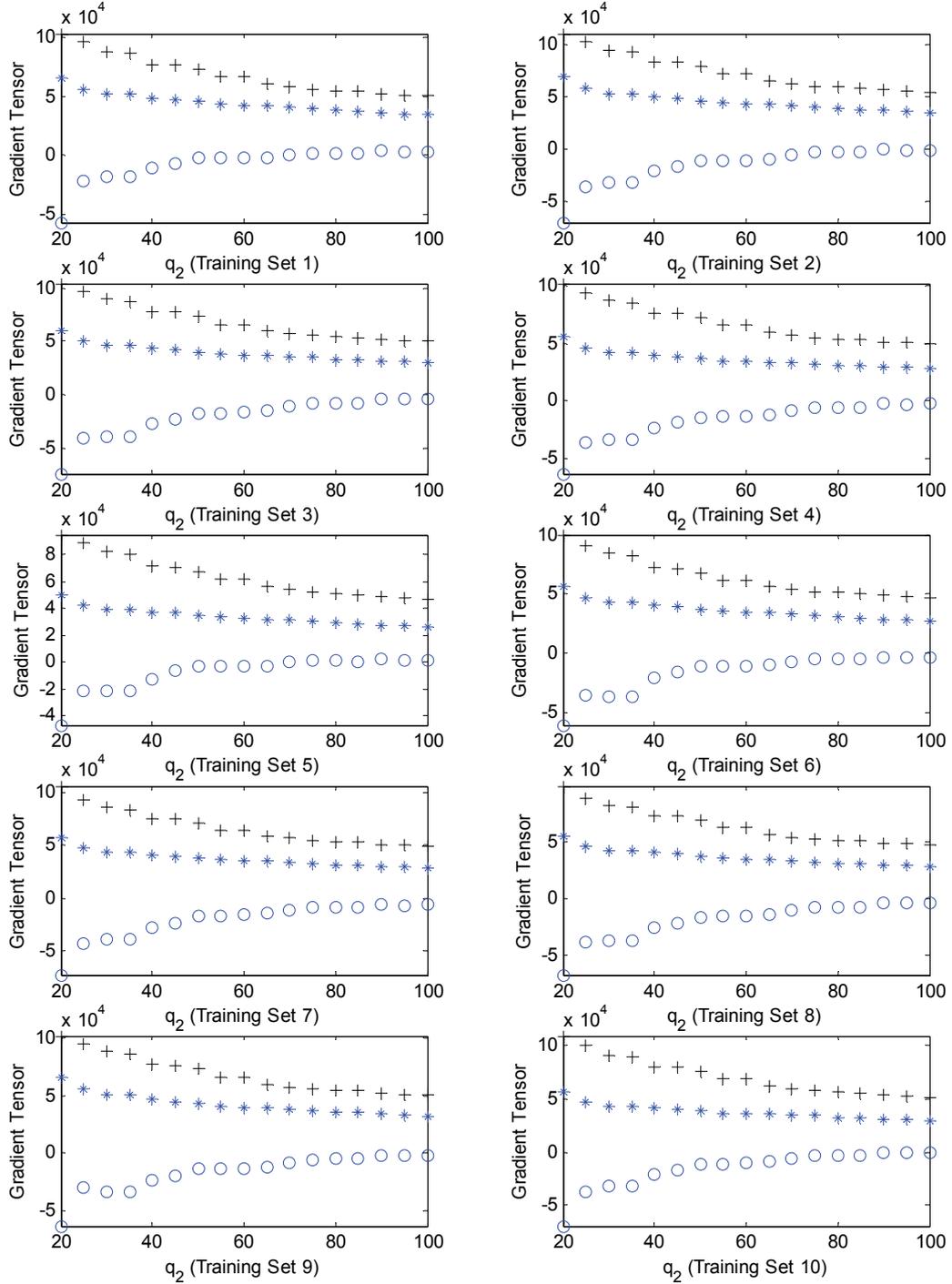
Figure 3. *Mean gradient tensor estimates versus q_2. The following symbols are used: $\hat{Q}_{xx}$ (*), $\hat{Q}_{yy}$ (+), $\hat{Q}_{xy}$ (o).*

Optimal anisotropy parameters: The behaviour observed in Figures 5 and 6 does not allow determining the anisotropic parameters unambiguously. After experimentation, we estimate *"optimal stationary"* $\hat{R}^*$ and $\hat{\theta}^*$ from the *median* of 70 values that involve the estimates for the seven values $q_2 = [0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55]$. This decision aims for a compromise between the number of points used in the estimation and smoothening effects. The resulting estimates are $\hat{R}^* \approx 1.66$ and $\hat{\theta}^* \approx 26.1°$. We base the estimate on the median rather than the mean, since the former is less sensitive to errors. The mean $\hat{R}$ based on the 70 values is approximately 1.73 and the standard deviation is 0.30. The minimum $\hat{R}$ is 1.25 and the maximum is 2.35.
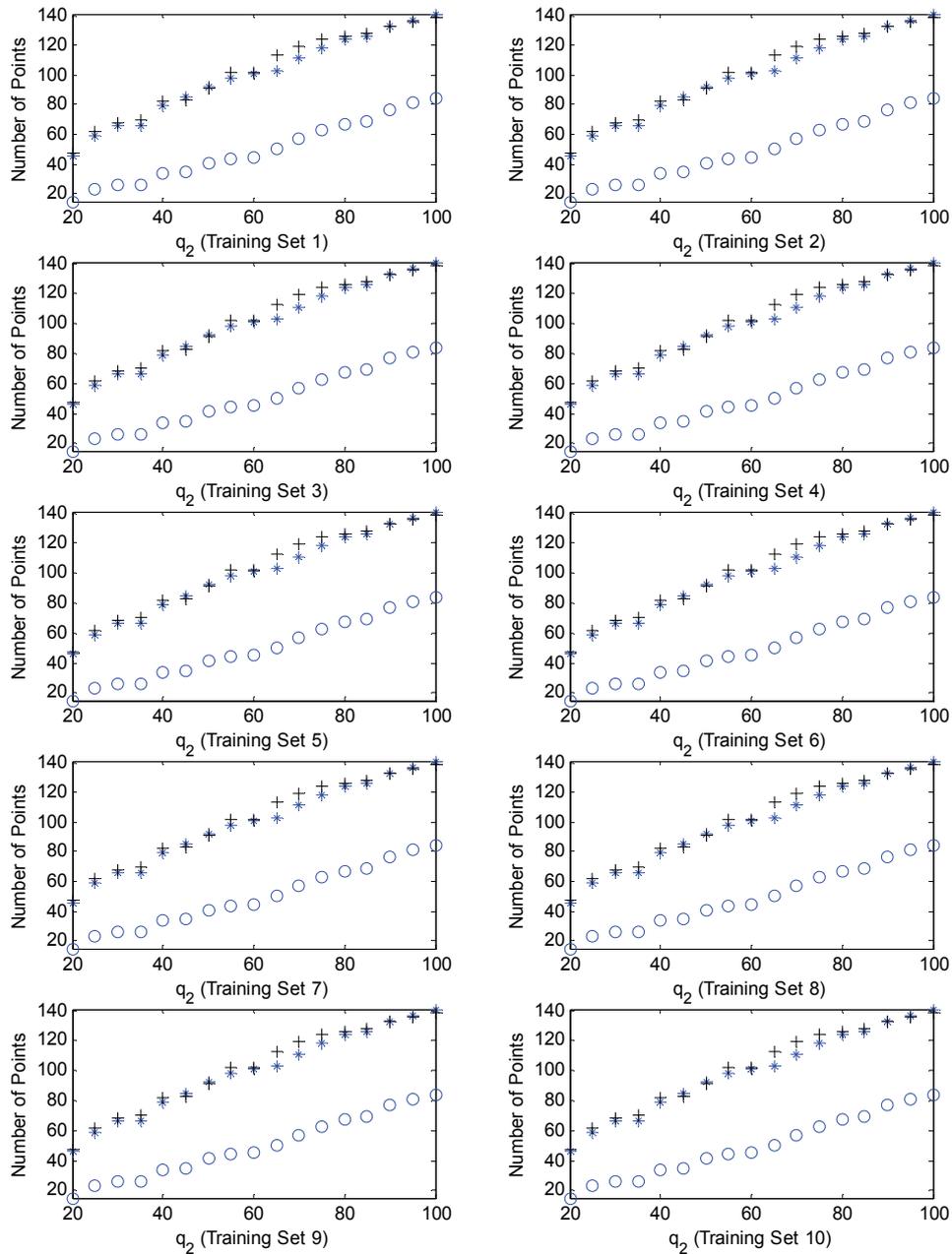
Figure 4. *Number of points used in the mean gradient tensor estimates versus q_2. The following symbols are used: n_{xx} (*), n_{yy} (+), n_{xy} (o).*

3.5. *IPC Transformation and Isotropy Verification*

The coordinates of the IPC system are obtained from Eq. (1) using $\hat{R}^*$ and $\hat{\theta}^*$ as the anisotropic parameters for *all the training sets*. Ideally, the anisotropic ratio should equal one in the IPC system.

Re-evaluation of anisotropic parameters: The mean GT in the IPC system is plotted in Figure 7. The plots exhibit marks of isotropy, e.g. reduced difference between the diagonal GT elements, and near-zero values for the off-diagonal elements. The estimated anisotropic ratio is plotted in Figure 8. The graphs show a clear reduction of anisotropy compared with Figure 5. The estimate of the "optimal" anisotropy ratio, based on the median of seventy values, is 1.11 (to be compared with the initial estimate of 1.66). Rotation angles are plotted in Figure 9. In contrast with the declining trend in Figure 6, the graphs exhibit oscillations of the rotation angle (some extending between positive and negative values). This behaviour is a mark of isotropy due to the absence of an optimal angle.
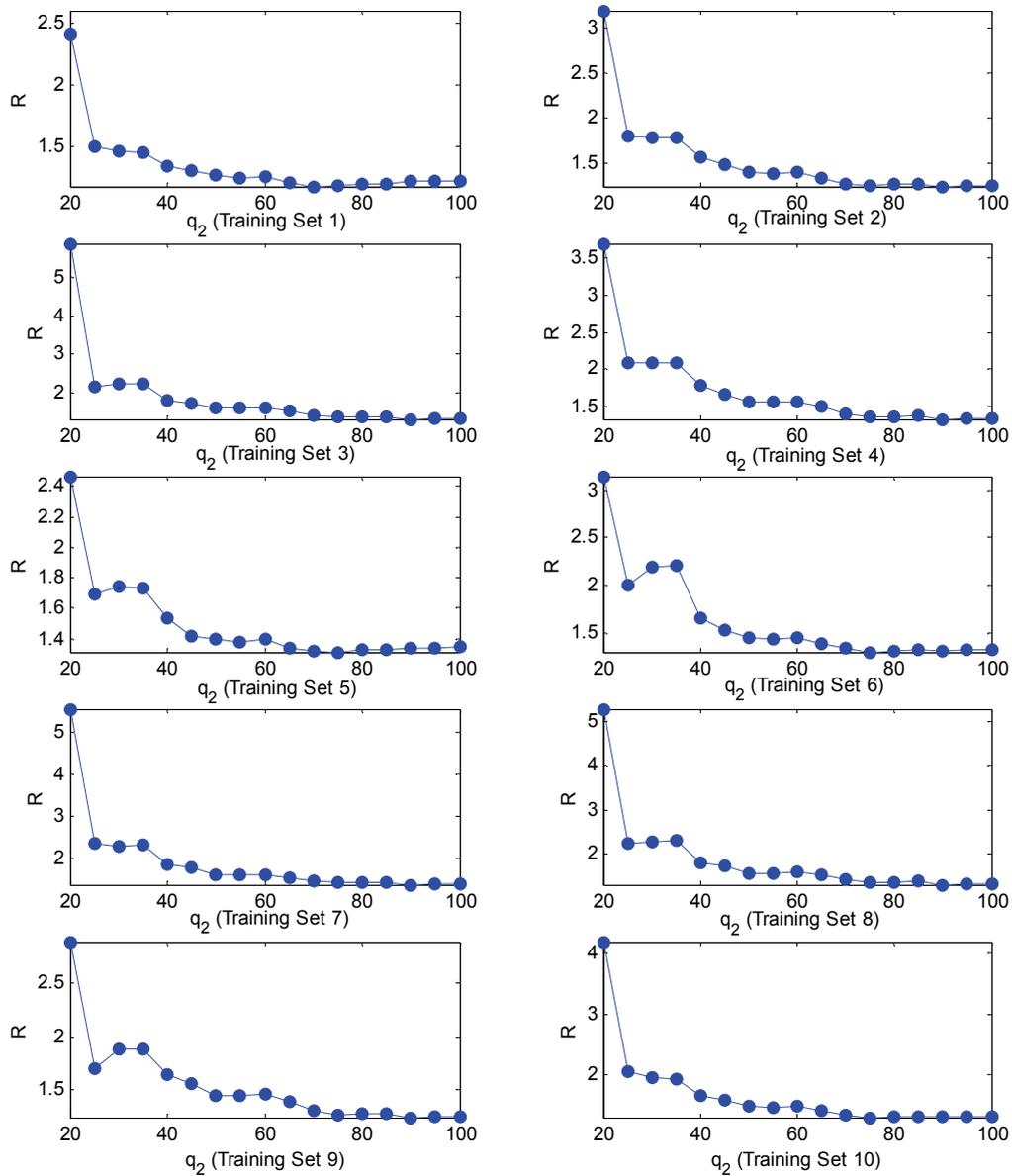
Figure 5. *Estimated anisotropic ratio versus q_2, for the ten training sets.*

Semivariograms in the IPC system: The semivariograms of the training sets in the E-W and the N-S directions are calculated in the IPC coordinates. The plots in Figure 10 exhibit improved isotropy, at least for certain lag distances, compared with their counterparts in the original system (Figure 2).
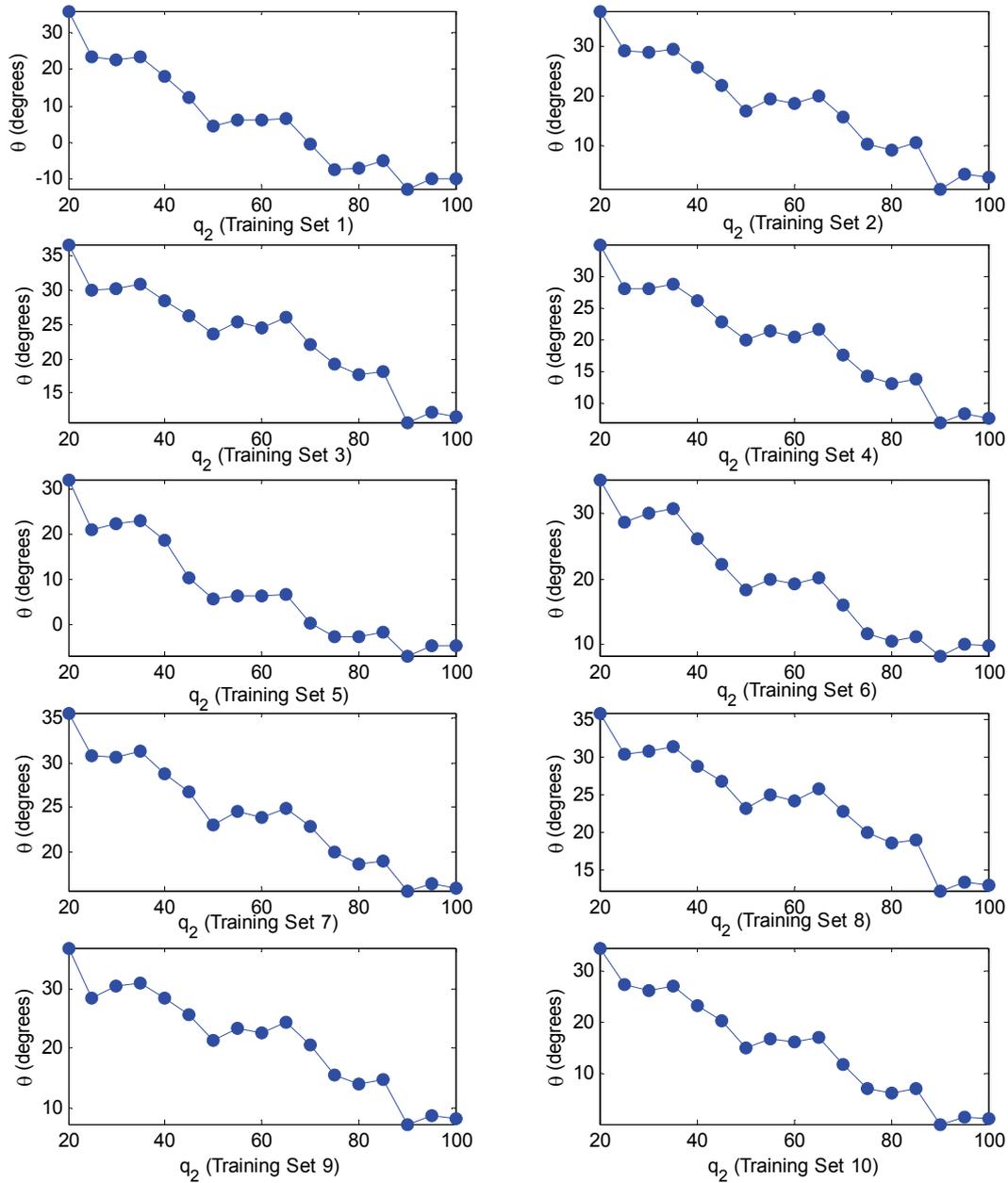
Figure 6. *Estimated rotation angle versus* q_2 *, for the ten training sets.*

3.6. *Application of the CTI to the Data*

First data set: The CTI analysis is applied to estimate the anisotropy as described above. The values of the mean GT elements versus q_2 are shown in Figure 11(a), and the anisotropic parameters in Figures 12(a) and 13(a). In the latter, only values of q_2 in the optimal range are shown. The anisotropic ratio varies between 1.37 and 1.63, and the rotation

angle between $14°$ and $26.5°$. Overall, there is no significant difference with the training sets. The new values are added to the parameters estimated from the training sets', and the "optimal estimates" based on the medians of the 77 values are updated to $\hat{R}^* \approx 1.65$ and $\hat{\theta}^* \approx 25°$.
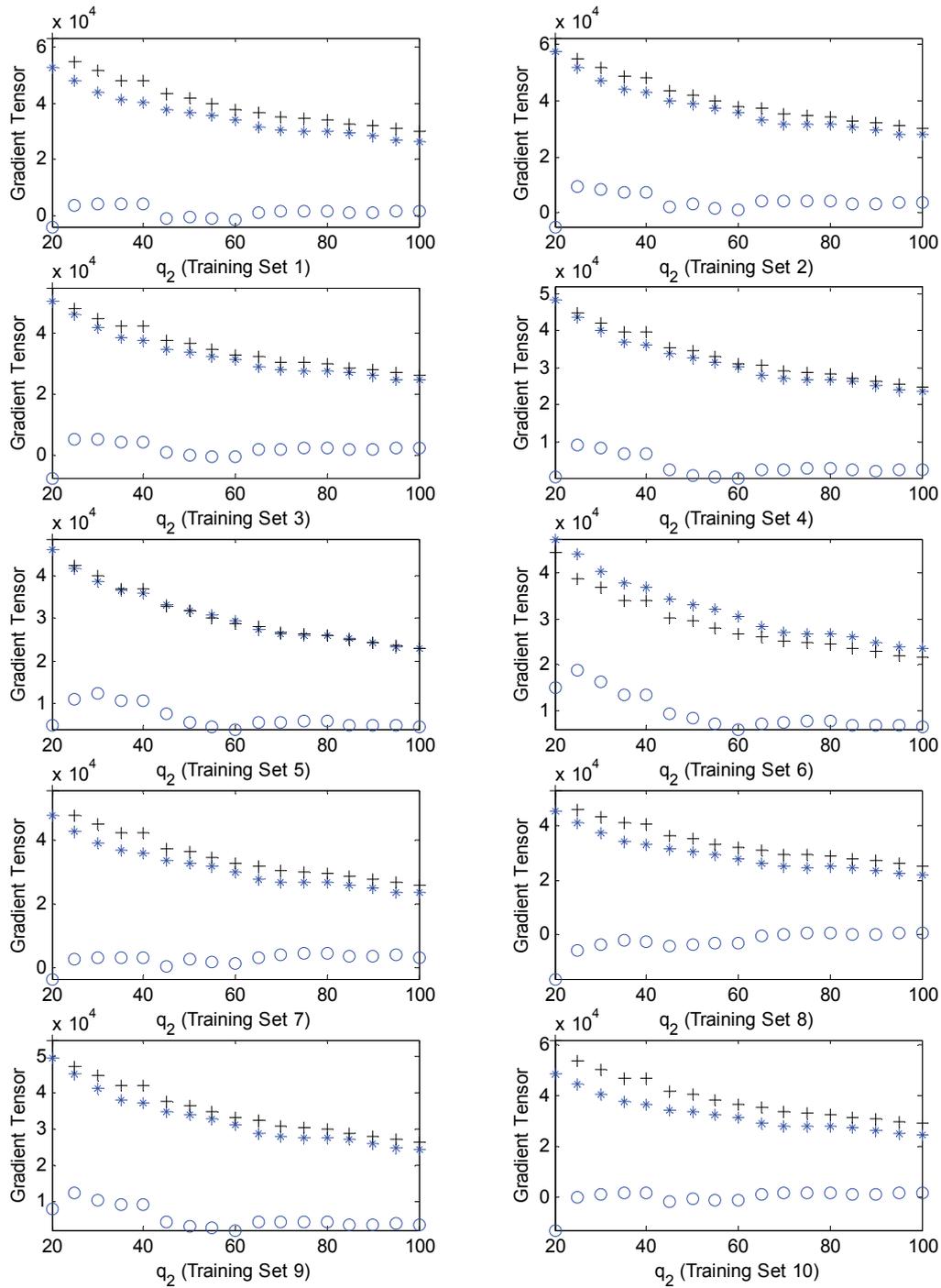
Figure 7. *Mean gradient tensor estimates versus* q_2 *in the IPC system for the ten training sets. The following symbols are used:* $\hat{Q}_{xx}$ *(*),* $\hat{Q}_{yy}$ *(+),* $\hat{Q}_{xy}$ *(o).*

Joker data set: Without pre-processing to remove outliers, the plots of the gradient tensor elements versus q_2, shown in Figure 11(b), exhibit discontinuities. These develop where an increase of q_2 brings the points with "extreme dose rate measurements" into the GT average. Such abrupt changes lead to instabilities in the solution (8), as explained in Section 2.5. The objective-function-minimization approach yields real but unrealistic values for $\hat{R}$ ($\approx 10^8$), corresponding to a poor minimum of the objective function (graph not shown in the paper); this is a warning that the CTI conditions are violated.

The discontinuities can be eliminated by removing the fluctuations with absolute value exceeding four times the sample's standard deviation. The resulting dependence of $\hat{R}$ and $\hat{\theta}^*$, shown in Figures 12(b) and 13(b) is well behaved. However, it is clear from the plots that the signature of anisotropy has changed. The anisotropic ratio takes values $\hat{R} < 1$, in the range from 0.72 to 0.86. Given the considerable uncertainty in $\hat{R}$ and without knowing the probability distribution of $\hat{R}$, it is not possible to rigorously define the statistical significance of this change. Nonetheless, given (1) that the $\hat{R}$ values are outside the range obtained from the training sets and the first data set, and (2) the opposite (compare Figs. 12(a) and 12(b)) trend dependence of $\hat{R}$ as a function of q_2, it is reasonable to conclude that the observed behavior represents a systematic change in the spatial distribution.

4. DISCUSSION

For a thorough evaluation of the CTI method, improved understanding of (1) the accuracy and uncertainty (precision) of the estimates and (2) the method's sensitivity to outliers and errors (e.g., ergodic, discretization, model) in the mean GT should be achieved. Quantification of nugget effects on the estimation of the mean GT will help to determine a range of nugget variance for reliable application of the CTI. Finally, it would be desirable to relax the conditions under which the method is valid and to extend it to three-dimensional data distributions. It needs to be verified if the IPC transformation improves the prediction accuracy compared to estimates based on explicitly directional models (e.g., kriging with anisotropic semivariogram model).

The time required for estimating the anisotropic parameters of the ten training sets, calculating the isotropic coordinates and re-evaluating the anisotropic parameters in the IPC system is approximately 28 sec on a 3GHz PC with 2GB of RAM. This does not include calculation of the experimental semivariograms; these are not necessary for application of the CTI method, but they are included in this study only for comparison purposes.

Regarding spatial interpolation, Spartan Spatial Random Fields (Hristopulos, 2003a; 2003b; 2004; 2005), currently investigated by the author, provide a new framework. The SSRF's exploit a 'pseudo-energy' functional (cost function), to define the field's probability density function. The latter is thus defined in terms of interactions, as opposed to an explicit, data-driven covariance matrix. If the interactions are restricted between neighbours, SSRF's are essentially Markov random fields. In spatial interpolation, it is possible to use differentiable Spartan covariance models in kriging algorithms. Other predictors based on a local definition of the error (in contrast with the ensemble-based definition used in kriging), are also possible in the SSRF framework. More importantly, the definition of the SSRF p.d.f. in terms of a 'pseudo-energy' functional is promising for the introduction of physics-based information in the geostatistical framework (Galmarini, 2005).
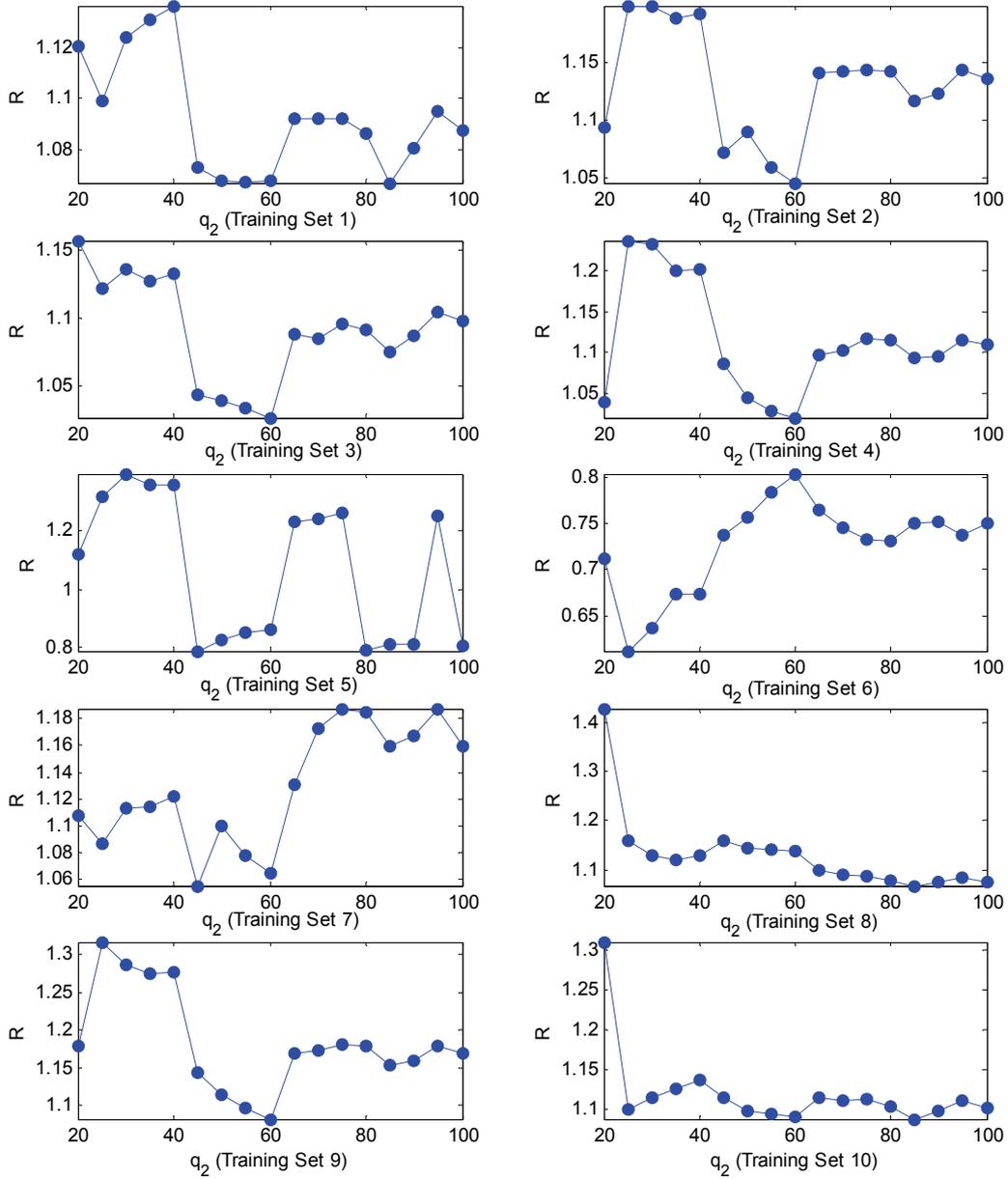
Figure 8. *Estimated anisotropic ratio versus* q_2 *(in the IPC system) for the ten training sets.*

Regarding application of the CTI method in an automated mapping system, human intervention is necessary in the training stage, where depending on the specific data experimentation may be required to determine the anisotropy parameters. If the anisotropic pattern is time-stationary, there is no need for further intervention, since it is possible to detect automatically outliers and systematic changes. Standard estimation methods can then be applied in the isotropic (IPC) system. Finally, if the generated maps need to be displayed in the original system, the coordinates can be restored by means of the inverse transformations:

$$
\begin{aligned}
x_i &= \left(\tilde{x}_i \cos\hat{\theta}^* - \tilde{y}_i \sin\hat{\theta}^* / \hat{R}^* \right) s_{\mathrm{x}} + \bar{x} \\
y_i &= \left(\tilde{x}_i \sin\hat{\theta}^* + \tilde{y}_i \cos\hat{\theta}^* / \hat{R}^* \right) s_{\mathrm{y}} + \bar{y}
\end{aligned}
\qquad i = 1, \ldots, N.
\tag{10}
$$
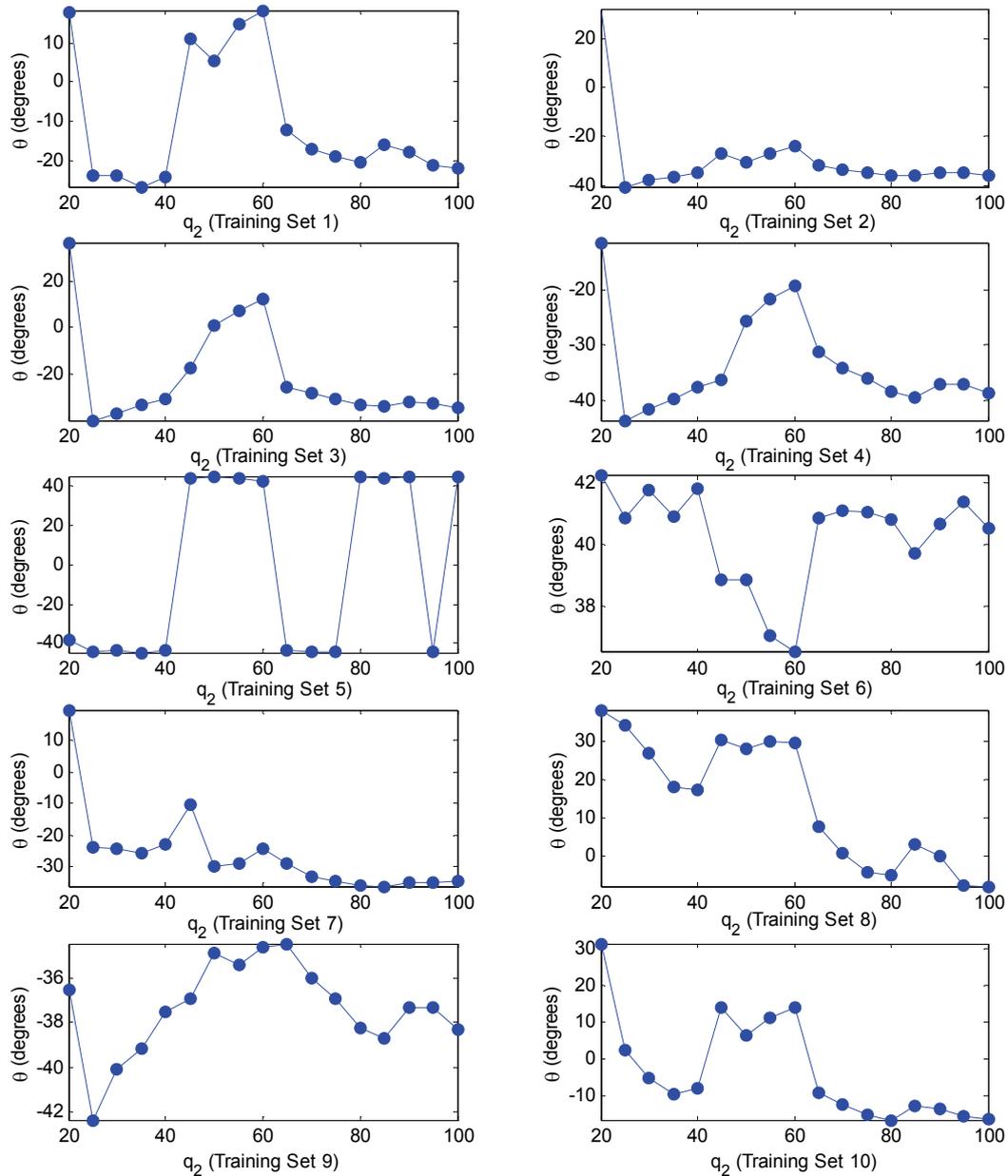
Figure 9. *Estimated rotation angle versus* q_2 *(in the IPC system) for the ten training sets.*

5. CONCLUSIONS

We employed the CTI method to identify anisotropies in the fluctuations of the gamma dose rate distribution for the ten training sets and two data sets of the SIC2004 test. The CTI method offers a systematic approach for determining anisotropy parameters in 2D spatial data. Systematic anisotropy identification serves two purposes: First, it enables coordinate system transformations that render the spatial dependence isotropic and thus simplifies the application of geostatistical methods for mapping.
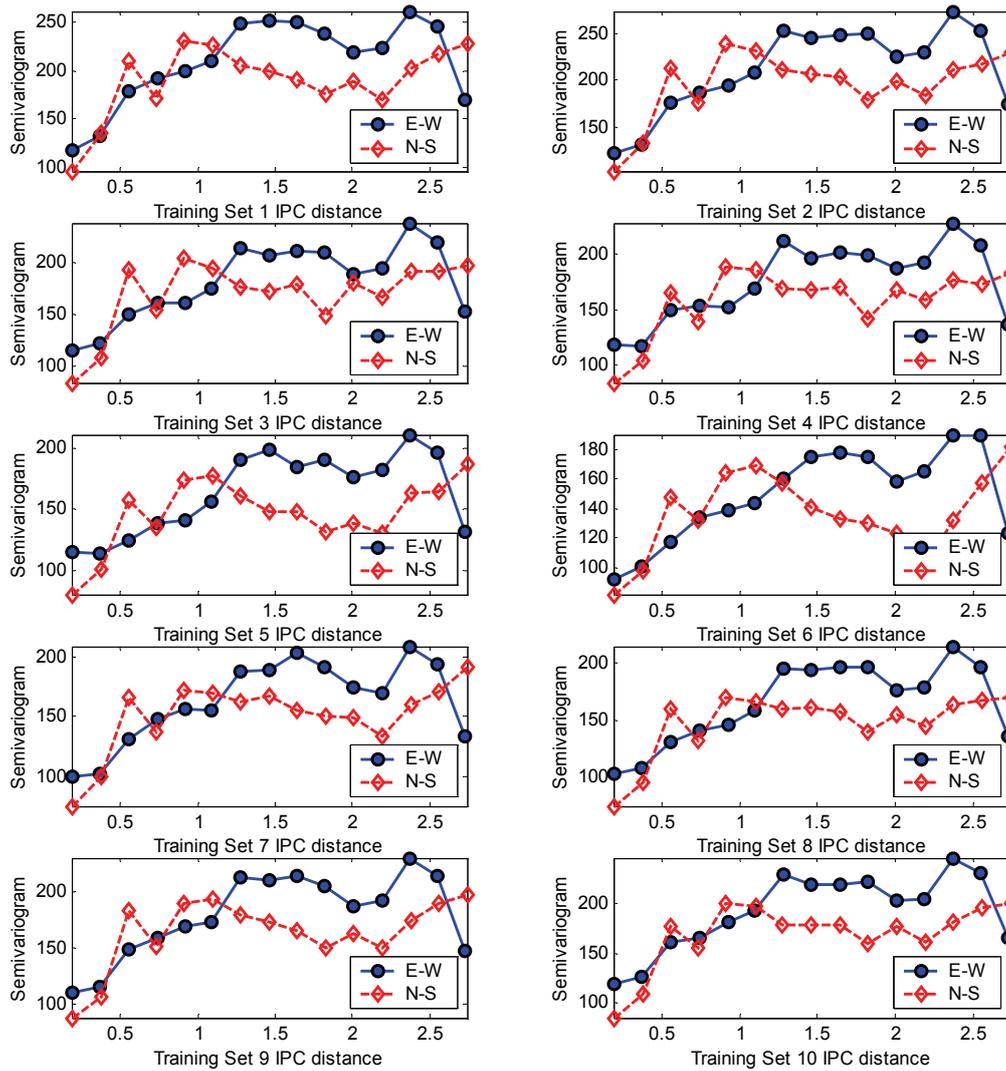
Figure 10. *Semivariograms for the ten training sets in the IPC system.*

Second, the CTI method permits detection of systematic changes in the spatial data; this capability can be used in an automatic mapping system to generate warnings in the event of such changes. It was found that the anisotropy in the training sets follows a stationary pattern, and optimal values of the anisotropic ratio and rotation angle were determined. It was verified that the isotropy transformation leads to significant reduction of the anisotropy. The first data set was found to follow the same anisotropy pattern as the training sets. In contrast, the joker set exhibited a distinctly different anisotropic pattern.

Codes

The MATLAB code used to determine the anisotropy parameters can be obtained by e-mailing the author.

Acknowledgments
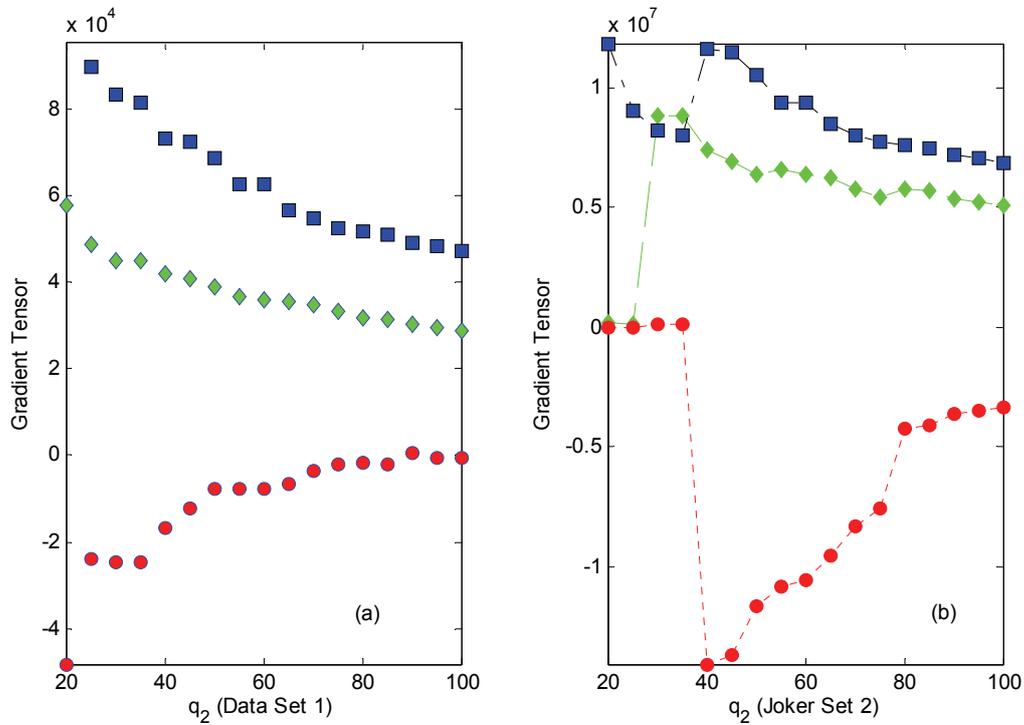
Figure 11. *Mean gradient tensor elements for the real (a) and joker (b) data sets. The following symbols are used: $\hat{Q}_{xx}$ (diamonds), $\hat{Q}_{yy}$ (squares), $\hat{Q}_{xy}$ (circles).*
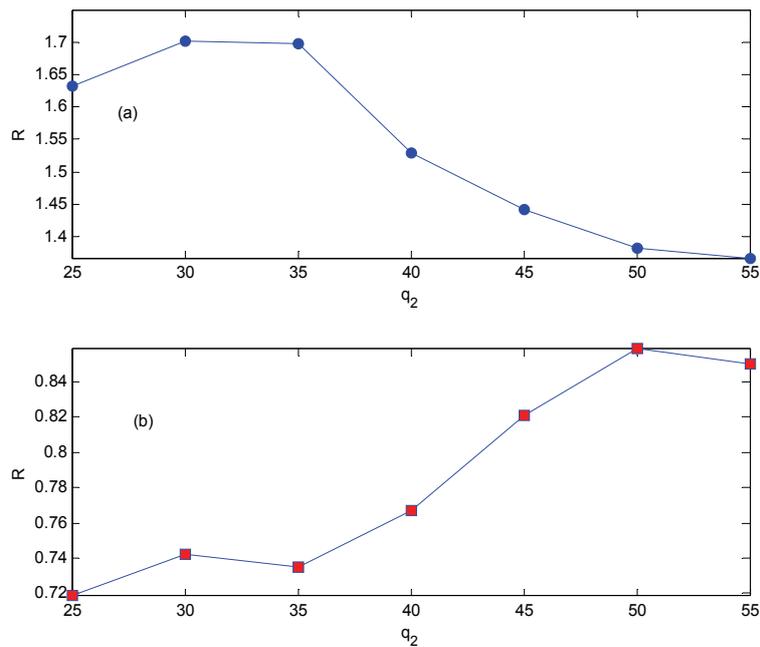
Figure 12. *Estimated anisotropic ratio for the real (a) and the joker (b) data sets.*
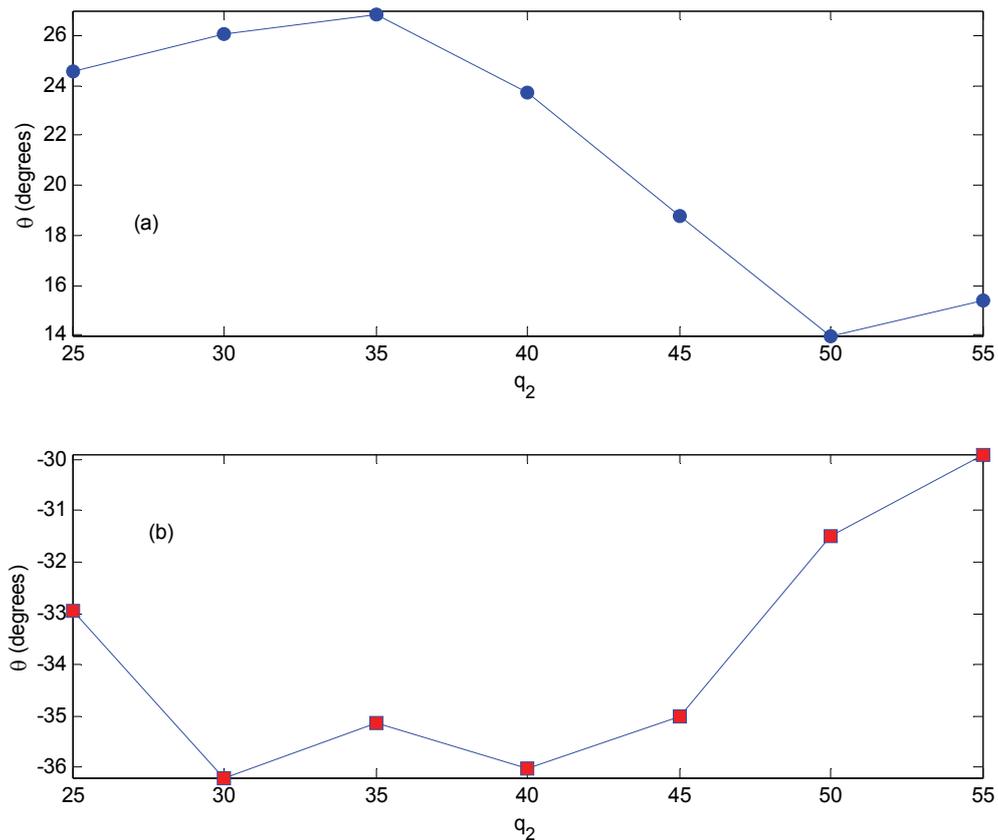
Figure 13. *Estimated rotation angle for the real (a) and joker (b) data sets.*

References

1. Abrahamsen, P. (1997). *A Review of Gaussian Random Fields and Correlation Functions*, 2nd Ed. Technical Report 917, Norwegian Computing Center, Box 114, Blindern, N-0314, Oslo, Norway.

2. Dubois, G., Galmarini, S. (2005). "*Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results*". This Volume.

3. Ecker, M. D., Gelfand, A. E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, 32(1): 67-82.

4. Galetakis, M., and Hristopulos, D.T. (2004). Prediction of Long-Term Quality Fluctuations in the South Field Lignite Mine of West Macedonia. In: "*Proceedings of Advances in Mineral Resources Management and Environmental Geotechnology.*" Agioutantis, Z. and Komnitsas, K. (Eds.), Heliotopos Conferences, Greece, pp. 133-138.

5. Galmarini, S. (2005). "*Real-Time Geostatistics for Atmospheric Dispersion Modelling and Vice-Versa*?" This Volume.

6. Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 481 p.

7. Hristopulos, D. T. (2002). New anisotropic covariance models and estimation of anisotropic parameters based on the covariance tensor identity. *Stochastic Environmental Research and Risk Assessment*, 16(1): 43-62.

8. Hristopulos, D. T. (2003a). Spartan Gibbs random field models for geostatistical applications. *SIAM Journal on Scientific Computing*, 24(6): 2125-2162.

9. Hristopulos, D. T. (2003b). Simulations of Spartan Random Fields. *In "Proceedings of International Conference of Computational Methods in Sciences and Engineering".* Simos, T.E. (Ed.). World Scientific, London, p. 242-247.

10. Hristopulos, D. T. (2004). Anisotropic Spartan random field models for geostatistical analysis. In: *"Proceedings of Advances in Mineral Resources Management and Environmental Geotechnology."* Agioutantis, Z. and Komnitsas, K. (Eds.), Heliotopos Conferences, pp. 127-132.

11. Hristopulos, D. T. (2005). Numerical Simulations of Spartan Gaussian Random Fields for Geostatistical Applications on Lattices and Irregular Supports. *Journal of Computational Methods in Sciences and Engineering* (in press).

12. Swerling, P. (1962). Statistical Properties of the Contours of Random Surfaces. *IEEE Transactions in Information Theory*, 8(4): 315-321.

13. Wackernagel, H. (2003). *Multivariate Geostatistics*. Springer Verlag, Berlin, 387 pp.

14. Yaglom, A.M. (1987). *Correlation Theory of Stationary and Related Random Functions I*. Springer, New York, 526 p.

Machine Learning for automatic environmental mapping: when and how?

Nicolas Gilardi[1], Samy Bengio[2]

[1]*Particle Physics Experiment Group, University of Edinburgh, Edinburgh EH9 3JZ, UK*
[2]*IDIAP Research Institute, CP 592, rue du Simplon 4, 1920 Martigny, Switzerland*

E-mail: ngilardi@ph.ed.ac.uk

Abstract: This paper discusses the opportunity of using Machine Learning techniques in an automatic environmental mapping context, as was the case for the SIC2004 exercise. First, the Machine Learning methodology is quickly described and compared to Geostatistics. From there, some clues about when to apply Machine Learning are proposed, and what outcomes can be expected from this choice. Finally, three well known regression algorithms: K-Nearest Neighbors, Multi Layer Perceptron and Support Vector Regression, are used on SIC2004 data in a Machine Learning context, and compared to Ordinary Kriging. This illustrates some potential drawbacks of SVR and MLP for applications such as SIC2004.

Keywords: Machine Learning, K-Nearest Neighbors, Multi Layer Perceptron, Support Vector Regression.

1. INTRODUCTION

It is a common mistake in some fields (such as Geostatistics) to confuse Machine Learning and Artificial Neural Networks (ANNs). Machine Learning can be seen as a methodology (similarly to Geostatistics), while ANNs are a family of modeling functions (like Kriging). Algorithms such as ANNs are very powerful estimators, and without a very careful methodology to control them, they can end up predicting something very different from the expectations. This is partly the origin of the bad reputation ANNs have in some scientific communities. Machine Learning allows to control efficiently the level of information an ANN is modeling. But it is a very general methodology and can also be used with other types of modeling functions, from the very simple and general K-Nearest Neighbors [5], to the more complex Support Vector Machines [17] (for static data), or Hidden Markov Models [13] (for sequential data). In fact, one can use almost any family of statistical modeling functions in the context of Machine Learning. However, now that the Machine Learning methodology is becoming more and more understood, a new tendency seems to be the application of complex algorithms on problems which do not always need them. Environmental mapping may be one of the fields concerned by such a tendency.

Thanks to the increasing concern about environmental issues, there is now a huge amount of data produced by nearly real time monitoring of pollution problems such as radioactivity or exhaust gas emissions. Fast and efficient estimators are thus necessary to produce high quality decision maps, and the objective of SIC2004 [3] was to make a review of the state-of-the-art in the domain. In such a situation it is natural to think about applying Machine Learning methodology. But before doing that, it is necessary to verify whether there is a need for it, and if so, which algorithm to choose for this specific application? In this paper, we try to give some hints about these two questions. First, we give an overview of the ideas behind the Machine Learning methodology. Then, we describe some advantages and drawbacks of the method by comparing it to Geostatistics, and give some advice on when it might be useful to apply Machine Learning instead of other modeling methods. Finally, we evaluate the relative efficiency of three algorithms applied to SIC2004 data in a Machine Learning context and conclude on the opportunity to use each of them for such application.

2. MACHINE LEARNING

Machine Learning can be seen as a branch of both statistics and computer science, which tries to develop computing methods with the aim to solve the so-called learning problem.

2.1. *Learning from data*

As defined by Vapnik [17], the learning problem consists of extracting meaningful information from a finite number of data, called the training set. This very general definition contains the underlying idea that data are the central source of available information. Any prior knowledge about the data and the problem at hand is of course useful but not mandatory.

Unfortunately, in general the training set not only contains the target meaningful information, but also contains less interesting information (such as various kinds of noise) which thus needs to be filtered out. A good model should then be able to extract only the general information from data, and filter out the specific one.

Vapnik showed [17] that it is possible to control the generalization performance of a model (how good it is expected to perform on unseen data, also known as test data) by controlling the capacity of the family of functions it is constructed from. This capacity roughly corresponds to the number of degrees of freedom of the family of functions (for instance the degree of a polynomial).

2.2. *The bias/variance dilemma*

The main principle behind capacity control is to be able to solve the famous bias/variance dilemma [6]. A family of functions with a high capacity should allow to fit almost any data set, i.e. it will have a low bias, defined as the error on the training set, which the learning procedure tries to minimize. Unfortunately, a new model based on a slightly different data set (with some examples that have been changed, but still coming from the same distribution) may end up with a solution quite different from the previous one. This reflects that this family of functions has a high variance, corresponding to the number of solutions in the family of functions that are compatible with the training set, and this is also to minimize.

On the other hand, a family of functions with low capacity will behave exactly the converse: it will produce almost always the same model whatever the exact content of the data set, but many of its predictions will be heavily biased.

It is then easy to understand that finding a family of functions with the optimal capacity (the one that simultaneously minimize the bias and the variance terms) will increase the chance to find a model with optimal generalization performance. Several techniques have been proposed in the literature to search for this optimal capacity, the most generic one, also widely used, being any flavor of the general cross-validation algorithm. In addition, some specific algorithms, such as large margin classifiers [17], contain an "automatic tuning" of their capacity during the construction of the model.

3. MACHINE LEARNING AND GEOSTATISTICS

Machine Learning and Geostatistics have many things in common. First, both are methods of statistical data analysis and modeling. They are also mostly "data driven" approaches, as opposed to physical modeling. One of their main differences is that Geostatistics has a very specific scope of application while Machine Learning is much more general.

During the last decade, many people tried, with more or less success, to use so called "Machine Learning algorithms" to model spatially distributed data [4]. The many contributions of SIC2004 applying such algorithms show that there is still a great interest in this domain. However, if Machine Learning has some advantages over Geostatistics, it may

also have some drawbacks.

3.1. *What are the differences?*

Machine Learning and Geostatistics have many differences, but some have very strong consequences on the usefulness of the methods in various situations.

Above all, there is a great conceptual difference in the way of interpreting data. In Machine Learning, the data samples are assumed to be independently and identically drawn from a single unknown random process. The training procedure tries to model, directly or indirectly, this distribution. In Geostatistics, each data sample is considered as a separate random variable, often assumed to follow a Gaussian distribution, which parameters depend on the rest of the data samples, following some unknown correlation parameters. A great confidence is given to the training data, which are used to find these correlation parameters.

The other big difference, as described before, is that Geostatistics has been developed specifically for spatial data analysis, while Machine Learning is a much more general method, often dealing with thousands of input dimensions, and up to millions of examples, with most of the research focused on classification tasks while Geostatistics is more interested in the estimation of continuous data.

These differences have various consequences on the usage that can be done of Machine Learning algorithms for spatial data analysis.

3.2. *The drawbacks of the Machine Learning approach*

For the "new-comer", the most disturbing aspect of a model created using a Machine Learning approach is that its parameters are often very difficult (or even sometimes impossible) to interpret physically. Machine Learning has been designed to retrieve information, not to interpret it. It is however sometimes possible to produce data using a good Machine Learning model and to fit a physical model on them.

A consequence of the data concept of Machine Learning is that there is no particular confidence about training measures. The main drawback is that the training procedure needs generally more data than a Geostatistical approach would. There exist methods which try to reduce the number of data needed [11] by maximizing the information contained in the training set, but they are only applicable in situations where modeling and sampling are closely related. Another important way to reduce the need for data is to constrain the family of functions with more prior knowledge about the problem to solve, as it is actually done in Geostatistics for spatial data.

Some other consequences of this are a strong smoothing effect in regression applications, and a total lack of robustness when trying to predict data drawn from a phenomenon with only a few or no examples in the training data.

3.3. The advantages of the learning approach

There are three main advantages to Machine Learning. First, it is a very generic approach and can be applied on any kind of data (keeping in mind the limitations described earlier), and to a very large panel of algorithms.

The second advantage of Machine Learning models is their robustness to noise. This is a direct consequence of the data concept and the real strength of the method.

Last but not least, Machine Learning has no constraint on data behavior, apart from the training set to contain the requested information and the expected function to being "smooth". In particular, there is no hypothesis on the shape of the data distribution, which is a very nice property for risk mapping.

3.4. *Summary: when should one use Machine Learning?*

Given the various advantages and drawbacks of Machine Learning, it is better not to use it blindly in all situations.

As presented qualitatively in Figure 1, Machine Learning should not be used with small datasets (i.e. less than 100), and would probably not give better results than other statistical methods with medium datasets (a few hundred data) unless some prior knowledge is available. However, with large datasets (a few thousands and more), Machine Learning is probably the best approach whatever the prior knowledge available.
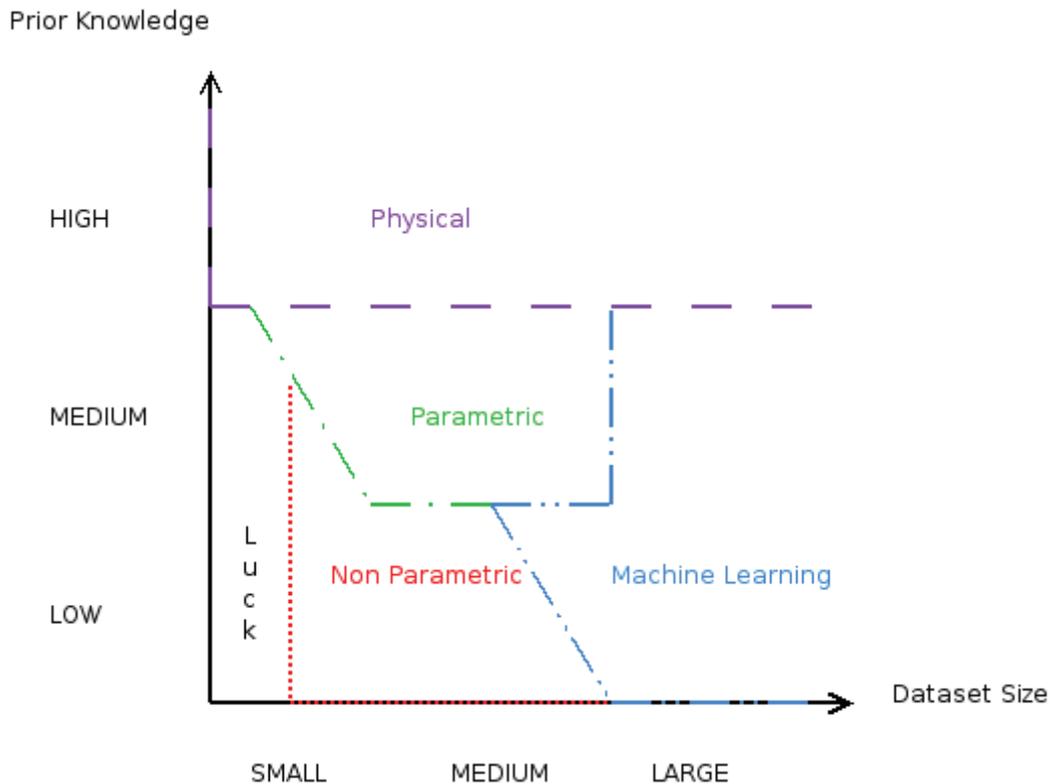
Figure 1: *Qualitative representation of the optimal usage of various modeling methods with respect to the size of the data set and the prior knowledge about the studied phenomenon.*

At the level of applications, Machine Learning is a good approach for all what concerns probabilities (e.g. risk mapping). On such problems, and still given a sufficient number of data, the absence of a priori on data distribution allows dedicated Machine Learning algorithms to be very reliable [8] [7].

Another domain where Machine Learning is a good choice is multivariate analysis. Dimensionality reduction techniques (such as principal component analysis) have benefited a lot from Machine Learning research, allowing complex non-linear transformations of the input space [15].

Finally, another application of interest for environmental mapping, and where Machine Learning research is very active, is time series prediction [1].

Machine Learning is thus applicable in many fields of interest for environmental mapping, provided enough data are available. However, as we will see in the next section, the fact that Machine Learning can be applied does not mean that any Machine Learning algorithm is suited for the job.

4. COMPARISON OF THREE MACHINE LEARNING ALGORITHMS

Even when Machine Learning can be applied, it is very important to use an algorithm that is suited for the application. In this section, using SIC2004 data, we apply Machine Learning methodology to construct models from three different algorithms, and compare their results to Ordinary Kriging.

The three algorithms chosen for this comparison are well known and often used in Machine Learning. Some contributions to SIC97 [4] and SIC2004 used them in different flavors. In the present paper, we only used their simplest versions.

4.1. *K-Nearest Neighbors*

It might sound odd to classify K-Nearest Neighbors (KNN) as a Machine Learning algorithm. Actually, it is the way the capacity parameter k is chosen that allows this label.

The principle of KNN is very simple: Let X be a set of l input data $\{x_1, \ldots, x_l\}$, $x \in \mathbb{R}^n$ and Y be their corresponding output values, $\{y_1, \ldots, y_l\}$, $y \in \mathbb{R}^n$. Let X_x and Y_x be sorted list versions of X and Y, according to their Euclidean distance, in the X space, to x. Thus, let X_x^i be the ith nearest component of X to x, and Y_x^i be its corresponding output. Given some x, the corresponding estimated output is then given by:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} Y_x^i$$

The problem is thus to find the value of k which minimizes the error $(y - \hat{y})^2$ on some unseen data. Hence, k is chosen by cross-validation.

4.2. *Support Vector Regression*

Support Vector Machines (SVM) [17] for classification problems were developed a decade ago. Later, the algorithm was extended to the regression case [16] and named Support Vector Regression (SVR).

For a given set of data $(x_i, y_i)_{1 \le i \le l}$, $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, the simplest linear SVR algorithm tries to find the function

$$f(x) = w \cdot x + b$$

minimizing the quadratic optimization problem

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} Q(y_i - f(x_i))$$

where C controls the trade-off between optimizing the criterion Q and the capacity of the resulting function. In general, $Q(x) = \max\{0, |x| - \epsilon\}$, which corresponds to Vapnik's ε-insensitive loss function, and does not penalize errors less than $\epsilon \ge 0$. After some reformulation and taking into account the case of non-linear regression, the optimization problem is then transformed into the minimization of

$$\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j) + \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

subject to

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0$$
$$0 \le \alpha_i, \alpha_i^* \le C, \text{ for } 1 \le i \le l$$

where the α_i, α_i^* are Lagrange multipliers, solutions of the optimization problem, C is the *soft margin* parameter, weighting the influence of the loss function against the regularization term, and $k(x_i, x_j)$ is a *kernel function*, defining the feature space in which the optimal solution of the problem will be computed in order to handle non-linear problems. In our experiments, we used the popular Gaussian Radial Basis Function (RBF) kernel:

$$k(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

To estimate a new point, we then use the following function f:

$$f(x) = \sum_{i=1}^{N} (\alpha_{s_i} - \alpha_{s_i}^*) k(x, x_{s_i}) + b$$

where the s_i, $1 \leq i \leq N$ are the indices of the data points for which either α_{s_i} or $\alpha_{s_i}^*$ is non zero. Those points are called *support vectors*.

In order to control the capacity of SVRs, one can tune the following hyper-parameters: C, the trade-off between large capacity and small error, ϵ the error we are ready to accept for each estimate, and σ, the bandwidth of the RBF kernel.

4.3. *Multilayer Perceptron*

A multilayer perceptron (MLP) is a particular architecture of artificial neural networks, composed of layers of non-linear but differentiable parametric functions. An MLP for regression can be written mathematically as follows:

$$f(x; \theta) = b + \sum_{n=1}^{N} w_n \cdot \tanh\left(b_n + \sum_{m=1}^{M} x_m \cdot w_{nm}\right)$$

where the estimated output $f(x;\theta)$ is a function of the input vector x (indexed by its M values x_m), and the parameters $\{\theta : w_n, w_{nm}, b_n, b$; with $n \in [1,N]$, $m \in [1,M]\}$. This MLP is thus a weighted combination of N hyperbolic tangents[3] of weighted combinations of the input vector. Given a criterion Q to minimize, such as the mean squared error,

$$Q = \sum_{i=1}^{l} (y_i - f(x_i; \theta))^2$$

between the desired output y_i and the estimated output $f(x_i; \theta)$, for a given training set of size l, one can minimize such criterion using a gradient descent algorithm [14]. This algorithm is based on the computation of the partial derivative of the criterion Q with respect to all the parameters θ of $f(x; \theta)$. Gradient descent can then be performed using

$$\theta = \theta - \lambda \cdot \frac{\partial Q}{\partial \theta}$$

for each parameter θ where λ is the *learning rate*. It has been shown that given a number of hyperbolic tangents N sufficiently large, one can approximate any continuous function using such MLPs [9].

The capacity of MLPs can be tuned in several ways, including the number of hidden units M, the learning rate λ, the number of training iterations, and the *weight decay* γ, a factor that can added to the criterion in order to push the parameters θ to 0 in order to linearize the resulting function[4].

---

[3] Other non-linear but differentiable functions could replace hyperbolic tangents, including the so-called *sigmoid*(x)-1/(1+exp(-x)).

[4] The new criterion then becomes $Q = \sum_{i=1}^{l} (y_i - f(x_i; \theta))^2 + \gamma \sum_j \theta_j^2$.

4.4. *Application to SIC2004*

4.4.1. The Datasets

We used the SIC2004 data sets to evaluate the adequacy of these algorithms to environmental mapping application. Two problems are to be solved. The first one is called the "Natural" one: data are real measurements of daily gamma radioactivity measured all across Germany. The second is called the "Joker": it is a similar set of measurements as in the Natural set to which a simulated radioactive accident was added.

For each of these two data sets, two case studies have been created. This will allow evaluating the influence of the training set size on Machine Learning algorithms performances:

- the "Small" case study: 80 measurements are available to estimate the radioactivity at 808 locations;
- the "Large" case study: the training set contains 808 measurements, and the test set only 200.

For the Large case, we took exactly the same data sets as proposed for SIC2004, but inverted the train and test sets. For the Small case, the 80 points were randomly extracted from the 200 training data of SIC2004. For the Small Joker set, we just made sure that the hot spot data were among the 80 training points.

4.4.2. Experimental Protocol

The context of these experiments is a bit different from the one of the SIC2004 competition as we are not using the spare data sets provided as prior information. For the rest, we tried to stay in the spirit of "automatic mapping", human intervention being limited to choosing the variation ranges of the hyper-parameters of the algorithms.

In practice, this means that the choice of the optimal hyper-parameters was computed via cross validation over the predefined list of values. In the case of Ordinary Kriging (OK), an omni directional variogram model was constructed by least square fitting of a spherical model.

The quality of the models is evaluated using the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), the Mean Error (ME), and the correlation coefficient of predictions versus real values (R). Furthermore, using the fact that the Mean Squared Error (MSE) and the MAE correspond to averages of some real values, we used a statistical test, the *Student's t-test*, in order to assess the statistical significance of the difference between all pairs of results, with a confidence of 95%.

In addition, as we are dealing with environmental mapping and a simulated emergency situation, we expect the models to be able to detect the "Joker spot" and *locate* it correctly, i.e. accurately identify an area of high radioactivity levels. This is why a qualitative observation of the output maps is also part of the quality evaluation process.

4.4.3. The OK Models

As explained before, the OK models are based only on multi-directional spherical variograms models which parameters are presented in Table 1. The fitting was straight-forward in the case of the Natural sets, but for the Joker, it was necessary to remove the outliers from the Joker spot in order to be able to construct the variogram model. The outliers were chosen based on the fact they were lying very far from the main data distribution. Such a procedure is easy to automatize. This ended up with 2 points removed from the Small set and 6 from the Large set. Of course, these points were put back into the dataset for the estimation.

OK	Nugget	Sill	Range
Natural & Small	41.49	377.8	545926
Natural & Large	65.73	474.0	461211
Joker & Small	117.8	469.4	412513
Joker & Large	0.0	1128.6	52653

Table 8: *Parameters of the spherical variogram models for the different data sets.*

4.4.4. The KNN Models

The number of neighbors was chosen between 1 and 20. The optimal value was obtained by cross-validation over the training data and are presented in Table 2.

KNN	K
Natural & Small	4
Natural & Large	8
Joker & Small	11
Joker & Large	3

Table 2: *Hyper-parameters of the KNN models for the different data sets.*

4.4.5. The SVR Models

The three hyper-parameters of the SVR models were chosen by cross-validation. The input and output data of each data set were normalized, with zero mean and unit standard deviation, estimated on the training sets. The range of values for the hyper-parameters were: $C \in \{1, \dots, 1000\}$, $\sigma \in [0.1, 2.0]$, $c \in [0.01, 1.0]$. The optimal values are presented in Table 3.

SVR	C	σ	c
Natural & Small	1	0.5	0.03
Natural & Large	1	0.3	0.03
Joker & Small	10	1.1	0.5
Joker & Large	10	0.1	0.01

Table 3: *Hyper-parameters of the SVR models for the different data sets.*

4.4.6. The MLP Models

The three hyper-parameters of the MLP models were chosen by cross-validation. The input and output data of each data set were normalized, with zero mean and unit standard deviation, estimated on the training sets. The range of values for the hyper-parameters were: number of hidden units (nhu) $\in [1, 40]$, learning rate (λ) $\in [10^{-5}, 10^{-3}]$, weight decay (γ) $\in [0, 0.001]$. The optimal values are presented in Table 4.

| MLP	nhu	λ	γ
Natural & Small	3	0.001	0.001
Natural & Large	35	0.001	0.0
Joker & Small	35	0.001	0.0
Joker & Large	30	0.001	0.001

Table 4: *Parameters of the MLP models for the different data sets.*

4.5. *Analysis of the results*

4.5.1. The Natural Datasets

   Figure 2 shows the two training sets for the Natural data. Both sets are relatively smooth with small variance, as can be observed in Table 5. The highest values are located in the East and South-West of the area, while the smallest are located in the North.
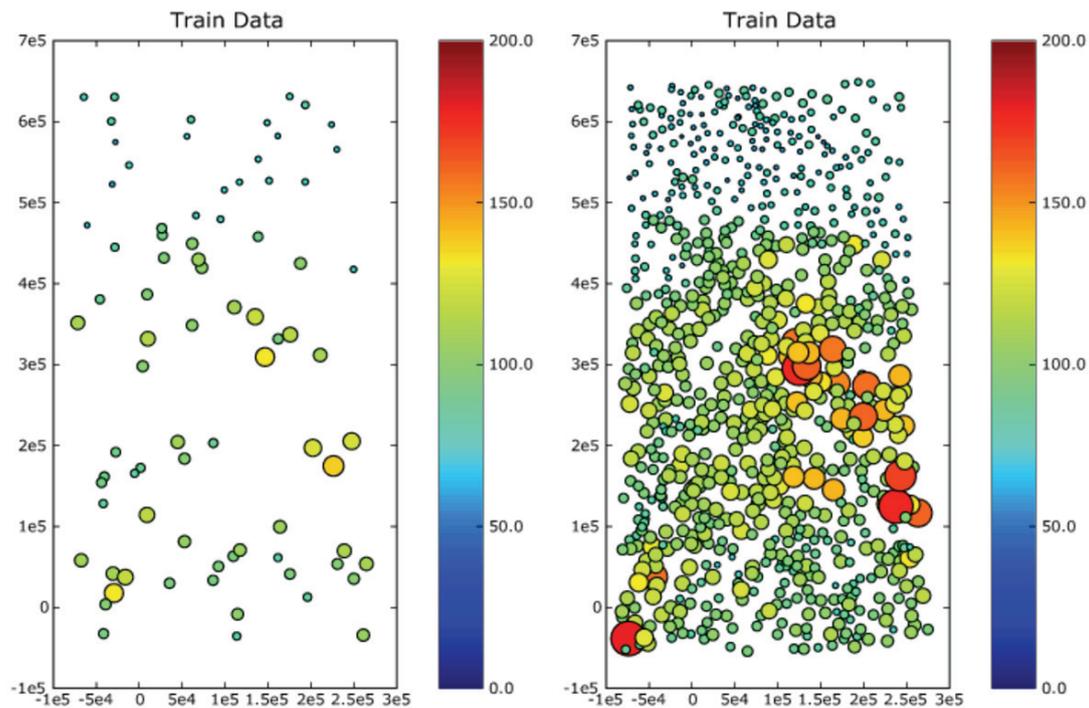
Figure 2: *The Natural Small (left) and Large (right) training data sets. Colors are proportional to radioactivity level at the central location. Point sizes are proportional to the fourth power of the radioactivity level.*

Natural	Min	Median	Max	Mean	Std. Dev.	Skewness	Kurtosis
Small	66.0	96.3	138.0	95.4	17.1	0.3	-0.5
Large	57.0	98.8	180.0	98.0	20.0	0.5	0.7

Table 9: *Natural training data sets summary statistics.*

The data sampling of the Small dataset is not very uniform, especially in the Center-South and Center-East. However, this should not influence too much the prediction results given the smoothness of data.

Results on the Small Set:

   Figure 3 presents the prediction maps of the 4 algorithms (Figures 3b to 3d) with respect to the real data (figure 3a). Visually, it is almost impossible to distinguish the 4 prediction maps from one another. It seems that all the algorithms managed to extract almost exactly the same information from the training set. They are all unable to reproduce exactly the high values from the East and South-West, but the general trend is very well reproduced.
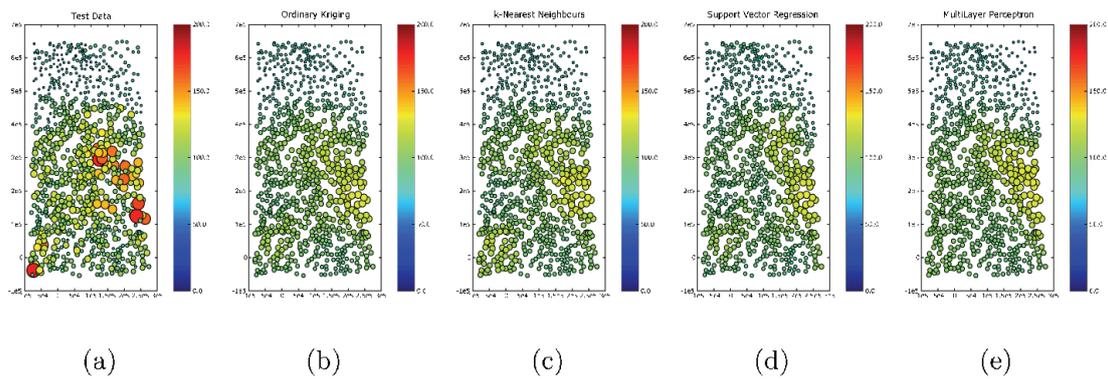
Figure 3: *Small Natural data set estimation maps. Figure a) is the real map. The others are the estimation from: b) Ordinary Kriging, c) K-Nearest-Neighbors, d) Support Vector Regression, and e) Multilayer Perceptron. Color and size scales are the same as in figure 2.*

Natural & Small	OK	KNN	SVR	MLP
RMSE	12.82	13.16	14.36	13.26
MAE	9.41	9.68	10.78	10.09
ME	1.52	1.25	2.63	1.5
R	0.77	0.76	0.71	0.75

Table 6: *Small Natural data set estimation errors.*

The prediction errors given in Table 6 lead to exactly the same conclusion: all the 4 models are performing equivalently, with a RMSE significantly smaller than the standard deviation of the data set. The relatively bad correlation coefficient is the sign that some high values were not correctly estimated. These errors are probably the highest contribution to the RMSE as well. In any case, the Student test did not show any significant difference between the performances.

Results on the Large Set:

The prediction maps of figure 4 look again very similar, although OK and KNN seem to give slightly better predictions in the Eastern area than SVR and MLP.
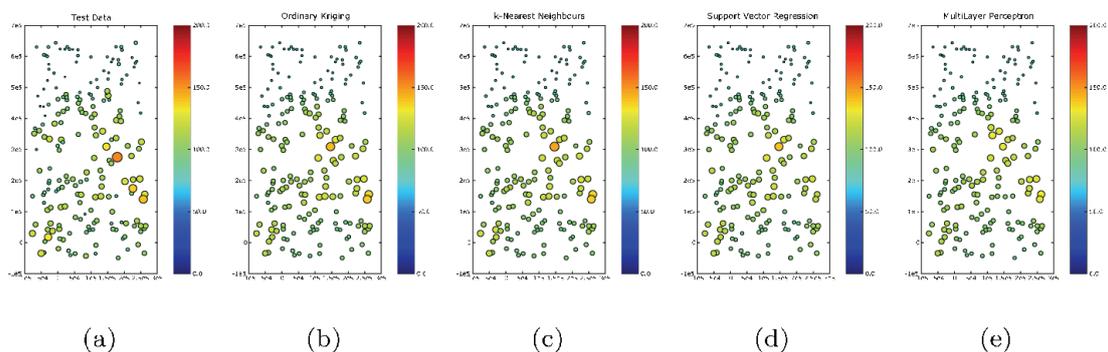
Figure 4: *Large Natural data set estimation maps. Figure a) is the real map. The others are the estimation from: b) Ordinary Kriging, c) K-Nearest Neighbors, d) Support Vector Regression, and e) Multilayer Perceptron. Color and size scales are the same as in Figure 2.*

Natural & Large	OK	KNN	SVR	MLP
RMSE	10.11	10.07	10.32	10.33
MAE	7.65	7.64	7.63	7.81
ME	-1.75	-1.70	-0.9	-0.63
Pearson's r	0.83	0.83	0.82	0.81

Table 7: *Large Natural data set estimation errors.*

Table 7 confirms this similarity between the predictions. The Student test also confirms that there is no significant difference between the models performances.

4.5.2. The Joker Datasets

The Joker data sets look far more complex than the Natural one. Basically, we have a very large flat area, with a very tiny hot spot. More precisely, it is a "hot trail", extending in a West-East direction, as we can see on Figure 5b. As a consequence, the general statistics of the data sets are strongly perturbed (cf. Table 8).
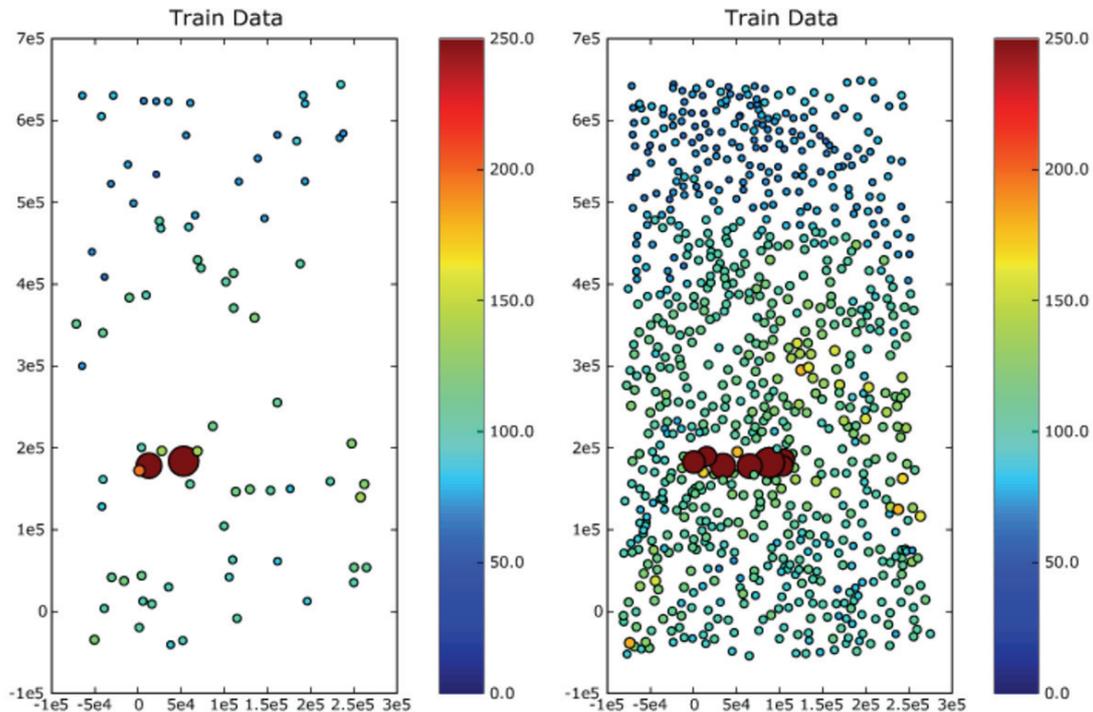

Figure 5: *The Joker Small (left) and Large (right) training data sets. Point sizes and colors are proportional to radioactivity level at the central location.*

Joker	Min	Median	Max	Mean	Std. Dev.	Skewness	Kurtosis
Small	58.2	99.5	1499.0	127.0	190.9	6.4	41.5
Large	57.0	99.0	1528.2	105.4	83.7	11.5	153.8

Table 8: *Joker training data sets summary statistics.*

Results on the Small Set:

Unlike the results for the Natural data set, Figure 6 shows very different prediction maps depending on the algorithm involved. Apart for SVR, all the other models manage to locate correctly the center of the hot spot, OK giving the tightest one. However, none is able to actually reproduce the real shape of the phenomenon, due to a lack of information. Interestingly, MLP and SVR show that they are able to build anisotropic models. Unfortunately, they completely missed the West-East shape of the hot spot.

(a)   (b)   (c)   (d)   (e)
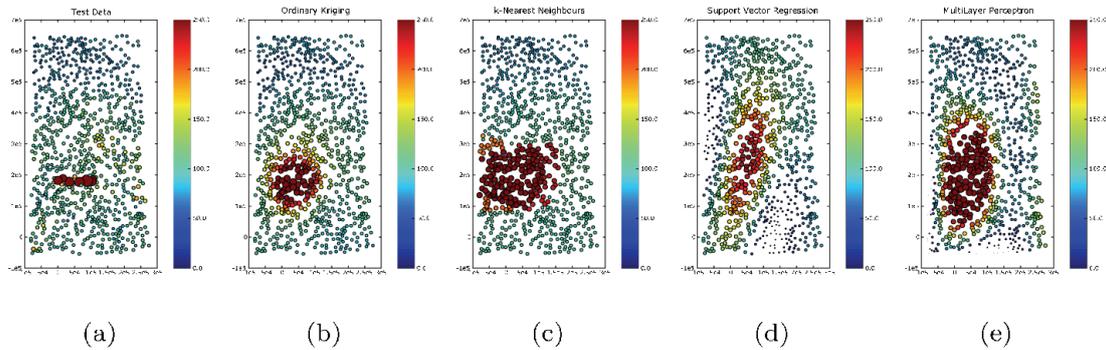
Figure 6: *Small Joker data set estimation maps. Figure a) is the real map. The others are the estimation from: b) Ordinary Kriging, c) K-Nearest Neighbors, d) Support Vector Regression, and e) Multilayer Perceptron. Color and size scales are the same as in Figure 5.*

Joker & Small	OK	KNN	SVR	MLP
RMSE	82.60	109.72	91.32	114.64
MAE	33.20	52.99	49.96	65.89
ME	-14.84	-34.63	-3.85	-27.12
Pearson's r	0.43	0.27	0.18	0.27

Table 9: *Small Joker data set estimation errors.*

Table 9 gives more details on the prediction values themselves. KNN and MLP are the worst predictors on this case study, while OK is significantly better than the rest. SVR is in between, but its very small correlation coefficient is the sign that something has going really wrong in the prediction of the hot spot.

Results on the Large Set:

Figure 7 presents the prediction maps for the Large Joker set. The performances of KNN and OK and SVR in localizing the hot spot are very good. MLP manages to locate the hot spot correctly, but its predictions are too smooth.

<table>
<tr><td>(a)</td><td>(b)</td><td>(c)</td><td>(d)</td><td>(e)</td></tr>
</table>
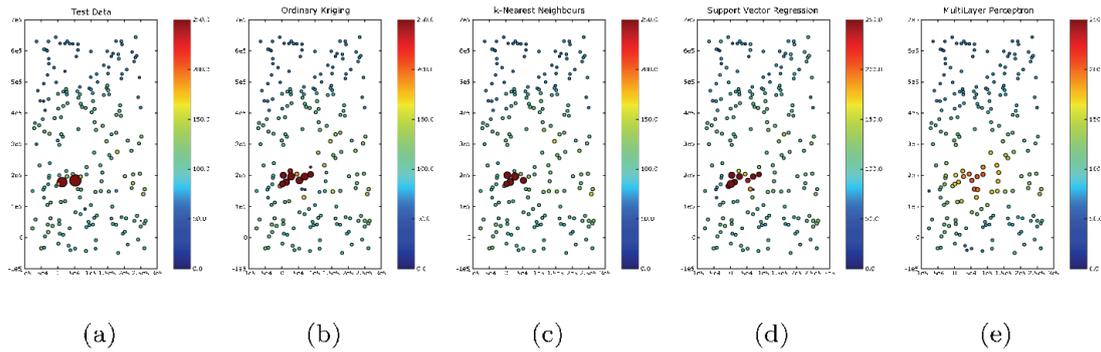
Figure 7: *Large Joker data set estimation maps. Figure a) is the real map. The others are the estimation from: b) Ordinary Kriging, c) K-Nearest Neighbors, d) Support Vector Regression, and e) Multilayer Perceptron. Color and size scales are the same as in Figure 5.*

Joker & Large	OK	KNN	SVR	MLP
RMSE	97.79	96.74	113.69	114.14
MAE	27.16	22.12	28.93	28.29
ME	-3.02	0.27	-1.27	5.61
R	0.60	0.61	0.38	0.36

Table 10: *Large Joker data set estimation errors.*

In the prediction errors of Table 10, OK and KNN are performing equivalently, a little bit better than SVR and MLP. SVR high RMSE is a bit astonishing given its apparently good performance in localizing the hot spot. However, a very careful look at the data shows that it actually predict a little bit off target, which is enough to produce very high errors and destroy the correlation coefficient. Performances of MLP are easier to understand given the strong underestimation of the hot spot values.

4.5.3. Discussion on the experiments

The introduction of the Joker set in SIC2004 was a very good idea. As the experiments on the Natural sets showed, it is almost impossible to distinguish prediction algorithms dealing with smooth data, even with a small data set. However, when an event such as the one simulated by the Joker arises, algorithms are under stress and some discrepancy appears.

First of all, Ordinary Kriging is very well adapted to such applications. This is of course not a scoop, but it is still interesting to notice for non Geo-statisticians. In the context of the Joker, the hot spot was small enough with respect to the studied area to allow its exclusion of the variogram modeling process. This might not always be the case however, and if no variogram is available, then OK is useless. Still, OK was very fast to compute (given the small number of data), and very efficient, both in terms of prediction error and of localization of the hot spot.

For what concerns Machine Learning, we have shown that it can be used for such applications, provided the data set size is somehow proportional to the complexity of the problem. On the smooth Natural data sets, only a few data were necessary to extract some relevant information. On the other hand, Machine Learning did not handle very well the Small Joker set, where the various algorithms were all giving too much importance to the hot spot extension. In such context, the lack of information from data is clearly to blame, while the prior information about the importance of spatial correlation is giving a great advantage to OK.

KNN gave very good performance despite (or thanks to) its simplicity. The poor results it gave on the Small Joker were clearly due to the size of the data set, as its good performance on the Large Joker attests. Most of all, it was never outperformed by SVR or MLP, while being a lot faster to train and use.

SVR relative performances were good on the Natural sets, but much worse on the Small Joker where it had some difficulties to locate the hot spot. On the Large Joker, its performances were a bit worse than KNN and OK, but the localization of the hot spot was good, showing that with enough data, this algorithm is able to handle complex phenomena. Its main drawback of the algorithm is the search for the optimal hyper parameters, which is very slow and thus difficult to do in a time critical situation.

Finally, MLP showed also its dependency to the data set size, but it showed its strong smoothing effect as well. This was particularly obvious for the Large Joker where localization was good, but the hot spot values were strongly under estimated. Algorithms such as MLP are more at ease on noisy datasets or large scale tendencies. Unfortunately, the context of the Joker data set is exactly the reverse. In addition, MLP has the strong drawback of being initialized randomly. The consequence is that two models trained on the same data, with the same hyper-parameters can give different results. This is a cause of high uncertainty on the optimality of the "best" hyper-parameters found.

As a conclusion, we can say that SVR and MLP are probably not the best choice for the type of application proposed for SIC2004. If the predictions on smooth data set were good, they were not better than the ones from OK or KNN, which are much simpler and faster to tune. And when a situation like the Joker arises, even given enough data, SVR was still not better than KNN, and MLP predictions were too smooth.

5. CONCLUSION

In this paper, we presented an overview of Machine Learning, making a strong distinction between the method itself and the applications and algorithms it can be applied on.

By comparing it to Geostatistics, we have shown that it was based on a very different philosophy, which is the source of many interesting properties, but also of some non negligible drawbacks. We have then proposed a qualitative representation of when Machine Learning should be preferred to other methods, with respect to the quantity and type of information available.

In the experiments presented in this paper, we rose the issue of applying "classical" Machine Learning algorithms on problems they were not designed to solve. Hence, we have shown that algorithms such as MLP and SVR were usually not a good choice for environmental mapping applications such as the one proposed for SIC2004. Simpler and faster algorithms such as KNN can give similar or better results for the same amount of data. And if the available measurements allow the construction of a variogram, OK is often much more reliable, whatever the number of data available.

Of course, this does not mean that Machine Learning has nothing to offer to the Environmental Mapping community. Time series analysis and risk mapping are examples of applications where the advantages of Machine Learning can have a significant impact on the quality of the results.

References

1. Aussem, A., F. Murtagh, and M. Sarazin (1995). Dynamical recurrent neural networks: towards environmental time series prediction. *International Journal of Neural Systems*, **6**(2):145–170.

2. Collobert, R., S. Bengio, and J. Mariéthoz. (2002). *Torch: a modular machine learning software library*. Technical Report IDIAP-RR 02-46, IDIAP, 2002.

3. Dubois, G., and S. Galmarini (2005). Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results. *This Volume.*

4. Dubois, G., Malczewski J. and M. De Cort (2003). *Mapping radioactivity in the environment. Spatial Interpolation Comparison 1997*. (Eds). EUR 20667 EN, EC. 268 p.

5. Duda, R. O., and P. E. Hart (1973). *Pattern Classification and Scene Analysis*. Wiley-Interscience.

6. Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, **4**(1):1–58.

7. Gilardi, N., S. Bengio, and M. Kanevski (2002). *Conditional Gaussian mixture models for environmental risk mappin*g. In: IEEE Workshop on Neural Networks for Signal Processing, NNSP, pages 777-786.

8. Gilardi, N., and T. Melluish (2001). *Confidence evaluation for risk prediction*. In: Proceedings of the 7th Annual Conference of the International Association for Mathematical Geology IAMG2001, Cancun, Sept. 6-12, 2001. [CDROM]

9. Hornik, K., M. Stinchcombe, and H. White (2989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**:359–366.

10. Hunter, J. (2005). Matplotlib, http://matplotlib.sourceforge.net

11. Krogh, A., Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (eds.): *Advances in Neural Information Processing Systems*, Vol. 7. MIT Press, Cambridge, MA: 231-238.

12. Minasny, B., A. B. McBratney, and B. M. Whelan (2005). *Vesper version 1.62*. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006, http://www.usyd.edu.au/su/agric/acpa .

13. Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. In: *Proceedings of the IEEE*, **77**(2):257–286.

14. Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning internal representations by error propagation. In: D. E. Rumelhart and James L. McClelland (eds.), *Parallel Distributed Processing*, Volume 1. MIT Press, Cambridge, MA.

15. Schoelkopf, B., A.J. Smola, and K.-R. Mueller (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**:1299–1319.

16. Smola, A. J., and B. Schlkopf (1998). *A tutorial on support vector regression.* Technical Report 30, NeuroCOLT2, October 1998.

17. Vapnik, V. N. *The nature of statistical learning theory*. Springer, second edition, 1995.

Real-time Geostatistics for Atmospheric Dispersion Forecasting, and vice versa?

S. Galmarini

Radioactivity Environmental Monitoring, Institute for Environment and Sustainability, Joint Research Centre, European Commission, Via E Fermi, 21020 Ispra (VA), Italy

E-mail: stefano.galmarini@jrc.it

Abstract: The possible role of geostatistics in real time atmospheric dispersion modelling is analysed and discussed from the atmospheric dispersion modelling view point. Expectations are presented with respect to the possible development of an objective analysis of real-time monitoring data for a rapid understanding of the problem and for a better use of data for atmospheric dispersion model evaluation and amelioration. The paper describes also a real *joker* case occurred in 1998 in Europe and tries to outline what could be done should a objective analysis have been available at that time. In the last section a speculation is presented that relates the possible synergies between geostatistical analysis and atmospheric dispersion modelling.

Keywords: Geostatistics, emergency response, atmospheric dispersion forecasting.

1. INTRODUCTION

Atmospheric dispersion forecasting for emergency response applications has gained increasing attention in the last 20 years, to the extent of becoming one of the vital elements of operational emergency preparedness and response procedures. The use of weather forecast and of numerical dispersion models at various scales for the prediction of the consequences of point or surface releases of harmful substances, can assist the decision making process in understanding the level of severity of the problem and to have a preliminary estimate of the consequences evolution with time. In this context no difference will be made with respect to the scale to which models apply (regional, medium, and short) neither to the release and dispersion case simulated (radioactive, chemical or biological agents) nor on the release causes (incidental, accidental, or intentional). In all such cases the common feature is the real time monitoring and prediction of atmospheric dispersion for emergency response and support to decision making.

As in the case of most environmental disciplines, model results are affected by a number of uncertainties. In the case of atmospheric dispersion modelling these may be summarised into three main categories:

- Uncertainty connected to the atmospheric circulation forecast.

- Those stemming from the modelling approach adopted to model the transport and the dispersion process.

- Under emergency conditions, uncertainties connected to the information on the release characteristics (namely source location, strength and height of the emission).

A priori, the weight of these uncertainties on the model results can be assumed equally important.

A vital element for the increase of model results value and usefulness for emergency response is the availability of real-time monitoring data. The latter can be used for immediate

validation of the model results, data assimilation for the improvement of the simulation, and the reduction of models' intrinsic uncertainties.

Real-time monitoring is performed for a wide range of environmental parameters, from meteorological data to atmospheric concentrations of pollutants and radioactive contaminants. There is no need in this context to outline as an example what is the relevance of the collection of meteorological data for weather prediction and how the former have improved weather forecast over the last decades. The same unfortunately can not be stated for atmospheric pollutants monitoring data which are in most of the cases used for a-posteriori validation studies and seldom used in real-time atmospheric dispersion forecasting.

In this context, geostatistics plays a crucial role as discipline that allows connecting data collection or monitoring to atmospheric modelling or forecasting. The rigorous approaches of geostatistics to the transformation of information sparse in space and time into a coherent picture of a physical phenomenon are well established and proven for an immense variety of applications beyond atmospheric dispersion.

From the atmospheric forecasting view point, however, what is expected is the possibility to use geostatistics in emergency response procedures, i.e. the possibility to analyse objectively a set of monitoring data in a short amount of time and reliable way.

This paper will try to outline the atmospheric modelling perspective on the role that geostatistics could play for the improvement of atmospheric dispersion forecasting and the paper will also speculate on a mutual assistance between the two disciplines.

2. FROM SPARSE INFORMATION TO OBJECTIVE ANALYSIS

R. Daley in his book *Atmospheric Data Analysis* (1991) effectively defines the four-dimensional atmospheric information structure as *vast*, *heterogeneous* and *asynchronous*. Such characterization definitely fits atmospheric variables such as wind speed and direction, temperature, humidity, precipitation and any other quantity that can be derived from those. Presently meteorological networks that constantly monitor the atmospheric status account for thousands of surface stations and radio soundings (over land and sea), aircraft reports, remote sensing data. The measurements collected routinely are efficiently used for assimilation in forecast models, verification and improvements as a reasonable (though never complete), space and time coverage of the atmospheric status is monitored.

The forecast of atmospheric dispersion cannot rely on a comparable amount of information that can be instrumental to the verification and improvement of the dispersion forecast. Atmospheric concentrations of pollutants and contaminants are indeed collected on a routine basis (e.g. air quality and radioactivity monitoring) but they can rarely be used for real-time evaluation of atmospheric dispersion forecasts. The reasons for that reside in:

1. the time required for data acquisition and analysis, quality check, and harmonisation, in other words the rather long time needed to reduce heterogeneity and asynchronicity of the data;

2. depending on the scale of the process in relation to the coverage of a monitoring network, the availability of small amount of measurements does not always allow for the reconstruction of a coherent and complete picture of dispersion process.

With specific reference to (2), geostatistics can definitely play an important role in trying to generalise the information collected by few monitors and to build a coherent picture of the event, provided however that the procedures adopted can be performed in real-time conditions. In this respect the applicability of geostatistics to real-time monitoring data of atmospheric dispersion that can subsequently be used for modelling applications, reduces to the distinction between *subjective* and *objective analysis*.

Atmospheric dispersion models are systems that work on numerical grid basis that is used for the solution of transport and dispersion equations or as a fixed system of reference over which the results of lagrangian calculations can be projected. The aspect of regularly spaced values and surface or volume representativity of the grid nodes is a crucial element for this discussion.

While several robust and extensively tested approaches are described by geostatistics to transfer sparse network information to regular grid distribution, little is known on the possibility to use the latter regardless of an expert (*subjective*) judgment and in an "automated" (*objective*) fashion. It should be clear that "automated" in this context does not exclude the role of an expert judgment and does not refer to the development of machine-based tools that should be working on their own. It more specifically relates to the possibility of using a procedure that in a limited amount of time can provide the best reconstruction of, for example, a concentration field and that can work with little a-priori information on the nature and scale of the event. While objective analysis is a well established procedure for the use of meteorological data in atmospheric circulation forecast or nowcast, no information is available on comparable approaches for atmospheric dispersion of pollutants at various scales and in particular for emergency response.

3. OBJECTIVE ANALYSIS FOR EMERGENCY RESPONSE

It goes without saying that during emergency conditions, should a monitoring network identify the exceedance of a specific variable background value (e.g. atmospheric concentration, dose) no time is available for a subjective analysis which by any means is the most accurate one can hope to obtain (though accuracy is in turn a function of the time invested in the analysis).

What an atmospheric modeller wishes to obtain from the analysis is essentially a clear indication of:

- the spatial scale of the process;

- the best possible reconstruction of the dispersion field that goes beyond the local measurements produced by the network.

Indeed point measurements can still be used to validate a preliminary model result by interpolating model data to the monitoring location (e.g. Goodin *et al.*, 1979) but yet the model simulation cannot be performed accurately if no indication on the process scale is available which consequently can provide hints on the possible source of the pollutant. If data from a monitoring network is the only information available, namely no indication on the source location is provided, nor on the amount of mass released and the time/duration of the release, little can be done from the atmospheric dispersion modelling view point a part from sensitivity analysis on scenarios. Although source-receptor models could be used together with backward trajectories in order to identify a possible origin of the pollutant or the contaminant and other techniques are available for the source term reconstruction (e.g. Issartel and Baverel, 2003; Hurdin and Issartel, 2000; Robertson, 2004), a rigorous analysis from the geostatistical view point becomes the very first assessment needed to investigate the data and to have a first estimate of the nature and the scale of the anomaly. The use of single point measurements could be limited by the number of non-zero values available while a more general reconstruction of the spatial distribution for locations that are not covered by the monitoring network can be very useful for the application of the abovementioned techniques. The availability of a reconstructed field can trigger immediate response from the modelling point of view. For example, a concentration field reconstructed form a rigorous analysis can subsequently be advected and dispersed by means of a dispersion model in order to determine its time evolution from that location (e.g. Galmarini *et al.*, 1993) while more detailed information are gathered on the actual causes of the emission.

The *joker* dataset presented for the analysis of SIC2004 (Dubois and Galmarini, 2004, this volume) was in fact intended as an example in this respect. The field presented in the *joker* is a sub sample of a pseudo dispersion process of Gaussian type that mimics a short time emission of a tracer. Should a dataset of the form of the *joker* be available one may wonder:

- Are there methodologies to perform a scientifically sound analysis in order to generalize the information obtained from few monitors?

- What can we learn about the process from the results produced by the analysis?

- How much time is needed to perform the analysis?

- To what extent can it be considered objective in the sense described above?

- Can the information produced by the analysis be used by dispersion models for validation and data assimilation?

The results obtained from the SIC2004 exercise clearly show that a "surprise factor" has played a role in the analysis of the *joker* case. All the training datasets as well as the main datasets where coherent and representing a similar situation. Algorithms were fine tuned on those datasets and in some cases they worked satisfactorily also on the *joker*. However, to what extent can these results be generalised and those algorithm work for any other *joker* case?

3.1 No joker, this is real

If the case represented by the *joker* would sound too theoretical or academic, a similar but real situation occurred in 1998 in Europe.

On June 9, the Swiss authorities announced that radiation levels reaching thousand times background were detected by their national monitoring network. At that time no information was available on a release of radioactive gases anywhere in Europe. Similar exceedance were also detected and announced by France and Italy.

Figure 1: *Location of the 1998 accidental release of ^{137}Cs (Algeciras, South of Spain)*
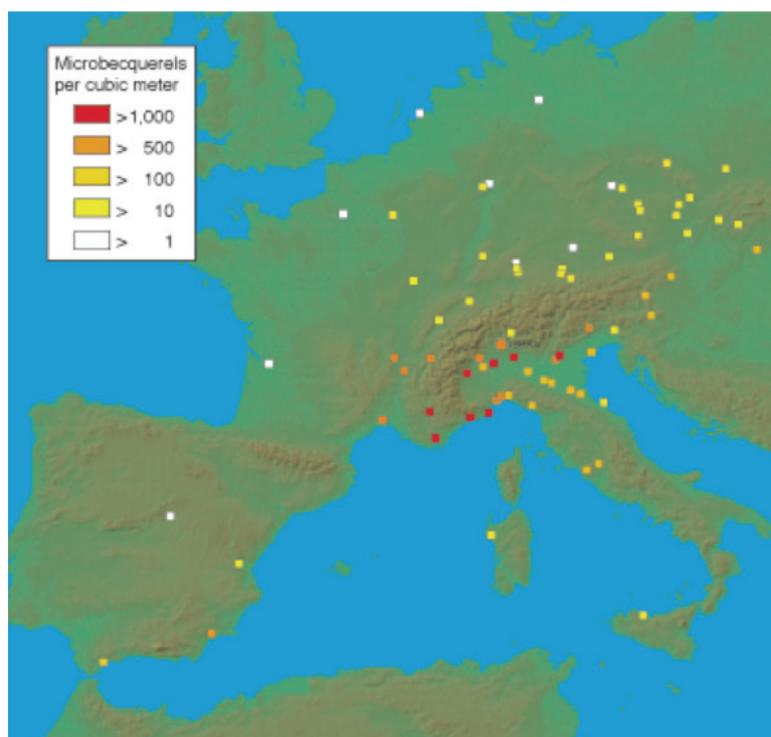
Figure 2: *Monitoring locations detecting the Algeciras release. (From Basket et al., 1999, courtesy of the author)*

On June 10 the Spanish Nuclear Security Agency received information from the Acerinox steel mill near Algeciras (Figure 1) which notified that radioactivity was detected in one of its plants. However the national agency, had not observed elevated radiation levels from its network. On June 12, the source of the release was identified as a medical radiotherapy device containing ^{137}Cs that was processed in the steel mills scrap metal furnace.

The amount of mass released and the time of the release, were unknown, but the incident was thought to have taken place during the last week of May. On the same day, the International Atomic Energy Agency also announced the incident and speculated on its possible connection to elevated levels of ^{137}Cs detected in southern Europe.

The data collected in various countries were obtained with different samplers, different averaging times (ranging from 1 to 14 days) and different radiological sensitivity thresholds.

After a detailed harmonization of the monitoring data (Figure 2) and the performance of a number of model simulations, a scenario of the possible dispersion event was reconstructed (Vogt *et al.*, 1998; Yamazawa, 1998; Pobanz *et al.* 1999; Buckley, 1999; Robertson, 2004). The modeled wind field over the Mediterranean basin was obtained from general circulation models, mesoscale models and WMO meteorological stations measurements (Figure 3 a-d, Vogt et al, 1998).

A sensitivity analysis was performed over a number of release conditions by means of dispersion models in order to reproduce the measured concentrations (Figure 4) and to obtain an estimate of the unknown source term. The release was finally estimated to 8 - 80 Ci over a few hours period in the morning of 30 May (00 - 03 UTC)[5].

---

[5]The health effect was concluded to be well below exposure limits.
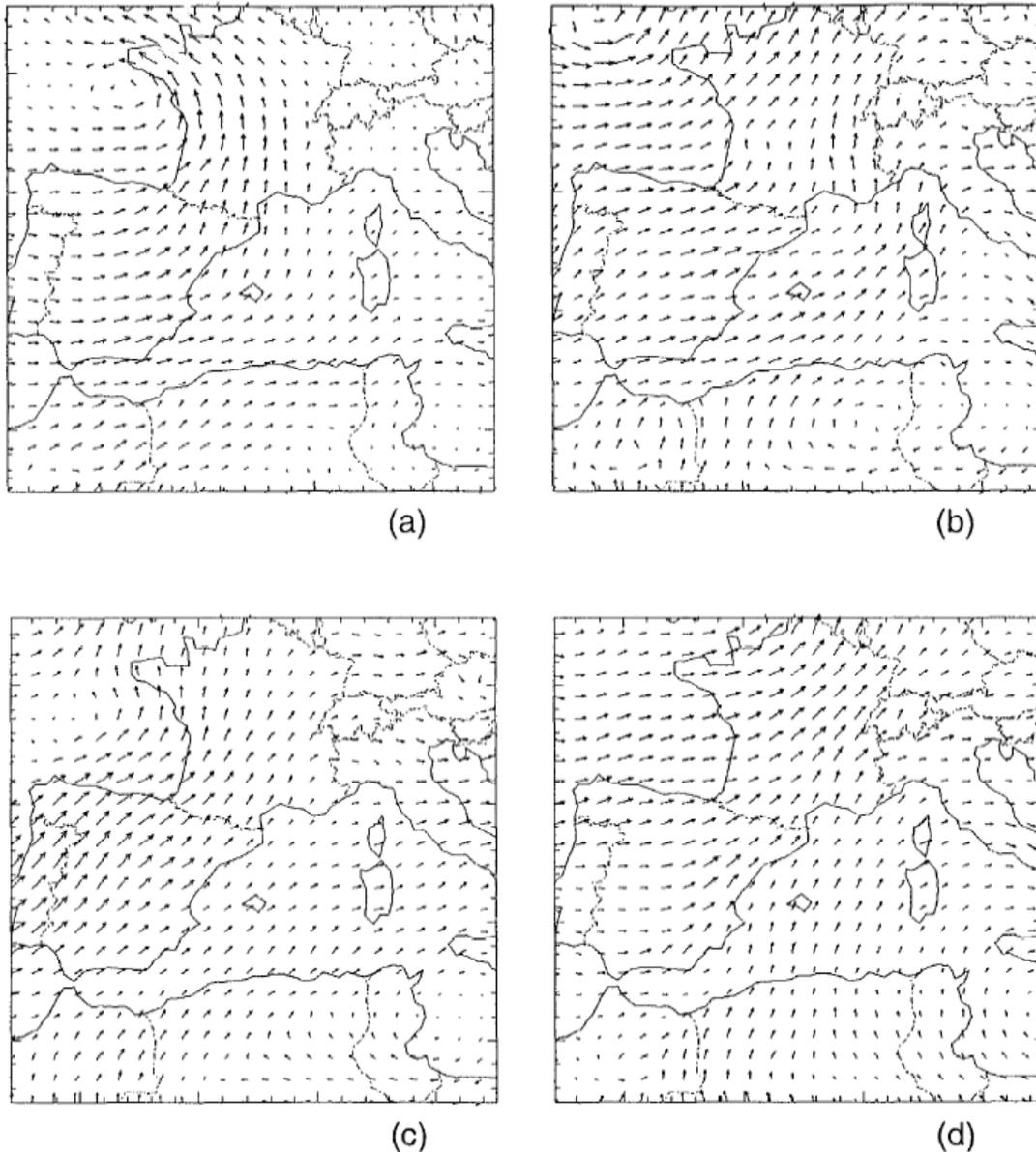
Figure 3(a-d). *Set 3 model wind vectors at 1500 m above ground at 0130 UTC for (a) 30 May, (b) 31 May, (c) 1 June, and (d) 2 June 1998 (from Vogt et al., 1998).*

Given the scenario presented one may wonder what an objective geostatistical analysis could have produced and how it would have helped in reducing the time to reconstruct the scenario.

The Algeciras release is particularly complex due to the fact that the cloud of ^{137}Cs traveled unnoticed for few days over the Western Mediterranean Sea before reaching the monitoring network in Southern France, Switzerland and Italy as shown in Figure 4. The radioactivity levels monitored on the eastern coast of Spain were most probably not considered anomalies and they did not trigger an early alert.

The example presented is particularly interesting as it outlines quite clearly what are the bottlenecks in the process of real-time reconstruction of a scenario and the amount of delay that can be accumulated before a clear idea of the actual situation is available.

Namely:

1. Data collection from different countries (this depends on the scale of the event it could also relate to a smaller scale than the one presented and therefore a single country);

2. Data harmonization in terms of sampling methodology, sampled variable, statistical treatments of the measured value (e. g. averaging times, threshold values);

3. Field reconstruction, scale identification;

4. Source location identification;

5. Source term estimate.
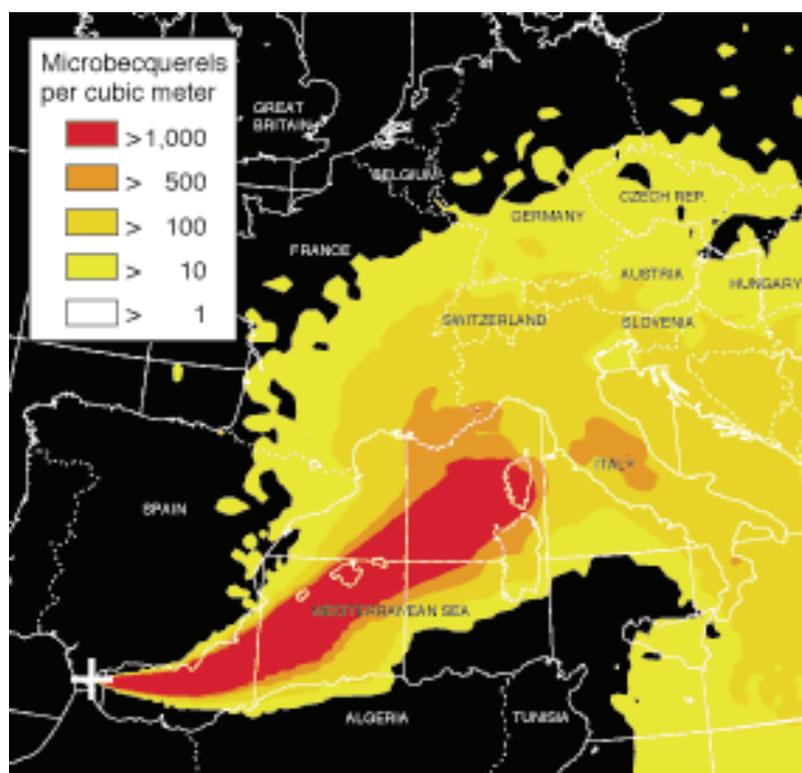
Figure 4. *Dispersion model simulation of the Algeciras release by the LLNL ARAC system. Average air concentration over 7 days from release. (From Basket et al., 1999, courtesy of the author)*

Points (1) and (2) have to be solved at the level of international collaboration and adoption of common monitoring standards. For radiological surveillance and monitoring in Europe, initiatives in this respect have been and are being taken by several European Commission activities and projects (e.g. EURDEP, AIRMON) as well as by European Union legislation (Euratom Treaty). Point (3) could be addressed by geostatistics should a sound objective analysis procedure be identified, and (4) and (5) by atmospheric dispersion modeling.

5. ATMOSPHERIC DISPERSION MODELLING AND GEOSTATISTICS, POSSIBLE SYNERGIES

At this stage of the discussion another important issue should be tackled in order to partly lift the burden of a very complicated task from geostatistics and to try to explore whether atmospheric dispersion modelling could assist synergistically the former towards the performance of the objective analysis.

The fundamental concept behind geostatistical analysis of the data is self-consistency between the information available and the one reconstructed. Statistical treatments are put at work in order to preserve the information available from the monitoring network and to generalise it to neighbouring points towards the reconstruction of the alleged field. In spite of the sophisticated techniques developed for geostatistical analysis, the starting point is always the information available which in some cases can be insufficient and limiting. This is very much connected to the relationship between the scale of the process and the scale of the monitoring network. The issue can become very tricky in this respect as networks constructed for large scale events will miss the early stages of development of the process while small scale networks will be saturated with extreme values in case of large scale events. Similarly atmospheric processes develop in three-dimensional space and networks for contaminants are always at the surface thus providing a limited perspective on the process. The geostatistical analysis potentials are therefore dimmed from start by the incompleteness of the information available.

One could however envisage the performance of a geostatistical analysis that takes into account physical aspects of the process. We refer more specifically to an analysis that is not only based on the measured information available from the monitoring network and how it relates to physical space but it considers additional elements such as the time evolution and the physics of the process that has generated the measured data. One example is connected to the role of the vertical structure of a dispersing cloud, the fact that monitoring data relate to surface measurements and that geostatistical techniques only rely on that information for the field reconstruction. The vertical structure of the atmosphere in the vicinity of the ground is governed by precise laws that define the depth of the so-called atmospheric boundary layer (ABL), namely the turbulent atmospheric layer responsible for the vertical distribution of a tracer. The question that can be addressed is for example, to what extent a parameter like the boundary layer depth could be included as binding element in a geostatistical analysis for the sake of a better reconstruction of a concentration field? The space and time evolution of the ABL significantly influence the concentration levels at ground and could be determined on the basis of available meteorological measurements or models. Can a geostatistical analysis be envisaged that includes physical constraints to the problem?

This is just a simple illustration according to which physical principles governing the process could provide an improved analysis of the monitoring data. More ambitiously, can one envisage the development of a geostatistical analysis that takes into account the very clear wind pattern presented in Figures 3(a-d)?

The counterpart of this approach is obviously data assimilation in the sense of using monitoring network data to bind the solution provided by a dispersion model. However the emphasis in this case is more on the modelling approach than on the role of the data. Several parameters are still needed in order to configure a model simulation, and several groups of parameters can lead to equivalent solutions in spite of the minimization of the model to observation error. No indication would yet be available to bind the solution in the vertical direction thus allowing for a large family of solutions to be equally probable. A physics-based geostatistics would however keep the data as central element of the analysis and make use of additional information for the sake of the improvement of the field reconstruction.

By continuing to speculate one could envisage a multiple level approach like the one of Figure 5. The figure depicts possible synergies between geostatistics (1), atmospheric modelling (2), and monitoring (3). From the most obvious and straightforward, data analysis

(5) to model validation and data assimilation (6), and the one described above, physics-based geostatistics (4).

All the synergies aim toward what can be described as *data simulation*. In particular if (4) could be developed within an objective analysis for emergency response, the result could represent the very first tool available for a preliminary screening of the situation that provides quantitative indications on the subsequent adoption of more sophisticated tools.

Some examples of this type of analysis exist although very much confined to the research activity and far from a practical implementation.
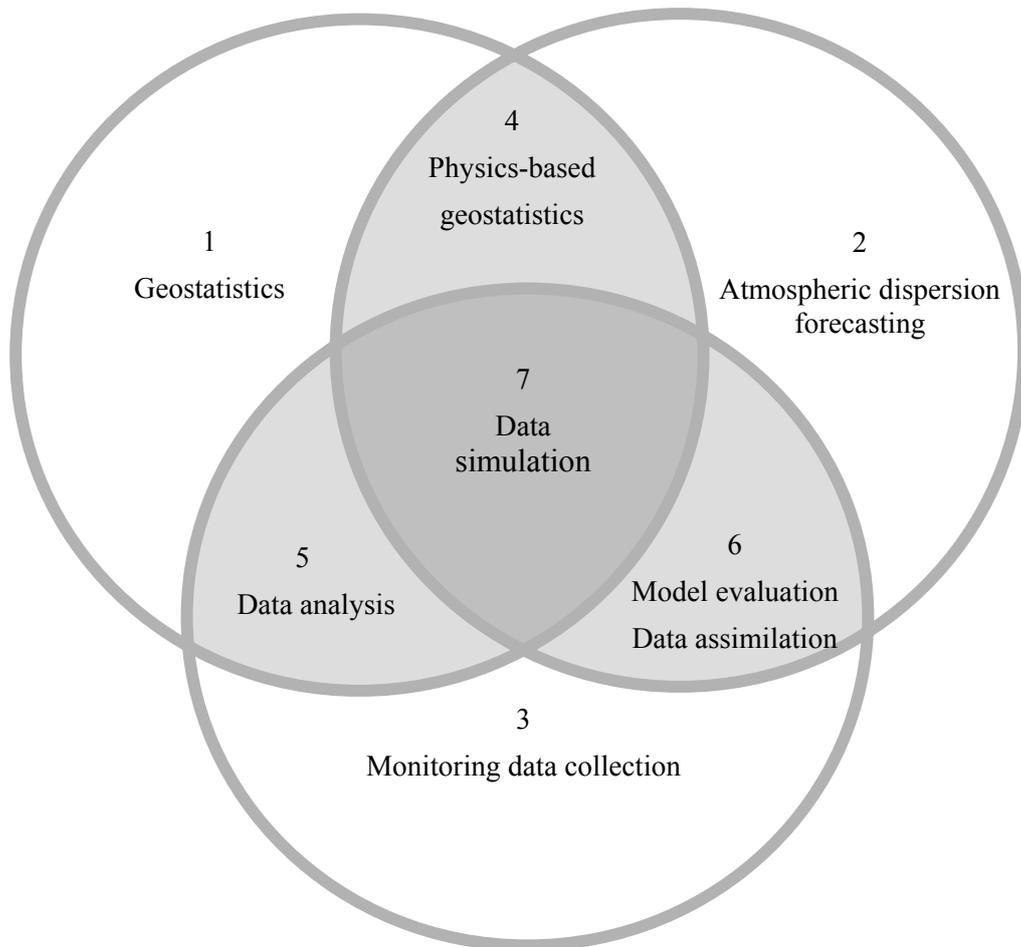
Figure 5. *Schematic representation of the synergies between monitoring data, geostatistics and atmospheric dispersion modelling.*

6. CONCLUSION

The paper presents a discussion on the possible role of geostatistics for the real-time analysis of atmospheric dispersion variables. No specific reference is made to the scale of the process or the type of pollutant. The case considered is the atmospheric dispersion of undefined pollutant that disperses over a monitoring network. No a-priori indication is available on the scale of the process, the amount of mass released and the location of the source. The only information available is the measurements from real-time monitoring.

Particular emphasis is give to the necessity of rapid and objective procedures that would allow the reconstruction of a concentration field and the determination of the scale of the problem. The atmospheric dispersion modelling perspective on what is expected from a real-time objective geostatistical analysis was presented.

The Algeciras accidental release of radioactive gas happened in 1998 was presented as real counterpart of the *joker* case studied in SIC2004. The case presented was used to outline the major bottlenecks in real-time atmospheric dispersion monitoring, data analysis, and modelling.

In the last section the issue of physics-based geostatistical analysis was presented where the synergic use of geostatistics and atmospheric dispersion models is proposed as a possible approach to the problem.

In this respect we propose that a future SIC exercise could be dedicated to this specific aspect for the sake of exploring and stimulating more research along this line.

References

1. Baskett R. L.(1999). Forewarnings of Coming Hazards, S&TR, 4-11

2. Buckley R. L. (1999). Modeling Atmospheric Deposition from a Cesium Release in Spain Using a Stochastic Transport Model, WSRC-MS-99-00660

3. Daley R., (1991). *Atmospheric Data Analysis*, Cambridge Atmospheric and Space Science Series, Cambridge University Press, pp. 457

4. Dubois G. and S. Galmarini, (2005). Spatial Interpolation Comparison (SIC) 2004: introduction to the exercise and overview of results. *This volume*.

5. Galmarini S., Grippa G., De Cort M., and Graziani G., (1993). A system for using the air radioactivity measurements in a long range model to forecast cloud evolution, *Radiation Protection Dosimetry*, 50, Nos 2-4, 339-341

6. Goodin W., R., G. J. McRa and J. H. Seinfeld. (1979). A Comparison of Interpolation Methods for Sparse Data: Application to Wind and Concentration Fields. *Journal of Applied Meteorology*: Vol. 18, No. 6, pp. 761–771.

7. Hourdin F., Issartel J.-P. (2000). Sub-surface nuclear tests monitoring through the CTCT xenon network. *J. Geoph. Res. Lett,* 27, 15: 2245-2248.

8. Issartel J.P., Baverel J. (2003). Inverse transport for the verification of the Comprehensive Nuclear Test Ban Treaty. *Atmos. Chem Phys.* 3: 475-486.

9. Yamazawa, H., (1998). Long-range Dispersion Analysis on Accidental Atmospheric Release of Cesium-137 at Algeciras, J. Nucl. Sci. Technol., 41, 114-116. *(in Japanese)*

10. Robertson L. (2004). Extended back-trajectories by means of adjoint equations, SMHI reports Meteorology and Climatology No 105

11. Pobanz B.M., P.P. Vogt, F.J. Aluzzi, J.C. Pace, (1999) Comparison of Gridded Versus Observation Data to Initialize ARAC Dispersion Models for the Algeciras, Spain Steel Mill Cs-137 Release, UCRL-JC-131200, Lawrence Livermore National Laboratory, Livermore, California.

12. Vogt P. J., B. M. Pobanz, F. J. Aluzzi, R. L. Baskett, and T. J. Sullivan, (1999). ARAC simulation of the Algeciras, Spain steel mill CS-137 release, 51 pp., UCRL-JC-131330, Lawrence Livermore National Laboratory, Livermore, California.

European Commission

EUR 21595 EN - Automatic mapping algorithms for routine and emergency monitoring data

Editor: G. Dubois

Luxembourg: Office for Official Publications of the European Communities

2005 – 149 pp. – 21.0 x 29.7 cm

Scientific and Technical Research Series

ISBN 92-894-9400-X

Abstract

The Spatial Interpolation Comparison (SIC) 2004 exercise was organised during the summer 2004 within the frame of the activities of the Radioactivity Environmental Monitoring (REM) group of the Institute for Environment and Sustainability at the Joint Research Centre (JRC) of the European Commission. Its purpose was to assess the current know-how in the field of "automatic' or "real-time" mapping. The underlying idea was to explore the way algorithms designed for spatial interpolation can automatically generate maps on the basis of information collected regularly by monitoring networks, this in situation of routine or emergencies. This EU report presents a collection of essays written by experts in the field of geostatistics and machine learning about the results obtained by the researchers who participated to SIC 2004. The report should provide the reader with a useful overview of the current state of the art in the field of automatic mapping of environmental variables.

The mission of the Joint Research Centre is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre