INTRODUCTION TO THE SPATIAL INTERPOLATION COMPARISON (SIC) 2004 EXERCISE AND PRESENTATION OF THE DATASETS

G. Dubois

Radioactivity Environmental Monitoring, Emissions and Health Unit, Institute for Environment and Sustainability, Joint Research Centre, European Commission, Italy
Correspondence to: gregoire.dubois@jrc.it

S. Galmarini

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

The Spatial Interpolation Comparison (SIC) 2004 exercise was organised during the summer 2004 to assess the current know-how in the field of “automatic 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. Participants to this exercise were invited to use some prior information to design their algorithms and to test them by applying the software code to two given datasets. Estimation errors were used to assess the relative performances of the algorithms proposed. Participants were not only invited to minimize estimation errors but also to design the algorithms so as to render them suitable for decision-support systems used in emergency situations. The data used in this exercise were daily mean values of gamma dose rates measured in Germany. This paper presents the exercise and the data used more in detail.

1. INTRODUCTION

1.1 SPATIAL INTERPOLATION COMPARISON EXERCISES

The Spatial Interpolation Comparison 2004 (SIC2004) exercise follows that organized in 1997 (SIC97; published in EC, 2003), which was inspired by a work of Englund (1990). The common denominator of these exercises is to provide participants with a subset of n measurements taken from N observations of a variable X and to ask them to estimate the values assumed by the variable at N-n locations. This concept is summarized in Figure 1. The real values observed at these N-n sampling places are revealed at the end of the exercise in order to assess the relative performances of the proposed algorithms.

Because Englund’s exercise in 1990 was limited to geostatistical techniques, it was felt that a new one involving any type of spatial interpolation function would have generated more applicable results. SIC97 was therefore set up; its outcome confirmed (see for example Bucher and Včkovski, 1995; EC 2003) the large impact of human factors in the decision process that leads to the choice of an interpolation function and to its associated parameters. As a result, even if all the participants faced the same problem to be solved, no two maps were identical among the 36 results submitted (Dubois and Shibli, 2003). The difficulty in interpolating spatial data is further aggravated when little freedom is given in exploring the data prior to deciding what algorithm to choose and what values to attribute to the parameters involved. This is clearly the case with systems that generate maps automatically and will be the topic covered by this spatial interpolation exercise.

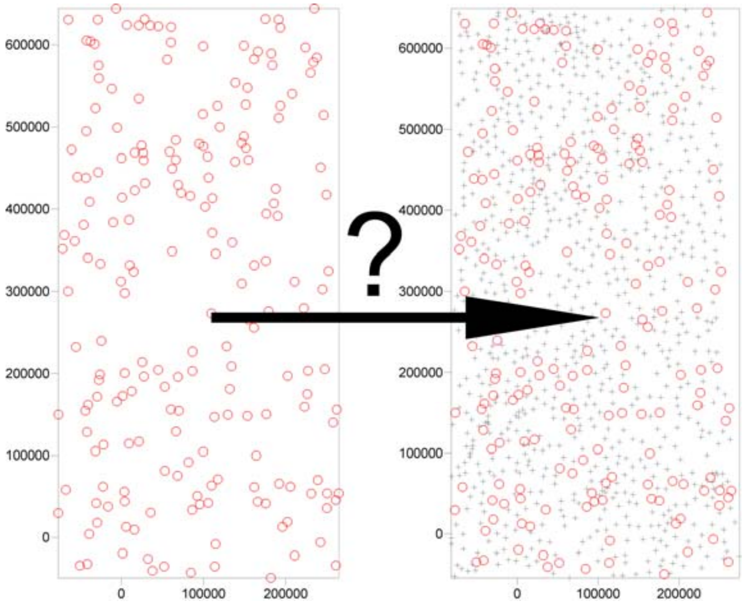

Figure 1
Description of the Spatial Interpolation Comparison (SIC) concept: participants are invited to minimize errors when using n observations (left, circles) to estimate values located at N locations (right, crosses). The left map shows the sampling locations of 200 training data and the right map shows 1008 point locations for which estimated values were requested

1.2 AUTOMATION OF THE SPATIAL INTERPOLATION STEP

Geographic Information Systems (GIS) are frequently linked to automated monitoring networks that collect measurements of various environmental variables in real time. For example, ozone, radioactivity, seismic activity, meteorological data are a few of these variables that need to be interpreted automatically and summarized through maps. Obviously, it is infeasible to have experts prepare new maps every 10 minutes, say, so one needs to automate the interpolation process. It is the purpose of SIC2004 to explore the approach taken by experts in designing such automatic mapping systems.

This need for further research in the field of GIS automation is particularly justified since many automatic networks do regularly report observations that are critical to the human environment. In practical terms, in the case of emergencies one would expect to have mapping systems that generate reasonable results, that are computationally efficient and that deal properly with so-called “outliers” of extreme events (e.g. high ozone levels, nuclear incidents or accidents, earthquakes, storms). To this end, SIC2004 emulated the design from the previous exercises but also involved a second dataset to better underline problems encountered in the case of extreme events.

1.3 COMPUTATIONAL OBJECTIVES OF SIC2004

The framework in which participants had to design their algorithms has been defined by the following computational objectives. The interpolation function should generate results that are:

- obtained in a minimum amount of time;
- "reasonable", that is the estimated errors should be kept to a minimum;
- and computed without any human intervention, in the sense that only manual downloading and uploading of data files were allowed.

The generation of information about the associated estimation uncertainties was strongly encouraged but nevertheless not mandatory.

To assess the reliability of the interpolation function, it was requested to minimize the root- mean-squared error (*RMSE*) of the estimated values computed at the *n* locations. The mean absolute error (*MAE*) and the bias (or mean error, *ME*) also needed to be reported. These errors are defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|a_i^* - a_i\right|,$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(a_i^* - a_i\right),$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i^* - a_i)^2},$$

where a^*_i is the estimated value at location *i* and where a_i is the true value. Pearson's correlation coefficient, *r*, between the estimated and the true values was also requested from the participants.

Another essential parameter that participants had to take into account was that the algorithms had to be able to deal with extreme events that would affect the statistical portrait of the monitored phenomenon. This may happen if stations malfunction or if accidents happen, that is, when measured values exceed background values by far.

2. PRIOR INFORMATION

2.1 DATA SOURCE

In order to allow participants to design their algorithms properly, information about the investigated variable and about the topology of the monitoring network had to be provided in advance. Two types of data were provided:

1. coordinates and observations made of the variable *X* at *n* fixed locations, and
2. geographical coordinates only of the *N-n* locations at which estimations calculated by the mapping algorithm are required.

The data used for SIC2004 are measurements of gamma dose rates that have been extracted from the EUropean Radiological Data Exchange Platform (EURDEP) database (De Cort and de Vries, 1997). EURDEP is a system developed by the Radioactivity Environmental Monitoring (REM) group (Institute for Environment and Sustainability, Joint

Research Centre of the European Commission, http://rem.jrc.cec.eu.int/) to make European radiological monitoring data available to decision-makers with a frequency that is close to a real-time. From this database, 10 sets of mean daily values collected during 2003, roughly one day randomly drawn from each month, were selected. 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 that country, 1008 were selected by drawing a rectangular window (their relative locations are shown in the right part of Figure 1). These stations were common to each of the 10 datasets and all reported values for each day selected. The data were distributed in text files, more precisely in comma-separated value (.csv) format in which the first column held an integer value used as a unique identifier for each station, the second and third column indicated the relative x and y coordinates (in metres) and the last column gave the gamma dose rate measured in nanoSievert/hour (nSv/h). An example of the first six lines (corresponding to six measurements) of such a file is:

```
13,99554,598199,76.9
21,60497,621464,74.7
30,99102,515625,72.2
31,119587,499850,74.0
35,116784,525378,74.5
43,110576,413471,113.0
```

2.2 DATA STATISTICS

TRAINING SETS: 200 INPUT DATA

From these 1008 monitoring locations, a single sampling scheme of 200 monitoring stations was selected randomly and extracted for each of the 10 datasets. These data were distributed to the SIC2004 participants to allow them to train and design their algorithms. The 200 sampling locations have a spatial distribution that is nearly random (Clark and Evans, 1954). Table 1 gives some statistics for nearest-neighbour distances. Table 2 and Figure 2 further summarise the statistics of the observations reported in 10 datasets, and the left part of Figure 1 shows the relative locations of the 200 stations.

NETWORK FOR ESTIMATED VALUES: 808 OUTPUT DATA

The remaining 808 (that is, 1008-200= 808) stations were used to define the network topology where estimated values would be requested. Table 1 gives some statistics for nearest-neighbour distances for these remaining stations. Thus only the coordinates of the remaining 808 monitoring stations were disclosed to the participants.

Figure 2 and Table 3 show exhaustive statistics for the 10 datasets.

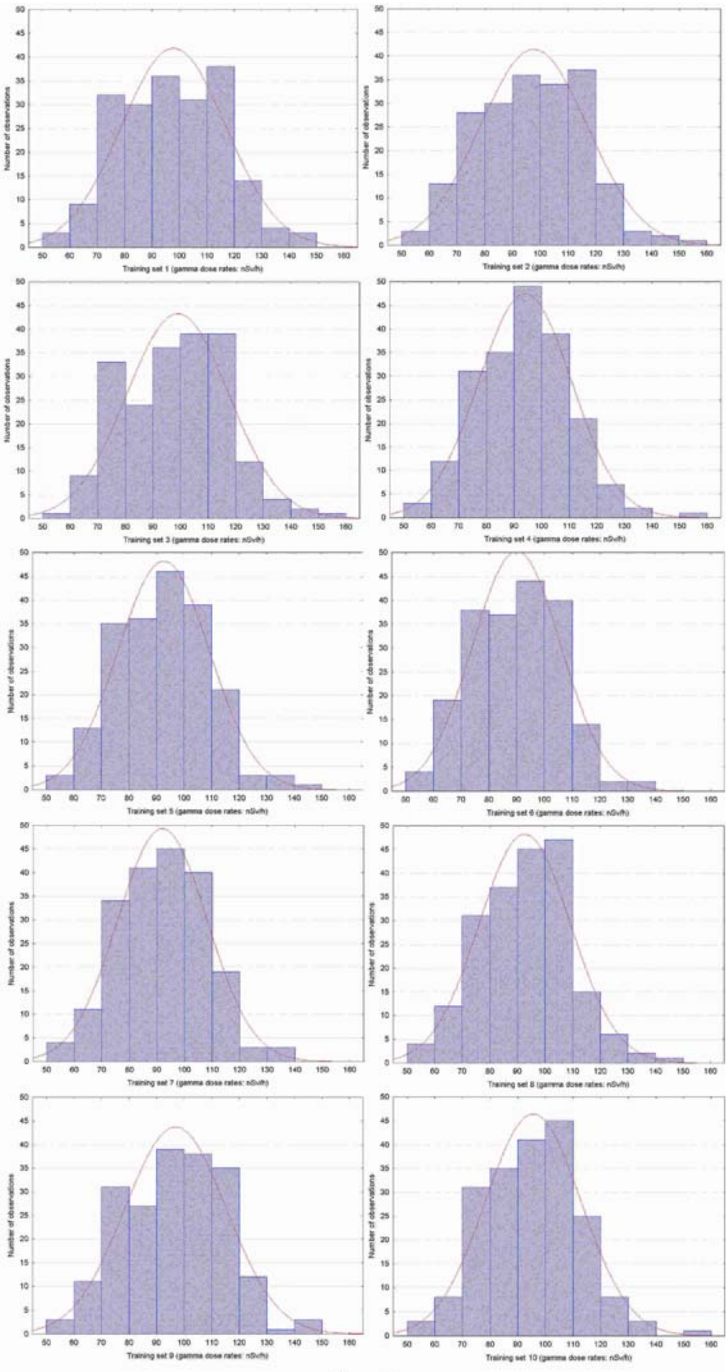

Figure 2

Frequency histograms for the first 10 training sets (from top to bottom, left to right, sets 1 to 10) described in Table 1. The smooth curve depicts the normal distribution.

Datasets:	N = 1008	n = 200	N-n = 808
Minimum distance (km)	1.6	5.0	1.6
Maximum distance (km)	21.0	53.0	23.5
Median distance (km)	10.7	16.0	11.3
Mean distance (km)	10.7	18.0	11.3
Standard deviation (km)	3.0	9.0	3.2
Coefficient of variation	0.3	0.5	0.3
Skewness	0.0	1.2	0.2
Clark & Evans' test for complete spatial randomness	1.4	1.1	1.3

Table 1
Nearest-neighbour distances statistics for the SIC2004 datasets.
Measurement units are km.

Training sets (n = 200)	Min.	Max.	Mean	Median	Std. dev.	Skewness	Kurtosis
Set 1	55.8	150.0	97.6	98.0	19.1	0.0	-0.5
Set 2	55.9	155.0	97.4	97.9	19.3	0.1	-0.5
Set 3	59.9	157.0	98.8	100.0	18.5	0.1	-0.3
Set 4	56.1	152.0	93.8	94.8	16.8	0.2	0.0
Set 5	56.4	143.0	92.4	92.0	16.6	0.2	-0.2
Set 6	54.4	133.0	89.8	90.4	15.9	0.1	-0.5
Set 7	56.1	140.0	91.7	91.7	16.2	0.1	-0.4
Set 8	54.9	148.0	92.4	92.5	16.6	0.1	-0.1
Set 9	56.5	149.0	96.6	97.0	18.2	0.0	-0.4
Set 10	54.9	152.0	95.4	95.7	17.2	0.1	-0.2

Table 2
Statistics for the 10 datasets used to train the algorithm used in SIC2004.
Measurement units are nSv/h.

Complete sets (N = 1008)	Min.	Max.	Mean	Median	Std. dev.	Skewness	Kurtosis
Set 1	55.0	193.0	98.9	99.5	21.1	0.4	0.7
Set 2	54.9	188.0	98.8	99.5	21.2	0.4	0.6
Set 3	59.9	192.0	100.3	101.0	20.4	0.5	0.7
Set 4	56.1	180.0	95.1	95.4	18.8	0.6	1.1
Set 5	56.1	168.0	93.7	94.0	18.1	0.5	0.6
Set 6	54.4	168.0	90.9	91.6	17.2	0.4	0.6
Set 7	56.1	166.0	92.5	92.9	16.9	0.4	0.4
Set 8	54.9	176.0	93.5	94.1	18.1	0.5	1.0
Set 9	56.5	183.0	97.8	98.7	19.9	0.4	0.6
Set 10	54.9	183.0	96.6	97.1	19.0	0.5	0.8

Table 3
Descriptive statistics for the full datasets from which the training data were extracted. Measurement units are nSv/h.

2.3 NATURAL BACKGROUND RADIATION LEVELS

If the type of variable used and the source of measurements were disclosed to the participants, no additional details were provided. Participants were, however, free to use any information they might have considered as relevant. For what concerns the physical properties of the variable, one will here briefly remind the reader that people have always been exposed to ionizing radiations originating from their natural environment: small amounts of uranium, thorium, radium and potassium in the terrestrial crust as well as cosmic radiations are the main contributors to the total annual dose we receive. These radiations differ in time and space and are usually in the level of magnitude of a hundred of nanoSievert per hour (nSv/h). Because the origin of these radiations is mainly geological, spatial fluctuations of dose rates are very similar to fluctuations observed in the concentration levels found for the related minerals. Small amounts of radon, a radioactive gas which comes from the radioactive decay of uranium, also seeps into the atmosphere from the soil and contributes to the total dose. Local differences can thus occur because of changes of atmospheric pressure (rainfall) or snow cover attenuating terrestrial radiations. One will find on the web pages of BFS that the natural exposure to natural radiations in Germany varies between 2 and 5 mSv/y (millisievert per year) depending on the locality and that it can present up to 10 mSv/y in a few locations. In average, the exposure of man is around 2.1 mSv/y in Germany, from which 1.1 mSv come from the radon gas, 0.3 mSv come from the natural radioactivity found in food and the remaining 0.7 mSv are from the cosmic and terrestrial radiations.

3. SIC2004 DATASETS

For a few months the above data were made available on the internet, after which SIC2004 participants had to submit, to the Editorial Board and before a given deadline, a clear description of the algorithm developed for the exercise. In exchange for this description, participants received a code to identify themselves and a password to access a web site where the data for the exercise could be loaded. Because the computing time was recorded, participants were authorized to download the data only once and had to send the results back in the shortest amount of time. This guaranteed that no unexpected interaction with the data would have occurred during the exercise.

The two datasets that were downloaded by the participants are described below.

3.1 FIRST DATASET

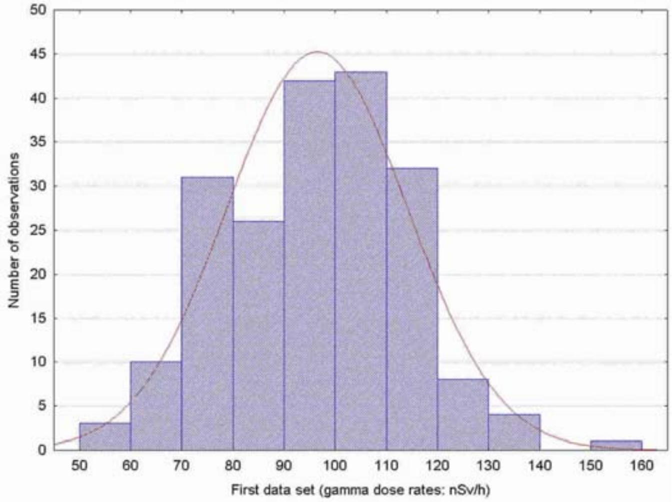

Figure 3
Frequency histogram of the first data used for estimations and theoretical normal distribution (line).

The first dataset consisted simply of an 11th dataset of 200 measurements which had co-ordinates 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 given in Table 4. A frequency histogram for the 200 measurements is given in Figure 3, above.

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

Table 4
Summary statistics of the first dataset used for the SIC2004 exercise.

3.3 SECOND DATASET

Together with the first dataset, participants to SIC2004 received an unexpected second one. The purpose of the so-called "joker" dataset was to assess the performance of the algorithms in case of emergencies and outliers. If participants were warned about the possibility of anomalies in datasets, the use of a second dataset distributed as a surprise allowed us to ensure that the same automatic mapping algorithm was applied to both sets. This second dataset was derived from the first dataset described above, except that an emission of an undefined radioactive substance was simulated in the South-West part of the monitored area. The simulated release occurred at a location that was distributed with the "output" datasets and had the ID = 911 and the following coordinates: x_0 = 202 and y_0 = 182665.

It was assumed that the emission had undergone an undefined dispersion process (e.g. Jacobson 1998). The mass per unit surface at every point in space and time instant, $c(x,y,t)$, could thus be given by the following expression:

$$c(x,y,t) = \left(\frac{2Q}{\sigma^2}\right) e^{-\frac{1}{2}\left(\frac{x-x_0-ut}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{y-y_0}{\sigma}\right)^2} \tag{1}$$

where

Q is total mass emitted,
u is the velocity of the advection process,
σ is a dispersion parameter,
t is time

The values used for the parameters were:

$Q = 1.10^7$ [undefined units]
$u = 7$ [m/s]
$\sigma = 5000$ [m]

The final field calculated at all points of the original set of locations is obtained by discretising (1) and by adding up fields relating to an emission time T=15000 s by means of the expression:

$$c(x_i, y_i, t_{j+1}) = c(x_i, y_i, t_j) + \left(\frac{2Q}{\sigma^2}\right) e^{-\frac{1}{2}\left(\frac{x-x_0-ut}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{y-y_0}{\sigma}\right)^2} \tag{2}$$

for *j*= 0,..., *t* with an increment of 1.

Thus, this expression simulates a time integrated concentration over duration *T*. The values derived from the simulated emissions have then been summed to the background values of the first dataset. As a result of this simulation, 5 out of 200 stations showed a significant increase in the dose rates, from which 2 had an increase on the order of 10 times the original background level reported initially. The summary statistics for this second set are given in Table 5. The related frequency histogram is shown in Figure 4 and shows, as expected, the strong impact of the simulated release on the statistical distribution of the data.

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

Table 5
Summary statistics of the second ("joker") dataset used for the SIC2004 exercise.

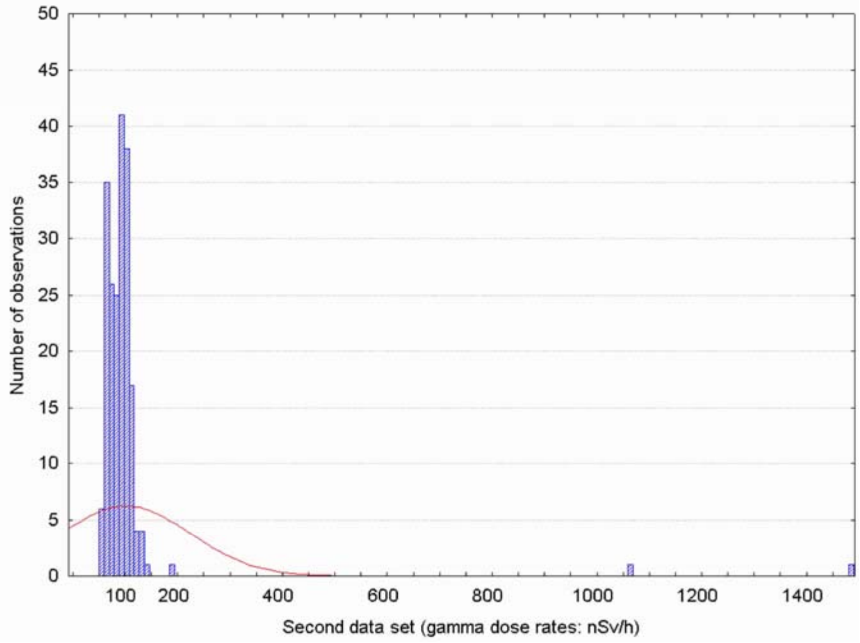

Figure 4
Frequency histogram of the subset used as the "joker" data set used for estimations and theoretical normal distribution (line).

Figure 5 shows a 3D model of the dispersion process on the top of the overall background structure of the dose rates which remained largely unaffected.

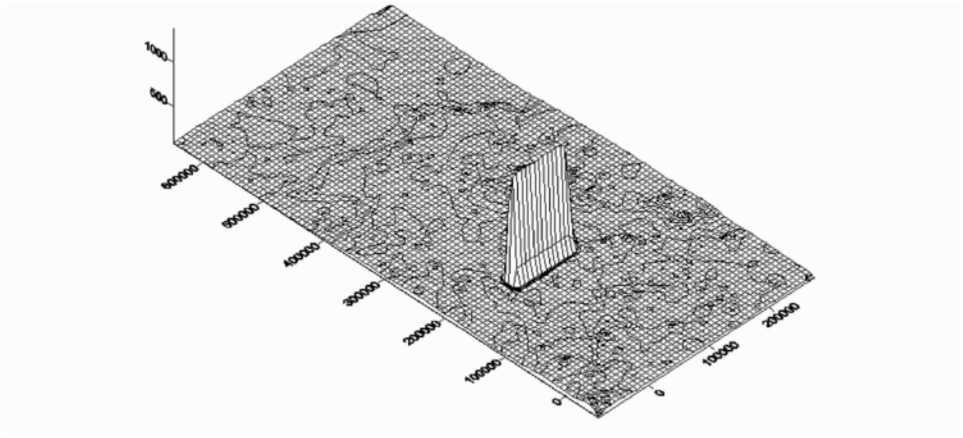

Figure 5
3D model of the dispersion process displayed on the top of the background radiation map (vertical scale in nSv/h).

3.4 ACCESSING THE DATASETS

Date used in SIC97 and SIC2004 can be downloaded from the AI-GEOSTATS web site, see (www.ai-geostats.org).

ACKNOWLEDGMENTS

We are very grateful to the German Federal Office for Radiation Protection for allowing us to use the dose-rate measurements in this exercise. We were also helped by two colleagues: Patrick Petit kindly provided us with the technical assistance needed to set up the web site which recorded download and upload times for the datasets; and Tore Tollefsen rendered this paper much more readable.

REFERENCES

Bucher, F., Vèkovski, A. Improving the selection of appropriate spatial interpolation methods. In: "*Spatial Information Theory: a theoretical basis for GIS; Proceedings of the COSIT'95 conference*". 1995; Frank, A. U., Kuhn, W. (Eds.), Springer Verlag, pp. 351-364.

Clark, P., Evans, F. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 1954; 35:445–453.

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

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

EC, European Commission. *Mapping Radioactivity in the environment. Spatial Interpolation Comparison 1997*. 2003; EUR 20667 EN, EC. Dubois, G., Malczewski, J., De Cort, M. (Eds), 268 pp.

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

Jacobson, M. Z. *Fundamentals of Atmospheric Modeling*, 1998; Cambridge University Press, 672 pp.

Cite this article as: Dubois, G; Galmarini, S. 'Introduction to the Spatial Interpolation Comparison (SIC) 2004 exercise and presentation of the datasets'. *Applied GIS*, Vol 1, No 2, 2005. pp. 09-01 to 09-11.
DOI: 10.2104/ag050009