Interactive Declustering of Spatial Environmental Data for Geostatistical Analyses

8th Annual Hawaii International Conference on Statistics, Mathematics, and Related Fields

Stanley M. Miller
Christopher P. Oertel
John R. Giles

January 2009

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint should not be cited or reproduced without permission of the author. This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the United States Government or the sponsoring agency.
Interactive Declustering of Spatial Environmental Data for Geostatistical Analyses

Stanley M. Miller, Department of Civil Engineering, University of Idaho, Moscow, Idaho
Christopher P. Oertel, Environmental Engineering, Idaho National Laboratory, Idaho Falls, Idaho
John R. Giles, Environmental Engineering, Idaho National Laboratory, Idaho Falls, Idaho

ABSTRACT

Clustered sampling often results during environmental site investigations when localized areas are over-sampled due to specific concerns in those areas or due to monitoring programs focused on specific zones. An interactive, nearest-neighbor method for efficient spatial declustering has been developed as part of the ongoing monitoring and assessment of Cs-137 concentrations in soils at the Idaho National Laboratory (INL) site. In situ field measurements of Cs-137 have been obtained with a field gamma-ray spectrometer using a sampling layout with points concentrated around several former nuclear processing facilities. The spatial declustering process allows for more useful and productive geostatistical studies focused on characterization of spatial dependence and subsequent spatial estimation to generate maps that depict Cs-137 concentrations across the site.

Key Words: spatial environmental data, sample declustering, geostatistics

INTRODUCTION

In many environmental sampling programs, clusters of closely spaced samples often result when localized areas are over-sampled due to specific concerns in those areas or due to monitoring programs focused on specific zones. Not only does this sampling provide biased estimates of the mean and distribution of the spatial attribute of interest, it also often makes it difficult to conduct meaningful geostatistical studies focused on characterization of spatial dependence and subsequent spatial estimation to produce maps that depict the spatial distribution of the environmental attribute.

Such is the case for the ongoing monitoring and assessment of Cs-137 concentrations in soils at the Idaho National Laboratory (INL) site, a government facility that covers 890 square miles in eastern Idaho. This particular radionuclide is a byproduct of nuclear processing activities and has a half-life of 30 years, decaying into Ba-137 by β emission with a concurrent release of gamma radiation. In situ field measurements of Cs-137 concentrations in soils have been obtained by using a field gamma-ray spectrometer. The original sampling layout was established by the U.S. Department of Energy in the 1970’s and 1980’s, with a significant proportion of the points clustered around several former nuclear processing facilities.

The annual sampling program in 2006 included 290 locations, and additional sampling locations were added during the 2007 monitoring campaign to bring the total number to 342. Measurement levels of Cs-137 concentrations in soils during 2006 ranged from 0.0003 to 17.3 pCi/g (picocuries per gram), with several measured values exceeding 0.23 pCi/g (i.e., the identified risk-based concentration for a 30-year residential exposure scenario). The sample clustering in high-valued zones caused difficulties with
geostatistical modeling (i.e., modeling of spatial dependence using variograms or spatial covariances), so additional sampling sites in non-clustered areas were established for the 2007 sampling program. However, the effects of the high-valued clusters were not sufficiently buffered. It became clear that some type of spatial declustering would be necessary, not only to pursue the geostatistical investigations (spatial dependence modeling and subsequent kriging interpolation to generate estimation maps of Cs-137), but also to provide guidance on future sampling protocols, especially in regard to eliminating some of the “redundant” sampling locations to enhance the economy and efficiency of the Cs-137 sampling program.

Investigators have proposed several methods for such declustering of spatial data sets that exhibit preferential sampling in both location and value. One option, known as cell declustering, relies on partitioning the sampling field into square cells and then estimating declustering weight as an inverse proportion of the number of data per cell (for example, Journel, 1983). A proposed variation of the cell method involves using approximated global kriging weights (from summed local krigings) as the declustering weights (Deustch, 1989). Another declustering approach relies on using area polygons that enclose data locations to weight each datum inversely proportional to the area of the polygon containing that datum (for example, Isaaks and Srivastava, 1989, Chap. 10). Finally, if covariance stationarity can be assumed for the random function used to model the data set, then declustering weights can be based on the spatial correlation of the spatial attribute using the ratio of determinants of correlation matrices (Bourgault, 1997). Each of these methods has disadvantages when trying to decluster a data set to obtain a histogram and a variogram (i.e., model of spatial dependence) more representative of the entire sampling population than that which can be obtained solely from the clustered, preferential spatial sample.

NEAREST-NEIGHBOR DECLUSTERING

A nearest-neighbor declustering method recently presented by Olea (2007) has been investigated for application to the INL Cs-137 data. In this approach, the original spatial sample set is divided into two subsets using the cumulative relative frequency (crf) plot of the calculated nearest-neighbor distances. One subset consists of the data found in spatial clusters, while the other subset consists of those data scattered about the study area in a non-clustered fashion. Repeated nearest-neighbor calculations for points in the clustered subset with other points in the non-clustered subset allow the user to sequentially add points from the first subset to the second one, thus gradually replacing the culled clustered zones one point at a time until representative statistics result and reasonable spatial-dependence models can be generated (i.e., variograms).

The example data set used by Olea (2007) to illustrate the declustering method contained 140 observations, including 54 of which were contained in high-valued clusters, with the remaining 86 sampling locations forming a fairly evenly-spaced, nonclustered subset. The mean of the clustered subset was approximately seven times that of the nonclustered subset.

The extent of preferential clustering in the INL Cs-137 data is far greater than Olea’s example. For instance, the radionuclide measurements from 2006 showed that approximately 250 out of the total 290 sampling locations were associated with high-valued clusters nearby former nuclear processing facilities. To investigate the applicability of nearest-neighbor declustering for this severe case of clustered spatial data, we developed a visually-based, interactive system using Mathcad® software (PTC, 2008). This allowed for a step-by-step visual assessment of using the nearest-neighbor method for sequentially transferring sampling points from the first subset to the second one, as described above.
INITIAL DATA ANALYSIS

Comprehensive annual sampling program for measuring Cs-137 concentrations in soils across the INL site was initiated several years ago. Field gamma-ray spectrometers have been used to measure insitu emission of gamma rays, measurements that then can be converted to concentrations reported in picocuries per gram (pCi/g).

Data analysis has been based primarily on Mathcad® worksheets developed specifically for this project. Two annual sampling programs have been analyzed, those completed in 2006 and 2007. A postplot of the 2006 Cs-137 sampling sites clearly illustrates spatial clustering at four locations across the site (Fig. 1A). Summary statistics for the 2006 data also are given in Fig. 1. The data distribution is strongly skewed right with 196 data values less than or equal to 0.5 pCi/g.

Spatial dependence modeling using variogram computations was attempted for all of the original 290 data from 2006. The isotropic (i.e., all-directional) variogram indicated very little spatial dependence, as indicated by the random fluctuations about the variance (Fig. 2A). In an effort to eliminate some of this “noise”, the data set was culled by removing 28 high-valued outliers (i.e., those data with Cs-137 values exceeding 2 pCi/g), and a new variogram was generated (Fig. 2B). This new variogram also displayed random fluctuations about the variance with little indication of spatial dependence (which would be shown by increasing variogram values from near the origin).

Summary statistics for the 2007 data are presented in Fig. 3. The data distribution is strongly skewed right with 190 data values less than or equal to 0.5 pCi/g. Attempts at modeling spatial dependence using variograms provided similar results to that for the 2006 data. That is, for both the original data and for a culled data set with 39 outliers removed, the computed variograms are quite noisy (Fig. 4).

Figure 1. Statistical summary of Cs-137 data collected in 2006: A. Postplot of all sampling sites; B. Histogram of all data.
Figure 2. All-directional (isotropic) variograms for the Cs-137 data collected in 2006: A. Variogram for all 290 data values; B. Variogram for culled data set with outliers exceeding 2 pCi/g removed. The horizontal line represents the sample variance.

Figure 3. Statistical summary of Cs-137 data collected in 2007: A. Postplot of all sampling sites; B. Histogram of all data except for an outlier of 49 pCi/g.
Figure 4. All-directional (isotropic) variograms for the Cs-137 data collected in 2007: A. Variogram for all 342 data values; B. Variogram for culled data set with outliers exceeding 2 pCi/g removed. The horizontal line represents the sample variance.

NEAREST-NEIGHBORDECLUSTERING

To implement the nearest-neighbor declustering method of Olea (2007), we developed an interactive Mathcad® worksheet to display sequentially the postplots of both Subset 1 (non-clustered data) and Subset 2 (clustered data). The first step is to identify the initial configuration of Subset 1 by inspection of the cumulative relative frequency plot (CRF) of all the calculated nearest-neighbor distances. Each data location has a nearest neighbor, and the distance to that neighbor was identified using a sorting routine. As shown in Fig. 5, the cumulative probability where the CRF plot effectively levels off at an

Figure 5. Partial CRF plot for 2006 Cs-137 data, with the “break” between the two subsets indicated by upper plateau of the graph at cumulative probability of 0.88 (thus, 0.12 x 290 data = 35 non-clustered data).
upper plateau (CRF value of 0.88) defines a reasonable break to subdivide the 2006 data into the two subsets. In this case, Subset 2 (clustered data) consists of the data with nearest-neighbor distances less than the 0.88 quantile value (i.e., 0.88 x 290 = 255 clustered data), and Subset 1 consists of the data with nearest-neighbor distances greater than this quantile value (i.e., 0.12 x 290 = 35 non-clustered data).

Relevant statistics for the two subsets are summarized in Table 1. It should be noted that Subset 2 (clustered data) does not exclusively contain only high values of Cs-137; in fact, the clusters include a large number of values less than or equal to 0.2 pCi/g. This helps to explain the erratic behavior in the variograms at short lag distances, because in some of the clustered zones small data values are located in close proximity to large data values, effectively causing large variogram values at short lags.

Repeated nearest-neighbor calculations for points in the clustered Subset 2 with other points in the non-clustered Subset 1 allow the user to sequentially add points from the clustered subset to the other one, thus filling in the previously clustered zones with one point at a time until representative statistics result and reasonable spatial-dependence models can be generated (i.e., variograms). In this case, we repeated this iteration until the noise in the computed variogram was reduced enough to be able to reveal some spatial structure in the non-clustered Cs-137 data. The Mathcad® worksheet displays a revised post plot for each iteration and an updated plot of the variogram for the new non-clustered data set. For the example shown in Fig. 6 based on the 2006 data set with outliers removed, a visual assessment of the iteration process clearly showed that the declustered data set to be used for subsequent spatial dependence modeling and geostatistical estimation (mapping) resulted after the 9-th iteration. That is, the number of data to use in the declustered data set is 44, obtained from the original 35 non-clustered values plus 9 other data added in (based on maximum nearest-neighbor distances to other data locations) to represent the original sampling cluster locations. Although Olea (2007) mentions several numerical or statistical checks to help identify when enough iterations have been completed, in our case that decision point was indicated clearly by a sudden change (degradation) in the variogram plot (see Fig. 7).

A similar declustering analysis was conducted for the 2007 data set. In this case, the number of original non-clustered data in Subset 1 was 55, using the 0.84 quantile value of the nearest-neighbor distance (Fig. 8). The 3-rd iteration provided good results and prescribed a new declustered set of 58 values (Fig. 9).

Table 1. Summary of statistics for initial subsets of non-clustered data (Subset 1) and clustered data (Subset 2).

<table>
<thead>
<tr>
<th>Cs-137 Value (pCi/g)</th>
<th>2006 Cs-137 Data</th>
<th>2007 Cs-137 Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subset 1</td>
<td>Subset 2</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.027</td>
<td>0.0003</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.344</td>
<td>17.31</td>
</tr>
<tr>
<td>Mean</td>
<td>0.084</td>
<td>1.071</td>
</tr>
<tr>
<td>Median</td>
<td>0.075</td>
<td>0.36</td>
</tr>
<tr>
<td>No. data ≤ 0.2 pCi/g</td>
<td>34</td>
<td>97</td>
</tr>
<tr>
<td>No. data &gt; 0.2 and ≤ 0.5 pCi/g</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>No. data &gt; 0.5 and ≤ 1.0 pCi/g</td>
<td>46</td>
<td>2</td>
</tr>
<tr>
<td>No. data &gt; 1.0 pCi/g</td>
<td>48</td>
<td></td>
</tr>
</tbody>
</table>

* Maximum if data value of 49 is excluded.
8th Annual Hawaii International Conference on Statistics, Mathematics, and Related Fields

User-specified spherical variogram model terms:

- nug = 0.00
- range = 34000

**Iteration No. 9**

N2 := length(a) = 218
N1 := length(z) = 44

**A. Post Plot**

**B. All-directional Variogram**

Figure 6. Postplot and variogram for 2006 Cs-137 data (with outliers removed), based on Iteration 9 for declustering the data set (i.e., the number of non-clustered data in Subset 1 is N = 44).

**Iteration No. 10**

N2 := length(a) = 217
N1 := length(z) = 45

**A. Post Plot**

**B. All-directional Variogram**

Figure 7. Postplot and variogram for 2006 Cs-137 data (with outliers removed), based on Iteration 10 for declustering the data set (i.e., the number of non-clustered data in Subset 1 is N = 45).
Figure 8. Partial CRF plot for 2007 Cs-137 data, with the “break” between the two subsets indicated by upper plateau of the graph at cumulative probability of 0.84 (thus, $0.16 \times 342 = 55$ non-clustered data).

User-specified spherical variogram model terms: $\text{nug} := 0.001 \quad \text{range} := 23000$

Iteration No. 3


Figure 9. Postplot and variogram for 2007 Cs-137 data (with outliers removed), based on Iteration 3 for declustering the data set (i.e., the number of non-clustered data in Subset 1 is $N = 58$).
CONCLUSIONS

The ongoing monitoring and assessment of Cs-137 concentrations in soils at the Idaho National Laboratory (INL) site is based on a spatially clustered sampling protocol with emphasis on areas in close proximity to nuclear processing facilities. This clustered sampling has made it difficult to develop spatial dependence models for the purpose of conducting geostatistical estimation (mapping) of Cs-137 concentrations across the entire site. An interactive, visual, nearest-neighbor method for efficient spatial declustering has been developed with Mathcad® to provide an efficient way to generate a declustered subset of the original data set that subsequently can provide reasonable spatial dependence models and enhanced spatial interpolation (kriging) and mapping. The removal of high-valued outliers (data values exceeding 2 pCi/g) also helped to improve the spatial dependence modeling process.

As expected when sampling clusters occur in preferentially high-valued locations, any declustered subset of the original data will have a sample variance much lower than that of the entire original sample. Hence, the variogram sills (plateaus) will have lower values, and the resulting kriging errors will be considerably lower than those obtained when using a variogram based on the entire data set. Thus, those kriging errors may be somewhat optimistic, and possibly unrealistic, when compared to the sample statistics of the original data. This should be noted and acknowledged when reporting the results of such studies. The main thing to remember when working with declustered subsets is to use the entire original data set during the kriging operation, so that all local spatial estimates will honor the original data values at the original sampling locations.

ACKNOWLEDGEMENTS

This work was conducted at the University of Idaho and was funded by the Idaho National Laboratory through research contract No. XXXXXX.

REFERENCES


