In "Environmental Modeling with GIS" M. Goodchild B. Parks L. Steyaert Oxford Univ. Press 1993

44

GRASS Used in the Geostatistical Analysis of Lakewater Data from the Eastern Lake Survey, Phase I

HANNAH RASMUSSEN RHODES AND
DONALD E. MYERS

This case study describes how GRASS was used in the geostatistical analysis of the ELS-I data. The objective was to obtain estimates of the spatial distributions of ion concentrations for chemical analytes most relevant to lake acidification processes. In applying geostatistical techniques the support of the data (e.g., the area associated with a particular lake sample) has a crucial effect. The original objective of the ELS-I sampling did not require incorporating the size of the lakes nor require more than one sample per lake. In order to examine the influence of lake size on the kriging analysis, pseudo sample positions within the large lakes (surface area greater than 500 ha) sampled during the ELS-I were compiled using GRASS. These pseudo sample positions allow for incorporating the differing sample supports into the analysis.

ELS-I DATA

The data analyzed were obtained from the ELS-I initiated in 1983 by the U.S. Environmental Protection Agency (EPA). The ELS-I was designed to describe the chemical status of lakes in areas of the eastern U.S. containing the majority of low-alkalinity lakes. These were defined as having acid neutralizing capacity (ANC) $\leq$ 400 μeq l^{-1}. Complete details on statistical design, lake selection, analytical methodologies, field methods, and quality-assurance protocols are given elsewhere (Best et al., 1986; Linthurst et al., 1986; Overton et al., 1986; Overton, 1986; Kanciruk et al., 1986); so only the key design features are presented here.

A stratified design was used wherein sample lakes were selected statistically from each stratum. The first level of stratification included three regions of the eastern U.S. (Northeast, Upper Midwest, Southeast) expected to be the most susceptible to change as a result of acidic deposition. The second stratification level was obtained by dividing each region into subregions exhibiting geographic homogeneity with respect to water quality, physiography, vegetation, climate, and soils. Finally, each subregion was further subdivided into three alkalinity map classes (ANC < 100, 100 – 200, and > 200 μeq l^{-1}). The population of lakes was screened to exclude lakes surrounded by or adjacent to anthropogenic activities such as intense urban, industrial, or agricultural land use. The survey was conducted in the Fall of 1984 with one water sample per lake collected from a depth of 1.5 m at the deepest part of the lake. 26 chemical analytes related to surface water acidification and several physical attributes were measured for each lake.

The geostatistical analysis was restricted to the Northeast, identified as region 1, which is divided into five subregions: Adirondacks (1A), Poconos/Catskills (1B), Central New England (1C), Southern New England (1D), and Maine (1E). Several chemical analytes from the PC version of the ELS-I data were examined, but due to space limitations only the results pertaining to ANC in subregion 1A are presented in this case study. Figure 44-1 shows the locations of the lakes sampled in subregion 1A.

COMPILATION OF PSEUDO SAMPLE POSITIONS USING GRASS

Consider the layout in Figure 44-2, where a lake of interest is located near one large lake and four smaller lakes. Let l_e denote the lake to be estimated, l_i sampled lake i, x the original sample location, and d_i the distance between l_e and the sample location for l_i. When the kriging estimate and variance are calculated, the number of samples in the search neighborhood must be specified. Suppose a search neighborhood encompassing four samples is chosen. Since $d_1 > d_2 > d_3 > d_4 > d_5$ the value corresponding to the large lake, l_1, will not be included in the kriging

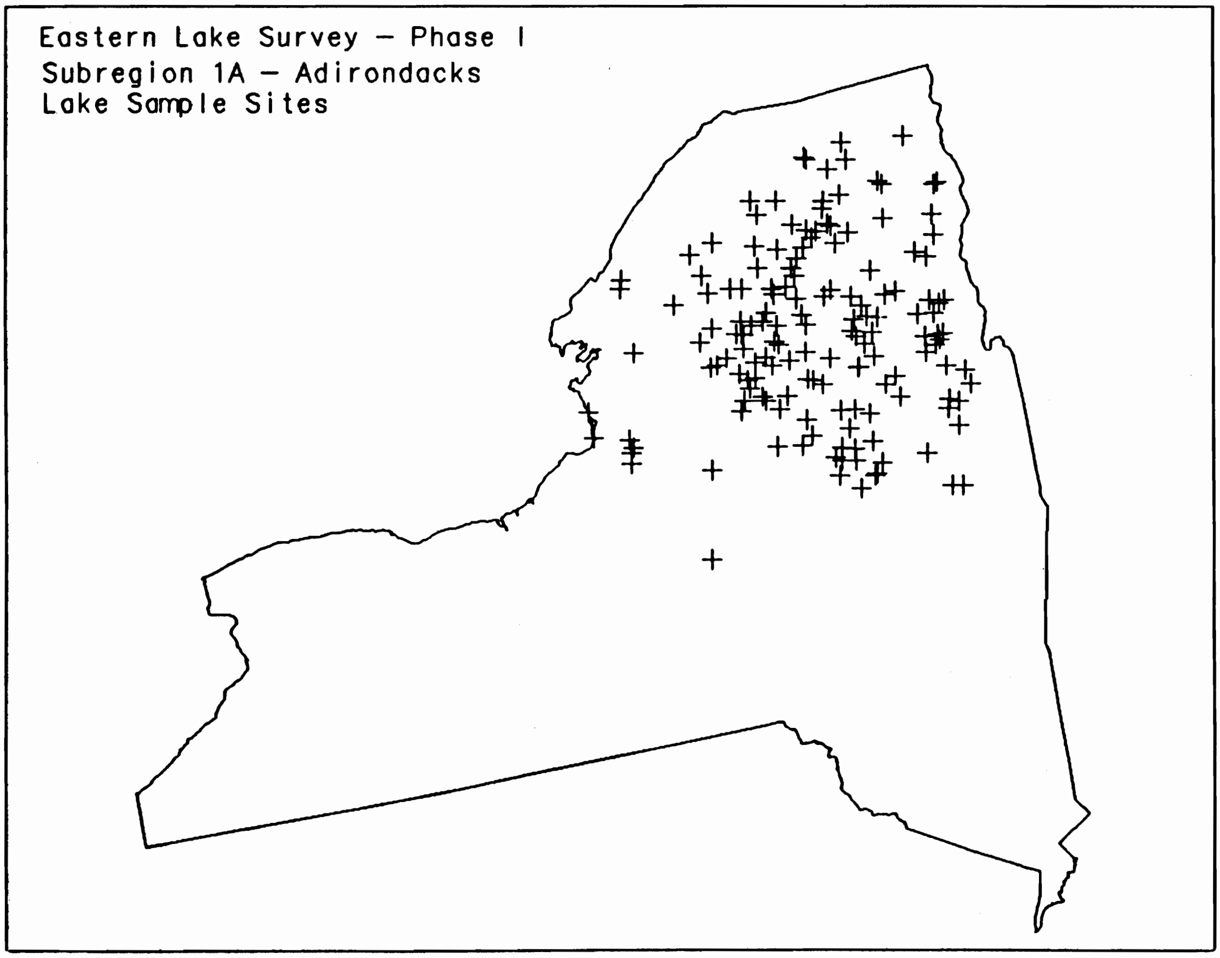

Figure 44-1: Locations of lakes sampled in Subregion 1A.

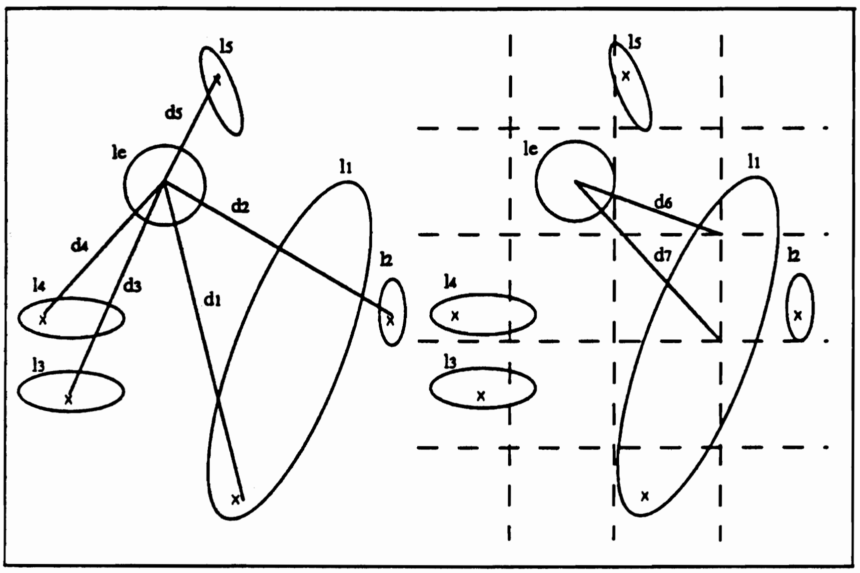

Figure 44-2: Theoretical example.

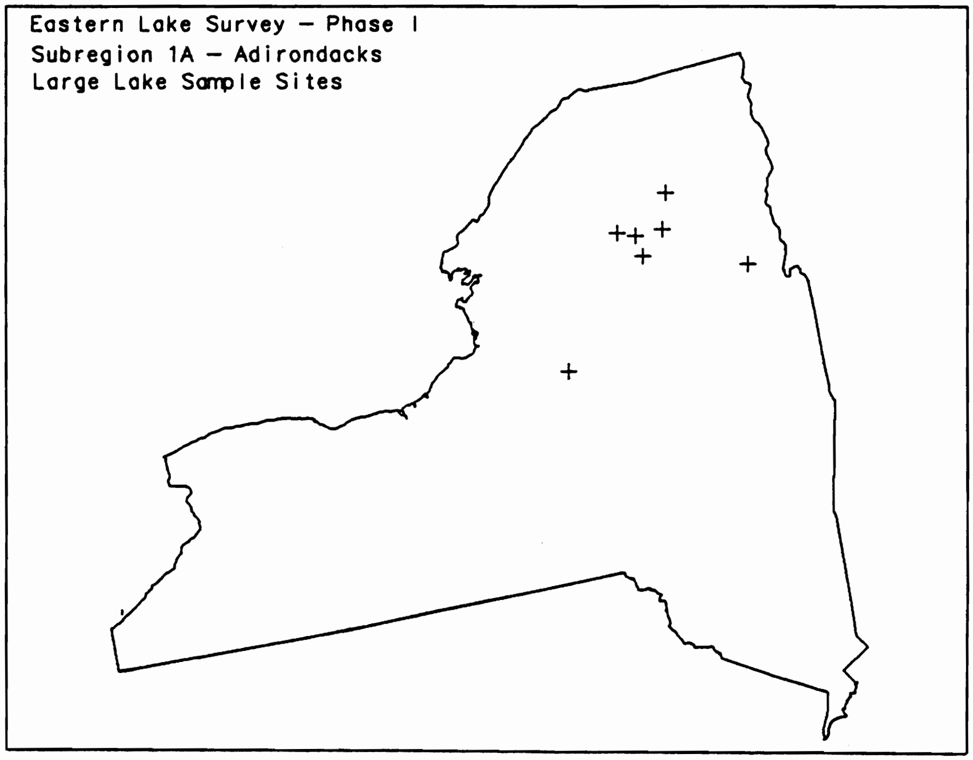

Figure 44-3: Location of large lakes sampled in Subregion 1A.

estimation of the value for l_e. Consider overlaying a grid and adding pseudo sample positions at the gridpoints lying within the boundary of l_1. Now $d_1 > d_2 > d_3 > d_7 > d_4 > d_6 > d_5$; so the values corresponding to lakes l_2 and l_3 will no longer influence the kriging estimate for l_e, but the value corresponding to l_1 will be included.

To obtain outlines of the large lakes sampled during the ELS-I, hydrological digital line graph (DLG) data displaying small-scale (1:2,000,000) streams and water bodies in the northeastern U.S. were purchased from the U.S. Geological Survey. It should be pointed out that the ground coordinate system for the DLG data is the Albers Equal-Area Conic projection, whereas the ELS-I sample locations have coordinates given in latitude and longitude.

The GRASS 3.1 programs used in the compilation process will be outlined here. First, a default window specifying the geographic coverage of the DLG data was created, and the *import.to.vect* program was used to convert the DLG data into the GRASS vector format. Windows for the five subregions of region 1 were defined with the *window* command. The locations for the lakes sampled in region 1 were converted from latitude and longitude to Albers coordinates in order to have a common coordinate reference. Then site files of the Albers coordinates for the large lakes sampled were created. *Dvect* was used to display the outlines of lakes in the northeastern U.S., and the large lakes sampled were marked by overlaying the site files using *d.sites*. Figure 44-3 shows the location of the large lakes sampled in subregion 1A. The *window* command was used to zoom in on the large lakes sampled, and *Dgrid* was used to overlay a 2000 m sized grid. A 1000 m sized grid was also overlaid in order to examine the effect of grid size. Pseudo sample positions were then obtained at each of the grid points falling within the boundary of the large lake with *Dwhere*. Figure 44-4 depicts one of the large lakes sampled in subregion 1A with the original sample location marked and a 2000 m sized grid overlaid. Table 44-1 lists the number of pseudo sample positions added to the large lakes sampled in subregion 1A with a 2000 m and a 1000 m sized grid overlaid. The lake name, surface area, and ANC value are also included.

The last step in the compilation process consisted of converting the Albers coordinates for the pseudo sample positions to latitude and longitude, adding them to the ELS-I data files, and assigning values to the pseudo sample positions. Three approaches for attributing values were considered: (1) original analyte value, (2) original analyte value + $\text{rand}*\sqrt{\text{nugget}}$, and (3) original analyte

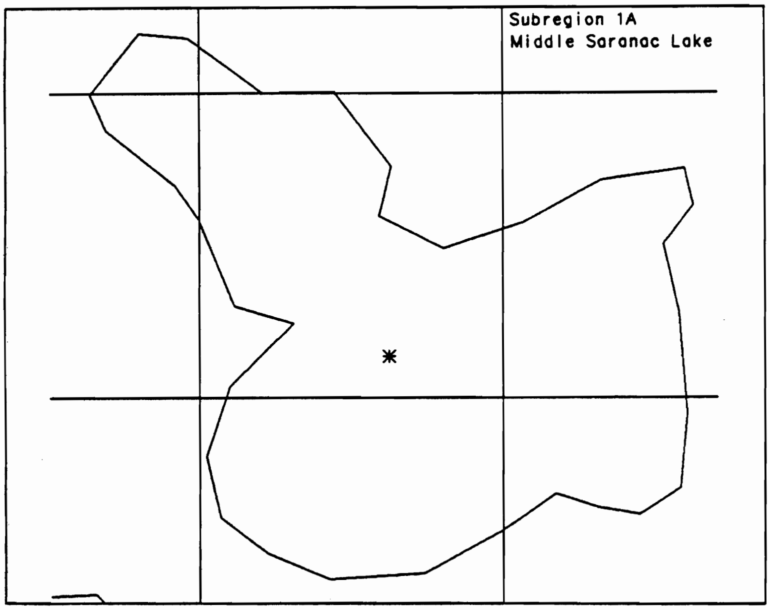

Figure 44-4: Middle Saranac Lake.

value + rand*std.dev., where rand is a random number between –1.0 and 1.0. The first approach treats the sampled lakes as perfectly mixed, whereas the latter two incorporate spatial variability within the sampled lakes. Three augmented data sets were thus created with varying values assigned to the pseudo sample positions obtained with the 2000 m sized grid. The first approach was used for the pseudo sample positions from the 1000 m sized grid. Table 44-2 lists various statistics for ANC computed from the original and the four augmented data sets.

GEOSTATISTICAL ANALYSIS

Kriging is a geostatistical estimation (interpolation) technique. The early theoretical work was done by Matheron, and the theory has been developed extensively (Journel and Huijbregts, 1978; Myers et al., 1982; Myers, 1991). A discussion of the kriging theory is beyond the scope of this case study. Examples of the development of the kriging theory and applications of its use to acid precipitation data can be found in Eynon and Switzer (1983), Finkelstein (1984), Bilonick (1985), Bilonick (1988), Venkatram (1988), Guertin et al., (1988), Marcotte (1989), or Seo et al. (1990 a, b).

Before the kriging analysis was performed, the degree of drift in the chemical analyte values was evaluated with the BLUEPACK-3D RECCO program. BLUEPACK-3D is a geostatistics software package developed jointly by the Centre de Geostatistique et de Morphologie Mathematique in Fontainebleau, France, and the BRGM (French Geological Survey). The actual kriging analysis

Table 44-1: Large lakes information

Lake ID	Lake name	Surface area	2000 m grid	1000 m grid	ANC μeq l^{-1}
1A1-044	Long Lake	1619	8	29	84.2
1A1-042	Lows Lake	1019	3	14	54.6
1A3-030	Delta Reservoir	958	3	10	1515.8
1A1-043	Little Tupper Lake	923	3	9	64.7
1A2-035	Brant Lake	595	1	6	381.4
1A1-074	Middle Saranac Lake	567	2	8	178.0
1A1-056	Forked Lake	514	3	12	56.7

Table 44-2: Statistics for ANC

Statistic	Original	2000 m grid	2000 m (nugget)	2000 m (std. dev.)	1000 m grid
Mean	251.910	255.846	252.261	259.005	255.702
Std. dev.	448.297	453.018	452.717	461.541	451.423
Minimum	–34.200	–34.200	–97.404	–281.370	–34.200
25th %tile	27.650	38.150	25.219	21.950	54.600
Median	115.700	107.400	111.850	117.750	84.200
75th %tile	234.725	231.200	232.850	273.350	212.750
Maximum	3226.800	3226.800	3226.800	3226.800	3226.800

was performed using Geo-EAS (Geostatistical Environmental Assessment Software), which is a collection of interactive software tools for performing geostatistical analyses. Geo-EAS was produced under the sponsorship of the EPA Environmental Monitoring Systems Laboratory at Las Vegas, Nevada.

First PREVAR was run to generate the pair comparison file containing the distances and relative directions between pairs of sample points. Variogram values were calculated with VARIO utilizing the pair comparison file. Pair distance intervals (lags) were specified with increments equal to the maximum interpair distance divided by 48. Omnidirectional, as well as four directional variograms, were generated in order to determine possible anisotropies. Variogram models were fitted visually to the sample variograms and the cross-validation program XVALID was run to test the models. In cross-validation each data location is suppressed one at a time, and an estimate is obtained using the other data locations, thus producing an estimation error for each sample location. In this manner XVALID was used to determine the number of sample points in the kriging search neighborhood yielding the smallest average estimation error. The actual kriged estimates were produced by KRIGE, and CONREC was executed to generate contour maps from the gridded data file produced by KRIGE.

It should be pointed out that when the variograms were computed at the alkalinity map class stratification level, the number of data pairs per lag was very small even for as few as five lags, making the variograms unsuitable for modeling. Therefore the subsequent kriging analysis was performed at the subregion stratification level. First the kriging analysis was performed on the original data set. No drift was evident in the ANC values, and the directional variograms revealed no significant anisotropies. A Gaussian model with a nugget of 25,000, a sill of 3,000,000, and a range of 6.9 was fitted to the omnidirectional variogram. Figure 44-5 shows the plot of the variogram values versus distance with the model overlaid. The kriging estimates were produced with the region divided up into 100 cells, and a kriging search neighborhood encompassing 14 points was used. 16 contouring levels ranging from –50 to

Table 44-3: Kriging estimates for ANC

Statistic	Original	2000 m grid	2000 m (nugget)	2000 m (std. dev.)	1000 m grid
Mean	251.366	255.458	251.728	257.695	246.495
Std. dev.	310.064	344.329	342.941	343.712	353.122
Minimum	4.790	4.790	2.408	4.790	4.790
25th %tile	74.259	70.247	59.996	74.761	58.315
Median	154.980	141.985	140.045	139.550	111.775
75th %tile	267.158	262.825	265.930	257.595	250.540
Maximum	1416.100	2079.700	2016.500	2105.900	1581.100

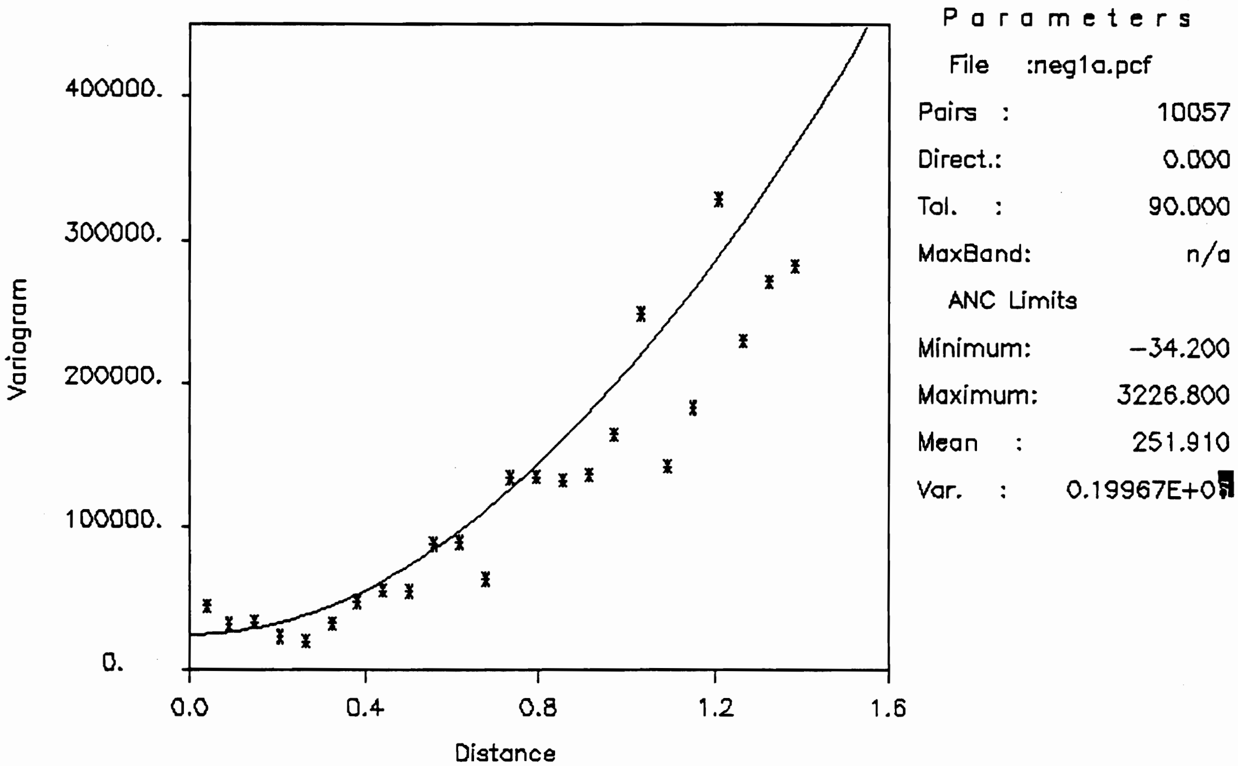

Figure 44-5: Variogram for ANC.

2200μeq l^{-1} were specified. The kriged estimates for ANC are shown in Figure 44-6 and contours for the kriging standard deviations are depicted in Figure 44-7.

In order to examine the effects of lake size on the kriging estimates, kriging analyses were performed again on the four augmented data sets following the steps outlined and using the model fitted to the variogram from the original data set. Cross-validation was then performed on the four augmented data sets in order to compare the various estimates. Table 44-3 lists statistics for the kriging estimates for ANC and Table 44-4 lists statistics for the difference between the kriging estimates for ANC and the actual ANC values. Figures 44-8 through 44-13 show the contour maps for the kriged ANC values and the kriging

Table 44-4: Difference between kriging estimates for ANC and actual ANC

Statistic	Original	2000 m grid	2000 m (nugget)	2000 m (std. dev.)	1000 m grid
Mean	–0.544	–0.388	–0.532	–1.310	-9.207
Std. dev.	335.024	295.203	297.010	308.807	259.017
Minimum	-2501.800	-2501.800	-2501.800	-2501.800	-2501.800
25th %tile	–66.355	–63.937	–66.807	–83.542	-20.340
Median	24.213	14.795	18.228	22.286	0.000
75th %tile	114.470	81.794	103.311	113.495	45.737
Maximum	837.880	860.480	852.330	861.640	744.220

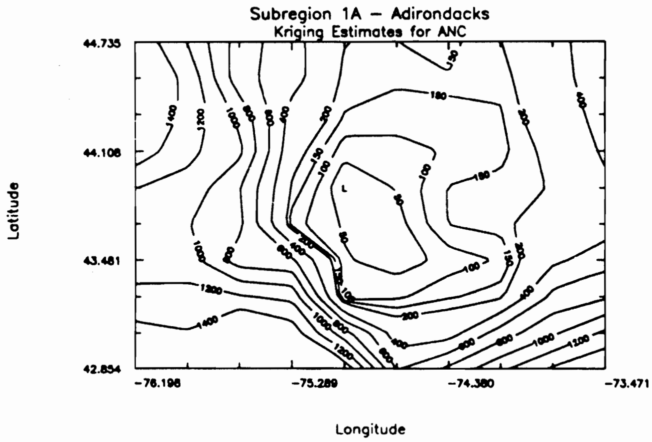

Figure 44-6: Kriging estimates for ANC.

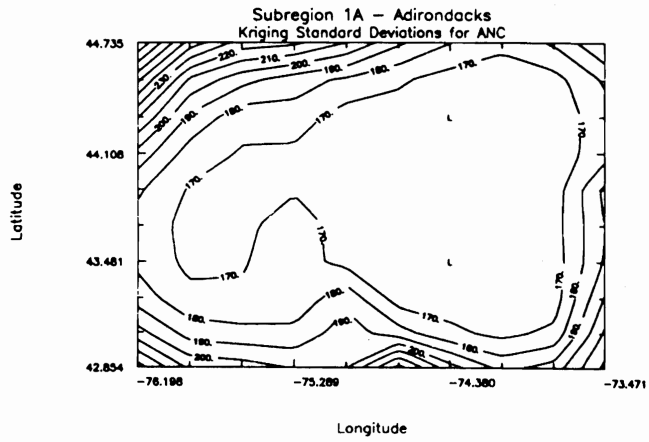

Figure 44-7: Kriging standard deviations for ANC.

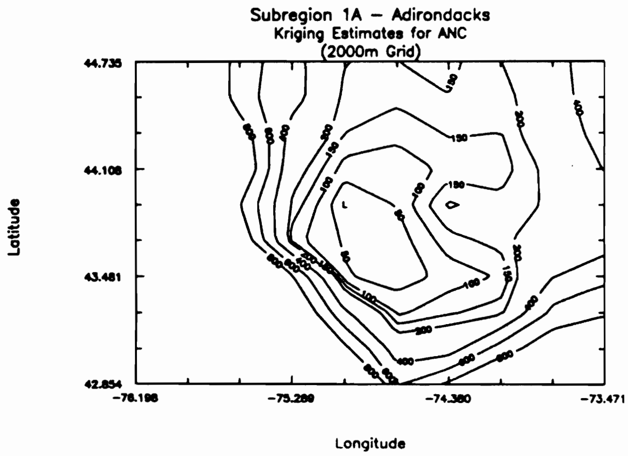

Figure 44-8: Kriging estimates for ANC, 2000 m grid.

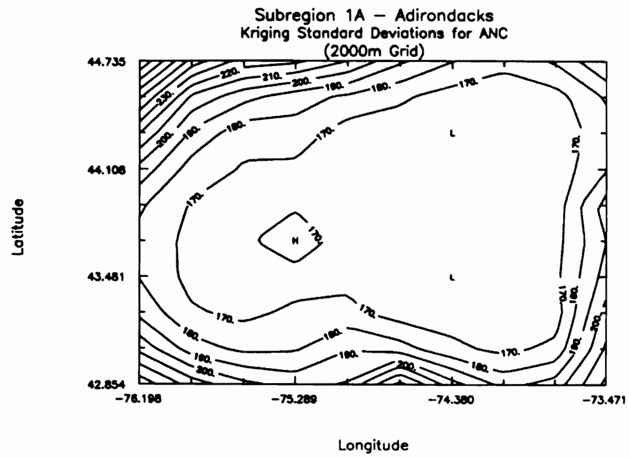

Figure 44-9: Kriging standard deviations for ANC, 2000 m grid.

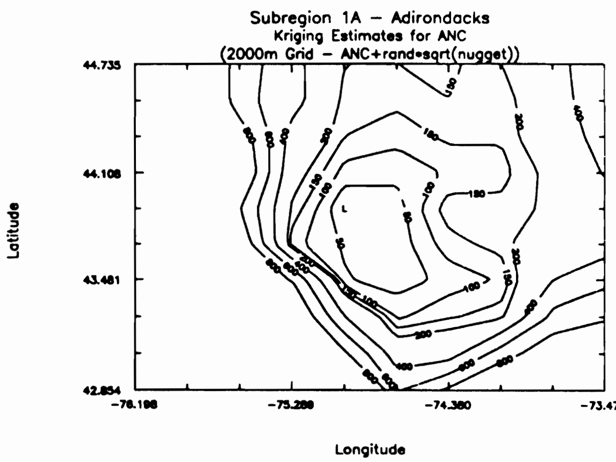

Figure 44-10: Kriging estimates for ANC, 2000 m grid (Nugget).

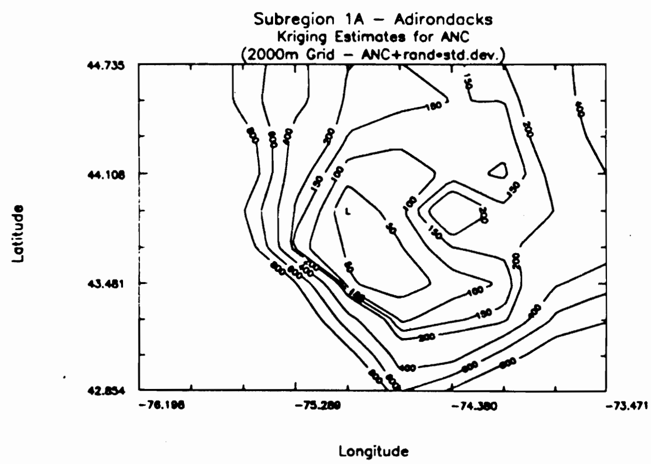

Figure 44-11: Kriging estimates for ANC, 2000 m grid (std. dev.).

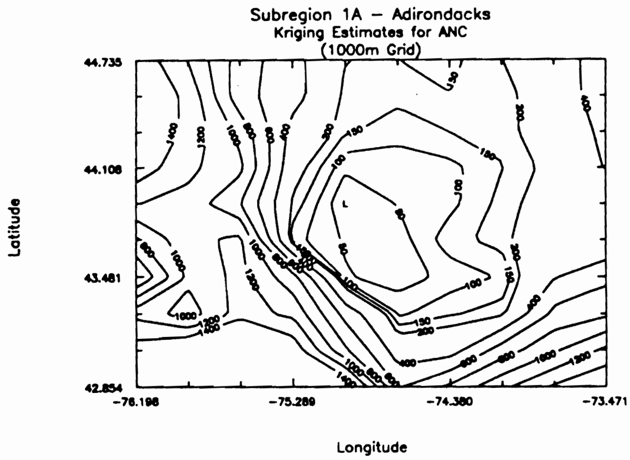

Figure 44-12: Kriging estimates for ANC, 1000 m grid.

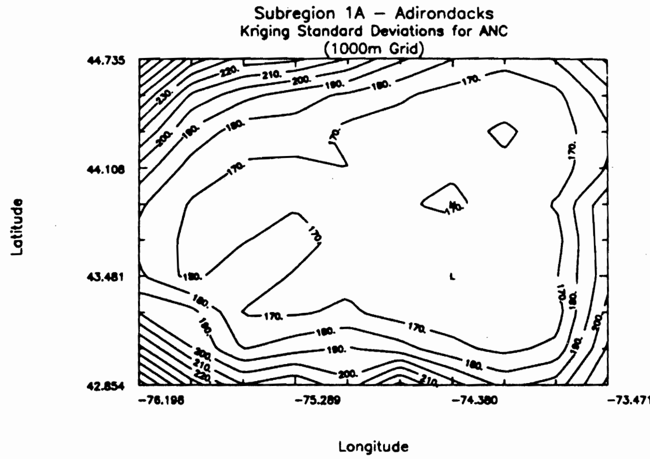

Figure 44-13: Kriging standard deviations for ANC, 1000 m grid.

standard deviations. The plots of the kriging standard deviations for the three augmented data sets obtained from the 2000 m sized grid were virtually identical and were therefore not included.

SUMMARY AND CONCLUSIONS

To examine the influence of lake size on the kriging analysis, GRASS was used to identify and assign pseudo sample positions within the large lakes sampled during the ELS-I. The inclusion of the pseudo sample positions had significant effects on the kriging analysis, as evidenced by the changes in the contour maps for the kriged ANC values. The results are sensitive to the size of the grid used to identify the pseudo sample positions and the values assigned to the pseudo sample positions.

NOTICE

Although the research described in this chapter has been funded wholly or in part by the EPA, it has not been subjected to Agency review and therefore does not reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

Best, M.D., Creelman, L.W., Drouse, S.K., and Chaloud, D.J. (1986) National Surface Water Survey, Eastern Lake Survey - Phase I, quality assurance report. *Technical Report EPA/600/4-86/011*, Las Vegas, NV: U.S. Environmental Protection Agency.

Bilonick, R.A. (1985) The space–time distribution of sulfate deposition in the Northeastern United States. *Atmospheric Environment* 19(11): 1829–1845.

Bilonick, R.A. (1988) Monthly hydrogen ion deposition maps for the Northeastern U.S. from July 1982 to September 1984. *Atmospheric Environment* 22(9): 1909–1924.

Eynon, B.P., and Switzer, P. (1983) The variability of rainfall acidity. *The Canadian Journal of Statistics* 11(1): 11–24.

Finkelstein, P.L. (1984) The spatial analysis of acid precipitation data. *Journal of Climate and Applied Meteorology* 23: 52–62.

Guertin, K., Villeneuve, J., Deschenes, S., and Jacques. G. (1988) The choice of working variables in the geostatistical estimation of the spatial distribution of ion concentration from acid precipitation. *Atmospheric Environment* 22(12): 2787–2801.

Journel, A.G., and Huijbregts, Ch. J. (1978) *Mining Geostatistics*, Orlando, FL: Academic Press Inc.

Kanciruk, P., Gentry, M.J., McCord, R.A., Hook, L.A., Eilers, J.M., and Best, M.D. (1986) National Surface Water Survey, Eastern Lake Survey – Phase I, data base dictionary. *Technical Report ORNL/TM-10153*, Oak Ridge, TN: Environmental Science Division, Oak Ridge National Laboratory.

Linthurst, R.A., Landers, D.H., Eilers, J.M., Brakke, D.F., Overton, W.S., Meier, E.P., and Crowe, R.E. (1986) Eastern Lake Survey – Phase I, characteristics of lakes in the Eastern United States. Volume I: population descriptions and physico-chemical relationships. *Technical Report EPA/600/4-86/007a*, Las Vegas, NV: U.S. Environmental Protection Agency.

Marcotte, D. (1989) Spatial estimation of frequency distribution of acid rain data using Bigaussian kriging. *Statistical Applications in the Earth Sciences*, pp. 287–296.

Myers, D.E. (1991) Interpolation and estimation with spatially located data. *Chemometrics and Intelligent Laboratory Systems* 11: 209–228.

Myers, D.E., Begovich, C.L., Butz, T.R., and Kane, V.E. (1982) Variogram models for regional groundwater geochemical data. *Mathematical Geology* 14(6): 629–644.

Overton, W.S. (1986) Working draft, analysis plan for the EPA Eastern Lake Survey, March 18, 1985. *Technical Report 113*, Corvallis, OR: Dept. of Statistics, Oregon State University.

Overton, W.S, Kanciruk, P., Hook, L.A., Eilers, J.M., Landers, D.H., Blick, D.J., Jr., Brakke, D.F., Linthurst, R.A., and DeHaan, M.D. (1986) Eastern Lake Survey—Phase I. characteristics of lakes in the Eastern United States. Volume II: lakes sampled and descriptive statistics for physical and chemical variables. *Technical Report EPA/600/4-86/007b*, Washington, DC: U.S. Environmental Protection Agency.

Seo, D., Krajewski, W.F., and Bowles, D.S. (1990a) Stochastic interpolation of rainfall data from rain gages and radar using cokriging 1. Design of experiments. *Water Resources Research* 26(3): 469–477.

Seo, D., Krajewski, W.F., and Bowles, D.S. (1990b) Stochastic interpolation of rainfall data from rain gages and radar using cokriging 2. Results. *Water Resources Research* 26(5): 915–924.

Venkatram, Akula (1988) On the use of kriging in the spatial analysis of acid precipitation data. *Atmospheric Environment* 22(9): 1963–1975.