Environmental Health Perspectives
Vol. 94, pp. 75–82, 1991

Estimating Exposure Using Kriging: A Simulation Study

by Daniel Wartenberg,* Christopher Uchrin,† and Patricia Coogan‡

Retrospective studies of disease often are limited by the resolution of the exposure measurements. For example, in a typical study of adverse health effects from contaminated groundwater, the number of wells sampled may range from only a few to as many as several dozen, while the number of cases and controls may be in the hundreds or more. To derive individual estimates of exposure for wells that were not sampled, investigators must extrapolate. In this study, we compare three methods of extrapolating from a limited number of observations to estimate individual exposures. Using two naive models of groundwater contamination, we compare nearest neighbor interpolation, inverse distance squared weighting, and kriging for estimating exposure based on a limited number of measurements. Our results show that although kriging is a statistically optimal method, it is not markedly better than simpler interpolation algorithms, though it is considerably more complex to use. Aberrant well measurements and discontinuities are problematic for all methods. We provide some guidance in interpolating data and outline a more comprehensive comparison of methodology.

Introduction

The quality and power of retrospective studies of disease, particularly in the context of community toxics, are often limited by the resolution of the exposure measurements. For example, one may wish to study a situation in which a groundwater contaminant that is suspected of causing disease has been detected in some private wells. If one is limited by available resources, as is most often the case, it may not be possible to conduct new field measurements to determine the extent of the contamination. Instead, one often has only the use of an extant set of data collected previously for another purpose on which to base the exposure assessment. The number of wells sampled in the region of concern may range from only a few to as many as several dozen, even though the population of concern may number in the hundreds to thousands to tens of thousands. And yet, to conduct an epidemiologic investigation, one must estimate exposure for each study subject and determine whether there is an association between this estimated exposure and disease. This study evaluates through simulation the effectiveness of a few methods of exposure extrapolation or estimation. We also apply the extrapolation methods to one real data set from an ongoing study of cancer incidence. Unlike studies that assess the accuracy of estimation, we evaluate the impact of the estimates on epidemiologic measures of effect.

*Department of Environmental and Community Medicine, UMDNJ-Robert Wood Johnson Medical School, Piscataway, NJ 08854.
†Department of Environmental Sciences, Rutgers University, New Brunswick, NJ 08903.
‡Department of Environmental Health, Boston University School of Public Health, Boston, MA 02118.

Extrapolating Exposure Data

A variety of strategies can be used to extrapolate or predict from a few samples to many subjects. In their most straightforward application, all assume a relatively (or locally) smooth contaminant surface. For the simulations considered in this study, we constrain the surface to meet this assumption. In one set of extrapolation methods, the investigator fits a geographic function to the entire data set and estimates values based on the fitted surface [e.g., global mean/median (*1*); trend surface analysis (*2,3*)]. In another set of methods, the investigator fits a geographic function to a local set of points and estimates values from the locally fitted surface [e.g., Akima interpolation (*4*); local trend surface analysis (*5,6*); Laplacian smoothing splines (*7*); natural neighbor interpolation (*8*)]. In the third set of methods, the investigator takes a weighted average of some or all observed points to give interpolated values [e.g., inverse distance squared weighting; kriging (*9*)]. Few studies have compared these methods [Laslett et al. (*1*) and Boufassa and Armstrong (*10*) are notable exceptions], although many cite specific weaknesses or limitations (*11–14*). For this study, we pick three methods for estimating contaminant concentrations at unmeasured locations: *a*) assign to the unknown point the value at the nearest observed point (nearest neighbor interpolation (NN); *b*) assign to the unknown point a weighted average (mean) of the nearest k points using an inverse squared distance (ISD) weighting rule; and *c*) assign to the unknown point an estimate derived from kriging (KRG).

Implementation of the first method is straightforward. One calculates the distance from each observed data value to the point to be estimated, chooses the shortest distance, and assigns the value from that closest point to the point to be estimated.

The second method is an approach used in many contouring

programs and geological applications. One calculates the distance from each observed data point to the point to be estimated, selects the nearest k points, calculates the weight for that point as the inverse square of that distance, sums the product of these weights times their respective values, and divides this weighted sum by the sum of the weights. This quotient is assigned to the point to be estimated. Intuitively, this is how we often evaluate maps visually. That is, mentally we take an average of values nearby the point to be estimated, giving more emphasis to those closer by and not allowing far away points to contribute substantially to our intuitive estimate.

The third method is considerably more complex. Kriging is another weighted-average method of estimation that assumes that geographically close samples are more similar than geographically distant ones. It is statistically optimal in that it is a BLUE (best, linear, unbiased estimator). Kriging is the preferred approach of geostatisticians to interpolation and prediction, although it requires investigator intervention and sophisticated programming. A thorough discussion of its use is beyond the scope of this paper. The reader is referred to Journel and Huijbregts (*9*) for an extensive discussion of the method and Jernigan (*15*) for a more elementary presentation.

In brief, to krige a data set, one must first estimate the variogram, or spatial covariance function, of the data. That is, one must estimate how similar each observation is to each of its neighbors, and then one must fit this set of similarities to a mathematical function that increases as the separation distance between point pairs increases. The variogram is then used to derive optimal weights for averaging observations nearby the point to be predicted into the estimated value. It is similar to the second method, ISD weighting, in that the weights decrease as a function of the distance of each observed datum to the point to be estimated. It differs in that each weight, rather than being arbitrarily the inverse of the squared distance, is derived from an observed property of the data, the spatial covariance.

Methods

Simulation Strategy

To compare the utility of these methods for estimating exposure, we consider a hypothetical epidemiological study. In this study, we postulate a contamination scenario, sample the groundwater quality based on this scenario, select a set of cases and controls from the region of the contamination, and calculate both the difference in mean exposure among cases and controls and the odds ratio.

First we postulate a contamination scenario. We do this to enable us to simulate sampling the groundwater and the disease process. We use this scenario to generate cases and controls via our model and to generate groundwater samples. But, we do not use data from this contamination scenario directly for our analyses because it is unrealistic to have a complete assessment of contamination in a study; instead we use samples. For this study, we consider two scenarios. The first scenario has a rectangular plume of contaminant in a rectangular study area, 11 by 11 (Fig. 1*A*). A tongue of contamination extends into the study area from the south, covering the middle half of the southern border and extending halfway through the study area toward the north. All contaminant levels within the contaminant tongue have a concentration of 1. All contaminant levels outside of the contaminant tongue have a concentration of 0.

Now we sample the contaminant field. For this study, we locate 25 randomly placed sampling locations within the study area, determine the "true" concentration based on our contamination scenario, and then include a term for the random inaccuracy associated with the measurement process. Each contamination value should be exactly 1 or 0, but we allow the measurements to vary a bit by adding a random number from a uniform distribution ranging from -1 to 1.

Next, we must pick our study subjects and assign them case or control status. For this study, we have decided to use 50 cases and 50 controls. To find them, we pick random locations throughout our study area. For each location, we evaluate the true contaminant level. Then, we determine if this location is a case or control. We set a disease cutoff based on a background rate of disease and select a dose-response model. For this study, we assume that the background rate of disease is 5%. The dose-response model for this study is linearly increasing risk with increasing dose, with those exposed to 1.0 units of contaminant experiencing a 15% incidence rate or a relative risk of 3. In other words, locations within the contaminant plume have a 3 in 20 chance of being a case and a 17 in 20 chance of being a control. Those outside the plume have a 1 in 20 chance of being a case and a 19 in 20 chance of being a control. Locations are collected randomly until 50 cases and 50 controls have been amassed.

Finally, we must estimate the exposure for each ease and each control based on our samples of the groundwater contamination. For our standard, which we call Truth, we use the true concentration of the contaminant which we know by virtue of having defined it analytically. For the NN method, for each study subject, we select the "measured" value of the contaminant at the closest of the 25 measurements. For the ISD method, we take a weighted average of the 25 measurements. For the KRG method, we first fit a variogram to the contamination scenario and then apply the kriging algorithm to the nearest 8 points. For each of these exposure estimation methods, we calculate the mean estimated exposure among cases and the mean estimated exposure among controls. We also assign each study subject an exposure status (exposed or unexposed) based on whether their measured exposure is greater than 0.5 (exposed) or less than 0.5 (unexposed), and calculate an odds ratio. We report means and standard deviations for 500 replications of each scenario.

For the second simulated scenario, we change the shape and size of the contamination plume (Fig. 2*A*). We assume it to be a paraboloid, with maximum of 1 near the middle of the south border, falling off to 0 halfway toward the southern corners and also toward the middle of the study area. Having defined the contamination scenario, the rest of the procedure follows that for the first simulation.

Real Data: Trichloroethylene in the Ashumet Valley, Massachusetts

This consideration of extrapolation methods for groundwater contamination data was motivated by an ongoing study of an apparent excess of cancer cases on Cape Cod and the possibility of environmental causation. One of the potential environmental agents to which some attribute the cancer excess is groundwater

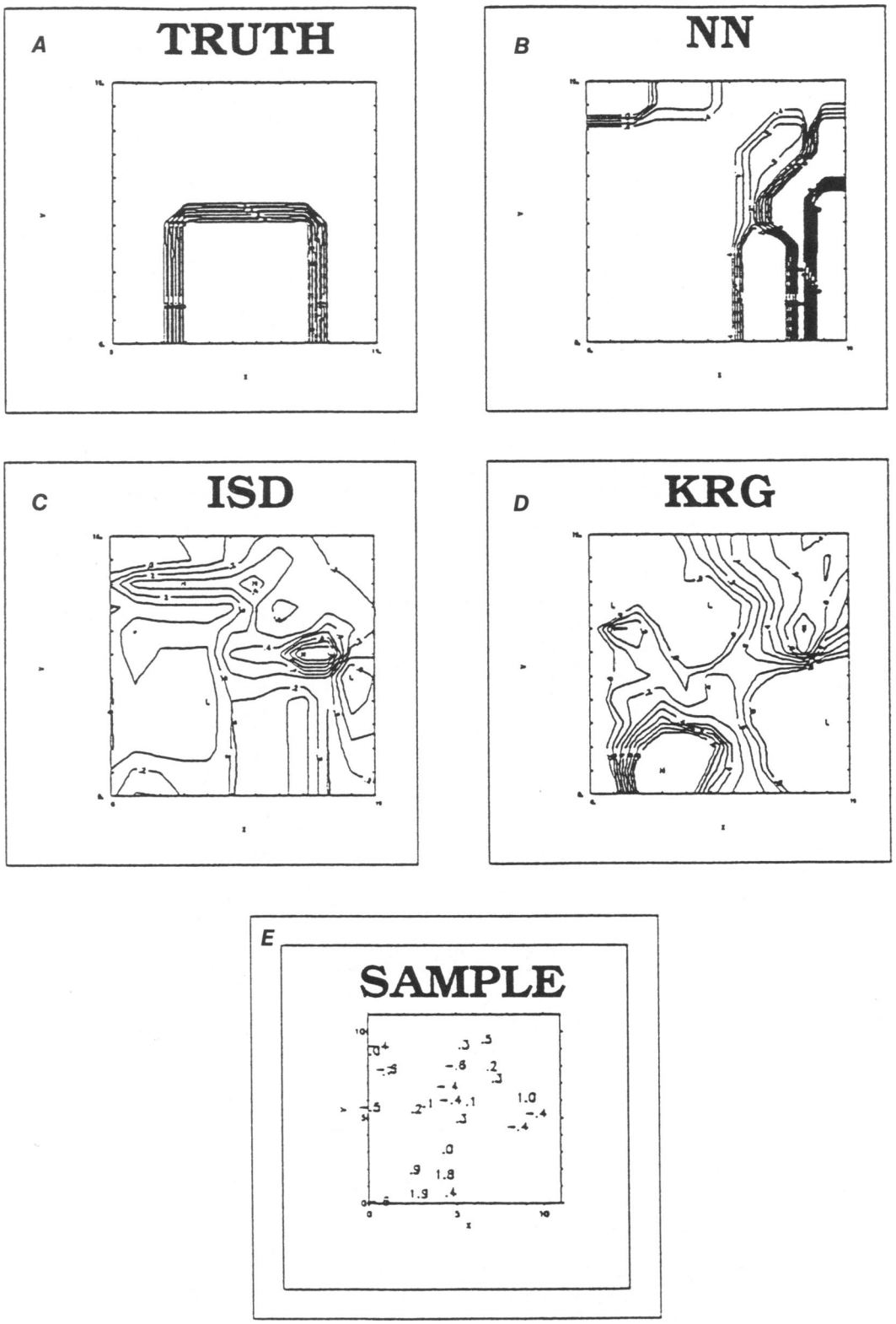

FIGURE 1. The discontinuous plume. This figure shows plots of groundwater concentrations for a contaminant for model 1, the discontinuous plume. (*A*) Plot of the model data. (*B*) Results of nearest neighbor interpolation. (*C*) Result of inverse distance squared weighting interpolation. (*D*) Result of kriging. (*E*) Plot of the sampled data values.

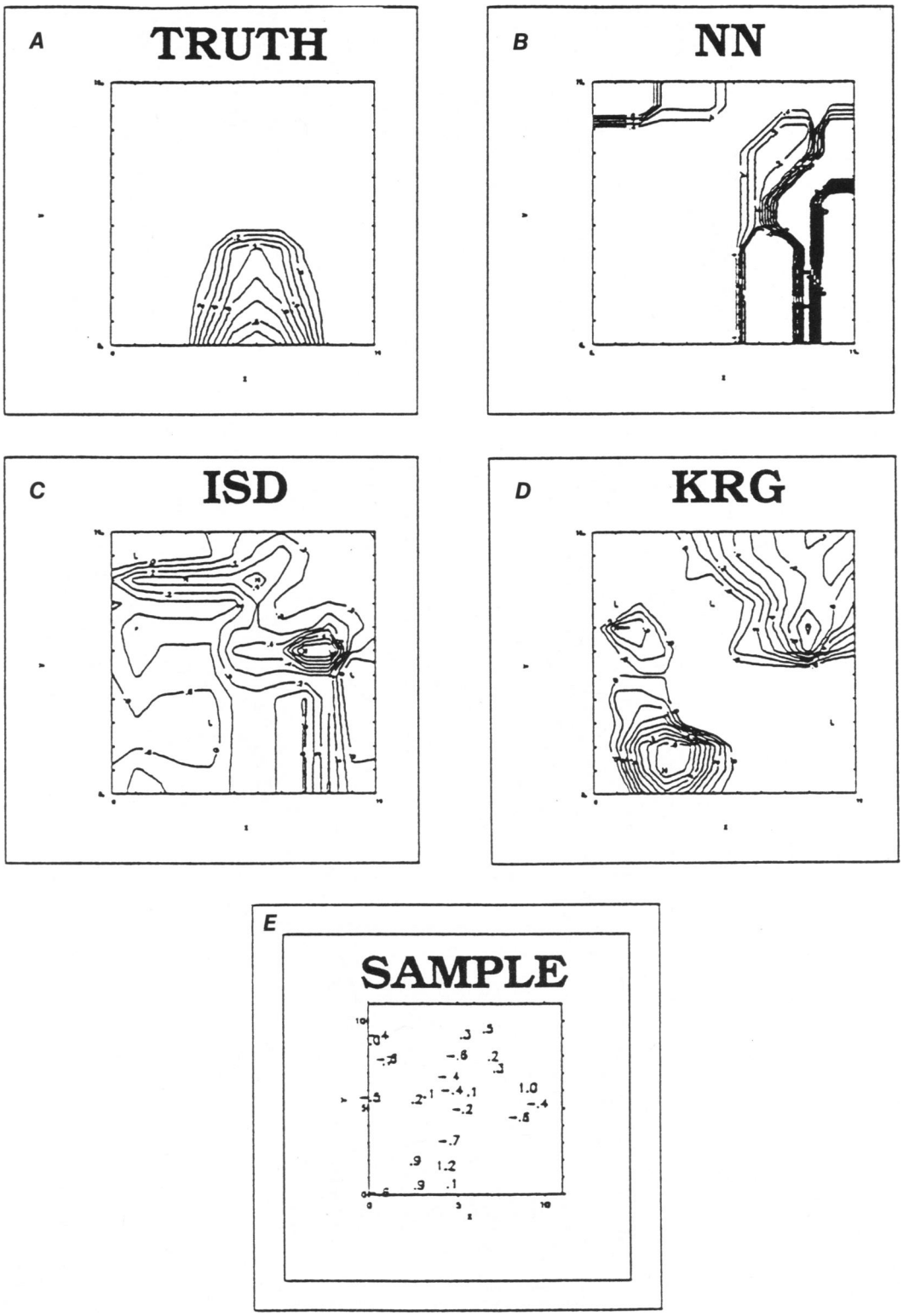

FIGURE 2. The continuous (or parabolic) plume. This figure shows plots of groundwater concentrations for a contaminant for model 2, the continuous (or parabolic) plume. (*A*) Plot of the model data. (*B*) Results of nearest neighbor interpolation. (*C*) Result of inverse distance squared weighting interpolation. (*D*) Result of kriging. (*E*) Plot of the sampled data values.

contamination. Since we were unable to sample the water ourselves due to budgetary constraints, we relied on a set of water quality measurements made by the United States Geological Survey. They have taken a series of measurements of a variety of organic contaminants throughout the region. Cases and controls were identified independently of this information, and exposures had to be assigned to each. There are 59 unique water quality measurements, 1200 cases, and 1500 controls. We apply these same three approaches to these data although in this part of the exercise we cannot assess the truth.

Results

Simulations

The results of the simulations are a bit surprising (Tables 1 and 2). Listed in the tables are mean exposure values for cases and their standard deviations, mean exposure values for controls and their standard deviations, and mean odds ratios and their standard deviations (defining content values of greater than 0.5 as exposed). The row labeled "Truth" is derived by calculating these indices for the sample points based on the analytically derived contamination values. The other three rows are the result of the estimation procedures using sampled contamination values (the true value plus a random error component). Variograms were fit to the contamination scenario rather than the sampled data because of the instability of the estimated variances based on a relatively small number of samples. The fitted variogram for the first situation had a nugget value of 0, a sill of 0.27, and a range of 6. The fitted variogram for the second situation had a nugget value of 0, a sill of 0.06, and a range of 5.5.

For the first simulation, we see that all of the methods underestimated case exposure, overestimated control exposure, and underestimated the odds ratio. Results for all three methods are relatively similar and more similar to one another than to the truth. This suggests, not surprisingly, that one inherent problem in all these estimation procedures is sampling, both the number of samples and their placement. Figures 1*B*, 1*C*, and 1*D* show maps of the contamination based on these procedures for one randomly selected replication. KRG most nearly captures the plume, although all do poorly. Figure 1*E* shows a map of the sampled contamination that was used to draw the three maps.

For the second simulation, again we see that all of the methods underestimated case exposure, overestimated (or equalled) control exposure and underestimated the odds ratio and the results for the three methods differed more from the truth than from each other. NN is noticeably worse than ISD or KRG. Figures 2*B*, 2*C*, and 2*D* show maps of the contamination based on these procedures for one randomly selected replication. Again, KRG most nearly captures the true surface. Figure 2*E* shows a map of the sampled contamination that was used to draw the three maps. It is interesting to note that of the six maps, each map is more similar to the other map made with the same method than to the other maps of the same contaminant scenario. This suggests a consistent bias in estimation. It is partially an artifact of sampling.

Table 1. Simulation results for a discontinuous contaminant plume.[a]

Method	Case exposure	Control exposure	Odds ratio
Truth	0.50 ± 0.07	0.23 ± 0.06	3.83 ± 1.95
NN	0.46 ± 0.19	0.25 ± 0.18	1.72 ± 0.86
ISD	0.42 ± 0.16	0.28 ± 0.15	2.26 ± 1.28
Krige	0.43 ± 0.17	0.24 ± 0.17	2.17 ± 1.21

[a]All values are means and standard deviations of 500 replicate runs.

Table 2. Simulation results for a continuous contaminant plume.[a]

Method	Case exposure	Control exposure	Odds ratio
Truth	0.16 ± 0.04	0.08 ± 0.03	2.31 ± 10.60
NN	0.15 ± 0.16	0.08 ± 0.16	1.33 ± 0.70
ISD	0.14 ± 0.13	0.10 ± 0.13	1.76 ± 1.56
Krige	0.14 ± 0.15	0.08 ± 0.15	1.72 ± 1.31

[a]All values are means and standard deviations of 500 replicate runs.

Real Data

Results of interpolating the real data are shown in Figure 3. The observed data values are shown in Figure 3*A*, the NN estimates in Figure 3*B*, the ISD estimates in 3*C*, and the KRG estimates in Figure 3*D*. Both ISD and KRG show an apparent plume entering from the left side of the map. KRG shows a more smoothly varying function, while ISD shows a steep gradient. NN shows a corresponding steep gradient but no source from the left. KRG has a peak somewhat centrally located from left to right while both ISD and NN show maxima toward the right-hand border. Looking at the data in Figure 3*A* one sees that the distribution of sample sites is somewhat clustered and that the distribution of data values is not smooth. Highs exist toward the bottom right and along a transect from the bottom left to the top right. But, as always, these data are confounded with sampling error, temporal variation, and different well depths, among other problems.

Discussion

At the outset of this discussion, it is important to note the limited scale of this study. This exercise is meant as a preliminary investigation of a problem that warrants more detailed study and as a vehicle for identifying relevant issues for consideration in further work. In subsequent studies, we plan to characterize this problem more thoroughly and make recommendations regarding particular strategies for handling specific data situations.

There are many problems in these simulations that affect all methods. These include the number and placement of samples, the number and placement of subjects, the measurement precision, and the nature of the contamination plume. There also are many assumptions, simplifications, and decisions built into the methodology as presented here that can be controlled by the investigator. Astute choices could improve the performance of any or all of these methods.

First, we review the simulation methodology. The models chosen for the contamination plume were arbitrary and unrealistic. Neither plume corresponds closely to true contaminations in shape or slope. In future studies, we plan to use more sophisticated groundwater models (*16,17*). Measurement variability was assumed independent of contamination value, which probably is inappropriate. And the sample localities were limited to 25 and were randomly placed. All methods would improve with more samples. All methods would give more accurate

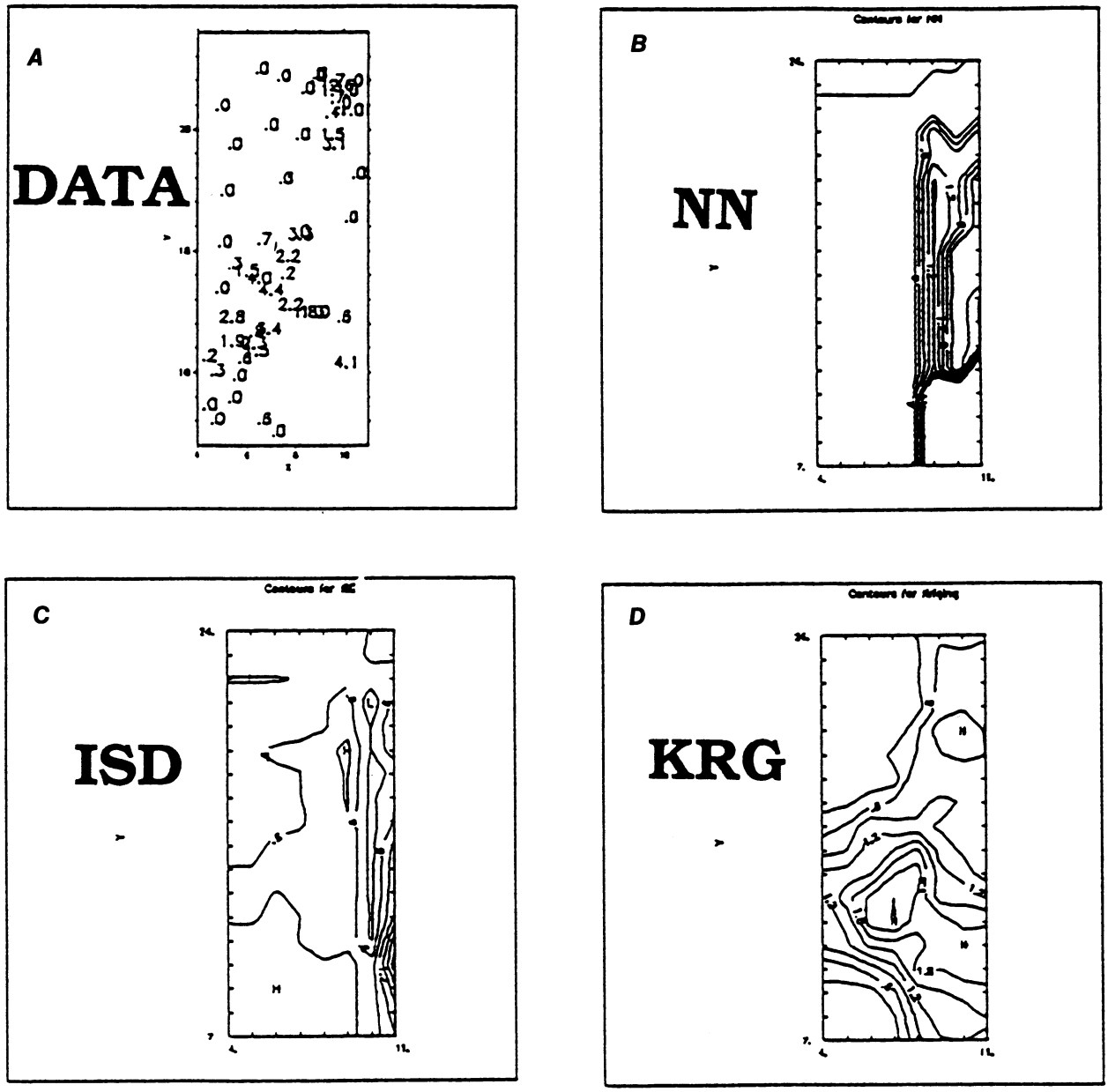

FIGURE 3. Trichloroethylene in the Ashumet Valley, Massachusetts. This figure show plots of groundwater contamination by trichloroethylene in the Ashumet Valley, Massachusetts. All data are plotted in log units. (*A*) Plot of the sampled data values. (*B*) Results of the nearest neighbor interpolation. (*C*) Result of inverse distance squared weighting interpolation. (*D*) Result of kriging.

maps if more samples were placed inside the plume. Unfortunately, we believe our choice of 25 random samples to be realistic of real groundwater problems. Other problems that are not within our control but which are real limitations of using groundwater data in general are the fluctuating groundwater table, differential well depths (unless one wishes to use three-dimensional estimation methods), and the time history of the contamination. The latter may be the most significant as our measures often coincide with or follow the disease events, even though a causative exposure would have to precede it, possibly by many years.

The disease process model we used and the measures of effect also were arbitrary. We could use more subjects. We could consider disease models with exponential or threshold effects. We could consider situations of large relative risks, although a relative risk of 3 seems appropriate for some environmental agents. We could consider a variety of ways for classifying subjects as exposed or unexposed [see, for example, Wartenberg and Northridge (*18*)], and we could use a logistic model in preference to an odds ratio. But we do not believe that these choices would affect the nature of the results reported here. We do believe that a more useful measure of effect would be the power to detect a

statistically significant odds ratio and plan to use this in future investigations. It is possible that even though ISD had a mean odds ratio close to the truth, it would have less power than KRG because the estimates are more variable (higher standard deviation). It is important to remember that the goal of this study is to assess the utility of these methods for use in epidemiologic investigations, not to reconstruct the true groundwater contamination map. Although we make reference to the maps (Figs. 1 and 2), the statistical power of the analysis is of primary interest.

Each estimation method could be tuned to provide better estimates for groundwater contamination. The NN method could be reformulated to take an average of the k nearest neighbors rather than only the value of the nearest, as in this study. But, unless the sampling grid is considerably more dense, this method is not likely to perform well.

The ISD method performed better in terms of the odds ratios than the other two methods for both contamination scenarios, although only marginally better than KRG. In the second scenario, one explanation for this may be that the contamination model was parabolic (i.e., had quadratic terms in which the function value changes as the square of the distance moved) and nearly isotropic (i.e., equal effects in all directions). In other words, the contamination scenario was designed to meet the assumptions of the ISD model. In the first scenario, given the sparse sampling, a similar assumption is not far off. Nonetheless, the estimates and the maps were not terribly accurate, and improvements could be made. For instance, ISD could be restricted so that only the k nearest neighbors are used for the exposure estimates. This restriction is placed on most contouring packages that use this model. It creates a more local estimate. But this should not affect the estimate markedly. One could try other, arbitrary weighting schemes, such as inverse distance or inverse distance cubed. Inverse distance squared sometimes is chosen as analogous to a diffusion process, where materials spread as the square of the distance from the source. Since groundwater flow is advective, inverse distance weighting might be more appropriate. One also could use different weights for different flow directions, for anistropic fields. But this would require a more detailed knowledge of the groundwater system than an epidemiologist is likely to have.

Kriging offers the most opportunity for improvement. The fitting of the variogram, to a large degree, determines the accuracy of kriging estimator. We have not discussed how this was done due to space limitation of this paper, but we can make some recommendations. First, implicit in the kriging model is the stationarity of the data field. Ideally, this can be modeled and removed using trend surface analysis or a related method and the residuals subjected to kriging (*9,19*). Outliers could be identified and removed before further analysis as these can impart undue influence on the variogram (*6,13,19*). In this study, the variogram function was fit to the contamination scenario rather than the samples because there were so few samples. More samples would help. If the variogram were fit to samples, a nugget effect could have been included to account for measurement variance. We can include a nugget effect artificially in further investigations. The neighborhood of points considered in this study was eight. This could be varied as well.

Conclusions

Exposure estimation is a difficult and challenging enterprise. Results are highly data dependent, and without accurate and plentiful samples one cannot get high resolution estimates. The performance of exposure estimators is highly dependent on the underlying surface to be estimated, the number and placement of the samples, and the estimation model used. We found that under particular scenarios, the statistically optimal estimation procedure, as used by a naive user, did not outperform another model more closely aligned to the data distribution. However, many improvements in specification and application could be made and will be explored in future studies.

We thank Hamilton Gilbert for assistance in data preparation. Initially, kriging results and all plots were calculated using the GEO-EAS program available from Evan Englund, Environmental Monitoring Systems Laboratory, U.S. EPA in Las Vegas, Nevada. Simulation programs were written by the authors. This work was conducted with support from the Massachusetts Department of Public Health.

REFERENCES

1. Laslett, G. M., McBratney, A. B., Pahl, P. J., and Hutchinson, M. F. Comparison of several spatial prediction methods for soil pH. J. Soil Sci. 38: 325–341 (1987).
2. Whitten, E. H. T. The practical use of trend-surface analyses in the geological sciences. In: Display and Analysis of Spatial Data (J. C. MuCullough, Eds.), Wiley, New York, 1975, pp. 282–297.
3. Haggett, P., Cliff, A., and Frey, A. Locational Analysis in Human Geography. Edward Arnold, London, 1977.
4. Akima, H. A method of bivariate interpolation and smooth surface fitting for irregularly spaced data points. Algorithm 526. ACM Transactions on Mathematical Software 4: 148–164 (1978).
5. Czegledy, P. F. Efficiency of local polynomials in contour mapping. Math. Geol. 4: 291–305 (1972).
6. Wartenberg, D. Exploratory spatial statistics: Outliers, leverage points and influence curves. In: Spatial Statistics: Past, Present and Future (D. A. Griffith, Ed.), Institute of Mathematical Geography, Ann Arbor, MI, 1990, pp. 133–156.
7. Wahba, G., and Wendelberger, J. Some new mathemaatical methods for variational objective analysis using splines and cross validation. Month. Weather Rev. 108: 1122–1143 (1980).
8. Sibson, R. A brief description of natural neighbour interpolation. In: Interpreting Multivariate Data (V. Barnet, Ed.), John Wiley and Sons, New York, 1981, pp. 21–53.
9. Journel, A. G., and Huijbregts, C. J. Mining Geostatistics. Academic Press, New York, 1978.
10. Boufassa, A., and Armstrong, M. Comparison between different kriging estimators. Math. Geol. 221: 331–345 (1989).
11. Krige, D. G., and Magri, E. J. Studies of the effects of outliers and data trasformations on variogram estimates for a base metal and a gold ore body. Math. Geol. 14: 557–564 (1982).
12. Starks, T. H., and Fang, J. H. The effect of drift on the experimental semivariogram. Math. Geol. 14: 309–319 (1982).
13. Cressie, N. Toward resistant geostatistics. In: Geostatistics for Resources Characterization (G. Verly, M. David, A. G. Journel, and A. Marechal, Eds.), D. Reidel, Dordrect, The Netherlands, 1984, pp. 21–44.
14. Armstrong, M., and Boufassa, A. Comparing the robustness of ordinary kriging and lognormal kriging: outlier resistance. Math. Geol. 20:447–457 (1988).
15. Jernigan, R. W. A Primer on Kriging. U.S. EPA Office of Policy, Planning and Evaluation, Washington, DC, 1986.
16. Uchrin, C. G., and Lewis, T. E. Modelling sorption and degradation of toxic and hazardous substances in ground water systems. In: Chemistry for Protection of the Environment (L. Pawlowski, E. Mentasti, W. J. Lacy, and C. Sarzanini, Eds.), Elsevier, New York, 1987, pp. 153–167.
17. Uchrin, G. G., and Lewis, T. E. A BASIC encoded model for one-dimensional ground water systems incorporating first-order degradation kinetics. J. Environ. Sci. Health A23: 469–482 (1988).

18. Wartenberg, D., and Northridge, M. Dichotomizing exposure in case control studies: a new approach. Am. J. Epidemiol, in press.

19. Cressie, N. Kriging nonstationary data. J. Am. Stat. Assoc. 81: 625–634 (1986).