Commentary

How (Not) to Lie with Spatial Statistics

Luc Anselin, MA, PhD

Taking a cue from Mark Monmonnier's classic,[1] I formulate some cautionary remarks related to the use of methods and software for spatial data analysis, with particular reference to empirical work dealing with cancer prevention and research. Due to length limitations, the discussion will have to be brief, incomplete, and largely nontechnical. For more comprehensive and technical reviews of some of the issues raised, see the articles by Anselin,[2] Greenland,[3] and Wakefield.[4] In this context, I define spatial data analysis broadly as consisting of three important components: exploratory spatial data analysis (ESDA), visualization, and spatial modeling. Although the dividing lines between these areas of interest are not precise, I consider ESDA as concerned with the search for interesting "patterns," visualization as consisting of methods to show these interesting patterns, and spatial modeling as the collection of techniques (also referred to as spatial regression analysis, spatial econometrics) to explain and predict these patterns. Recent overviews of the methodology of spatial statistics and spatial econometrics can be found in Lawson,[5] Anselin et al.,[6] Banerjee et al.,[7] Waller and Gotway,[8] and Schabenberger and Gotway.[9]

The focus on patterns highlights the importance of location and distance, two central concepts in spatial data analysis. Recent methodologic advances in spatial statistics, combined with the ready availability of cheap and powerful desktop geographic information systems (GIS) have brought spatial analysis within reach of many nonspecialists. The array of techniques available can be bewildering, especially because many of them are easily applied through the use of commercial off-the-shelf point-and-click software, without much guidance as to what is appropriate for the situation at hand. This is further complicated by the fact that spatial data can be represented in many different ways (e.g., as discrete spatial objects, such as counties, or as continuous surfaces, such as a risk surface). In addition, a given representation does not necessarily provide insight into the type of spatial process at hand. For example, a point could represent an event, such as the address of a person undergoing cancer screening, or a discrete object, such as the location of a noxious facility, or even a sample point designed to measure a continuous phenomenon, such as an air quality monitoring station. Although these are all points on a GIS map, they each require a distinct statistical approach, respectively referred to as point pattern analysis (events), lattice data analysis (discrete objects), and geostatistics (continuous surfaces). Methods and properties that are appropriate for one type of analysis do not readily transfer to other types of spatial processes. Unfortunately, the GIS (and spatial analysis software) remains largely ignorant about the nature of the underlying process, and simply deals with the data as "points," thereby not preventing meaningless analyses (such as the application of geostatistical analysis to discrete lattice data).

The upshot of this situation is that care is needed in the range of activities involved in spatial data analysis, from the collection of data and the use of software to the interpretation of results and their application in policy analysis. Along the way, choices that yield different results must be made, offering the temptation to tailor the method to the desired result (to lie with statistics). I will briefly comment on a few salient points and important tradeoffs, starting with data problems, methodologic challenges, and software issues, and closing with some remarks on interpretation and policy.

Data Problems

Spatial data include the location of the observation as an essential attribute. This is either recorded in a coordinate system as an absolute location (such as latitude–longitude, or some projected x,y coordinates), or referred to as an administrative entity, such as a census tract or ZIP code zone. In practice, the geographic information about patients or hospitals, for example, is not necessarily available in such form, but is more likely recorded as a street address. The process of translating street addresses to the formal spatial location information is referred to as **geocoding**. Although straightforward to carry out in most commercial GIS, it is also fraught with problems such as inaccurate address information or flaws in the spatial database on street locations. This will result in errors that need to be accounted for in any spatial statistical analysis. Unfortunately, in practice, such errors do not tend to be random, but show systematic spatial variation. For example, more inaccuracies will tend to be found in

From the Spatial Analysis Laboratory, University of Illinois, Urbana, Illinois

Address correspondence and reprint requests to: Luc Anselin, MA, PhD, Spatial Analysis Laboratory, University of Illinois, 333 Davenport Hall, MC 150, 607 South Mathews Avenue, Urbana, IL 61801-3671. E-mail: anselin@uiuc.edu

0749-3797/06/$–see front matter
doi:10.1016/j.amepre.2005.09.015

recently developed suburban neighborhoods than in long established urban blocks, resulting in systematic spatial patterns of failure to match, or missing observations.

Systematic errors in the allocation of street addresses to the proper location have repercussions for the computation of rates (of incidence or mortality) for geographic areas. The rates serve as estimates of the underlying risk by dividing the number of events of interest (cancer cases, number of screenings) by the population at risk (i.e., the population to which the events pertain). The inaccuracy in rates is especially important for small geographic areas where changes of only a few counts in the numerator may yield significant changes in the rate, and thus in the estimate of the underlying risk. Not only is the numerator important in this respect, but also the denominator. Apart from the decennial census years, information on the population residing in a given small areal unit (census tract, ZIP code zone, county) is not regularly available and must be estimated. For larger units, such as counties, this can be fairly reliably done with some degree of detail on gender, ethnicity, and age category. This uses established demographic techniques based on birth and death records and models to estimate net migration. However, the smaller the geographic unit, the less reliable these estimates become. In practice, it is nearly impossible to obtain estimates with any degree of detail about age, gender, and ethnicity in a cost-effective manner at a smaller spatial scale than the county level. By necessity, these estimates have to rely on models (and their assumption), and the prediction error will have to be accounted for in statistical analyses. This has obvious implications for the accuracy of computed rates, especially in areas undergoing rapid demographic transition.

The use of explicit information on the location of individuals, such as their addresses, raises concerns about protecting privacy. Techniques exist to avoid the identification of individuals while retaining important geographic information, but there is no consensus on how this can be used in the reporting of results, such as summary maps. As a result, due to legal and institutional constraints, spatial analysis often has to be carried out at aggregate spatial scales that may not be meaningful for the research question at hand (e.g., the effect of a noxious facility on elevated cancer incidence). Such analyses may suffer from the ecologic fallacy (or modifiable areal unit problem) that conclusions obtained at aggregate levels do not translate to meaningful behavioral interpretations at the micro scale.

Methodologic Challenges

The methodologic focus in the spatial analysis of cancer risk and prevention is on detecting, visualizing, and explaining instances where the distribution of risk is heterogeneous in space. Particular interest lies in identifying locations of significantly elevated risk, trying to relate these patterns to salient explanatory variables (risk factors) and incorporating these insights into policies of prevention and care provision. Several challenges are faced in this endeavor.

The risk estimate, as a rate or proportion, is inherently unstable in the sense that the precision of the estimates is not uniform. This precision is directly related to the size of the population at risk, yielding a high variance in estimates for small areas (in the sense of having a small population). In practice, this means that a high rate does not necessarily imply a similarly high risk, but could be due to the larger variance of the estimate, especially in small areas. This could indicate spurious outliers or suggest heterogeneity in risk when in fact it is uniform. A large statistical literature is devoted to addressing variance instability in rates. Approaches can be divided roughly between spatial aggregation and smoothing. In the former, the instability is "corrected" by grouping small areal units until they reach a threshold population at risk. This is often implicit in agency policies that preclude the disclosure of rates or risk estimates for areas that do not reach a critical population size (e.g., 25,000 or 50,000). Clearly, this approach is impractical if the focus is on processes that operate at small spatial scales. The alternative is to adjust the original "raw" rate estimate, by "borrowing" information or **smoothing**. This is often based on notions from Bayesian statistics, in which a (posterior) estimate is obtained by combining the data with "prior" information. In so-called Empirical Bayes methods, the prior information is extracted from the data itself, for example, using a national or regional risk estimate as a prior in small-area estimation. A large number of methods have been proposed, which can yield sometimes drastically different results. This potentially leads to confusion among uninformed practitioners. An important aspect to keep in mind is that smoothing is essentially a form of modeling, in which a delicate balance is obtained between assumptions imposed by the model (such as a prior distribution, the inclusion of explanatory variables, or functional form) and what the data support. Whereas one method often is selected at the expense of others, sensitivity analysis is important to gain insight into the respective tradeoffs involved. Also, when data are scarce (as in the case of rare events in small areas), the information extracted from them will by necessity be limited and unreliable.

A related issue pertaining to the estimation of rates is the practice of controlling for the effect of known risk factors, such as gender and age distribution. In epidemiologic practice, the **standardization** of rates to a common age/gender distribution is the rule. This so-called standard population is typically estimated at the national level and based on a specific census year. Its use corrects for apparent heterogeneity in rates due

solely to differences in age distribution (e.g., a county dominated by elderly males would, ceteris paribus, have a much higher prostate cancer incidence rate than the national average age profile would suggest). Standardization either applies a reference age distribution to the age-specific rates obtained at the location of interest (direct standardization), or computes the local rate by multiplying its age distribution with a reference risk (indirect standardization). Both methods result in a loss of information (of age-specific risk estimates). More importantly, they assume homogeneity of the relation between risk and age distribution across space (and time), which may be unrealistic. Extending the standardization to other risk factors becomes more tenuous, because the assumption of a constant or proportional relationship across space between the risk and the risk factor may not be warranted. A similar issue is encountered in the use of model-based smoothing, where more complex models inevitably imply stronger assumptions. In the exploratory stage of a spatial analysis, it is better to avoid imposing too many assumptions and instead let the data speak for themselves. Again, different standardization methods may yield different suggestion of outliers or clusters and sensitivity analysis is in order.

Once interesting patterns are identified, it remains a challenge to relate the spatial heterogeneity in (relative) risk to meaningful explanatory variables (risk factors). This is typically carried out by means of regression analysis. Due to the presence of spatial heterogeneity as well as spatial correlation, standard methods do not apply, and one has to make use of specialized techniques of spatial regression analysis (or spatial econometrics). These methods are complex and still constitute a very active area of research in statistics and econometrics. This is particularly the case when it comes to modeling phenomena in both space and time.

An additional complication encountered in spatial data analysis is the so-called change of support problem (COSP). This is present when the spatial scale of measurement for the variables of interest is different, such as point observations on air quality and health statistics collected by census tract. The solution to this spatial mismatch requires the application of spatial interpolation to bring all variables to a common spatial unit of observation (e.g., the census tract). Such spatial interpolation induces measurement errors with complex spatial structure. The resulting additional uncertainty must be properly accounted for in the regression model and other analyses.

Software

The days are long gone when the dearth of spatial analysis software was seen as a major impediment for the application of these techniques in practice. Significant progress has been made, especially during the last decade. Although mainstream commercial statistical software is still limited in its spatial functionality, a large number of freestanding niche packages, applets, macros, and scripts developed for statistical toolboxes and GIS software fill the need. Most of these implementations are noncommercial, developed in the academic world, with considerable research support from agencies such as the National Science Foundation (NSF), the National Institutes of Health (NIH), the Centers for Disease Control and Prevention (CDC), the National Institute of Justice (NIJ), and the Environmental Protection Agency (EPA). A notable private-sector exception is the recently released spatial statistics toolbox in the leading commercial GIS software, ArcGIS 9.0, which consists of a collection of functions written in the open-source Python language.

The multitude of available software tools may confuse the nonspecialist. All too often the inclusion of a technique in a software package suggests that it is the state of the art, which, in a rapidly changing field like spatial statistics, is not always the case. Furthermore, software tends to be limited in the scope of techniques included, which may misrepresent the range of methodologic options (and pitfalls) available to the analyst. There is a tension between user-friendliness of the software and the technical sophistication needed to properly appreciate the assumptions and limitations of the various techniques. Moreover, there are few standards, with little interoperability between the different packages (and GIS software), and considerable duplication.

Software that implements advanced or specialized methods, although it most likely exists, is often hard to find and not always fully documented. The usefulness of software clearinghouses, such as that maintained by the NSF-funded Center for Spatially Integrated Social Science (www.csiss.org), should not be underestimated. However, much needs to be done to provide further standardization and quality control. In this respect, the growing presence of spatial analytical software tools developed in an open-source environment is encouraging. Efforts such as RGeo, organized around the R statistical programming environment (sal.uiuc.edu/csiss/Rgeo), and PySAL, a library of spatial analytical routines written in Python (sal.uiuc.edu/projects_pysal.php), involve a growing community of developers and allow the latest methods to be included in a transparent manner (the source code serving as the ultimate documentation). Still, much remains to be done, especially to provide effective tools to carry out the exploration and modeling of space–time dynamics in a GIS environment.

From Statistics to Policy

In the context of cancer prevention and control, spatial statistical analysis and GIS are only a means and not the end. The use of these methods varies depending on the

policy goals. For example, whereas the spatial analysis of cancer incidence and mortality tends to focus on the etiology of the various diseases, this may be only a tangential goal for a prevention policy. In the latter context, the insights gained from a statistical analysis of spatial distributions can be very fruitful in the implementation of a spatial decision support system. For example, applications of so-called "geo-marketing" techniques can be useful in identifying underserved populations (markets), assessing the spatial distribution of future demand for care, locating medical facilities (e.g., for screening), and targeting messages effectively to change behavior (e.g., to promote screening). The importance of the statistical insights lies in the quantification of the uncertainty associated with various estimates and in exploiting the spatial characteristics of this uncertainty in the decision process.

The current state of the art in spatial statistics is impressive, and substantial progress has been made. Although it may be tempting to translate this into "best practice methods" and institutionalized guidelines for applied research, such as the identification of a cluster, there is little hope for a satisfactory solution along these lines. New techniques are constantly being suggested as well as insights gained into the tradeoffs among different approaches. Any best practice, be it a method or software tool, is likely out of date by the time its approval has passed all the institutional hurdles. Fortunately, the new communication tools facilitated by the Internet allow the development of a community of scholars and practitioners where insights can be shared, clearinghouses provided to methods and software tools, and ongoing training provided. Such a dialogue between practice and research is likely to push the field forward, to result in effective public health policy, and to lessen the scope for "lying with spatial statistics."[9]

This commentary is based in part on a presentation made at a meeting on "GIS Research Priorities for Comprehensive Cancer Control," organized by the Centers for Disease Control and Prevention (CDC)'s Division of Cancer Prevention and Control, Santa Barbara, CA, November 17-18, 2004. This research was supported in part through NSF Grant BCS-9978057 to the Center for Spatially Integrated Social Science (CSISS), and by a Cooperative Agreement between the Centers for Disease Control and Prevention and the Association of Teachers of Preventive Medicine (ATPM), award number TS-1125. The contents of the note are the responsibility of the author and do not necessarily reflect the official views of NSF, CDC, or ATPM.

No financial conflict of interest was reported by the authors of this paper.

References

1. Monmonnier M. How to lie with maps. Chicago: University of Chicago Press, 1996.
2. Anselin L. Under the hood. Issues in the specification and interpretation of spatial regression models. Agric Econ 2002;27:247–67.
3. Greenland S. A review of multilevel theory for ecologic analyses. Stat Med 2002;21:389–95.
4. Wakefield J. A critique of statistical aspects of ecological studies in spatial epidemiology. Environ Ecol Stat 2004;11:31–54.
5. Lawson A. Statistical methods in spatial epidemiology. Chichester: John Wiley and Sons, 2001.
6. Anselin L, Florax R, Rey S. Advances in spatial econometrics, methodology, tools and applications. Berlin: Springer-Verlag. 2004.
7. Banerjee S, Carlin B, Gelfand A. Hierarchical modeling and analysis for spatial data. Boca Raton, FL: Chapman & Hall/CRC, 2004.
8. Waller L, Gotway C. Applied spatial statistics for public health data. Chichester: John Wiley and Sons, 2004.
9. Schabenberger O, Gotway C. Statistical methods for spatial data analysis. Boca Raton, FL: Chapman & Hall/CRC, 2005.