Journal of Official Statistics, Vol. 23, No. 3, 2007, pp. 291–297

Discussion

Linda Williams Pickle[1]

I would like to second Sarah's thanks to Dr. Goodchild for a thought-provoking presentation. It is always valuable to hear a different point of view regarding our statistical data and tools, and it is especially so to hear from someone as distinguished in the field of geography as our speaker. I certainly agree with Mike's assertion that there has been a "spatial turn" in several disciplines, including the social sciences and public health, in the past decade. For example, in the late 90s, papers started to appear demonstrating the importance of community characteristics with regard to disease rates, over and above individual patient effects, leading to much more interest in geographic analyses recently.

I would like to begin my remarks by discussing a simplified process for spatial data analysis. Then I will show a few examples of spatial data in cancer research, but I will focus my remarks on spatial statistical analysis and end with a few comments on future directions.

1. A Process for Spatial Data Analysis

A simplified process for spatial data analysis is to first collect, format and store geographic (i.e., location-specific) data, usually in a GIS, then apply statistical methods to analyze that data, interpret the results and then display them using geovisualization techniques. Social science theory can inform the study design at each of these steps by suggesting data, hypotheses and interpretations of results. Mike's presentation addresses these steps separately, but in fact this process is usually nonlinear, with interplay between the steps (Figure 1). For example, a typical analysis would include the following steps:

- Use a GIS to store the geographic data.
- Perform an exploratory analysis of the data, including mapping the original values and perhaps a test for clustering at different spatial resolutions and a check of the stationarity assumption.
- Following that, we often need to transform some of the variables or we may want to calculate new variables using the GIS.
- After constructing the final dataset, apply the appropriate statistical methods, e.g., a confirmatory hypothesis test, model-based estimation or prediction.
- An important component of model-based methods is plotting and mapping the residuals to see if there is any remaining pattern in them that suggests an inadequate model. Usually we need to modify the model and rerun several times, looping

[1]StatNet Consulting, LLC, 20203 Goshen Road, No. 189, Gaithersburg, MD 20879, U.S.A. Email: Linda@statnetconsulting.com

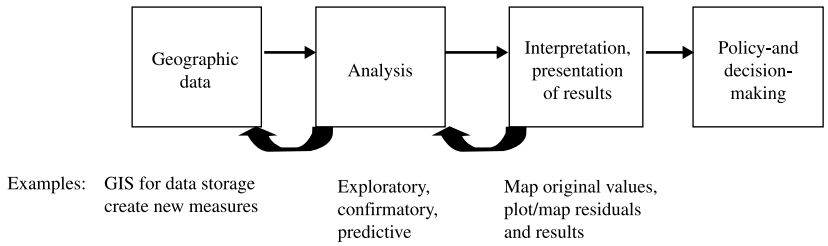


Fig. 1. A functional diagram of spatial statistical analyses

between the analysis and presentation steps until we are satisfied that we have the most accurate results.

- A recent addition to this analytic process, as Mike said, is the application of this new scientific knowledge to policy- and decision-making.

Mike demonstrated how interest in geographic data analysis (first and second points above) has greatly increased in the social sciences recently. At the same time, there has been a parallel "explosion" of advances in statistical methods for spatial data. For example, ten years ago, only a handful of papers were presented on spatial statistical methods at the Joint Statistical Meetings; this year there were at least five entire sessions on this topic.

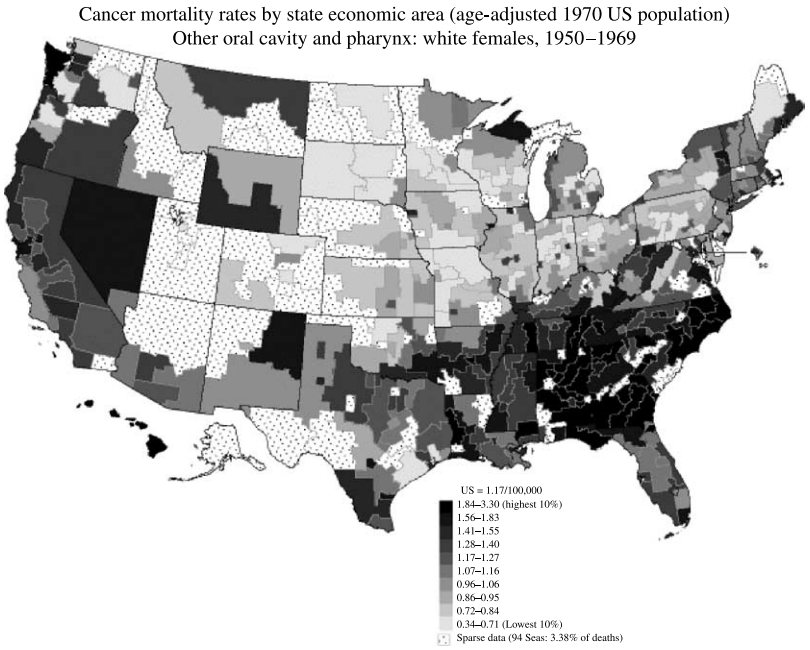
2. Examples of Place-based Analysis

First, I would like to address questions raised by the discussion of nomothetic versus idiographic science, using examples from cancer epidemiology to illustrate:

- Can we generalize from knowledge at distinct locations, or is every place unique?
- Are descriptive methods second-rate or can they be useful in the analytic process?
- How can results of spatial statistical analyses inform policy-making?

Figure 2 is a map of oral cancer mortality rates for white women, 1950–69 (Devesa, Grauman, Blot, Pennello, Hoover, and Fraumeni, Jr. 1999). This map is descriptive but the first NCI cancer atlas published in 1975 also included results of a statistical test comparing the local rate to the U.S. rate (Mason, McKay, Hoover, Blot, and Fraumeni, Jr. 1975). The darkest areas have rates in the highest 10% of the mapped areas. There is clear geographic clustering of high rates in the southeastern states.

Can we generalize to the entire U.S. from findings only in the S.E., or is the S.E. unique? The original map identified specific areas where NCI epidemiologists could conduct interview studies with oral cancer cases and controls. The working hypothesis for the study design was that the higher risk was due to exposure to textile mill dust. However, the generalization resulting from the study was that the carcinogenic components of smokeless tobacco (specifically snuff) cause extremely high cancer risk among long-term users at the exact site where the tobacco was in contact with gum tissue (Winn, Blot, Shy, Pickle, Toledo, and Fraumeni, Jr. 1981). Policy changes implemented after this study were (1) a ban of sales of smokeless tobacco to minors, and (2) campaigns to stop smokeless tobacco use among role models for young people, such as baseball players.

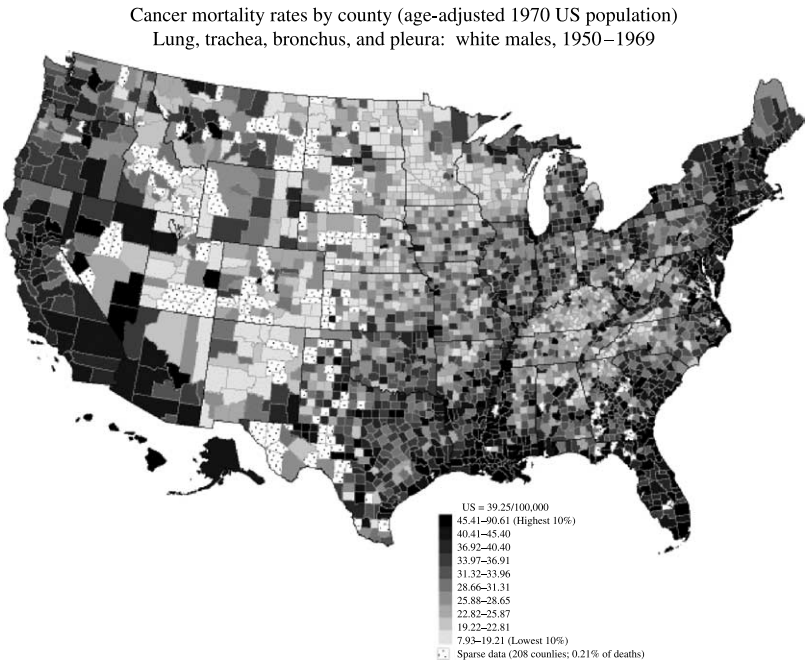


Source: Devesa, Grauman et al. (1975),
Atlas of cancer mortality for U.S. counties: 1950–1969.

Fig. 2. Average annual age-adjusted mortality rates for oral cancer among white females, 1950–1969

A similar map of white male lung cancer mortality (Figure 3) did not have obvious geographic clusters of high rates. However, there was clustering in ATTRIBUTE space – most of the "hot spots" along the Gulf and south Atlantic coasts were port cities during World War II. The working hypothesis for a series of follow-up studies was that the high risks in these port cities were partly due to airborne pollution from the petrochemical industry. However, statistical analyses found that, after controlling for smoking, a greater risk was associated with occupational exposure to asbestos during shipyard work (Blot, Davies, Brown, Nordwall, Buiatti, Ng, and Fraumeni, Jr. 1982; Blot, Harrington, Toledo, Hoover, Heath, and Fraumeni, Jr. 1978; Blot, Morris, Stroube, Tagnon, and Fraumeni, Jr. 1980; Harrington, Blot, Hoover, Housworth, Heath, and Fraumeni, Jr. 1978). The consistency of risk factors identified in multiple geographic areas strengthened the findings from the individual studies by providing evidence that the statistical analyses were correct. Eventually this body of evidence led to laws for asbestos containment and abatement.

So, to summarize, descriptive maps can identify places where follow-up epidemiologic studies can be conducted and can suggest causal factors to be tested in these studies. Although some would argue that each place has a unique, sometimes nonquantifiable, set of characteristics, the effects that many risk factors have on subsequent disease are common across geography, making generalization possible. That is, unless there are complex interactions at work, what makes a place unique is its set of characteristics, not the effect of each of these characteristics on disease, and so statistical models can be used to quantify disease risks across diverse localities. Finally, we have illustrated how spatial statistical analyses can result in policy decisions.



Source: Devesa, Grauman et al. (1975),
Atlas of cancer mortality for U.S. counties: 1950–1969.

Fig. 3. Average annual age-adjusted mortality rates for lung cancer among white males, 1950–1969

We now have GIS and related tools to add value to these sorts of health studies. A GIS can provide information about potential exposures that cannot be obtained through traditional epidemiologic methods, such as by personal interview. The first example is a study in Nebraska that demonstrated the use of remote sensing to reconstruct historical crop patterns for environmental exposure assessment (Ward, Nuckols, Weigel, Maxwell, Cantor, and Miller 2000).

Landsat satellite images were matched to farmers' crop reports at a particular time point to determine the reflectance signatures on the images that identified certain crops. Images from other places at other times could then be translated to crop maps. Results are estimates of the likelihood of exposure to particular pesticides at each location; for example, likely exposure to atrazine. The underlying assumption that each farmer uses the same type and "dose" of pesticide for each crop has been questioned. This method, though, does provide an estimate of the probability of pesticide exposure that cannot be obtained in any other way, especially back in time. Landsat data are now available for 30 years, making this a useful resource for diseases with long exposure-to-disease lag times, such as cancer.

3. Spatial Statistical Analysis

Moving to more analytic issues, Mike mentioned a number of sources of uncertainty: measurement error in mapped outcome variables, lack of replicability in defining classes (interrater disagreement), imprecise boundary definitions, and location variability due to

the earth's axis wobble and tectonic movement. For most public health studies, measurement error of the outcome variable is most important. Variation due to uncertainty of class assignment could be a problem, e.g., classification of stage or other clinical definitions that are somewhat subjective. Because of privacy concerns, much of our data is aggregated to predefined administrative units, so boundary definitions and location variability are not usually a concern. I would suggest adding other sources of error to the list:

- Random (or statistical) error, for example variations in disease rates year by year, that is usually easily accounted for in the analysis.
- Uncertainty due to choice of the appropriate statistical model, not so easily quantifiable.
- Measurement errors in the covariates. If these uncertainties can be quantified, errors-in-covariates statistical models can take this into account.
- Location errors due to geocoding problems, which I will illustrate.

A statistician is often unaware of some sources of error that would be obvious to a geographer. For example, an analyst may accept the dataset as having exact locations of residence of cases, but in fact errors in geocoding (that is, assigning a location to an address) may confound the results. Maps of prostate cancer incidence in Virginia from a recent study showed only slight differences over time (1990-94 vs 1995-99) but there were major pattern differences between county- and census tract-level maps (Oliver, Matthews, Siadaty, Hauck, and Pickle 2005). At first, we thought that this was a scale effect. We knew the county of residence from medical records of all the patients, but 26% of the addresses could not be geocoded to tract, mostly because they had Rural Route addresses. We found that these tract assignments were not missing at random – the 'geocodability' varied over space (lower success rate in rural western counties) and over time (there was a reduction of Rural Route addresses over time). When we restricted the county rate maps to include only those cases whose address could be geocoded to tract, the county and tract patterns looked much more similar. This example of geographic confounding argues strongly for a team approach to spatial data analysis, including a geographer as well as a statistician.

Of the other characteristics of geographic data mentioned that affect statistical analyses, spatial dependence and stationarity concerns are important for most applications. The relationship between spatial dependence and scale is a concern – for example, if I see clustering at a regional level, is there necessarily clustering at a local level? If not, what does this tell me about the underlying reasons for the clustering?

Strong stationarity means that the joint distribution of a process only depends on the relative, not the actual, locations of observations. I agree with Mike that this rarely holds. However, I disagree that this precludes statistical analysis. In contrast, weak stationarity requires a constant mean over all locations $\{s\}$, and that the covariances and variances of pairs of observations only depend on the distances between them, not on their actual locations. That is, $\mathrm{Var}[Y(s+h) - Y(s)] = 2\gamma h$. This leads to a semivariogram plot of 1/2 of the pairwise variances versus the binned distances h. Weak stationarity is assumed for many statistical methods, but for model-based analysis we only need this assumption to hold for the residuals, even if it does not for the original outcome variable. This is much easier to achieve.

There are new geostatistical tools available to help us assess whether weak stationarity is a reasonable assumption. For example, ESRI's Geostatistical Analyst™ presents a color-coded diagram of the semivariogram, along with the usual semivariogram plot, and allows us to try fitting the plot with one of many functional forms. This is an example of the use of a GIS tool within a statistical analysis.

Spatial heterogeneity of the geographic pattern due to population variation is usually not of interest, so we can adjust or weight the analysis to remove it. For example, disease counts (d_i) that are Poisson distributed over various population units (n_i) are probably not stationary, but the rates (d_i/n_i) probably are.

I mentioned earlier that there have been many recent advances in spatial statistical methods. This is particularly true for model-based methods. Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

There are two covariates, and i indexes the locations. The simplest form of this model is to assume that the errors are independently and identically distributed: $\varepsilon_i \sim N(0, \sigma^2)$. This fails to account for the spatial autocorrelation in the data, so a better choice is to assume an error variance-covariance matrix that describes the spatial dependencies between pairs of observations: $\varepsilon_i \sim N(0, \Sigma)$. A common goal of the analysis is to include the necessary covariates to account for the spatial autocorrelation in the data, removing the need to estimate the complex spatial covariance matrix.

We can include additional uncertainty in the model by adding random effects. For example, if the X_2 covariate is more uncertain than X_1, we could let its regression coefficient be a random effect, specifying its presumed distribution:

$$y_i = \beta_0 + \beta_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \text{ where } b_2 \sim N(\beta_2, \Omega) \text{ and } \varepsilon_i \sim N(0, \Sigma)$$

Random effects can also be used to allow parameter estimates to vary over space. Spatio-temporal statistical models are extensions of spatial statistical models, but with possible temporal autocorrelation as well as spatial autocorrelation.

4. Future Directions

Finally, thinking about future directions, I certainly agree that the public is increasingly familiar with geographic information and presentation methods, such as isopleth weather maps, Google Earth-type animation which is being used for local news and traffic reports, and more sophisticated maps in newspapers and magazines, such as the dot density map of population that was on the front page of *U.S.A. Today* in October 2006.

Social science applications are expanding in a number of areas of public health. For example, the cancer control continuum describes activities from the prevention of cancer through detection, diagnosis and treatment to survivorship. There are many areas across this continuum where location is beginning to play a major role. One area of interest is the examination of effects of neighborhoods on the utilization of cancer screening services, such as availability of public transportation and distance to the nearest screening facility. One of the major crosscutting issues receiving a lot of interest and funding at NIH is the social determinants of health disparities.

Again, I would like to thank Mike for a stimulating talk that I hope will encourage more collaboration between geographers and statisticians on data analyses, not only in social sciences, but in other disciplines as well.

5. References

Blot, W.J., Davies, J.E., Brown, L.M., Nordwall, C.W., Buiatti E., Ng, A. and Fraumeni, J.F. Jr. (1982). Occupation and the High Risk of Lung Cancer in Northeast Florida. Cancer, 50, 1982 364–371.

Blot, W.J., Harrington, J.M., Toledo, A., Hoover, R.N., Heath, C.W. and Fraumeni, J.F. Jr. (1978). Lung Cancer after Employment in Shipyards During World War II. The New England Journal of Medicine, 299, 620–624.

Blot, W.J., Morris, L.E., Stroube, R., Tagnon, I. and Fraumeni, J.F. Jr. (1980). Lung and Laryngeal Cancers in Relation to Shipyard Employment in Coastal Virginia. Journal of the National Cancer Institute, 65, 571–575.

Devesa, S.S., Grauman, D.J., Blot, W.J., Pennello, G.A., Hoover, R.N. and Fraumeni, J.F. Jr. (1999). Atlas of Cancer Mortality in the United States: 1950–1994. NIH Publication No. 99–4564. Bethesda (MD), National Cancer Institute.

Harrington, J.M., Blot, W.J., Hoover, R.N., Housworth, W.J., Heath, C.W. and Fraumeni, J.F. Jr. (1978). Lung Cancer in Coastal Georgia: A Death Certificate Analysis of Occupation: Brief Communication. Journal of the National Cancer Institute, 60, 295–298.

Mason, T.J., McKay, F.W., Hoover, R.N., Blot, W.J. and Fraumeni, J.F. Jr. (1975). Atlas of Cancer Mortality for U.S. Counties: 1950–1969. Bethesda, MD: U.S. Department of Health, Education, and Welfare.

Oliver, M.N., Matthews, K.A., Siadaty, M.S., Hauck, F.R. and Pickle, L.W. (2005). Geographic Bias Related to Geocoding in Epidemiologic Studies. International Journal of Health Geographics, 4, 29.

Ward, M.H., Nuckols, J.R., Weigel, S.J., Maxwell, S.K., Cantor, K.P. and Miller, R.S. (2000). Identifying Populations Potentially Exposed to Agricultural Pesticides Using Remote Sensing and a Geographic Information System. Environmental Health Perspectives, 108, 5–12.

Winn, D.M., Blot, W.J., Shy, C.M., Pickle, L.W., Toledo, A. and Fraumeni, J.F. Jr. (1981). Snuff Dipping and Oral Cancer among Women in the Southern United States. The New England Journal of Medicine, 304, 745–749.

Received April 2007