Journal of Official Statistics, Vol. 23, No. 3, 2007, pp. 285–289

Discussion

Using Geospatial Information Resources in Sample Surveys

Sarah M. Nusser[1]

1. Introduction

Geographic information plays an integral role in conducting sample surveys. Most sampling and observation units are connected in some way with geographic space, and it is often of interest to summarize results by geographic regions. In recent years, digital geospatial data have been introduced into several components of the survey process. While advances have primarily focused on revamping existing methods, new opportunities are now being explored for creating sampling frames and collecting geospatial data objects that describe human behavior over time and space.

At the same time, the bridge between survey methods and geoscience is not fully developed, and as statisticians, we have much to learn about the meaning of a characteristic associated with an individual location or small region on the earth. The total survey error framework (see e.g., Biemer and Lyberg 2003) has been useful in describing the specific meaning of a variable as it relates to the data generation process and the theoretical constructs that motivate measurements. This perspective is applicable to thematic variables stored in geospatial data layers, but models may need to incorporate error in location coordinates. Positional error affects the quality of the attribute variable in the geospatial data layer and any variables derived from manipulating source layer(s).

This discussion extends the comments offered by Goodchild by considering current developments in the use of geospatial data for sampling and data collection in sample surveys, with an emphasis on the interface between geoscience and statistics. We begin with a brief discussion of error in geospatial locations, and then review uses of digital geospatial data in sampling and data collection.

2. Errors in Geospatial Locations

Researchers have long used place identifiers (e.g., Census tract, county, state) in sampling and analysis. For example, county-level identifiers are used to merge socio-economic and

[1] Department of Statistics and Center for Survey Statistics and Methodology, Iowa State University, Ames, IA 50011-1210, U.S.A. Email: nusser@iastate.edu

Acknowledgments: I thank the Morris Hansen Committee for their invitation to discuss Professor Goodchild's article. Some research cited in this discussion was supported in part by Cooperative Agreement No. 68-3A75-4-122 with the USDA Natural Resources Conservation Service and by the National Science Foundation under Grant No. EIA-9983289, a portion of which was funded by the U.S. Bureau of Labor Statistics, National Agricultural Statistics Service, and U.S. Census Bureau. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the U.S. Bureau of Labor Statistics, National Science Foundation, National Agricultural Statistics Service, Natural Resources Conservation Service or U.S. Census Bureau.

health statistics from different information resources to create a frame or to augment survey data for analysis. Because place identifiers are usually well-defined, our view of measurement error has largely been restricted to the characteristics associated with place names.

With geospatial data sources and the tools available via Geographic Information System (GIS) software, more complex linkages can be pursued. In this setting, characteristics may be associated with a specific location represented by a coordinate, and combining data sources is conceptually as simple as draping two sources of information over each other in a GIS. However, the coherency of the integrated data is strongly affected by factors such as the choice of the coordinate system, the method used to project the earth's surface onto a plane, the resolution of the source information, and so on. Steinberg and Steinberg (2006) provide an introduction to these factors in the context of social science.

For many geospatial data sources, models describing the error structure of analysis variables must also consider errors associated with an attribute's location. Methods for estimating spatial models in the presence of positional errors are an active topic of statistical research (e.g., Cressie and Kornak 2003; Zimmerman 2007). For many data sources, limited information is available to assist the modeling process. Metadata standards have been developed by geographic scientists for describing numerous features of a geospatial data source such as its origin, date, scope, location parameters, attribute parameters, and so on (e.g., Content Standard for Digital Geospatial Metadata FGDC-STD-001-1998 created in the U.S. by the Federal Geographic Data Committee, ISO 19115 created by the International Organization for Standards). For statisticians, however, the geography-centric specification may be insufficient for statistical uses such as modeling of errors in the positions and thematic content of geospatial databases.

3. Geospatial Data and Probability Sampling

Area sampling is inherently tied to geographic space. Even if we think of the sample frame as lists of area segments, the root of the area sample frame is a map. The most familiar example comes from the U.S. Census Bureau, which creates area sampling materials by linking its Master Address File with its TIGER (Topologically Integrated Geographically Encoded and Referenced) database, a geospatial representation of Census geography and other features such as transportation networks, streams and water bodies. Less familiar is the use of digital aerial photographs in GIS to delineate the borders of area sampling units using the boundaries of physical features. The U.S. Department of Agriculture (USDA) National Agricultural Statistics Service uses this approach to area sampling. The resulting segment boundaries are printed on an aerial photograph for each segment to assist interviewers in finding the area segment and communicating with a respondent about land that he or she farms within the area segment.

For social surveys, most area sampling frames rely on whatever geographic regions are provided in the source database (e.g., data from the latest decennial census). Detailed geospatial coverages offer the potential to create sampling units and size measures that are more tailored to survey objectives and target population characteristics. For example, unique characteristics of the survey topic or study population can be addressed by developing models that use geospatial data layers as covariates to predict more relevant values for size measures or the determination of geographic units.

For surveys where a point on the land is a link to an observation unit, individual coordinates can be sampled irrespective of ground features. If sample points are repeatedly observed over time via remote sensing or field visitation, however, difficulties arise due to errors inherent in Global Positioning System (GPS) locations or in materials from different time points on which sample point coordinates are displayed. For example, because of errors in coordinates, overlaying the original sample point coordinate on an image for a new survey can indicate a slightly different physical location than it did on a prior image. Although it is tempting to think of the digital coordinate as the definition of the sample point, in repeated observations, the data collected in prior years are a more important determinant of where to collect new data for that sample point. That is, the features on the land define the physical location of data collection, not the digital representation of the coordinate. This can be difficult to understand if one (incorrectly) assumes that coordinates are accurate in the absolute sense. Similarly, because the quality of a GPS location is affected by several environmental factors (e.g., satellite configuration, physical barriers), using a coordinate from a GIS database as the target for a GPS receiver will not necessarily lead a data collector to the same physical location over repeated visits. Once again, the context of the prior survey provides the definition of where the repeated observation should be taken.

Geocoding of addresses is another use of point-level data in sampling. In this process, the coordinates of addresses are obtained via fieldwork or via overlays with digital maps or imagery. For example, the U.S. Census Bureau plans to capture geographic coordinates using GPS receivers as part of the 2010 Decennial Census address canvassing effort. Survey organizations also use geocoding services to attach coordinates and other location characteristics (e.g., county, zip code) to addresses sampled from U.S. Postal Service delivery sequence files (Iannacchione et al. 2003; Link et al. 2006). When the location of the dwelling is primarily used for navigating to the sampled unit, errors in the coordinate are less troublesome than if the intent is to model survey data associated with the geocoded households or to combine these data with continuous space geospatial layers. Zimmerman and colleagues consider a class of models for positional errors for geocoded addresses based on a study in a rural county (Zimmerman et al. 2007), along with an analysis approach that uses coarsened locations to address the problem (Zimmerman 2007).

4. Survey Field Data Collection

Field data collection for social surveys increasingly relies on digital maps and GPS on mobile computers for planning field work and finding sample units. The availability of GPS and digital maps on field computers improves the availability of maps and the flexibility in how they can be used. Small studies have shown that these resources may lead to improved accuracy and efficiency in address canvassing settings (Murphy and Nusser 2003).

Thus far, capture of land characteristics is of limited interest in social surveys, although this may become important for surveys that focus on environmental issues. Research in computer-assisted survey methods for a USDA natural resource survey called the National Resources Inventory (NRI) has uncovered several challenges in creating survey instruments to collect geographic objects on computers (e.g., delineate boundaries of features) (Nusser 2004, Nusser 2005). The flow from section to section can easily be controlled for a GIS-based survey instrument in a way that facilitates the survey process. It is also tempting to implement a highly constrained flow for delineation of individual features, as is done in a questionnaire-based computer-assisted survey instrument. However, using GIS software to capture feature boundaries is inherently nonlinear and interactive, and the utility and usability of a GIS survey instrument is severely compromised if a strict linear flow is imposed. Another issue is that proprietary GIS software systems (e.g., ESRI ArcGIS) are designed to accommodate a wide range of ad hoc analyses, offering a variety of tools that are far broader and often more complicated than is desirable in a survey setting. Custom-developed GIS survey instruments are necessary to create a simple and usable interface for consistent use of a limited set of tools for survey data collection.

If repeated observations of geographic features on imagery are of interest, a geographically unorthodox approach may be required to meet the statistical criterion that data be collected in the same physical location. For example, the NRI involves repeated observations of area features (e.g., streams, water bodies, transportation networks) on a sequence of high resolution digital aerial photographs. In the first survey, features are delineated onto an aerial photograph that has been orthorectified (a process that standardizes the representation of the image on a plane). In subsequent years, if the feature overlaid on the next survey's orthorectified image does not fall in the same physical location depicted by initial survey's image, then the new image is aligned to the delineated features from the prior observation so that they are properly located on the new survey's image. This preserves the statistical criterion that feature areas are measured in the same coordinate system from survey to survey (i.e., areas are measured in the same metric), but it is unappealing for geographers because it corrupts the quality of the orthorectification transformation.

An emerging area for social scientists is the use of location sensors to study human behaviors. This ranges from large-scale travel patterns (e.g., Kwan and Lee 2004) to small-scale social interactions (e.g., Chen et al. 2007). This kind of methodology is still rare in sample surveys, but examples of surveys that incorporate sensor-based methods are beginning to emerge. For example, in the U.S., the 2005–2006 National Health and Nutrition Examination Survey collected seven days of physical activity data using an accelerometer from a sample of individuals.

5. Summary

As the availability of geospatial data and analysis tools continues to expand, new and exciting opportunities will arise for both the data collection and analysis phase of sample surveys. Sampling is being modernized and innovative methods are emerging for data collection, including a new class of objects that capture human behavior. Perhaps the most significant problem we face is the wide gap in the terminology and modeling approaches used by statisticians and geographers. Effective use of geospatial resources in sample surveys will require statisticians and survey methodologists to become more familiar with the underlying characteristics of geospatial information. A stronger link between geoscience and statistics will ultimately benefit both fields.

6. References

Biemer, P.P. and Lyberg, L.E. (2003). Introduction to Survey Quality. New York: Wiley.

Chen, D., Yang, J., Malkin, R., and Wactlar, H.D. (2007). Detecting Social Interactions of the Elderly in a Nursing Home Environment. ACM Transactions in Multimedia Computing, Communications and Applications, 3(10). Available at: http://portal.acm.org.

Cressie, N. and Kornak, J. (2003). Spatial Statistics in the Presence of Location Error with an Application to Remote Sensing of the Environment. Statistical Science, 18, 436–456.

Iannacchione, V.G., Staab, J.M., and Redden, D.T. (2003). Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey. Public Opinion Quarterly, 67, 202–210.

Kwan, M.-P. and Lee, J.-Y. (2004). Geovisualization of Human Activity Patterns using 3D GIS. In Spatially Integrated Social Science, M.F. Goodchild and D.G. Janelle (eds). New York: Oxford University Press, 48-66.

Link, M.W., Battaglia, M.P., Frankel, M.R., Osborn, L., and Mokdad, A.H. (2006). Address-based versus Random-Digit-Dial Surveys: Comparison of Key Health and Risk Indicators. American Journal of Epidemiology, 164, 1019–1025.

Murphy, E.D. and Nusser, S.M. (2003). Evaluating User Interfaces for Accommodation of Individual Differences in Spatial Abilities and Way-Finding Strategies. Proceedings of the UAHCI 2003: 2nd International Conference on Universal Access In Human-Computer Interaction, 4, 1005–1009.

Nusser, S.M. (2004). Computer-Assisted Data Collection for Geographic Features. Proceedings of the American Statistical Association, Section on Survey Research Methods. [CD-ROM].

Nusser, S.M. (2005). Digital Capture of Geographic Feature Data for Surveys. Proceedings of the 2005 Federal Committee on Statistical Methodology Research Conference. Available at: http://www.fcsm.gov/05papers/Nusser_IXA.pdf.

Steinberg, S.J. and Steinberg, S.L. (2006). Geographic Information Systems for the Social Sciences: Investigating Space and Place. Thousand Oaks, CA: Sage Publications, Inc.

Zimmerman, D.L. (2007). Estimating Spatial Intensity and Variation in Risk from Locations Coarsened by Incomplete Geocoding. Provisionally accepted by Biometrics.

Zimmerman, D.L., Fang, X., Mazumdar, S., and Rushton, G. (2007). Modeling the Probability Distribution of Positional Errors Incurred by Residential Address Geocoding. International Journal of Health Geographics, 6, 1. Available at: http://www.ij-healthgeographics.com/content/6/1/1

Received April 2007