STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 30
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography (Neuchâtel, Switzerland, 10-12 April 2000)

Topic (iv): Spatial analysis in a statistical context and disclosure control procedures

SELECTED ISSUES CONCERNING DISCLOSURE AVOIDANCE IN THE CONTEXT OF USER-DEFINED GEOGRAPHY

Submitted by U.S. Bureau of the Census [1]

Contributed paper

ABSTRACT

The Information Age has inundated people with more statistics than ever before. The casual interest or the "need to know" demand for facts about a geographic area specific to the interests of the researcher or data gatherer create a disclosure avoidance challenge of a new dimension for data providers. The potential for finer resolution in data tabulations becomes possible with improvements in data collection tools; advances that result largely from innovations in technology combined with improved data collection procedures. The finest resolution for capturing data, that is, the spatial location of the respondent, has become a reality. Demands for more information at more precise resolutions, including spatial resolution, do not necessarily result in recognition by official statistical offices to the increased capacity made possible by new technical offerings.

The convergence of data, users, and technology has raised interesting disclosure questions and dilemmas for statistical organizations. Users increasingly compare data with geography and they oftentimes want to perform these analyses for different time periods. Changes in geography over time, availability of data for a given area only at specific points in time, and availability of data for the same geography over different time periods are examples of queries that impact data availability, data usability, and disclosure avoidance concerns.

Higher resolutions of data collection and dissemination and the potential for using data at the smallest levels of geography, raise relevant questions concerning the feasibility of meeting these demands while conforming with concerns about confidentiality. These issues are further complicated by the availability of technology such as geographic information systems (GIS) which provide tools to add, subtract, compare, and analyze geography and data in which the use of software functions raise legitimate questions about the impact of these tools on disclosure avoidance of statistical data.

Keywords: geography, user-defined, respondent location, data integration, disclosure

---

[1] Prepared by Timothy Trainor and Kaile Dougherty

I. INTRODUCTION

1. User interest in accessing and using more data about smaller levels of geography has evolved through the decades and is nearing its limit. Generally, the level of geography for which data had been collected was the same level of geography used to tabulate and publish the data. In the United States, data from a decennial census was limited to the county level until the early 20th century when data were made available for smaller geographic areas. The 1990 decennial census was the first to make census block-level data available nationwide. The agency's TIGER System, a national geospatial database that was used to support the data collection and tabulation activities, made this accomplishment possible. Unknowingly, the TIGER System also spawned a multi-billion dollar geographic information system (GIS) business sector. This system was an early resource to the Information Age.

2. This paper attempts to identify some of the issues concerning improved spatial and statistical data availability and user expectations which may affect disclosure avoidance. This paper is not intended to address issues of confidentiality, nor to identify solutions or potential solutions for avoiding disclosure with respect to geographic conditions.

II. IMPROVEMENTS IN COMPUTER TECHNOLOGY

3. Computer technology continues to develop at an unpredictable pace. The time period for upgrades and modernization for organizations continues to shorten. Data storage technology improves while increasing cost efficiency. The computing power of hardware grows, complemented by continual development of rules-based and semi-intelligent software.

4. The availability of data has an impact on the speed of technical development. The growth of GIS technology emerged quickly once spatial data were within the reach of its user community. Statistical organizations worldwide have made significant contributions to the creation of spatial and statistical data. Widening of the user base of data and technology influenced further growth in the functionality and use of GIS.

II.1 Advances in technology for improved positional accuracy

5. Traditionally, census organizations predominantly were interested in relative accuracy, that is, the correct location of one or more respondents in relation to an assignment area and a data reporting area. Users of spatial data have not limited their application of the data to the same intent as planned by the data producers. Two results have occurred. Data users want improved positions to meet their needs for integration with other data sets, such as natural resources data, and statistical agencies are realizing benefits from improved spatial data precision.

II.2 Access to better precision of spatial data sources

6. Until recently, updates to the TIGER data base have been based on annotations by local governments and census field operations to paper copies of TIGER-produced census maps. Relative accuracy oftentimes was the best that could be achieved. In order to maintain a more consistent geospatial framework, the use of remote sensing technologies is now feasible. Satellite imagery and digital orthophotos from aerial photography at various elevations are two imagery sources which are increasingly available. Competition will continue thereby reducing the cost of these once unaffordable spatial data sources.

7. Global Positioning System (GPS) data offer very precise locational information and have notable potential for enumeration and update operations. While the information is limited to latitude and longitude positions, users can enter attribute data as they encode the location. In at least one developing country, GPS was used

successfully to create a spatial data base for the statistical agency in the absence of maps and the high cost of other sources, such as aerial photos, surveys, and imagery.

II.3 Accelerated mobility creates increases in spatial information

8. Improvement in automating field operations with laptop computers and related technologies add to spatial content through use of digital maps, GPS, and map update by field staff. Technology developments are creating new sources of spatial data. For example, telecommunications development provides new sources of locations. Mobile cellular telephones have GPS functionality and Internet protocol (IP) addresses have a latitude/longitude coordinate location.

III. IMPROVEMENTS IN DATA COLLECTION PROCEDURES

9. There appears to be a direct correlation between the availability of finer resolution census data and the size of a growing user community. The potential for finer resolution in data tabulations is made possible by improvements in data collection operations which result in part from advances in technology combined with improved data collection procedures. The United States Census 2000 plans called for building a Master Address File (MAF). Addresses were acquired through a variety of programs, including Postal Service support and local government participation, to prepare the housing unit address list used for questionnaire mailout/delivery for the decennial census. Improvements in the address coverage and the ability to geographically combine areas of similar mailing methods aided the geographic delineation of the types of enumeration areas for Census 2000. Following Census 2000, the MAF will be used to support other censuses and surveys, such as the quinquennial Economic Census and the monthly demographic surveys.

III.1 Respondent Location in Data Collection

10. The spatial location of the respondent is a viable spatial feature for census use. Historically, housing unit locations were captured manually by enumerators and affixed to a map as a means of controlling field operations. This spatial information was not maintained for a variety of reasons, including cost to capture and maintain the data and the limited use beyond a decennial census.

11. In previous U.S. censuses, questionnaire results were aggregated minimally at the block level or enumeration assignment area and ultimately at higher geographic levels for data tabulation. The location of an address, including enumerated housing units in rural areas, serve as the spatial control point for respondent information. While each respondent is referenced to a specific location, one or more respondents can be associated with a single location, for example, multi-unit households or apartments. The respondent location spatial control points will anchor other geospatial data such as features and boundaries that may have varying levels of accuracy. If the respondent location is now the anchor, then the accuracy of locations of other "relative" features will be measured differently than in the past.

III.2 American Community Survey

12. The U.S. Census Bureau has an interest in maintaining and updating respondent locations through the MAF and TIGER to fulfill data collection requirements for a new national survey, the American Community Survey (ACS). In the ACS, data typically collected from the sample long-form questionnaire in a decennial census will be acquired monthly and accumulated annually to disseminate much more timely information about the nation, states, communities, and over five-year averages, neighborhoods. The concept of a neighborhood is not a census term, but it is a geographic area that has local meaning and may be user-defined depending on its use. In relation to census data, a neighborhood oftentimes is defined by one or more census tracts. The census tracts are defined (delineated) by local planning organizations and the Census Bureau.

IV. IMPROVEMENTS IN DATA DISSEMINATION - THE AMERICAN FACTFINDER

13. The American FactFinder (AFF), an Internet-based data dissemination system, is the Census Bureau's primary data delivery vehicle for Census 2000 and beyond. Currently, the system offers data and products from the 1997 Economic Census, the American Community Survey, and the 1990 decennial census. In addition to the tools that empower users to access and display statistical data, the AFF offers interface utilities to access and view geography and related products. The legal and statistical areas for which data are available are viewable through a geography browser. Mapping functions allow users to create their own reference or thematic maps. The AFF also provides access to a variety of predefined statistical and geographic products.

14. The AFF is not designed as a full GIS, but it does demonstrate some of the potential for GIS functions. The spatial data source for the AFF is the Census Bureau's TIGER data base and related geographic support files. Features, boundaries, names, and codes are accessible by users of the system.

15. Traditionally, census data users were offered predefined products, usually in the form of printed reports or computer files on tape or CD-ROM. The AFF opens new opportunities for users to customize products from selections of geography and data that conform with the Census Bureau's non-disclosure policies to protect the confidentiality of respondents. The availability of geography and data add a new set of challenges for consideration in order to determine to what extent these two levels of information will be available independently and in combination.

V. USER-DEFINED GEOGRAPHY

16. Data are available from censuses and surveys at various levels of geography. Decisions about the availability of geographic levels are based on many factors, such as past use, amount of effort to offer the data (staff resources and cost), customer studies and user demand, appropriateness of the data, and conformance with confidentiality guidelines. Whatever decisions are made, the level(s) of geography is selected from a set list of possible choices, such as county (legal area) and census designated place (CDP-statistical area). The Census Bureau works closely with local planning organizations in defining statistical areas, such as census tracts, to make available meaningful data for small and large geographic areas. Nevertheless, users increasingly are interested in data for areas that do not conform to predefined geography. Users increasingly want to define their own areas of interest.

17. As Openshaw notes, "Today user expectations are increasing. They want more relevant but less data, they want accurate derived statistics to be used in resource allocation, and they expect to obtain it for finer scaled geographies that are most appropriate to their own needs rather than what the Census Offices decree they can have."[2]

V.1 Limitations of Feature-Based Delineation

18. Geographic areas are defined in terms of boundary features. For many areas (unincorporated communities and postal places, for instance), boundaries are indistinct or not well known, although place identity is apparent. Some CDPs currently are over-bounded or under-bounded because of reliance on visible features for boundaries. When CDPs are over-bounded, data about the population and their housing units not associated with the place are included, sometimes skewing data for the place. When CDPs are under-bounded, data about the population and their housing units associated with the place are excluded, resulting in a loss of

[2] Openshaw, S., Duke-Williams O., and Rees, P. "Measuring Confidentiality Risks in Census Data", Working Paper 97/8. http://www.geog.leeds.ac.uk/wpapers/97-8.htm

information about the place. Identification of geographic areas in terms of associated residential and commercial address locations could provide a more precise means of delineating places and other statistical entities.[3]

V.2 Data Integration

19. Users continue their interest in merging assorted data sets. Data are integrated either as a view (draping of data sets over one another), or they are combined and reprocessed, thereby possibly taking on a different meaning and form. Such data sets could include entities that are based on subjects very different from demographic or economic data such as watersheds, soil conservation districts, health care regions, irrigation districts, or special taxation districts. Integrating spatial data is fraught with a variety of problems that span technical as well as data issues. With decentralization of spatial data to all sectors and levels of users, interest in integration is complicated by the lack of sophisticated software tools and the uncertainty and incompatibility of information concerning data quality.

V.3 Non-Standard Geographic Entities and User-Defined Geographic Areas

20. The ability to provide data for geographic entities not recognized within the standard census geographic hierarchy is of interest to some data users. One means of meeting user needs would be to permit tabulation of data for user-defined geographic areas. Data associated with respondent locations allows precise delineation of user-defined areas. Potentially, this would allow users to acquire data relevant to their geographic area of interest.

21. The definition of a neighborhood based on interest or need from a collection of respondent locations is an example of local use of census data. From a technological perspective, and from a data perspective, the capability exists to define and redefine areas of interest as aggregations of respondents. One significant assurance that cannot be compromised is adherence to the safeguard of respondent confidentiality.

V.4 Special Tabulations and User Defined Geography

22. As Marsh et al. indicate in their examples from the 1991 census in the United Kingdom:

> "Although the census offices publish an increasingly large number of tables of census data, they can never satisfy all users. The option of asking for non-standard tabulations of census data customized to the user's own requirements is slow and costly and does not allow for non-tabular statistical techniques to be used on the data. The normal interactive process of moving from analysis to interpretation to refined analysis is inhibited. There will be users who eventually need to ask for a special tabulation, but most could benefit from using an Samples of Anonymized Records (SAR) as a test-bed for their request."[4]

23. The above example specifically addresses census data; however, data and geography go hand-in-hand. Therefore, if procedures to ensure disclosure avoidance are based on established census geography, such as the block, block group, and census tract, what implications exist if the data selection is not based on established census geography? For example, if a user is allowed to delineate their own user-defined area, how would the procedures of nondisclosure be applied in these cases? Do similar approaches apply where the source file

---

[3] Cotton, N., Hoy, E., King, G., Ratcliffe, M, and Trainor, T. "County Level Integration (CLI) Issues Paper". Unpublished. U.S. Census Bureau. 1998.

[4] Marsh, C.; Skinner, C.; Arber, S.; Penhale, B.; Openshaw, S.; Hobcraft, J.; Lievesley, D.; and Walford, N. "The Case for Samples of Anonymized Records from the 1991 Census", Journal of the Royal Statistical Society, Series A, Volume 154, Part 2, 1991, pp 305-340.

used to return an answer contains the necessary avoidance techniques or does this type of a query affect the way in which procedures for disclosure avoidance are applied?

VI. GENERAL ISSUES OF GIS AND DISCLOSURE – THE DIFFERENCING PROBLEM

24. Openshaw et al. discuss basics of adding and subtracting geographic area:

> "Differencing involves overlaying data reported for different geographies and arises when simple arithmetic can be used to provide estimates of census data for small areas of a size less than the thresholds. In the context of the 1991 census disclosure control strategy this represents a potential breach of confidentiality. In fact all that can safely be deduced is that it is a breach of an arbitrary minimum size rule, the effectiveness of which as a risk of disclosure ameliorative device is unknown. Furthermore, the data as published were still protected by the secondary mechanism of random value perturbation."

25. The current Census Bureau proposal starts first with the data and when conditions of nondisclosure pass the test, then a check of the geography is applied. What would be the result if the process were reversed, that is, selecting geography, then data, and then check for nondisclosure?

VII. GENERAL ISSUES CONCERNING RESPONDENT LOCATION AND DISCLOSURE

26. The central issue of respondent locations in confidentiality concerns centers on whether a respondent's spatial location is an issue of confidentiality and, if so, what is the impact on the confidentiality concerns raised by statistical data for the location:

- Collection of spatial locations of respondents – is this operation an unnecessary burden or benefit? There were changes to the 1986 Freedom of Information Act of the U.S. in response to new user demands on electronic data providers. Several reform acts and amendments have evolved to address new issues concerning user rights to information and data providers responsibilities. As the amount of improved spatial data grows, complications arise about data availability and restrictions, equal access, acquisition and maintenance costs, and misuse. Other questions include the quality of the data and the limit of how accurate data need to be.

- If respondent locations are treated similar to statistical data with disclosure concerns, the need for clear and evident documentation concerning the manipulations of spatial data locations are critical to warrant against abuse and misuse of the data, such as misassociations of data to a given location.

- Do the data need to be better? If respondent locations offer more precise information for data collection and tabulation, is that information an improvement on past practices, or is the resolution of the spatial data too fine for any meaningful use of the data?

- What impact would spatial data swapping have on distortion of statistical analysis?

- The issue of threat to privacy is of great concern with regard to GIS because of its ability to integrate data with geographic location.[5] According to Devi and Richardson, one of the major concerns with respect to the privacy of government geographic information relates to the accuracy and completeness of the GIS

[5] Devi, T.S. Gayathri and Richardson, Jesse J., "Public Access to Government Geographic Information in the Electronic Age" URISA Proceedings 1999.

database, which according to the authors, is distinct in GIS, since the addition or deletion of information could compromise the accuracy of the information.

- In discussing the need for an SAR in relation to an existing longitudinal survey, one of the arguments made by Marsh et al. for not using the longitudinal survey is the "fine grain geography" needed for research. This argument, along with others, resulted in "more restricted procedures for accessing the data to protect confidentiality."

- It is not clear what impact group quarters and other non-standard enumeration groups (overseas residents, military, homeless, etc.) will have on use of individual address locations for collection, processing, dissemination, and use of census data.

- Having respondent location information increases the use and usefulness of data. For data, more precise functions, such as radius calculations from a single or collection of respondent locations, yield more accurate results. Data at a finer resolution, with accurate location, have even greater potential utility when integrated with other spatial data sets.

- Respondent locations provide finer resolution for housing structure information. Locations can contain multiple structures and one structure can contain multiple resident locations. This creates the potential for interest in "z" spatial values where knowing the level of respondent locations can provide a new type of data for spatial and statistical research and analysis.

- Respondent locations allow for reaggregations of user-defined geography, including those of similar geographic rules, to encourage comparability across national borders and contribute to efforts of global spatial data infrastructures. The positive use of respondent locations is that locations can be aggregated to any other geographic area without having to conform to census administrative or statistical geographic areas.

VIII. CONCLUSION

27. The speed with which finer resolutions of geography are used for census data collection far exceed the application of its potential use, particularly in light of controlled restrictions of data dissemination in adhering to confidentiality requirements. There are many geographic issues that must be researched, studied and evaluated as a direct result of these advances. They emerge from different areas such as operations, policies, technological developments and data use. Meanwhile, development continues and adds more topics to explore.

28. The issues raised in this paper require further thought and study to determine their effect on disclosure concerns. These include, but are not limited to, the degree to which each of these issues influences the direction of disclosure avoidance based on policy and requirements. Beyond that, scenarios describing the characteristics of each issue will be examined. These results then will be compared to current disclosure rules and, where nonconformance exists, alternative options will be studied.

Bibliography

Cotton, N., Hoy, E., King, G., Ratcliffe, M, and Trainor, T. "County Level Integration (CLI) Issues Paper." Unpublished. U.S. Census Bureau. 1998.

Devi, T., Gayathri, S., and Richardson, Jesse J., "Public Access to Government Geographic Information in the Electronic Age." URISA Proceedings. 1999.

Lynch, M., and Foote, K., "Legal Issues Relating to GIS, " The Geographer's Craft Project, Department of Geography, University of Texas at Austin, [WWW Document] URL: http://www.utexas.edu/depts/grg/gcraft/notes/legal/legal.html

Marsh, C.; Skinner, C., Arber, S., Penhale, B, Openshaw, S.; Hobcraft, J., Lievesley, D., and Walford, N. "The Case for Samples of Anonymized Records from the 1991 Census, " Journal of the Royal Statistical Society, Series A, Volume 154, Part 2, pp 305-340. 1991.

Openshaw, S., Duke-Williams O., and Rees, P. "Measuring Confidentiality Risks in Census Data, " Working Paper 97/8. [WWW Document] URL: http://www.geog.leeds.ac.uk/wpapers/97-8.htm

Robinson, G. (1950) Ecological correlation and the behavior of individuals. *Am. Sociol. Rev*., 15, 351-357.

Zayatz, L., Steel, P., and Rowland, S. "Disclosure Limitation for Census 2000. " Working Paper. U.S. Census Bureau. 1999.