Working Paper No. 34
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (iv): Confidentiality issues for small areas

ZIP CODE TABULATION AREA AND CONFIDENTIALITY

Contributed paper

Submitted by the National Center for Health Statistics, Centers for Disease Control and Prevention, United States[1]

1. INTRODUCTION

1. This paper provides a presentation and preliminary examination of the advantages and disadvantages of releasing U.S. statistical data by postal code while preserving confidentiality. The situation is, to say the least, complicated and uneven.

2. FIPS

2. Federal Information Processing Standard (FIPS) was developed by the Bureau of the Census to standardize identification of states, counties and other minor civil divisions (MCD) within 50 states. Two digit FIPS codes for states are complemented by three digit county codes.

3. The term "counties" refers to the "first-order subdivisions" of each State and statistically equivalent entity, regardless of the local terminology (county, parish, borough, etc.). The county equivalent entities include the parishes of Louisiana; the boroughs and census areas of Alaska; the independent cities of Maryland, Missouri, Nevada and Virginia; and the portion of Yellowstone National Park in Montana.[2] The District of Columbia and Guam have no first-order subdivisions, and therefore these same areas serve as first-order subdivisions.[3] Currently there are 3,141 counties in the 50 states. (New York City is made of 5 counties that are boroughs.)

4. The second-order subdivisions that are townships, boroughs, villages, etc., also have their own FIPS codes that follow county codes. All MCDs are subdivided into census tracts that are identified by their unique numbers as block groups and blocks within census tracts.

---

[1] Prepared by Jay H. Kim (jkim@cdc.gov) and Lawrence Cox (lcox@cdc.gov).

[2] National Technical Information Service. Counties and Equivalent Entities of the U.S. Its Possessions and Associated Areas Category: Federal General Data Standard Representations and Codes. U.S. Department of Commerce. August, 1990. p.4.

[3] Ibid., p. 4.

5. In FIPS code states are alphabetically ordered and numbered from Alabama to Wyoming, as counties are within states.

3. ZIP Code

6. Zone Improvement Plan code, better known as ZIP code, has been developed solely for the purpose of mail delivery by the U.S. Postal Service. It constantly changes to meet the need for mail delivery to the newly developing communities. "ZIP code areas" are not polygons, but rather a collection of mail delivery routes sharing a common ZIP code. Therefore, ZIP code areas are in fact not two-dimensional areas but instead collections of one-dimensional line segments. Consequently, there are no official boundary lines between two ZIP code areas.

7. The first digit of ZIP code runs westward from 0 in New England states and New Jersey to 9 in states in the west coast including Alaska and Hawaii. As mail volumes and recipients grew, U.S. Postal Service added 4 extra digits to existing 5 digit codes to aid efficiency in sorting and delivery.

8. FIPS code and ZIP code systems are totally independent of each other as the purpose and development of the two systems are entirely different. One ZIP code area can cross the boundaries of several MCDs in some geographic areas of the U.S. while there are many ZIP code areas in one MCD in most urban areas. As ZIP code follows streets for the convenience of mail delivery, it often cuts through "census" blocks. Therefore, the boundaries of ZIP code hardly coincide with those of MCDs.

9. Recently business communities imposed a different meaning to ZIP code. Since the ZIP code areas define, in a sense, arenas of community life, business wants information by ZIP code area for marketing purposes. Though there have been some efforts for conversion of ZIP code to FIPS code, none of them gives accurate result. One cannot be converted to the other because the boundary lines do not match.

4. ZCTA

10. ZIP code Tabulation Areas, ZCTAs (zik'tahs), developed by the Census Bureau, are generalized area representing five-digit ZIP code areas built by aggregating the Census 2000 blocks whose addresses use a given ZIP code that is same as assigned ZIP codes. The arrangement of whole blocks to belong to one ZCTA made possible at least in theory to convert street address to FIPS code, and then to ZCTA. Some ZCTA codes will be different from actual ZIP codes, since some ZIP codes represent one or very few addresses.

11. The Bureau of the Census published census results by ZCTA beginning 2000 census and made data from Census 2000 including population available in both coding systems; ZCTA and FIPS code. This eliminates the reason to convert the information from one system to the other. However, data from other sources, such as survey results, health data, have yet to be converted from one geocoding system to the other.

12. The Bureau of the Census is preparing ZCTA Outline Map series that will be available in the near future. The Bureau, however, does not and will not provide conversion table showing the relationship between ZCTAs and other types of geographic areas leaving the task to the private industry. With a complete set of address items that includes state, county, town, house number, street, ZIP code, census tract, block group and block on a single record, and a table that will interpret those items into proper ZCTAs, conversion from one geocoding to the other will be possible. Census Bureau does not plan to make and publish this "conversion file" leaving the opportunity to make it based on Bureau's TIGER/Line files[4] to private industry.

[4] A file that will convert common street address to FIPS code.

13. Of 32,038 five-digit ZCTAs, 42 cross state boundaries and 27 percent of ZCTAs cross county lines.[5] Some ZCTAs are physically discontinuous due to ZIP code characteristics that follow street lines. The physical size of ZCTAs varies from 0.0019 square miles to 5,498.4 with average being 83.2.

14. There are 885 three-digit ZCTAs of which 40 cross state lines. The population of 3-digit ZCTAs ranges from 64 to 2.8 million with 199,298 being the median.

5. Confidentiality Issues

15. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) is intended to be a strong protection for patient's medical information while improving access to care. Yet, it's rules define the information cannot be released for an area whose population is less than 20,000 for the protection of privacy.[6]

16. The Bureau of Census has originally determined that a geographical region must have at least 100,000 people to release information without violating personal confidentiality. As simulation studies using census data support this figure, this standard has been used for many years. These studies indicated that increasing the size of a geographic area does not significantly decrease the probability of being identified if sampled, after a certain point, but the number and type of variables play important role.[7]

17. Of two types of data, tabular and micro data that Census Bureau releases, micro data includes data on individuals, households, businesses or other entities that are sensitive to confidentiality. The percent of records that can be uniquely identified in a micro data sample varies as the geographic detail on the file varies. Using Public-Use Micro data Areas files with sizes ranging from 20,000 to 500,000 records, Hawala found that the percentage of uniquely identifiable population increases as the population size decreases.[8] The uniquely identifiable characters are, for example, specific occupation and very high ages. The percentages ranged from 15.7 in a short form of 21,000 to 3.0 in a file of 496,000. With files of long form characteristics, the percentages were 22.7 in a file of 21,000 records and 3.3 percent in a file of 522,000. Hawala concluded that this study supports the validity of the use of the 100,000 population limit for most demographic micro data files.

18. There are two ways of de-identification (removing identification) with differing levels. One is removal of 'direct' identifiers such as name, address and ID numbers. The other is removal of 'indirect' identifiers such as ZIP code and birth date. Even if there is a reasonable basis to believe that protected health information can be used after de-identification to identify the individual, it must be balanced between risk of identification and usefulness of the information.[9] With three-digit ZIP code having less than 20,000 people there are only 18 ZIP code areas by 1990 census and 14 ZCTAs by Census 2000. The census bureau suggested that the initial three digits of ZIP codes might constitute proper level of de-identification.

19. The size of population of ZCTA is dependent on the number of personal information to protect confidentiality of individuals providing information to the Bureau of the Census. The Bureau modifying

---

[5] Census 2000 Summary File 1, Census of Population and Housing. U.S. Department of Commerce. Issued September 2001 on DVD.

[6] Final Modification to the Privacy Rule, Federal Register, August 14, 2002. p. 82710.

[7] Greenberg B, Voshell L. The Geographic Component of Disclosure Risk for Microdata. Bureau of the Census Statistical Research Division Report: Census/SD/RR-90-13. October, 1990.

[8] Hawala S, On the Variation of the Percent of Uniques in a Microdata Sample and the Sample Size. Unpublished Paper. June, 2000.

[9] Final Modifications to the Privacy Rule, Federal Register, August 14, 2002. p. 82708.

its long standing 100,000 population standard suggested the smallest size of ZCTA population to be kept at 20,000 for a small number of demographic variables, 50,000 for ten variables, and 80,000 for 15 variables.[10]

20. NCHS simulation study using national survey data took random samples from populations then compared the samples to the whole population to see how many records were matched uniquely to a unique person on the basis of 9 demographic variables.[11]
About 2.5% was uniquely identifiable from the population of 500,000, about 14% for a population of 100,000 and about 25% for a population of 25,000. This percentage dropped significantly when occupation was eliminated from the list of variables to 0.4%, 3% and 10%, respectively. The probability of these samples to match uniquely is highly dependent on the number of variables and the level of geographical detail.

6. Geographic Units Used by Studies

21. Two studies critically viewed ZCTA as analytic unit. In a recent studies discussing the disclosure risk, Steel and Sperling stated that the more the selection criteria for ZCTA are used, the greater the chance for disclosure. Since ZCTAs having less than 500 households are most numerous, this geographic unit has more risk of disclosure.[12]

22. In a public health disparities geocoding in Massachusetts and Rhode Island Krieger and others discussed the nature of ZIP code in length. They noted that all cancer cases between 1987 and 1993 were geocoded by ZIP code in two New England states. When they traced back the cases by minor civil divisions, 10.4% of the cancer cases in Massachusetts were recorded in code areas that were not accounted in 1990 census. It was 0.7% in Rhode Island.[13] It could be due to the diversity and size of the human society reflected on ZIP code changes, they stated. They noted that the ZIP code level analyses yielded socioeconomic gradients contrary to those observed from data compiled by block group or census tract.

23. Many studies have used ZIP code area as unit of analysis taking the analyzability of this geographic unit as granted without discussions or questions about it. Some researches treated ZIP code area as an equal bases as census tract. Some made up another level of unit of analysis such as socioeconomic groups based on characteristics of ZIP code area.

24. Some even compared ZIP code area against Health Service Area (HSA) and concluded that "smaller areas seem suitable for studying primary care manpower, and larger areas are appropriate for highly specialized tertiary care."[14] Kransner and others stated that counties are too small and HSAs are too large and for many purposes analysis by "ZIP code" area would be conceptually superior and relatively simple. It should be noted that three-digit ZIP code area as they used in their study varies widely in size of

---

[10] Final Modifications to the Privacy Rule, Federal Register. August 14, 2002, p. 82711.

[11] Horm J. A Simulation Study of the Identifiability of Survey Respondents when their Community of Residence is Known. National Center for Health Statistics, 2000.

[12] Steel P., Sperling J. The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code Tabulation Geography. Unpublished monograph. 2001.

[13] Krieger N, Waterman P, Chen JT, Soobader M, Subrmanian SV. ZIP Code Caveat: Bias Due to Spatiotemporal Mismatches Between ZIP Codes and US Census-Defined Geographic Area – The Public Health Disparities Geocoding Project. American Journal of Public Health; July 2002: 92(7). 1100-1102.

[14] Krasner KM, Ramsay DI, Weary PE. Physician Distribution Analysis Based on ZIP Code Areas Applied to Dermatologists, AJPH. 1977 Oct; 67(10):974-977.

physical area and population. This study aimed primarily at the distribution of dermatologists in Iowa and Pennsylvania and the HSAs were redefined since then.[15]

25. Acevedo-Garcia related tuberculosis rates with exposures to poverty, crowded housing, dilapidated housing, contact with immigrants, residential isolation and partial density in 5-digit ZIP code areas in New Jersey.[16]

26. "ZIP code areas" and counties were geographic units of studies in pediatric cancer by White and Aldrich.[17] Setting socioeconomic status based on education and income levels using 1990 census and ascribed by ZIP code, Merkin and others studied the association between race, area SES, and advanced breast cancer stage.[18] And Koch and Denike's comment on GIS approaches to the problem of disease clusters was directed toward "ZIP code / postal code" as a unit of analysis without further questioning its validity.[19]

7. Discussion

27. There are also studies using county as basic geographic unit. Hanson and others compared spatial clustering methods of alcohol mortality by county[20] and Kulldorff and others for the study of breast cancer clusters in the Northeast United States.[21]

28. The analyzability of topics in public health by ZIP code has been proven. As in many literatures, ZIP code area was taken granted and has been used as unit of data gathering and analysis without any discussions on advantage or disadvantage over other geographic units as if it is just another geographic unit for research.

29. ZCTA, the new form of ZIP code, now has logical boundaries that does not cut through blocks thus compatible to other geographic units such as census tract, minor civil divisions and more that can be represented by FIPS code. As long as data confidentiality and usability is preserved by regrouping the 14 ZCTAs that have less than 20,000 population to hold more residents, and as long as the rapidly changing boundary of ZIP code area is recorded, ZCTAs can easily be used for any studies and surveys. Its strength and advantage over traditional geographic units are yet to be tested.

---

[15] Makuc DM, Haglund B, Ingram DD, et al. Health Service Areas for the United States. National Center for Health Statistics. Vital and Health Statistics. 1991;2(112).

[16] Acevedo-Garcia D. ZIP Code-Level Risk Factors for Tuberculosis: Neighborhood Environment and Residential Segregation in New Jersey. AJPH. 2001 May, 91(5): 734-741.

[17] White E, Aldrich TE. Geographic Studies of Pediatric Cancer near Hazardous Waste Sites. Archives of Environmental Health. 1999; 54(6):390-397.

[18] Merkin SS, Stevenson L, Powe N. Geographic Socioeconomic Status, Race, and Advanced-Stage Breast Cancer in New York City. AJPH. 2002 January: 32(1): 64-69.

[19] Koch T, Denike K. GIS Approaches to the Problem of Disease Clusters: A Brief Commentary. Social Science & Medicine. 2001; 52: 1751-1754.

[20] Hanson CE, Wieczorec WF. Alcohol Mortality: A Comparison of Spatial Clustering Methods. Social Science & Medicine. 2002; 5: 791-802.

[21] Kulldorff MJ, Feuer EJ, Miller BA, Freedman LS. Breast Cancer Clusters in the Northeast United States: A Geographic Analysis. American Journal of Epidemiology. 1997; 146(2): 161-170.