United Nations ECE/CES/2010/12

Economic and Social Council

Distr.: General
30 March 2010

Original: English

Economic Commission for Europe

Conference of European Statisticians

Fifty-eighth plenary session
Paris, 8–10 June 2010
Item 6 of the provisional agenda
Spatial statistics

Combining variable spatial data with grids to improve data visualization

Note by the United States Census Bureau

Summary

The note gives a general overview of legal, statistical and administrative geography used by the United States Census Bureau. The note analyses the similarities and differences between the spatial data and statistical grids and discusses the advantages and disadvantages of their use. In some cases both approaches are combined as an integrated solution. Two case studies are presented as examples of different approaches of using spatial statistics: population analysis of Haiti and agricultural statistics of the United States.

I. Introduction

1. Greater availability of spatial data combined with statistics at lower levels of geography has led to an increased interest in the use of spatial statistics in recent years. Statistics related to area usually are polygons represented by administrative units or geometric grids. Decisions about which form to choose are driven by the purpose of the data presentation, the analysis, the characteristics of the statistical and spatial data, and the graphic output that the user will present. There are advantages and disadvantages to using both the statistical grid and administrative polygons. In some cases, both approaches are combined as an integrated solution.

2. The nature of the spatial data elements and characteristics that underlie census geography impacts their use. Spatial statistics concentrate on patterns and clusters of activity. In this paper, a general overview of legal, statistical and administrative geography used by the United States Census Bureau is presented and compared to characteristics of geometric grids. Two case studies are also presented as indicators of different approaches to the use of spatial statistics. Questions are posed to stimulate further discussion and future development.

II. Spatial data

3. Spatial data can be classified into three groups: geostatistical data, point pattern data and lattice data (Cressie, 1993). Geostatistical data are data collected over a continuous spatial domain that is referenced to the earth. Geostatistical data are characterized by "observations associated with a continuous variation over space, typically in the function of distance" (Anselin, 1992). An example of geostatistical data would be soil samples gathered across a region.

4. When interest lies in where events occur, the data referenced are referred to as point pattern data. Point pattern data are concentrated on the location of the individual data points and specifically the spatial pattern created (Cressie, 1993). Instances of spatial point pattern data are spaced irregularly. With this spatial data type, it is not possible to predict the location of occurrence with confidence. Housing unit locations may be one example of point pattern data.

5. Lattice data is collected on a regular or irregular lattice with some defining neighborhood structure (Cressie, 1993). The region where points are collected has a finite number of sites. The extent of the space is contained. Values are assigned for data points and the locations of data points are known. An example of lattice data would be the number of people within each county by state.

III. Spatial statistics

6. Spatial statistics, also referred to as geostatistics, is a form of statistics analyzing spatiotemporal datasets. Spatial statistics are distinguished from other forms of statistics in that they are concerned with the location of the data values. All data have spatial and temporal attributes. The closeness of these data is often an indicator of the similarity of the data. As Waldo Tobler indicated in his First Law of Geography, "everything is related to everything else, but near things are more related than distant things" (1970). Therefore, the data that are closer together, spatially or temporally, are more likely to be similar than those that are farther apart. Spatial statistics use various techniques to study data and their

topological, geometric, and geographic attributes. The goal of spatial statistics is to determine the amount of spatial variations between point data that vary in space and/or time. Spatial statistics can be used to describe the spatial features of a data set and to interpolate from a given set of data to areas where little or no information is available. In spatial statistics every location displays a spatial pattern, whether in the form of the environment, climate, pollution, urbanization or human health.

7. People have always been trying to find patterns in the world around them. Therefore, spatial analysis can be found as far back in history as the beginning of geography, mapping, and surveying. However, the formalized study of spatial statistics did not begin until the later part of the twentieth century. Today, spatial analysis is dependant on computer-based techniques because of the enormous amounts of geographic data, the complex statistical and geographic analysis programs, and advanced spatial modeling. These data-rich environments are a result of emerging technologies. Today, data can be gathered from remotely sensed imagery, intelligent transportation systems, and mobile devices equipped with global positioning systems (GPS) that can report location in near-real time. With the abundance of Geographic Information Systems (GIS), the management of vast stores of data becomes commonplace. As a result, spatial analysis has become a tool that is available to a wide audience. This allows a greater number of people to become analysts and compute and analyze the relationships and patterns between and among data.

8. With the over-abundance of available data, new challenges are being realized in the areas of data storage, representation, retrieval, transmission and most importantly summarization. The study of spatial statistics drives the need for automated methods of summarization, classification, and prediction or modeling. While spatial statistics can be found in many disciplines, the commonality is found in the pervasiveness of data patterns. Data occurring over space and time are often connected by interactions and are evident in spatial patterns, which guide the study of spatial statistics. The overall purpose of spatial statistics is to unearth and study these interactions and the resulting patterns, classify them and then model the interactions and patterns for future data.

9. There are four fundamental questions spatial statistics try to answer:

(a) How are the data distributed?

(b) What is the pattern created by the data?

(c) Where are the clusters?

(d) What are the relationships between sets of data or values?

IV. Statistical grids

10. Statistical grids are rectangular spatial data containers that usually have equal dimensions and have a consistent size for a specific use. Creation of a grid surface begins with an origin of regular spacing in both the *x* and *y* directions (horizontal and vertical). Grids provide relational perspective within the extent of the gridded surface, one grid cell to another. Because of their geometrical configuration, grids are scalable for higher or lower data resolution. A grid is a place holder, a space for storing data instances. The grid space itself does not have definition or meaning.

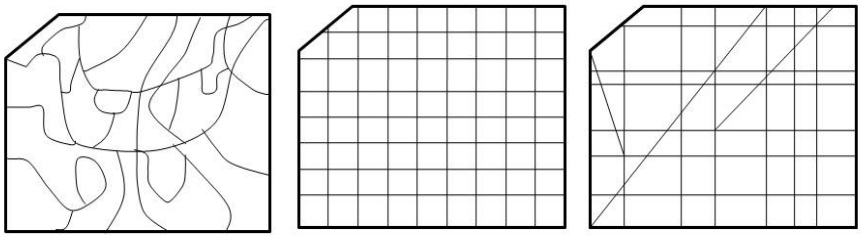
11. Statistical data are applied to the grid cell. Point data are dispersed in a regular pattern, or in an irregular random pattern. One goal is to assure that more than one data point does not occupy the same location. For a real association, classed data are applied to the entire grid cell surface and are related to surrounding cells that either have the same

class or different classes. Similarities result in clusters and patterns while differences show unique or outlier events.

12. Overlaying a grid to an irregular spatial network (such as patterns of mobility) may allow the user to see relationships between the two geographic references. Generally, larger grids compared to smaller scale spatial features (greater area) results in better comparative data relationships. Man-made features like transportation networks (roads and railroads) have irregular and amorphous shapes. Natural features like mountain ranges and rivers share similar properties of irregular orientation.

13. The boundaries of administrative areas are often irregular. In some cases, rectangular administrative areas, for example, those defined by the historic township and range system in the United States, may be coextensive with a statistical grid (Figure 1). The likelihood of being coextensive depends on many factors such as the origin, size, and purpose of the grid.

Figure 1
Comparison of varying geograpic areas



14. Grids offer a moveable and variable sized window into the data. They also provide a mechanism for integrating data from other sources. Data integration currently is a complex challenge with irregular spatial data. While grids offer well-defined properties, they do not relate to the irregular nature of real world phenomenon such as spatial data.

V. Similarities and differences between spatial data and statistical grids

15. Spatial data represent real world phenomena. Some data are natural and are predominantly unpredictable. Man-made data are usually unpredictable; however, some features are created based on patterns which follow plans such as specifications. Planted trees in an orchard, open space between opposite directions of a limited access road and mile marker signs along a road at regular intervals are examples of the latter.

16. Spatial data varies over space and time. According to Griffith and Paelinck, real world data are noisy, dirty and messy (2007). Data uncertainty (or unpredictability) makes data noisy. Ground conditions could show a relatively arid landscape with an unexpected body of water fed by natural springs.

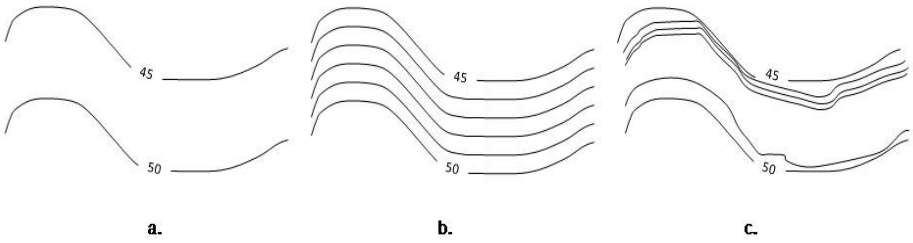
17. Dirty data comprises different inconsistencies that may include incomplete data or outliers and anomalies. Maintaining complete and accurate spatial data in real time as events occur and/or change is not possible for national level data sets. Even with continual sensors, very little spatial data are collected in real time. Incomplete data exist due to factors such as limitations in sources, data quality and resource availability. Outliers are the extremes. For example, temperature data at either end of the temperature scale based on local and unusual microclimate conditions may indicate the extremes. Anomalies are the

exceptions. The term ‘county’ has a common definition for a second order level of government. The term ‘parish’ in Louisiana is equivalent to a county and is an exception to the standard term.

18. Messy data usually have observational dependencies. Single family homes are identifiable at ground level and usually from a satellite image. However, a garage that is converted into a housing unit is not normally discernable without further information or inspection.

19. The white space on a map has spatial characteristics that are not shown. This condition becomes more pervasive as the scale is enlarged and more white space is shown on the map image. Interpolating is difficult due to the nature of spatial data. Predicting spatial data imputation to fill gaps or holes in the map base may be closer to guesswork. Take the common example of interpolating contours. Known contour intervals, the elevation value along a contour line, are determined by relatively precise photogrammetric and geodetic or surveying processes. The elevation values between one index contour line and the next are unknown (Figure 2(a)). One can interpolate the orientation of the line that shows gradual, regular changes in elevation (Figure 2(b)). The fact is that the steepness or relative stable elevation is not known without precise calculation (Figure 2(c)).

Figure 2
Comparison of varying geographic areas



20. In the spatial data world, metadata is the single development that allows improved data use and most importantly assists in efforts to integrate geospatial data. Metadata standards offer a process to document information such as data quality, vintage, source and other needed information at each feature instance. While it could be voluminous, metadata provides for accurate and informed decisions about using and integrating discreet data.

21. Geographic coordinates alone are not sufficient in defining spatial data. Attributes give added meaning and imply purpose of georeferenced data. Examples of attributes such as a classification schema, a geographic name, and a multitude of other descriptors establish a full feature definition and add value to the geographic phenomenon.

22. Managing geospatial data is a complex enterprise. There are numerous opportunities to introduce error into spatial data. As geographic data are referenced to the earth, one of the early encounters with error centers on inaccurate data locations. Other complications can be introduced through the multitude of attributes that define the feature instance. Furthermore, geospatial processes lend themselves to error propagation.

23. Spatial data accuracy is a desirable outcome. Assuring correct geographic relationships in the context of statistical data is paramount. Applying concepts of topology in geospatial processes assures compliance to the requirement for maintaining correct relationships among geospatial point, line and area primitives. Irregularity in the shape or condition of a geographic area does not pose a problem in assuring correct geographic relationships. For example, topological principles assure the relationship between a point

representing a housing unit and its census block, regardless of the shape or geographic extent of the block. The housing unit is within its correct block and is relatively accurate even if the locations of the boundaries of the census block are not positionally accurate.

24. There are various differences between the application of spatial statistics and the use of statistical grids. Visualizing these approaches leads to a better understanding of their characteristics, differences and similarities. Tools have evolved that help in determining the best approach according to the purpose and use of the data.

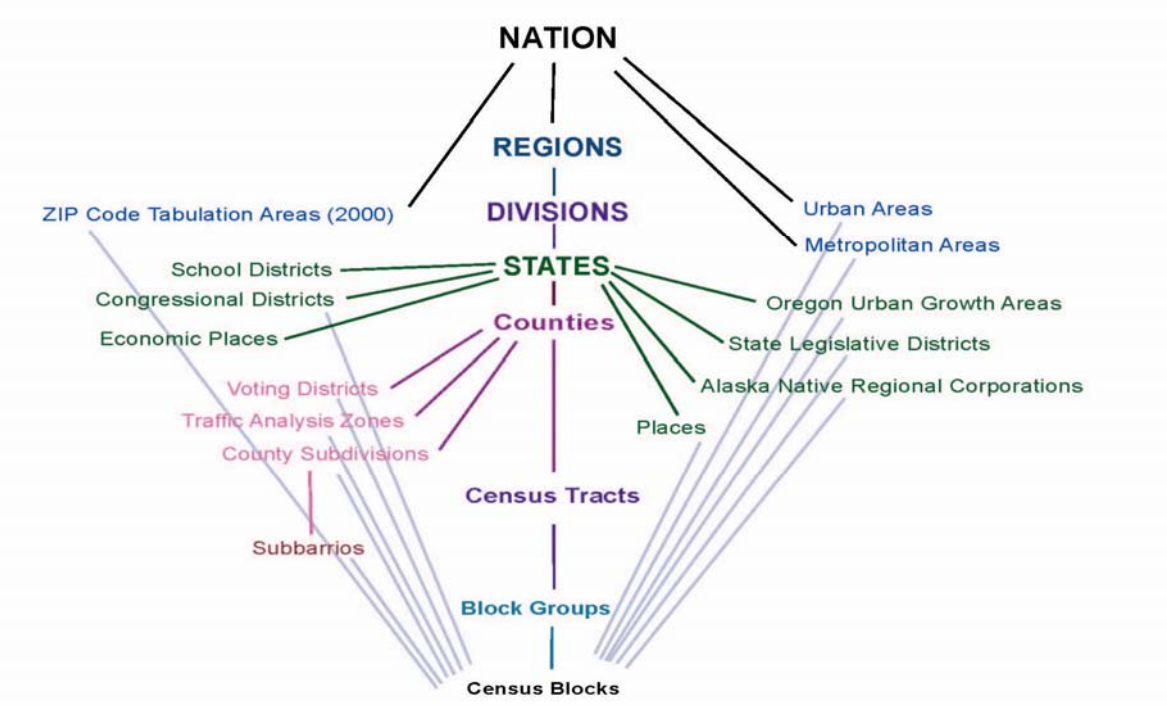
VI. Geographic framework at the United States Census Bureau

25. The United States Census Bureau maintains geography at the census block level. It is the smallest atomic geographic area for which data are collected and tabulated as part of the decennial census. Roads and other visible features form the boundaries of census block polygons. Other features such as the limits of a city or town boundary (that may be a surveyed boundary that is not visible on the ground) are also used as census block boundaries. There are millions of census blocks that cover the United States and its Territories. All area is assigned to a census block.

26. Compared to the consistent design of statistical grids, census blocks are irregularly shaped. The selection of features to serve as block boundaries adhere to specified criteria. While shapes are irregular, particularly outside of inner city street patterns, the size of block polygons are within a general tolerance. Blocks that are too large are difficult to manage for field operations. With all spatial data, there are “dirty” blocks that have anomalous characteristics. For example, a road and/or railroad may follow the shore of a meandering river along a valley. The narrow polygons formed by these continuous features often result in elongated census blocks.

27. Other census geographic areas are comprised of blocks. It is one characteristic that is common to all levels of census geography. Many other levels of geography share some form of nested relationships, that is, a higher level has as part of its definition lower levels of geography that nest within its boundaries. One common example involves the county level. Counties are comprised of census tracts, and then block groups within tracts, and then finally census blocks as indicated in Figure 3 of the standard hierarchy of census geographic entities.

Figure 3
Standard hierarchy of Census geographic entities



28. There are three types of census areas: legal, statistical, and administrative. Legal areas are determined by other levels of government. For example, a boundary of an incorporated city is approved by elected officials. Statistical areas are those defined by the United States Census Bureau often in concert with partners from planning organizations and similar groups in order to offer meaningful areas for data tabulations. An example of administrative areas includes delineation of school districts within a city or town.

29. The level of accuracy for a particular geographic area is dependent on various factors. To delineate legal areas like cities, a boundary and annexation survey is conducted to receive boundary changes as well as new incorporations. The quality of the information is based on the provider's sources and the quality of their process. Statistical area delineations are based on criteria. As interpretation and local interest varies, so the results are variable.

30. The small size and utility of census blocks makes them good candidates to serve as a unit for acquiring, managing and using spatial statistics. Use of grids to substitute for this finite level of geography likely offers less capability and more complexity than that available with blocks. In general, the smaller the level of census geography, the greater the expectation is for greater precision. The use of the census block has raised expectations on the availability and quality of data.

VII. Improved visualization and analysis — case studies

31. In a soon-to-be-published paper, the United States Census Bureau's Population Division undertook a population analysis of Haiti (Azar et al., forthcoming). In this project, the goal was to map population at the scale of 100-meter grid cells, using census data and

satellite image analysis. Census population counts were distributed to grid cells based on the quantity of human-created impervious surface (built-up areas such as buildings and roads) in each cell. The gridded map was then enhanced with online mapping tools (United States Census Bureau, 2010), allowing for customizable data views and creating a common ground for analysis across the country.

32. In a previous publication, the 1992 Agricultural Atlas of the United States, a combination of grids and administrative geography were used to more accurately display the locations of data instances in a series of dot distribution maps (USDA, 2010). The publication contained approximately 190 dot distribution maps, as well as over 120 choropleth maps. Dot maps of the United States were produced using county level data.

33. The agricultural topics varied, and it was clear that within the boundaries of many counties, a random dot placement would lead to agricultural activity in areas where it was impossible or highly unlikely, for example, rangeland in urban areas or crops in tundra. Sources included separate generalized files of county boundaries, shorelines and urban areas. A raster grid file for the land use/land cover data set was used. At smaller scales, these land use/land cover data themes were created as coverage polygon files and were integrated with the administrative county boundaries.

34. Agricultural data were classed into one of five general groupings: general farm; crop; rangeland and grazing; livestock; and orchard. For each land use category (e.g. forest and woodland grazed) a value was assigned for the likelihood of occurrence for each agricultural group (high for rangeland and grazing). The resulting ranking was assigned a percentage of the number of dots placed within that portion of the county with that specified land use. Each dot for a given map represented a specified number of data instances (dot value).

35. The effect of joining two disparate data types resulted in a more accurate representation of the mapped data. Analysts were provided with data patterns that reflected the impact of land use on various agricultural activities. Similar efforts where grid data combined with administrative geography can yield more favorable results.

VIII. Conclusion

36. The relatively recent convergence of data, technology, software tools and computing power opened new opportunities for analysts to study statistical data as it relates to location. While basic functions emerged in earlier geographic information systems, interest has grown in evaluating patterns and clusters of occurrence as well as predicting future activity.

37. In situations where lower level geographic areas like census blocks are available, the case is made to consider this data type as an alternative to statistical grid as the nature of spatial data influences statistical data instances. Opportunities exist in merging administrative units with statistical grids that could result in new uses of integrated data for analysts. Efforts during the last three decades have focused on building spatial data sets and using features in conventional ways. Visualizing the effect of spatial statistics in various forms offers different views for analysts to pursue and opens the door for expanded analysis.

IX. References

Anselin, L. (1992), Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences, Technical Report 92-10, National Center for Geographic Information and Analysis

Azar D., et al. (Forthcoming), Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti, *International Journal of Remote Sensing*

Cressie, N.A.C., (1993), *Statistics for Spatial Data*, Wiley: New York

Griffith, D.A. and Paelinck, J.H.P. (2007), An equation by any other name is still the same:

Spatial econometrics and spatial statistics, *Annals of Regional Science*, Vol. 41

Tobler W. (1970) A computer movie simulating urban growth in the Detroit region, Economic Geography, 46(2)

United States Census Bureau. 2010. Haiti Earthquake: United States Census Bureau Population Data. Online: https://www.geoint-online.net/community/haitiearthquake/default.aspx.

USDA. 2010. Agricultural Atlas of the United States. Online: http://www.agcensus.usda.gov/Publications/1992/Agricultural_Atlas/textfile/introduc.asc