UNIT 6 - SAMPLING THE WORLD

UNIT 6 - SAMPLING THE WORLD

Compiled with assistance from Charles Parson, Bemidji State University and Timothy Nyerges, University of Washington

---

For Information that Supplements the Contents of this Unit:

[Links to the following resources have been omitted.]

- Errors in Maps (Chrisman/U of Washington) -- US data quality standards.
- Error, Accuracy and Precision (Geographer's Craft) -- (A few graphics); types of errors; sources of inaccuracy and imprecision; problems of propagation and cascading; beware of false precision and false accuracy; dangers of undocumented data; principles of managing error.
- Measurement Basics (Chrisman/U of Washington) -- Graphics and description for: levels of measurement (nominal, ordinal, interval, ratio); other forms of measurement (absolute, counts, cyclical, multi-dimensional).
- Managing Error (Foote and Huebner/Geographer's Craft) -- Managing problems of error, accuracy and precision; setting standards for procedures and products (establishing criteria, training and testing); documenting, measuring and testing; calibrating a data set to ascertain how error influences solutions; etc.

This unit begins the section on data acquisition by looking at how the infinite complexity of the real world can be discretized and sampled.

UNIT 6 - SAMPLING THE WORLD

Compiled with assistance from Charles Parson, Bemidji State

University and Timothy Nyerges, University of Washington

A. INTRODUCTION

- the world is infinitely complex
- the contents of a spatial database represent a particular view of the world
- the user sees the real world through the medium of the database
    - the measurements and samples contained in the database must present as complete and accurate a view of the world as possible
    - the contents of the database must be relevant in terms of:
        - themes and characteristics captured
        - the time period covered
        - the study area

- this unit looks at techniques for sampling the world, and associated issues of accuracy, standards

B. REPRESENTING REALITY

- a database consists of digital representations of discrete objects
- the features shown on a map, e.g. lakes, benchmarks, contours can be thought of as discrete objects
    - thus the contents of a map can be captured in a database by turning map features into database objects

- many of the features shown on a map are fictitious and do not exist in the real world
    - contours do not really exist, but houses and lakes are real objects

- the contents of a spatial database include:
    - digital versions of real objects, e.g. houses
    - digital versions of artificial map features, e.g. contours
    - artificial objects created for the purposes of the database, e.g. pixels

Continuous variation

- some characteristics exist everywhere and vary continuously over the earth's surface
  - e.g. elevation, atmospheric temperature and pressure, natural vegetation or soil type

- we can represent such variation in several ways:
  - by taking measurements at sample points, e.g. weather stations
  - by taking transects
  - by dividing the area into patches or zones, and assuming the variable is constant within each zone, e.g. soil mapping
  - by drawing contours, e.g. topographic mapping

- each of these methods creates discrete objects
  - the objects in each case are points, lines or areas

- a raster can be thought of as:
  - a special case of a point sample where the points are regularly spaced
  - a special case of zones where the zones are all the same size

- each method is approximate, capturing only part of the real variation
  - a point sample misses variation between points
  - transects miss variation not on transects
  - zones pretend that variation is sudden at boundaries, and that there is no variation within zones
  - contours miss variation not located on contours

- several methods can be used to try to improve the success of each method
- e.g. for zones:
  - map the boundaries as fuzzy instead of sharp lines
  - describe the zones as mixtures instead of as single classes, e.g. 70% soil type A, 30% soil type B

C. SPATIAL DATA

- phenomena in the real world can be observed in three modes: spatial, temporal and thematic
  - the spatial mode deals with variation from place to place
  - the temporal mode deals with variation from time to time (one slice to another)
  - the thematic mode deals with variation from one characteristic to another (one layer to another)

- all measurable or describable properties of the world can be considered to fall into one of these modes - place, time and theme
- an exhaustive description of all three modes is not possible
- when observing real-world phenomena we usually hold one mode fixed, vary one in a controlled" manner, and measure"the third (Sinton, 1978)
  - e.g. using a census of population we could fix a time such as 1990, control for location using census tracts and measure a theme such as the percentage of persons owning automobiles

- holding geography fixed and varying time gives longitudinal data
- holding time fixed and varying geography gives cross- sectional data
- the modes of information stored in a database influence the types of problem solving that can be accomplished

Location

- the spatial mode of information is generally called location

Attributes

- attributes capture the thematic mode by defining different characteristics of objects
- a table showing the attributes of objects is called an attribute table
  - each object corresponds to a row of the table
  - each characteristic or theme corresponds to a column of the table

thus the table shows the thematic and some of the spatial modes

Time

- the temporal mode can be captured in several ways
  - by specifying the interval of time over which an object exists
  - by capturing information at certain points in time
  - by specifying the rates of movement of objects

- depending on how the temporal mode is captured, it may be included in a single attribute table, or be represented by series of attribute tables on the same objects through time

D. SAMPLING REALITY

Scales of measurement

- numerical values may be defined with respect to nominal, ordinal, interval, or ratio scales of measurement
- it is important to recognize the scales of measurement used in GIS data as this determines the kinds of mathematical operations that can be performed on the data
- the different scales can be demonstrated using an example of a marathon race:

1. Nominal

- on a nominal scale, numbers merely establish identity
  - e.g. a phone number signifies only the unique identity of the phone

- in the race, the numbers issued to racers which are used to identify individuals are on a nominal scale
  - these identity numbers do not indicate any order or relative value in terms of the race outcome

2. Ordinal

- on an ordinal scale, numbers establish order only
  - phone number 9618224 is not more of anything than 9618049, so phone numbers are not ordinal

- in the race, the finishing places of each racer, i.e. 1st place, 2nd place, 3rd place, are measured on an ordinal scale
  - however, we do not know how much time difference there is between each racer

3. Interval

- on interval scales, the difference (interval) between numbers is meaningful, but the numbering scale does not start at 0
  - subtraction makes sense but division does not
  - e.g. it makes sense to say that 200C is 10 degrees warmer than 100C, so Celsius temperature is an interval scale, but 200C is not twice as warm as 100C
  - e.g. it makes no sense to say that the phone number 9680244 is 62195 more than 9618049, so phone numbers are not measurements on an interval scale

- in the race, the time of the day that each racer finished is measured on an interval scale
  - if the racers finished at 9:10 GMT, 9:20 GMT and 9:25 GMT, then racer one finished 10 minutes before racer 2 and the difference between racers 1 and 2 is twice that of the difference between racers 2 and 3
  - however, the racer finishing at 9:10 GMT did not finish twice as fast as the racer finishing at 18:20 GMT

4. Ratio

- on a ratio scale, measurement has an absolute zero and the difference between numbers is significant
  - division makes sense

e.g. it makes sense to say that a 50 kg person weighs half as much as a 100 kg person, so weight in kg is on a ratio scale
    - the zero point of weight is absolute but the zero point of the Celsius scale is not

- in our race, the first place finisher finished in a time of 2:30, the second in 2:40 and the 450th place finisher took 5 hours
    - the 450th finisher took twice as long as the first place finisher (5/2.5 = 2)

- note these distinctions, though important, are not always clearly defined
    - is elevation interval or ratio? if the local base level is 750 feet, is a mountain at 2000 feet twice as high as one at 1000 feet when viewed from the valley?

- many types of geographical data used in GIS applications are nominal or ordinal
    - values establish the order of classes, or their distinct identity, but rarely intervals or ratios

- thus you cannot:
    - multiply soil type 2 by soil type 3 and get soil type 6
    - divide urban area by the rank of a city to get a meaningful number
    - subtract suitability class 1 from suitability class 4 to get 3 of anything

- however, you can:
    - divide population by area (both ratio scales) and get population density
    - subtract elevation at point a from elevation at point b and get difference of elevation

Multiple representations

- a data model is essential to represent geographical data in a digital database
- there are many different data models
- the same phenomena may be represented in different ways, at different scales and with different levels of accuracy
- thus there may be multiple representations of the same geographical phenomena
- it is difficult to convert from one representation to another
    - e.g. from a small scale (1:250,000) to a large scale (1:10,000)

- thus it is common to find databases with multiple representations of the same phenomenon
    - this is wasteful, but techniques to avoid it are poorly developed

E. DATA SOURCES

Primary data collection

- some of the data in a spatial database may have been measured directly
    - e.g. by field sampling or remote sensing

- the density of sampling determines the resolution of the data
    - e.g. samples taken every hour will capture hour-to- hour variation, but miss shorter-term variation
    - e.g. samples taken every 1 km will miss any variation at resolutions less than 1 km

- a sample is designed to capture the variation present in a larger universe
    - e.g. a sample of places should capture the variation present at all possible places
    - e.g. a sample of times will be designed to capture variation at all possible times

- there are several standard approaches to sampling:
    - in a random sample, every place or time is equally likely to be chosen
    - systematic samples are chosen according to a rule, e.g. every 1 km, but the rule is expected to create

      no bias in the results of analysis, i.e. the results would have been similar if a truly random sample had been taken

    - in a stratified sample, the researcher knows for some reason that the universe contains significantly different sub-populations, and samples within each sub-population in order to achieve adequate representation of each
        - e.g. we may know that the topography is more rugged in one part of the area, and sample more densely there to ensure adequate representation
        - if a representative sample of the entire universe is required, then the

subsamples in each subpopulation will have to be weighted appropriately

Secondary data sources

- some data may have been obtained from existing maps, tables, or other databases
  - such sources are termed secondary

- to be useful, it is important to obtain information in addition to the data themselves:
  - information on the procedures used to collect and compile the data
  - information on coding schemes, accuracy of instruments

- unfortunately such information is often not available
  - a user of a spatial database may not know how the data were captured and processed prior to input
  - this often leads to misinterpretation, false expectations about accuracy

F. STANDARDS

- standards may be set to assure uniformity
  - within a single data set
  - across data sets
  - e.g. uniform information about timber types throughout the database allows better fire fighting methods to be used, or better control of insect infestations

- data capture should be undertaken in standardized ways that will assure the widest possible use of the information

Sharing data

- it is not uncommon for as many as three agencies to create databases with, ostensibly,

the same information
  - e.g. a planning agency may map landuse, including a forested class
  - e.g. the state department of forestry also maps forests
  - e.g. the wildlife division of the department of conservation maps habitat, which includes fields and forest

- each may digitize their forest class onto different GIS systems, using different protocols, and with different definitions for the classes of forest cover
- this is a waste of time and money
- sharing information gives it added value
- sharing basic formats with other information providers, such as a department of transportation, might make marketing the database more profitable

Agency standards

- state and national agencies have set standards for certain environmental data
  - the Soil Conservation Service (SCS) has adopted the "seventh approximation"as the national taxonomy
  - the US Geological Survey has set standards for landuse, transportation, and hydrography that are used as guidelines in many states
  - forest inventories are not standardized; agencies may use different systems while managing a contiguous region of forest land

- Unit 69 covers standards for GIS in greater depth

G. ERRORS AND ACCURACY

- note: Units 45 and 46 discuss this topic in detail
- there is a nearly universal tendency to lose sight of errors once the data are in digital form
- errors:
  - are implanted in databases because of errors in the original sources (source errors)
  - are added during data capture and storage (processing errors)
  - occur when data are extracted from the computer
  - arise when the various layers of data are combined in an analytical exercise

Original Sin - errors in sources

- are extremely common in non-mapped source data, such as locations of wells, or lot descriptions
- can be caused by doing inventory work from aerial photography and misinterpreting images
- often occur because base maps are relied on too heavily
    - a recent attempt in Minnesota to overlay Department of Transportation bridge locations on USGS transportation data resulted in bridges lying neither beneath roads, nor over water, and roads lying apparently under rivers
    - until they were compared in this way, it was assumed that each data set was locationally acceptable
    - the ability of GIS to overlay may expose previously unsuspected errors

Boundaries

- boundaries of soil types are actually transition zones, but are mapped by lines less than 0.5 mm wide
- lakes fluctuate widely in area, yet have permanently recorded shorelines

Classification errors

- are common when tabular data are rendered in map form
- simple typing errors may be invisible until presented graphically
    - floodplain soils may appear on hilltops
    - pastureland may appear to be misinterpreted marsh

- more complex classification errors may be due to the sampling strategies that produced the original data
- timber appraisal is commonly done using a few, randomly selected points to describe large stands
    - information may exist that documents the error of the sampling technique
    - however, such information is seldom included in the GIS database

Data capture errors

- manual data input induces another set of errors
- eye-hand coordination varies from operator to operator and from time to time
    - data input is a tedious task - it is difficult to maintain quality over long periods of time

Accuracy standards

- many agencies have established accuracy standards for geographical data
    - these are more often concerned with accuracy of locations of objects than with accuracy of attributes

- location accuracy standards are commonly decided from the scale of source materials
    - for natural resource data 1:24,000 scale accuracy is a common target
    - at this scale, 0.5 mm line width = 12 m on the ground

- USGS topographic information is currently available in digital form at 1:100,000
    - 0.5 mm line width = 50 m on the ground

- higher accuracy requires better source materials
    - is the added cost justified by the objectives of the study?

- accuracy standards should be determined by considering both the value of information and the cost of collection

REFERENCES

Berry, B.J.L and A.M. Baker, 1968. "Geographic sampling. In B.J.L. Berry and D.F. Marble, editors, Spatial Analysis. Prentice Hall, Englewood Cliffs NJ, 91-100. A classic paper on sampling geographical distributions.

Hopkins, Lewis D., 1977, "Methods for generating land suitability maps: A comparative evaluation,"AIP Journal October 1977:386-400. An excellent discussion of the different measurement scales is given in an appendix.

Sinton, D., 1978. "The inherent structure of information as a constraint to analysis: mapped thematic data as a case study, Harvard Papers on Geographic Information Systems, Vol. 7, G. Dutton (ed.), Addison Wesley, Reading, MA. A classic paper on the relationships between the database and reality.

Standard sampling theory is covered in many texts on scientific measurement.

EXAM AND DISCUSSION QUESTIONS

1. Take an example map showing the observed occurrences of some rare event, and discuss the factors influencing the sampling process. Good examples are maps of tornado sightings, herbarium records of rare plants.

2. Using a topographic map, discuss the ways in which the contents and design of the map influence the user's view of the real world.

3. Review the accuracy information available for several different scales and types of maps, and spatial databases if available.

4. The Global Positioning System (GPS) will soon be capable of providing latitude and longitude positions to the nearest meter using portable receivers weighing on the order of 1 kg, in no more than one minute. This is significantly more accurate than the best base mapping generally available in the US (1:24,000). Discuss what effect this system might have on map makers and map users.

Last Updated: August 30, 1997.