The Nature of Geographic Data

Types of spatial data

- Continuous spatial data: geostatistics
  - Samples may be taken at intervals, but the spatial process is continuous
  - e.g. soil quality
- Discrete data
  - Irregular: zonal data, regions, states, districts, postcodes, zipcodes
  - Regular lattice data: constructed grid, 'raster' representation

Types of spatial data

- Point patterns
  - Events occurring in a continuous (or finely discretised) space
- Polygons, and lines (GIS)

- **Q**: Where would a continuous space–based model be appropriate in social sciences?

Spatial Autocorrelation

- First law of geography: "everything is related to everything else, but near things are more related than distant things" – Waldo Tobler

- Many geographers would say "I don't understand spatial autocorrelation" Actually, they don't understand the mechanics, they do understand the concept.

Spatial autocorrelation

- Lattice or zone data
- Variable (x) recorded at places s
- Is the data random or are there similarities between neighbours?
- Does a high value of x tend to be associated with a high value of x in neighbouring places (and low values with low)?

Spatial Autocorrelation

- Spatial Autocorrelation – correlation of a variable with itself through space.
  - If there is any systematic pattern in the spatial distribution of a variable, it is said to be spatially autocorrelated
  - If nearby or neighboring areas are more alike, this is *positive spatial autocorrelation*
  - *Negative autocorrelation* describes patterns in which neighboring areas are unlike
  - Random patterns exhibit *no spatial autocorrelation*

Why spatial autocorrelation is important

- Most statistics are based on the assumption that the values of observations in each sample are independent of one another
- Positive spatial autocorrelation may violate this, if the samples were taken from nearby areas
- Goals of spatial autocorrelation
  - Measure the strength of spatial autocorrelation in a map
  - test the assumption of independence or randomness

Spatial Autocorrelation

- Spatial Autocorrelation is, conceptually as well as empirically, the two-dimensional equivalent of redundancy
- It measures the extent to which the occurrence of an event in an areal unit constrains, or makes more probable, the occurrence of an event in a neighboring areal unit.

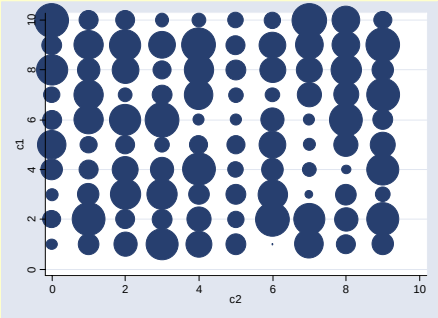
Spatial Autocorrelation

- Non-spatial independence suggests many statistical tools and inferences are inappropriate.
  - Correlation coefficients or ordinary least squares regressions (OLS) to predict a consequence assumes that the observations have been selected randomly.
  - If the observations, however, are spatially clustered in some way, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise.
  - They are biased because the areas with higher concentration of events will have a greater impact on the model estimate and they will overestimate precision because, since events tend to be concentrated, there are actually fewer number of independent observations than are being assumed.

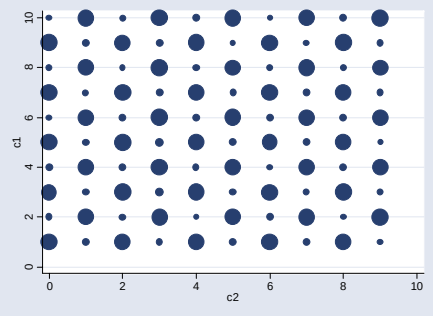
Indices of Spatial Autocorrelation

- Moran's I
- Geary's C
- Ripley's K
- Join Count Analysis

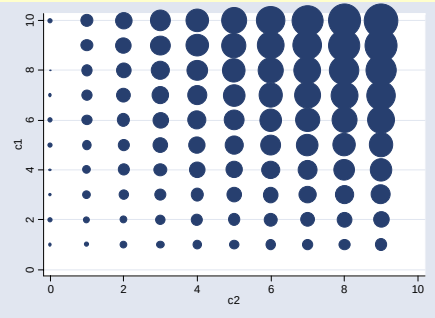
Random - no spatial autocorrelation



Overly dispersed - negatively autocorrelated

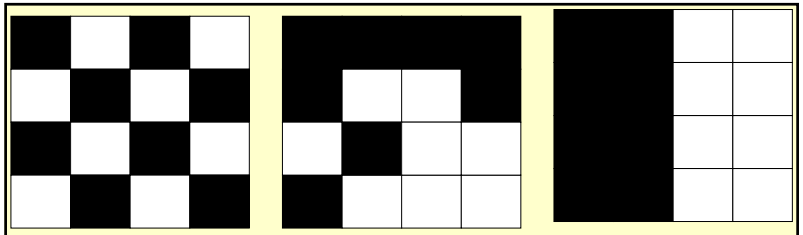


Positive spatial autocorrelation



Why does spatial auto correlation occur?

- Reaction functions?
- Spillovers, externalities?
- Unobserved similarities between places?
- Diffusion (disease spread)?
- Common activity in neighbouring areas (crime)?
- Common policy across neighbouring areas?



Join Counts

Map

	BB	WW	BW	Tot.
A	0	0	24	24
B	5	7	12	24
C	10	10	4	24

Computing the Joins

- When using a computer to compute joins, we will have twice as many joins (ex. WB, BW). Just divide by two…

$$L = \sum_{i=1}^{n}$$

- If we have n_b Black cells and $n_w = n - n_b$ white cells the probability p of a black cell is

$p_b = n_b / n$

$p_w = n_w / n$

Computing the probabilities

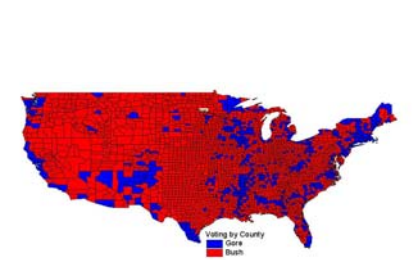
- Start with first cell. The probability it is black is p_b and the probability of white is p_w
- The probability of BB in two adjacent cells is

  $p_b * p_b$ or p_b^2
- Probability of BW is

  $p_b * p_w + p_b * p_w$ or $2\ p_b\ p_w$

Example from Election 2000

- A spatial analysis of Election 2000
  - Did the "blue and red map" really say something significant about the locations where people voted by County?
  - *ArcView example*



Probability of Bush winning a County

$\mathbf{p} := .78$

Probability of Gore winning a County

$\mathbf{q} := .22$

Total County Joins

$\mathbf{L} := 9068$

$\mathbf{K} := 47065$

Bush/Bush Joins

$\mathbf{BB} := 6253$

Gore/Gore Joins

$\mathbf{WW} := 879$

Bush/Gore Joins

$\mathbf{BW} := 1936$

$\mathbf{BB} + \mathbf{WW} + \mathbf{BW} = 9068$

Expected Bush/Bush Joins

$\mu\mathbf{BB} := \mathbf{p}^2 \cdot \mathbf{L}$

$\mu\mathbf{BB} = 5516.971$

Expected Gore/Gore Joins

$\mu\mathbf{WW} := \mathbf{q}^2 \cdot \mathbf{L}$

$\mu\mathbf{WW} = 438.891$

Expected Bush/Gore Joins

$\mu\mathbf{BW} := 2 \cdot \mathbf{p} \cdot \mathbf{q} \cdot \mathbf{L}$

$\mu\mathbf{BW} = 3112.138$

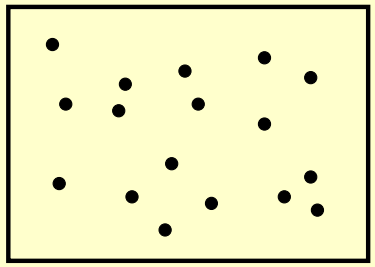
$\mu\mathbf{BB} + \mu\mathbf{WW} + \mu\mathbf{BW} = 9068$

Sampling

- The sampling density determines the resolution of the data
- Samples taken at 1 km intervals will miss variation smaller than 1 km
- Standard approaches to sampling:
  - Random
  - Systematic
  - Stratified

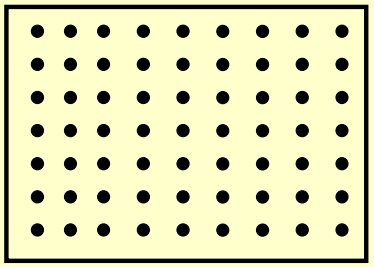
Random samples

- Every location is equally likely to be chosen



Systematic samples

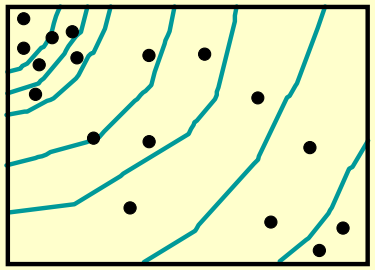
- Sample points are spaced at regular intervals



Stratified samples

- Requires knowledge about distinct, spatially defined sub-populations (spatial subsets such as ecological zones)
- More sample points are chosen in areas where higher variability is expected

Stratified samples

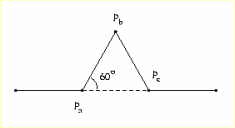


Selfsimilarity and fractals

The Koch Snowflake

$$Length = 1$$

First iteration

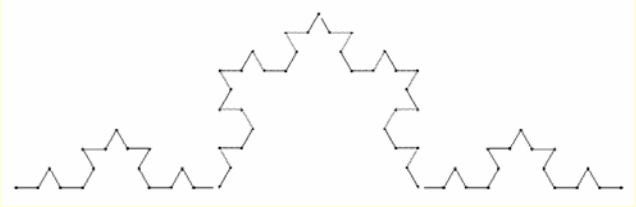


$$Length = \frac{4}{3}$$

After
2 iterations

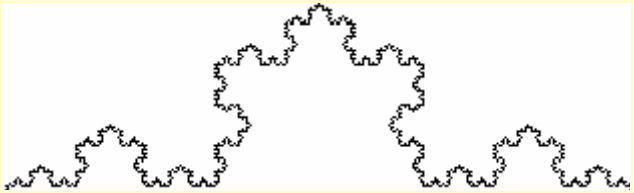
$$Length = \left(\frac{4}{3}\right)^2$$

After 3 iterations



$$Length = \left(\frac{4}{3}\right)^3$$

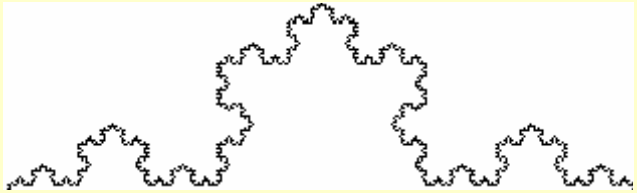
After *n* iterations



$$Length = \left(\frac{4}{3}\right)^n$$

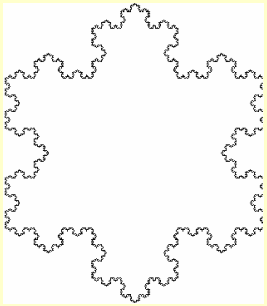
After ∞ iterations

(work with me here, people)



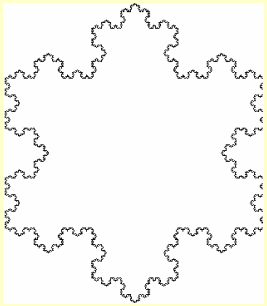
$$Length = \left(\frac{4}{3}\right)^{\infty} = \infty$$

The **Koch snowflake** is six of these put together to form . . .



. . . well, a snowflake.

Notice that the perimeter of the Koch snowflake is infinite . . .



. . . but that the area it bounds is finite (indeed, it is contained in the white square).

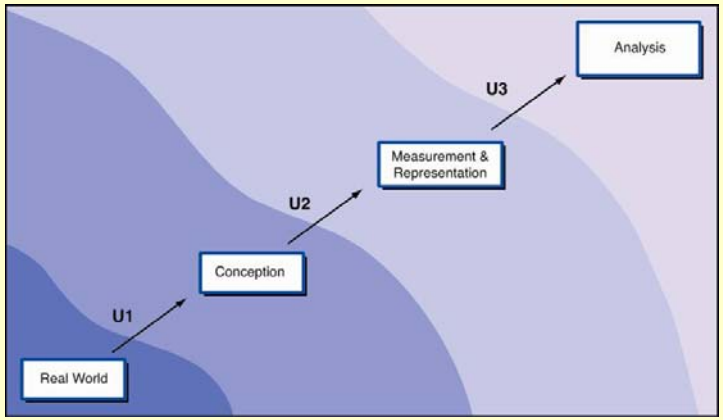
Uncertainty

Introduction

- Imperfect or *uncertain* reconciliation
  - [science, practice]
  - [concepts, application]
  - [analytical capability, social context]
- It is impossible to make a perfect representation of the world, so uncertainty about it is inevitable

Sources of Uncertainty

- Measurement *error*: different observers, measuring instruments
- Specification *error*: omitted variables
- *Ambiguity*, *vagueness* and the *quality* of a GIS representation
- A catch-all for 'incomplete' representations or a 'quality' measure



U1: Conception

- Spatial uncertainty
  - *Natural* geographic units?
  - Bivariate/multivariate extensions?
  - Discrete objects
- Vagueness
  - Statistical, cartographic, cognitive
- Ambiguity
  - Values, language

Scale & Geographic Individuals

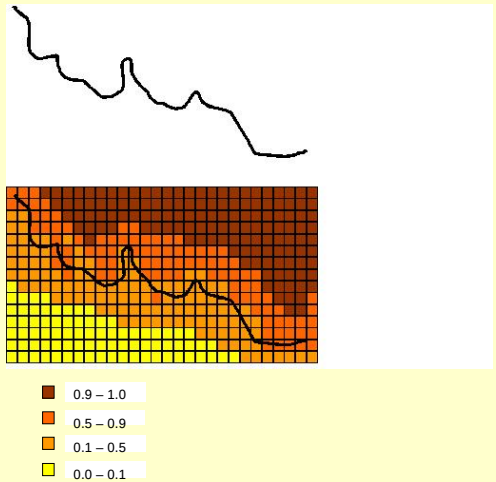
- Regions
  - Uniformity
  - Function
- Relationships typically grow stronger when based on larger geographic units

Scale and Spatial Autocorrelation

No. of geographic areas	Correlation
48	.2189
24	.2963
12	.5757
6	.7649
3	.9902

U2: Measurement/representation

- Representational models filter reality differently
  - Vector
  - Raster



Statistical measures of uncertainty: nominal case

- How to measure the accuracy of nominal attributes?
  - e.g., a vegetation cover map
- The confusion matrix
  - compares recorded classes (the *observations*) with classes obtained by some more accurate process, or from a more accurate source (the *reference*)

Example of a misclassification or confusion matrix. A grand total of 304 parcels have been checked. The rows of the table correspond to the land use class of each parcel as recorded in the database, and the columns to the class as recorded in the field. The numbers appearing on the principal diagonal of the table (from top left to bottom right) reflect correct classification.

	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	0	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

Confusion Matrix Statistics

- Percent correctly classified
  - total of diagonal entries divided by the grand total, times 100
  - 209/304*100 = 68.8%
  - but chance would give a score of better than 0
- Kappa statistic
  - normalized to range from 0 (chance) to 100
  - evaluates to 58.3%

Sampling for the Confusion Matrix

- Examining every parcel may not be practical
- Rarer classes should be sampled more often in order to assess accuracy reliably
  - sampling is often stratified by class

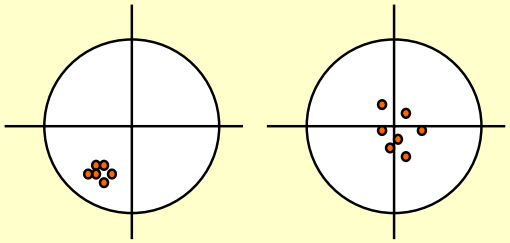
Per-Polygon and Per-Pixel Assessment

- Error can occur in both attributes of polygons, and positions of boundaries
  - better to conceive of the map as a field, and to sample points
  - this reflects how the data are likely to be used, to query class at points

An example of a vegetation cover map. Two strategies for accuracy assessment are available: to check by area (polygon), or to check by point. In the former case a strategy would be devised for field checking each area, to determine the area's correct class. In the latter, points would be sampled across the state and the correct class determined at each point.

Interval/Ratio Case

- Errors distort measurements by small amounts
- Accuracy refers to the amount of distortion from the true value
- Precision
  - refers to the variation among repeated measurements
  - and also to the amount of detail in the reporting of a measurement



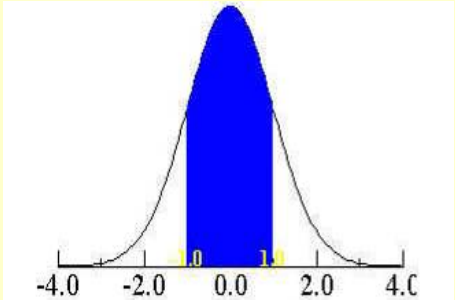
The term *precision* is often used to refer to the repeatability of measurements. In both diagrams six measurements have been taken of the same position, represented by the center of the circle. On the left, successive measurements have similar values (they are *precise*), but show a bias away from the correct value (they are *inaccurate*). On the right, precision is lower but accuracy is higher.

Reporting Measurements

- The amount of detail in a reported measurement (e.g., output from a GIS) should reflect its accuracy
  - "14.4m" implies an accuracy of 0.1m
  - "14m" implies an accuracy of 1m
- Excess precision should be removed by rounding

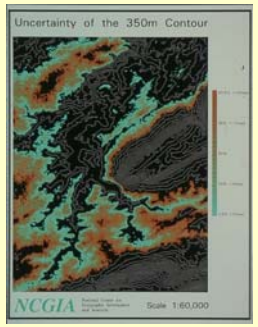
Measuring Accuracy

- Root Mean Square Error is the square root of the average squared error
  - the primary measure of accuracy in map accuracy standards and GIS databases
  - e.g., elevations in a digital elevation model might have an RMSE of 2m
  - the abundances of errors of different magnitudes often closely follow a Gaussian or normal distribution



The Gaussian or Normal distribution. The height of the curve at any value of *x* gives the relative abundance of observations with that value of *x*. The area under the curve between any two values of *x* gives the probability that observations will fall in that range. The range between –1 standard deviation and +1 standard deviation is in blue. It encloses 68% of the area under the curve, indicating that 68% of observations will fall between these limits.

Plot of the 350 m contour for the State College, Pennsylvania, U.S.A. topographic quadrangle. The contour has been computed from the U.S. Geological Survey's digital elevation model for this area.



Uncertainty in the location of the 350 m contour based on an assumed RMSE of 7 m. The Gaussian distribution with a mean of 350 m and a standard deviation of 7 m gives a 95% probability that the true location of the 350 m contour lies in the colored area, and a 5% probability that it lies outside.

A Useful Rule of Thumb for Positional Accuracy

- Positional accuracy of features on a paper map is roughly 0.5mm on the map
  - e.g., 0.5mm on a map at scale 1:24,000 gives a positional accuracy of 12m
  - this is approximately the U.S. National Map Accuracy Standard
  - and also allows for digitizing error, stretching of the paper, and other common sources of positional error

A useful rule of thumb is that positions measured from maps are accurate to about 0.5 mm on the map. Multiplying this by the scale of the map gives the corresponding distance on the ground.

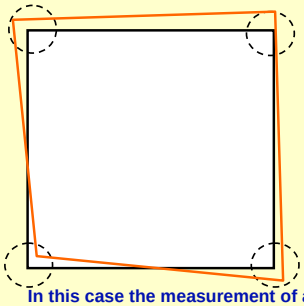
Map scale	Ground distance corresponding to 0.5 mm map distance
1:1250	62.5 cm
1:2500	1.25 m
1:5000	2.5 m
1:10,000	5 m
1:24,000	12 m
1:50,000	25 m
1:100,000	50 m
1:250,000	125 m
1:1,000,000	500 m
1:10,000,000	5 km

Correlation of Errors

- *Absolute* positional errors may be high
  - reflecting the technical difficulty of measuring distances from the Equator and the Greenwich Meridian
- *Relative* positional errors over short distances may be much lower
  - positional errors tend to be strongly correlated over short distances
- As a result, positional errors can largely cancel out in the calculation of properties such as distance or area

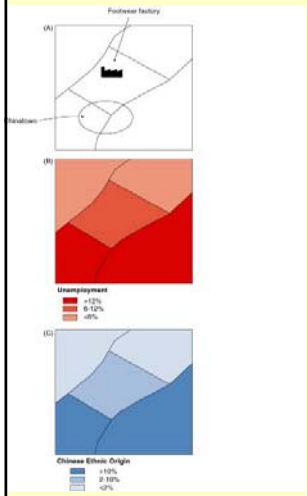
U3: Analysis. Error Propagation

- Addresses the effects of errors and uncertainty on the results of GIS analysis
- Almost every input to a GIS is subject to error and uncertainty
  - In principle, every output should have confidence limits or some other expression of uncertainty



Error in the measurement of the area of a square 100 m on a side. Each of the four corner points has been surveyed; the errors are subject to bivariate Gaussian distributions with standard deviations in *x* and *y* of 1 m (dashed circles). The red polygon shows one possible surveyed square (one *realization* of the error model).

In this case the measurement of area is subject to a standard deviation of 200 sq m; a result such as 10,014.603 is quite likely, though the true area is 10,000 sq m. In principle, the result of 10,014.603 should be rounded to the known accuracy and reported as as 10,000.



Ecological Fallacy

assuming a summarized variable is true for the entire population

MAUP

- Scale + aggregation = MAUP
  - can be *investigated* through simulation of large numbers of alternative zoning schemes

Living with Uncertainty

- It is easy to see the importance of uncertainty in GIS
  - but much more difficult to deal with it effectively
  - but we may have no option, especially in disputes that are likely to involve litigation

Some Basic Principles

- Uncertainty is inevitable in GIS
- Data obtained from others should never be taken as truth
  - efforts should be made to determine quality
- Effects on GIS outputs are often much greater than expected
  - there is an automatic tendency to regard outputs from a computer as the truth

More Basic Principles

- Use as many sources of data as possible
  - and cross-check them for accuracy
- Be honest and informative in reporting results
  - add plenty of caveats and cautions

Consolidation

- Uncertainty is more than error
- Richer representations *create* uncertainty!
- Need for *a priori* understanding of data and sensitivity analysis