Chapter 5 – Uncertainty

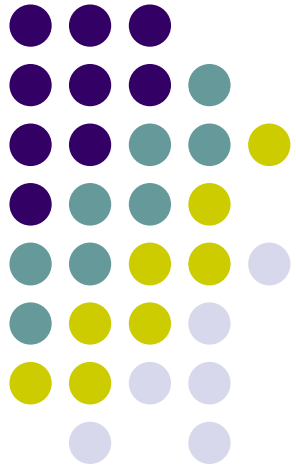
§ 1 Introduction

§ 2 U1: Uncertainty in Conception

§ 3 U2: Uncertainty in Measurement & Representation

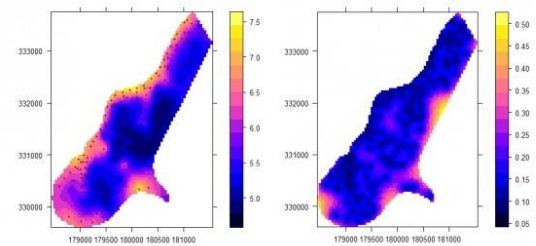
§ 4 U3: Uncertainty in Analysis

§ 5 Consolidation

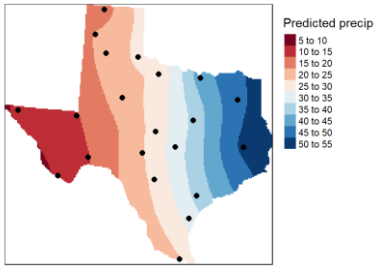


§ 1 Introduction

- Impossible to make a perfect representation, so uncertainty is inevitable
  - Representations are used to reconcile , which is imperfect
    - science with practice
    - concepts with applications
    - analytical methods with social context
  - Inherent complexity of the world makes it impossible to capture every single facet
- Representations depend upon inherently vague definitions and concepts
  - Error in measurement, multivariate statistics, accuracy
  - Components of quality: attribute accuracy, positional accuracy, logical consistency(一致性), completeness（完备性）, and lineage（沿袭）
- The term *uncertainty* is the catch-all term to describe situations in which the digital representation is simply incomplete, and as a measure of the general quality of the representation



Uncertainty Visualization



Uncertainty of Interpolation

A conceptual view of uncertainty

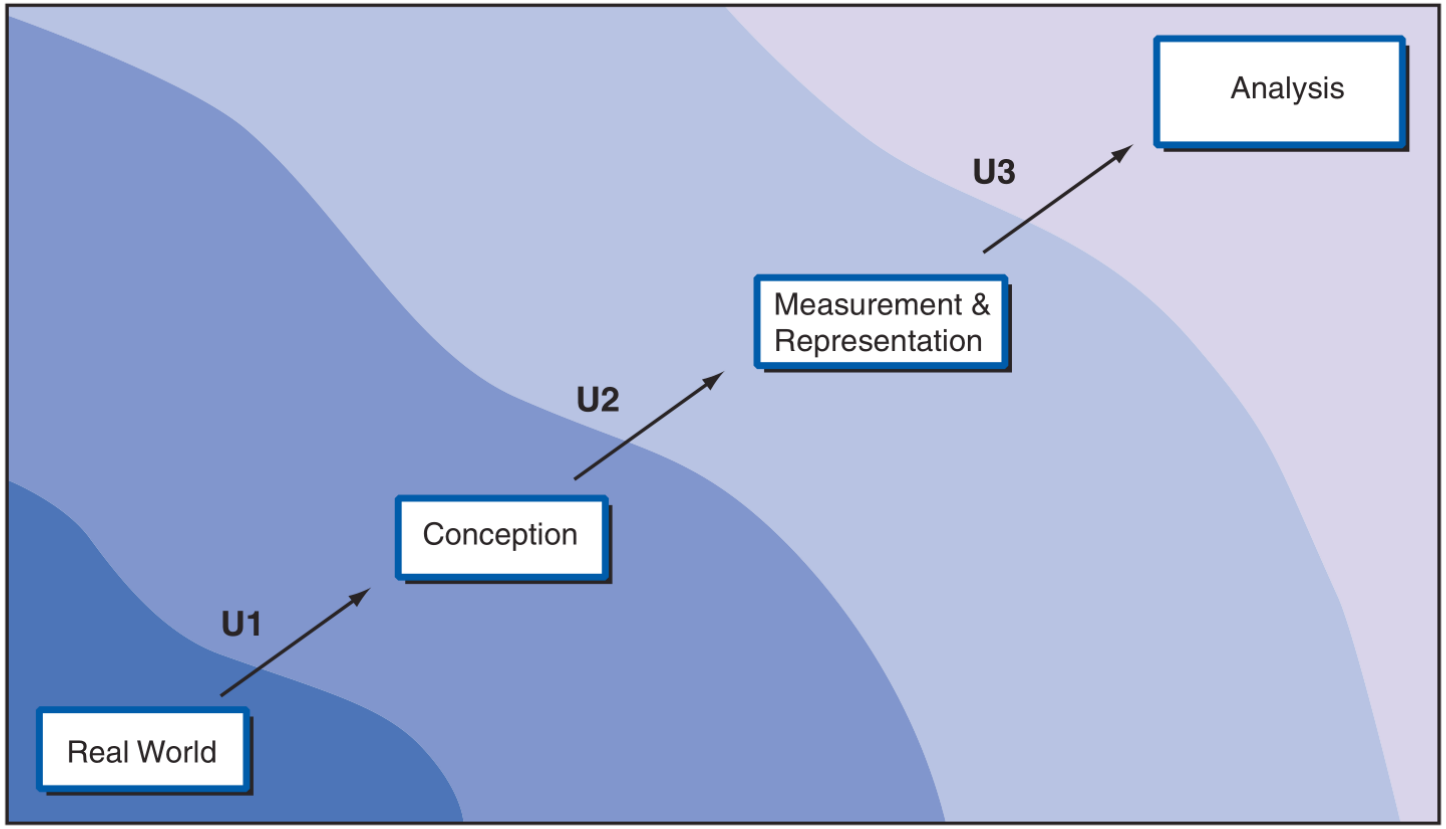


Figure 6.1 A conceptual view of uncertainty. The three filters, U1, U2, and U3 can distort the way in which the complexity of the real world is conceived, measured and represented, and analyzed in a cumulative way

§ 2 U1:Uncertainty in Conception

§ 2.1 Units of analysis

- geographic information science is only rarely founded upon natural units of analysis(not like a house , a pensil, a car etc.)
  - natural unit of measurement for a soil profile
  - spatial extent of a cluster of cancer cases
  - An environmental impact scope of spillage of an oil tanker
  - More difficult in bivariate and multivariate studies （e.g. ecological devision)
- The decisions about units of analysis is inherently subjective

Uncertainty of natural units

Geo-objects with Uncertainty

§ 2 U1:Uncertainty in Conception(cont.)

§ 2.2 Vagueness and ambiguity

- *Vagueness* both in the *positions* of the boundaries and in its *attributes*
  - Is the defining boundary of a zone crisp and well-defined?
  - Is our assignment of a particular label to a given zone robust and defensible?
- *ambiguity* Linguistic terms used to convey geographic information are ambiguous
  - Perception, behavior, language, and cognition all play a part in the conception of real-world entities and the relationships between them
  - Ambiguity is introduced when imperfect indicators of phenomena are used instead of the phenomena themselves
  - Differences in definitions are a major impediment to integration of geographic data over wide areas

Oak Woodland

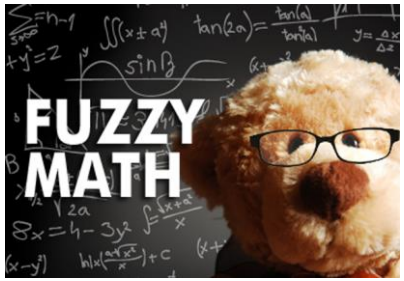
Strasse vs Street

Pond vs Lake

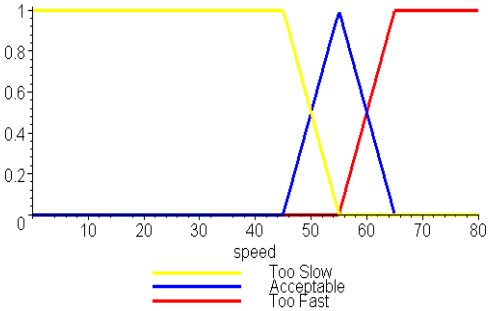
§ 2 U1:Uncertainty in Conception(cont.)

§ 2.3 Fuzzy approaches

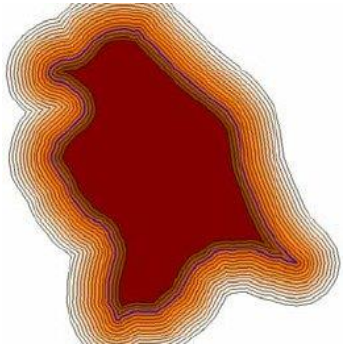
- *One way of resolving the assignment process is to adopt a probabilistic interpretation*
- In fuzzy logic, an object's degree of belonging to a class can be partial
- they appear to let us deal with sets that are not precisely defined
- Classes used for maps are often fuzzy, such that two people asked to classify the same location might disagree, not because of measurement error, but because the classes themselves are not perfectly defined and because opinions vary



Fuzzy mathematics



Fuzzy speed of a car



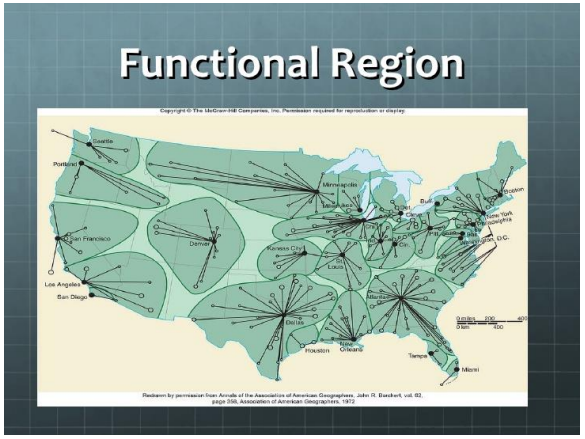
Fuzzy boundary

§ 2 U1:Uncertainty in Conception(cont.)

§ 2.4 The scale of geographic individuals

- Identification of homogeneous zones and spheres of influence lies at the heart of traditional regional geography as well as nowaday data analysis
- Relationships grow stronger when based on larger geographic units
  - Other geographers have tried to develop functional zonal schemes
  - Zone boundaries delineate the breakpoints between the spheres of influence of adjacent facilities or features

Seismic Zoning

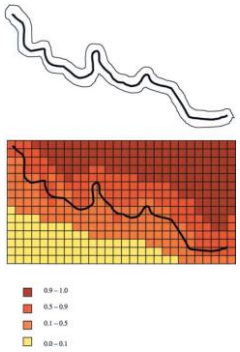


Founctional Regions

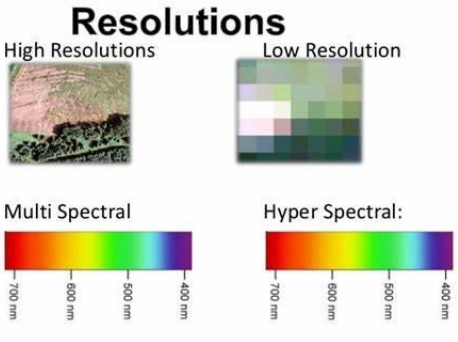
§ 3 U2:Uncertainty in Measurement & Representation

§ 3.1 Measurement and Representation

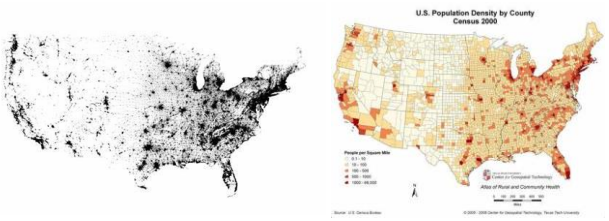
- Uncertainty can be measured differently under field and object views
  - In object view the line becomes an area delimiting where the true coastline might be
  - In field view a continuous surface defines the probability that any point is land
- A pixel whose area is divided among more than one class is termed a mixel
- In vector aggregation can be driven by the need to preserve confidentiality of individual records, or simply by the need to limit data volum



Uncertainty in Vector vs Raster



Mixel with loss of classification



Aggregation of population data

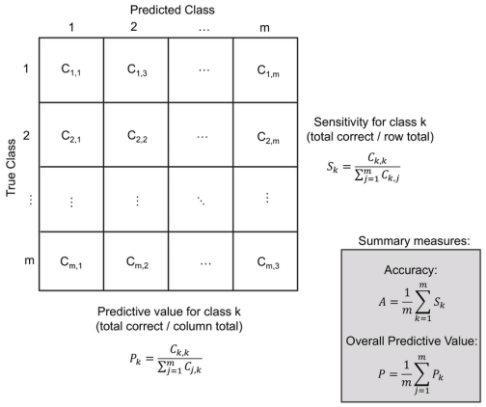
§ 3 U2:Uncertainty in Measurement & Representation(cont.)

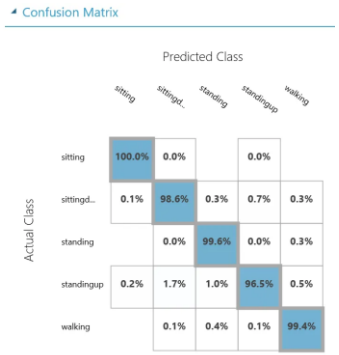
§ 3.2 Statistical models of uncertainty

- A geographic database is a collection of measurements of phenomena on or near the Earth's surface
- Nominal case(class):
  - Users and producers look at misclassification in distinct ways(*Confusion matix*)

misclassification or confusion matrix

	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



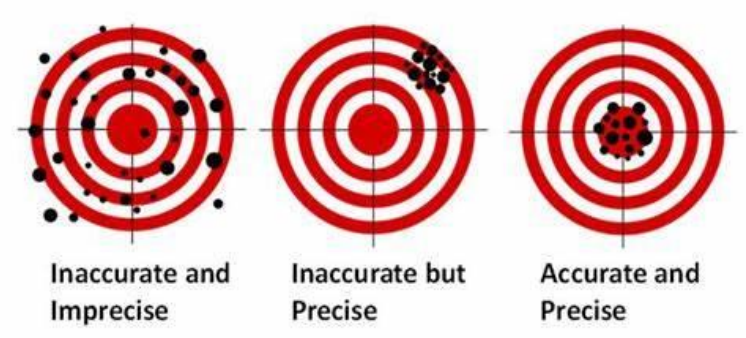


  - Sampling for accuracy assessment should pay greater attention to the classes that are rarer on the ground
  - Errors in land cover maps can occur in the locations of boundaries of areas, as well as in the classification of areas

§ 3 U2:Uncertainty in Measurement & Representation(cont.)

§ 3.2 Statistical models of uncertainty(cont.)

- Interval/ratio case(value):
  - Error in measurement can produce a change of class, or a *change of value*
  - accuracy refers to the difference between reality and our representation of reality
  - Precision refers to the number of significant digits used to report a measurement
  - It can also refer to a measurement's repeatability
    - Root mean square error (RMSE) – magnitude of error
    - Gaussian/Normal/Bell – error distribution

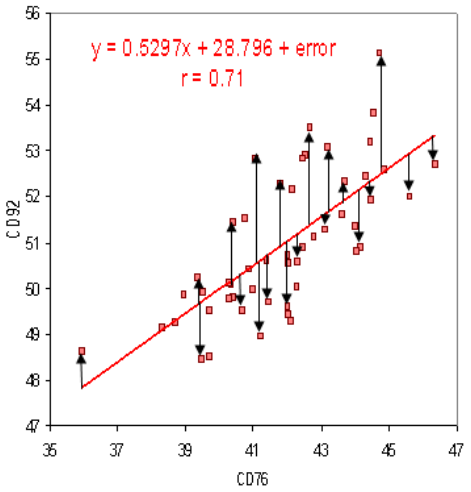


Precision vs Accuracy

Statisitc Models

- *Root Mean Square Error (RMSE,均方根误差,标准误差)*
  - Root Mean Square Error (RMSE) measures how much error there is between two data sets
  - It compares a predicted value and an observed or known value
  - The smaller an RMSE value, the closer predicted and observed values are

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$



Predicted vs observed value

Mean squared error	$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$MAE = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$MAPE = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

Different measures of error

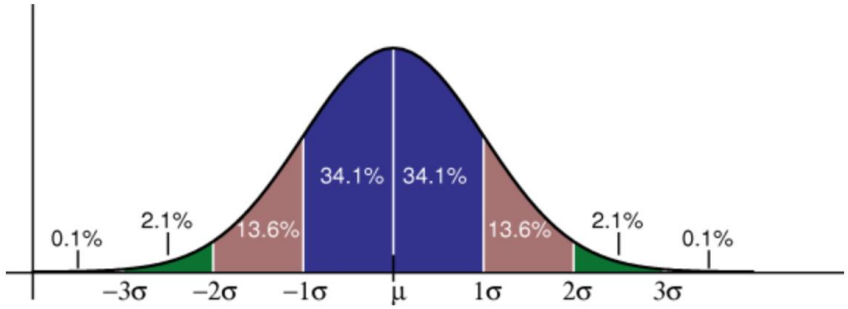
Statisitc Models（cont.）

- *Gaussian distribution*（高斯分布，正态分布，钟形曲线）
  - The Gaussian (normal) distribution was historically called the *law of errors*
  - It was used by Gauss to model errors in astronomical observations, which is why it is usually referred to as the Gaussian distribution
  - The *probability density function* for the standard Gaussian distribution (mean 0 and standard deviation 1) and the Gaussian distribution with mean μ and standard deviation σ is given by the following formulas

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$\phi(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$



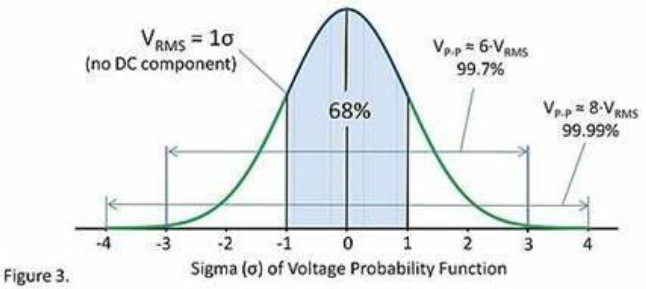


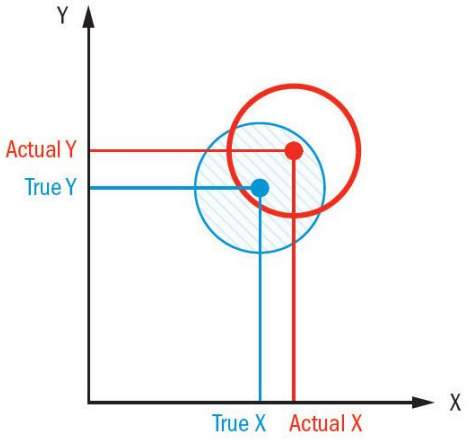
Figure 3.

Gaussian/Normal/Bell distribution

§ 3 U2:Uncertainty in Measurement & Representation(cont.)

§ 3.3 Positional Error

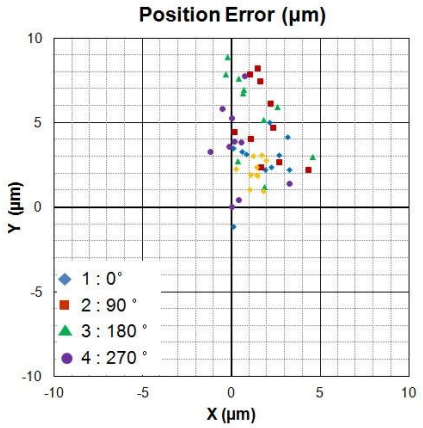
- In the case of measurements of position, Every coordinate subjects to error
- National Map Accuracy Standards often prescribethe positional errors that are allowed in databases
- A useful rule of thumb is that features on maps are positioned to an accuracy of about 0.5 mm



Measure X and Y location
and compare to the true position.

$$2 \cdot \sqrt{(\text{Actual X} - \text{True X})^2 + (\text{Actual Y} - \text{True Y})^2}$$

This formula must be less than
the ⌀ True Position tolerance



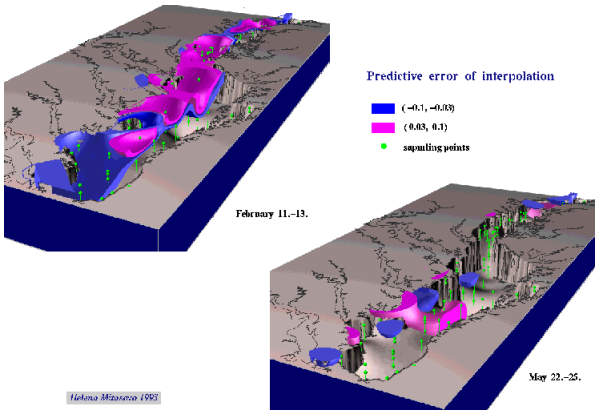
Loading Position	RMS Error (µm)
1 : 0°	3.779
2 : 90°	5.885
3 : 180°	6.423
4 : 270°	4.430
Average	**5.129**

Positional Error

§ 3 U2:Uncertainty in Measurement & Representation(cont.)

§ 3.4 The spatial structure of errors

- The spatial autocorrelation of errors can be as important as their magnitude in many GIS operations
- The spatial autocorrelation between errors helps to minimize their impacts on many GIS operations
- Spatial autocorrelation acts to reduce the effective number of degrees of freedom in geographic data



§ 4 U3:Uncertainty in analysis

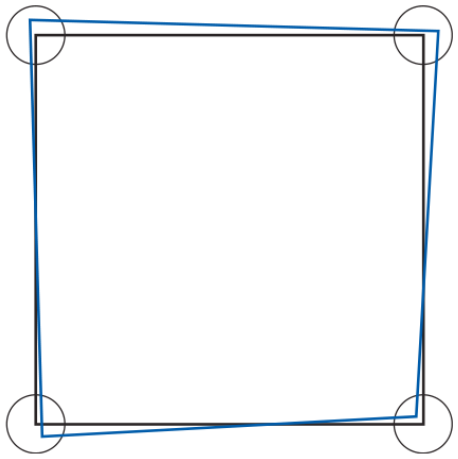
§ 4.1 Internal and external validation through spatial analysis

- Uncertainties in data lead to uncertainties in the results of analysis
- GIS gives us maximum flexibility when working with aggregate data
- Helps us to validate our data with reference to other available sources
- Three ways of dealing with the uncertainty:
  - although we can only rarely tackle the source of distortion, we can quantify the way in which it is likely to operate (or propagates) within the GIS, and can gauge the magnitude of its likely impacts
  - although we may have to work with aggregated data, GIS allows us to model within-zone spatial distributions in order to ameliorate the worst effects of artificial zonation
  - GIS allows us to gauge the effects of scale and aggregation through simulation of different possible outcomes
- This is internal validation of the effects of scale, point placement, and spatial partitioning

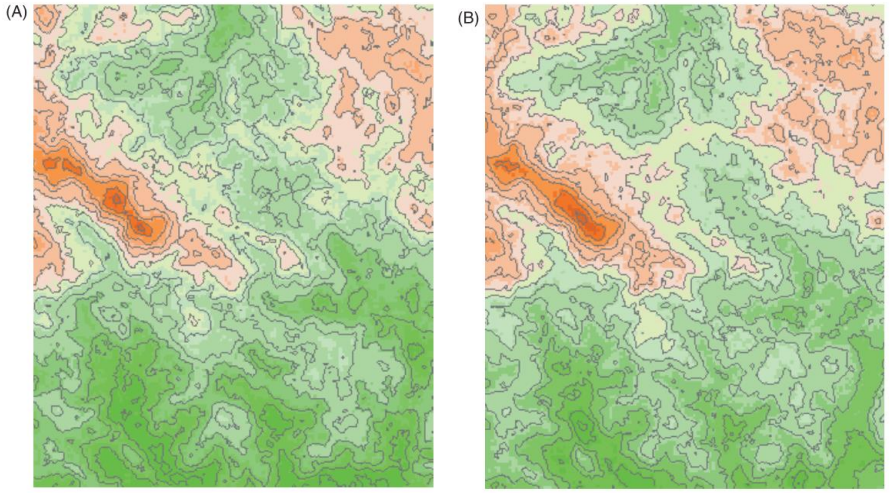
§ 4 U3:Uncertainty in analysis (cont.)

§ 4.2 Internal validation: error propagation

- Error propagation measures the impacts of uncertainty in data on the results of GIS operations
- Simulation is an intuitively simple way of getting the uncertainty message across



Error in the measurement of the area of a square



realizations of a model simulating the effects of error on a digital elevation model.

§ 4 U3:Uncertainty in analysis (cont.)

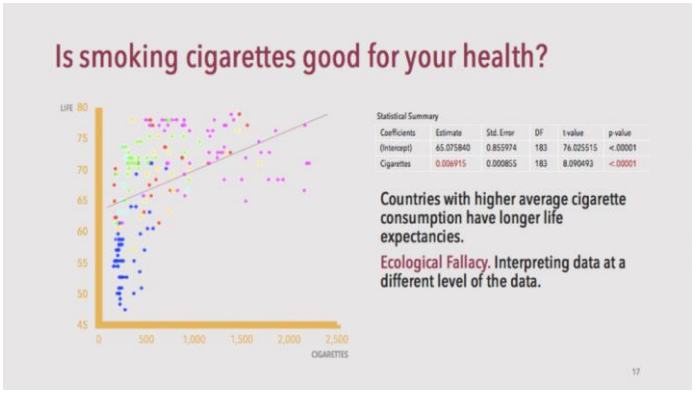
§ 4.3 Internal validation: aggregation and analysis

- In socio-economic GIS objects of study are usually aggregations
- Confidentiality restrictions usually dictate that uniquely attributable information must be anonymized in some way
- We cannot be certain in ascribing even dominant characteristics of areas to true individuals or point locations in those areas
- Inappropriate inference from aggregate data about the characteristics of individuals is termed the *ecological fallacy*(生态谬误)
- The effects of scale and aggregation are generally known as the *Modifiable Areal Unit Problem (MAUP,*可变面元问题*)*
- The *ecological fallacy* and the *MAUP* have long been recognized as problems in applied spatial analysis and, through the concept of spatial autocorrelation

Ecological fallacy

- often called an *ecological inference fallacy*, is an error in the interpretation of statistical data in an ecological study, whereby inferences about the nature of specific individuals are based solely upon aggregate statistics collected for the group to which those individuals belong
- Ecological fallacies can be grouped into a few divisions:
  - Confusion between aggregate correlations and individual correlations
  - Confusion between the group average and the likelihood of the individuals to reflect that average
  - Simpson's paradox
  - Confusion between the idea that a group having a higher average of something occurring will also have a higher likelihood of happening again

Complexity of Ecology

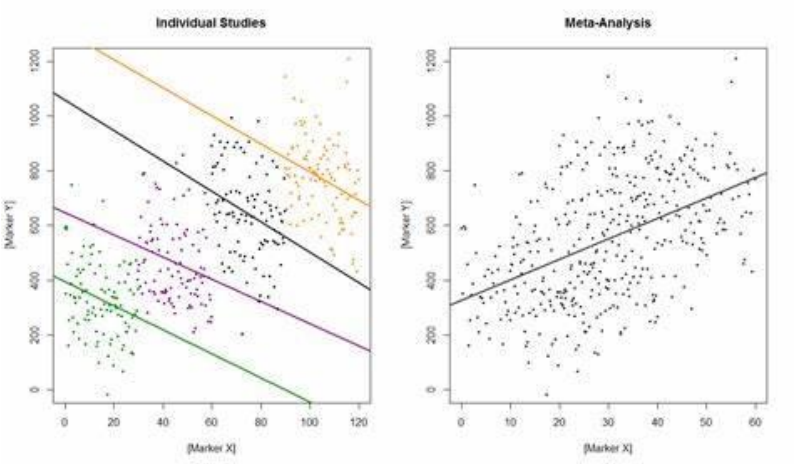


Ecological Fallacy in Cigarette Consumption

Simpson's paradox(辛普森悖论)

- **Simpson's Paradox**, also known as the **amalgamation paradox**, **reversal paradox**, or **Yule-Simpson Effect**, is a problem studied in statistics. It describes how when looking at data in groupings, one trend may be observed, but when looking at the aggregation of the group data, that trend may be the opposite
- The record for on-time-completion of city road projects is being compared between two cities over three years

Year	City A		City B	
1	1/4	25%	3/9	33%
2	5/6	83%	2/2	100%
3	9/15	60%	3/3	75%
	9/15	60%	8/15	53%



Individual Studies vs Meta-analysis

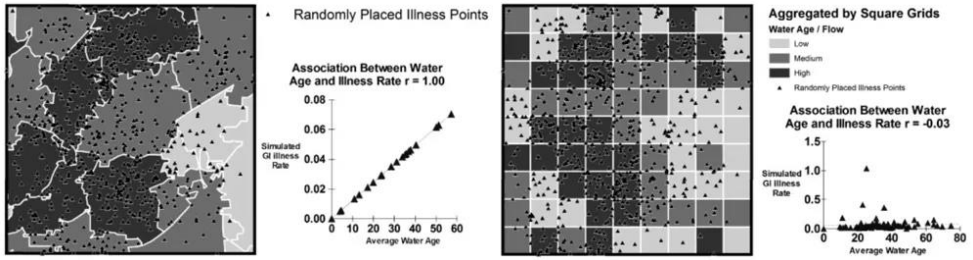
	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Department	# of Men	# of Women	Men Accepted	Women Accepted
A	825	108	62%	82%
B	560	25	63%	68%
C	325	593	37%	34%
D	417	375	33%	35%
E	191	393	28%	24%
F	373	341	6%	7%
Total	8442	4321		

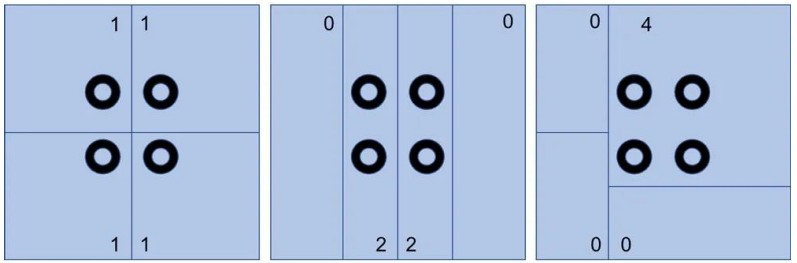
UC Berkeley Applicants

Modifiable Areal Unit Problem (MAUP)

- discovered in 1934, but the term was first described Openshaw, the areal units used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating
- The problem is especially apparent when the aggregate data are used for cluster analysis for spatial epidemiology etc
- MAUP is closely related to the topic of ecological fallacy and ecological bias
- Ecological bias as two separate effects:
  - Causes variation in statistical results between different levels of aggregation. Generally, correlation increases as areal unit size increases
  - The zone effect describes variation in correlation statistics caused by the regrouping of data into different configurations at the same scale



different levels of aggregation

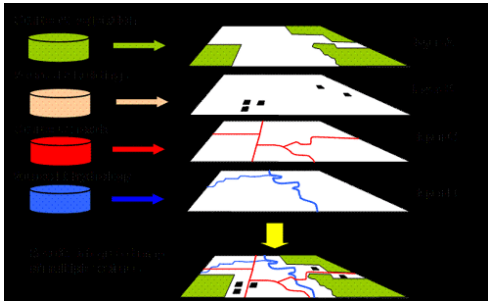


regrouping of data

§ 4 U3:Uncertainty in analysis (cont.)

§ 4.4 External validation: data integration and shared lineage

- *Conflation* combines the information from two data sources into a single source
- Because neighboring points are more likely to share *lineage* than distant points, errors tend to exhibit strong positive spatial autocorrelation
- Data sets with different lineages often reveal unsuspected errors when overlaid
- There is an emergent tension within the socio-economic realm,for there is a limit to the uses of inferences drawn from conventional, scientifically valid data sources which are frequently out-of-date, zonally coarse, and irrelevant to what is happening in modern societies
- Yet the alternative of using new rich sources of marketing data may be profoundly unscientific in its inferential procedures



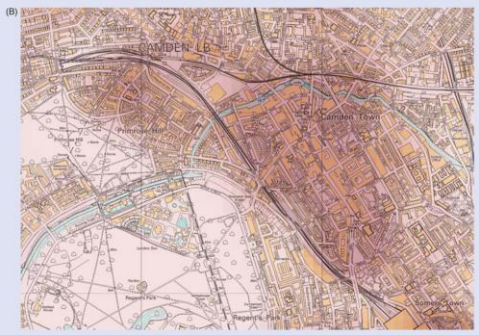
Integration of multi-source data sets

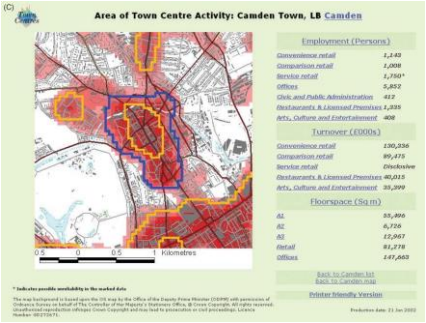
Overlay of data sets with different lineages

§ 4 U3:Uncertainty in analysis(cont.)

§ 4.5 Internal and external : induction and deduction

- The Modifiable Areal Unit Problem can be investigated through simulation of large numbers of alternative zoning schemes
- Zoning seems similar to sampling, but its effects are very different
- The way forward seems to be to complement our new-found abilities to customize zoning schemes in GIS with external validation of data and clearer application- centered thinking about the likely degree of within-zone heterogeneity that is concealed in our aggregated data
- MAUP will disappear if GIS analysts understand the particular areal units that they wish to study
- There is also a practical recognition that the areal objects of study are ever-changing, and our perceptions of what constitutes their appropriate definition will change
- within the socio-economic realm, the act of defining zones can also be self-validating if the allocation of individuals affects the interventions they receive





§ 5 Consolidation

- Richness of representation and computational power only make us more aware of the range and variety of established uncertainties, and challenge us to integrate new ones
- The fathoming of uncertainty requires a combination of the cumulative development of *a priori*, external validation of data sources, and inductive generalization in the fluid, eclectic data-handling environment that is contemporary GIS