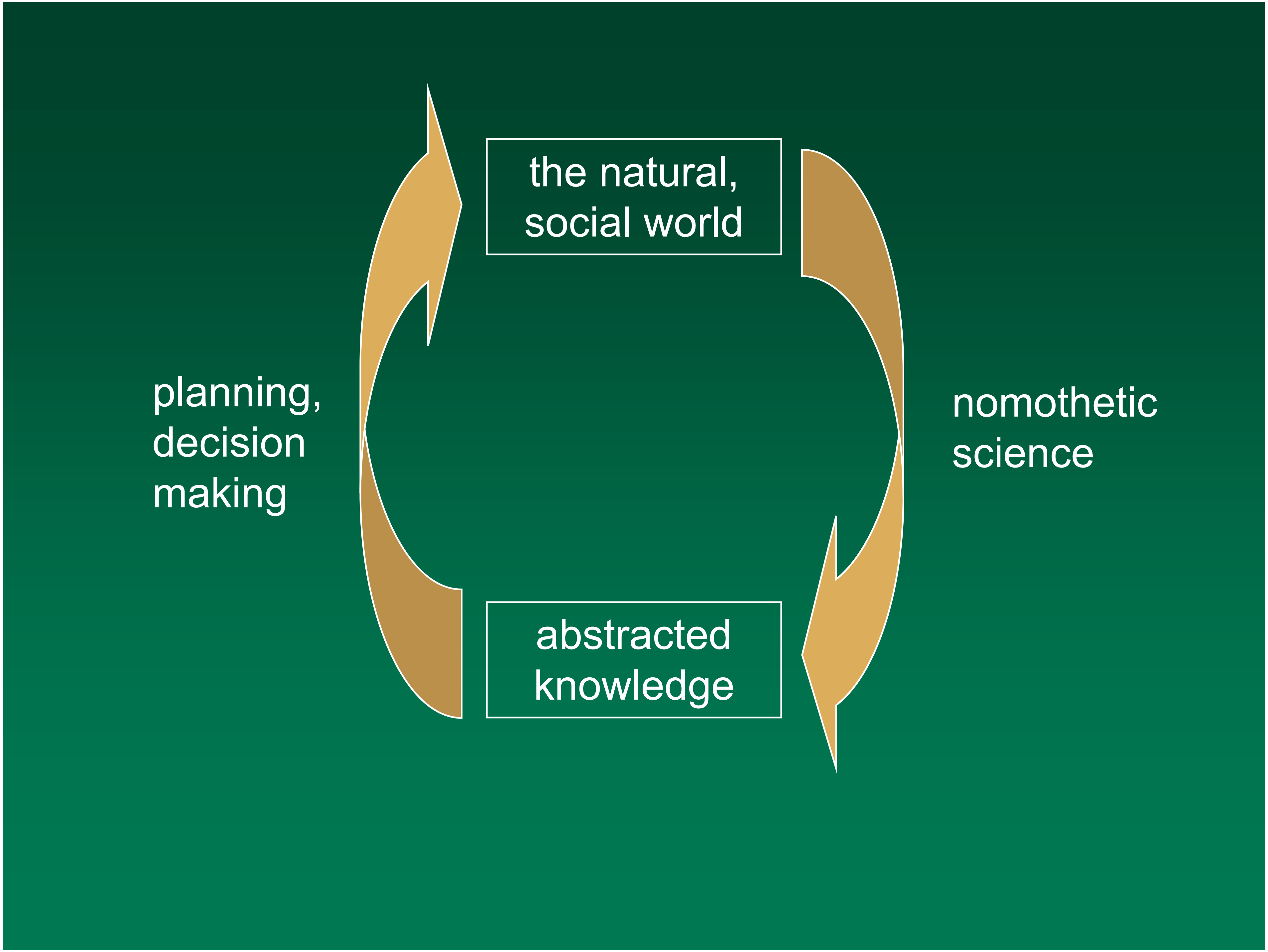
Statistical Perspectives on Spatial Social Science

Michael F. Goodchild
University of California
Santa Barbara

A conceptual framework

- Nomothetic science
  - knowledge that is true everywhere in space and time
- Idiographic science
  - the study of the unique
  - new planets
  - liquid lakes of Antarctica
  - *descriptive*, *anectodal* can be pejorative


the natural, social world
planning, decision making
nomothetic science
abstracted knowledge

Where does this leave geography (or history)?

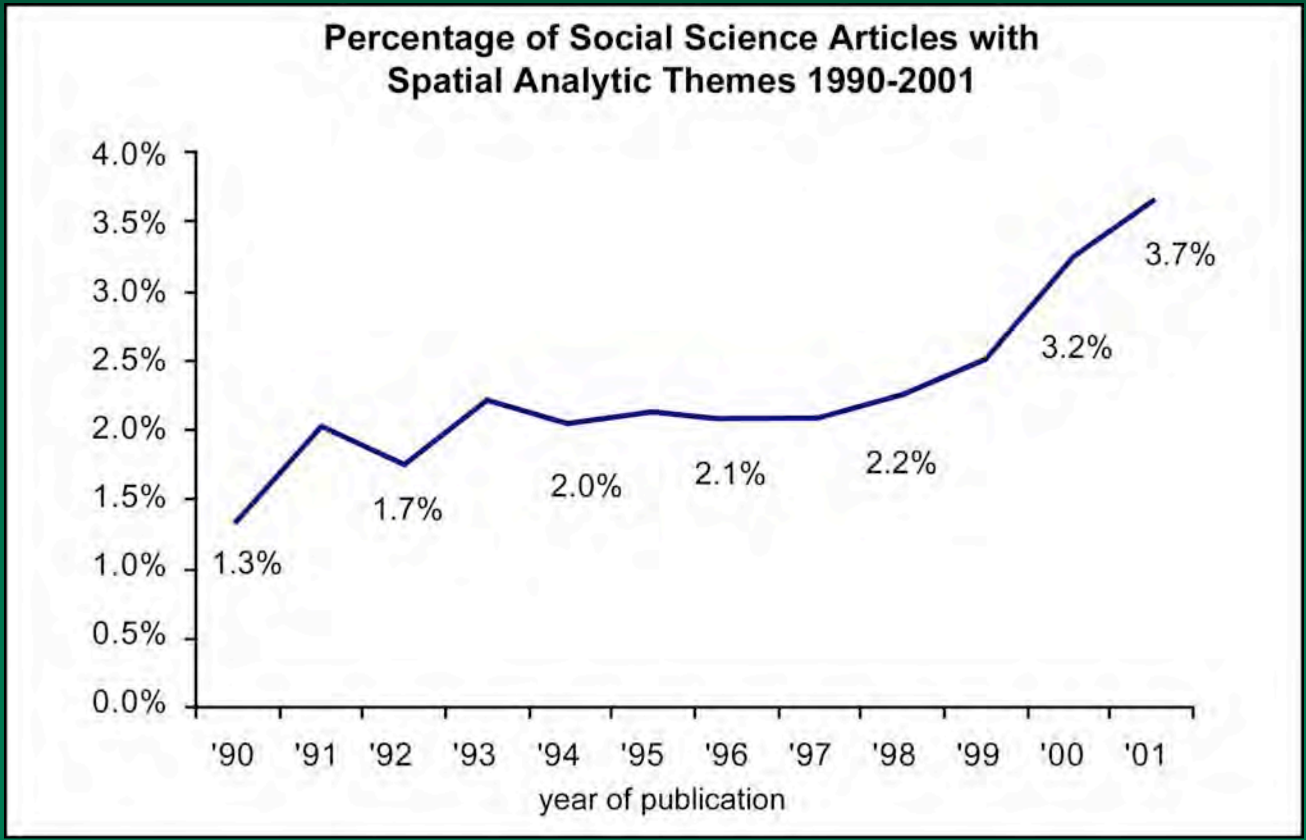
- An ongoing debate
  - the nomothetic ascendancy of the 1960s
  - Bunge's *Theoretical Geography*
- Varenius, 17th Century
  - *General Geography*
  - *Special Geography*
- Modern technology
  - the database as description
  - the software, models, analytic methods supporting nomothetic science
    - and planning, decision making

A *spatial turn* in science

- Adding space to theory
  - the *New Economic Geography*
    - space impeding flows of information, operation of markets
    - transport costs
  - *Spatial Ecology*
    - a heterogeneous resource base
    - space impeding interactions, breeding
    - metapopulations
- Reasoning from spatial data
  - cross-sectional
  - new tools to overcome methodological problems
  - impacts in all social, environmental disciplines

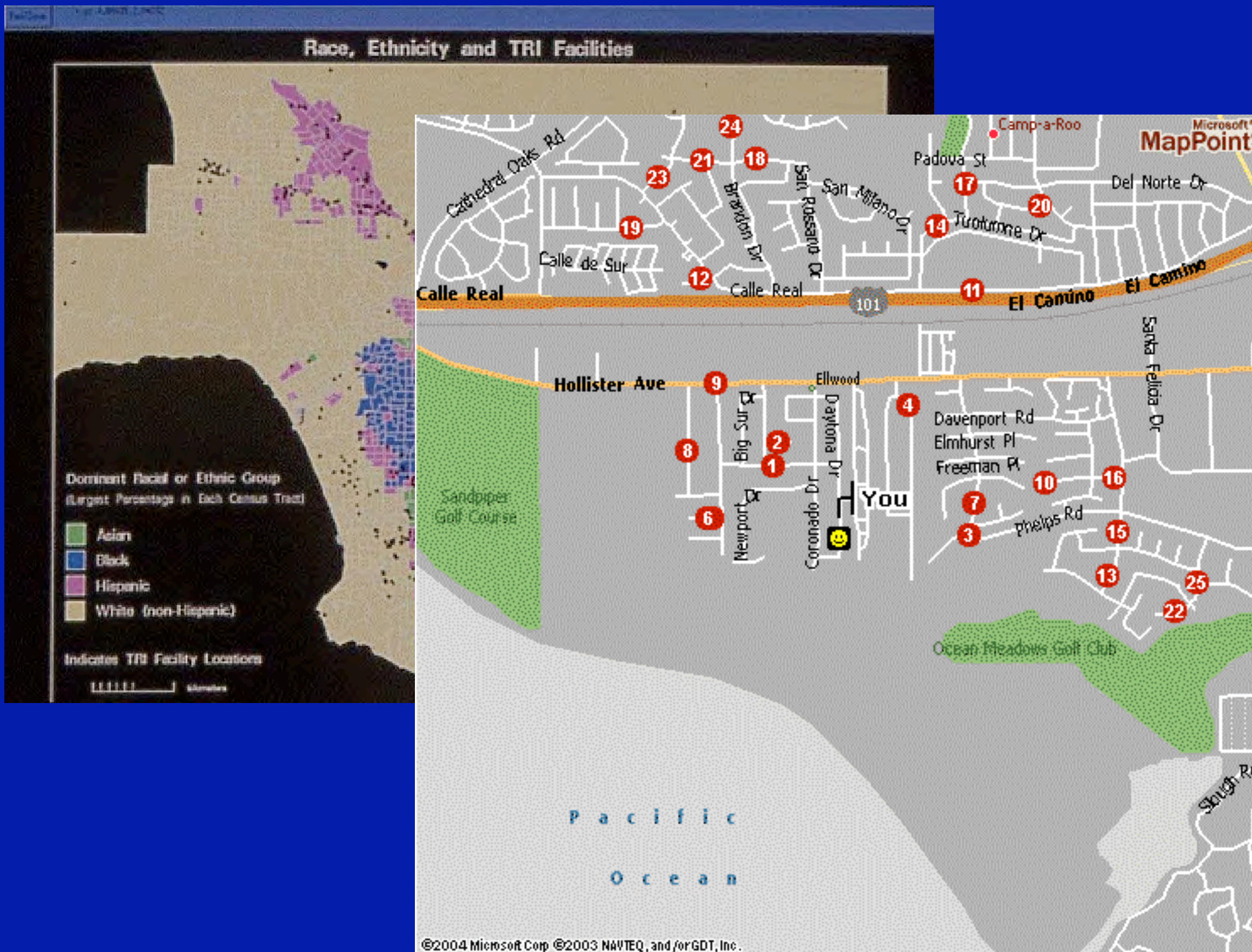
A growing literature

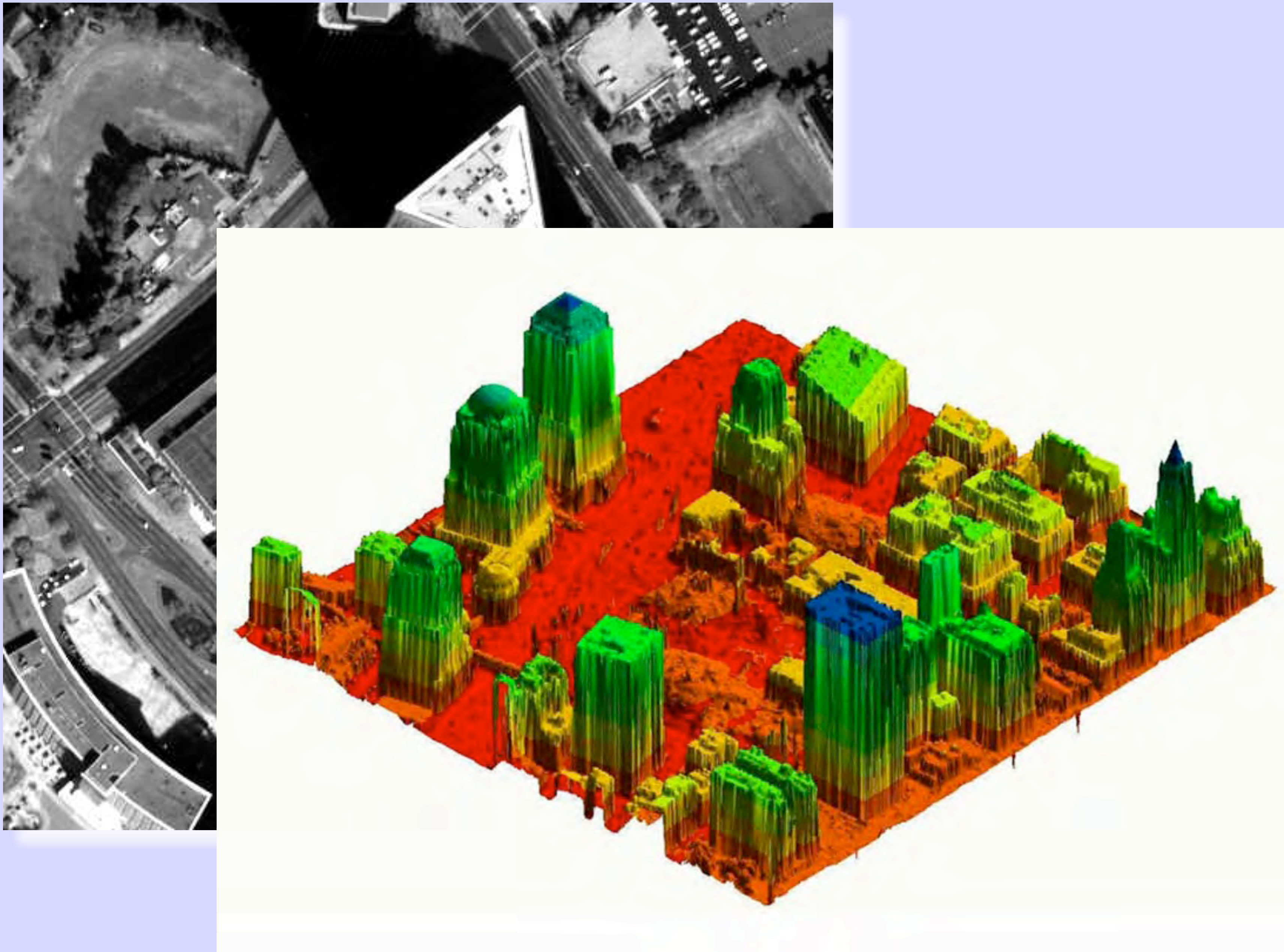
Spatially Integrated Social Science (Goodchild and Janelle, OUP, 2004)

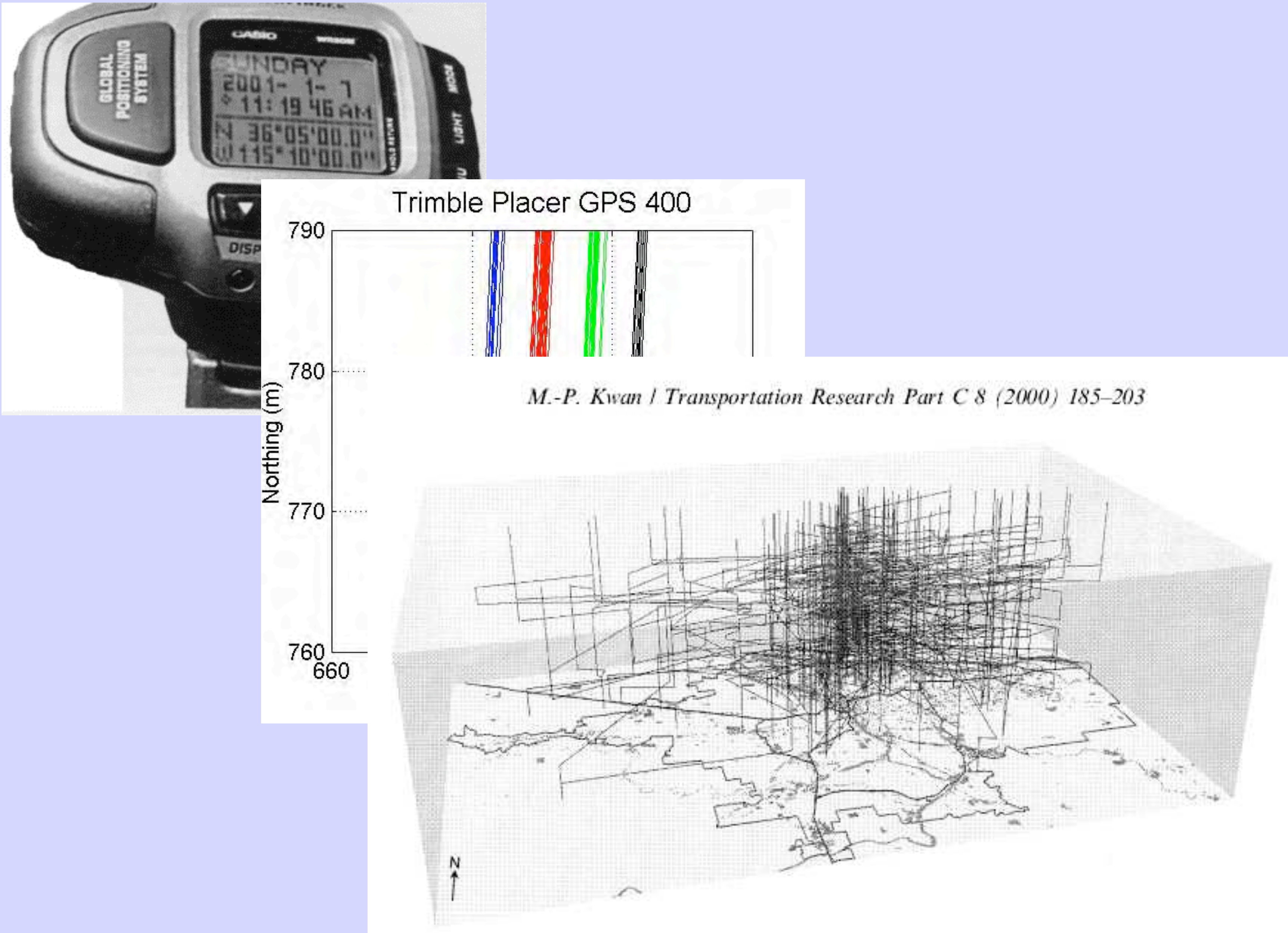


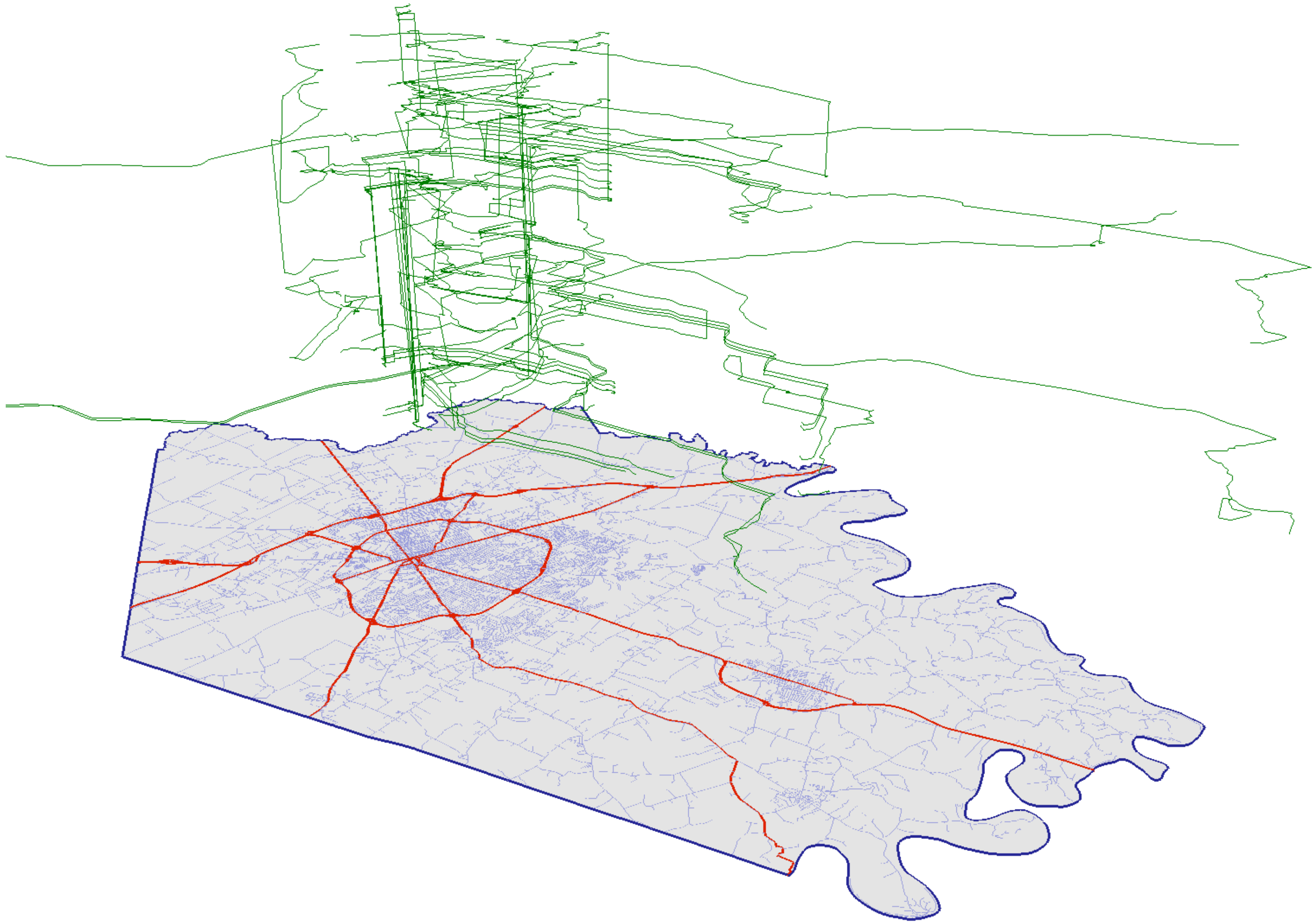
The drivers

- New technologies, new data
  - geographic information systems (GIS)
  - remote sensing
  - positioning (GPS)
  - delivery mechanisms
- Place-based analysis
- Applications of science in policy, decision making

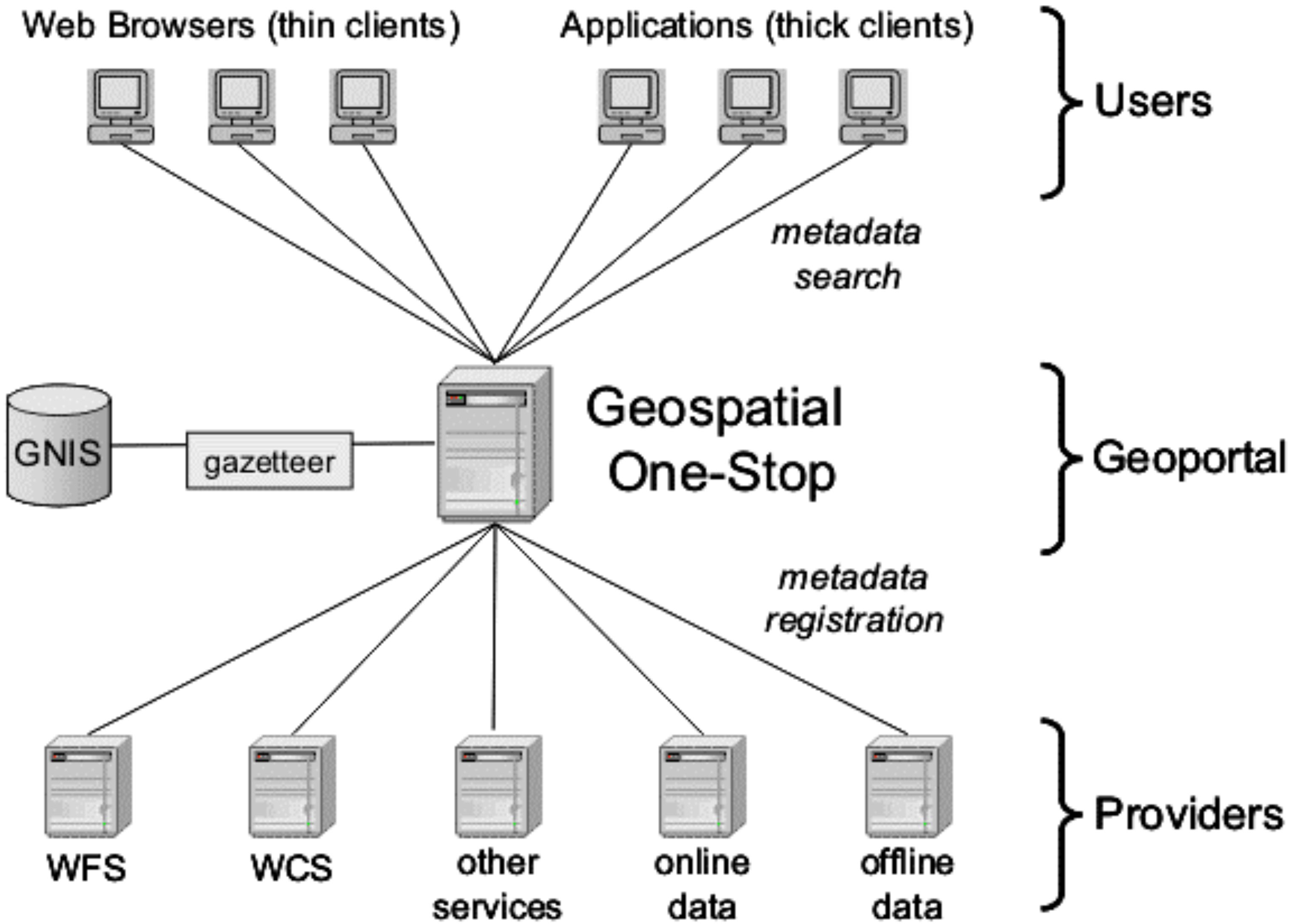

Race, Ethnicity and TRI Facilities
Dominant Racial or Ethnic Group
(Largest Percentage in Each Census Tract)
Asian
Black
Hispanic
White (non-Hispanic)
Indicates TRI Facility Locations
Microsoft MapPoint
Camp-a-Roo
Cathedral Oaks Rd
Padova St
Del Norte Dr
Brandon Dr
San Rossano Dr
San Milano Dr
Tuolumne Dr
Calle de Sur
Calle Real
El Camino
101
Hollister Ave
Ellwood
Big Sur Dr
Daytona Dr
Davenport Rd
Elmhurst Pl
Freeman Pl
Santa Felicia Dr
Newport Dr
Coronado Dr
You
Phelps Rd
Sandpiper Golf Course
Ocean Meadows Golf Club
Pacific Ocean
Slough Rd
©2004 Microsoft Corp ©2003 NAVTEQ, and/or GDT, Inc.

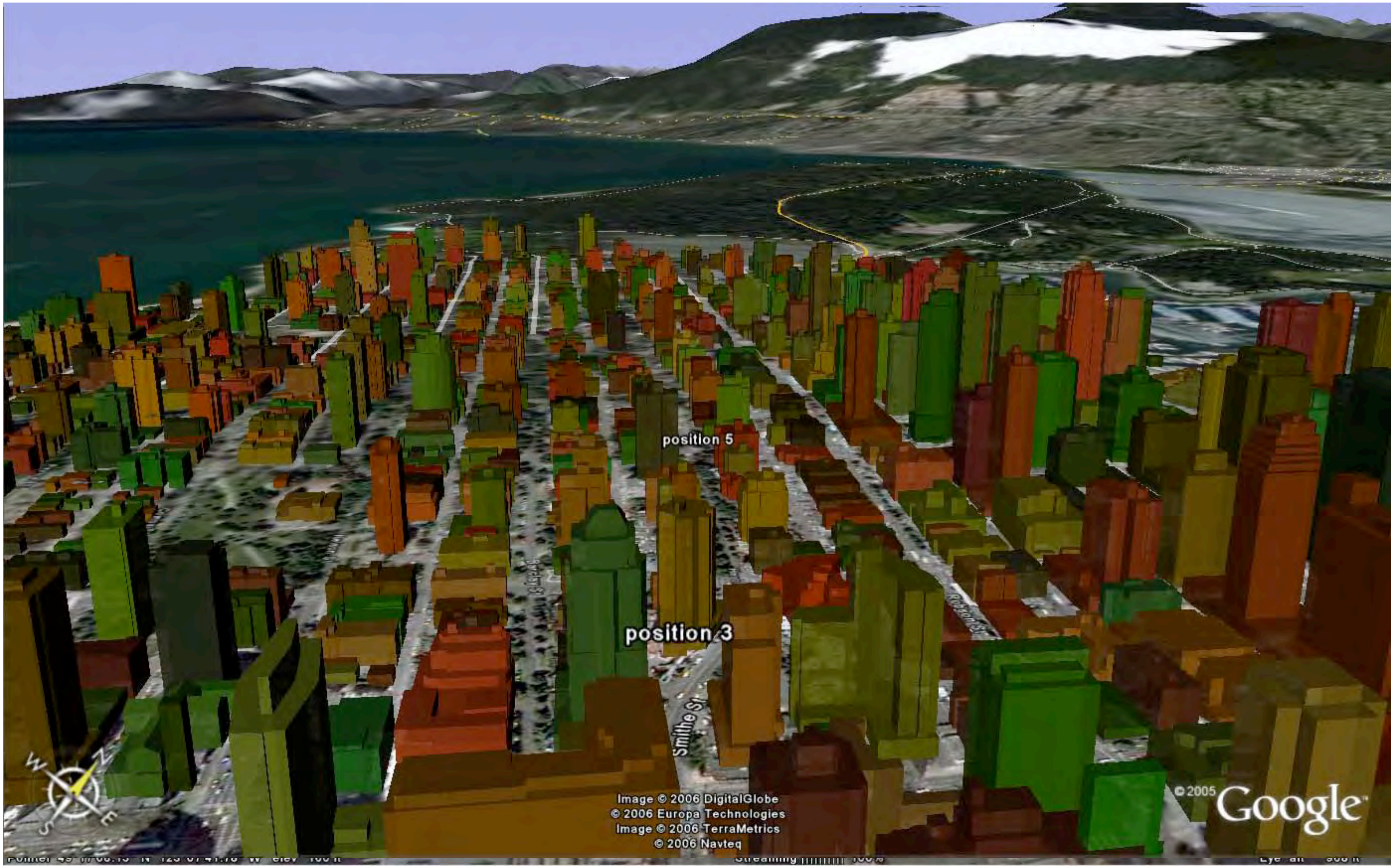


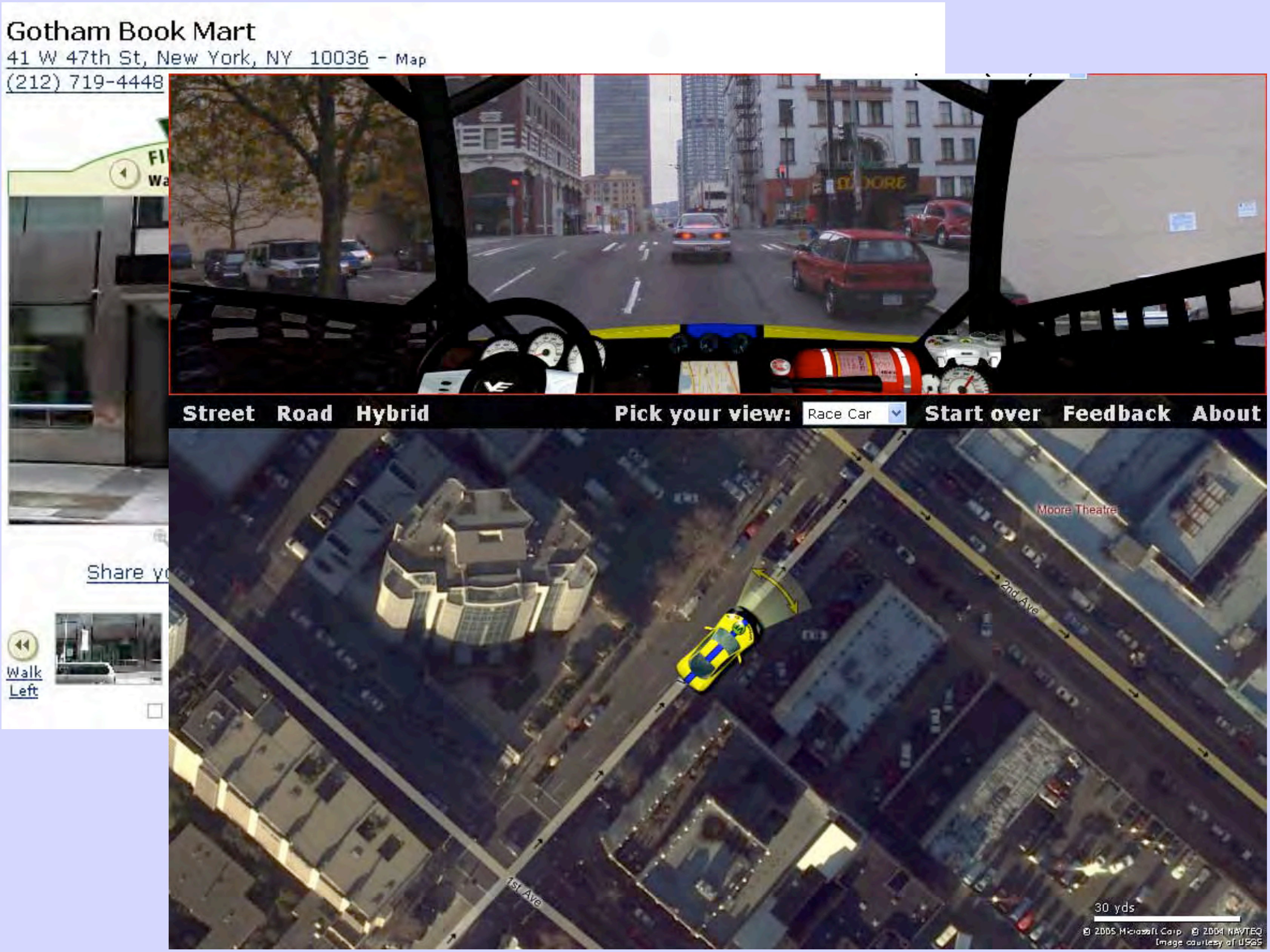

GLOBAL POSITIONING SYSTEM
CASIO
2001- 1- 7
11:19 46 AM
N 36°05'00.0"
W 115°10'00.0"
MODE
LIGHT
Trimble Placer GPS 400
790
780
770
760
660
Northing (m)
M.-P. Kwan / Transportation Research Part C 8 (2000) 185–203
N

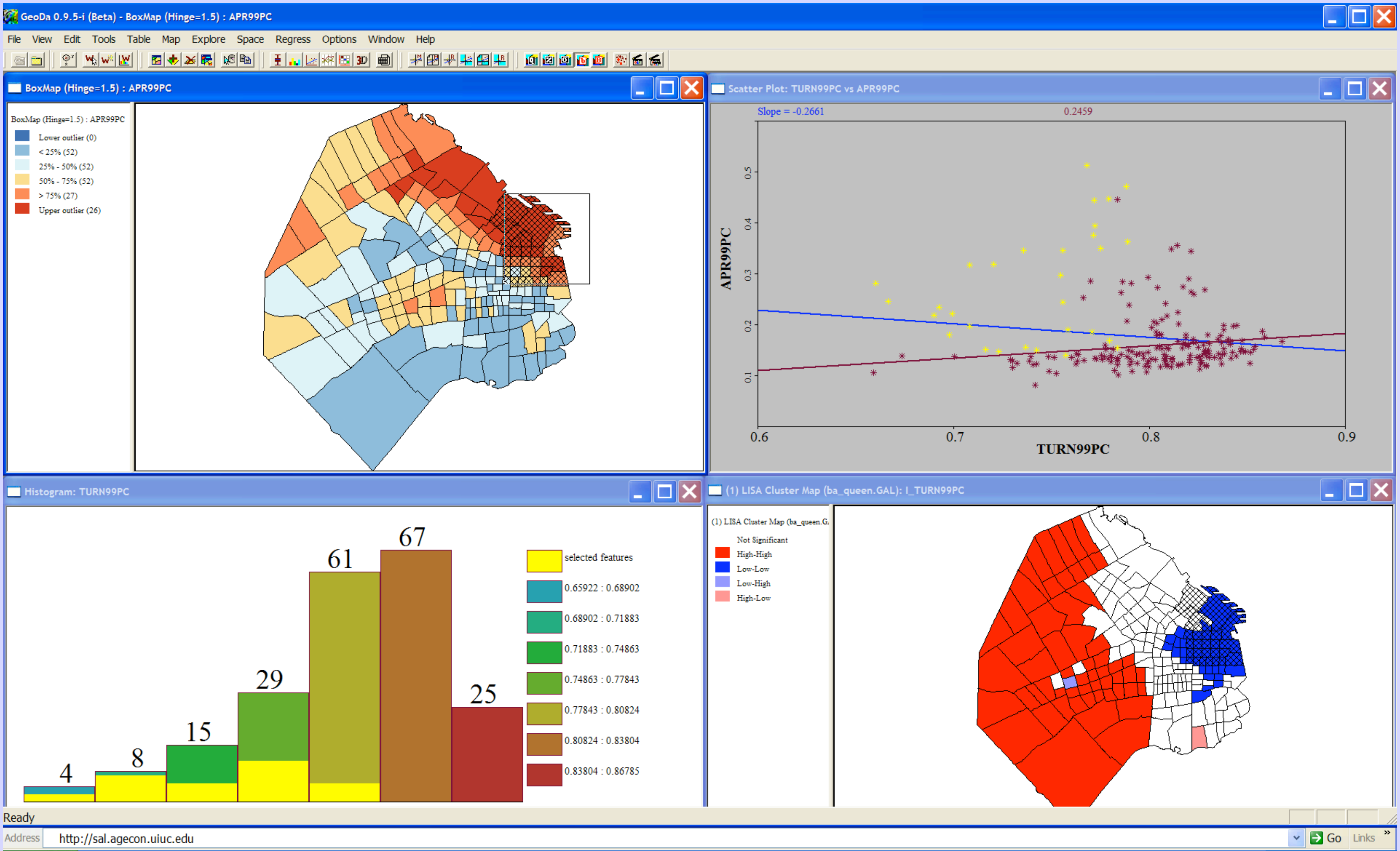


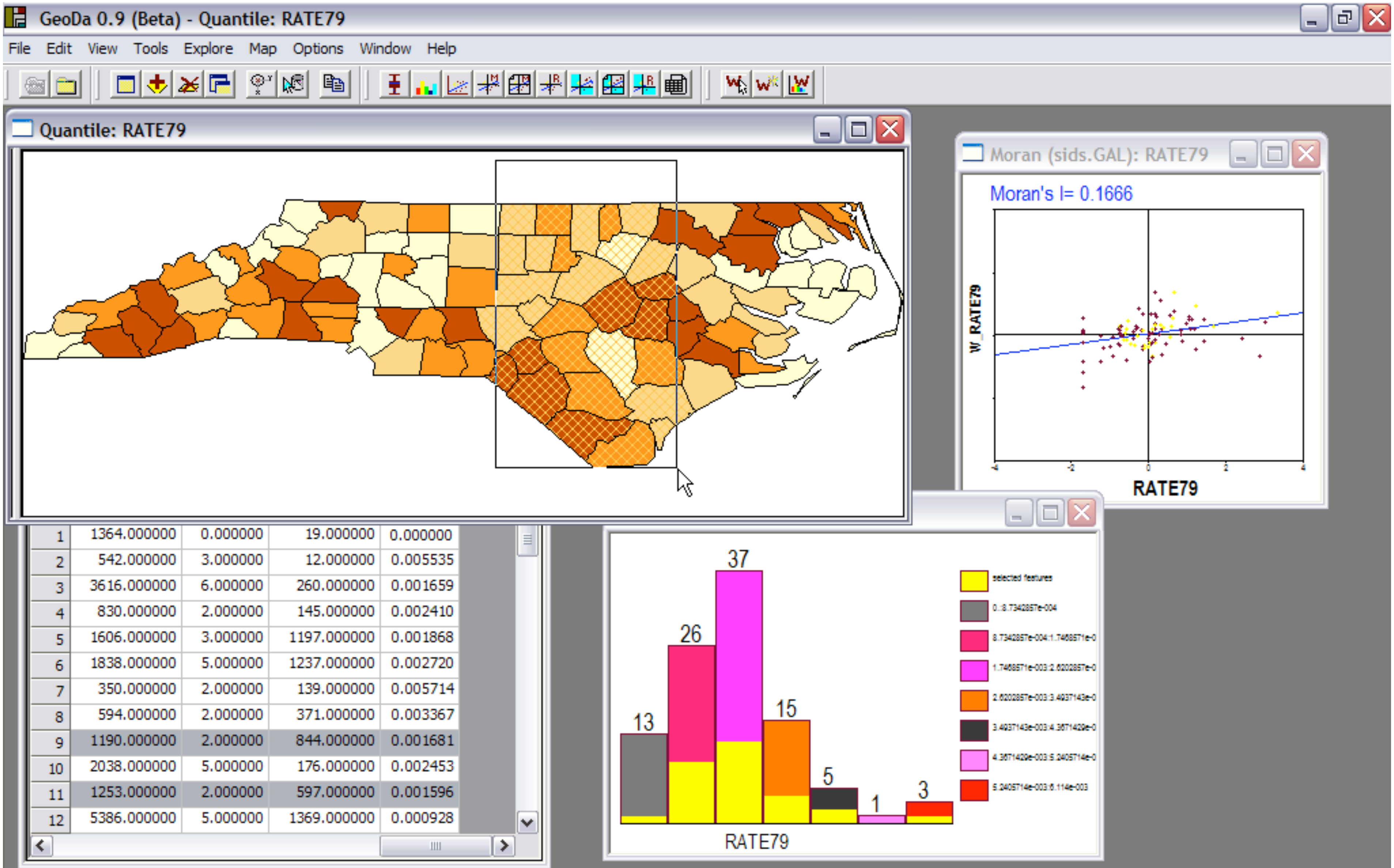

THE KENNEDY GROUP
Serial No. :
RETURN THIS
WAREHOUSE
HB 2


Web Browsers (thin clients)
Applications (thick clients)
Users
metadata search
GNIS
gazetteer
Geospatial One-Stop
Geoportal
metadata registration
WFS
WCS
other services
online data
offline data
Providers

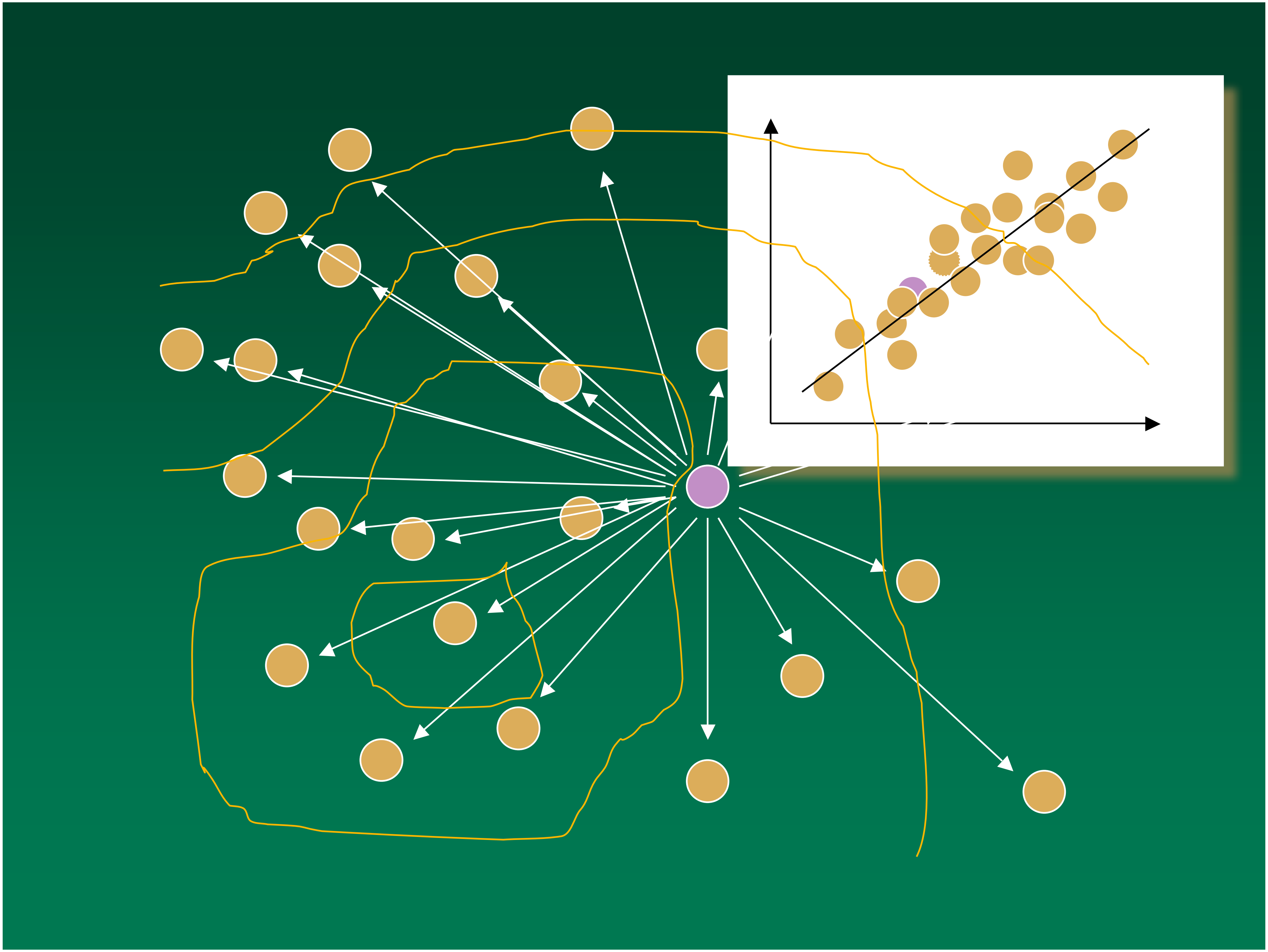

position 5
position 3
Smithe St
Image © 2006 DigitalGlobe
© 2006 Europa Technologies
Image © 2006 TerraMetrics
© 2006 Navteq
©2005 Google


Gotham Book Mart
41 W 47th St, New York, NY 10036 - Map
(212) 719-4448
Walk
Left
Street
Road
Hybrid
Pick your view:
Race Car
Start over
Feedback
About
Moore Theatre
2nd Ave
1st Ave
30 yds
© 2005 Microsoft Corp © 2004 NAVTEQ
Image courtesy of USGS


GeoDa 0.9.5-i (Beta) - BoxMap (Hinge=1.5) : APR99PC
File View Edit Tools Table Map Explore Space Regress Options Window Help
BoxMap (Hinge=1.5) : APR99PC
Lower outlier (0)
< 25% (52)
25% - 50% (52)
50% - 75% (52)
> 75% (27)
Upper outlier (26)
Scatter Plot: TURN99PC vs APR99PC
Slope = -0.2661
0.2459
APR99PC
TURN99PC
0.6
0.7
0.8
0.9
Histogram: TURN99PC
4
8
15
29
61
67
25
selected features
0.65922 : 0.68902
0.68902 : 0.71883
0.71883 : 0.74863
0.74863 : 0.77843
0.77843 : 0.80824
0.80824 : 0.83804
0.83804 : 0.86785
(1) LISA Cluster Map (ba_queen.GAL): I_TURN99PC
Not Significant
High-High
Low-Low
Low-High
High-Low
Ready
Address
http://sal.agecon.uiuc.edu
Go
Links



www.csiss.org



Characteristics of geographic data

- Are there general properties?
  - perhaps with the status of laws
  - though not deterministic
- What problems do they present for the application of statistical methods?

Tobler’s First Law

- “All things are related, but nearby things are more related than distant things”
  - W.R. Tobler, 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234-240
  - implies process as much as form
  - “nearby things are more similar than distant things”

Validity

- “Nearby things are less similar than distant things”
  - negative spatial autocorrelation
  - possible at certain scales
    - the checkerboard
    - retailing
  - but negative a/c at one scale requires positive a/c at other scales
  - smoothing processes dominate sharpening processes

Formalization

- Geostatistics
  - variogram, covariogram
  - measuring how similarity decreases with distance
  - parameters vary by phenomenon
    - does this make TFL less of a law?

Utility

- Representation
  - GI is reducible to statements of the form **<x,z>**
  - the atomic form of GI is unmanageable, encountered only in point samples
  - all other GI data models assume TFL
- Spatial interpolation
  - all methods implement TFL

If TFL weren’t true

- GIS would be impossible
  - a point sample is useful only with interpolation
- Life would be impossible

Statistical implications

- Independence is difficult to achieve
  - space observations beyond the phenomenon's range
  - cull observations
- Model spatial dependence explicitly
  - spatial lag models
  - replace $y = f(\mathbf{x})$ with $y = f(\mathbf{Wx})$
    - where the elements of **W** measure proximity

Expanding the horizons

- Other spaces
  - are there spaces for which TFL is not true?
  - digits of π
  - genome
- Other laws of GIScience

Candidate laws

- All important places are at the corners of four map sheets
- Montello and Fabrikant, “The First Law of Cognitive Geography”
  - “People think closer things are more similar”

A second (first) law

- TFL describes a second-order effect
  - properties of places taken two at a time
  - a law of spatial dependence
  - is there a law of places taken one at a time?
- Spatial heterogeneity
  - non-stationarity
  - uncontrolled variance

Corollaries of the second law

- There is no average place on the Earth’s surface
- Sampling is problematic
  - one must visit or map all of it to understand its full complexity
- Results depend explicitly on the bounds of the study
- The Noah effect
  - there is a finite probability of an event of any magnitude
  - to observe an event of a given magnitude it is simply necessary to wait long enough

A GIScientist's Noah effect

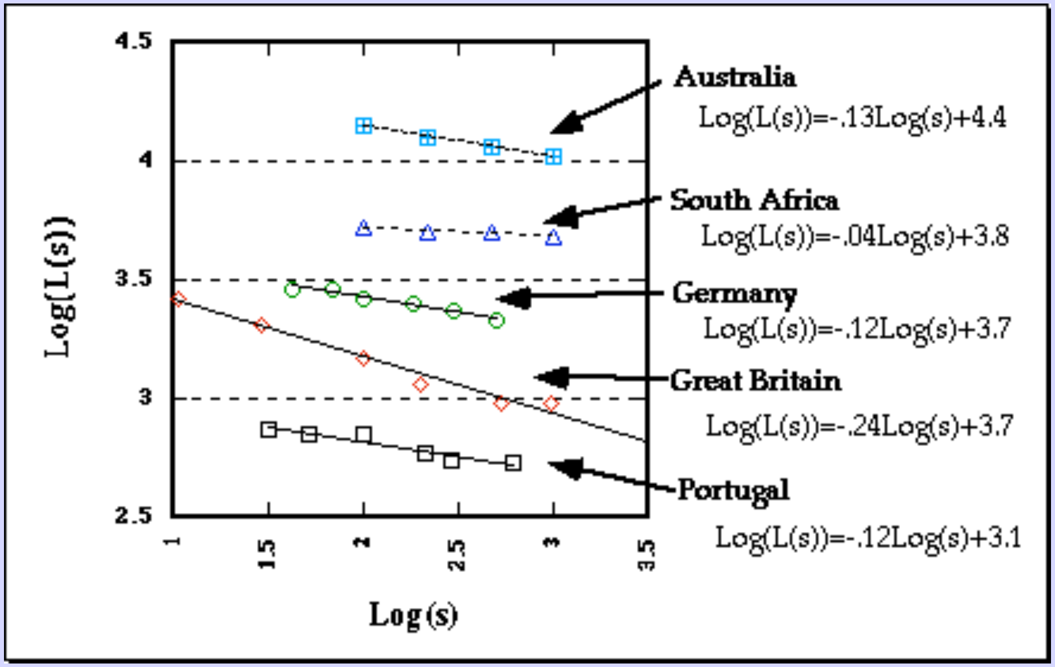
- The Eden effect
  - El Dorado
  - to find a feature of any magnitude it is sufficient to look far enough
    - but unlike time, the Earth's surface is limited

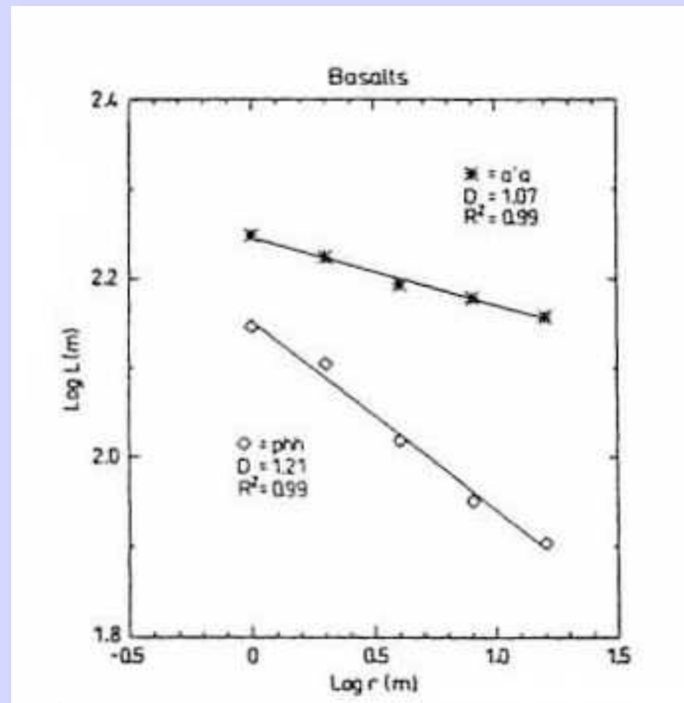
Practical implications of the second law

- A state is not a sample of the nation
  - a country is not a sample of the world
- Classification schemes will differ when devised by local jurisdictions
- Figures of the Earth will differ when devised by local surveying agencies
- Global standards will always compete with local standards

A fractal principle

- The closer you look the more you see
  - and for many natural phenomena the rate is orderly
  - Richardson plots
  - lengths of national boundaries
    - Spain and Portugal
    - context of 1920s


Log(L(s))
Log(s)
Australia
Log(L(s))=-.13Log(s)+4.4
South Africa
Log(L(s))=-.04Log(s)+3.8
Germany
Log(L(s))=-.12Log(s)+3.7
Great Britain
Log(L(s))=-.24Log(s)+3.7
Portugal
Log(L(s))=-.12Log(s)+3.1


Basalts
Log L (m)
Log r (m)
* = a'a
D = 1.07
R² = 0.99
◇ = phh
D = 1.21
R² = 0.99

Practical implications

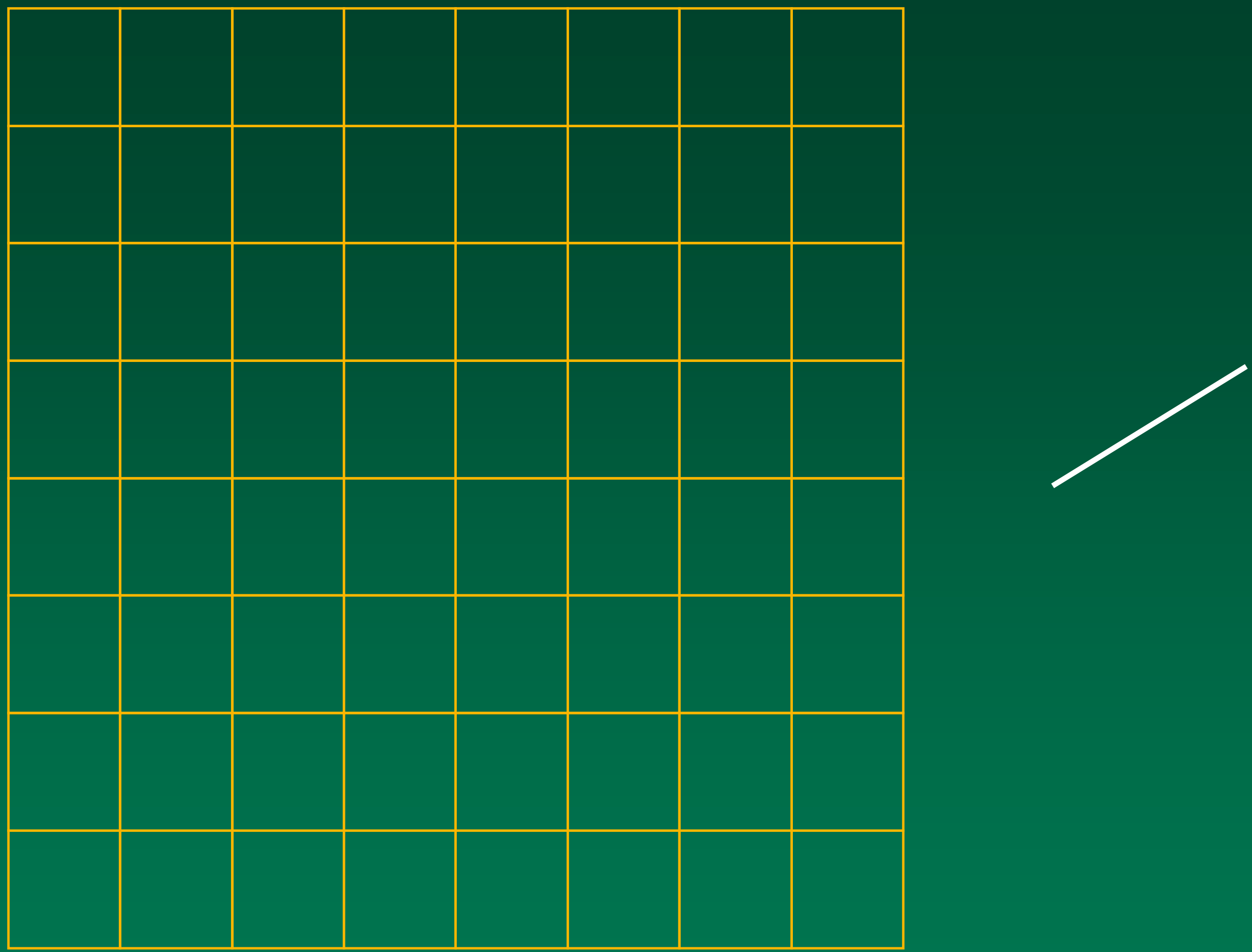
- Indexing schemes, quadtrees
  - partitioning of information at different scales
- Length is a function of spatial resolution
  - and variously under-estimated in GIS
  - as are many other properties
    - slope
    - soil class
    - land cover class
  - spatial resolution should always be explicit in GIS analysis
    - easy in raster
    - much more difficult in vector

Buffon’s needle

- Consider a needle of unit length
  - dropped randomly onto a set of parallel lines unit distance apart
  - probability that the needle will intersect a line?

Analytical results

- p(intersection) = $2/\pi$ = 0.6366
- Experimental determination of π
  - 5^{th} decimal place
  - $\sqrt{npq}/np = 10^{-5}$
  - $n \sim 10^{10}$
- lines s units apart, needle length l
  - $p = 2l / \pi s$
- relevance to GIScience?



$l \geq s\sqrt{2}$ E(number of cells intersected) = $4l / \pi s$

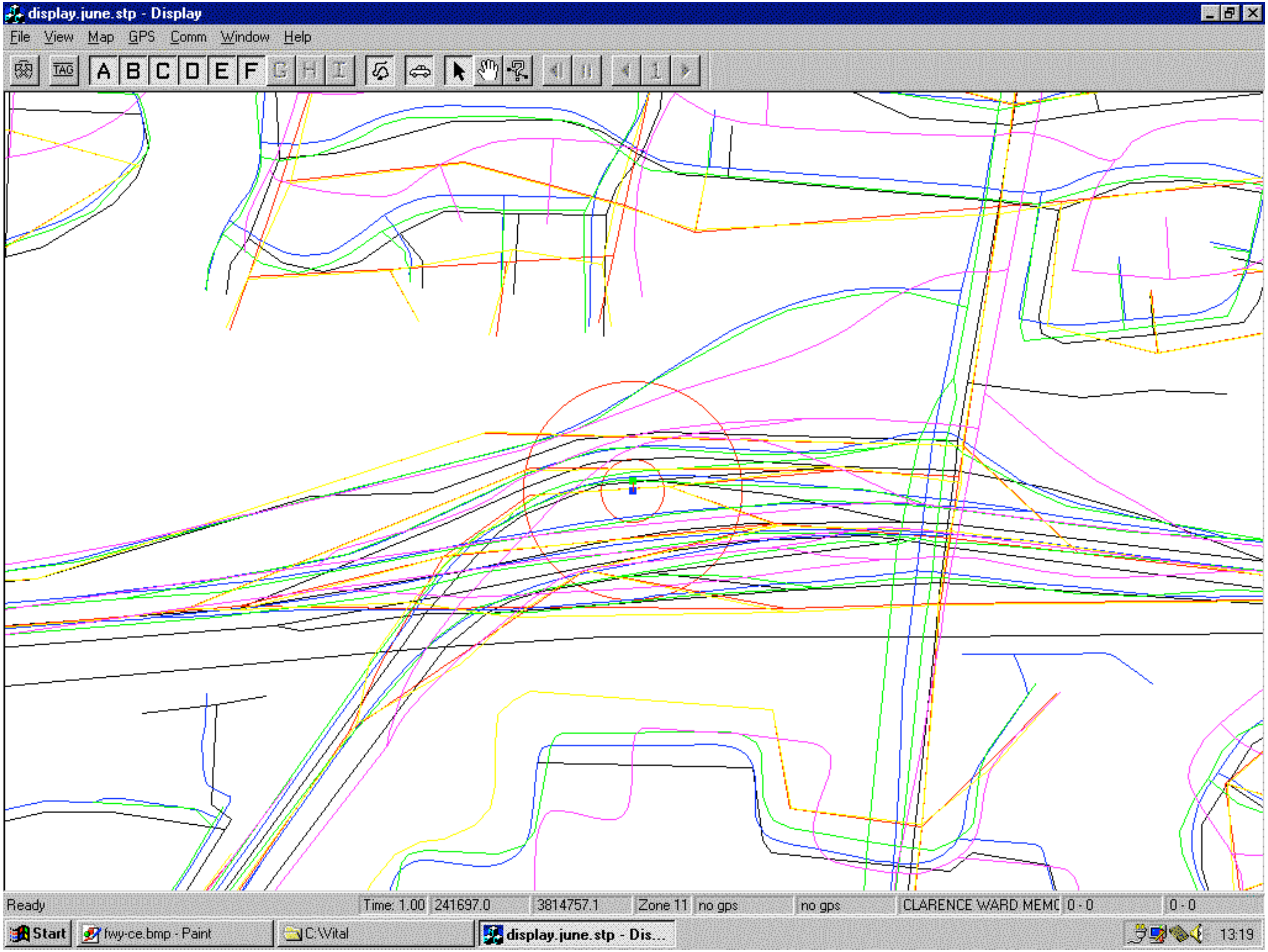
$l < s$ p(ends in different cells) = $(4ls - l^2) / \pi s^2$

Applications

- Short needle
  - quadrat-based experiments
    - avoid missing point-to-point interactions when pairs are in different cells
  - databases partitioned into tiles
    - avoid having to access multiple tiles when e.g. computing distance
- Long needle
  - operations on raster databases, e.g. intervisibility
    - depend on number of intersected cells

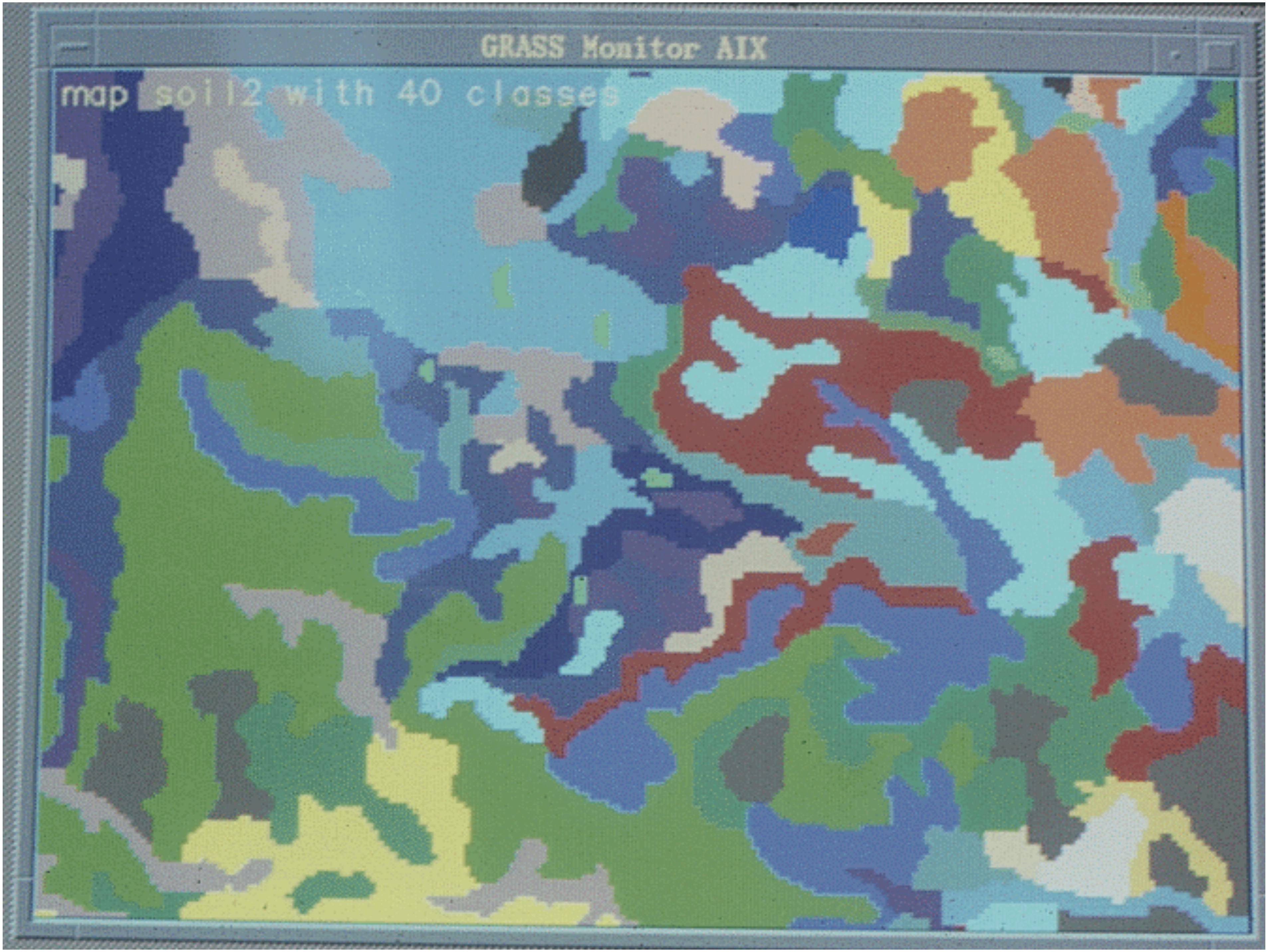
Uncertainty in spatial data

- All spatial data leave the user uncertain to some degree about the exact nature of the real world
  - no representation can be exact
  - all representations involve some combination of approximation, measurement, generalization


display.june.stp - Display
File View Map GPS Comm Window Help
TAG A B C D E F
Ready
Time: 1.00
241697.0
3814757.1
Zone 11
no gps
no gps
CLARENCE WARD MEMC
0 - 0
0 - 0
Start
fwy-ce.bmp - Paint
C:\Vital
display.june.stp - Dis...
13:19

Patterns of error

- Strong positive spatial autocorrelation
  - local shapes preserved
  - relative errors << absolute errors
  - derived properties minimally affected
    - distances
    - directions
  - suggests relative positioning better than absolute
  - measurement-based GIS

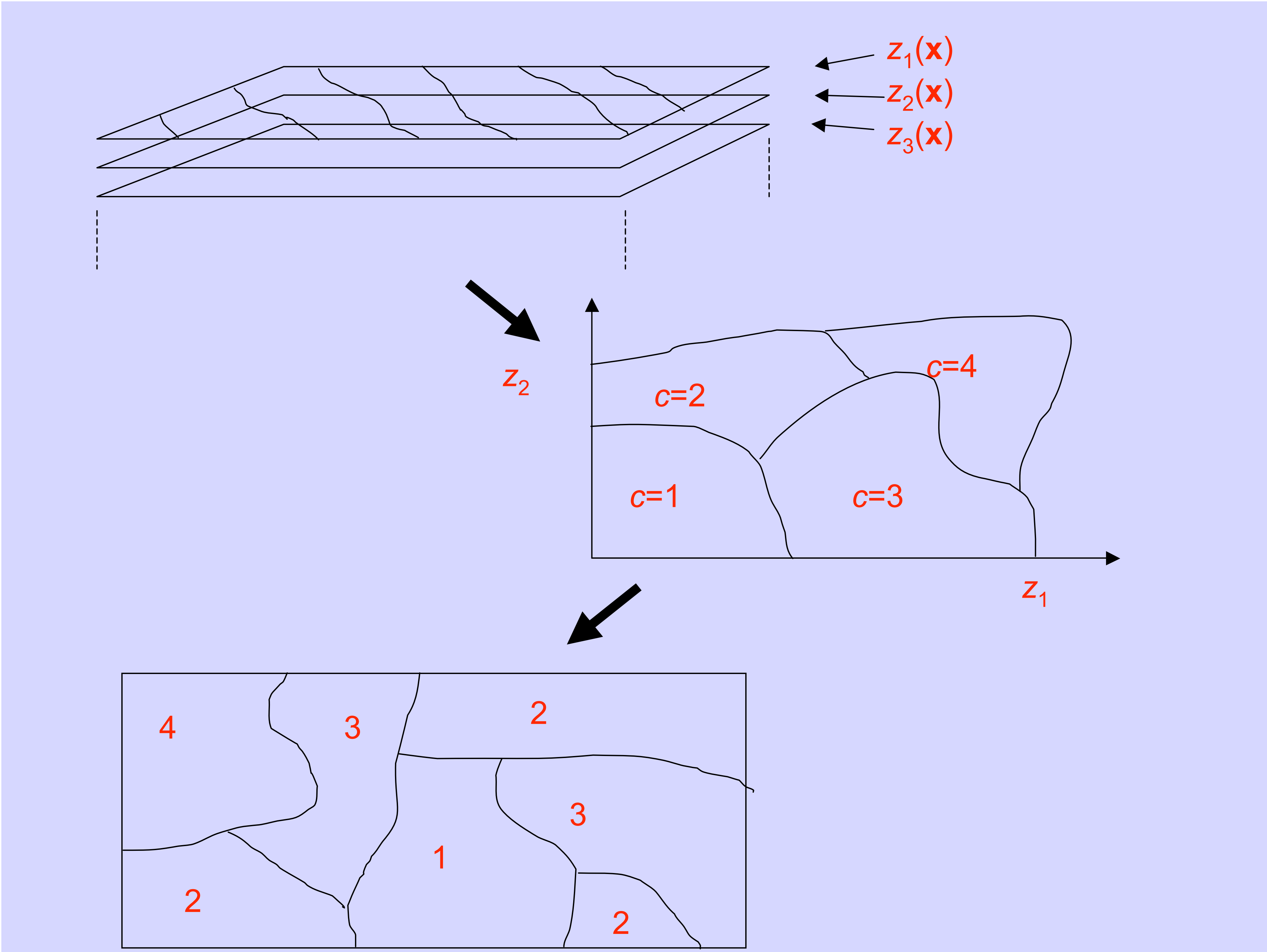

GRASS Monitor AIX
map soil2 with 40 classes

Uncertainty modeling

- Area-class maps are made by a long and complex process involving many stages, some partially subjective
- Maps of the same theme for the same area will not be the same
  - level of detail, generalization
  - vague definitions of classes
  - variation among observers
  - measuring instrument error
  - different classifiers, training sites
  - different sensors

Phase-space model

- *m* dimensional "phase" space defined by field variables
  - partition into *n* regions
- Generate *m* random fields to locate **x** in phase space
- Assign **x** to one of *n* classes



$z_1(\mathbf{x})$
$z_2(\mathbf{x})$
$z_3(\mathbf{x})$
z_2
$c=2$
$c=4$
$c=1$
$c=3$
z_1
4
3
2
3
1
2
2

Properties of the model

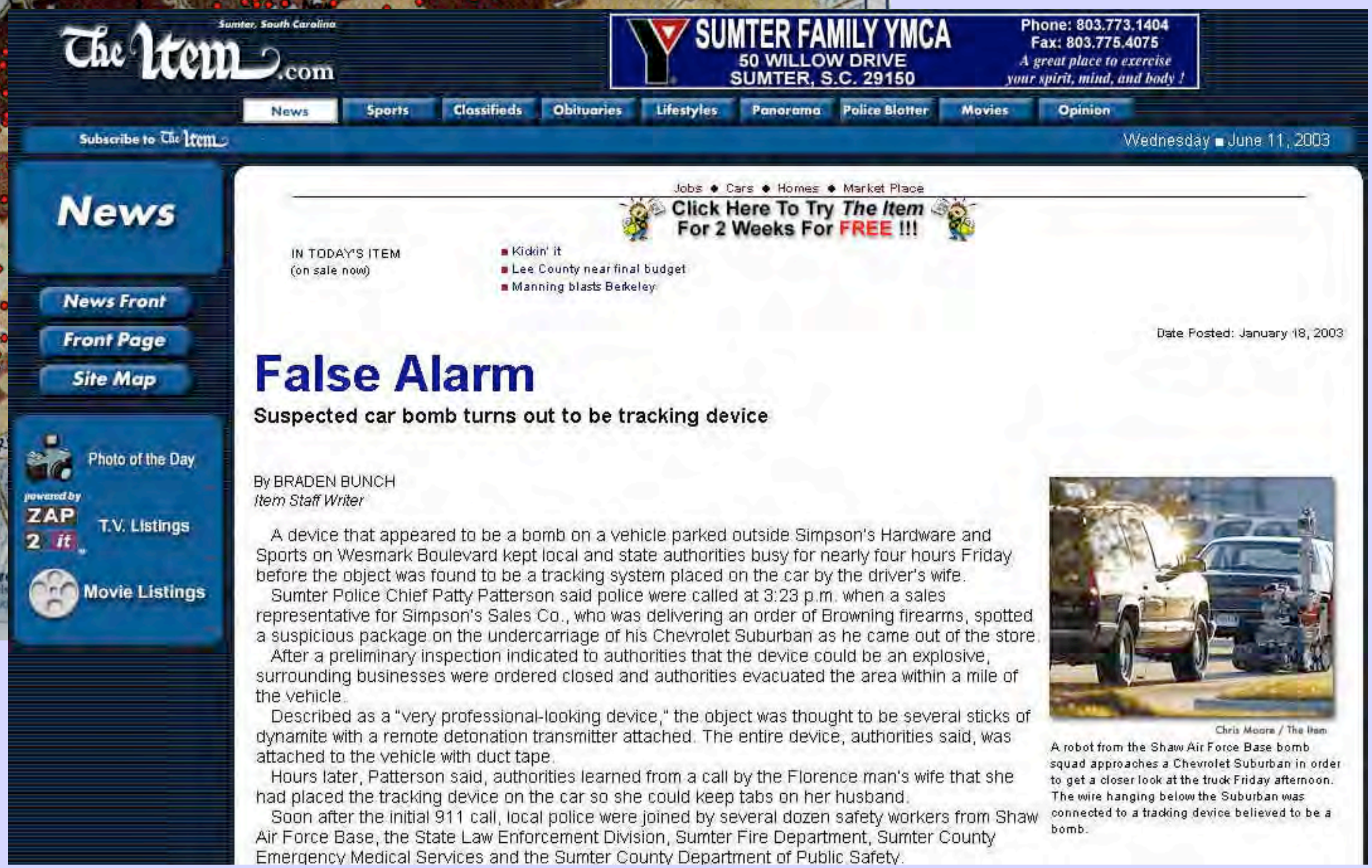
- Appearance matches
  - real map is a possible realization
- Variation in
  - positions of boundaries
  - classes assigned to areas
  - numbers of boundaries, areas
  - homogeneity of areas
- Fits well with theory in e.g. ecology
- Model is vastly overspecified

Four challenges

- Maps and dynamics
  - flows, events, transactions, change
  - spatiotemporal analysis
  - analysis of tracks
  - null hypotheses
  - a dynamic TFL
- Software
  - keeping track of methods, scripts
  - *process objects*

Challenges

- New data types
  - primarily commercial
    - geodemographics
    - Web usage
    - email traffic
    - tracking
- Ethical issues
  - privacy
  - security



The Item .com
Sumter, South Carolina

SUMTER FAMILY YMCA
50 WILLOW DRIVE
SUMTER, S.C. 29150
Phone: 803.773.1404
Fax: 803.775.4075
A great place to exercise your spirit, mind, and body !

News | Sports | Classifieds | Obituaries | Lifestyles | Panorama | Police Blotter | Movies | Opinion

Subscribe to The Item

Wednesday ■ June 11, 2003

News

News Front
Front Page
Site Map

Photo of the Day
powered by ZAP 2 it T.V. Listings
Movie Listings

Jobs ◆ Cars ◆ Homes ◆ Market Place

Click Here To Try *The Item* For 2 Weeks For FREE !!!

IN TODAY'S ITEM
(on sale now)

- Kickin' it
- Lee County near final budget
- Manning blasts Berkeley

Date Posted: January 18, 2003

False Alarm

Suspected car bomb turns out to be tracking device

By BRADEN BUNCH
Item Staff Writer

A device that appeared to be a bomb on a vehicle parked outside Simpson's Hardware and Sports on Wesmark Boulevard kept local and state authorities busy for nearly four hours Friday before the object was found to be a tracking system placed on the car by the driver's wife.

Sumter Police Chief Patty Patterson said police were called at 3:23 p.m. when a sales representative for Simpson's Sales Co., who was delivering an order of Browning firearms, spotted a suspicious package on the undercarriage of his Chevrolet Suburban as he came out of the store.

After a preliminary inspection indicated to authorities that the device could be an explosive, surrounding businesses were ordered closed and authorities evacuated the area within a mile of the vehicle.

Described as a "very professional-looking device," the object was thought to be several sticks of dynamite with a remote detonation transmitter attached. The entire device, authorities said, was attached to the vehicle with duct tape.

Hours later, Patterson said, authorities learned from a call by the Florence man's wife that she had placed the tracking device on the car so she could keep tabs on her husband.

Soon after the initial 911 call, local police were joined by several dozen safety workers from Shaw Air Force Base, the State Law Enforcement Division, Sumter Fire Department, Sumter County Emergency Medical Services and the Sumter County Department of Public Safety.

Chris Moore / The Item

A robot from the Shaw Air Force Base bomb squad approaches a Chevrolet Suburban in order to get a closer look at the truck Friday afternoon. The wire hanging below the Suburban was connected to a tracking device believed to be a bomb.

They are always watching you. Use cash. Do not give your phone number, social security number or address. Do not fill in questionnaires. Demand that credit firms remove you from marketing lists. Check your medical records often. Keep your telephone number unlisted. Never leave your mobile phone on. Do not use credit or discount cards. If you must use the Internet, use someone else's computer. Assume that all calls, voice mail, email and computer use are monitored.

The Economist, 1 May 1999