Geography and GIS as Middleware

John Wilson

8 July 2010

Spatial Science

1

Outline

- Changing face of geographic information science
- Recent projects (examples)
  - USC geocoding platform
  - Modeling hydrologic systems in metropolitan regions
  - Characterizing health impacts of environmental exposures
- Final thoughts & questions

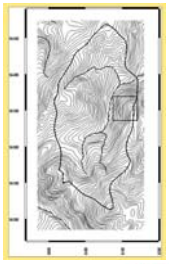
Slide Courtesy of Dean Gusch

Spatial Science

2

Geographic Information Systems

- A series of toolboxes to support geoprocessing
  - To help answer new questions or simply tackle old ones more efficiently
- Specialized (i.e. standalone) hardware and software
- NCGIA Core Curriculum – three new & in many ways separate geography courses

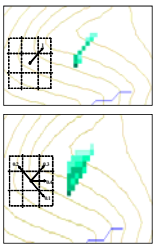


Spatial Science

3

Geographic Information Science

- Geoprocessing is now part of mainstream Web & database platforms
  - Google, Microsoft, Oracles
- Migration from desktop applications to web services
  - ArcGIS 10, ArcServer
- Increasing interest in enterprise solutions
- More open source solutions



Slide Courtesy of David Tarboton

Spatial Science

4

Geographic Information Science (2)

- Much broader & richer set of applications
- Standards & sharing have assumed important roles
  - OGC, spatial data infrastructures, many forms of remote & participatory sensing
- GIS certification & emergence of geospatial job classifications
- Renewed focus on underlying representational & technical issues
  - Ontologies, semantics, new data streams (sources) & analytical methods

Spatial Science

5

The immediate challenge …

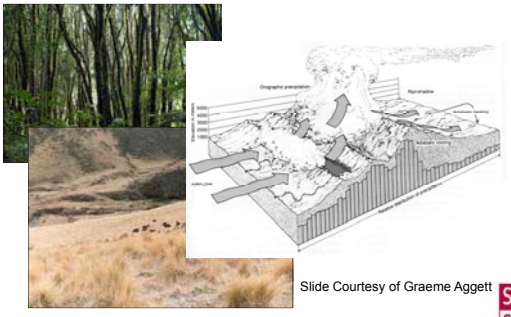
The development of relevant theory and concepts, and the cultivating of spatial intelligence through education, has lagged far behind, however, and it is clear that a wide gap exists between the power and accessibility of tools on the one hand and the ability of researchers, students, and the general public to make effective use of them on the other

Quote Courtesy of USC Geography Department Review Committee

Spatial Science

6

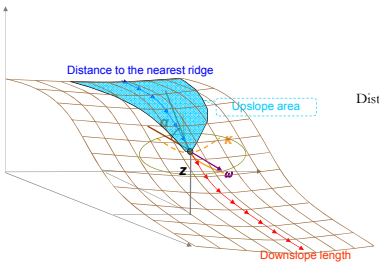
Terrain modeling



Slide Courtesy of Graeme Aggett

Spatial Science

Topographic attributes



Elevation (**z**)
Slope gradient (**α**)
Slope aspect (**ω**)
Curvatures (**κ**)
Distance to the nearest ridge
Downslope length
Upslope area

Slide Courtesy of Bard Romstad

Spatial Science

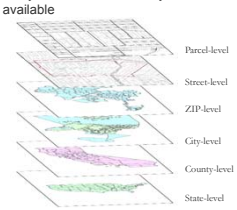
USC Geocoding Platform

- Collaborative work with Daniel Goldberg
- Provides extensible system for identifying and reducing spatial uncertainty & error
- This is a production system …
  - Supports > 3,000 users
  - Has produced >15 million geocodes thus far
- .Net implementation on top of SQL server for reference data
  - TIGER/Lines, Assessor files (parcel data)
- https://webgis.usc.edu/

Spatial Science

Motivation

Many different reference layers available



NAACCR GIS Coordinate Quality Codes

Code	Description
1	GPS
2	Parcel centroid
3	Complete street address
4	Street intersection
5	Mid-point on street segment
6	USPS ZIP5+4 centroid
7	USPS ZIP5+2 centroid
8	Assigned manually
9	USPS ZIP5 centroid
10	USPS ZIP5 centroid of PO Box or RR
11	City centroid
12	County centroid

Hierarchy-based best match criterion

Spatial Science

Best-match candidate selection

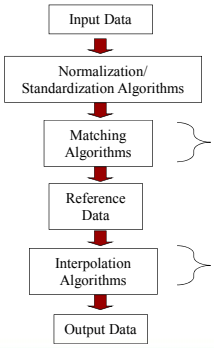
- All reference features of the same class do not have the same accuracy or certainty

90089 ~1:10,000 scale | 90011 ~1:60,000 scale | 90275 ~1:300,000 scale

- Hierarchy-based approaches will not produce the optimal output

Spatial Science

Theoretical framework



3620 South Vermont Avenue

Transform input to match reference data format

3620 S VERMONT AVE

Find a matching geographic feature in reference data

SELECT FromX, FromY, ToX, ToY
FROM SOURCE WHERE
(Start >=3620 AND End <= 3620) AND
(Pre = S) AND
(Name = VERMONT) AND
(Suffix = AVE)

Use matched geographic feature to derive output

Output Point = (20% * △X, 20% * △Y)

Spatial Science

Component: Input data

Input Data → Normalization/ Standardization Algorithms → Matching Algorithms → Reference Data → Interpolation Algorithms → Output Data

Error Contribution

Many different types, forms, and formats:

Street Addresses: 3620 South Vermont Ave
Postal Codes: Los Angeles, CA 90089-0255
Named Places: USC Kaprielian Hall
Intersections: Vermont & 36th Place
Relative Descriptions: b/w Bakersfield & Shafter

Different levels of information/certainty:

Street Addresses: Somewhere on street
Postal Codes: Somewhere on postal route
Named Places: Absolute location
Intersections: Somewhere near intersection
Relative Descriptions: Somewhere near locations

Incompleteness:
3260 S Vermont ___
3620 _ Vermont Ave
____ _ Vermont Ave

Inaccuracy:
3620 S Verment Ave
362_ S Vermont ___
3260 _ Vermont St

Spatial Science

13

Component: Input data cleaning

Input Data → Normalization/ Standardization Algorithms → Matching Algorithms → Reference Data → Interpolation Algorithms → Output Data

Error Contribution

- Parsing – Separating components of the address
  Token-based: relies on formatting

- Normalization – Identifying components of the address
  Substitution-based: relies on the token ordering
  Context-based: relies on position and schema knowledge
  Probability-based: relies on likelihood of occurrence

- Standardization – Formatting components of the address
  Schema mapping: must exist for all reference sources

3620	South	Vermont	Ave	Los Angeles	90089
Street Address				City	Zip

90089		St Los Angeles	St	Los Angeles	90089
Street Address				City	Zip

23	E	South	St	South Los Angeles	90089
Street Address				City	Zip

Spatial Science

14

Component: Matching algorithms

Input Data → Normalization/ Standardization Algorithms → Matching Algorithms → Reference Data → Interpolation Algorithms → Output Data

Error Contribution

- Multiple Match Types – Feature selected from reference set
  Exact: A single perfect match
  Non-exact: A single non-perfect match
  Exact ambiguous: Multiple perfect matches
  Non-exact ambiguous: Multiple non-perfect matches
  None: No matches

- Multiple Matching Methods – Ways of selecting features
  Deterministic: Rule-based, iterative
  Probabilistic: Likelihood-based, attribute weighting

- Multiple Fuzzifying Techniques – Alter input data
  Word Stemming: Porter Stemmer
  Phonetic Algorithms: Soundex
  Attribute Relaxation: Remove attributes and retry match

- Multiple Scoring Methods – compute a candidate score
  Relative attribute weighting
  Match-Unmatch weighting

Spatial Science

15

Component: Reference data

Input Data → Normalization/ Standardization Algorithms → Matching Algorithms → Reference Data → Interpolation Algorithms → Output Data

Error Contribution

- Multiple Data Types
  Point-based: ZCTA and Place Centroids
  Linear-Based: Street Centerlines
  Areal Unit-Based: Parcels, ZCTA and Place Boundaries

- Wide spectrum of accuracies/completeness
  Commercial vs. Public
  - Attribute accuracy – spatial and non-spatial
  - Attribute completeness – spatial and non-spatial
  - Feature complexity – simple vs. polylines
  Local Scale vs. National Scale
  - Census Place Boundaries vs. Local Neighborhoods

- Wide spectrum of cost/availability
  Free vs. Costly: TIGER/Lines vs. TeleAtlas
  Available vs. Not: Address points – CA. vs. N. Carolina

Spatial Science

16

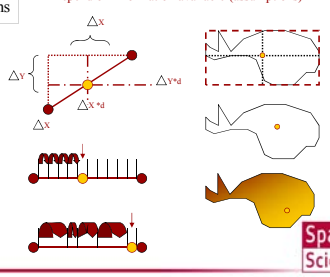
Component: Interpolation algorithms

Input Data → Normalization/ Standardization Algorithms → Matching Algorithms → Reference Data → Interpolation Algorithms → Output Data

Error Contribution

- Many methods of interpolation

Depend on reference feature type
Depend on information available (assumptions)



Spatial Science

17

Component: Interpolation algorithms

Input Data → Normalization/ Standardization Algorithms → Matching Algorithms → Reference Data → Interpolation Algorithms → Output Data

Error Contribution

- Lack of Process Transparency
  - Nothing reported about the decisions made or alternatives

- Output Data Type: Only Geographic Coordinates
  - Lose data required for determining true accuracy

- Output Accuracy: Feature Match Type + Probability
  - Nothing that indicates direction
  - Nothing that indicates distance
  - Nothing that indicates certainty area or surface

Spatial Science

18

Results

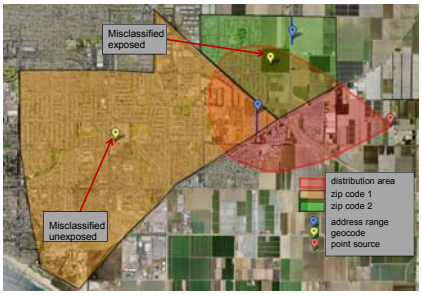
- Our geocoding platform provides an open source extensible approach that offers following benefits
  - Improves match rates using nearby candidates instead of reverting to a lower level of geographic match
  - Reduces spatial error & uncertainty by using an uncertainty-driven approach to pick the most likely location given the candidates available & their spatial / topological relationships
  - Uses intelligent tie breaking strategies to identify most likely outcome by interrogating the region around ambiguous matches & investigating the relationships between their attributes

Spatial Science

19

Significance of this work …

- May end up with exposure misclassification because of erroneous geocoding outputs

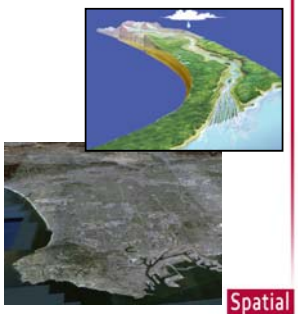


Spatial Science

20

Modeling hydrologic systems

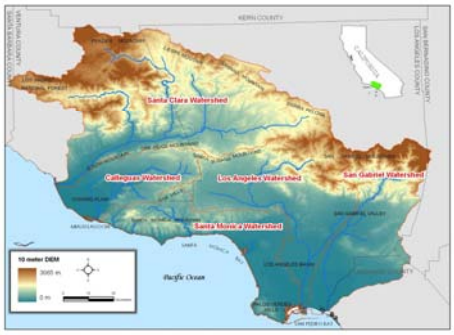
- Determine quantities of contaminants entering stream network
- Simulate transport of contaminants in reservoirs, rivers & groundwater
- Predict water quality by stream catchment



Spatial Science

21

Green Visions Project study area

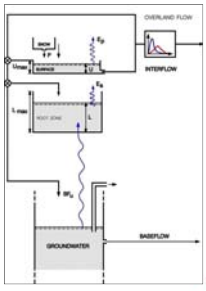


Spatial Science

22

MIKE BASIN

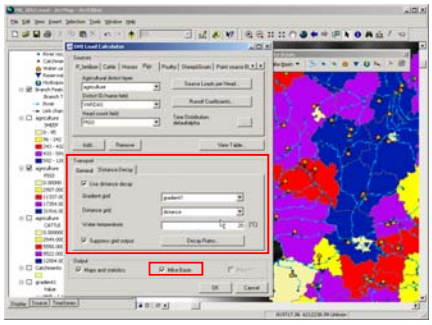
- Simple rainfall-runoff & pollution loading model authored by Danish Hydraulic Institute
- Runs on top of ArcGIS toolbox
- Incorporates easy-to-use model calibration & validation tools



Spatial Science

23

WQ module user interface



Spatial Science

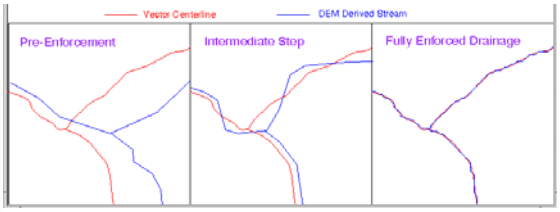
24

National Hydrography Dataset

- Stream address system for linking water-related information to national drainage network
- Supports upstream / downstream modeling along drainage network
- Provides rich cartographic feature content for making maps

Spatial Science

Drainage enforcement



Slide Courtesy of Pete Steeves

Spatial Science

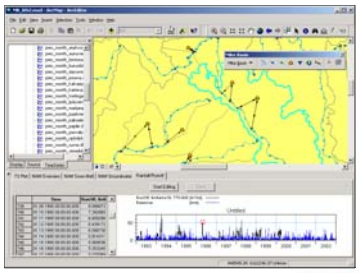
Model discretization

- Drains 2,200 km^2
- 1,783 unique stream segments (links) in NHD Plus
- 171 tributaries and sub-catchments used for MIKE BASIN model runs
- 11.73 km^2 (1,173 ha) minimum map unit

Spatial Science

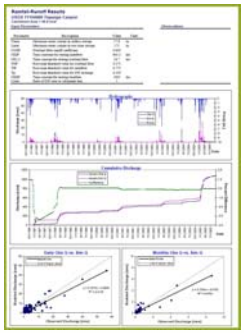
Rainfall-runoff model analysis

- Key inputs
  - NHD Plus
  - Rainfall, ET & temperature time series
  - Stream flow data for model calibration & validation



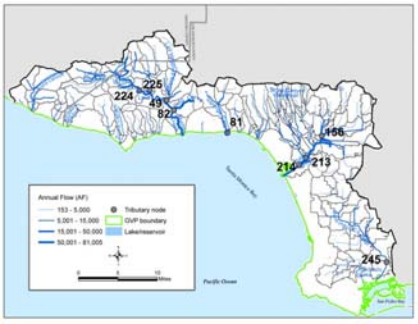
Spatial Science

Rainfall-runoff model results



Spatial Science

Santa Monica Bay flows



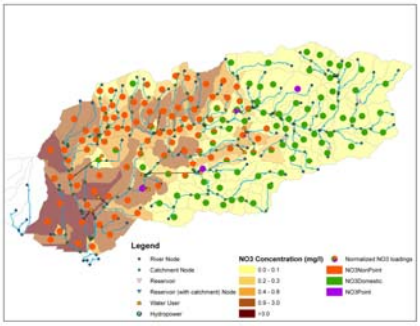
Spatial Science

Contaminant sources

- Point sources
  - Five major NPDES polluters
- Domestic sources
  - Population data
  - Sewage treatment plants
  - Distance decay calibration factor
- Fertilizer sources
  - Crop data
  - Fertilizer application rates
  - Distance decay calibration factor
- Livestock sources

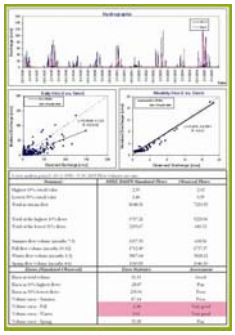
Spatial Science

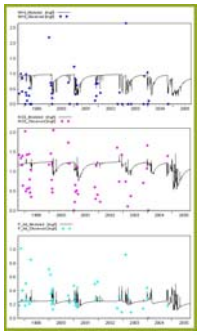
MIKE BASIN NO_3-N predictions



Spatial Science

Water quality modeling results





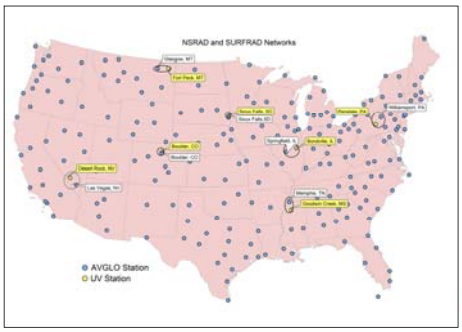
Spatial Science

Incidence of melanoma

- One of most rapidly increasing cancers among white population in U.S.
- Studies consistently point to UV exposure as most important risk factor
- Individual sun exposure is difficult to quantify
- Collaborative work with Myles Cockburn & Zaria Tatalovich
  - How well can we model spatial variations in UV radiation given measurement network & interpolation techniques available (in 2005)?

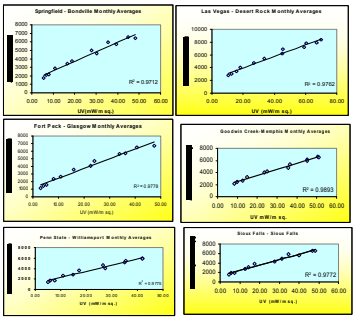
Spatial Science

Measurement network



Spatial Science

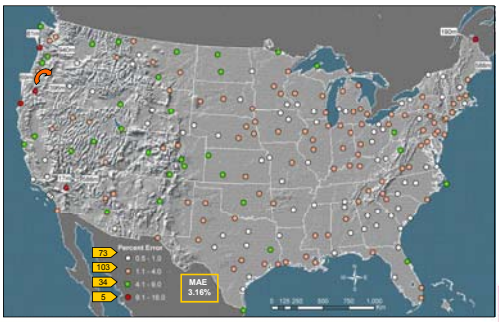
Radiation data correlations



Springfield – Bondville, IL
Las Vegas – Desert Rock, NV
Boulder – Boulder, CO
Fort Peck – Glasgow, MT
Goldwin Creek – Memphis, TN
Penn State – Williamsport, PA
Sioux Falls – Sioux Falls, SD

Spatial Science

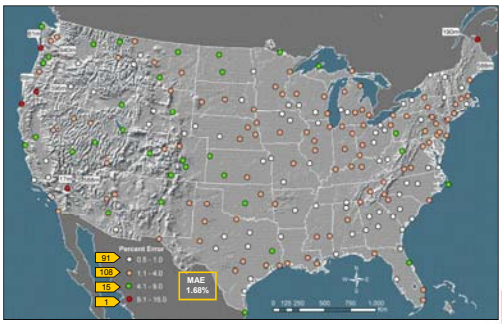
Thiessen polygon error map



Spatial Science

37

Kriged error map



Spatial Science

38

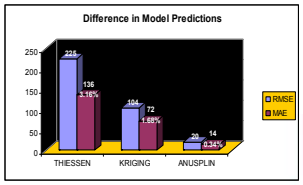
Spline error map

Spatial Science

39

Model differences

- Spline predictions 11 & 2.2 times better than Thiessen polygon & kriged predictions
- ANUSPLIN …
  - Smallest RMSE, MAE & VE
  - Highest R (observed / predicted)
  - Smallest R (observed / error)



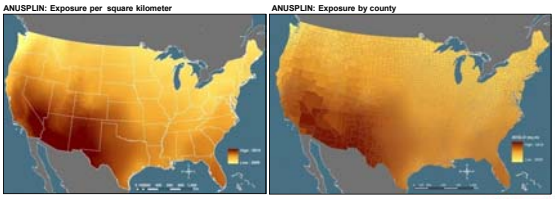
	RMSE	MAE	VE	Correlation OBS/PRED	Correlation OBS/ERR
THIESSEN	225	136 (3.16%)	50906	0.90	0.40
KRIGING	104	72 (1.68%)	10885	0.98	0.31
ANUSPLIN	**20**	**14 (0.34%)**	**384**	**0.99**	**0.15**

Spatial Science

40

ANSPLIN UV exposure model

- Used Arc/Info GRID tools to calculate zonal means
- Generated map of UV exposures by county



Spatial Science

41

Cognitive decline in older women

- Collaborative work with Dan Goldberg plus JC Chen & Rhonda Spencer, USC Keck School of Medicine
- National study tracing health of 7,500 women spread across 22 states
- Subjects undergo cognitive function testing annually (40 clinical centers)
- We plan to calculate mean solar light exposures for 0.5, 1, 3, 6 & 12 month intervals preceding each annual test

Spatial Science

42

Geography & GIS

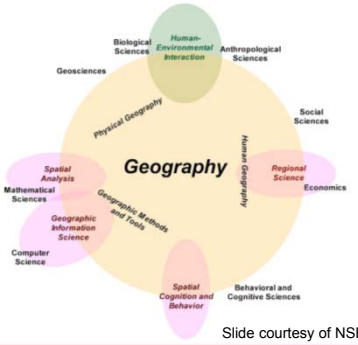
- Key intellectual contributions …
  - Ability to look at phenomena through a spatial lens
  - Ability to integrate knowledge from both social & physical realms
- Both rely on social-scientific basis of geography …
  - Quantitative revolution
  - Shifting theoretical concerns (feminism, post-modernism, globalization)
  - Revolution in geospatial technologies & global environmental crisis



Spatial Science

43

A geography-centric view …



Slide courtesy of NSF (2008)

Spatial Science

44

Final thoughts

- Rise of GIS has proceeded along two axes …
  - As a *science* concerned with principles & concepts associated with efficient acquisition, management & utilization of digital geospatial data (often in collaboration with computer scientists)
  - As suites of *analytical tools* that have been developed and applied in specific knowledge domains (public health and resource management)
  - Geography & the spatial turn …

Spatial Science

45

Selected References

- Goldberg D W, J P Wilson, C A Knoblock, B Ritz, and M G Cockburn (2008) An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics* 7: 60
- Goldberg D W, C A Knoblock, and J P Wilson (2007) From text to geographic coordinates: The current state of geocoding. *Journal of the Urban and Regional Information Systems Association* 19(1): 33-46
- Sheng J, J P Wilson, N Chen, J M Sayre, and J S Devinny (2007) Evaluating the quality of the National Hydrography Dataset for watershed assessments in metropolitan regions. *GIScience & Remote Sensing* 44(3): 1-22
- Ghaemi P, Swift J N, Sister C E, Wilson J P, and J Wolch (2009) Design and implementation of a Web-based platform to support interactive environmental planning. *Computers, Environment and Urban Systems* 33: 482-491
- Tatalovich Z, J P Wilson, and M Cockburn (2006) A comparison of Thiessen polygon, kriging, and ANUSPLIN models of potential UV exposure *Cartography and Geographic Information Science* 33(3): 217-237
- Tatalovich Z, J P Wilson, T Mack, Y Ying, and M Cockburn (2006) The objective assessment of lifetime cumulative ultraviolet exposure for determining melanoma risk. *Journal of Photochemistry and Photobiology B: Biology* 85(3): 198-204

Spatial Science

46