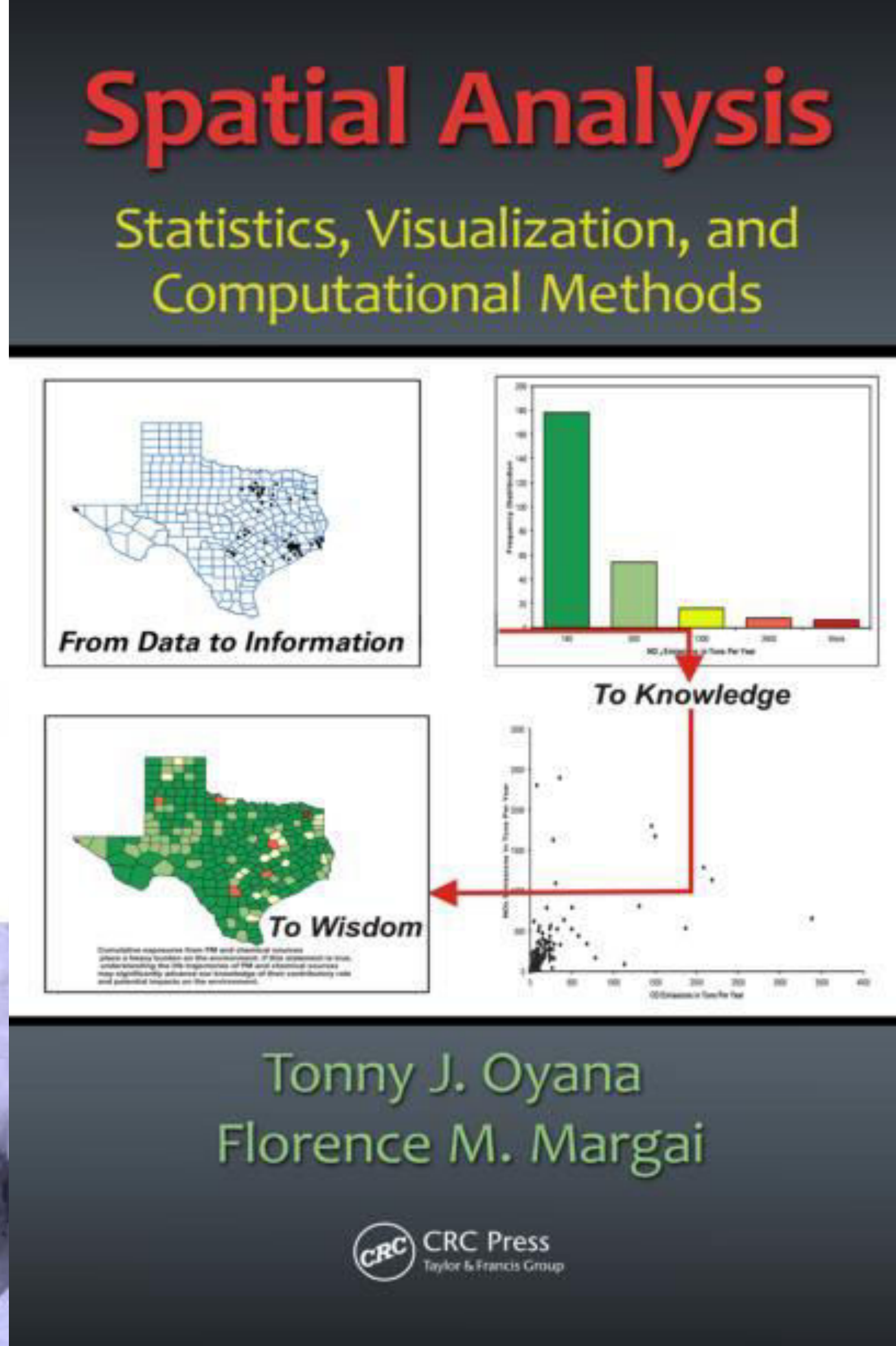
Optimising biodiversity data science for societal benefits in developing countries.

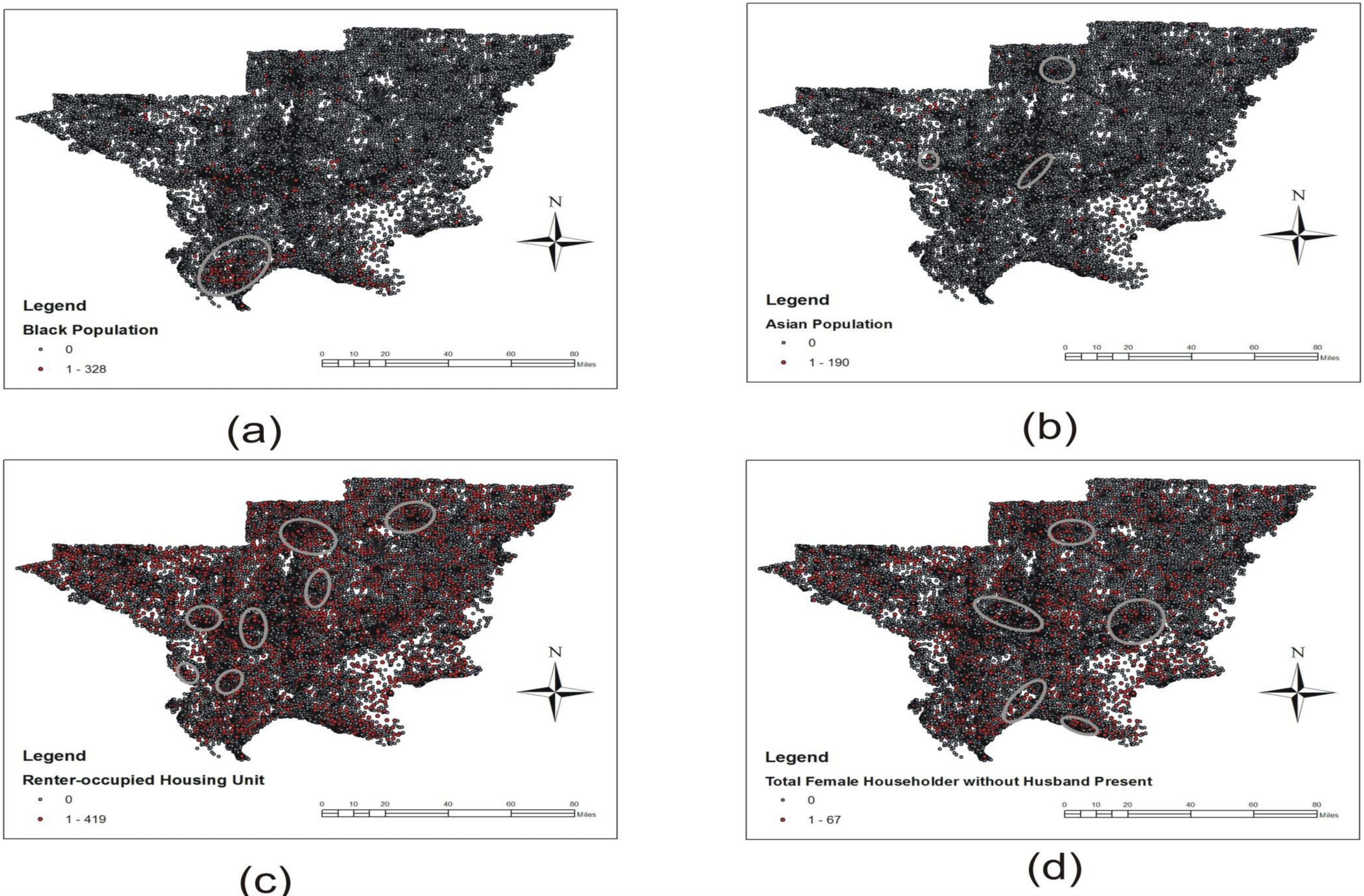
It is the data science revolution: So what are the opportunities and challenges?

Tonny J. Oyana, PhD

Principal & Professor, CoCIS, Makerere University

Why is this currently a very big deal?


Spatial Analysis
Statistics, Visualization, and Computational Methods
From Data to Information
To Knowledge
To Wisdom
Tonny J. Oyana
Florence M. Margai
CRC Press
Taylor & Francis Group



Four maps (a through d) illustrating census demographics from Southern Illinois for twenty counties consisting of 34,218 census blocks. Clusters were delineated for locations where there were higher-than-expected values of the measured variables

Contents

- Review of Key Concepts
  - What is biodiversity?
  - What is data science?
  - Then what is biodiversity data science
- How can it be optimise for societal benefits, especially for development? Principally, through
  - Education and training
  - Requires Significant Investment
  - Strategic use and applications

Contents

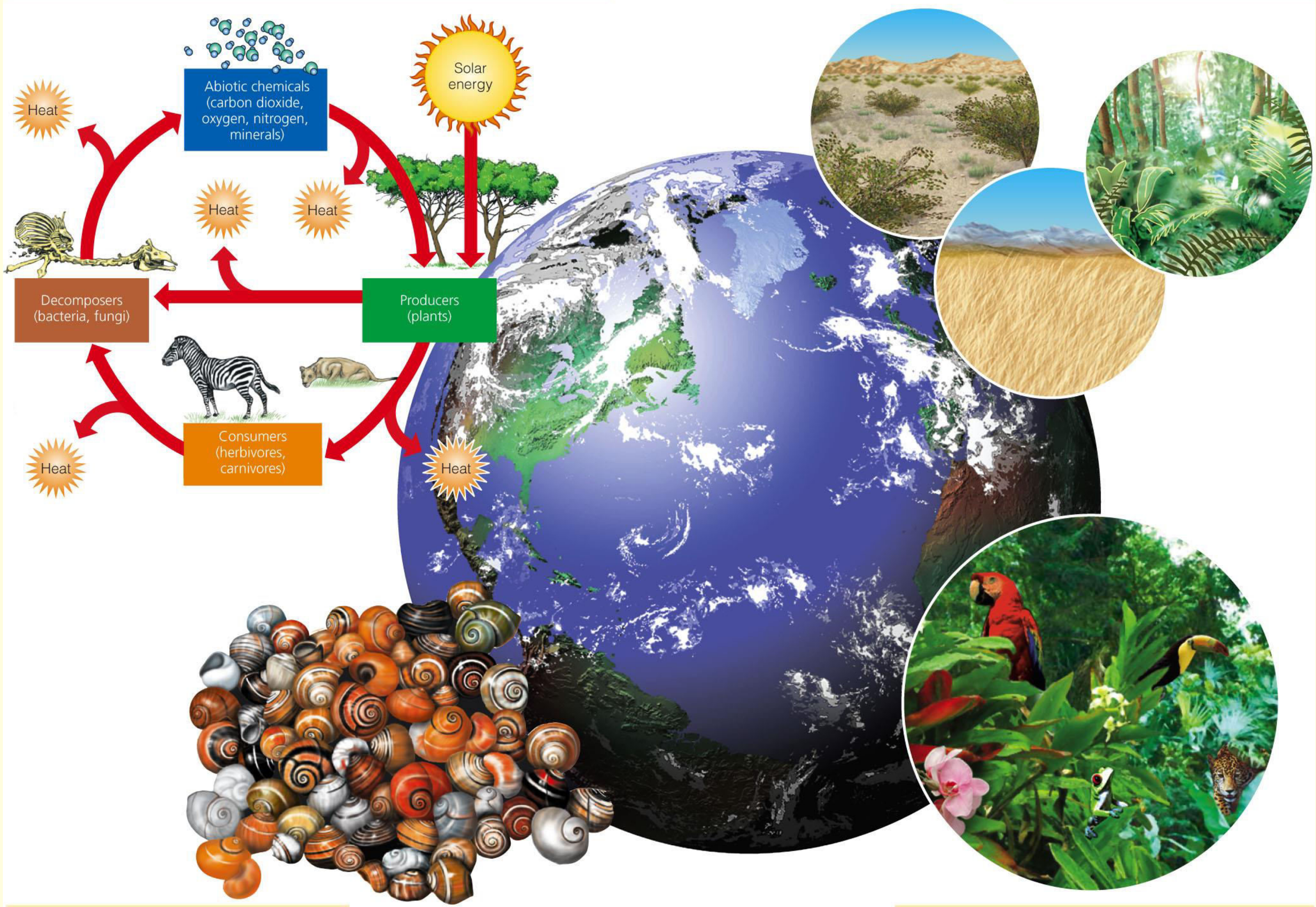
- Concepts in Data Analytics and Strategies
- 3 Example Applications:
  - Study I: Ensuring high quality and protection of Individual-level Geocoded Data
    - Geomasking Optimized Under Space-time and Exposure Constraints (GOUSTEC)
  - Study II: Using an External Exposome Framework to Study Life Course Exposure
  - Study III: Understanding the food environment in an Urban Setting
- Concluding Remarks

What is biodiversity?

- The father of biodiversity Edward O. Wilson (an eminent entomologist) first coined this term in 1986.
- Biodiversity is the variety of life on Earth and the essential interdependence of all living things
- Diversity is a vast concept refers to the range of variations or differences among some set of entities; biological diversity thus refers to varieties within the living world.
- There are 3 components of biodiversity
  - Diversity of genes (sample size and high dimensions)
  - Diversity of number of species (large sample size & HD)
  - Variety of ecosystems

Functional Diversity
The biological and chemical processes such as energy flow and matter recycling needed for the survival of species, communities, and ecosystems.

Ecological Diversity
The variety of terrestrial and aquatic ecosystems found in an area or on the earth.



Genetic Diversity
The variety of genetic material within a species or a population.

Species Diversity
The number and abundance of species present in different communities.

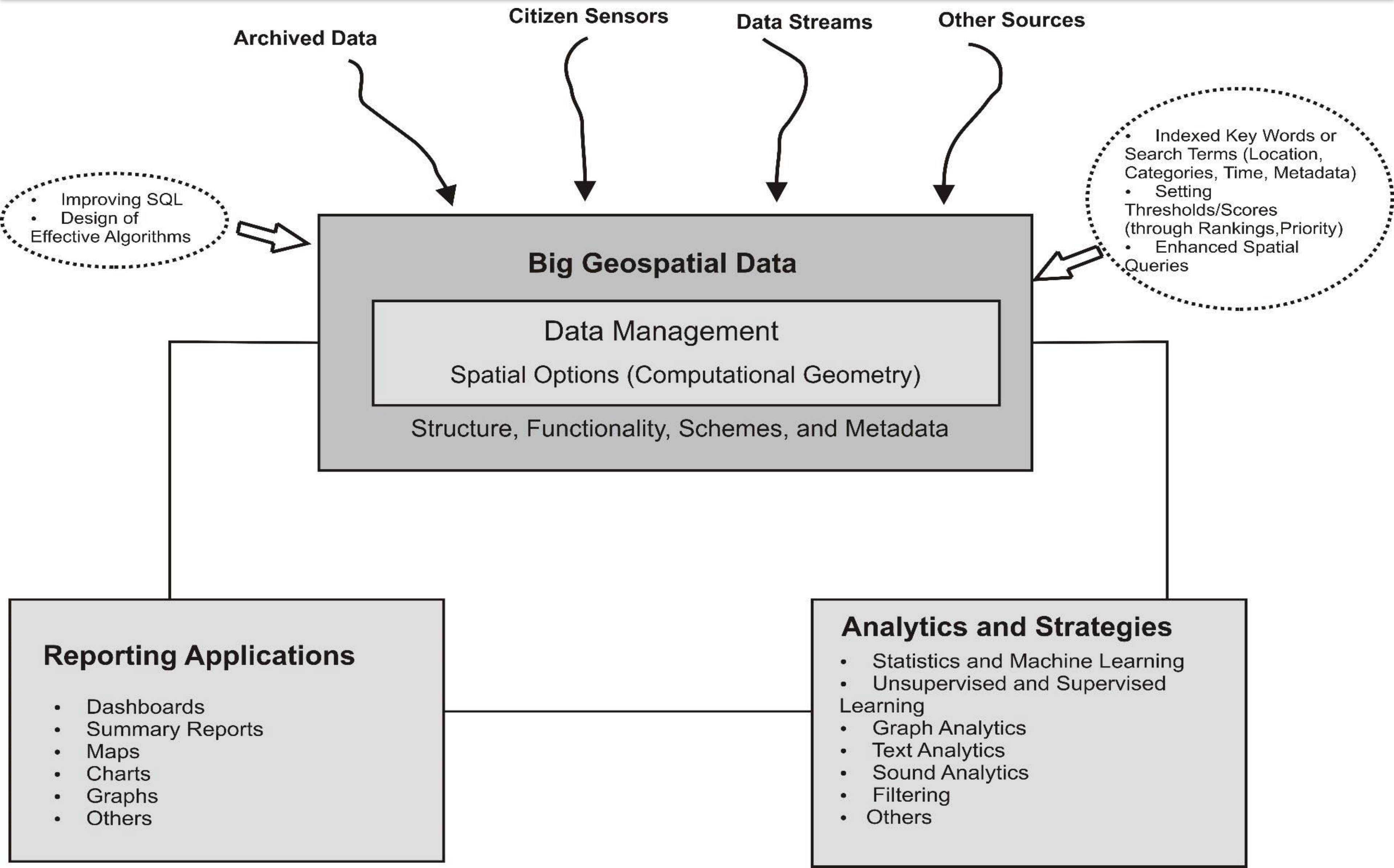
Fig. 4-2, p. 61

What is data science?

- **Data science is most recent field that has evolved with the concurrent growth of large-scale datasets and emerging technologies to handle the volume and variety of information from multiple sources and formats.**
- **There are 3 components of data science**
  - **Data management**
  - **Analytics and Strategies**
  - **Communication of the results/reporting appplications**

data
experience
business
analysis
analytics
work
team
tools
statistical
skills
technical
research
develop
new
modeling
statistics
strong
knowledge
algorithms
machine
must
processing
language
programming
recommendations
production
field
java
customers
degree
customer
distributed
familiarity
analytical
development
database
insights
preferred
products
platforms
understanding
well
using
years
test
big
and/or
communicate
e.g
implementation
quality
product
working
technologies
plus
least
hadoop
python
sets
proven
mining
use
techniques
sql
systems
building
communication
candidate
datasets
management
solutions
software
support
complex
engineering
company
computer
greenplum
design
environment
learning
including
requirements
deliver
models
problems
projects
science
help
improve
languages
large
marketing
prototypes
related

What is Data Science?



Concepts & Example Applications

- **Data** = facts, figure, and statistics
- **Knowledge** = facts, information, and skills
- **Strategies** = a plan of action/overall aim/design to achieve
- **Tools** = a device/implement/perform a specified function
- **Methods** = form of procedure for accomplishing/ established approach/systematic [order, structure, form, system, logic, design]
- **Our motivation** = making sense of data/exploring all angles/uncover the underlying data structure
- **Success** = know/understand the scientific approach

Uncover data secrets/unleash the power of data

- **Data Science** = #1 Data Management (DM)… Follow Best Practices, Standards, & Principles, but okay to break new ground

- **Data Science** = #2 Analytics & Strategies (AS)…"The key to understanding what the data says is to attack it from all angles"

- **Data Science** = #3 Communication Strategies & Reporting Applications (CSRA)

Components of Big Data wrt Geospatial Data

- Extends beyond very large size definition to include:
  - H-Volume = #1 DM "Constantly increasing in quantity"
  - H-Variety = #1 DM "Text, image, sound, video…structured, semi-structured, & unstructured…requires data fusion/put in data lake"
  - H-Velocity = #1 DM "Speed & growth in real time, processing of data streams"
  - Veracity = #2 AS "Quality of data…QAQC"
  - Value = #2 and 3 AS/CSRA "Potential value is huge but contingent upon the AS/CSRA success"

Platform for #DM and #AS

Computational Resources for Handling Big Geospatial Data

Main Types of Computing Platforms

Cluster Computing: Computers are linked through a fast local area network and function as a single unit.

Cloud Computing: Computers are linked together through the Internet to provide a shared pool of computing resources for accessing and storing data and programs.

Grid Computing: A loosely coupled network of computers from multiple locations that work together on common computing tasks.

Heterogenous Computing: Specialized computing system that use more than one kind of processor, for example central processing units and graphics processing units.

A List of Currently Available Software Kits

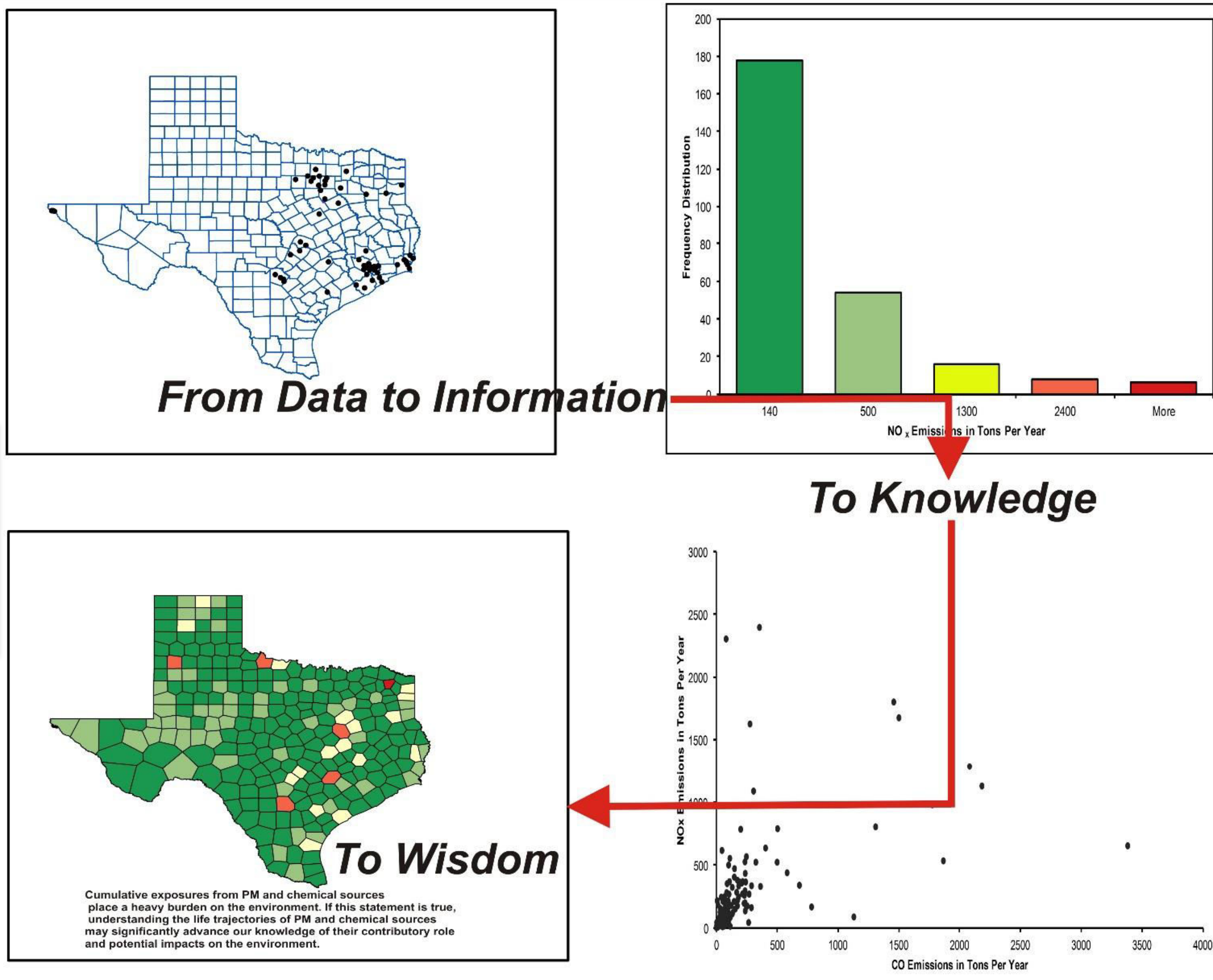
Spatial Analytical Tools and Methods	GISolve	GeoDa/ PySAL	Open-Topography	PGIST	pd-GRASS	R
Agent-based Modeling	X					X
Choice Modeling				X		
Domain-specific Modeling	X	X				X
Geostatistical Modeling	X					X
Local Clustering Detection	X	X				X
Spatial Interpolation	X	X				X
Spatial Econometrics		X				X
Visualization and Map Operations	X	X	X	X	X	X
Spatial Middleware	X					
Generic Cyberinfrastructure Capabilities	X	X			X	X
Online Problem-solving	X	X	X			X

Compiled from Schadt et al. (2010) and Wang et al. (2013)

The Scientific/Data Science Approach

- Knowledge Gap
- Problem Statement
- Literature Review
- Framework
- Hypothesis
- Study Design
- Analytical Elements
- Results
- Discussion & Implications
- Future Directions

Challenges



Big Data is fueled by **5Vs**: volume, variety, velocity, veracity, & value

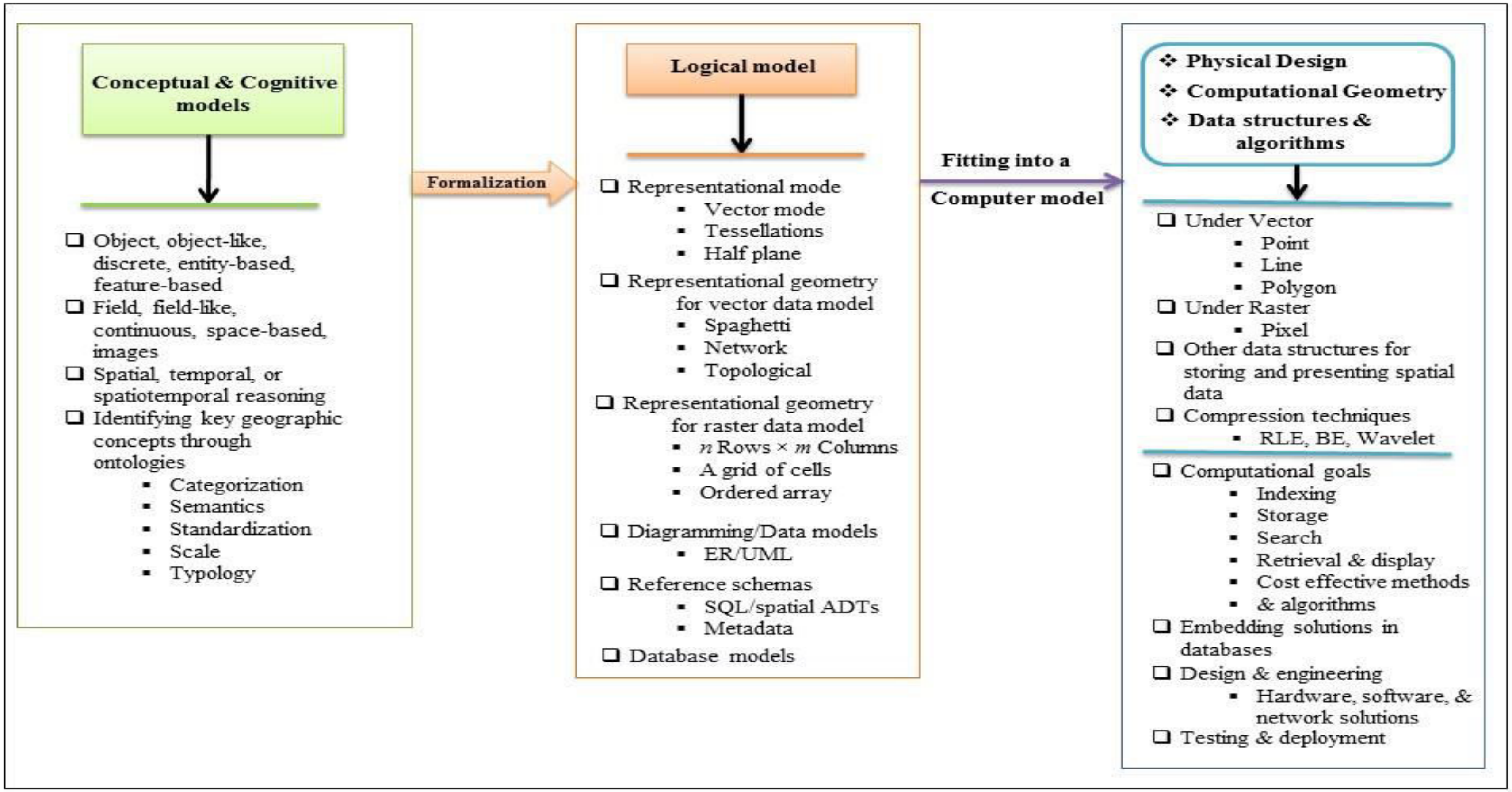
Opportunities & Challenges

- **From a Geospatial perspective,** understanding the unique attributes of spatial data, the spatial structure of the data, computational geometry, and the challenges that accompany the analysis of such data
- **Domain applications** must not be one-dimensional focusing on only the Descriptive or predictive Analytics, but include prescriptive analytics.
- **Develop a frontend software application or one-stop shop center/web-based tool** for pipelining backend computing technologies with geospatial data warehouses and data stream mining.

Opportunities & Challenges

- Improve methods for pipelining backend computing technologies with large-scale data warehouses and data stream mining.
- Ideas to consider:
  - Integration of ontological domain knowledge into spatial databases and domain applications
  - Data representation/spatial structure knowledge
  - Decomposition/scaling of methods and computational algorithms from desktop computing to heterogeneous computing and cloud computing environments
  - Development of frontend web-based technologies and interfaces for domain-specific science
  - Create forward-looking education and training opportunities in Data Science

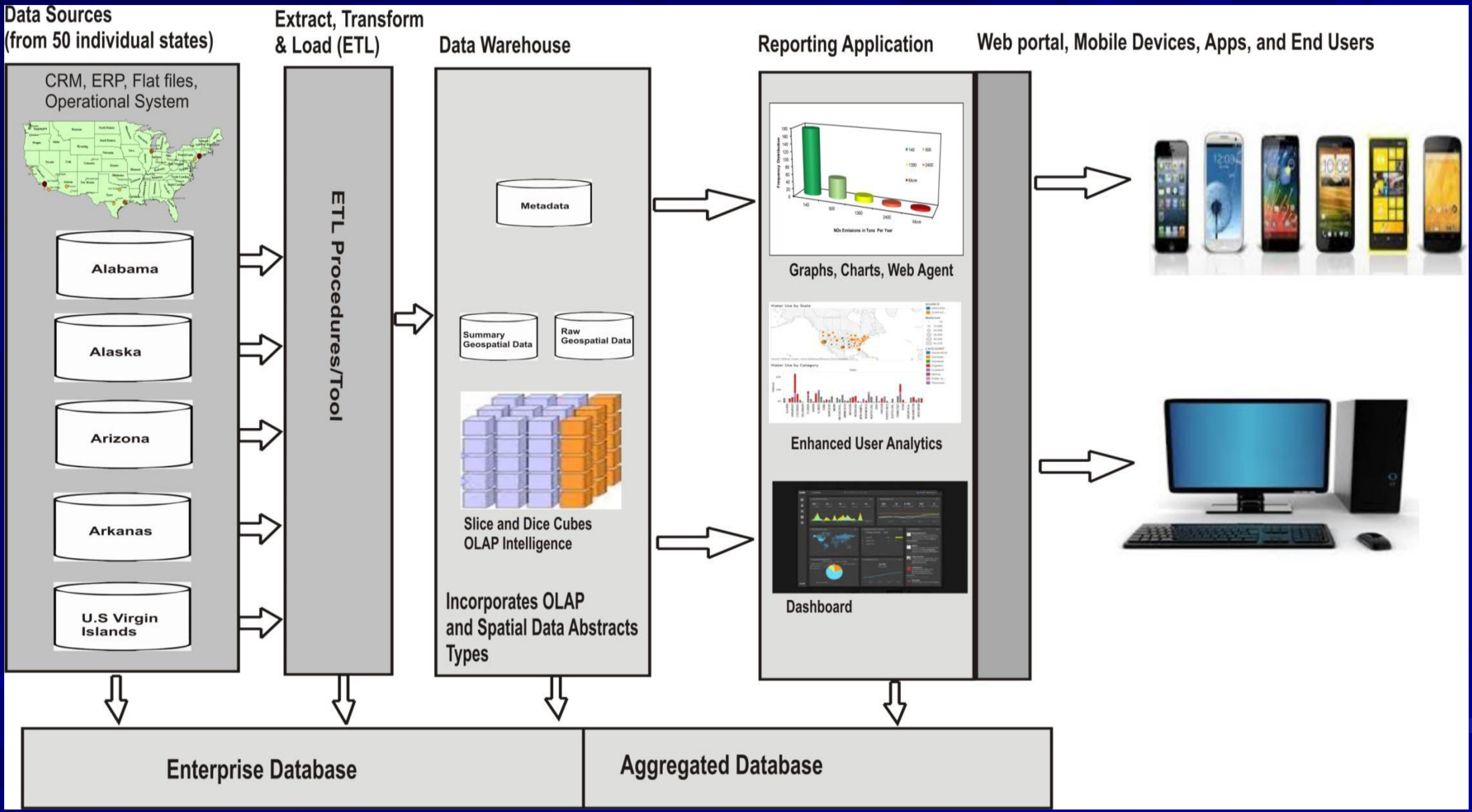
Data Representation Concepts



Schematic View of GIS Representation

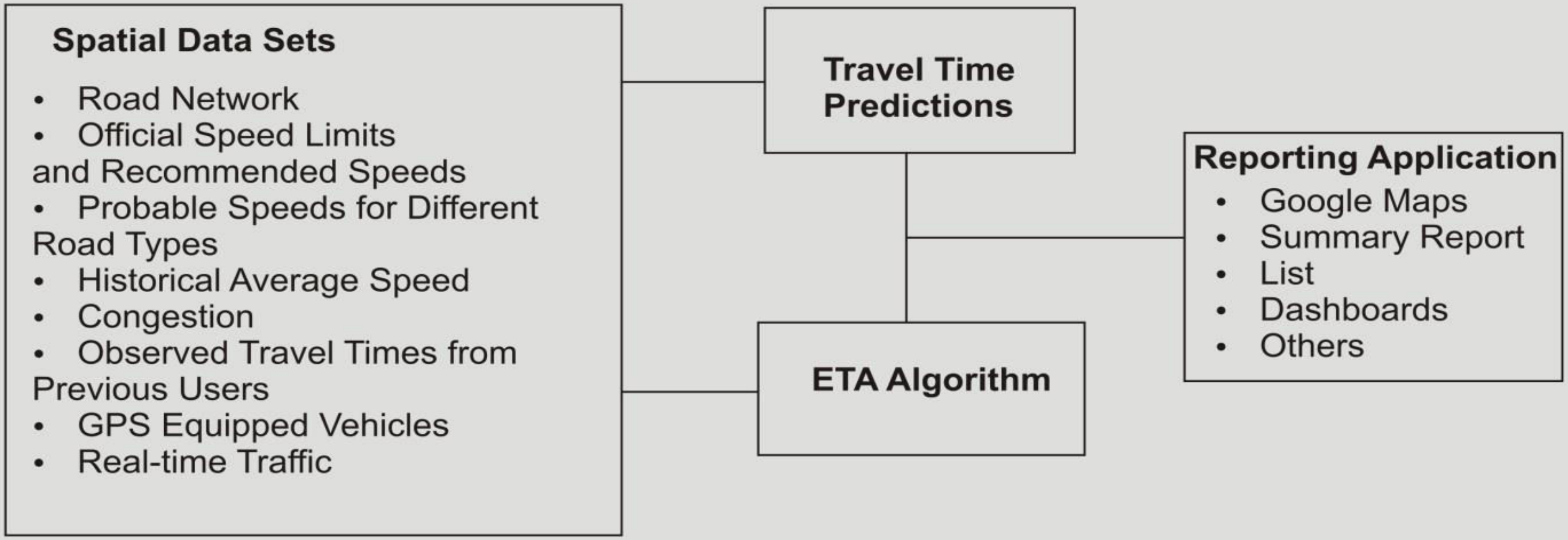
Data management (DM)

Analytics and Strategies (AS) & CSRA



A Data Intensive Application: Google ETA Algorithm

How Google Maps Estimated Time of Arrival (ETA) Algorithm Determines Travel Time for a Trip



Example Application I: Ensuring high quality and protection of Geocoded Health Data

Inspiration: ‘A scientist’s work is never complete, always evolving, learning, validating, and investigating better ideas/methods in pursuit of the scientific truth and a fine language to communicate the truth to a broad audience’

Rationale and Select Literature

- Geomasking techniques, such as **Random Direction and Fixed Radius, Random Perturbation within a Circle, Gaussian Displacement, *P*-sensitive *k*-anonymity algorithm, Donut Masking,** and **Bimodal Gaussian Displacement** are used to introduce noise and protect the privacy of individual-level information. But gaps persist, in the terms of the need to preserve spatial patterns, preserve space-time patterns, preserve temporal trends, and derive true environmental exposure measures.
- Only one study published in PNAS has approached geomasking as an optimization problem; however, the scope of the paper was limited to optimization for privacy protection and a few predefined set of locations for post-geomasking data. Our current study broadens this perspective.

Revealing the spatial distribution of a disease while preserving privacy

Shannon C. Wieland[a,b], Christopher A. Cassa[b], Kenneth D. Mandl[b,c,1], and Bonnie Berger[a,d,1]

[a]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307; [b]Children's Hospital Informatics Program at the Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Children's Hospital, Boston, MA 02115; [c]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115; and [d]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139-4307

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved August 19, 2008 (received for review February 1, 2008)

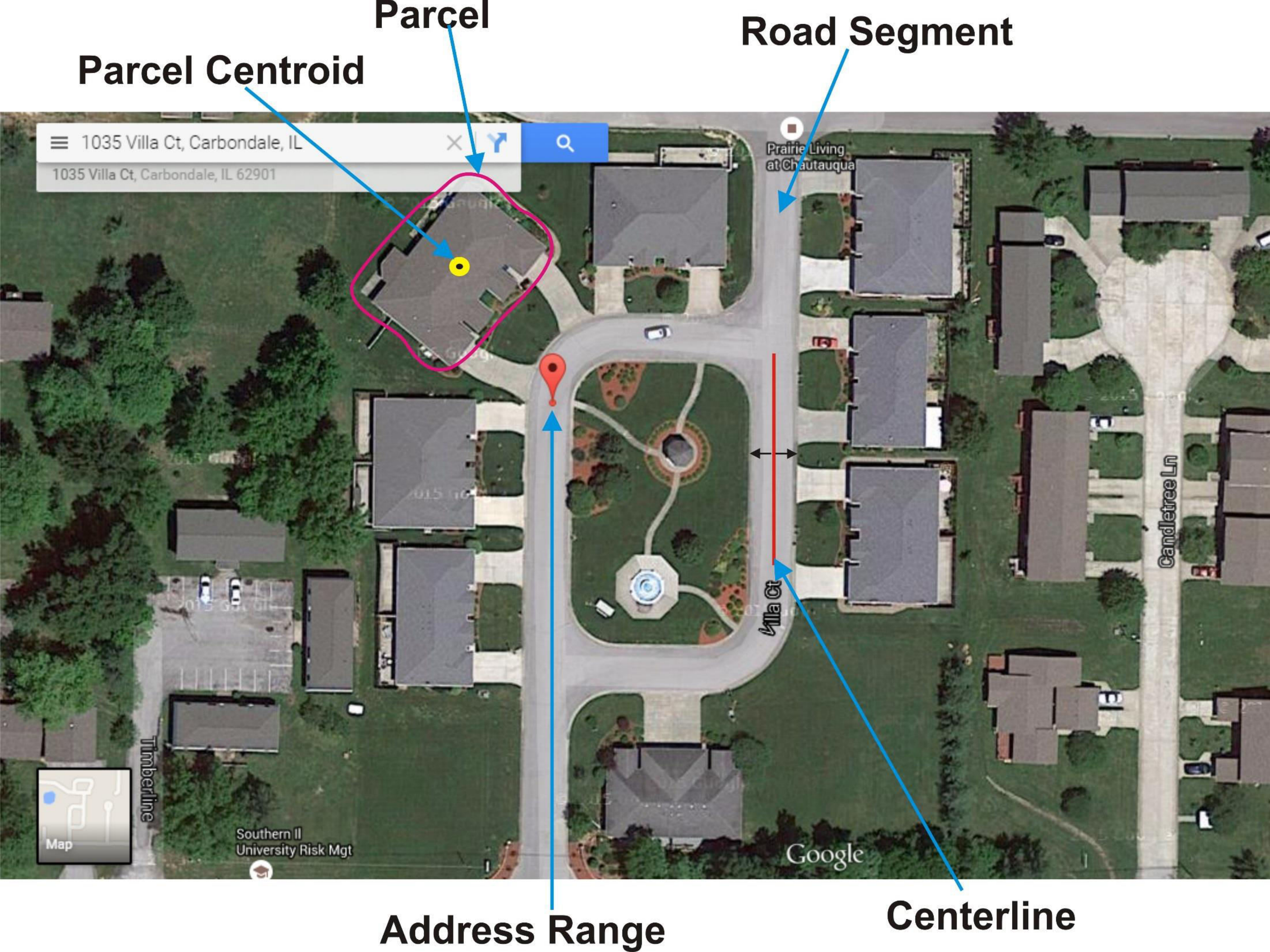
Datasets describing the health status of individuals are important for medical research but must be used cautiously to protect patient privacy. For patient data containing geographical identifiers, the conventional solution is to aggregate the data by large areas. This method often preserves privacy but suffers from substantial information loss, which degrades the quality of subsequent disease mapping or cluster detection studies. Other heuristic methods for de-identifying spatial patient information do not quantify the risk to individual privacy. We develop an optimal method based on linear programming to add noise to individual locations that preserves the distribution of a disease. The method ensures a small, quantitative risk of individual re-identification. Because the amount of noise added is minimal for the desired degree of privacy protection, the de-identified set is ideal for spatial epidemiological studies. We apply the method to patients in New York County, New York, showing that privacy is guaranteed while moving patients 25–150 times less than aggregation by zip code.

patient privacy | spatial epidemiology | linear programming | data aggregation

qualified individual determines "that there is a very small risk that the information could be used by others to identify a subject of the information" (5).

The prevailing method for preserving privacy in spatial data is aggregating by predefined administrative regions, such as counties or census enumeration districts. These areas must be larger than the zip code level to comply with HIPAA. However, aggregation may compromise subsequent research by erasing useful spatial information (6); for example, the detection of spatial clusters is significantly less sensitive and specific when data are aggregated even by zip code (7). Furthermore, the level of privacy protection depends on the number of patient records. For example, if it is revealed that 20 patients having a certain disease reside in a region containing 20,000 people, then there is a $\frac{1}{1,000}$ chance that a randomly selected individual from the region is one of the patients. However, if 200 patients with the disease live in the region, then the probability that a random individual from the region is among the set of patients increases to $\frac{1}{100}$.

An alternative to aggregation is moving each patient to a new location to ensure privacy (8), formalized by the family of "geographical masks" proposed by Armstrong *et al.* (9). Each is a deterministic or stochastic function of geographical identifiers



Parcel
Road Segment
Parcel Centroid
1035 Villa Ct, Carbondale, IL
1035 Villa Ct, Carbondale, IL 62901
Prairie Living at Chautauqua
Villa Ct
Candletree Ln
Timberline
Southern Il University Risk Mgt
Map
Google
Address Range
Centerline

Methods for Coordinate Location Matching

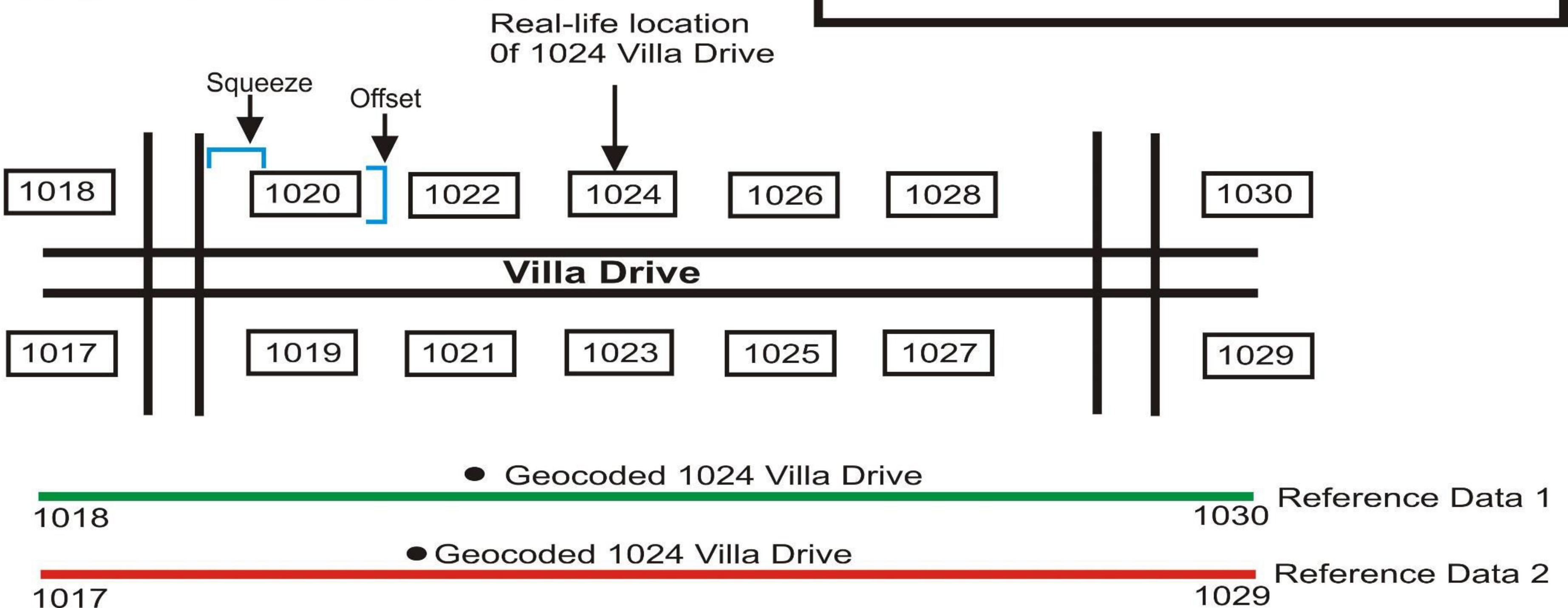
- Exact parcel centroid location
- Areal interpolation (e.g. parcel, building, building entrance, building access point, ZIP Code, village, city centroid)
- Address range interpolation

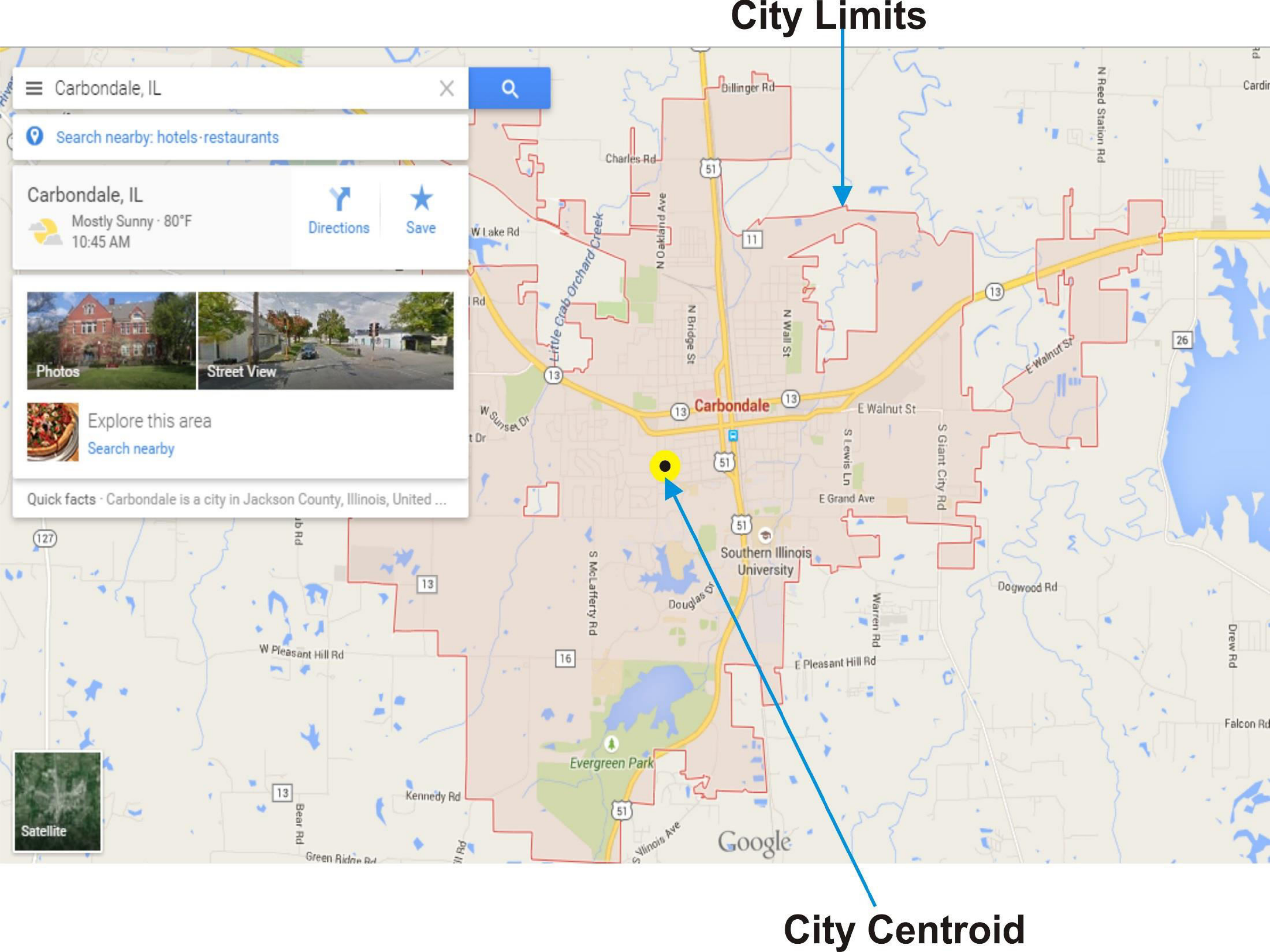
Sample Address Matching Algorithm

- Input Address (address & reference data)
- Test Similarity Measure (street, parcel measure etc.)
- Derive parameters (read addresses, rule-based for matching, create outputs)
- Find potential matches
- Assign match scores
- Decide best match
- End

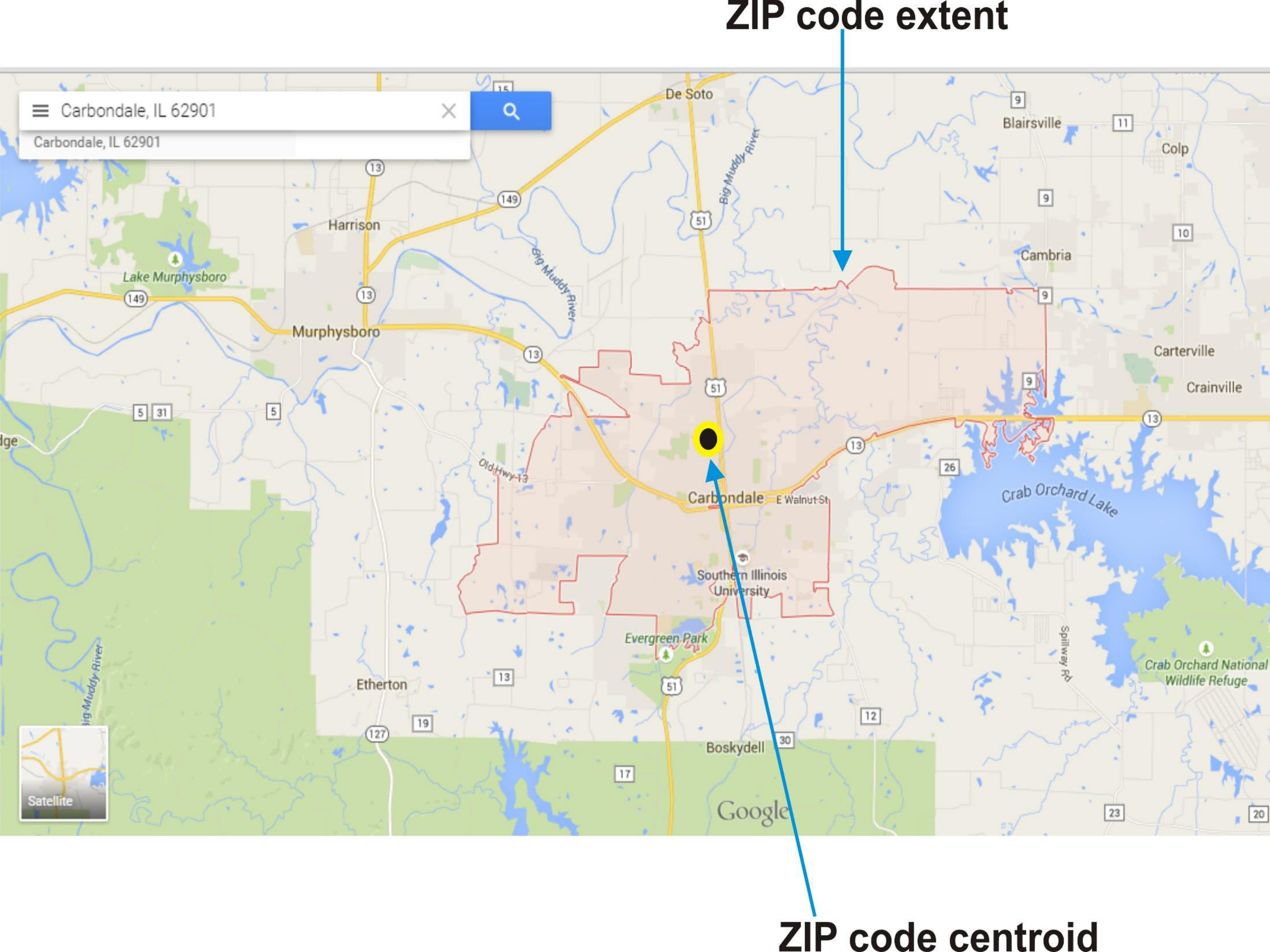
<u>Offset parameter</u> prevents a geocoded point being in the center of street where the address is located.

<u>Squeeze parameter</u> prevents a geocoded point residing in an intersection or too close to the end of a street.



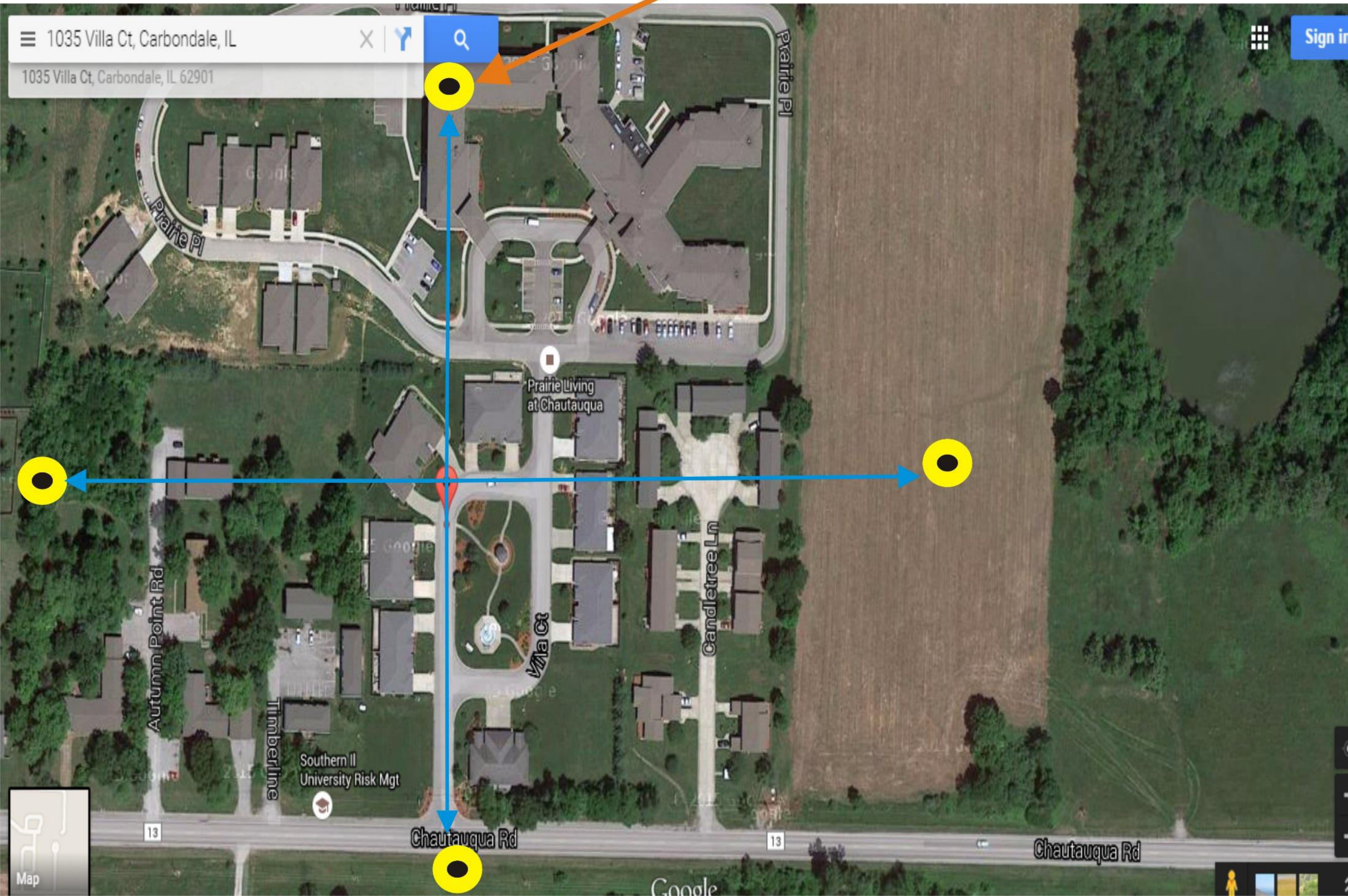


City Limits
City Centroid
Carbondale, IL
Search nearby: hotels·restaurants
Carbondale, IL
Mostly Sunny · 80°F
10:45 AM
Directions
Save
Photos
Street View
Explore this area
Search nearby
Quick facts · Carbondale is a city in Jackson County, Illinois, United ...
Carbondale
Southern Illinois University
Evergreen Park
Satellite
Google

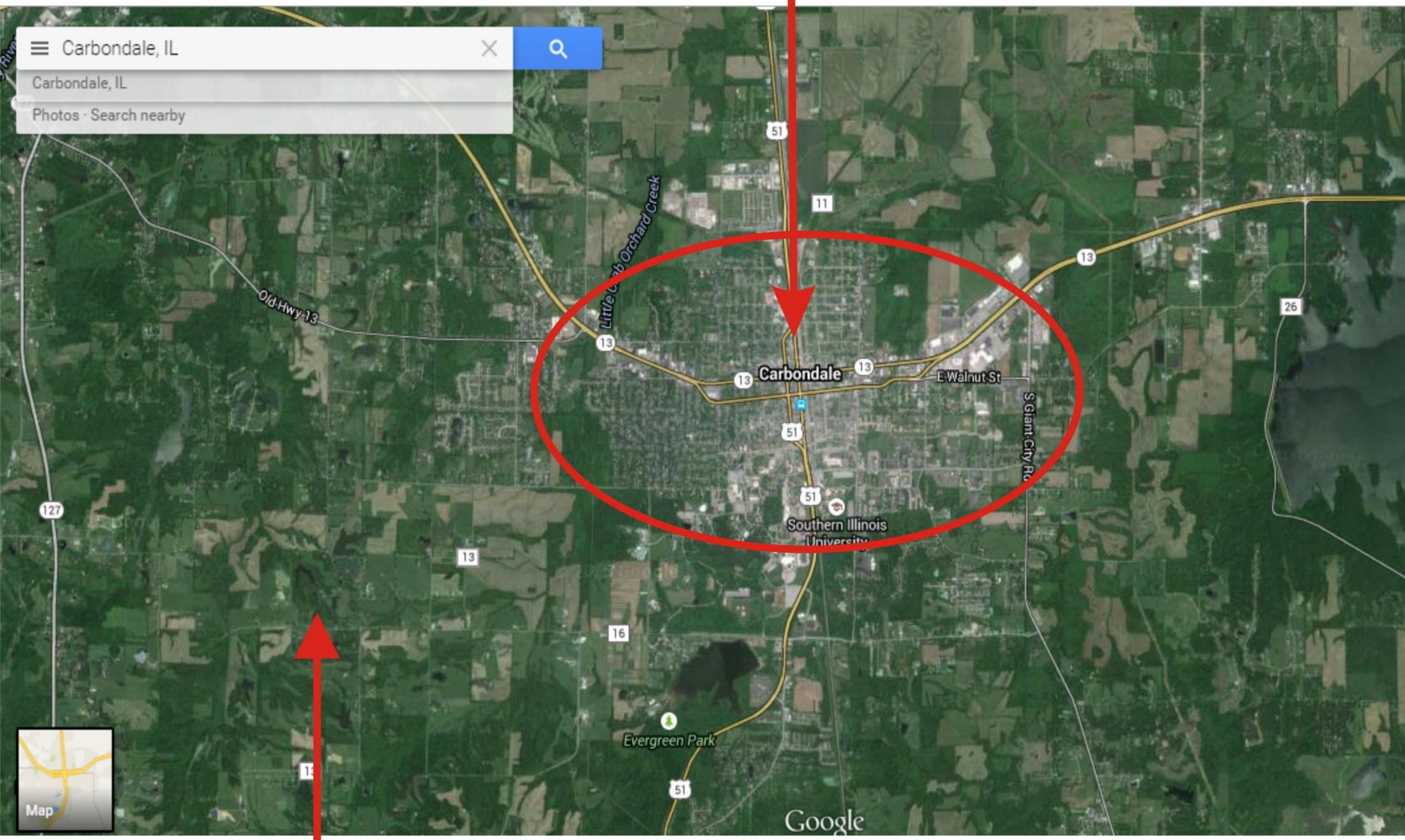


ZIP code extent
ZIP code centroid
Carbondale, IL 62901
Carbondale, IL 62901
De Soto
Blairsville
Colp
Harrison
Lake Murphysboro
Murphysboro
Cambria
Carterville
Crainville
Big Muddy River
Old Hwy 13
Carbondale
E Walnut St
Crab Orchard Lake
Southern Illinois University
Evergreen Park
Etherton
Boskydell
Spillway Rd
Crab Orchard National Wildlife Refuge
Satellite
Google

Potential Geomasking locations



Urban Setting



Rural Setting

The IDEA – Spatial analytics and Informatics (SAIC) research group

Geomasking Optimized Under Space-time and Exposure Constraints = GOUSTEC

Goal

- Enabling quality reproducible locational information to improve patient care

4 Multi-Objectives for optimization

*

- Identify five key socioeconomic measures [Race, Age, poverty, education, crime rate]
- Identify five key environmental measures [LULC, elevation, slope, temperature, precipitation, proximity to pollution source]
- Apply network distance
- Incorporate spatial and temporal patterns/dimension/clusters.

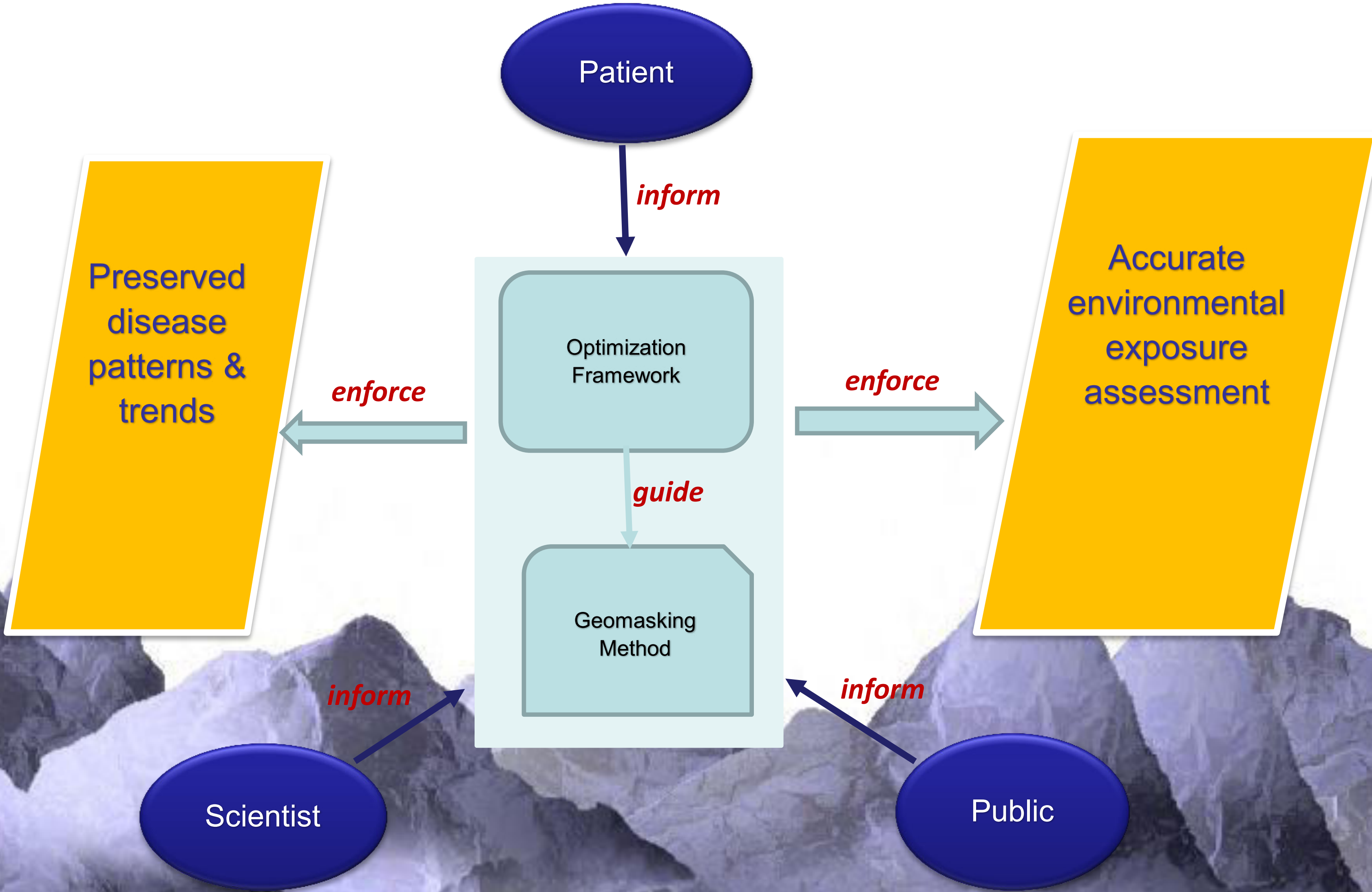
Research Team: Pierre Goovaerts, Yongmei Lu, Luke Achenie, and Tonny Oyana

First Step: Understanding pre-mask data by exploring all angles

Second Step: Geomasking under five constraints

Third Step: Exploring post-mask data to determine quality reproducible locational information to improve patient care/enhance targeted interventions

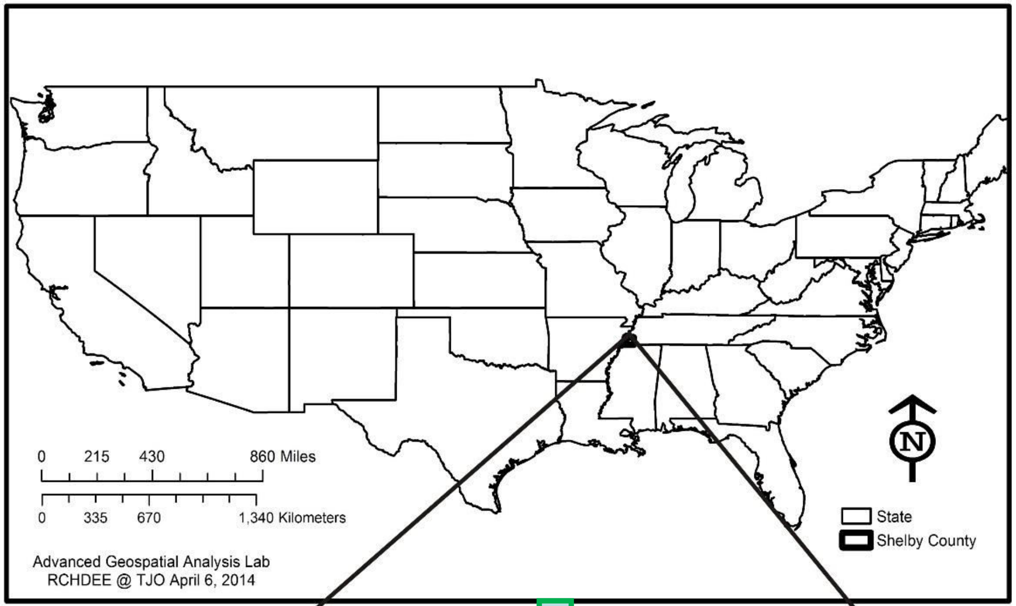
Geomasking Optimized Under Space-time and Exposure Constraints (GOUSTEC)

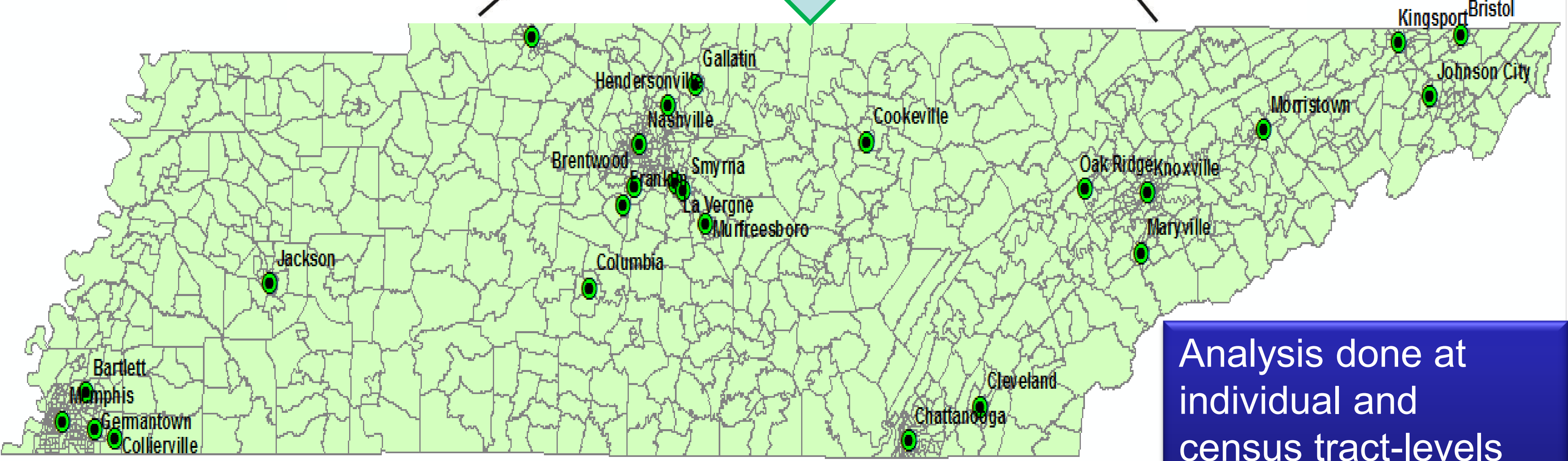


First Step: Understanding pre-mask/post mask data by exploring all angles

Published Results: Lindley LC, Oyana TJ. Geographic variation in mortality among children and adolescents diagnosed with cancer in Tennessee: Does race matter? (*In Press*, Journal of Pediatric Oncology Nursing).

Location of Study Area, TN, USA





Analysis done at individual and census tract-levels

Second Step: Geomasking under five constraints

Preliminary Results

Second Step: Geomasking under five constraints—Formalizing concepts

Define a set of constraints for the optimization framework in order to manage uncertainty introduced by geomasking. These constraints include g_1 through g_5:

1. Magnitude of displacement ensures *k-anonymity* for patient privacy protection, whereby a true health outcome case cannot be distinguished from at least *k-1* individuals to prevent re-identification (g_1),
2. Magnitude of displacement is spatially adaptive and varies as a function of land use and land cover (g_2), including the distribution of residential addresses and street network,
3. Spatial patterns in pre-mask data (measured by spatial statistics and variograms) are preserved (g_3),
4. Temporal trends in pre-mask data (modeled by time series or joinpoint regression at multiple scales are preserved (g_4),
5. Impact on exposure assessment (i.e. exposure to certain environmental hazards) is minimal (g_5).

$$\text{Model (1)}: \quad \min_x J(x) -> objective$$
$$s.t.\ g_j(x) \le 0,\ j = 1,\dots,K$$

$$\text{Model (2)}: \quad \min_x J(x) + \sum_{j=1}^{K} \lambda_j g_j(x) -> objective$$

$$\text{Model (3)}: \quad \min_x J(x) + \lambda_1 g_1(x) -> objective$$
$$s.t.\ g_j(x) \le \varepsilon_j\ (>0),\ j = 2,\dots,K$$

where *J(x)* is the default objective function
- *x* is the decision variable vector
- Minimizing *J(x)* leads to the optimal value of *x*

pyDonutGeomask1.0 - Notepad

File Edit Format View Help

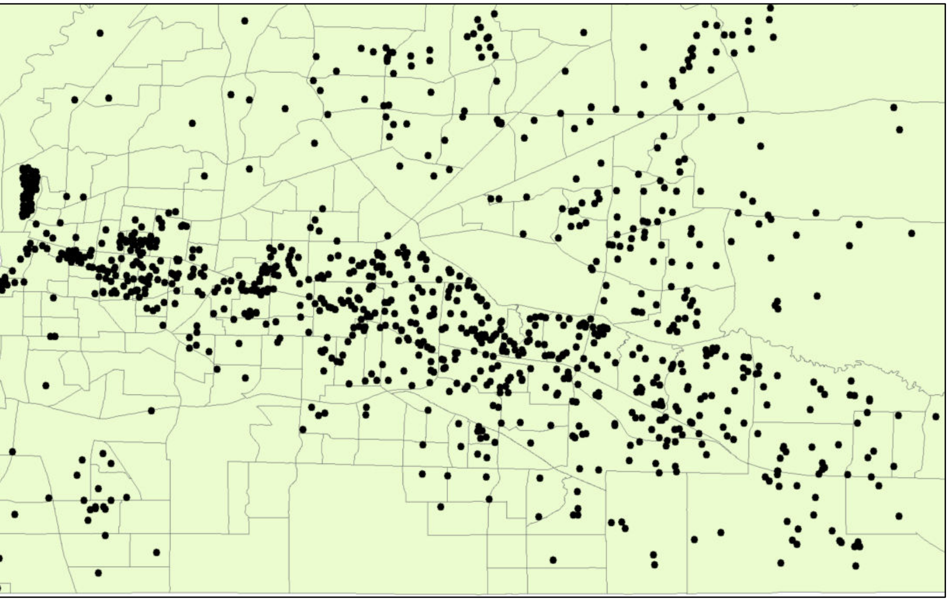
```
inputflag = 0
# Define User Input or Get As Parameters? (0=Define in Program, 1=Get As Parameters)
if inputflag==0:
    # Set EXTERNAL variables (local variables to be inputed by User)
    workspace = r"C:\Users\Tonny\Documents\ArcGIS\Geomasking.gdb"
    # Set local variables (local variables to be inputed by user) of Case File
    CaseFile = "OrangeCoord"
    CaseIDfieldname = "ID"
    CaseXfieldname = "X"
    CaseYfieldname = "Y"
    # Set local variables (local variables to be inputed by user) of Area File
    AreaFile = "Orange_blkgrp"
    AreaIDfieldname = "FIPS"
    Popfieldname = "POP2000"
    Areafieldname = "SQMI"
    DistanceUnit = "Miles"
    # Set local variables (local variables to be inputed by user) of Analysis Field
    kmin = 10
    kmax = 100
    # Set Random Seed Flag (Seed random generator? Y or N)
    rseed = "Y";
if inputflag==1:
    # Set EXTERNAL variables (local variables to be inputed by User)
    workspace = arcpy.GetParameterAsText(0)
    # Set local variables (local variables to be inputed by user) of Case File
    CaseFile = arcpy.GetParameterAsText(1)
    CaseIDfieldname = arcpy.GetParameterAsText(2)
    CaseXfieldname = arcpy.GetParameterAsText(3)
    CaseYfieldname = arcpy.GetParameterAsText(4)
    # Set local variables (local variables to be inputed by user) of Area File
    AreaFile = arcpy.GetParameterAsText(5)
    AreaIDfieldname = arcpy.GetParameterAsText(6)
    Popfieldname = arcpy.GetParameterAsText(7)
    Areafieldname = arcpy.GetParameterAsText(8)
    DistanceUnit = arcpy.GetParameterAsText(9)
    # Set local variables (local variables to be inputed by user) of Analysis Field
    kmin = arcpy.GetParameterAsText(10)
    kmax = arcpy.GetParameterAsText(11)
    # Set Random Seed Flag (Seed random generator? Y or N)
    rseed = arcpy.GetParameterAsText(12);

# --------------------------------------------
# Set INTERNAL variables (Defined for use in Program)
fieldPrecision = 18
fieldScale = 11
# Step1
```

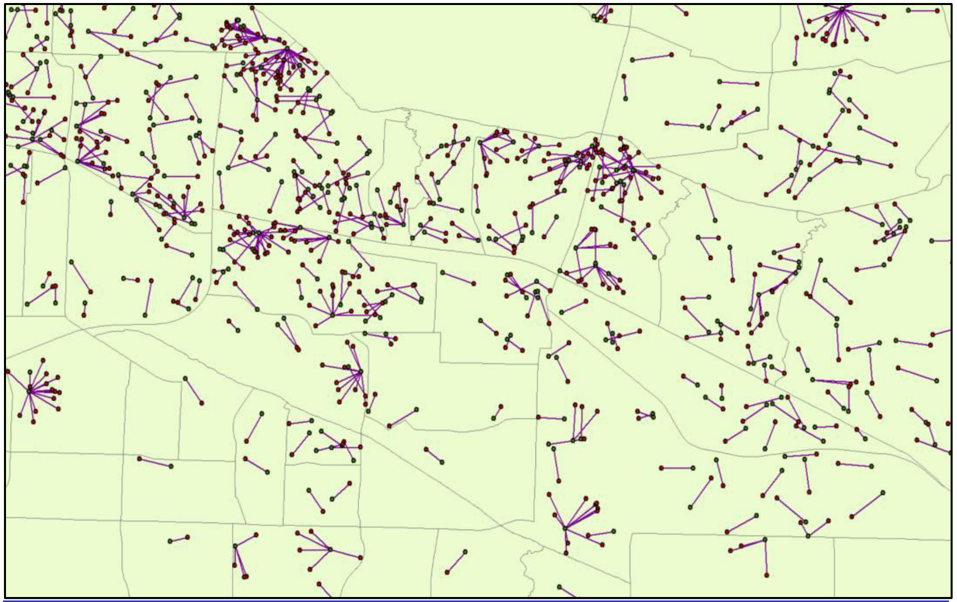
External parameters of code

Internal parameters of flexible optimization code

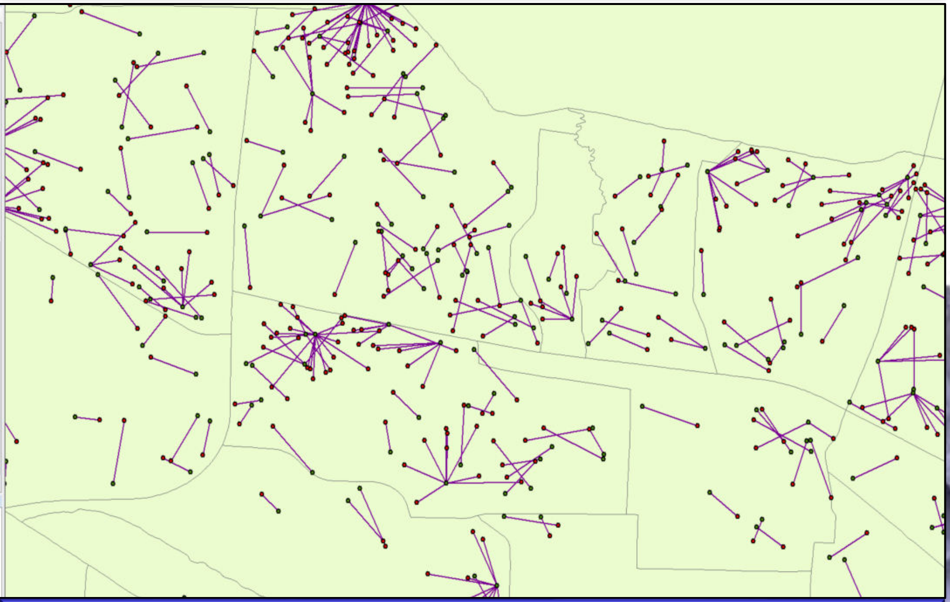
Implementation of a Test Code as a Python Script: Optimization Algorithm covers Objectives *g1* through *g3*



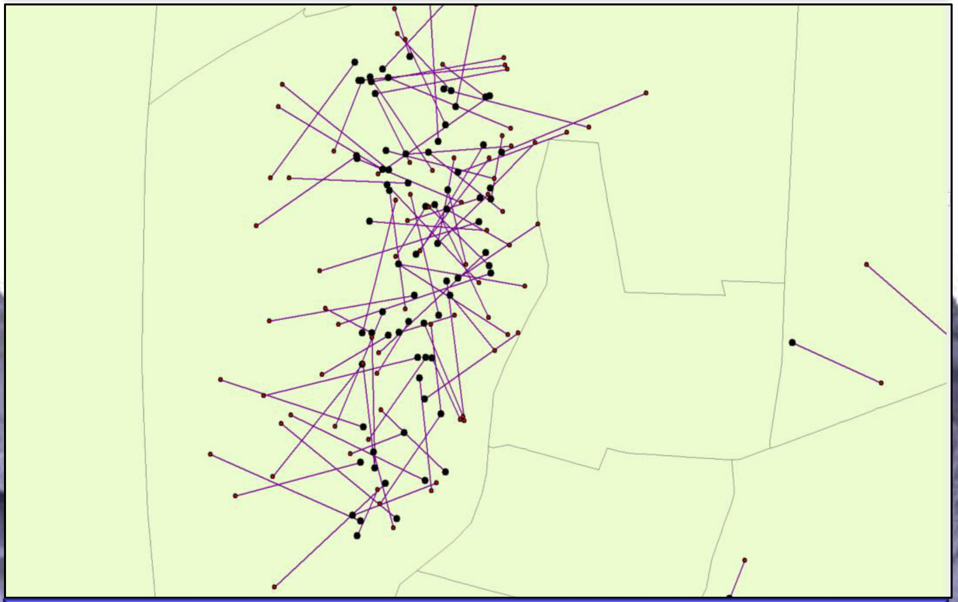
Spatial distribution of coordinate locations of individuals at census level



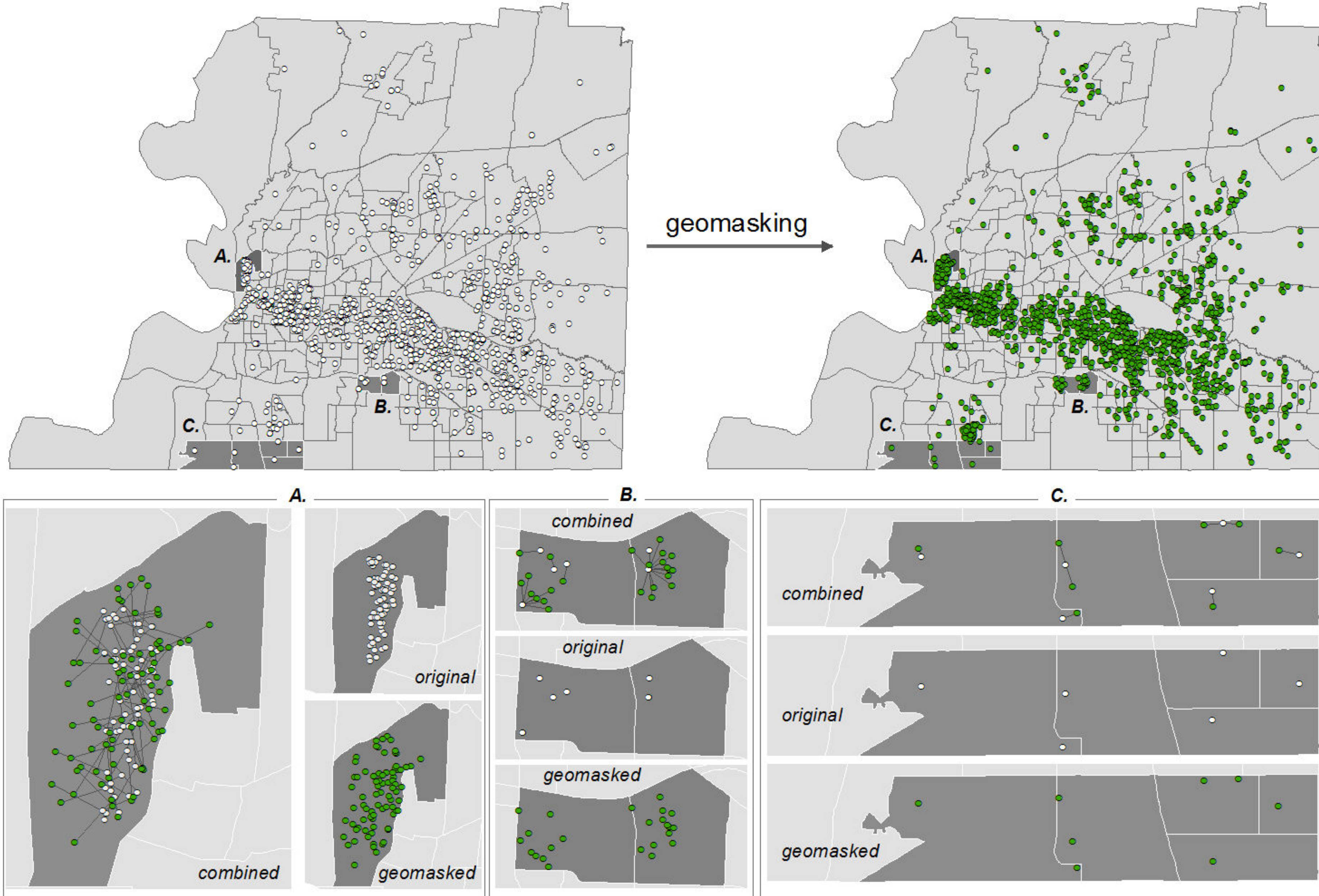
Spatial distribution of geomasked coordinate locations of individuals within census tracts (preserves *g1*–*g3*)



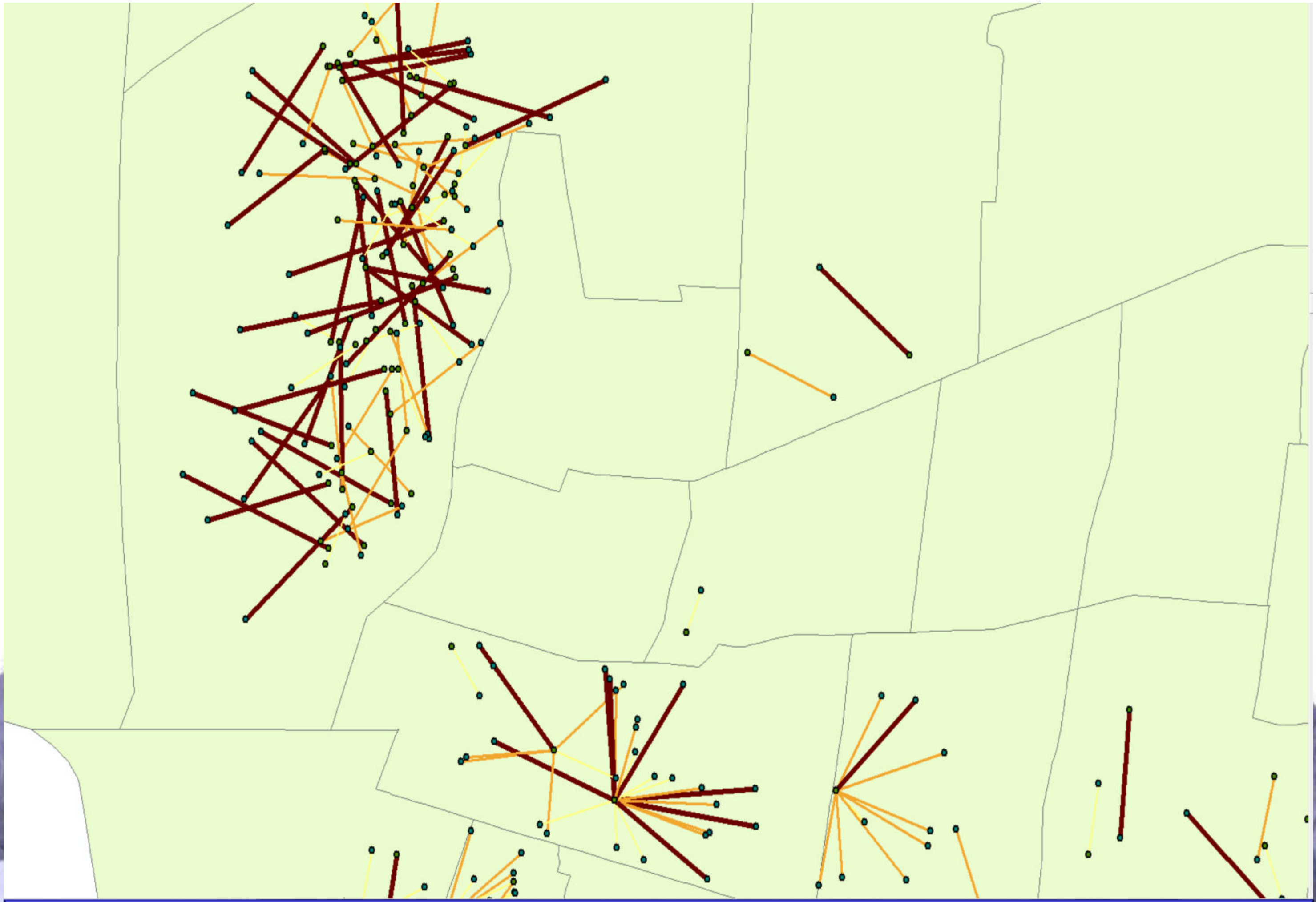
Geomasked data patterns: Zoomed in to 6 census tracts



Geomasked data patterns: Zoomed in to 1 census tract


geomasking
A.
B.
C.
A.
combined
original
geomasked
B.
combined
original
geomasked
C.
combined
original
geomasked

Third Step: Exploring post-mask data to determine quality reproducible locational information to improve patient care/enhance targeted interventions



Analyzing and visually communicating noise in a post-mask dataset: *k*-anonymity ranges from 0.1 to 0.5 km within the census tract

The Dilemma of Geocoded Data

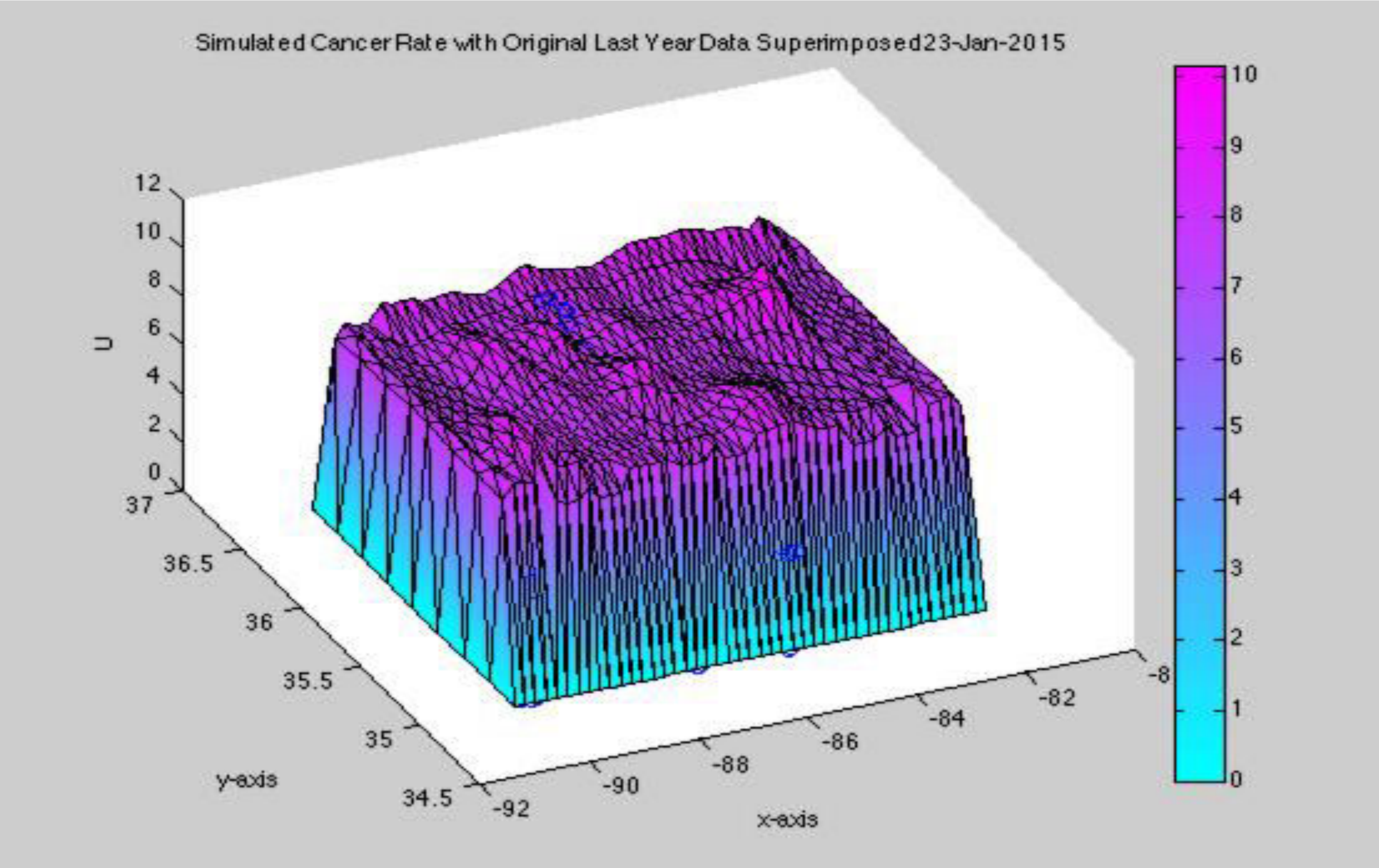
	Reference Data 1				Reference Data 2							
SCORE	X	Y	DATA_ID	Score_GDT	X_1	Y_1	Difference	Distance	STREETCODI	CITY	STATE	ZIP
100	-78.838704225	42.937843995	32	100	-78.8392602905	42.9375732500	0.0006184753	54.3636189181	2346 FILLMORE AVENUE	BUFFALO	NY	14214
100	-78.863993612	42.889849543	34	100	-78.8644835484	42.8902414256	0.0006273832	59.0927810642	96 ASH STREET	BUFFALO	NY	14204
100	-78.881021135	42.892655081	39	100	-78.8813898851	42.8928224717	0.0004049645	35.3390191139	222 CAROLINA STREET	BUFFALO	NY	14201
95	-78.855521445	42.912803457	53	100	-78.8558951231	42.9127571583	0.0003765354	30.8641939463	60 VERPLANK STREET	BUFFALO	NY	14208
100	-78.891221838	42.907529326	92	100	-78.8912777218	42.9075309905	0.0000559086	4.5549160541	379 PLYMOUTH AVENUE	BUFFALO	NY	14213
100	-78.891221838	42.907529326	94	100	-78.8912777218	42.9075309905	0.0000559086	4.5549160541	379 PLYMOUTH AVENUE	BUFFALO	NY	14213
100	-78.861991210	42.873638894	99	100	-78.8624435291	42.8735560297	0.0004598467	37.9935759371	93 HAYWARD STREET	BUFFALO	NY	14204
100	-78.845980472	42.820726705	103	100	-78.8463618657	42.8207183399	0.0003814854	31.1202482587	16 WOODYARD WAY	LACKAWANNA	NY	14218
80	-89.934806000	35.231674000			-89.9347765912	35.2320090635	0.0003363516	37.3529079657				
85	-89.992777000	35.077503000			-90.0755865165	35.1080123836	0.0882509972	8262.7474544959				
80	-90.038229000	35.148301000			-90.0395599455	35.1482048816	0.0013344117	121.4809077338	Undisclosed Memphis Locations			
85	-89.987627000	35.215920000			-90.0755865165	35.1080123836	0.1392154094	14418.9418625529				
85	-90.040407000	35.113239000			-90.0669814997	35.0906514227	0.0348769649	3486.0755663428				

Microsoft Visual Basic for Applications - asthma_test_model_data.xls

asthma_test_model_data.xls - Sheet1 (Code)

```
Public Function CalculateDistance(Lat1, Lon1, Lat2, Lon2) As Double
        Dim a, b, c As Double
        Const PI = 3.14159265358979
        Const RadiusEarth = 6371000
        ' 1 Degree is 69.096 miles, 1 mile is 1609.34 m
        a = Cos(Lat1 * PI / 180) * Cos(Lat2 * PI / 180) * Cos(Lon1 * PI / 180) * Cos(Lon2 * PI / 180)
        b = Cos(Lat1 * PI / 180) * Sin(Lon1 * PI / 180) * Cos(Lat2 * PI / 180) * Sin(Lon2 * PI / 180)
        c = Sin(Lat1 * PI / 180) * Sin(Lat2 * PI / 180)
        If (a + b + c) >= 1 Or (a + b + c) <= -1 Then
            CalculateDistance = 0
        Else
            CalculateDistance = Application.WorksheetFunction.Acos(a + b + c) * RadiusEarth 'Distance will be in meters
        End If
End Function
```

Validation of Matched Addresses: Testing positional accuracy between two reference data sets. Upper rows are asthma data set in Buffalo, New York; and lower rows are from undisclosed cohort data set of Memphis, TN



Reaction-Diffusion Mechanistic Models (RDMM) of Pediatric Cancer Estimates in Tennessee

The Dilemma: What do you preserve?

- Do you preserve distance metrics, directional metrics, SES metrics, space-time metrics, spatial patterns, temporal patterns or environmental exposure metrics?
- Your results should inform your decision
- Our goal is to build a flexible spatially-adaptive optimization algorithm that can accommodate true locational identity [should reflect core objectives *g1–g5*]. Thus enabling high quality & reproducible locational info to improve patient care.

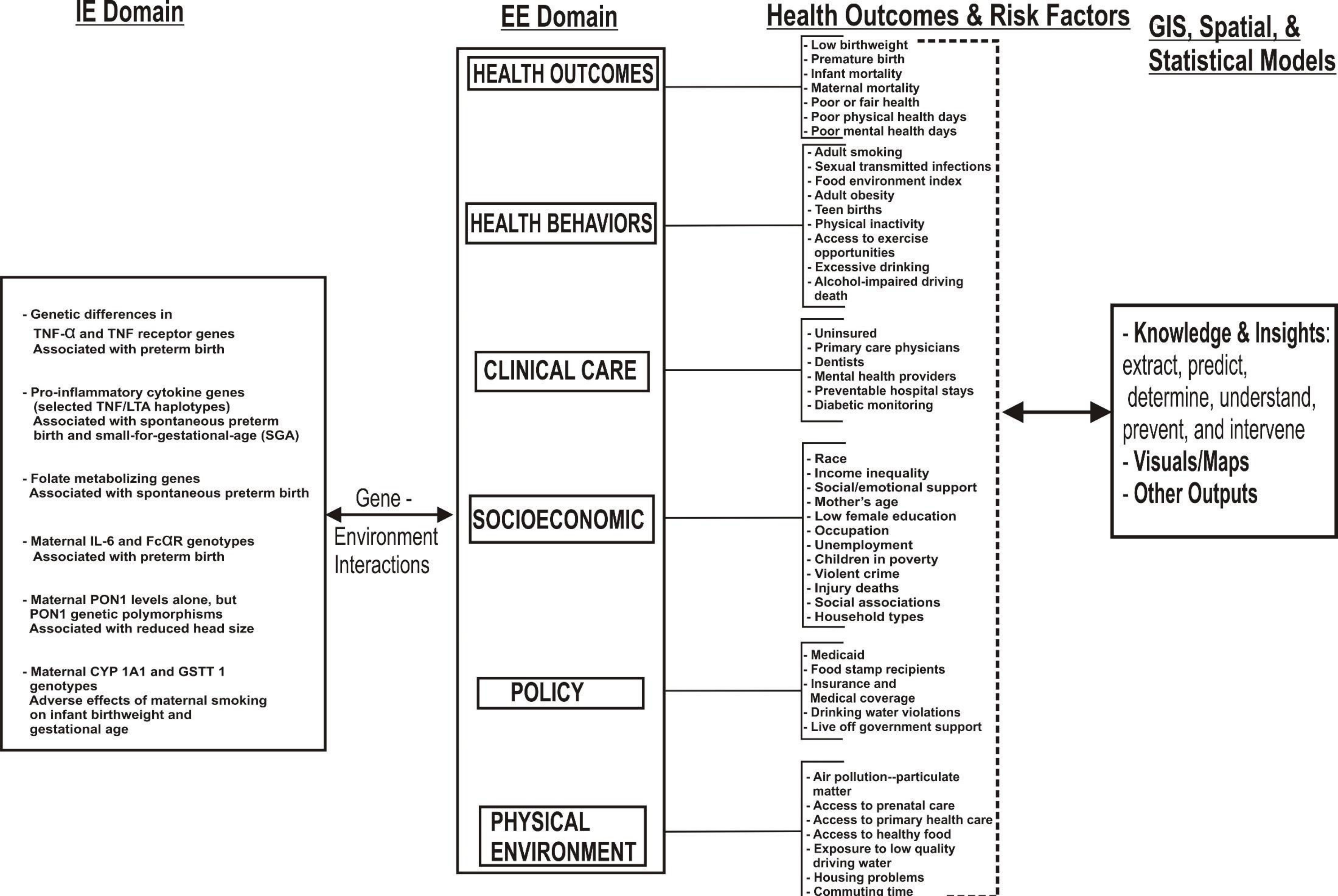
Research Team: Tonny J. Oyana, Patricia Matthews-Juarez, Stephania A. Cormier, Xiaoran Xu, and Paul D. Juarez

Example Application II: Using an External Exposome Framework to Examine Pregnancy-Related Morbidities and Mortalities: Implications for Health Disparities Research

Inspiration: ‘A scientist’s work is never complete, always evolving, learning, and investigating better ideas/methods in pursuit of the scientific truth and a fine language to communicate the truth to a broad audience’

Rationale and Select Literature

- A few recent studies exist on this emerging exposome topic, but there is still little information on geographically-integrated health measures.
- Genetic factors specific to the internal exposome domain or are reported to be associated with preterm birth, reduced head size, infant birthweight, and premature birth have been established through an extensive r of the Lit.
- Non-genetic factors that make up the external exposome domain and are specific to this application include, health outcomes, health behaviors, clinical care, socioeconomic, policy and programs, and the physical environment.
- Current framework combines exposome, GIS and spatial analysis, spatiotemporal models, computational and traditional statistical analytics to study the complex relationships of LBW across U.S. counties.



A conceptual framework for understanding adverse birth outcomes synthesized from over 50 research articles. IE refers to internal exposome while EE is the external exposome

International Journal of
Environmental Research and Public Health

Article

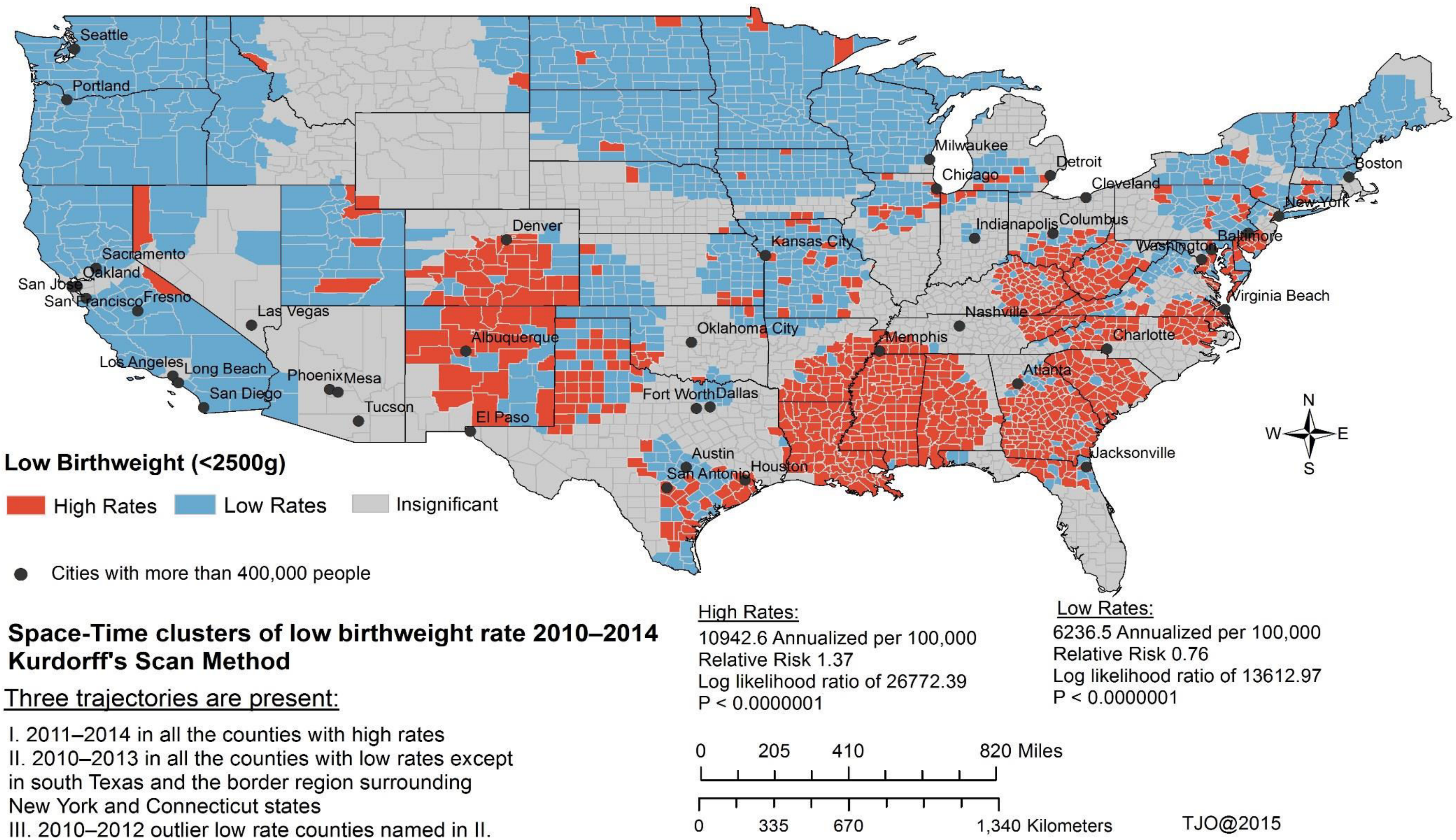
Using an External Exposome Framework to Examine Pregnancy-Related Morbidities and Mortalities: Implications for Health Disparities Research

Tonny J. Oyana [1,*], Patricia Matthews-Juarez [2,3], Stephania A. Cormier [2], Xiaoran Xu [2] and Paul D. Juarez [2,3]

Received: 12 August 2015; Accepted: 17 November 2015; Published: 22 December 2015
Academic Editors: Mark Edberg, Barbara E. Hayes, Valerie Montgomery Rice and Paul B. Tchounwou

1 Research Center on Health Disparities, Equity & the Exposome, Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA
2 Pediatrics, Infectious Disease and Microbiology, Immunology & Biochemistry, University of Tennessee Health Science Center, Le Bonheur Children's Medical Center, Memphis, TN 36163, USA; pmatthews-juarez@mmc.edu (P.M.-J.); scormier@uthsc.edu (S.A.C.); xxu24@uthsc.edu (X.X.); pjuarez@mmc.edu (P.D.J.)
3 Department of Family and Community Medicine, Meharry Medical College, Nashville, TN 37208, USA
* Correspondence: toyana@uthsc.edu; Tel.: +1-901-448-2829; Fax: +1-901-448-2701

Abstract: *Objective*: We have conducted a study to assess the role of environment on the burden of maternal morbidities and mortalities among women using an external exposome approach for the purpose of developing targeted public health interventions to decrease disparities. *Methods*: We identified counties in the 48 contiguous USA where observed low birthweight (LBW) rates were higher than expected during a five-year study period. The identification was conducted using a


Seattle
Portland
Sacramento
Oakland
San Jose
San Francisco
Fresno
Las Vegas
Los Angeles
Long Beach
San Diego
Phoenix
Mesa
Tucson
Denver
Albuquerque
El Paso
Kansas City
Oklahoma City
Fort Worth
Dallas
Austin
San Antonio
Houston
Milwaukee
Chicago
Detroit
Cleveland
Indianapolis
Columbus
Nashville
Memphis
Atlanta
Charlotte
Jacksonville
Washington
Baltimore
Virginia Beach
New York
Boston
N
W
E
S
Low Birthweight (<2500g)
High Rates
Low Rates
Insignificant
Cities with more than 400,000 people
Space-Time clusters of low birthweight rate 2010–2014
Kurdorff's Scan Method
Three trajectories are present:
I. 2011–2014 in all the counties with high rates
II. 2010–2013 in all the counties with low rates except
in south Texas and the border region surrounding
New York and Connecticut states
III. 2010–2012 outlier low rate counties named in II.
High Rates:
10942.6 Annualized per 100,000
Relative Risk 1.37
Log likelihood ratio of 26772.39
P < 0.0000001
Low Rates:
6236.5 Annualized per 100,000
Relative Risk 0.76
Log likelihood ratio of 13612.97
P < 0.0000001
0
205
410
820 Miles
0
335
670
1,340 Kilometers
TJO@2015

Specific Aims for Effects of $PM_{2.5}$ Lifetime Exposure on Child Health

- Examine the spatiotemporal relationship between
  - hospitalization rate
  - ER visits

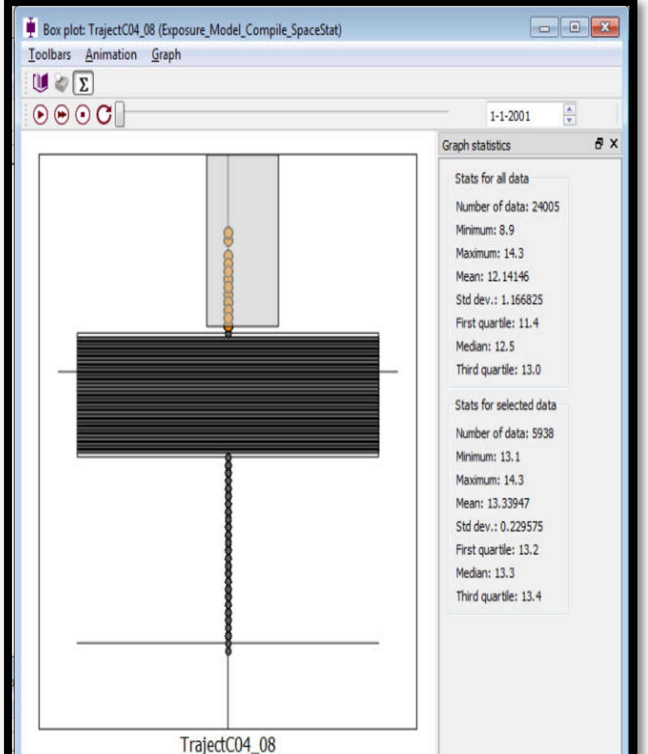
  of children with an asthma diagnosis and $PM_{2.5}$ exposure

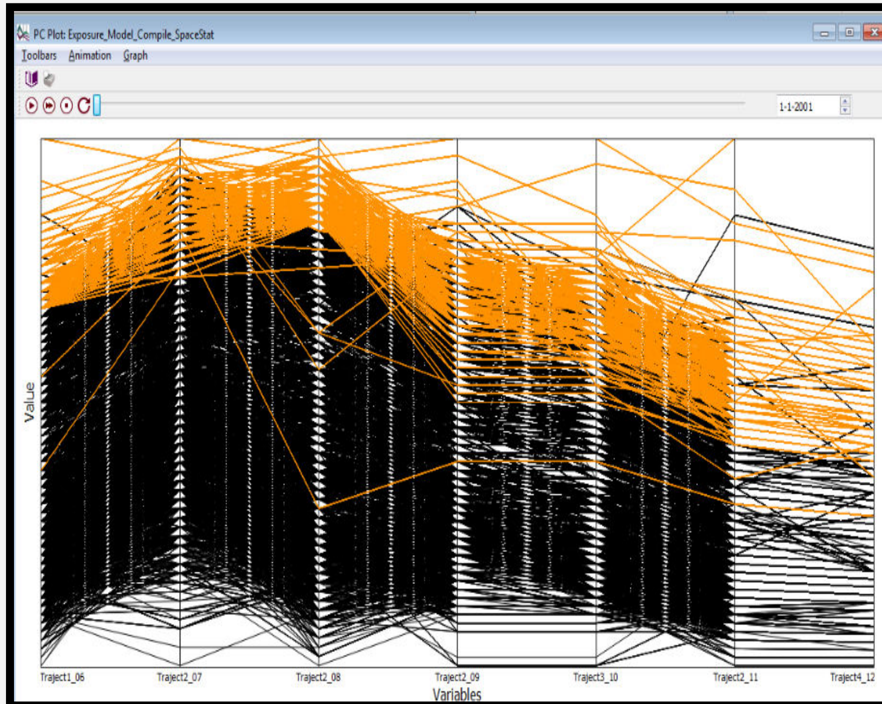
- Examine the temporal relationship between
  - hospitalization rate
  - ER visits

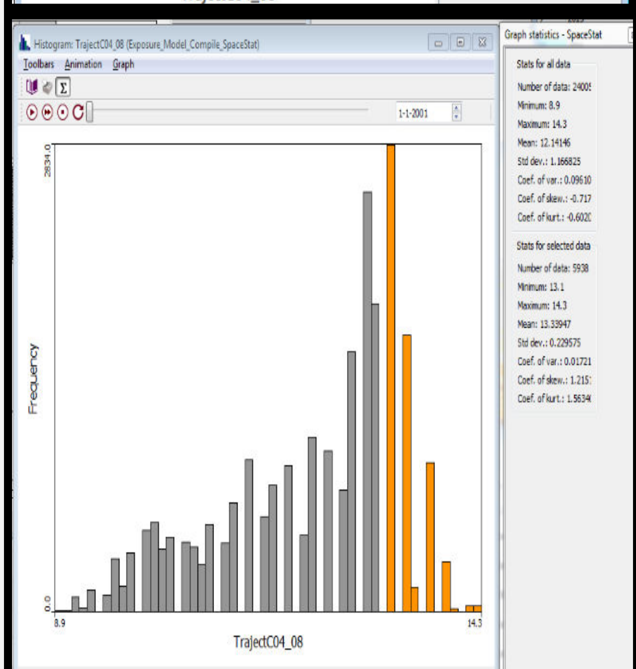
  of children with an asthma diagnosis and $PM_{2.5}$ exposure

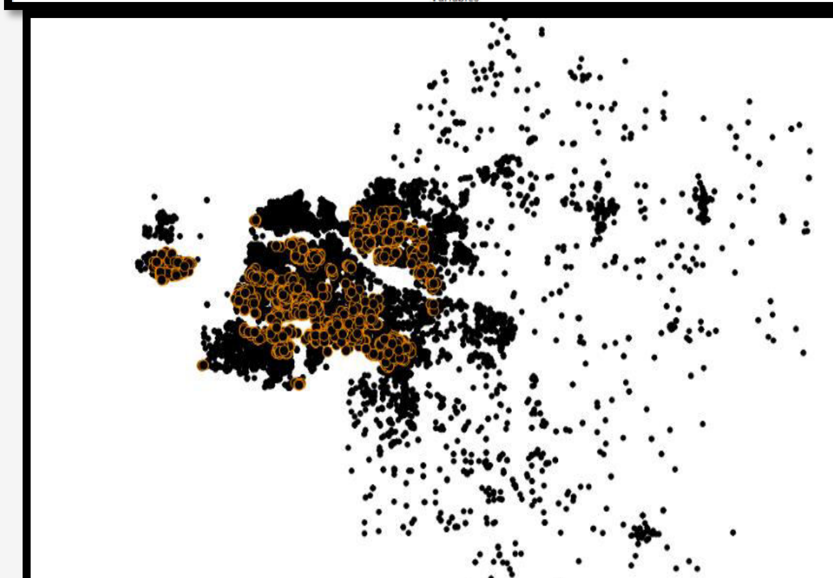
Accomplishing the Aims

- Spatiotemporal
  - Measuring short/long-term effects of drivers of asthma hospitalization/ER visits
- Temporal
  - Measuring trajectories of asthma hospitalization/ER visit drivers





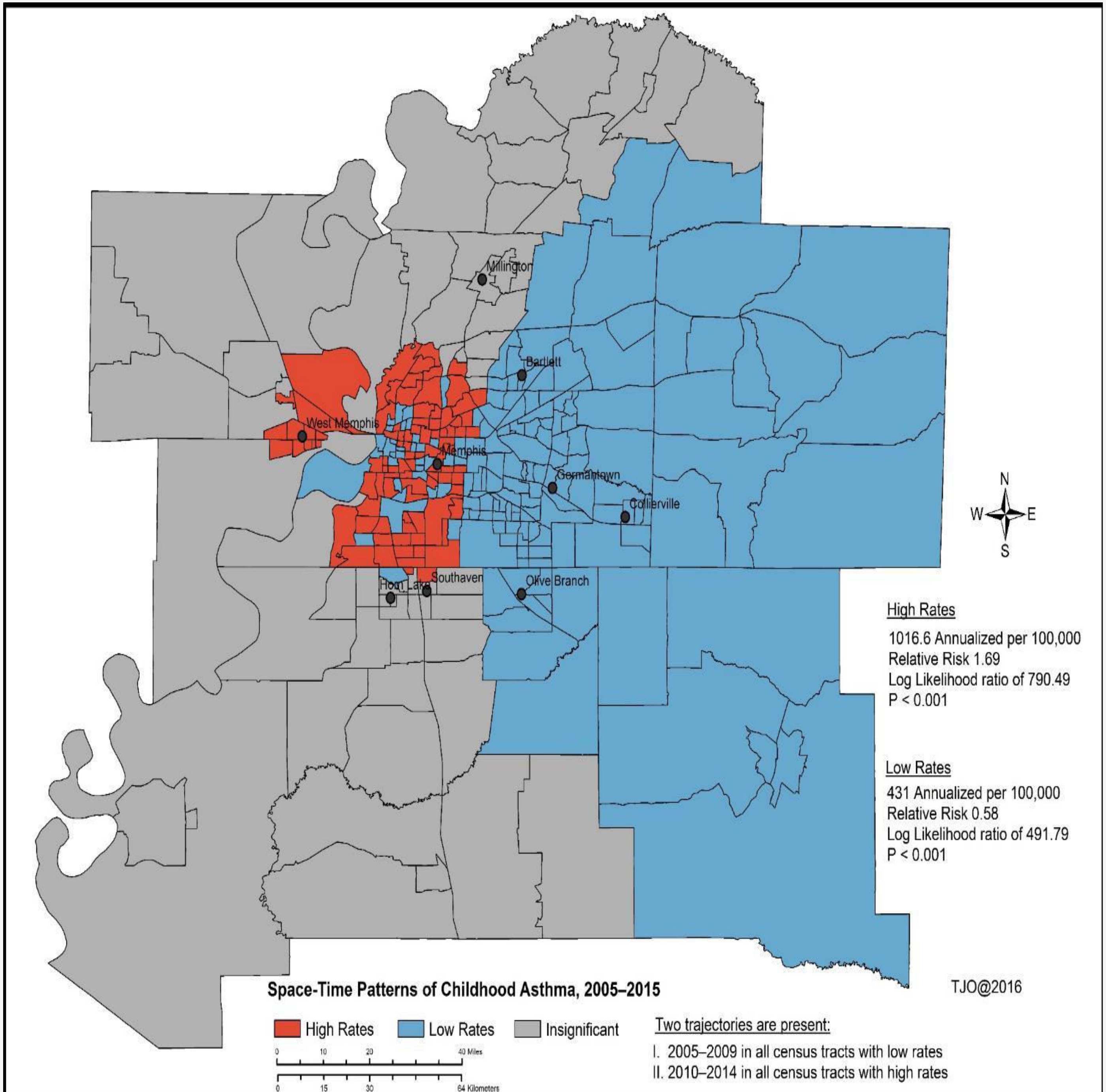




Temporal Exploration using Data Brushing Techniques

	Fayette	Shelby	Tipton	Benton	Desoto	Marshall	Tate	Tunica	
State	TN	TN	TN	MS	MS	MS	MS	MS	
Core urbanized county or outlying?	Outlying	Core	Outlying	Outlying	Core	Outlying	Out-lying	Outlying	Core
SO2 Emissions (tpy)	429	20,010	308	36	52	63	58	107	125
Primary Sulfate Emissions (tpy)	10	140	8	10	29	6	6	8	11
$PM_{2.5}$ emissions (tpy)	790	4,042	874	475	1,419	1,064	651	1,471	1,854
Population	38,413	928,792	61,160	8,712	161,732	37,098	28,970	10,741	50,952
Population (% of CBSA)	3%	70%	5%	1%	12%	3%	2%	1%	4%
Population growth (2000 – 2010)	33%	4%	19%	9%	51%	6%	14%	16%	0.2%
VMT (Millions)	540	8,562	417	190	1,798	683	365	237	866
VMT (% of CBSA)	4%	63%	3%	1%	13%	5%	3%	2%	6%

The EPA also evaluated the meteorology in the area by evaluating wind data collected at the



Millington
Bartlett
West Memphis
Memphis
Germantown
Collierville
Southaven
Horn Lake
Olive Branch
N
W
E
S
High Rates
1016.6 Annualized per 100,000
Relative Risk 1.69
Log Likelihood ratio of 790.49
P < 0.001
Low Rates
431 Annualized per 100,000
Relative Risk 0.58
Log Likelihood ratio of 491.79
P < 0.001
Space-Time Patterns of Childhood Asthma, 2005–2015
TJO@2016
High Rates
Low Rates
Insignificant
0 10 20 40 Miles
0 15 30 64 Kilometers
Two trajectories are present:
I. 2005–2009 in all census tracts with low rates
II. 2010–2014 in all census tracts with high rates

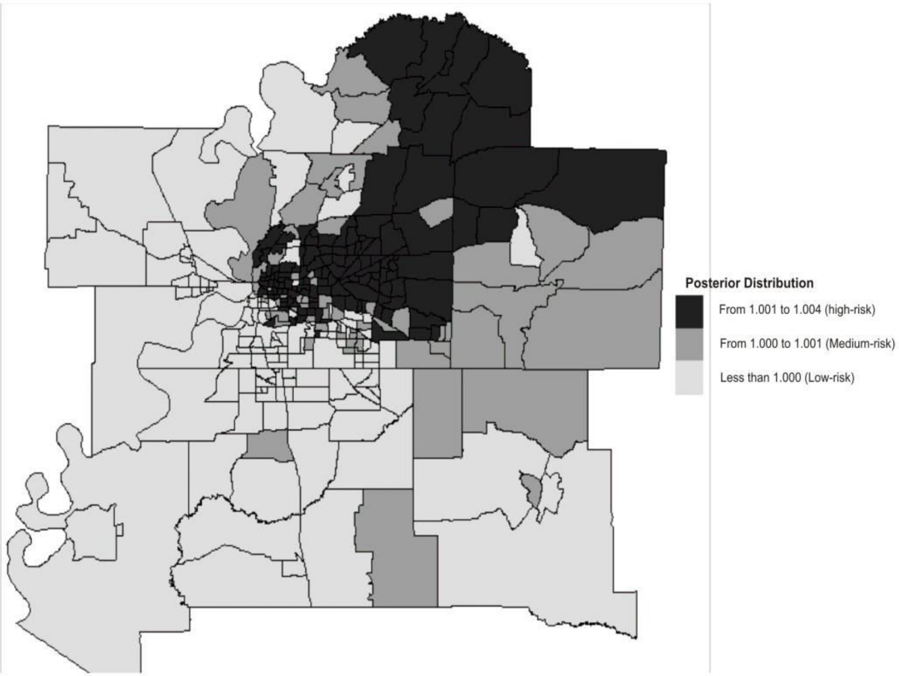


Figure 5a. Spatial main effects. The posterior distribution in the Bayesian spatiotemporal asthma disease-mapping shows 40% of the geographical areas/census tracts in the MMA region at high-risk, 20% medium-risk areas, and 40% low-risk areas.

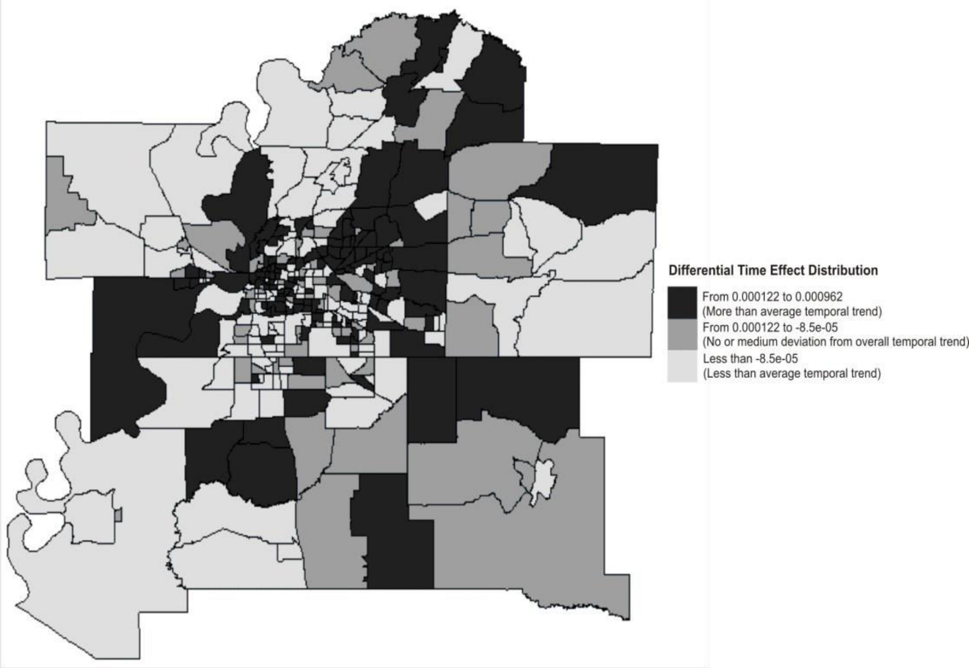
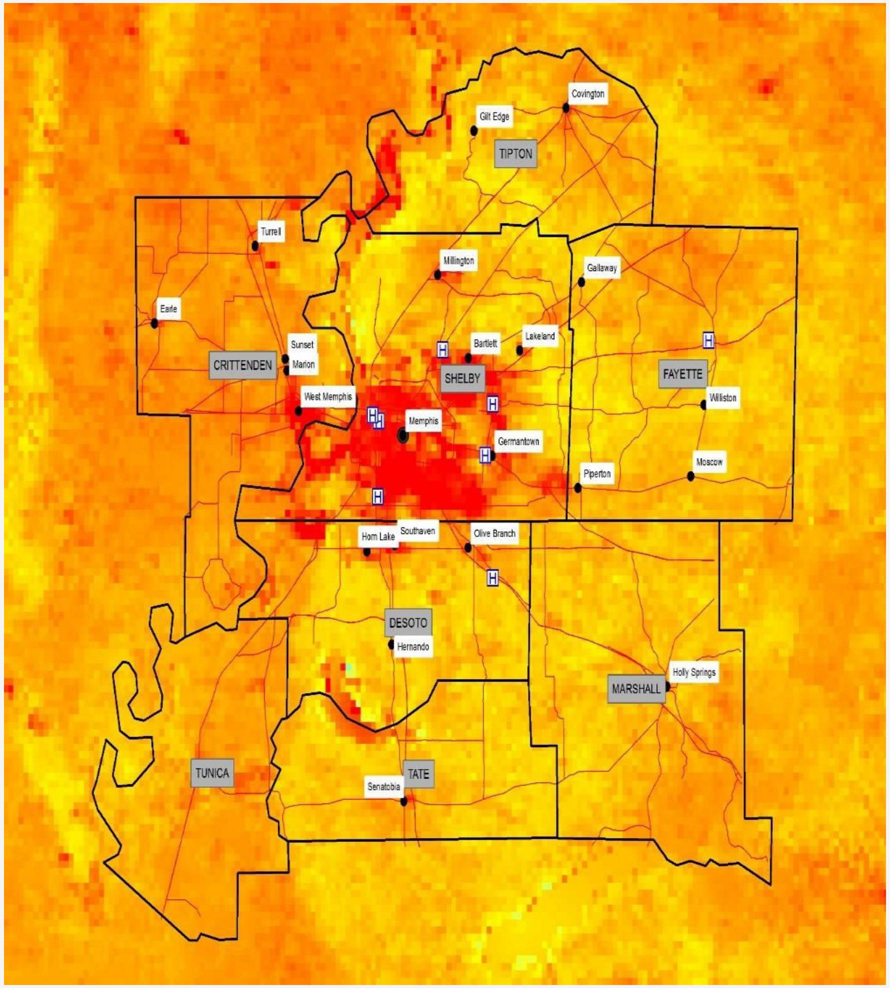
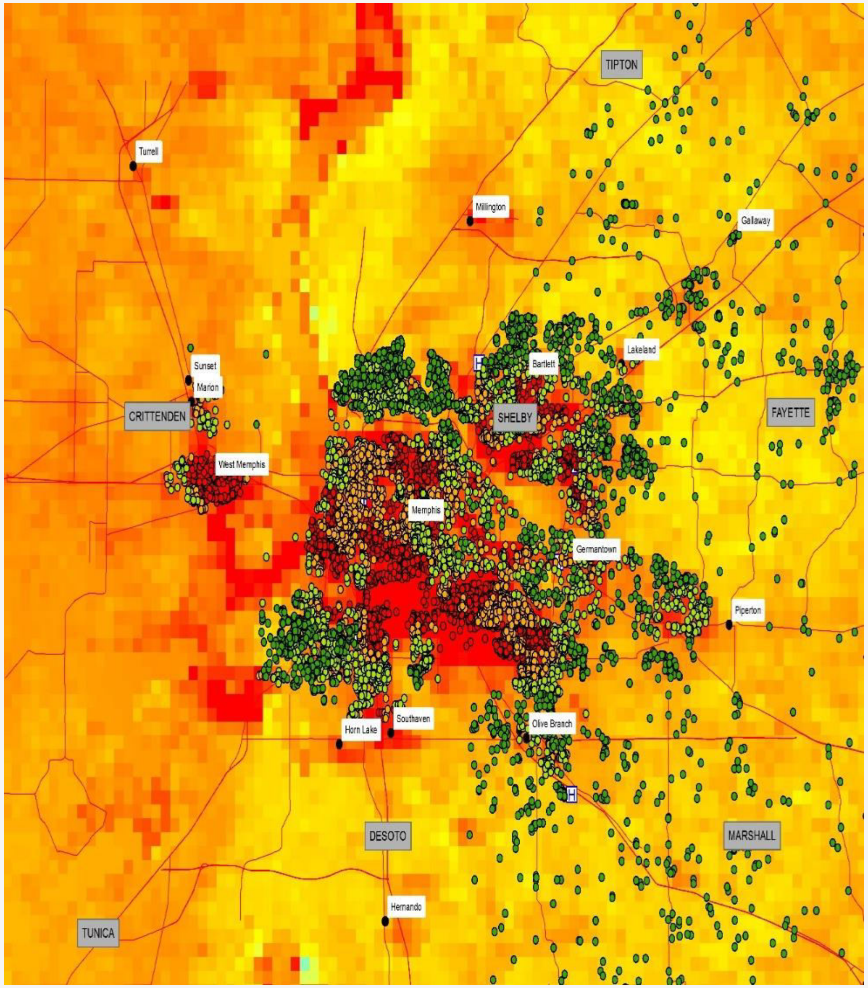


Figure 5b. Differential time effect 2005-2014. The effect in the Bayesian spatiotemporal asthma disease-mapping model shows 40% of the geographical areas/census tracts in the MMA region had a more than average temporal trend, 20% no or medium deviation from the overall temporal trend, and 40% a less than average temporal trend in disease risk.



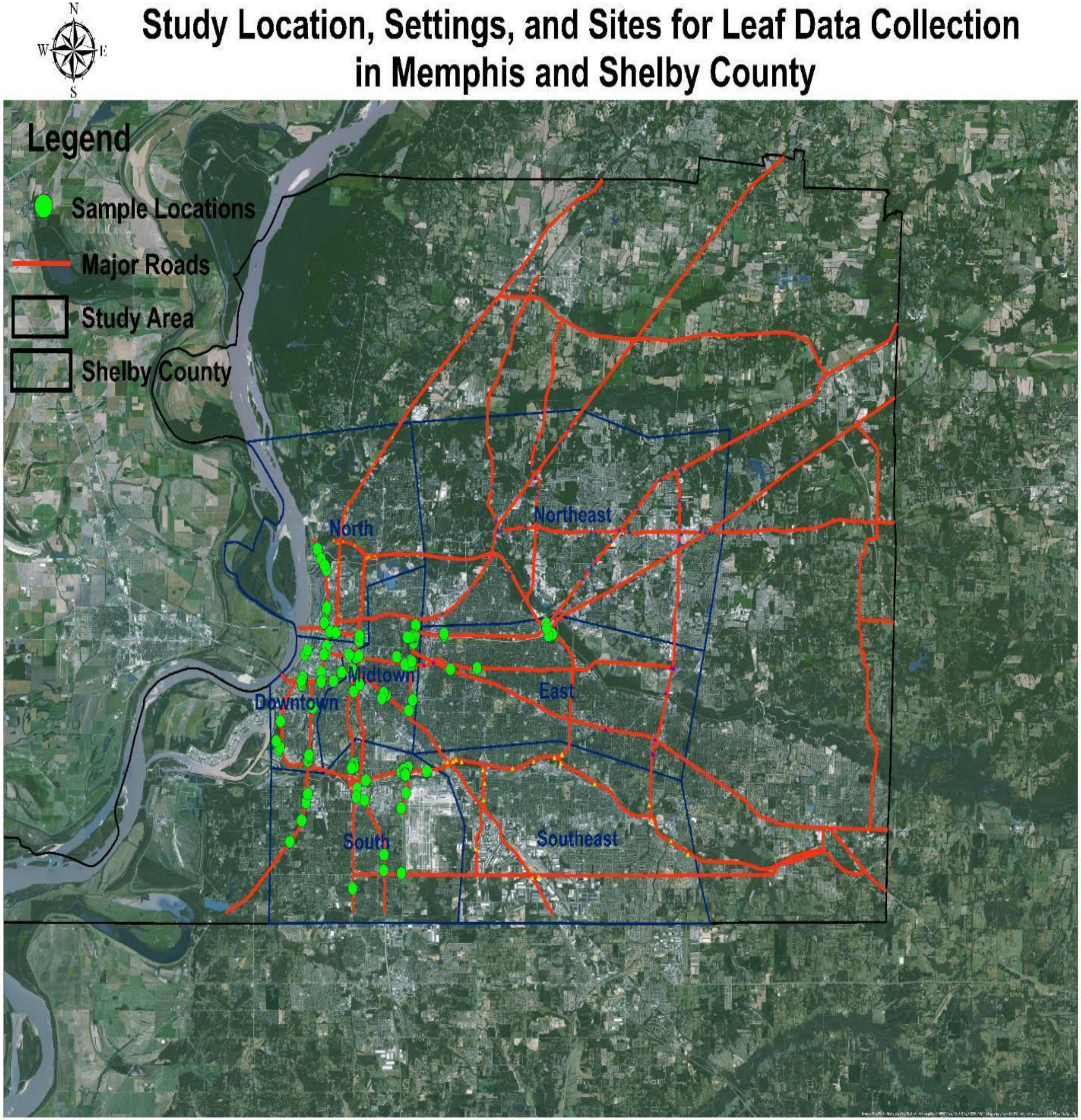
Map 7: High-resolution Satellite-derived average $PM_{2.5}$ Patterns between 2004 and 2008

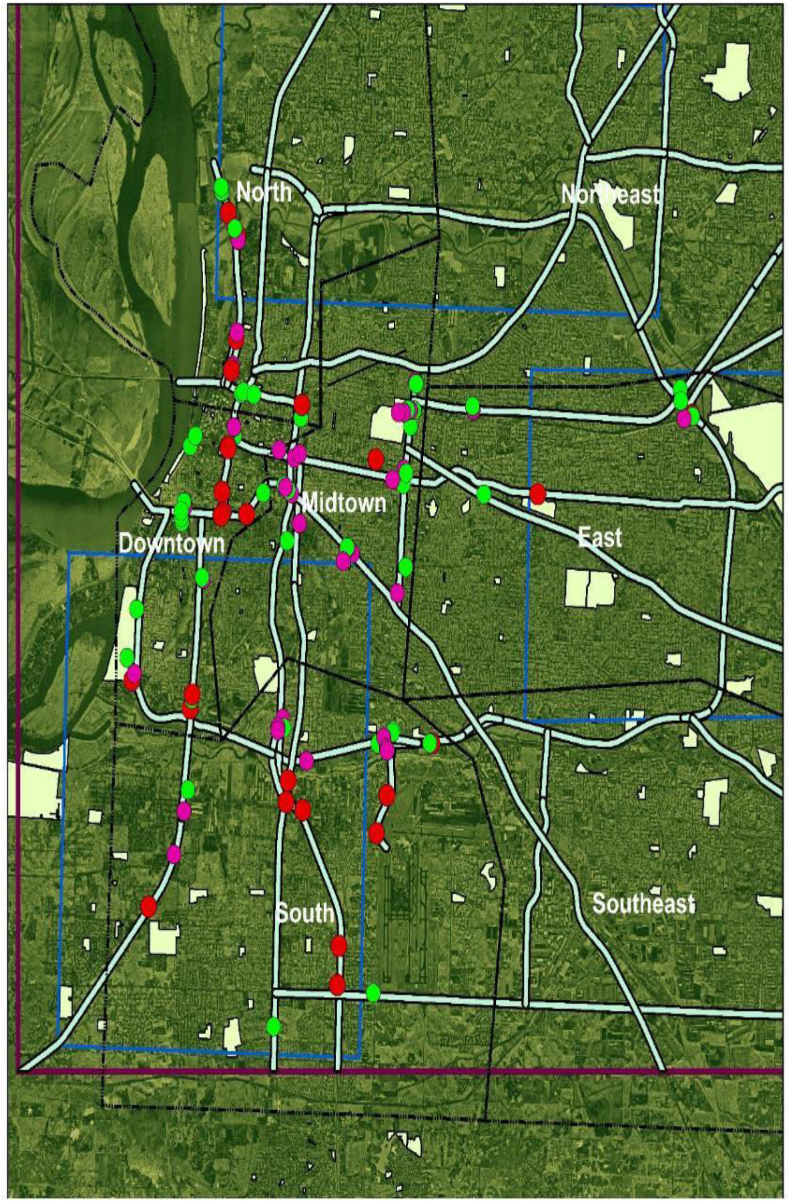


Map 8: $PM_{2.5}$ Statistically Linked to Asthma ER Visits and Hospitalization Encounters

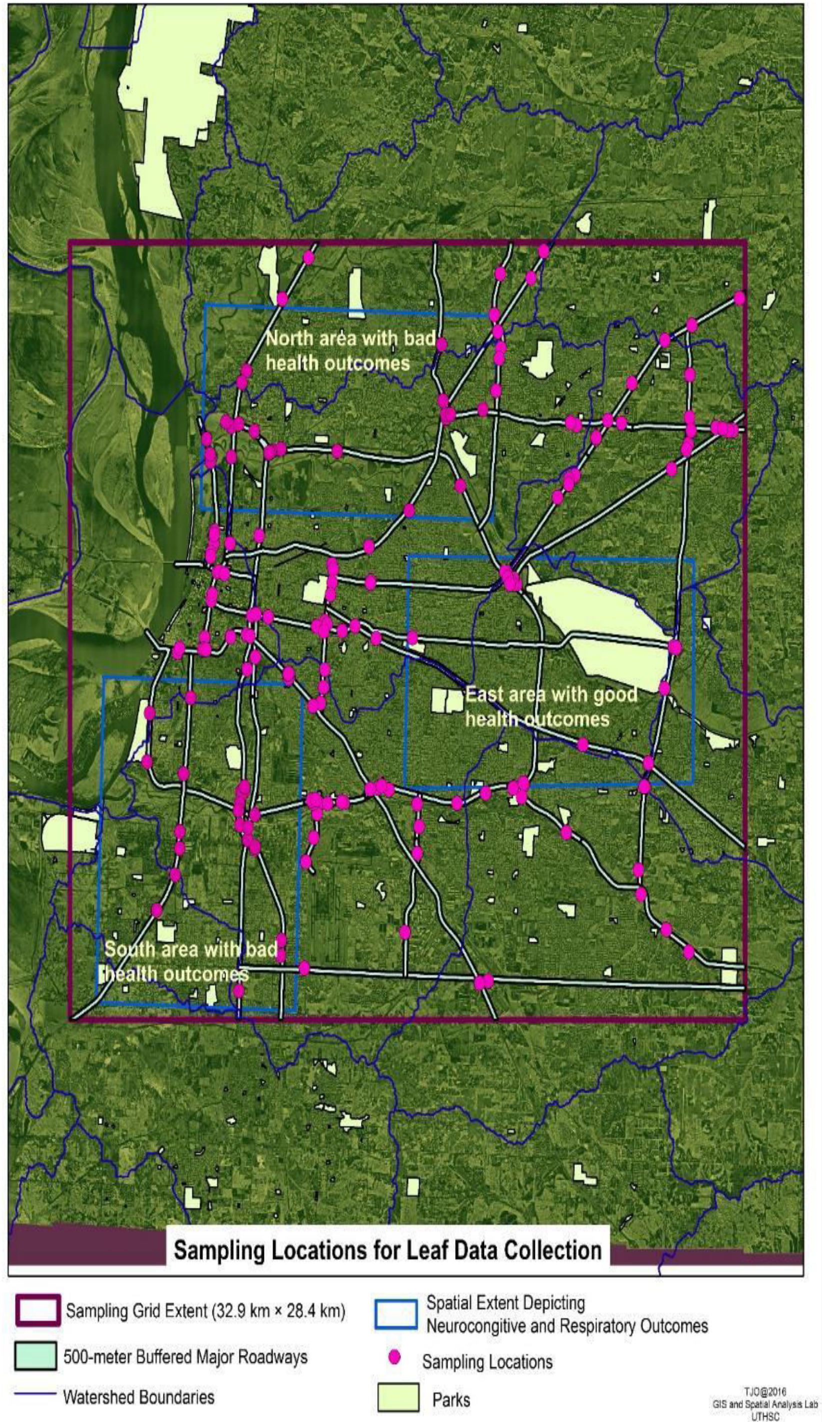
Data Source: Asthma Data was obtained from the Methodist Le Bonheur Healthcare System.

Study Location, Settings, and Sites for Leaf Data Collection in Memphis and Shelby County



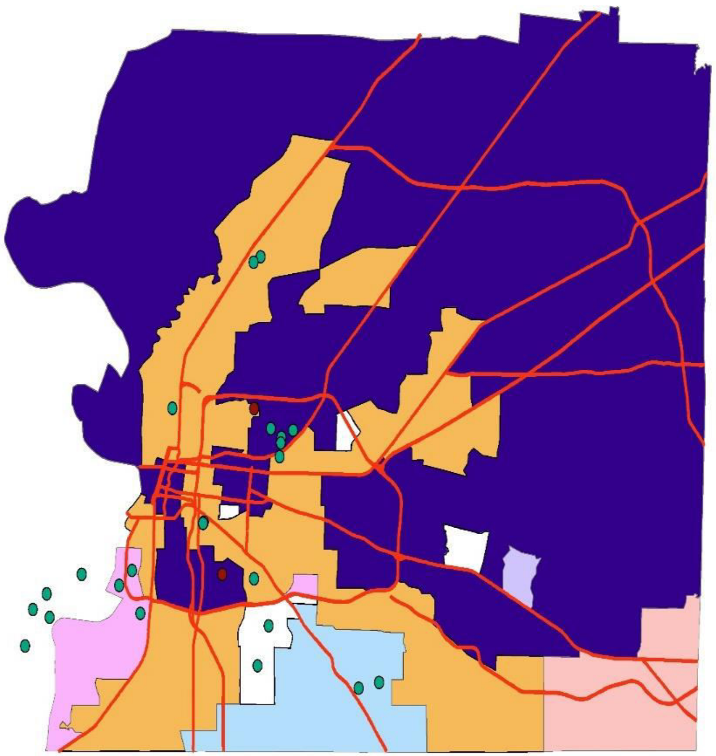


Spatial distribution of g-value of environmentally persistent free radicals (EPFRs)/PM2.5

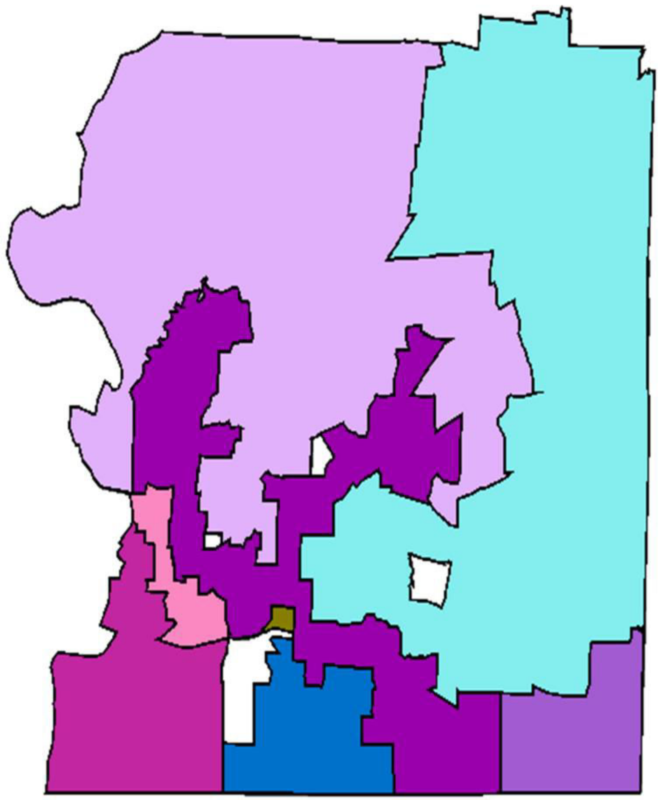


North area with bad health outcomes
East area with good health outcomes
South area with bad health outcomes
Sampling Locations for Leaf Data Collection
Sampling Grid Extent (32.9 km × 28.4 km)
500-meter Buffered Major Roadways
Watershed Boundaries
Spatial Extent Depicting Neurocongitive and Respiratory Outcomes
Sampling Locations
Parks
TJO@2016
GIS and Spatial Analysis Lab
UTHSC

A regionalization and partitioning algorithm searches and finds similar structures in a multidimensional dataset. The algorithm utilizes a powerful self-organizing map, adaptive kernel, and clustering methods with spatial contiguity constraints to identify areas with a similar profile.



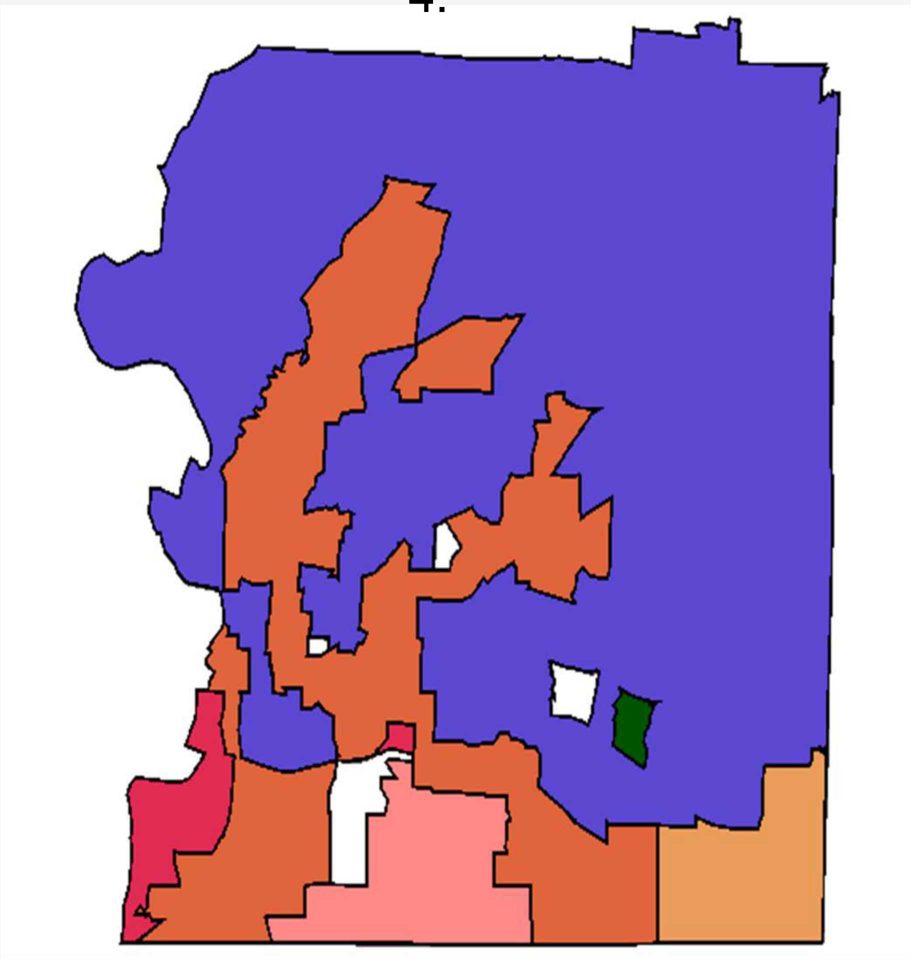
Map 1: Areas in Shelby County with Similar VOCs (71 compounds) overlaid with major roadways and individual pollution sources.



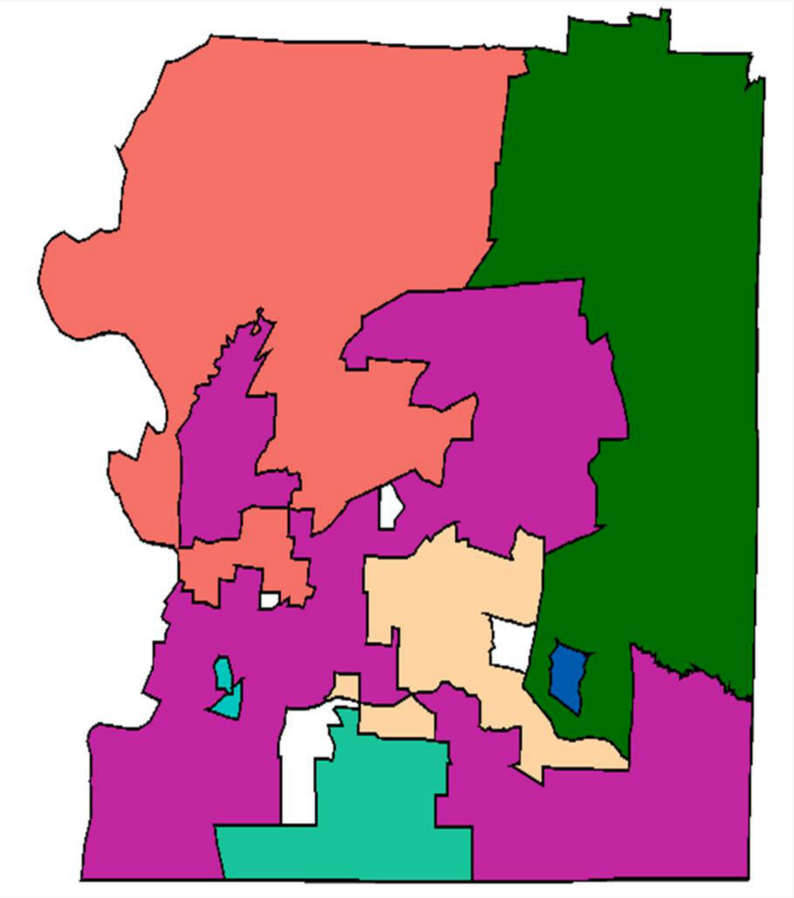
Map 2: Areas in Shelby County with Similar VOCs-BTEX Profile. BTEX = benzene, toluene, ethyl benzene, and o-, m- and p-xylene.

Data Source: The 2014 VOC data (112 monitoring sites) was obtained from Dr. Chunrong Jia, School of Public Health, University of Memphis.

Note the similarity in spatial patterns in Maps 2, 3, and 4.

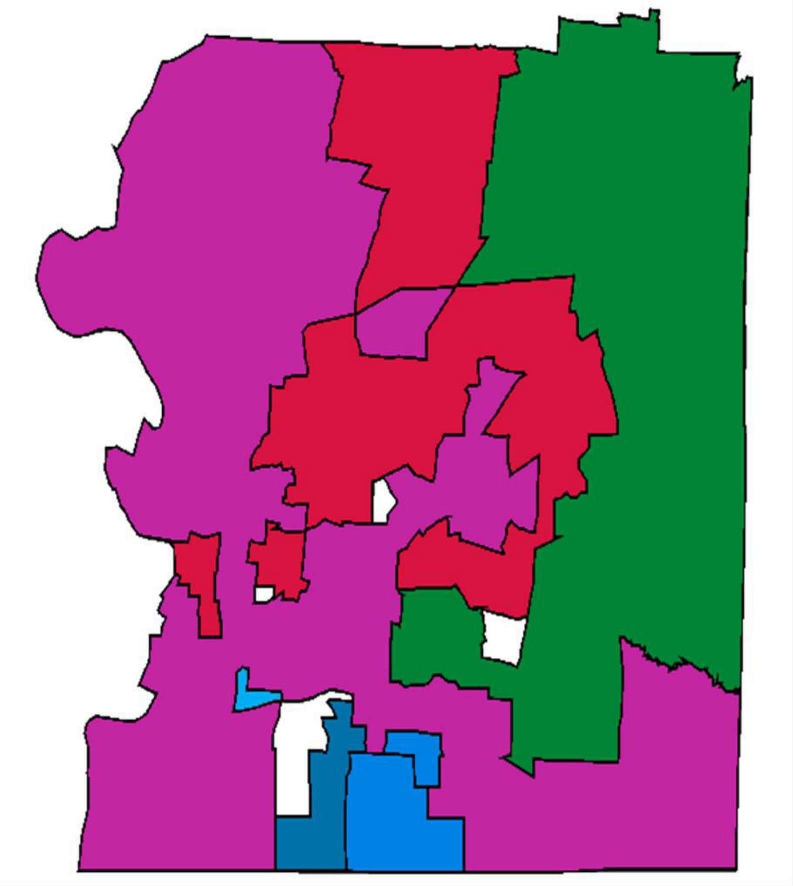


Map 3: Areas in Shelby County with Similar VOCs (71 compounds) minus background information.

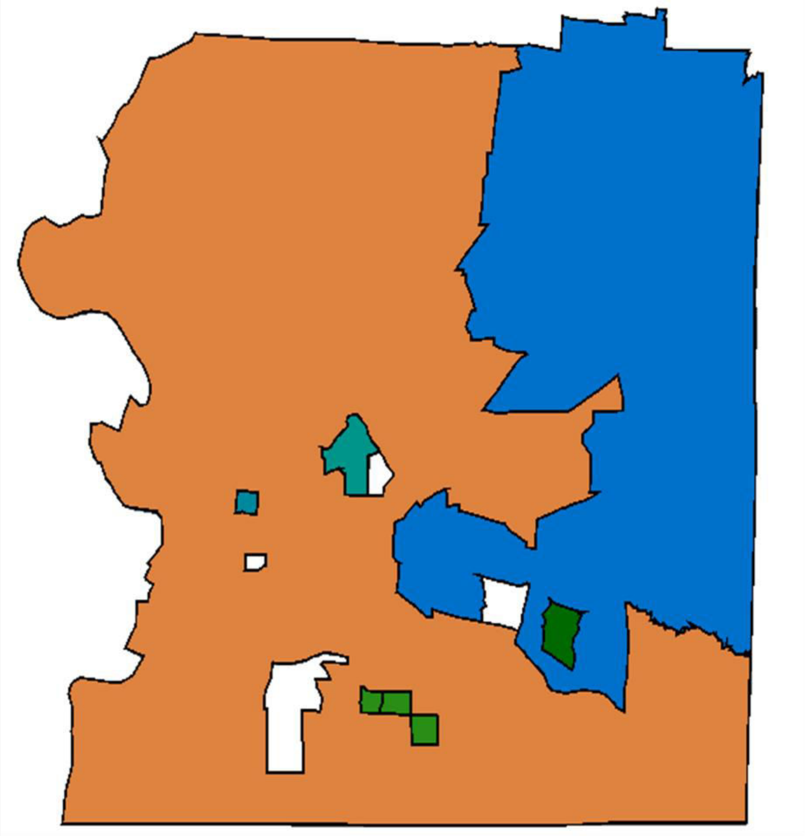


Map 4: Areas in Shelby County with Similar Highest Average VOCs-EAAN Profile. EAAN = Ethanol, Acetone, Allyl chloride, and Naphthalene.

Data Source: The 2014 VOC data (112 monitoring sites) was obtained from Dr. Chunrong Jia, School of Public Health, University of Memphis.



Map 5: Areas in Shelby County with Similar VOCs with combustion-BTT related compounds. BTT= benzene, toluene, and 1,2,4-trimethylbenzene.



Map 6: Areas in Shelby County with Similar VOC compounds with high cancer potency-BTHC profile. BTHC = Benzyl chloride, 1,1,2,2-Tetrachloroethane, Hexachloro-1,3-butadiene, and Chloroform Profile.

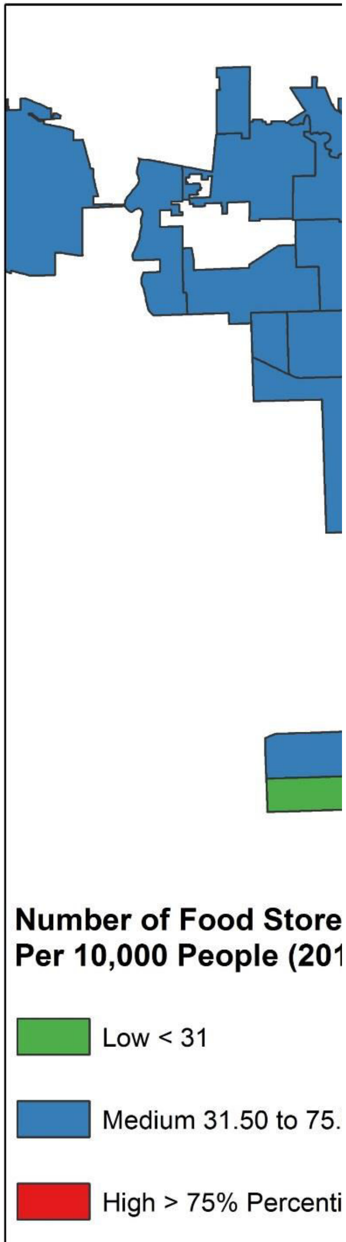
Data Source: The 2014 VOC data (112 monitoring sites) was obtained from Dr. Chunrong Jia, School of Public Health, University of Memphis.

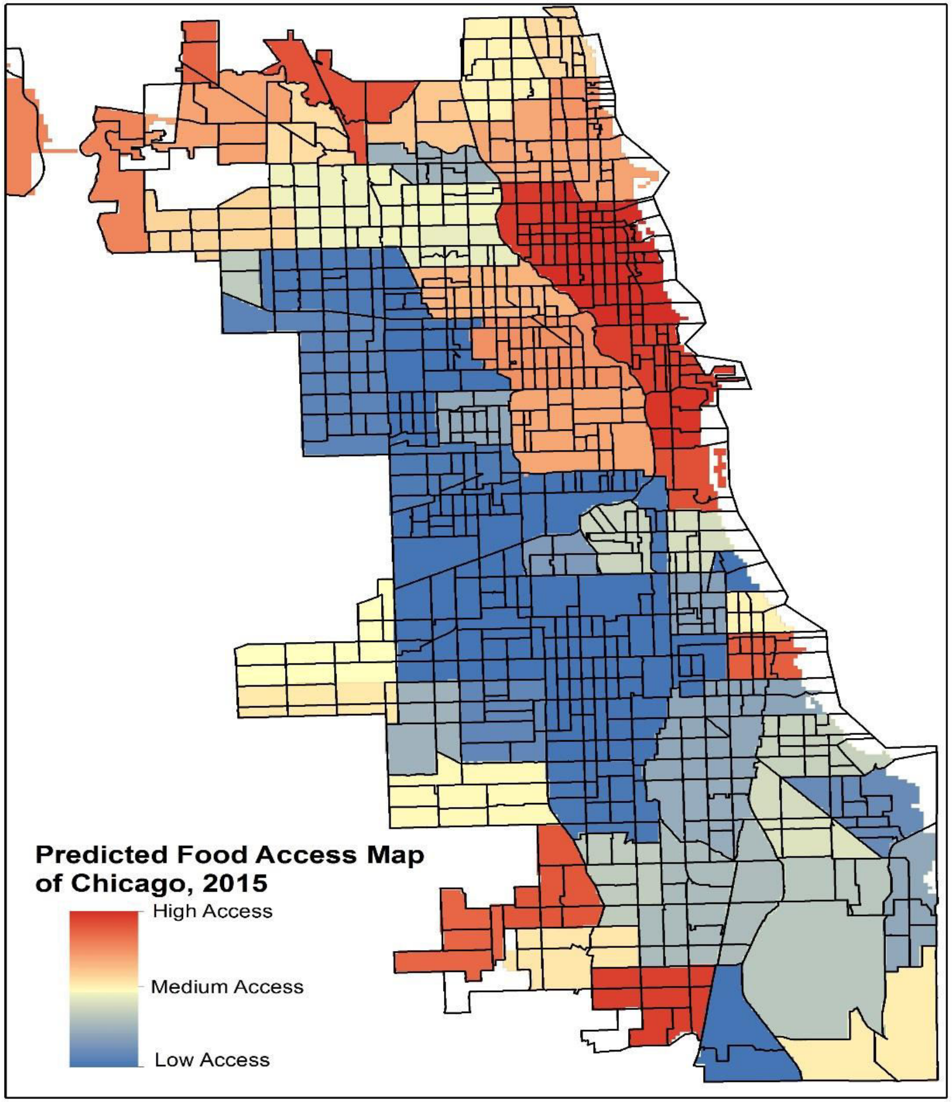
Example Application III: Understanding the food environment in low- and high-income settings

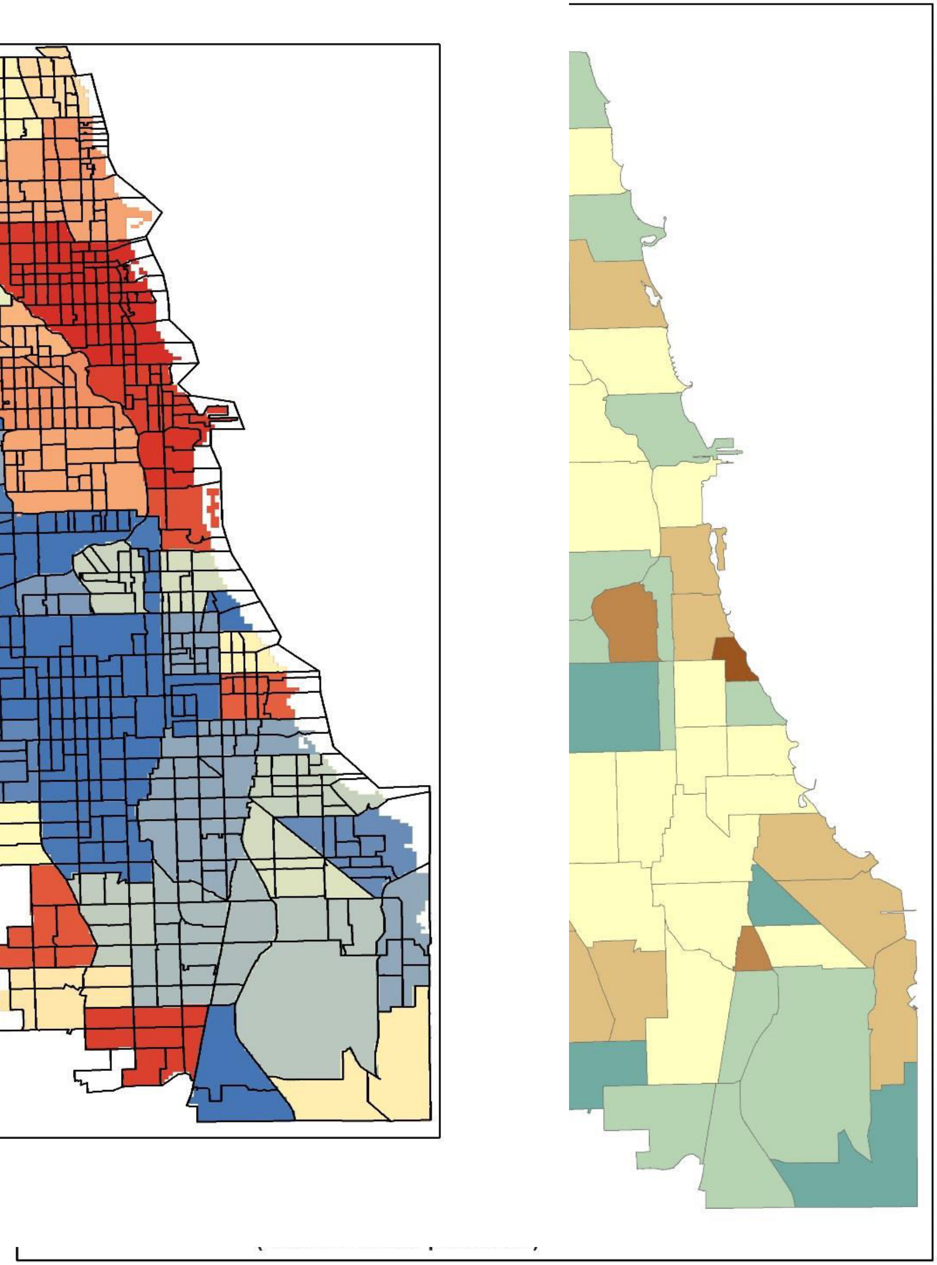
Inspiration: ‘A scientist’s work is never complete, always evolving, learning, and investigating better ideas/methods in pursuit of the scientific truth and a fine language to communicate the truth to broad audience’

Rationale and Select Literature

- Need to have a strong basis when designing place-based, targeted health interventions
- H_0: There are no significant differences in food access between low- and high-income settings in the city of Chicago, IL.
- Food access measures:
    1. Spatial access
    2. Temporal access
    3. Spatiotemporal access
    4. Other determinants of food choices and diet quality: food prices, food and nutrition assistance programs, and community socioecological characteristics. Most focus on 1 & 4 so need to…
- Include 2 & 3 measures to study the food environment
- Why: access to healthy food matters if we are to succeed in eliminating health disparities/achieve health equities.







Understanding the urban food environment in low- and high-income settings in Chicago

Concluding Remarks & Future Directions

- Incorporate *g4* and *g5* objectives in the model
- Test our geomasking algorithm on a wide spectrum of cohorts with a diverse activity pattern and environmental exposure over a life course and make the algorithm more robust
- Investigate strategies (e.g. decision science, uncertainty visualization methods) for incorporating uncertainty when reporting and visualizing post-mask health outcome data.
- Some examples: paired maps, bivariate and multivariate maps, automated systems, dashboards, and interpretive uncertainty

Concluding Remarks & Future Directions

- The art of human progress vs. technological advances. Should the race be framed with political-social notions or technological advances?
- How do we arrest biodiversity decline, especially in Africa? We have situation of a depleted or rapidly declining environment.
- I think: Data science can facilitate the production of data and new knowledge that can be used to support policy development and the design of most appropriate interventions.

Acknowledgments

- Support by the UT-Knoxville and UTHSC/Research Center on Health Disparities, Equity and the Exposome
- The Spatial Analytics and Informatics Core Research Team
- NSF award# 224702 CNS-0855221 for SIHPC
- IRB Human Subjects Approval Committee