Spatial Statistical Analysis Methods

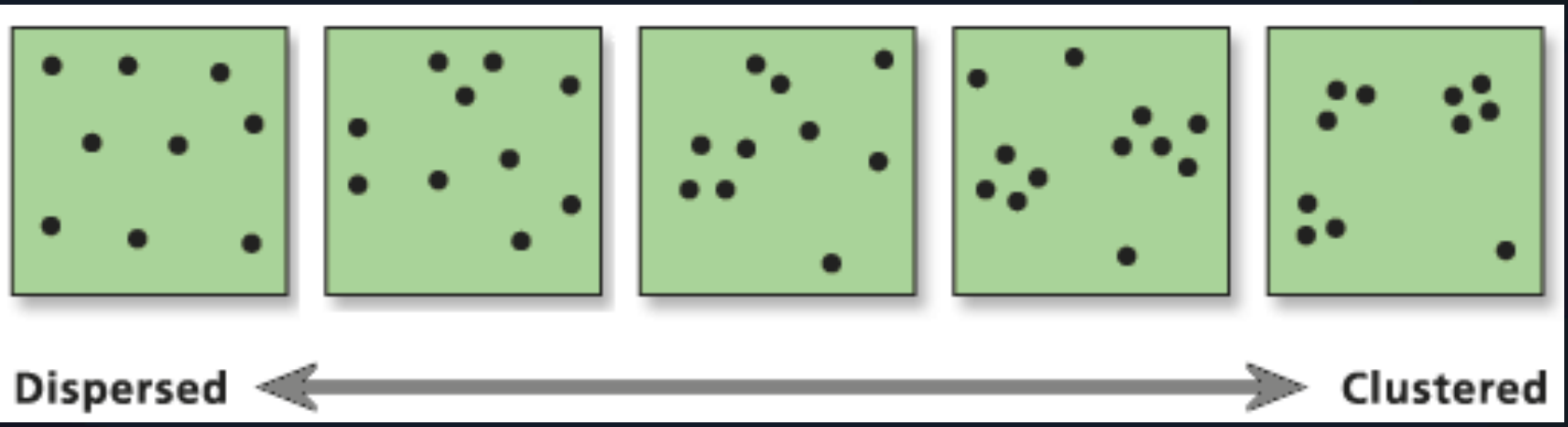
Jennifer A Miller, UT Dept. of Geography and the Environment

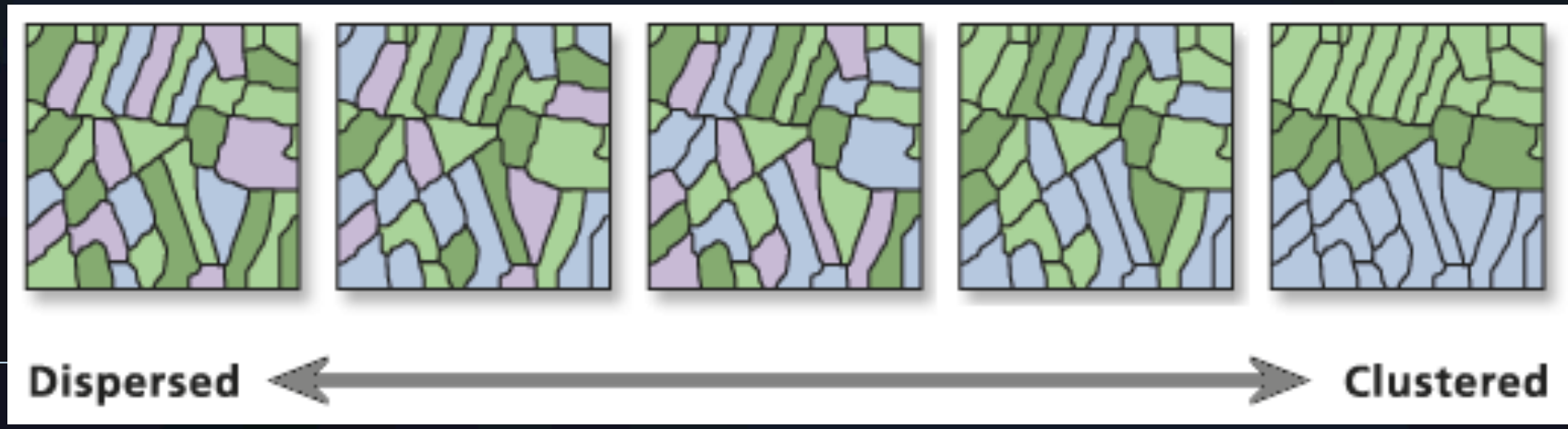
Overview

- Quantifying spatial patterns (spatial autocorrelation analysis)
- Exploring potential explanatory factors (regression)
  - Global & local
  - Effects of scale

Spatial autocorrelation

- Points (location) vs attributes

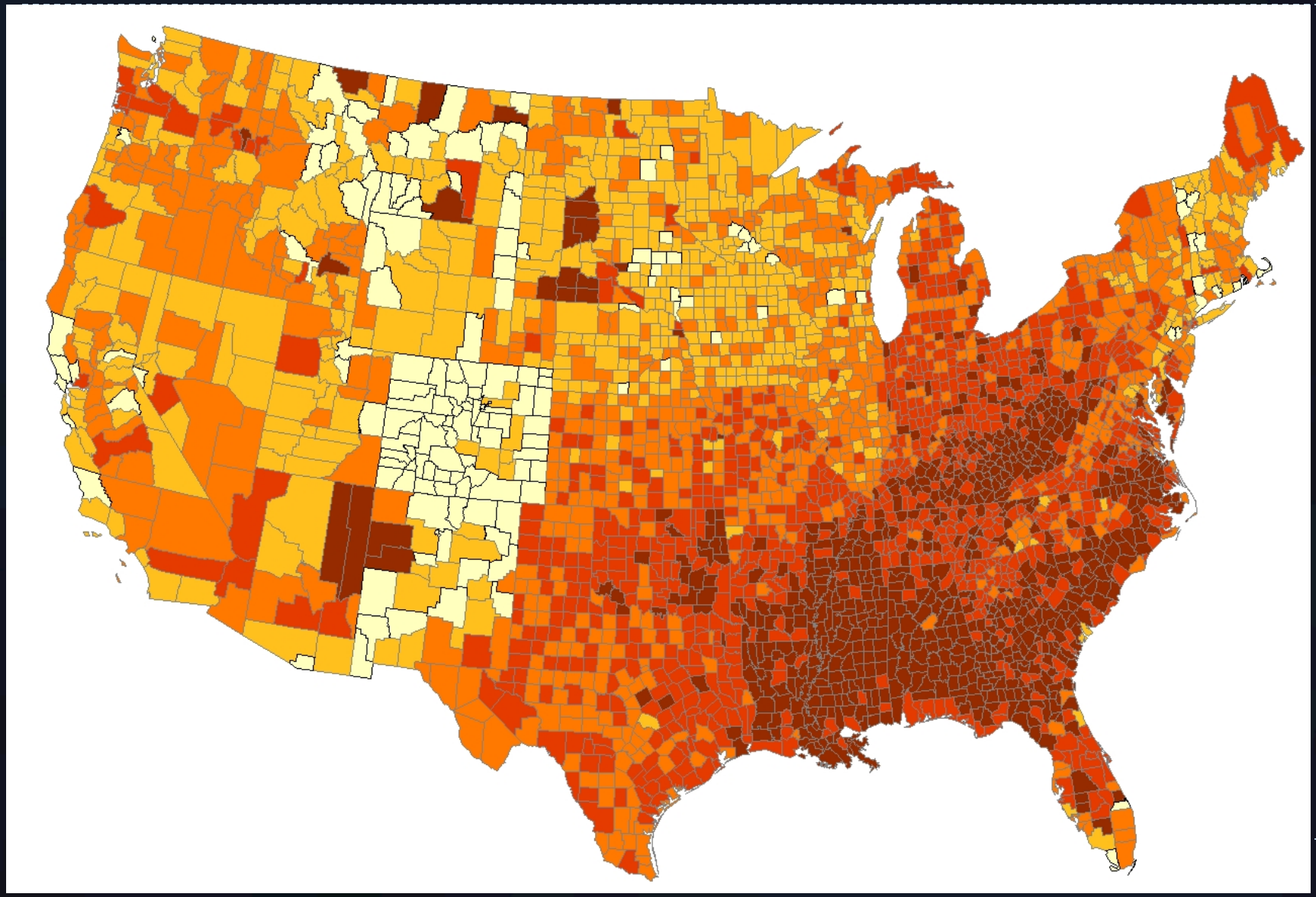




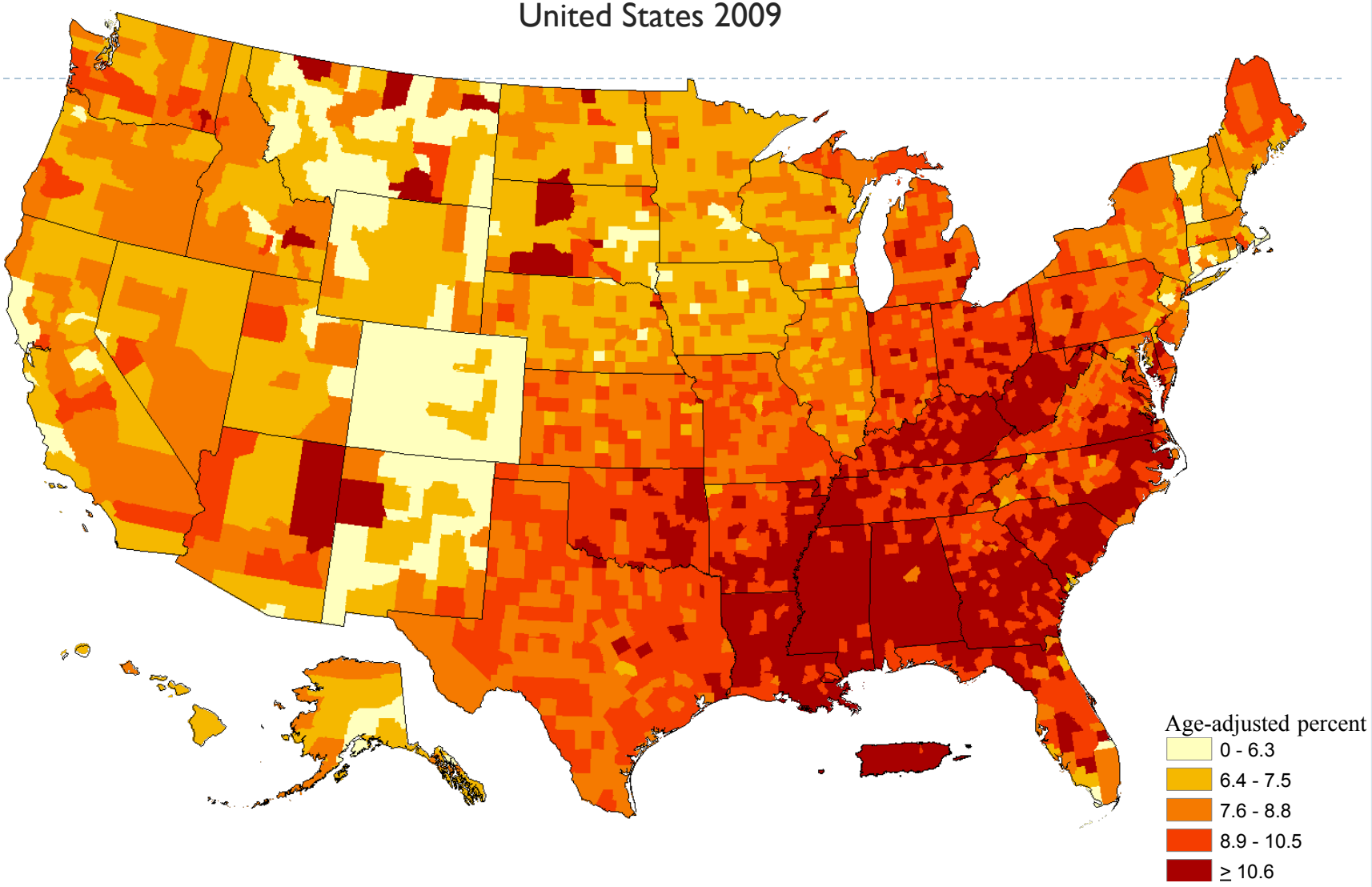
Spatial analysis: attributes

- Measuring spatial autocorrelation:
  - Based on both feature locations and feature values simultaneously.
  - **(x,y, z), where z is measured at interval or ratio level.**

Why measure SAC?



County-level Estimates of Diagnosed Diabetes among Adults aged ≥ 20 years: United States 2009



American Journal of Preventive Medicine

Volume 40, Issue 4, April 2011, Pages 434–439

Research article

Geographic Distribution of Diagnosed Diabetes in the U.S.: A Diabetes Belt

Lawrence E. Barker, PhD, Karen A. Kirtland, PhD, Edward W. Gregg, PhD, Linda S. Geiss, MA, Theodore J. Thompson, MS

CDC, Atlanta, Georgia

Background

The American "stroke belt" has contribute
diabetes have not been as specifically cha

Purpose

This study identifies a geographically c
diabetes is especially high, called the "diab

Articles

The Geography of Stroke Mortality in the United States and the Concept of a Stroke Belt

Douglas J. Lanska, MD, MS; Lewis H. Kuller, MD, DrPH

Author Affiliations

Correspondence to Douglas J. Lanska, MD, Department of Neurology, Rm E124, Kentucky Clinic, University of Kentucky, Lexington, KY 40536-0284. E-mail djlansva@ukcc.uky.edu.

Key Words:

- cerebrovascular diseases
- epidemiology
- risk factors
- geography
- mortality

Since at least 1940 there has been a
variation in stroke mortality rates within
reported in the Southeast, and particula
low rates are reported in the Mountai
Atlantic coast.[1] [2] [3] These general pa
observed for both sexes and for whites
been consistently declining in all geogr
over this interval.[1] Comparable age-adj
earlier periods are not available because
Death Registration Area, which did not i
until 1933,[4] [5] [6] and (2) US mortality da
sex, cause of death, and state in
mortality rates for whites suggest that
mortality was not in place in 1920; ins

Stroke.
1995;26:1145-1149
doi: 10.1161/01.STR.26.7.1145

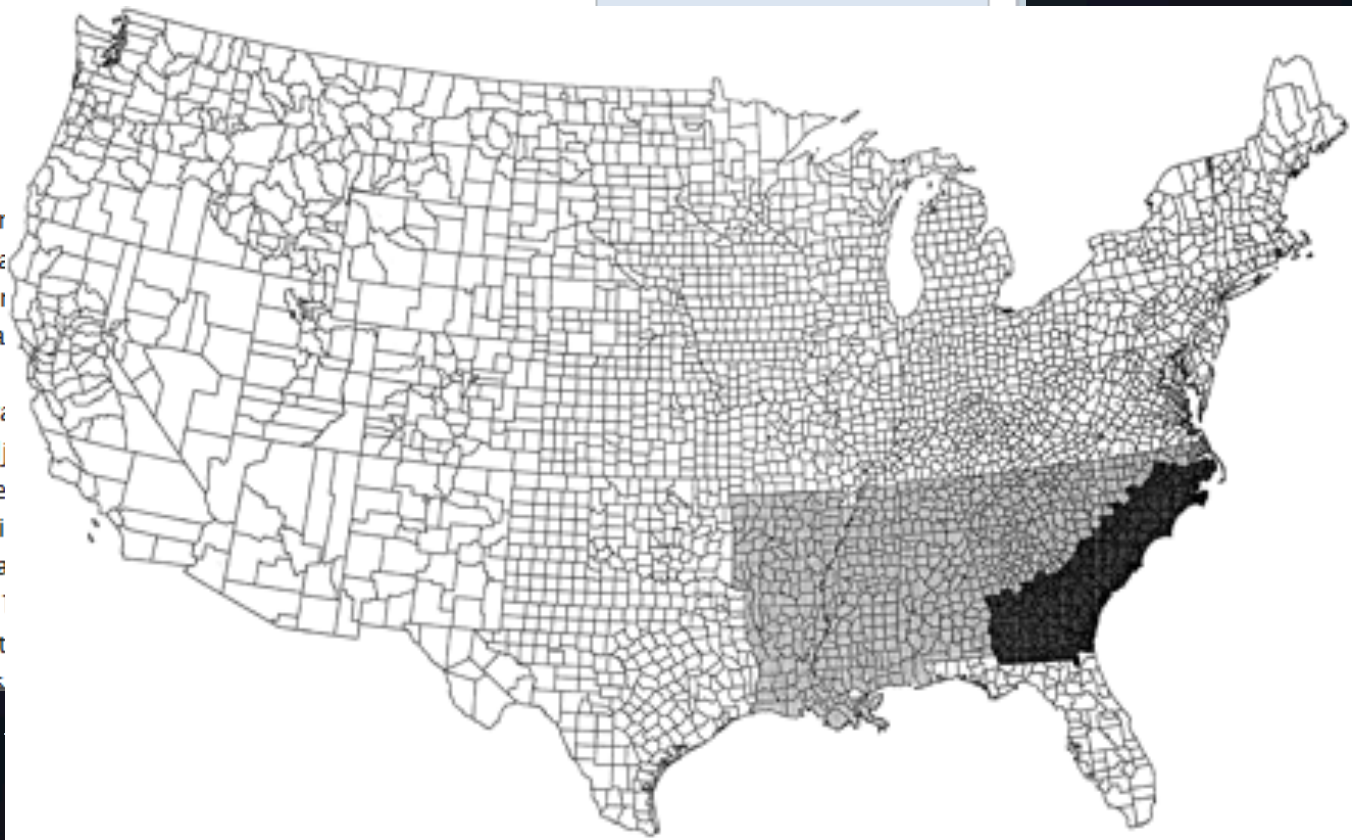
- Extract *Free*
- » Full Text

Classifications

Articles

Services

- E-mail this article to a friend
- Alert me when this article



The American Journal of Cardiology

Volume 107, Issue 6, 15 March 2011, Pages 935–937

Heart failure

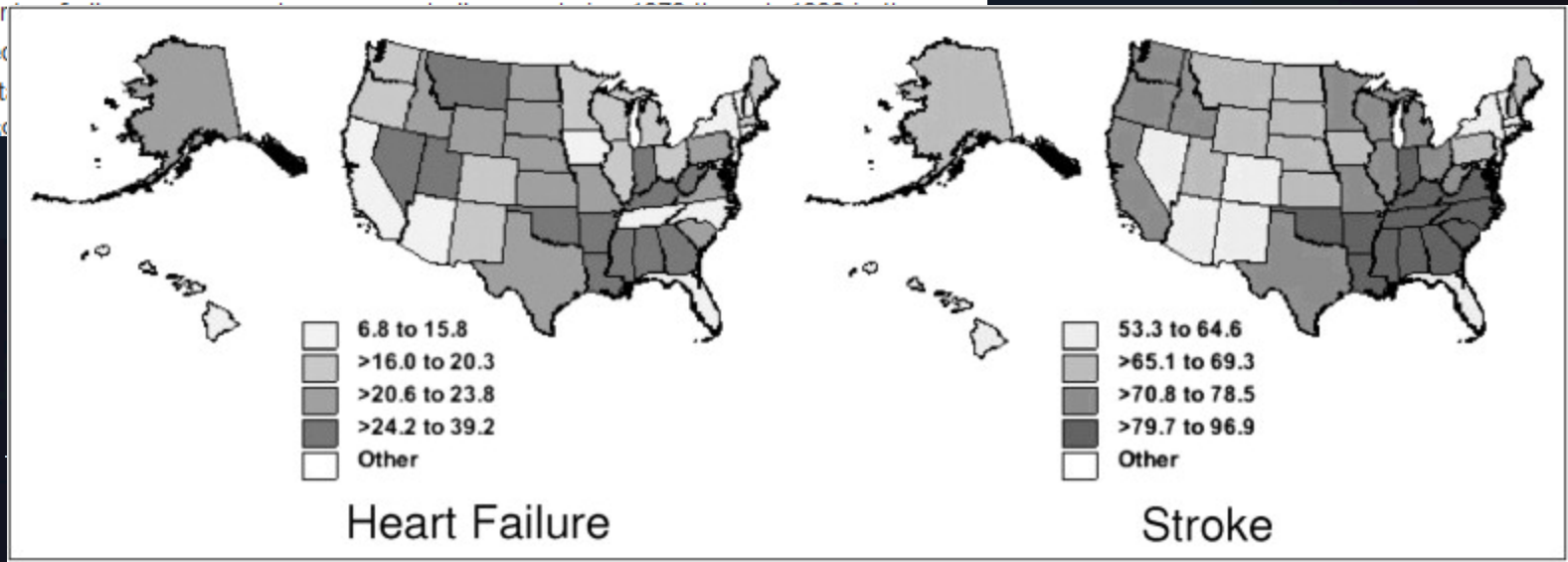
Evidence of a "Heart Failure Belt" in the Southeastern United States

Marjan Mujib, MBBS, MPH[a], Yan Zhang, MS, MSPH[a], Margaret A. Feller, MPH[a], Ali Ahmed, MD, MPH[a, b]

[a] University of Alabama at Birmingham, Birmingham, Alabama

[b] Veterans Affairs Medical Center, Birmingham, Alabama

The southeastern region of the United States is known as the "stroke belt" because of excess stroke mortality in this region compared to the rest of the country. However, whether a similar geographic variation in heart failure mortality exists is unknown. Using the Center for Disease Control and Prevention Wide-ranging Online Data for Epidemiologic Research publicly available compressed mortality data files and 2000 United States population as the standard, we estimated age-adjusted heart failure and stroke mortality rates per 100,000 for patien
United States and mapped
contiguous southeastern st
31.0/100,000, which was 6



Cancer Epidemiology

Volume 36, Issue 5, October 2012, Pages 436–438

Short communication

Implementing a weighted spatial smoothing algorithm to identify a lung cancer belt in the United States

David Blackley[a], Shimin Zheng[a], Winn Ketchum[b]

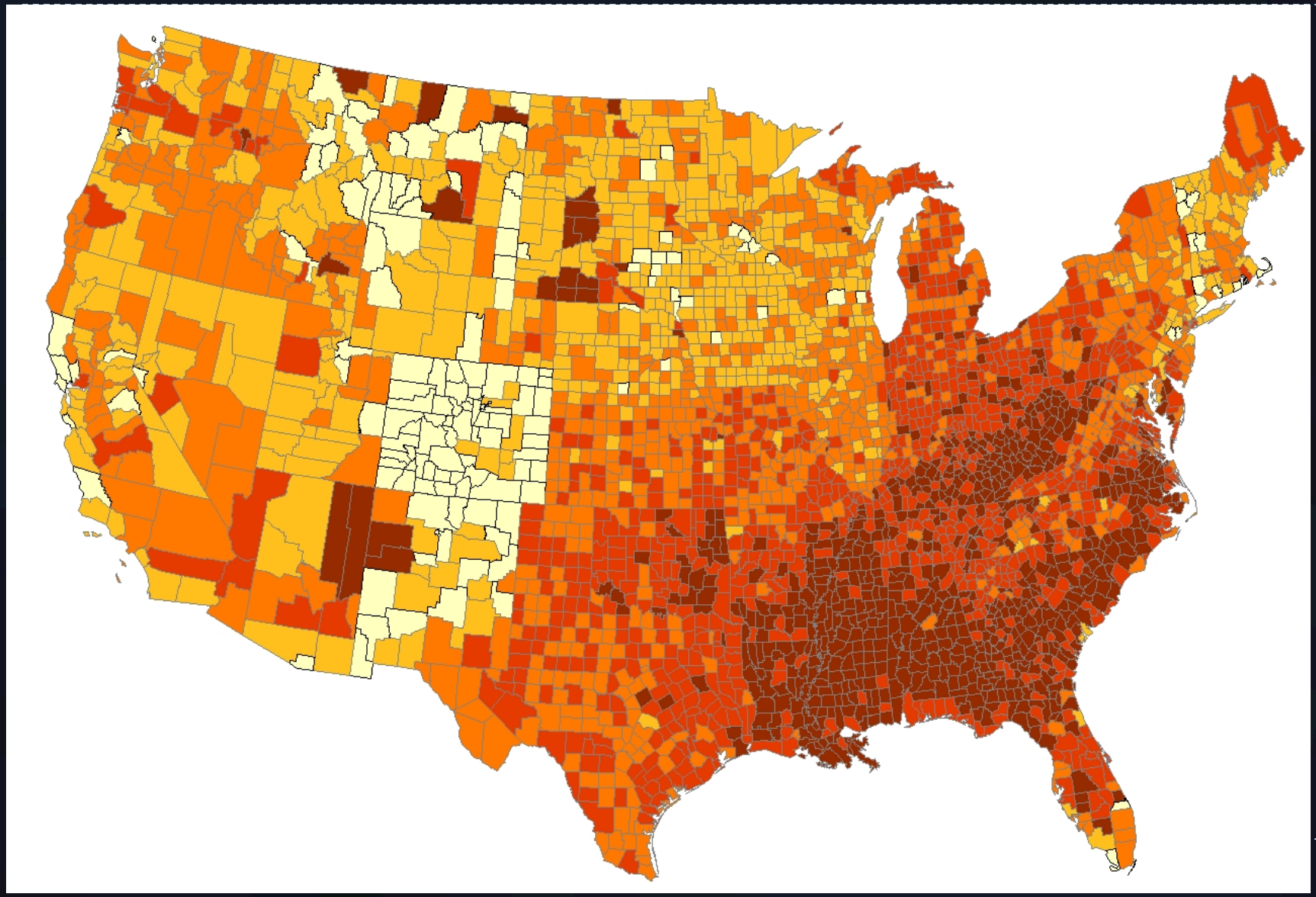
[a] Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Box 70259, Johnson City, TN 37614, USA

[b] Department of Geosciences, College of Arts and Sciences, East Tennessee State University, Box 70357, Johnson City, TN 37614, USA

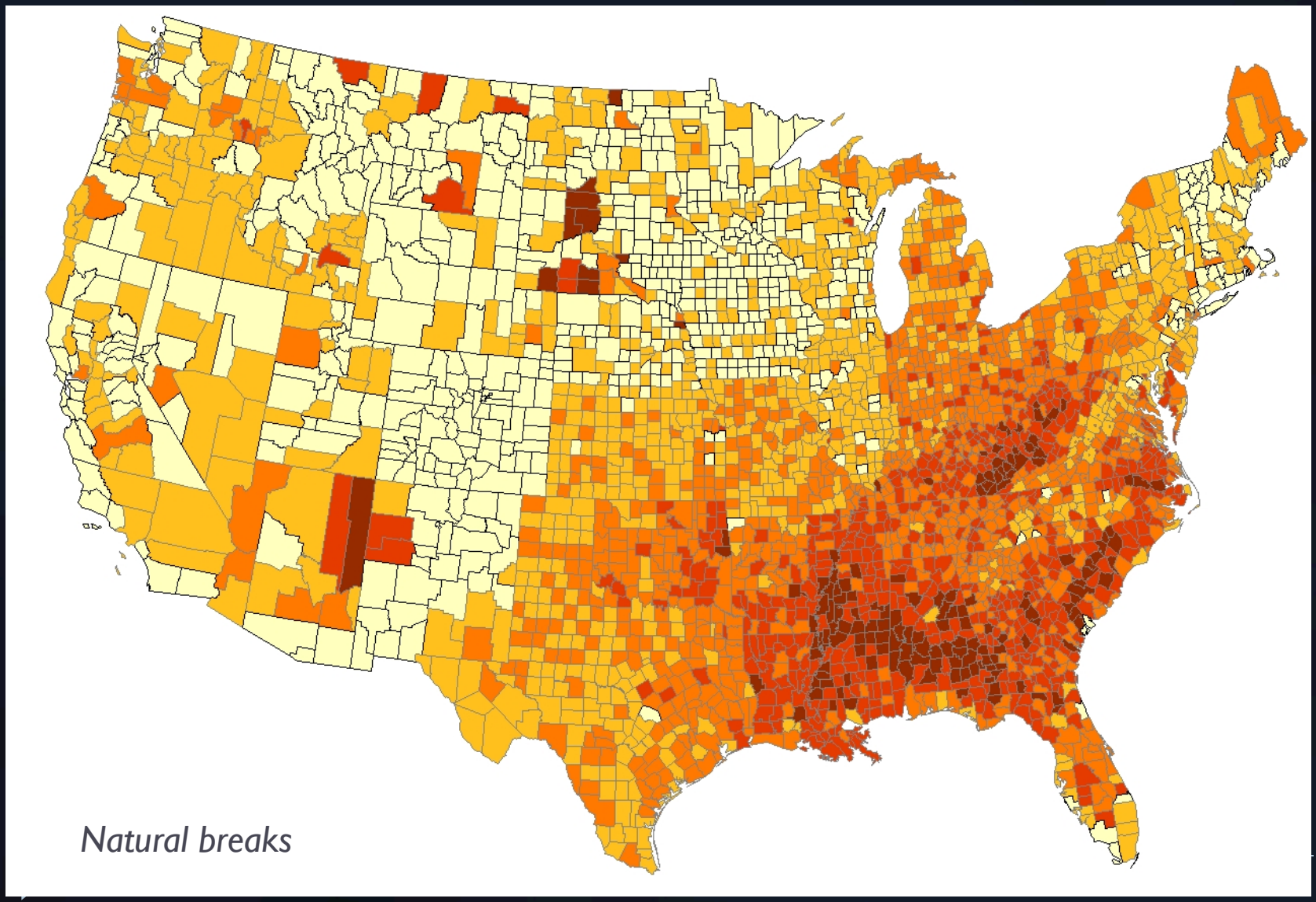
Abstract

Lung cancer is the leading cause of cancer death in the U
preventable. We use a spatial smoothing algorithm to ider
mortality, primarily in the Southeast, which we call a lung can
conveying patterns of high incidence or mortality; formally
increased public dialogue and more focused research. Pub
population lung cancer data with this information to help inform

Why measure SAC?

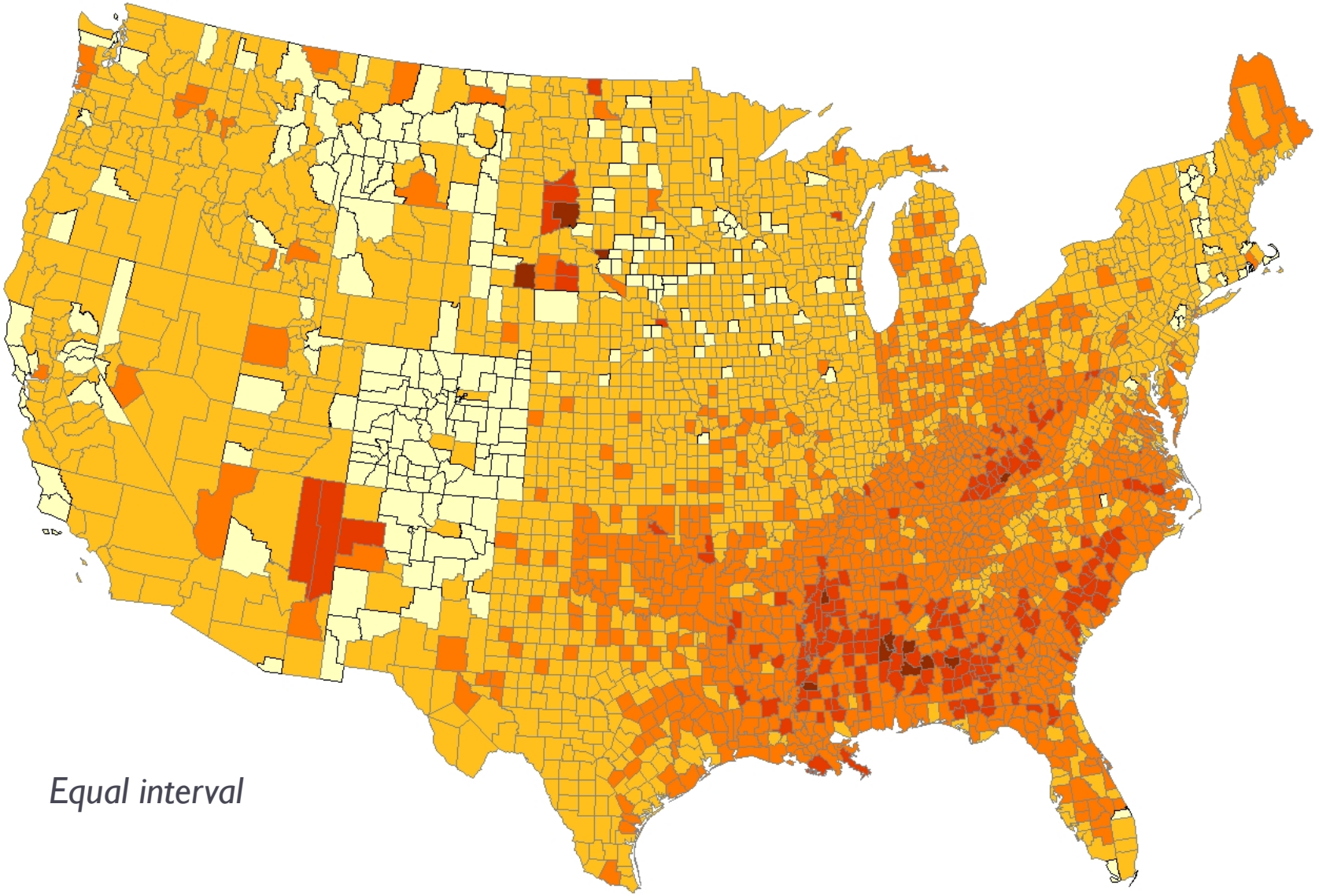


Why measure SAC?



Natural breaks

Why measure SAC?



Equal interval

Measuring Spatial Autocorrelation

- How similar or different are attribute values relative to associated geographic locations?
  - Comparison of attribute values?

  - **Comparison of geographic locations?**
    - Concepts of distance, adjacency, interaction, neighborhood
    - Spatial relationship between **all pairs** of locations
      - Spatial weights matrix- contiguity, distance, threshold, nearest neighbors, etc.
      - **W** represents a hypothesis about the spatial structure of phenomena under study

Moran's I

$$I = \frac{n}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$$

Moran's *I*

Covariance:
If *i* and *j* are on same side of mean = +
If *i* and *j* are on different side of mean = -

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Moran's I

Covariance weighted by spatial weights matrix

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Moran's *I*

Normalize *I* relative to n, spatial relationships, and range of values in *y*

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Moran's *I*

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

- Most often used with polygons with interval/ratio data
- Based on the product of 2 polygons' differences from the overall mean ("how much do they **vary together?**") (http://www.spatialanalysisonline.com/output/html/MoranlandGearyC.html)
  - The larger the value of *I*, the stronger the spatial autocorrelation.
    - *I* is not strictly limited to the range -1 to 1.

Geoda

- https://geodacenter.asu.edu/

Moran's *I* example

- *Z* score?
- Expected index?

Randomization null hypothesis

- Exp(*I*)= -1/(*n* – 1),

n = number of features

- So Exp(*I*) is always _____
- As n increases, Exp (*I*) _____

Randomization null hypothesis

- Exp(I)= -1/(n – 1),

n = number of features

- So Exp(I) is always **negative**
- As n increases, Exp (I) **decreases**

Randomization null hypothesis

- Exp(I)= -1/(n – 1), n = number of features
  - -1/(5226-1) = -0.0002
- $Z = (I_o – I_E)/(SD_{I_E})$

Randomization null hypothesis

- Exp(I)= -1/(n – 1), n = number of features
  - -1/(5226-1) = -0.0002
- $Z = (I_o – I_E)/(SD_{I_E})$
- *(Use Bonferroni corrections, Monte Carlo permutations)*

The Open Diabetes Journal, 2012, *5*, 29-37

Spatial Clusters of County-Level Diagnosed Diabetes and Associated Risk Factors in the United States

Sundar S. Shrestha*, Karen A. Kirtland, Theodore J. Thompson, Lawrence Barker, Edward W. Gregg and Linda Geiss

Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Abstract: *Introduction*: We examined whether spatial clusters of county-level diagnosed diabetes prevalence exist in the United States and whether socioeconomic and diabetes risk factors were associated with these clusters.

Materials and Methods: We used estimated county-level age-adjusted data on diagnosed diabetes prevalence for adults in 3109 counties in the United States (2007 data). We identified four types of diabetes clusters based on spatial autocorrelations: high-prevalence counties with high-prevalence neighbors (High-High), low-prevalence counties with low-prevalence neighbors (Low-Low), low-prevalence counties with high-prevalence neighbors (Low-High), and high-prevalence counties with low-prevalence neighbors (High-Low). We then estimated relative risks for clusters being associated with several socioeconomic and diabetesrisk factors.

Results: Diabetes prevalence in 1551 counties was spatially associated ($p<0.05$) with prevalence in neighboring counties. The rate of obesity, physical inactivity, poverty, and the proportion of non-Hispanic blacks were associated with a county being in a High-High cluster versus being a non-cluster county (7% to 36% greater risk) or in a Low-Low cluster (13% to 67% greater risk). The percentage of non-Hispanic blacks was associated with a 7% greater risk for being in a Low-High cluster. The rate of physical inactivity and the percentage of Hispanics or non-Hispanic American Indians were associated with being in a High-Low cluster (5% to 21% greater risk).

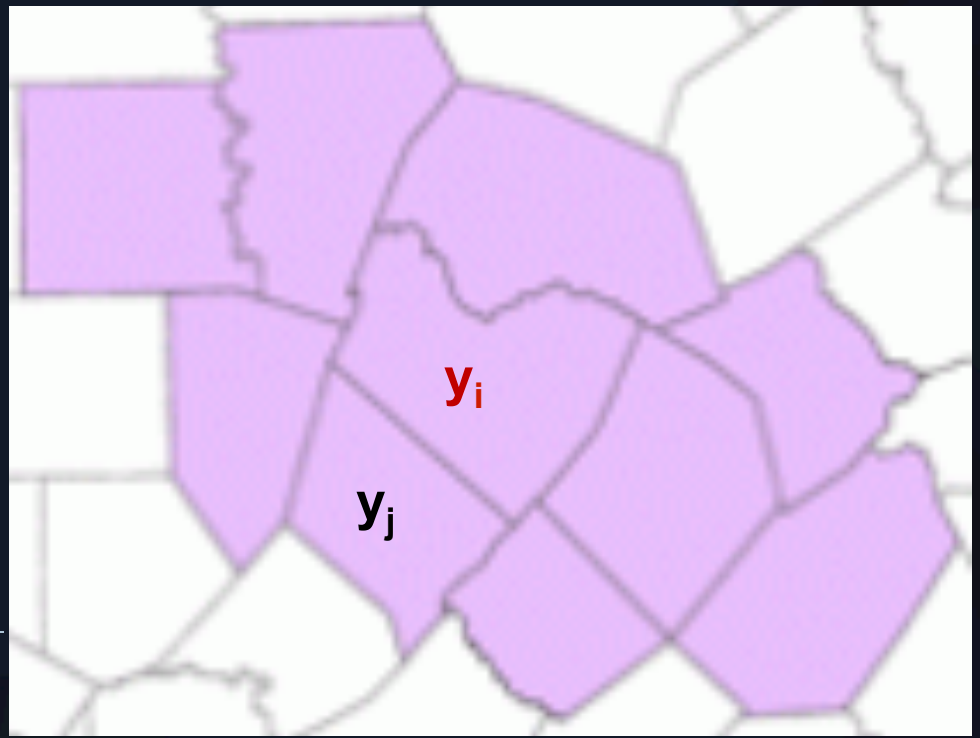
Discussion: Distinct spatial clusters of diabetes prevalence exist in the United States. Strong association between diabetes clusters and socioeconomic and other diabetes risk factors suggests that interventions might be tailored according to the prevalence of modifiable factors in specific counties.

- http://benthamopen.com/todiaj/articles/V005/29TODIAJ.pdf

Moran's *I* illustration

y_i	y_j	$(y_i - \bar{y})$	$(y_j - \bar{y})$	$(y_i - \bar{y})(y_j - \bar{y})$
15	18			

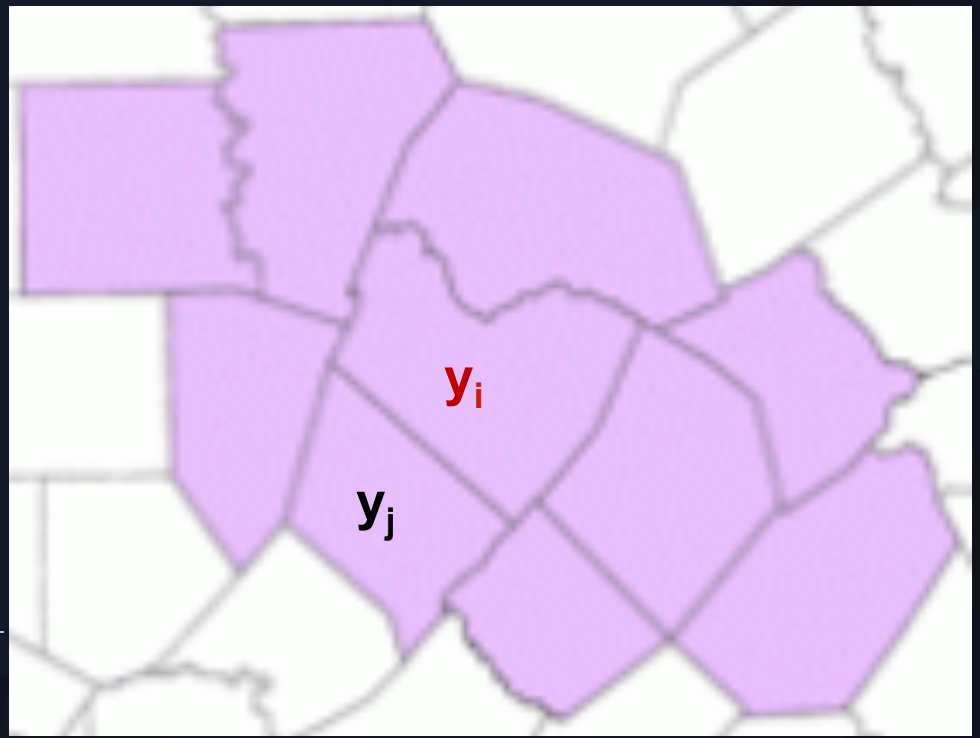
Mean of y = 10



Moran's *I* illustration

y_i	y_j	$(y_i - \bar{y})$	$(y_j - \bar{y})$	$(y_i - \bar{y})(y_j - \bar{y})$
15	18	5	8	40
12	8	2	-2	

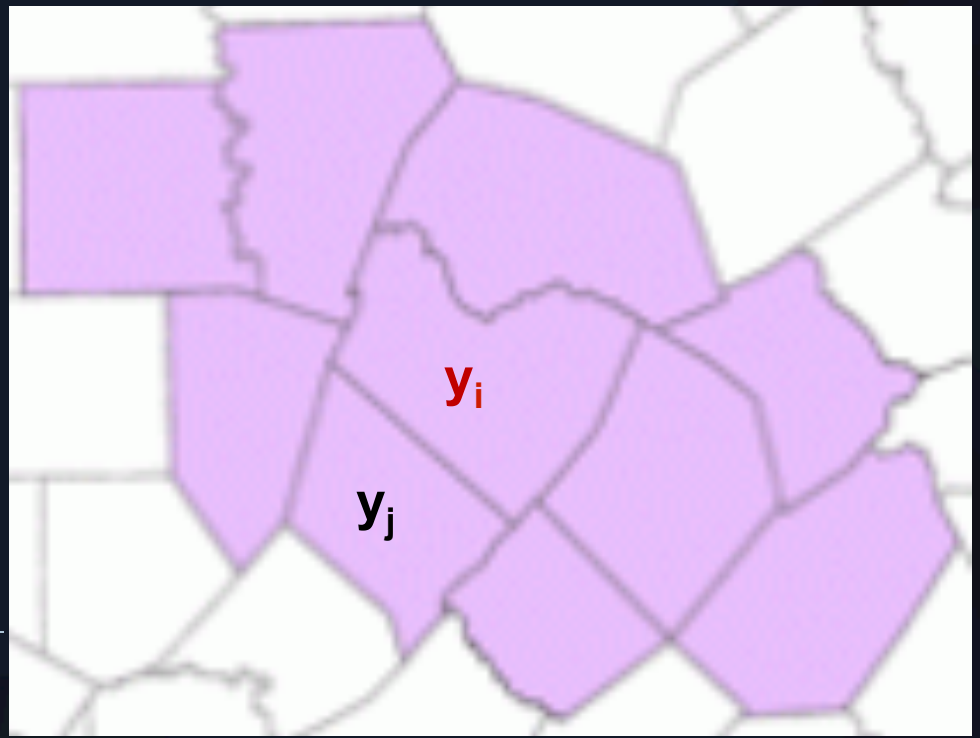
Mean of y = 10



Moran's *I* illustration

y_i	y_j	$(y_i - \bar{y})$	$(y_j - \bar{y})$	$(y_i - \bar{y})(y_j - \bar{y})$
15	18	5	8	40
12	8	2	-2	-4
8	12			

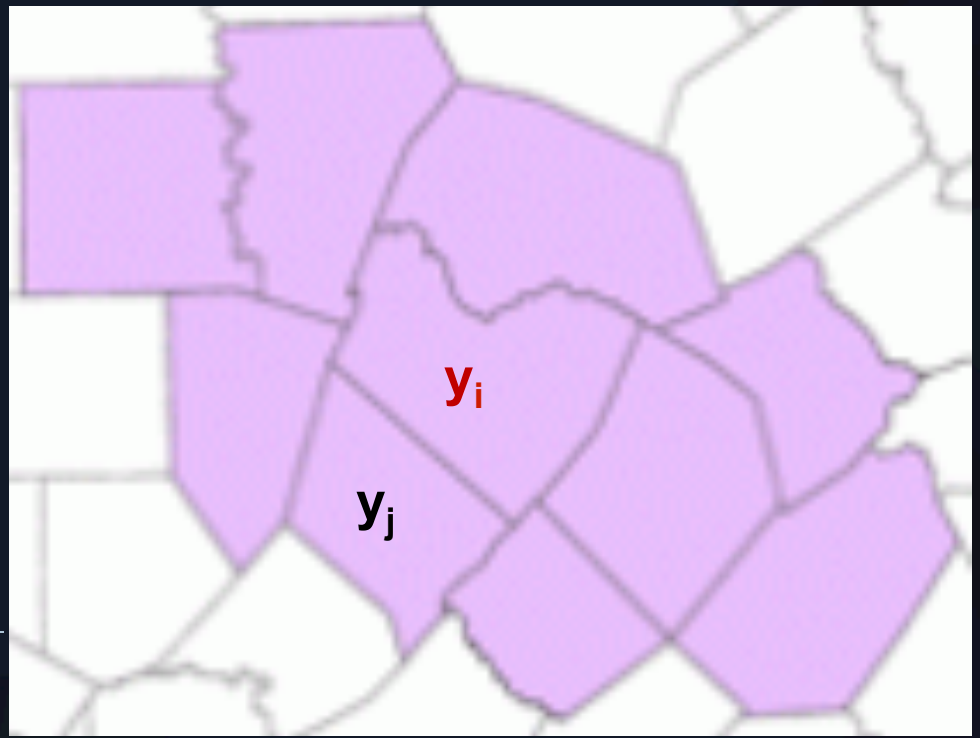
Mean of y = 10



Moran's *I* illustration

y_i	y_j	$(y_i - \bar{y})$	$(y_j - \bar{y})$	$(y_i - \bar{y})(y_j - \bar{y})$
15	18	5	8	40
12	8	2	-2	-4
8	12	-2	2	-4
2	5	-8	-5	40

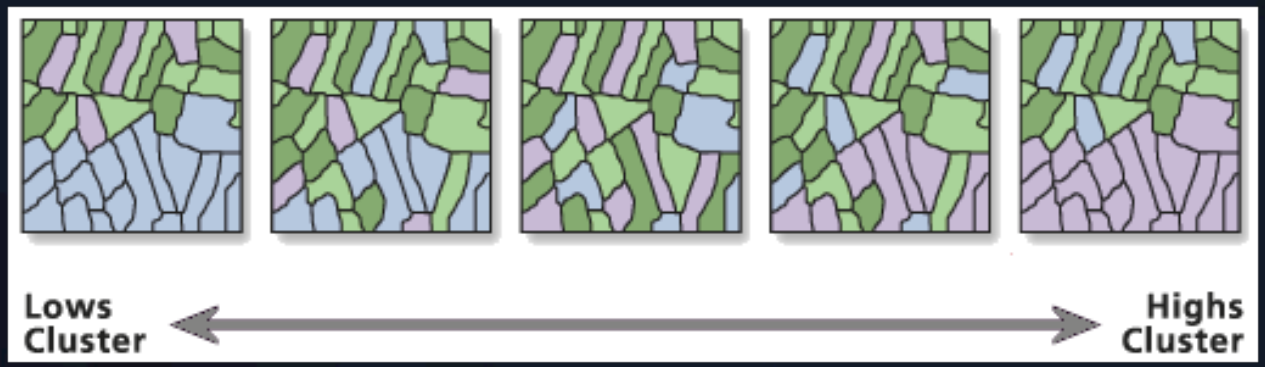
Mean of y = 10



General G-statistic

- Moran's *I* show whether nearby features are similar, **not** whether the similarity is among high or low values

- General G-statistic measures concentration of values over an area
  - h-h clusters ("hot spots")
  - l-l clusters ("cold spots")



Getis-Ord General *G*

Multiplies attribute values for each pair within neighborhood; HIGH product = H-H; LOW product = L-L

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \forall j \neq i \qquad (1)$$

where x_i and x_j are attribute values for features i and j, and $w_{i,j}$ is the spatial weight between feature i and j.

Divided by ***unweighted*** sum of products of all features

When obs G > exp G = hot spot;
When obs G < exp G = cold spot

Getis-Ord General *G*

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \quad \forall j \neq i \qquad (1)$$

where x_i and x_j are attribute values for features i and j, and $w_{i,j}$ is the spatial weight between feature i and j.

- *G* identifies whether local clusters have statistically significant high or low attribute values.
  - When obs G > exp G = hot spot; When obs G < exp G = cold spot

- *E(G)* = $\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}}{n(n-1)}$

Result	High/Low Clustering (G)	Spatial Autocorrelation (I)
The p-value is **not** statistically significant.	You cannot reject the null hypothesis. the observed spatial pattern of values could result from CSR/IRP	
The p-value **is** statistically significant, and the z-score is positive (>1.96).	Hot spot	Positive spatial autocorrelation
The p-value **is** statistically significant, and the z-score is negative (<-1.96).	Cold spot	Negative spatial autocorrelation

Spatial Statistics

Global

- Summarize data for whole regions
- Similarities across space
- Single statistic
- Unmappable
- Search for regularities

Local

- Local disaggregations of global statistics
- Differences across space
- Multi-valued statistic
- Mappable
- Search for exceptions (ex. ‘hotspots’)

SAC

- Local indicators of spatial association (LISA)- refer to local versions of Moran's *I* (I_i) and Geary's *c* (c_i).

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

- How to convert global statistic to local statistic?

SAC

- Local indicators of spatial association (LISA)- refer to local versions of Moran's *I* (I_i) and Geary's *c* (c_i).

Difference between target and mean

Sum of differences between each neighbor and mean

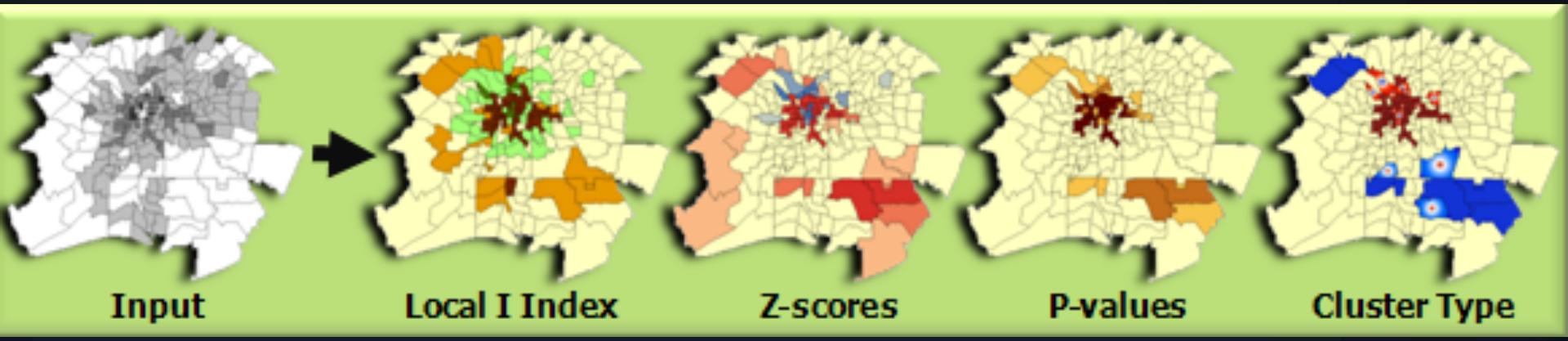
$$I_i = \frac{(x_i - \bar{x})}{s^2} * \sum_j w_{ij}(x_j - \bar{x})$$

Variance

LISA example

- Bird species richness in NY

Local Moran's *I*



- High + z score (> 1.96)? = H-H, L-L (cluster)
- High – z score (< -1.96)? = H-L, L-H (outlier)

(Local) statistics 'best practices'/issues

- Different swm definitions
  - All features should have at least 1 neighbor
  - No feature should have **all** other features as neighbor
  - Standardize rows (usually) with polygons
- $n > 30$
- Input field must be a count, rate, or similar numeric value
  - no negative values
- Edges, small distances can cause problems
- Use corrections (Bonferroni, etc) or Monte Carlo simulations to determine statistical significance

Review

- Is there a spatial pattern (clustering/dispersal)?
  - GLOBAL stats (Moran's *I*, Geary's *c*, General *G*)
- Where is the spatial pattern?
  - LOCAL stats
  - Where are clusters are high values? Where are clusters of low values? G_i *
  - Where are outliers (clusters of high surrounded by low or clusters of low surrounded by high)? Moran's I_i

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Information Economics and Policy 18 (2006) 423–448
INFORMATION ECONOMICS AND POLICY
www.elsevier.com/locate/iep

A spatial taxonomy of broadband regions in the United States

Tony H. Grubesic *

Abstract

The steady growth of broadband penetration in the United States is indicative of a major shift in advanced data services and last-mile infrastructure in the deregulated telecommunication environment. Although there are concerns with the equitable provision of broadband services in urban, rural and remote areas, the diffusion process has also created a unique landscape of broadband availability that reflects elements of competition, federal policy, local government initiatives, technological limitations and location. This paper explores the dynamic and diverse spatial landscape of broadband availability in the United States at the zip code level, for 2004. In addition, this study provides a multivariate, spatial taxonomy of broadband regions, highlighting their socioeconomic and demographic differences.

1. *High–high "broadband core":* Where zip codes displaying high levels of broadband availability and competition are surrounded by other zip codes with similar values. These regions correspond to the greatest levels of broadband availability and competition in the United States and are primarily located in urban areas.
2. *Low–low "broadband periphery":* Where zip codes displaying low levels of broadband availability and competition are surrounded by other zip codes with similar values. These regions are largely devoid of broadband options and are primarily located in the most rural and remote areas of the United States.
3. *Low–high "islands of inequity":* Where zip codes displaying low levels of broadband availability and competition are surrounded by zip codes displaying relatively high values. These locations are typically found adjacent to, or inside of, the broadband core zones.
4. *High–low "islands of availability":* Where zip codes displaying high levels of broadband availability and competition are surrounded by zip codes displaying relatively low levels. These locations are scattered throughout the U.S., with many of them found on the outskirts of MSAs or CMSAs.

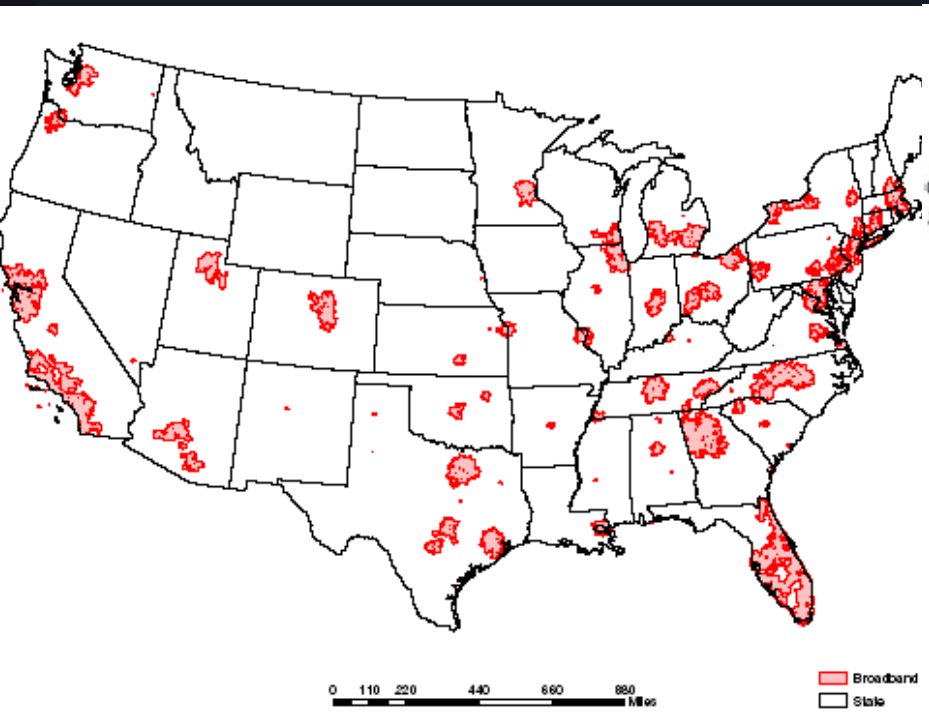


Fig. 4. Broadband core.

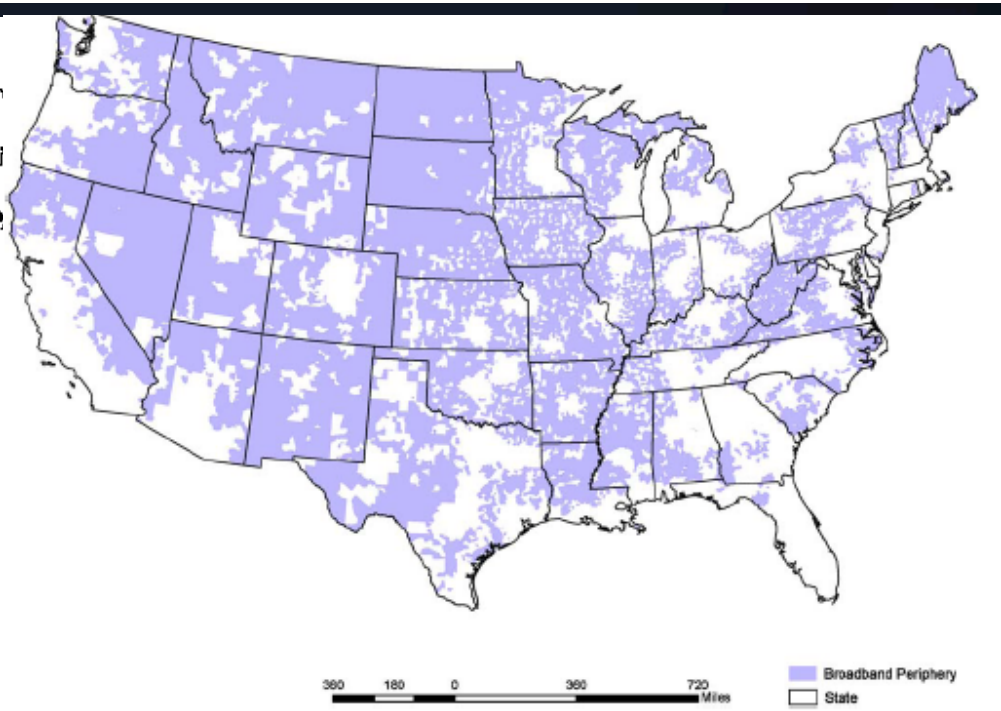


Fig. 6. Broadband periphery.

Fig. 7. Islands of inequity.

Fig. 8. Islands of availability.

Forest Ecology and Management 284 (2012) 107–115

Contents lists available at SciVerse ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

Remote detection of small wetlands in the Atlantic coastal plain of North America: Local relief models, ground validation, and high-throughput computing

Paul B. Leonard [a,*], Robert F. Baldwin [a], Jessica A. Homyack [b], T. Bently Wigley [c]

[a] School of Agricultural, Forest, and Environmental Sciences, Clemson University, Clemson, SC 29634, USA
[b] Weyerhaeuser NR Company, Vanceboro, NC 28586, USA
[c] National Council for Air and Stream Improvement Inc., Clemson, SC 29634, USA

ARTICLE INFO

Article history:
Received 17 May 2012
Received in revised form 7 July 2012
Accepted 24 July 2012
Available online 24 August 2012

Keywords:
Isolated wetlands
Ephemeral wetlands
LiDAR
High-throughput computing
Sustainable forestry certification

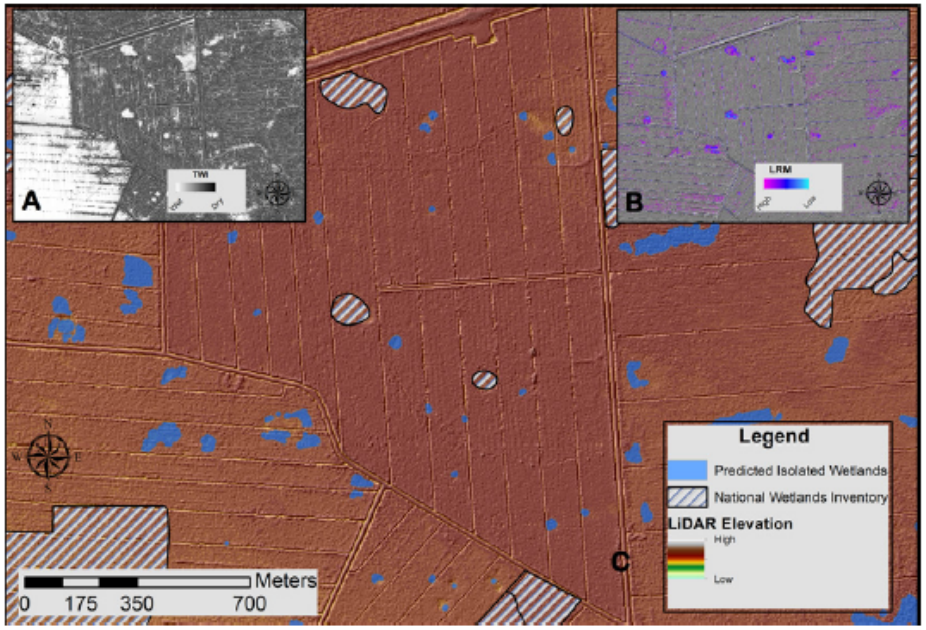
ABSTRACT

Isolated wetlands are ecologically important freshwater ecosystems that occur frequently throughout the Atlantic coastal plain ecoregions of North America. Known to support 86 species recognized by the US Fish and Wildlife Service as threatened or endangered, isolated wetlands are a conservation priority in the United States and elsewhere. They are often obscure and methods to detect them at the spatial scales necessary for systematic conservation planning and forest management have been time consuming, cost ineffective, or too coarse-filter. To fill existing information gaps and develop a repeatable, high-resolution methodology, we subjected LiDAR elevation data to custom relief models designed to elucidate fine-scale geomorphology, specifically small, localized changes in concavity, as a location predictor. Because fine grain size and large spatial extent can impose processing limits in landscape-level analysis, we executed our workflow in a high-throughput computing (HTC) environment, which achieved a 91× time-savings over our 55,000 ha study area. We conducted field validation at 114 randomly selected sites to measure model commission (14.9%), approximate omission (5.3%) error rates and estimate wetland boundaries. Depressional wetlands predicted in this study ($n = 4610$) were mostly small ($\bar{x} = 0.37 \pm 0.69$ ha) and previously unmapped sites. The mapping accuracy of this effort (85.1%) suggests that local relief models captured slight geomorphologic changes that successfully predict wetland boundaries in low-relief ecosystems. Many small wetlands are centers of biodiversity in forested landscapes and such analyses will provide information and improved methods for landscape-scale management and conservation.

3.1. Interpretation of LISA clusters

Based on model outputs, we describe the five aforementioned possible outcomes (i.e., cluster types) from LISA analysis (spatially autocorrelated LRM results), the implication of each cluster type, and the possible landforms delineated by each.

1. LrLr clusters were the most likely predictor of small, depressional wetlands. These points fell inside a depression and were surrounded by similar points.
2. LrHr clusters most obviously delineated ditches, but also described skidder ruts, small narrow pools (e.g., depressions immediately abutting convex, planted beds), coves, or spill points, which connected a larger complex of wetlands. The study area was highly intersected with a network of drainage ditches and points falling inside these areas created linear clusters while most surrounding points fell outside of a ditch.
3. HrLr clusters were around the boundaries of depressional areas where high-relief areas were surrounded by low-relief. No error sites contained these clusters although they may be found in small peninsulas, hummocks, or islands of vegetation commonly seen in pine flats.
4. HrHr clusters signified large flat areas with little micro-relief. Only one ground-verified wetland (1%) displayed this type of cluster and it was much larger (>0.10 ha) than targeted features although still omitted from NWI. This particular omission suggests the selected neighborhood may have been too small to correctly characterize some wetlands of this size (e.g., large pine flats).
5. NA clusters were typically found in areas where local relief values were >1 or where points displayed a non-significant, non-autocorrelated spatial arrangement. One of three omitted sites displayed this clustering. This inundated area was part of a larger pine flat wetland, correctly mapped nearby and likely hydrologically connected by an overspill point.



hy-based wetland models with the National Wetlands Inventory. For the same scene, (A) LiDAR-derived top
sive (i.e., high commission) compared to (B) LiDAR-derived local relief model. Frame (C) illustrates omissio

ELSEVIER

Health & Place 13 (2007) 368–380

www.elsevier.com/locate/healthpla

Spatial patterns of malaria in the Amazon: Implications for surveillance and targeted interventions

Marcia Caldas de Castro[a,*], Diana Oya Sawyer[b], Burton H. Singer[c]

Abstract

A measure of local spatial association, $G_i^*(d)$, is applied to test for the presence of malaria clusters in a colonization area in the Brazilian Amazon. Clusters of high and low malaria rates at different moments in time are identified. They suggest unambiguous spatial patterns of transmission, most likely linked to the social and natural habitat. Results imply that a comprehensive identification of the determinants of malaria transmission requires a spatial framework of analysis, and that control strategies must be spatially targeted and guided by a surveillance system that constantly learns the specificities of local transmission and adapts interventions to them.

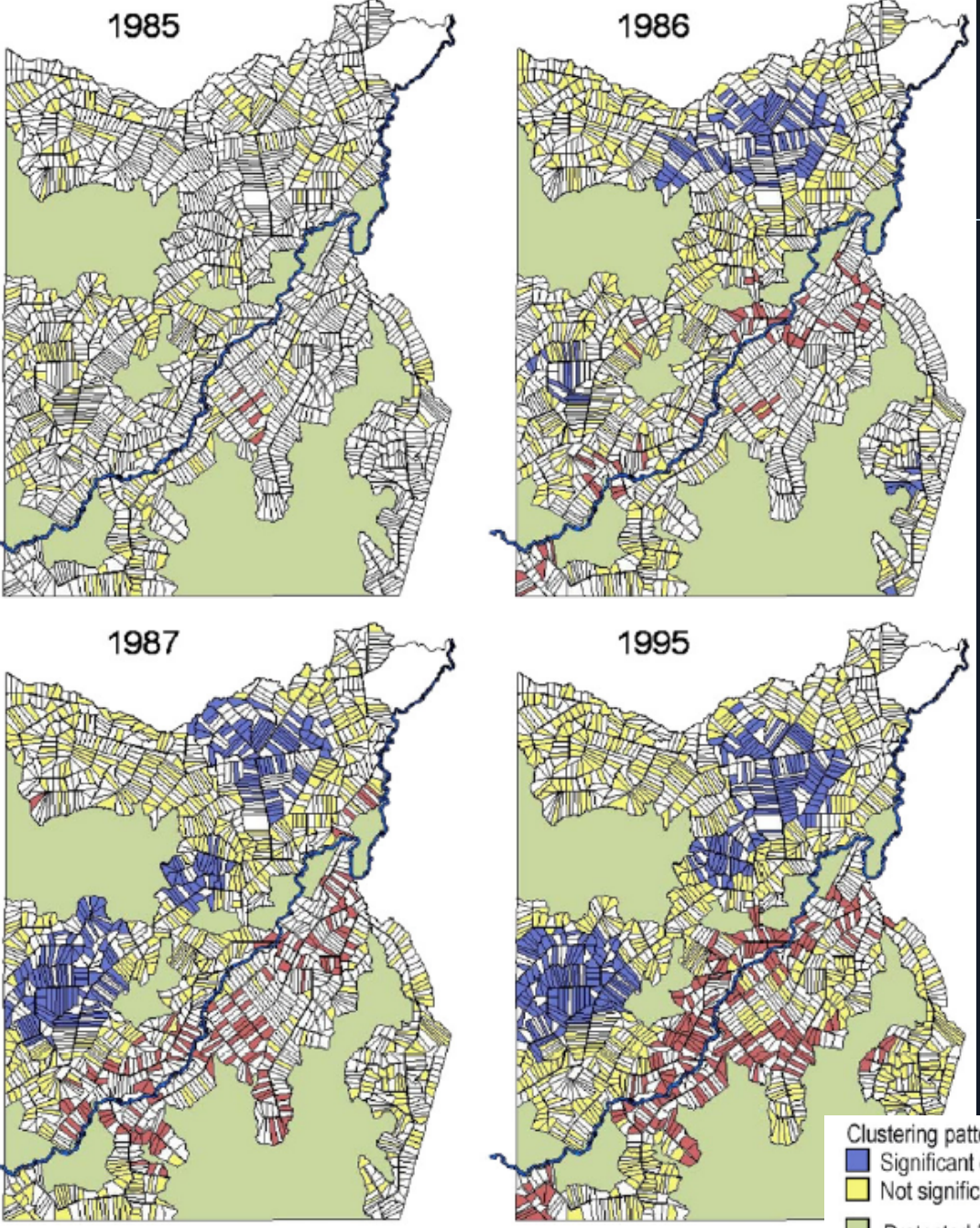


Fig. 2. Statistical significance of the $G_i^*(d)$ statistic ($d = 3500$ m)—Machadinho, 1985/1995

Besides facilitating the identification of clusters, we argue that the observed trajectories of the Gi* statistic at different distances can reveal the patterns of malaria transmission, and the importance of choosing the most appropriate distance. Next we show a series of graphs and maps that highlight the trajectories of the Gi* statistic for distances ranging from 500 to 8000 m. Each line in the graph represents the trajectory of Gi* values for one single plot, and the thick black lines are the cutoff values at a 5% level given by the FDR procedure for multiple testing.

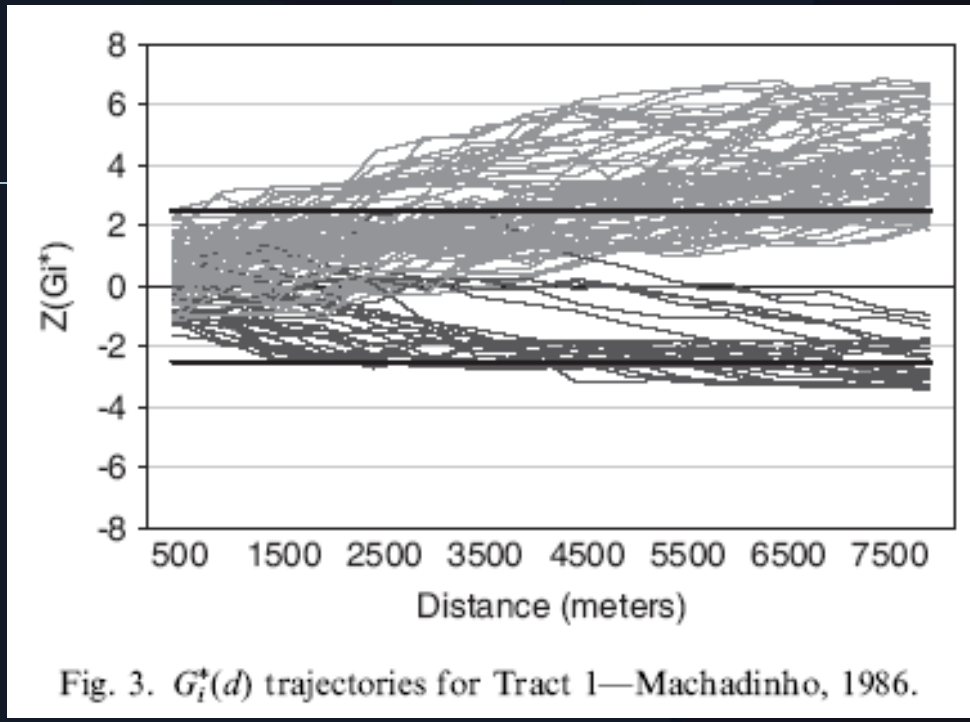


Fig. 3. $G_i^*(d)$ trajectories for Tract 1—Machadinho, 1986.

Fig. 3 shows the trajectories for Tract 1 in 1986. A few plots are associated with a downward pattern (darker lines), and they are all located in the Southeastern portion of Tract 1. In fact, this is the area where the small cluster of low malaria rates is
observed in that year, with an average rate slightly above 8%. The remaining plots in Tract 1 reveal an upward trend in the local statistic, suggesting that as distance increases the additional plots added to the neighborhood contribute to the pattern of clustering of high malaria rates in the area. The cloud of increasing lines is an indication of the dramatic increase of malaria transmission registered in Machadinho in 1986.

Two patterns are observed for Tract 2 in 1986, as shown in Fig. 4. The first is a tendency for clustering of low values, represented by graph (a) and given by plots colored as light gray in the map. The few lines that after a certain distance show an increase in the indicator are associated with locations closer to the border between Tracts 1 and 2. So, as distance increases, the set of neighbors for these plots will include locations in Tract 1 that have very high rates

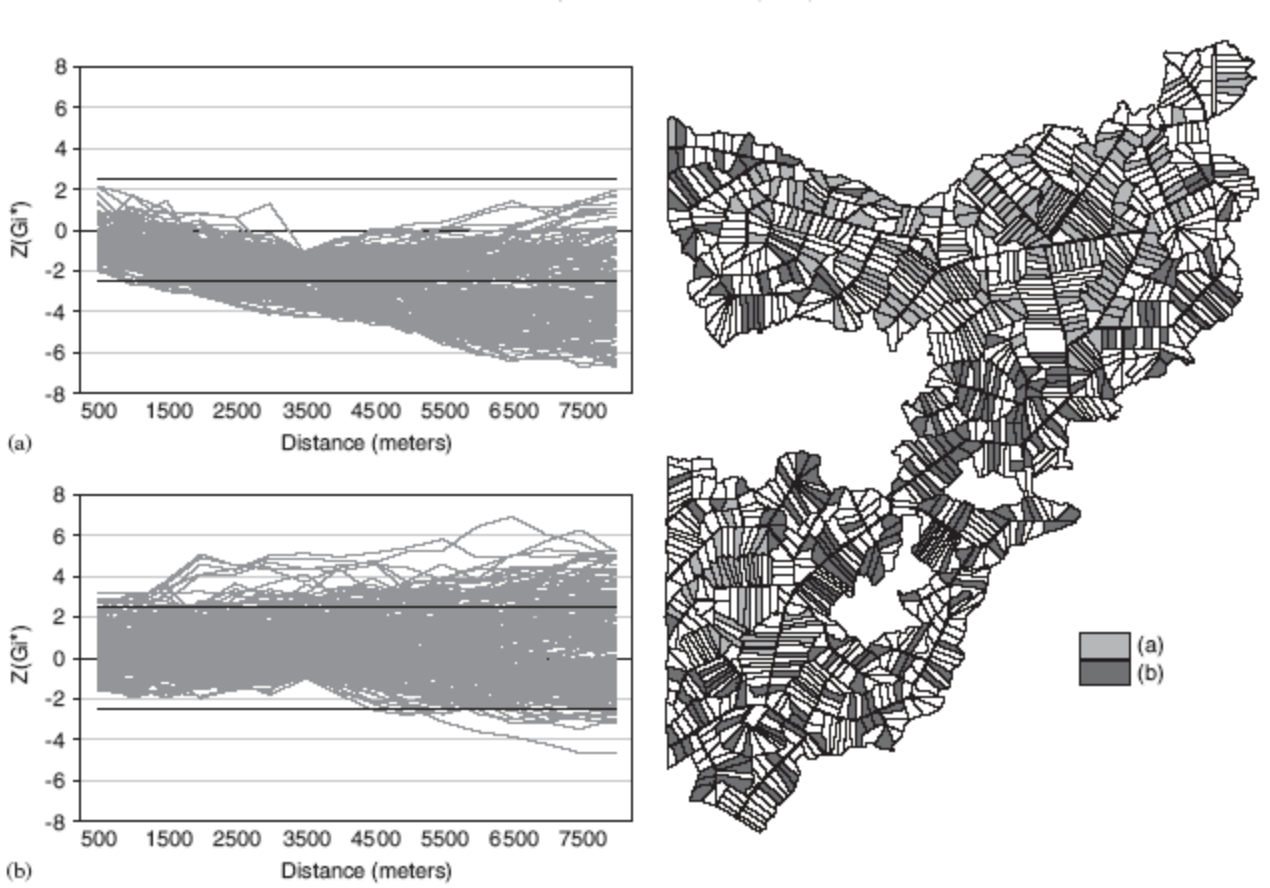


Fig. 4. $G_i^*(d)$ trajectories for Tract 2—Machadinho, 1986.

of malaria, changing the direction of the statistic. The second pattern, shown in graph (b) and associated with plots colored as dark gray in the map, has the opposite behavior. The tendency for clustering around high values is mainly restricted to the border between Tracts 1 and 2. Both graphs suggest a threshold at 3500 m, which is the distance used to compute the Gi statistic. Although this is by no means a proof that the correct distance was chosen, it is encouraging to detect underlying spatial processes acting at that distance.

- Bivariate LISA

$$I_i = \frac{(x_i - \bar{x})}{s^2} * \sum_j w_{ij}(x_j - \bar{x})$$

$$I_i = \frac{(x_i - \bar{x})}{s^2} * \sum_j w_{ij}(z_j - \bar{z})$$

Different variable

Applied Geography

Volume 40, June 2013, Pages 161–170

A GIS-based risk rating of forest insect outbreaks using aerial overview surveys and the local Moran's I statistic

Christopher Bone[a], Michael A. Wulder[b], Joanne C. White[b], Colin Robertson[c], Trisalyn A. Nelson[d]

[a] Department of Geography, University of Oregon, PO Box 1251, Eugene, OR 97403, USA

[b] Canadian Forest Service (Pacific Forestry Centre), Natural Resources Canada, 506 West Burnside, Victoria, BC V8Z 1M5, Canada

[c] Department of Geography & Environmental Studies, Wilfrid Laurier University, Waterloo, Ontario, N2L 3C5, Canada

[d] Spatial Pattern Analysis & Research (SPAR) Laboratory, Department of Geography, University of Victoria, PO Box 3060, Victoria, BC V8W 3R4, Canada

Abstract

The objective of this study is to provide an approach for assessing the short-term risk of mountain pine beetle *Dendroctonus ponderosae* Hopkins (Coleoptera: Scolytidae) attack over large forested areas based on the spatial-temporal behavior of beetle spread. This is accomplished by integrating GIS, aerial overview surveys, and local indicators of spatial association (LISA) in order to measure the spatial relationships of mountain pine beetle impacts from one year to the next. Specifically, we implement a LISA method called the

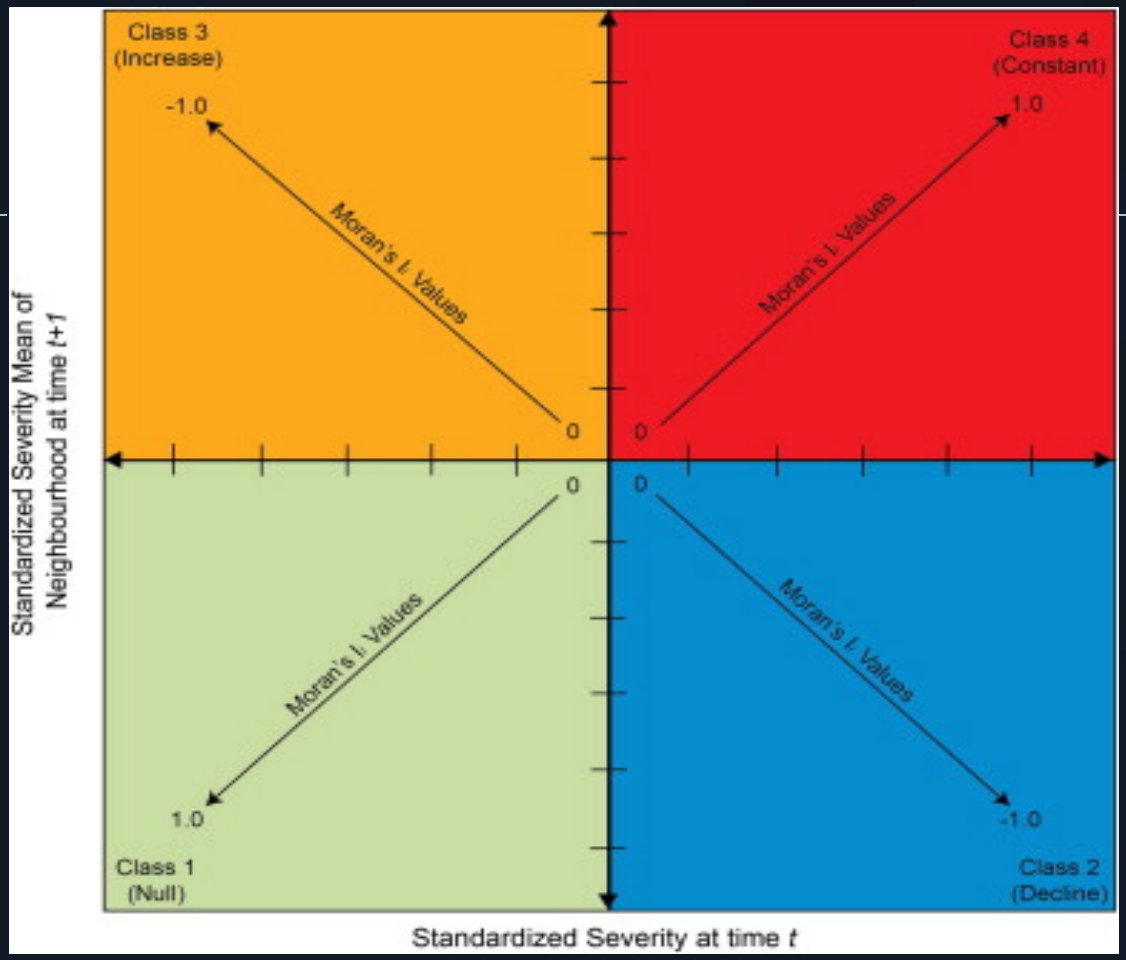
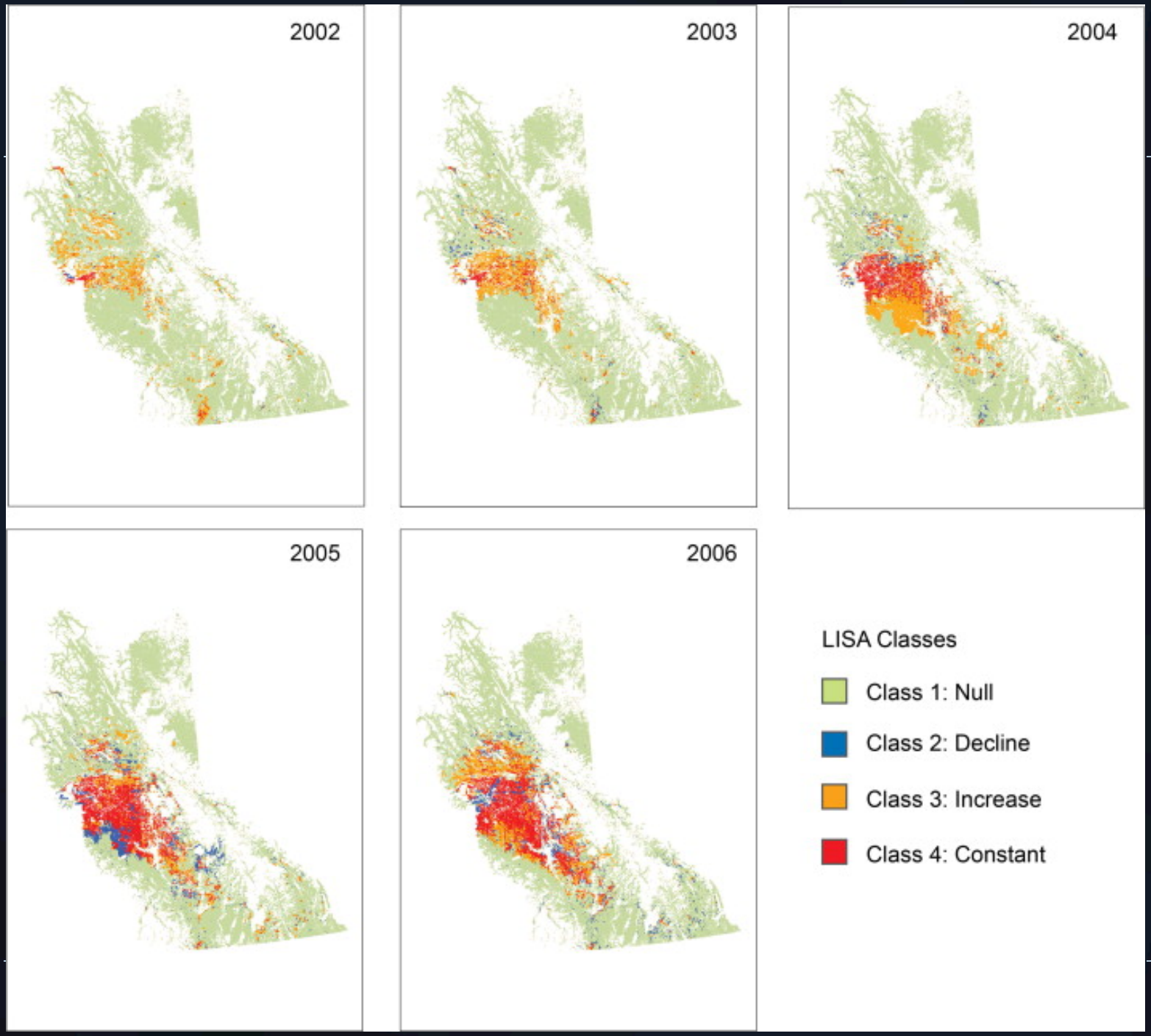


Fig. 3 A conceptual schematic of the Moran's I i scatterplot. Arrows within each class represents the direction of Moran's I values from 0 to 1. Each quadrat represents the relationship between observation i and its neighborhood. In addition, each qua...

A GIS-based risk rating of forest insect outbreaks using aerial overview surveys and the local Moran's I statistic

Applied Geography Volume 40 2013 161 - 170

http://dx.doi.org/10.1016/j.apgeog.2013.02.011


2002
2003
2004
2005
2006
LISA Classes
Class 1: Null
Class 2: Decline
Class 3: Increase
Class 4: Constant

Bivariate LISA example

Univariate LISA

$$I_i = \frac{(x_i - \bar{x})}{s^2} * \sum_j w_{ij}(x_j - \bar{x})$$

$$I_i = \frac{(x_i - \bar{x})}{s^2} * \sum_j w_{ij}(z_j - \bar{z})$$

Different variable

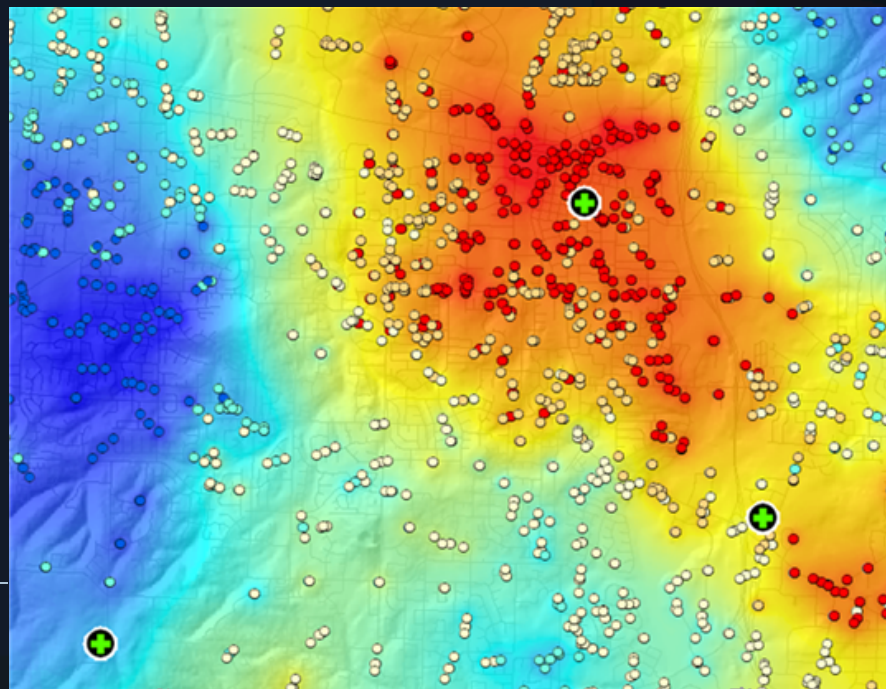
Bivariate LISA

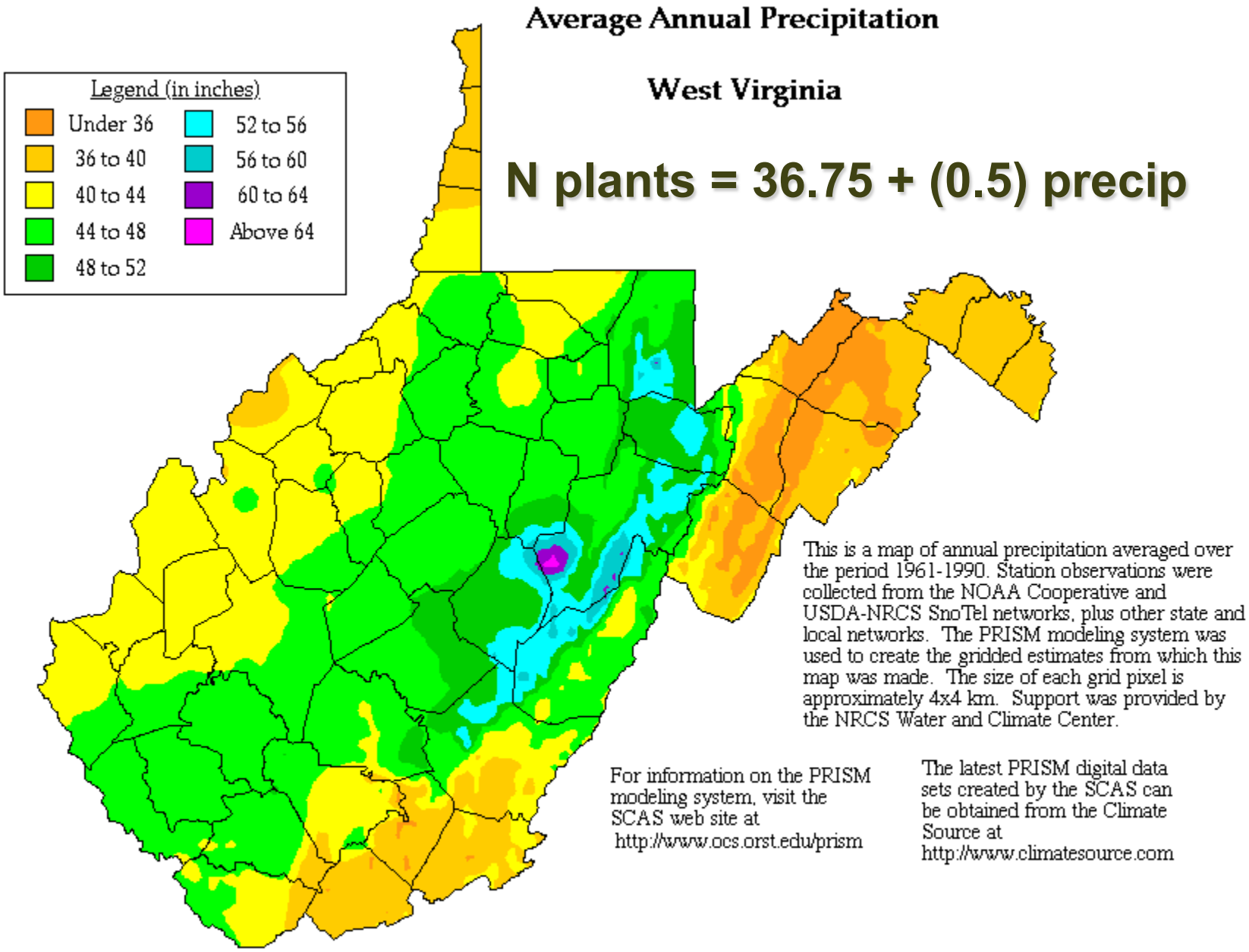
Regression slope

$$b = \frac{\sum \left((X_i - \bar{X})(Y_i - \bar{Y})\right)}{\sum (X_i - \bar{X})^2}$$

Where?

- **Where** are incidences of cancer higher than other places?
- **Where** are the hot spots for crime, 911 emergency calls, fires?
- **Where** are intersections that have high numbers of traffic accidents?



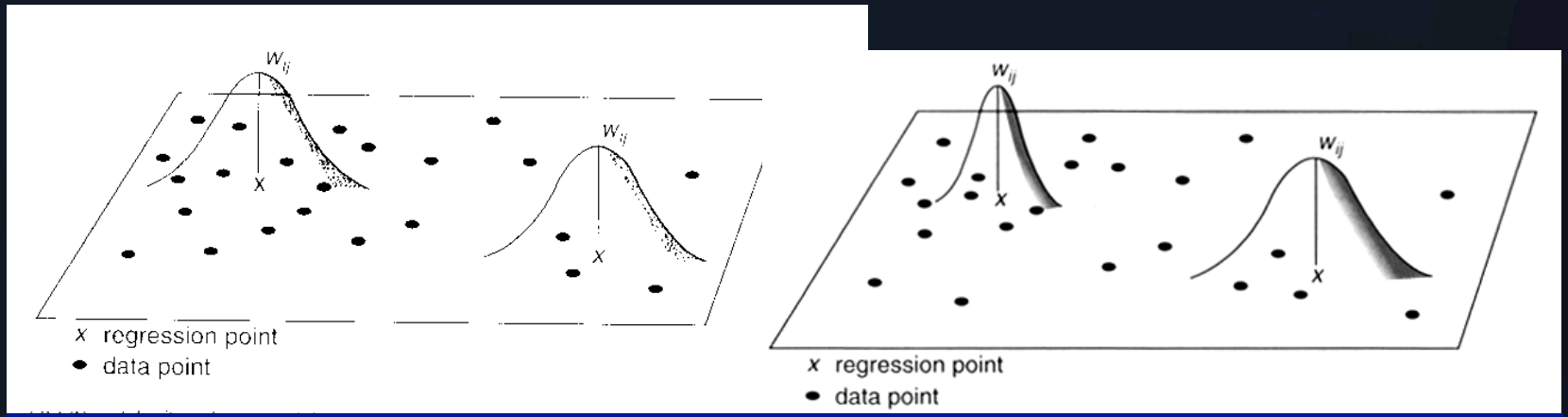

Average Annual Precipitation
West Virginia
Legend (in inches)
Under 36
36 to 40
40 to 44
44 to 48
48 to 52
52 to 56
56 to 60
60 to 64
Above 64
N plants = 36.75 + (0.5) precip
This is a map of annual precipitation averaged over the period 1961-1990. Station observations were collected from the NOAA Cooperative and USDA-NRCS SnoTel networks, plus other state and local networks. The PRISM modeling system was used to create the gridded estimates from which this map was made. The size of each grid pixel is approximately 4x4 km. Support was provided by the NRCS Water and Climate Center.
For information on the PRISM modeling system, visit the SCAS web site at http://www.ocs.orst.edu/prism
The latest PRISM digital data sets created by the SCAS can be obtained from the Climate Source at http://www.climatesource.com
Copyright 2000 by Spatial Climate Analysis Service, Oregon State University

Geographically weighted regression

- Addresses the non-stationarity directly
  - Allows the relationships to vary over space, i.e., βs do not need to be the same
  - $y_i=\beta_{i0} + \beta_{i1}x_{1i}+ \beta_{i2}x_{2i}+\ldots\ldots+ \beta_{in}x_{ni}+\varepsilon_i$

Instead of remaining the same everywhere, βs now vary by locations (i) (and R^2)

GWR kernel



GWR with fixed kernel GWR with adaptive kernel

From Fotheringham, Brundson and Charlton. 2002. *Geographically Weighted Regression*

Points are weighted based on distance from center of kernel
e.g. Gaussian kernel where weighting is given by:

$$w_i(\mathbf{g}) = \exp[-1/2(d_{ij}/b)^2$$ where b is bandwidth

OLS coefficient (slope)

1) $$b = \frac{\sum\left(\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right)}{\sum\left(X_i - \overline{X}\right)^2}$$

Sum of crossproducts

Sum of squares of X

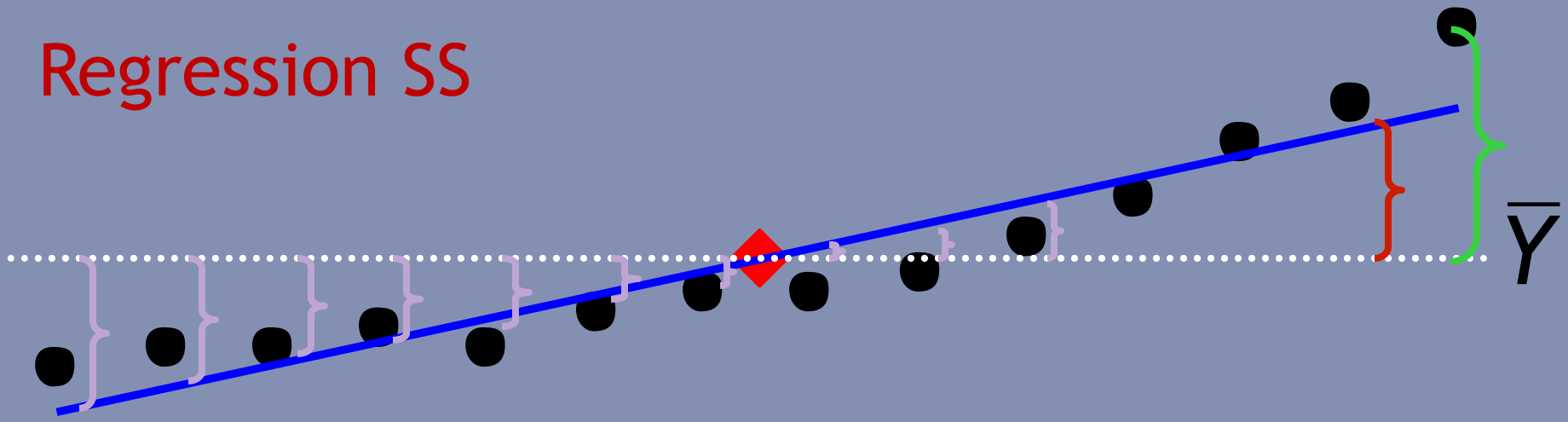
So, proportion of the total variation in Y accounted for by the regression,

$$r^2 = \frac{\text{regression SS}}{\text{total SS}}$$

$$\frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

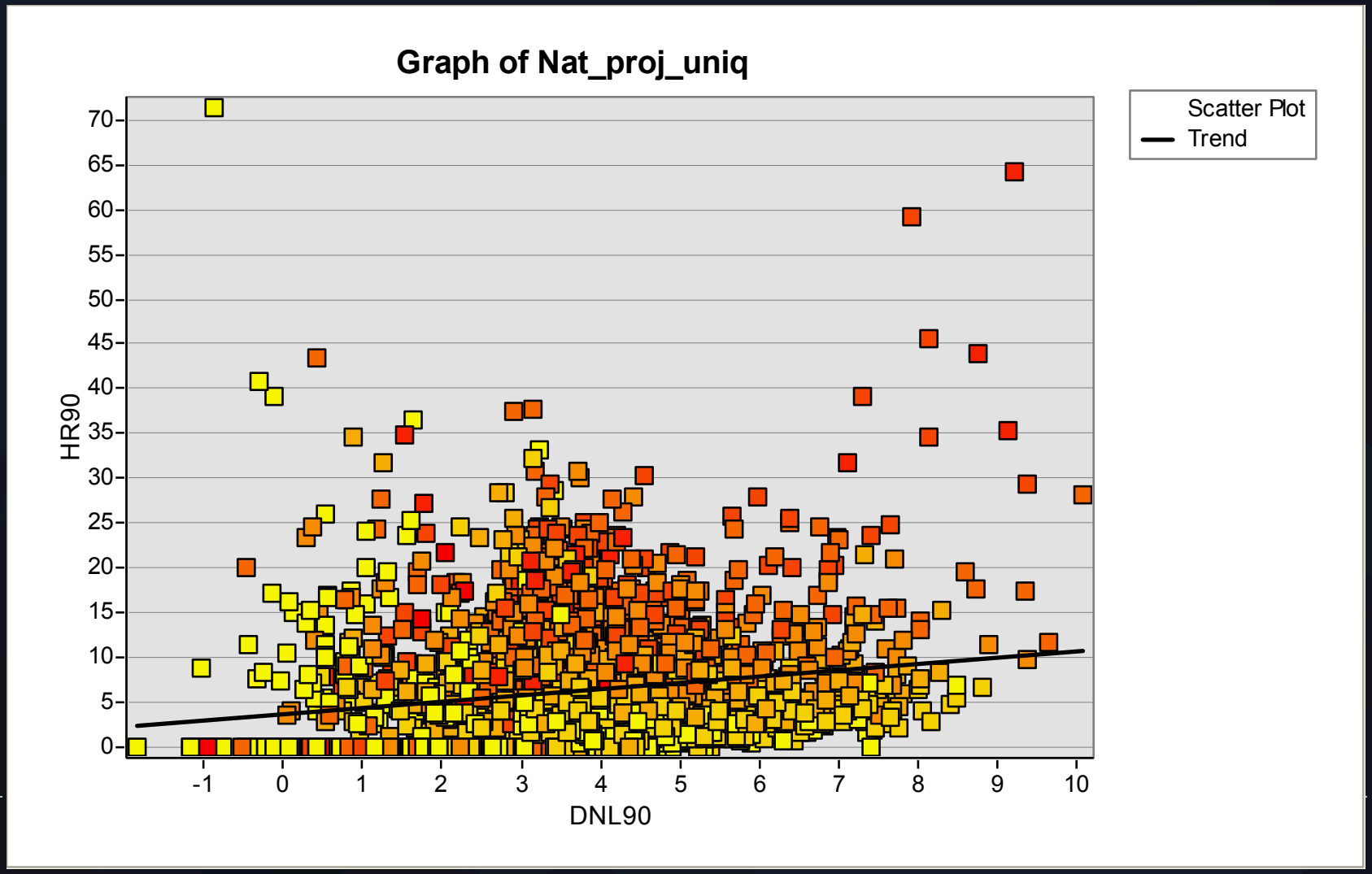
Total SS

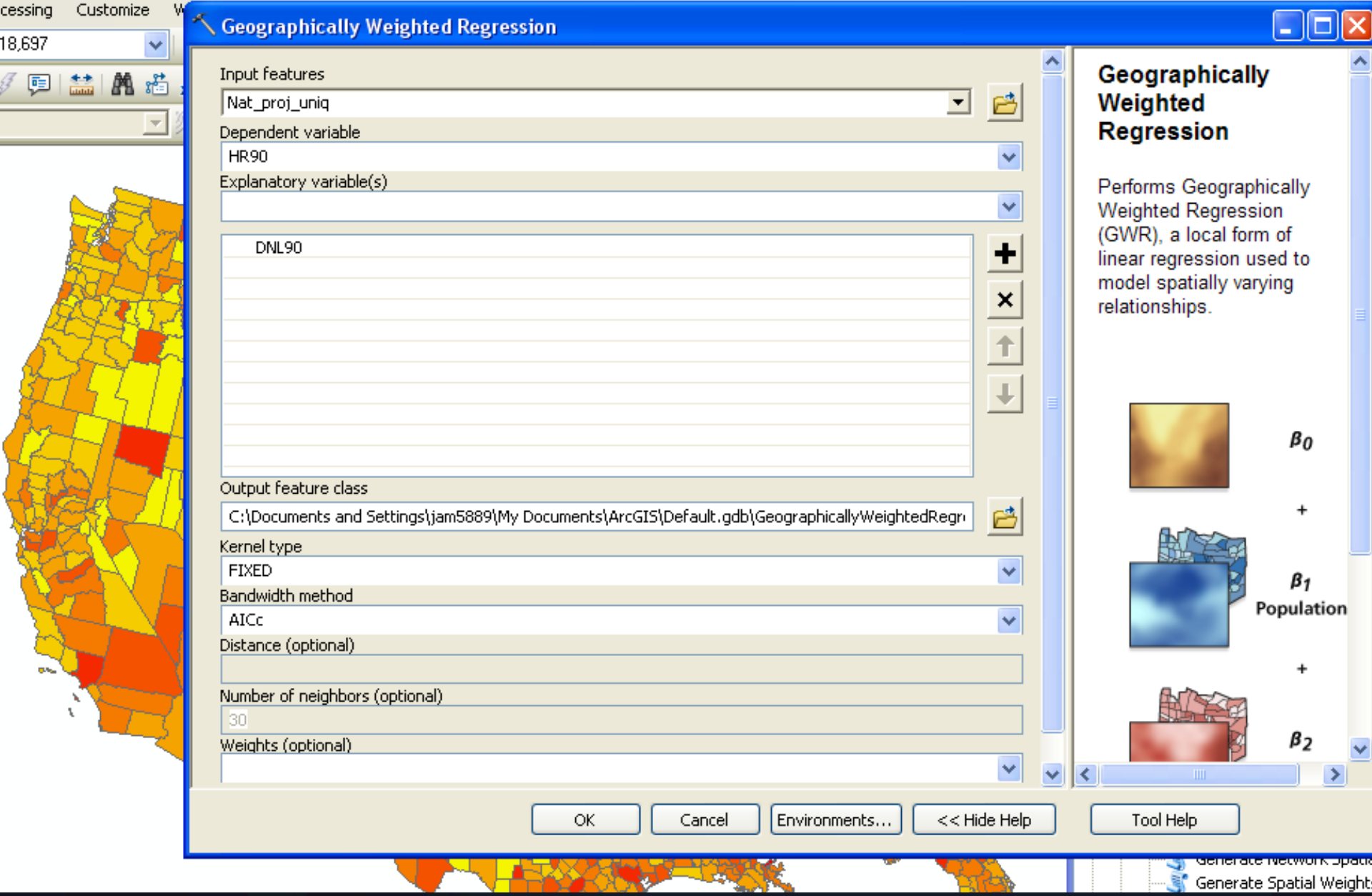
Regression SS

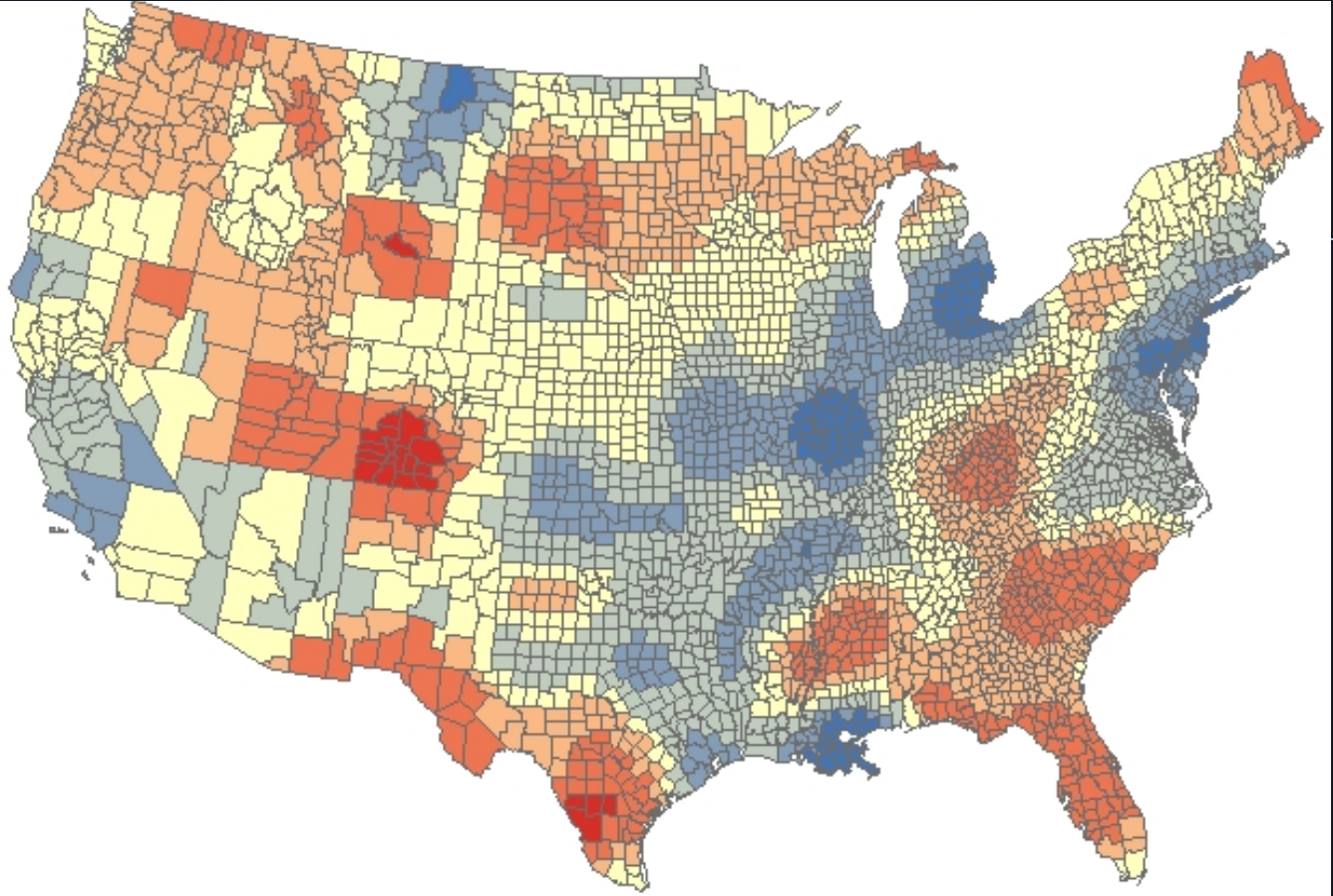


- Homicide rate = f(pop density)

- Homicide rate = f(pop density)




cessing
Customize
18,697
Geographically Weighted Regression
Input features
Nat_proj_uniq
Dependent variable
HR90
Explanatory variable(s)
DNL90
Output feature class
C:\Documents and Settings\jam5889\My Documents\ArcGIS\Default.gdb\GeographicallyWeightedRegr
Kernel type
FIXED
Bandwidth method
AICc
Distance (optional)
Number of neighbors (optional)
30
Weights (optional)
Geographically Weighted Regression
Performs Geographically Weighted Regression (GWR), a local form of linear regression used to model spatially varying relationships.
β0
+
β1
Population
+
β2
OK
Cancel
Environments...
<< Hide Help
Tool Help
Generate Spatial Weights



Legend

GeographicallyWeightedRegression3

C1_DNL90

- 2.466550 - 5.339189
- 1.560441 - 2.466549
- 0.859449 - 1.560440
- 0.192890 - 0.859448
- -0.652443 - 0.192889
- -2.719717 - -0.652444
- -5.405655 - -2.719718

Global coeff = 0.7002

Homicide rate = f(population density)

More GWR info

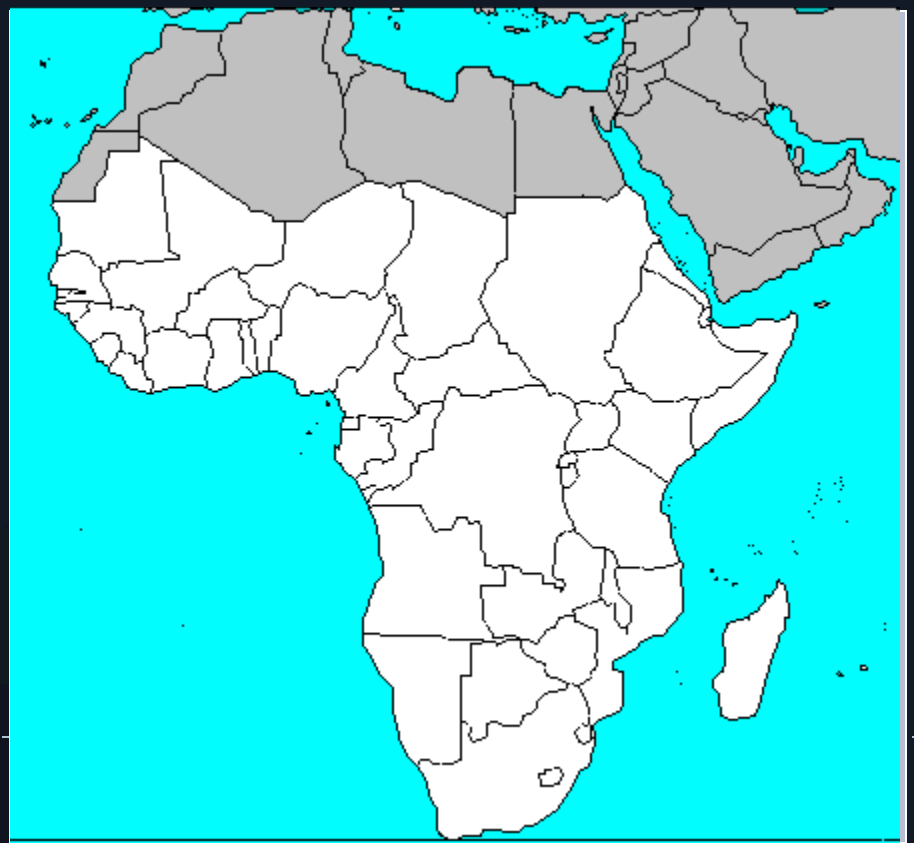
- http://www.st-andrews.ac.uk/geoinformatics/gwr/
- Fotheringham et al, 2002
- Foody, 2004
- Osborne, et al. 2007: Non-stationarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions 13*, 313–23.

Example applications:

Foody, G. M. (2004b) Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography*, 13, 315-320.

Sub-Saharan Africa:

(bird) Species richness = f(NDVI, Temp, Precip)



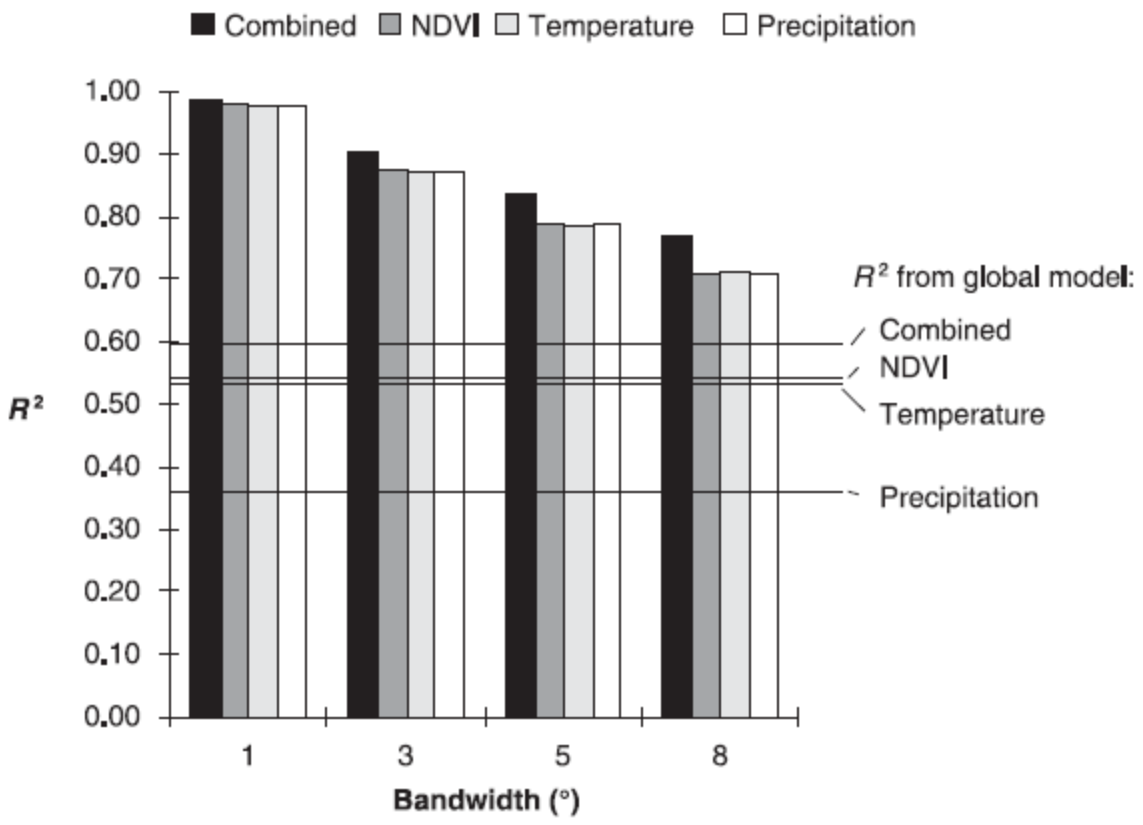


Figure 1 Amount of variance explained (R^2) by relationships between species richness and the three explanatory variables alone and combined, determined using all cases with paired species richness and environmental data ($n = 1729$). The GWR analyses were undertaken with four different bandwidths. In each GWR analysis the model parameters were significantly nonstationary (Monte Carlo permutation test, $P < 0.0001$). Every GWR model also had a lower Akaike Information Criterion than the corresponding global model (minimum difference = 735.98) indicating a closer fit to the data, accounting for the differences in the complexity and degrees of freedom of the models compared. For comparative purposes the magnitude of the R^2 estimated from the corresponding conventional (global) regression analyses are indicated by horizontal lines.

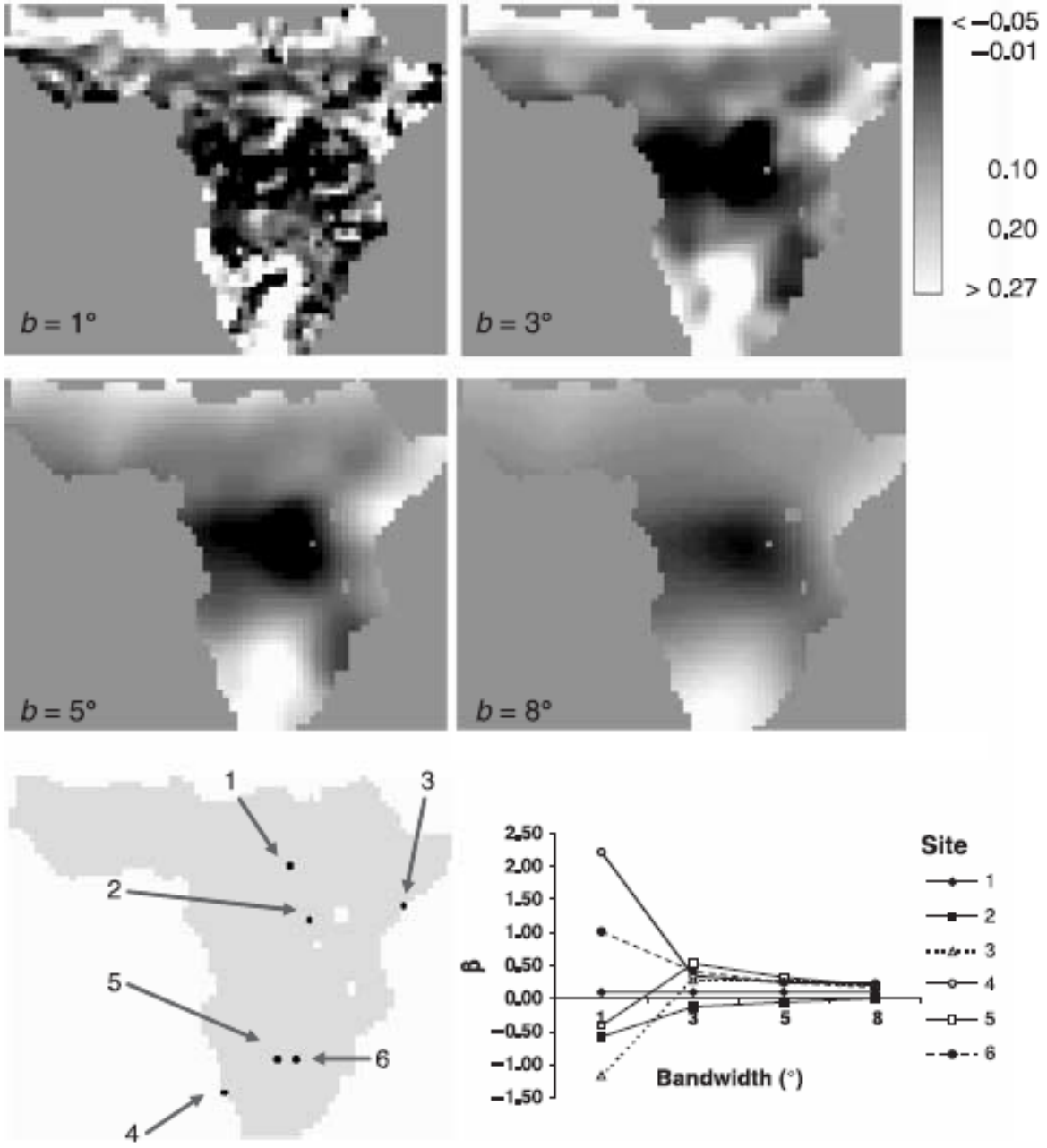


Figure 3 Spatial variation in the slope of the relationship between species richness and total annual precipitation at four bandwidths (b). Spatial detail increases with a decrease in bandwidth. The effect of scale variation is evident in the relationship between bandwidth and the estimate of the slope parameter (β) highlighted for six locations. Those selected include examples that are uni-directionally positive (site 2), uni-directionally negative (sites 4 and 6), relatively stable (site 1) as well as a location at which there is a major change in the direction and magnitude of the estimate with scale change (site 5), with the estimate for each location gradually converging towards the global model estimate of 0.1016. For the images depicting the slope of a relationship, a common grey-scale has been used throughout and the background set to mid-grey.

OPEN ACCESS Freely available online

PLoS one

Testing the Water–Energy Theory on American Palms (Arecaceae) Using Geographically Weighted Regression

Wolf L. Eiserhardt[1], Stine Bjorholm[1], Jens-Christian Svenning[1], Thiago F. Rangel[2], Henrik Balslev[1]*

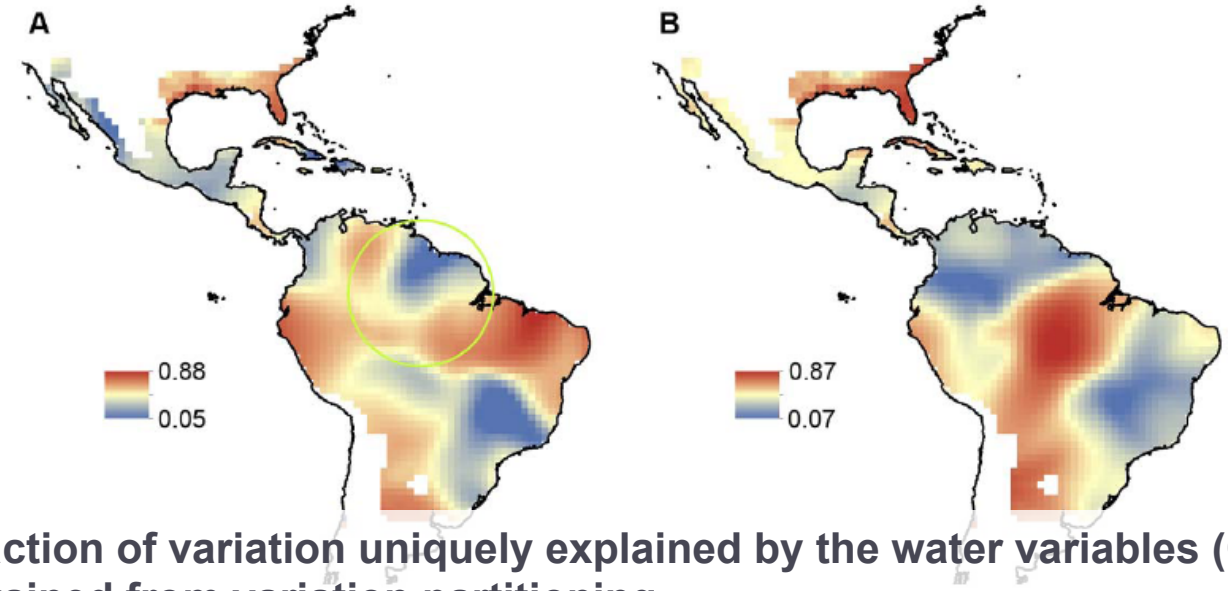
1 Ecoinformatics and Biodiversity Group, Department of Bioscience, Aarhus University, Aarhus, Denmark, **2** Departamento de Ecologia, ICB, Universidade Federal de Goiás, Goiânia, GO, Brazil

Abstract

Water and energy have emerged as the best contemporary environmental correlates of broad-scale species richness patterns. A corollary hypothesis of water–energy dynamics theory is that the influence of water decreases and the influence of energy increases with absolute latitude. We report the first use of geographically weighted regression for testing this hypothesis on a continuous species richness gradient that is entirely located within the tropics and subtropics. The dataset was divided into northern and southern hemispheric portions to test whether predictor shifts are more pronounced in the less oceanic northern hemisphere. American palms (Arecaceae, $n = 547$ spp.), whose species richness and distributions are known to respond strongly to water and energy, were used as a model group. The ability of water and energy to explain palm species richness was quantified locally at different spatial scales and regressed on latitude. Clear latitudinal trends in agreement with water–energy dynamics theory were found, but the results did not differ qualitatively between hemispheres. Strong inherent spatial autocorrelation in local modeling results and collinearity of water and energy variables were identified as important methodological challenges. We overcame these problems by using simultaneous autoregressive models and variation partitioning. Our results show that the ability of water and energy to explain species richness changes not only across large climatic gradients spanning tropical to temperate or arctic zones but also within megathermal climates, at least for strictly tropical taxa such as palms. This finding suggests that the predictor shifts are related to gradual latitudinal changes in ambient energy (related to solar flux input) rather than to abrupt transitions at specific latitudes, such as the occurrence of frost.

Citation: Eiserhardt WL, Bjorholm S, Svenning J-C, Rangel TF, Balslev H (2011) Testing the Water–Energy Theory on American Palms (Arecaceae) Using Geographically Weighted Regression. PLoS ONE 6(11): e27027. doi:10.1371/journal.pone.0027027

"There was strong spatial heterogeneity in palm richness–climate relationships, as evidenced by a minimum AICC difference between GWR and OLS models of ΔAIC_C = 663 (median 1515, maximum 3043)."



raction of variation uniquely explained by the water variables (C) and energy variables (D) btained from variation partitioning

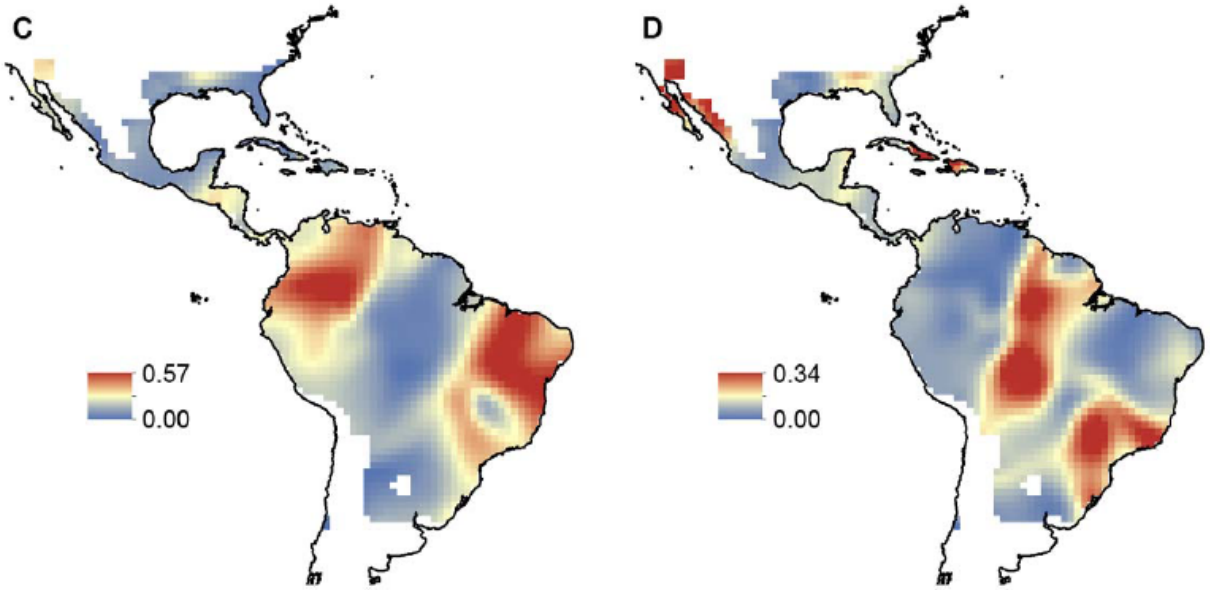


Figure 2. Variation in American palm species richness locally explained by water and energy. Local R^2 values obtained from geographically weighted regression (GWR) of palm species richness on annual precipitation, precipitation of the driest month, and water deficit (A) and mean annual temperature, minimum temperature of the coldest month, and potential evapotranspiration (B). Fraction of variation uniquely explained by the water variables (C) and energy variables (D) obtained from variation partitioning. The green circle in (A) shows the GWR bandwidth for a cell situated at the equator.

doi:10.1371/journal.pone.0027027.g002

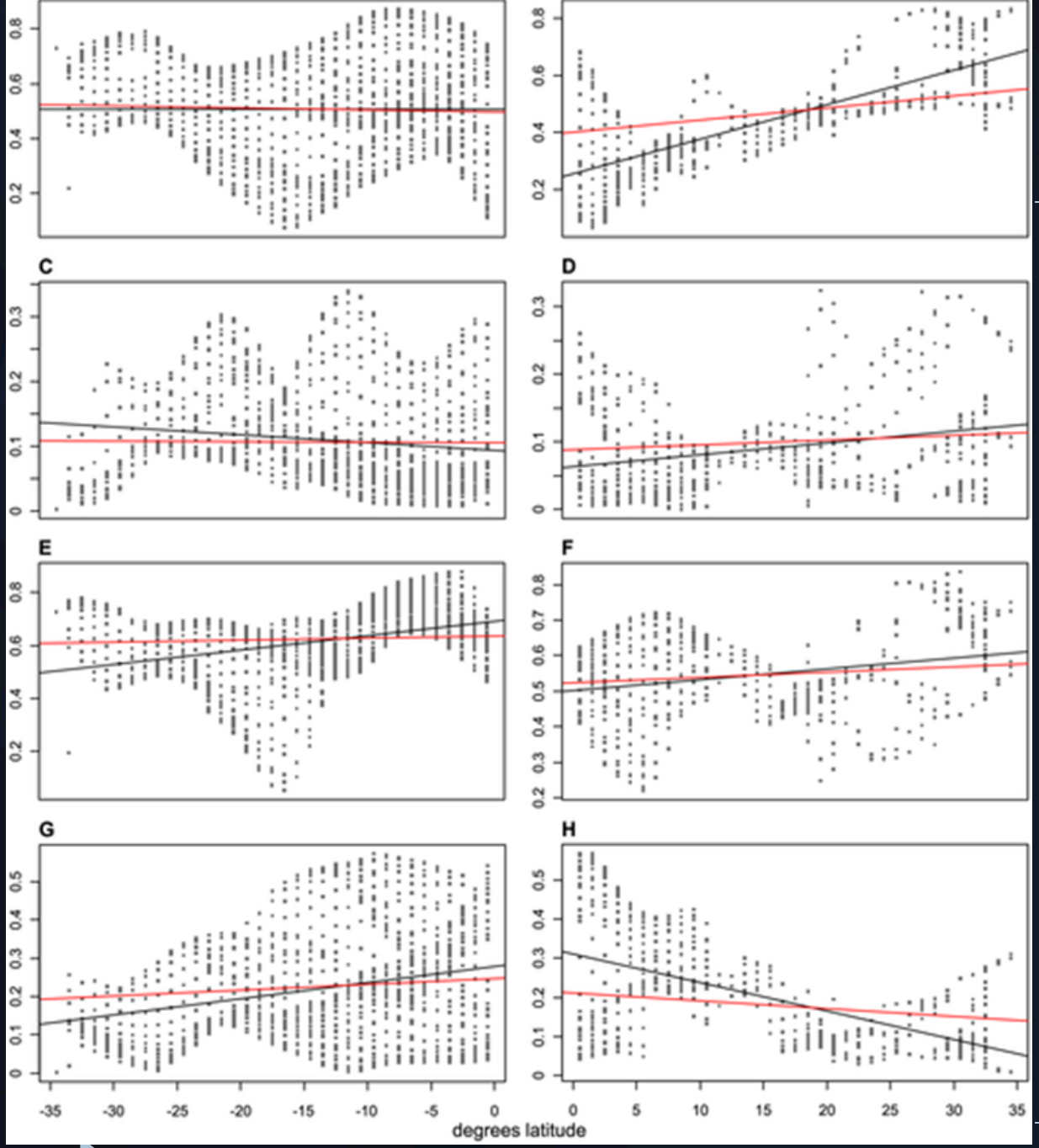


Figure 3. Latitudinal trends in the ability of water and energy to explain American palm species richness.

The amount of variation in palm species richness locally explained by energy variables (A–D) and water variables (E–G) plotted against latitude. A, B: total energy (Re). C, D: pure energy (Rpe). E, F: total water (Rw). G, H: pure water (Rpw). Regression lines obtained from OLS regression (black) and SAR regression (red).

International Journal of Geographical Information Science
Vol. 25, No. 3, March 2011, 423–438

Taylor & Francis
Taylor & Francis Group

Spatial nonstationarity and the scale of species–environment relationships in the Mojave Desert, California, USA

Jennifer A. Miller[a]* and Robert Q. Hanham[b]

[a]*Department of Geography and the Environment, The University of Texas at Austin, Austin, TX, USA;*
[b]*Department of Geology and Geography, West Virginia University, Morgantown, WV, USA*

(*Received 13 July 2009; final version received 19 August 2010*)

Species distribution models (SDM) have become a fertile area of research interest at the confluence of spatial ecology and GIScience and have been used to study a wide range of biogeographical phenomena, including invasive species, vector-borne diseases, and biological diversity. Scale is one of the most important considerations in any spatial analysis study because different spatial patterns emerge at different scales. An issue related to the 'extent' concept of scale that has more recently been recognized as important is spatial nonstationarity, which exists when processes or models of processes vary across space. This research examined the scale of species–environment relationships by a relatively new (in SDM) statistical method, geographically weighted regression (GWR). We tested four different types of species and 10 different types of environmental (climate and topography) variables in univariate GWR models to explore how stationarity and explanatory power varied with scale (as a function of GWR bandwidth size). The results suggest that the scale of species–environment relationships varies for both different types of species and different types of environmental variables. The two metrics used here – stationarity index and explained variance – did not show congruity in terms of a 'characteristic scale.' Species' relationships with climate and elevation variables became stationary at broader scales, and in some cases the models did not become stationary at the largest bandwidth tested. The complex topographic variables used here operate at finer scales and were often stationary across all scales or became stationary at small bandwidths. In addition to being instrumental for examining the effects of scale on spatial nonstationarity and a model's explanatory ability, GWR can also be used to explore potential geographical factors that result in nonstationarity.

Keywords: species distribution model; nonstationarity; scale; geographically weighted regression

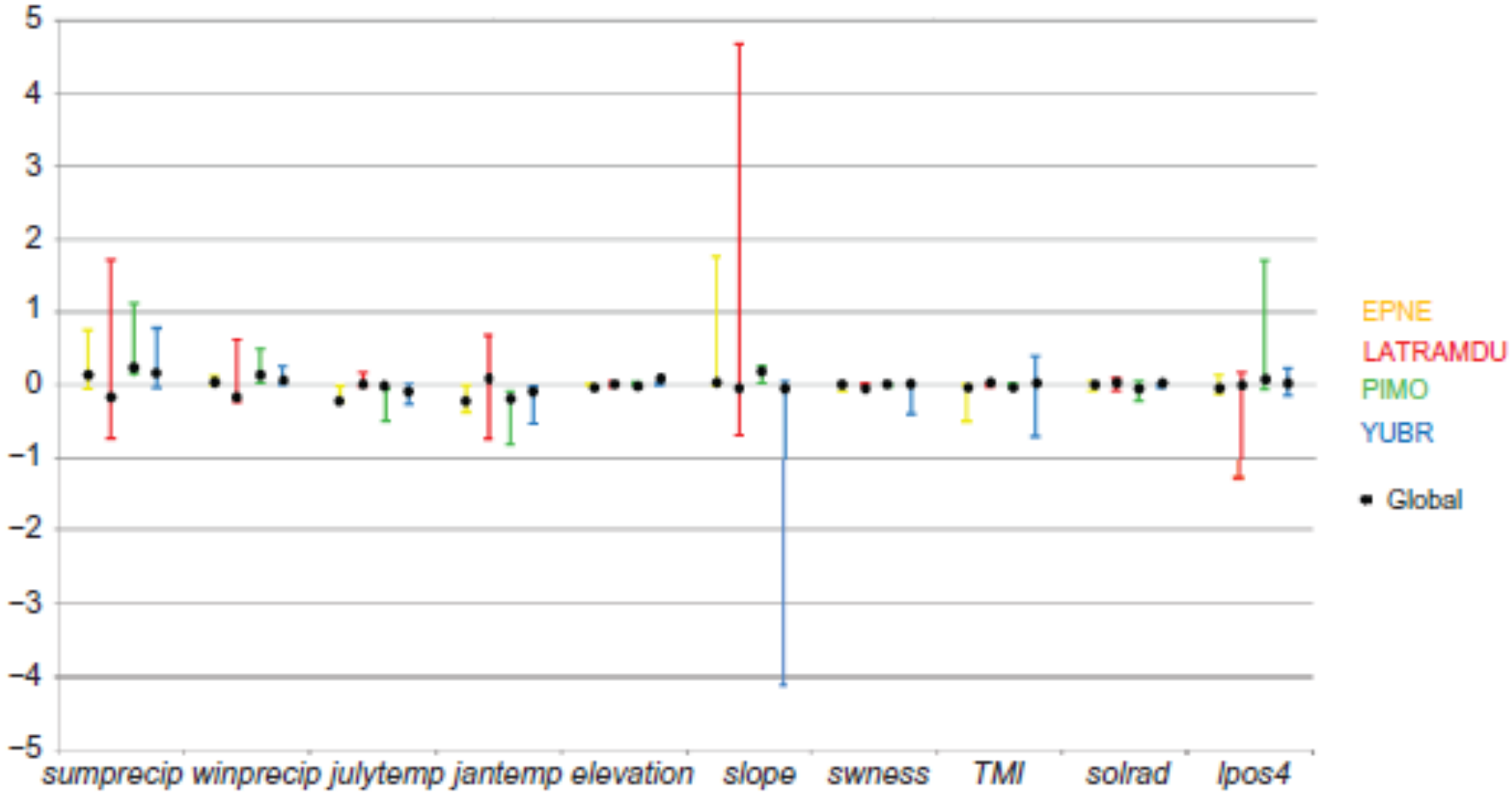


Figure 3. Coefficients for global and GWR models (linear only). GWR model coefficients are given as ranges at the smallest available bandwidth. Global coefficient values are indicated by black circles.

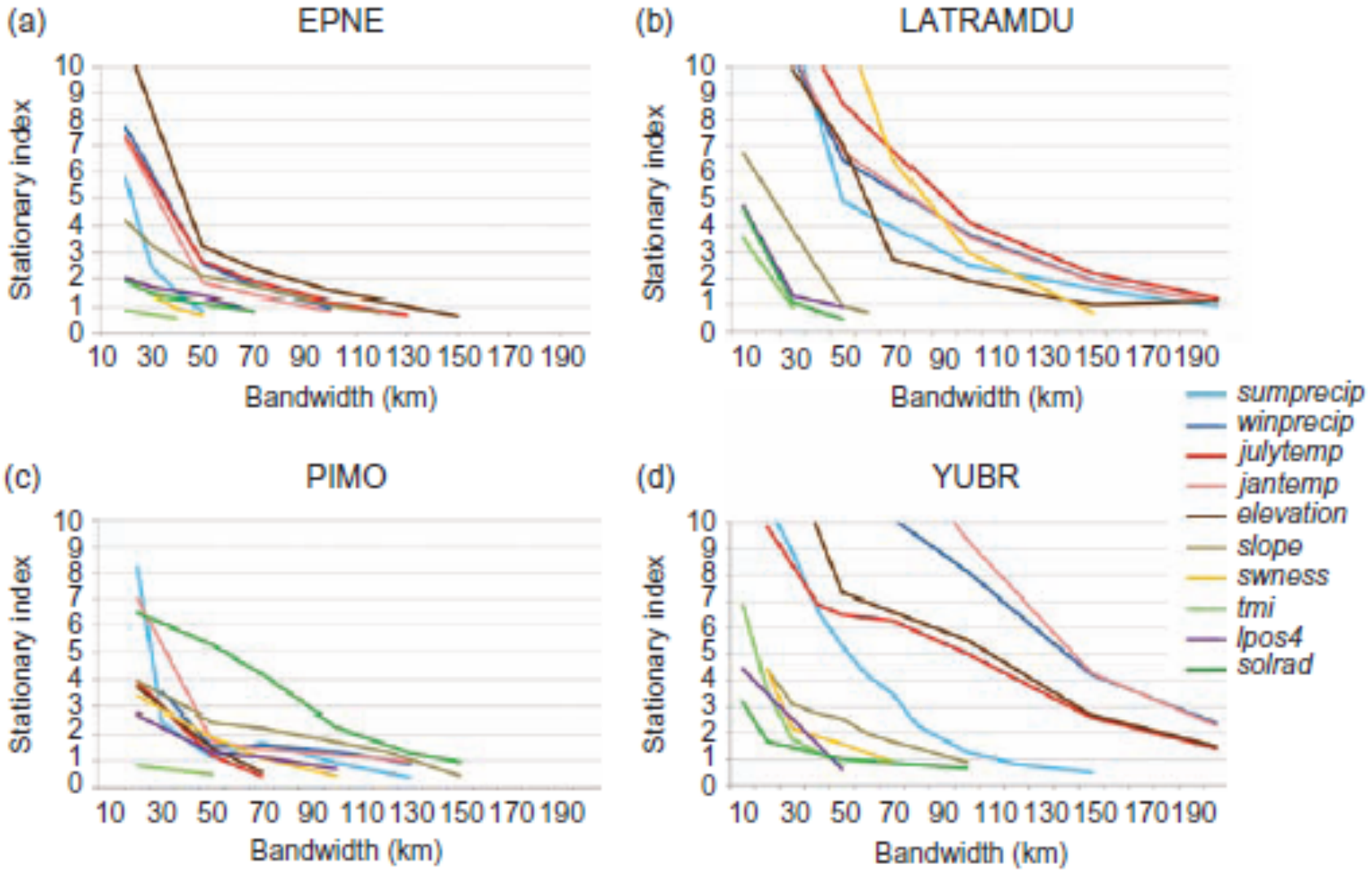


Figure 4. Stationarity index for each species–environment model combination across all bandwidths for (a) EPNE, (b) LATRAMDU, (c) PIMO, and (d) YUBR. SI < 0 indicates stationarity.

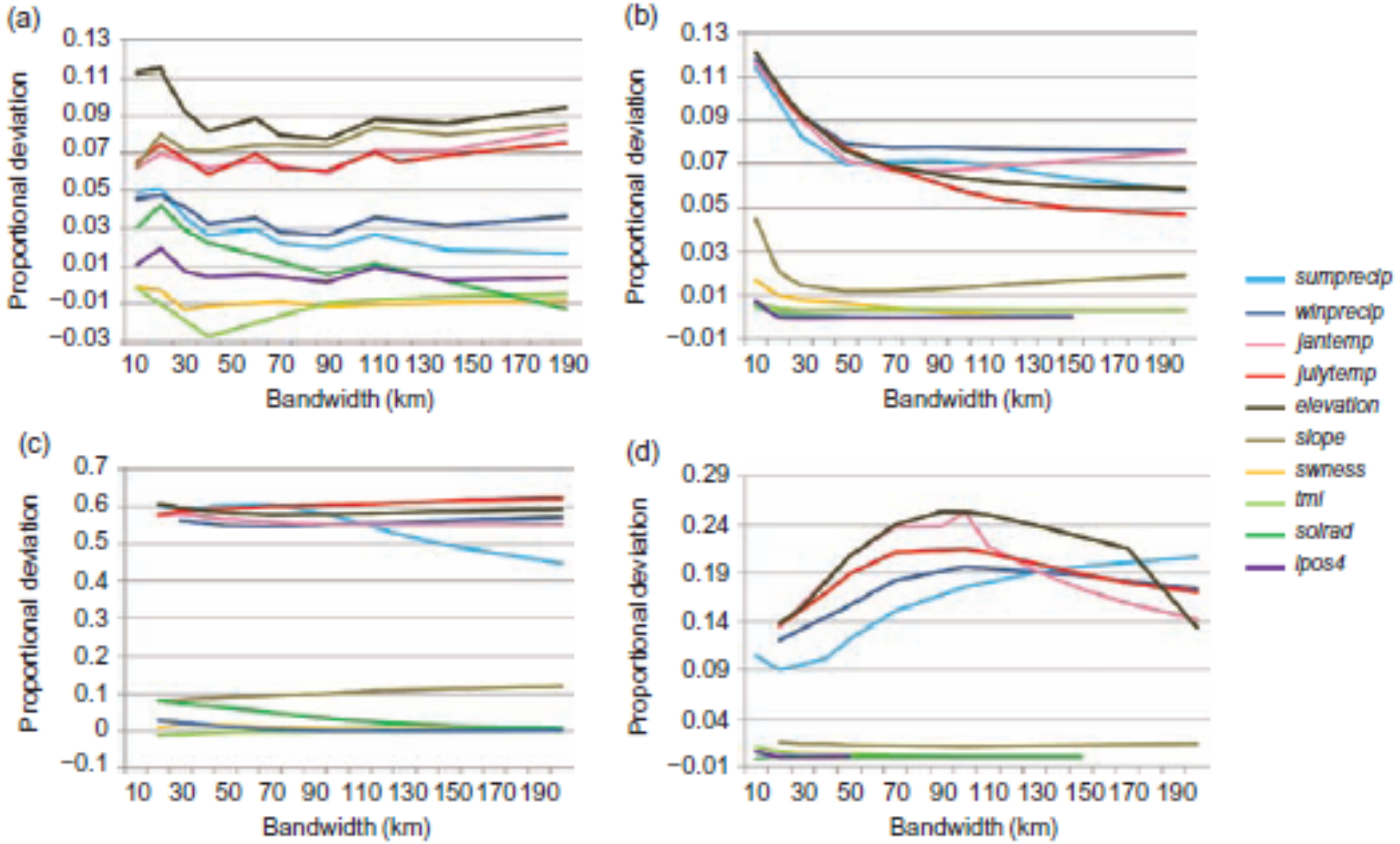


Figure 5. Proportional deviation from the null expectation based on average R^2 of 50 random predictors for (a) EPNE, (b) LATRAMDU, (c) PIMO, and (d) YUBR.

Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic

Matthew J. S. Windle, George A. Rose, Rodolphe Devillers, and Marie-Josée Fortin

Windle, M. J. S., Rose, G. A., Devillers, R., and Fortin, M-J. 2010. Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. – ICES Journal of Marine Science, 67: 145–154.

Analyses of fisheries data have traditionally been performed under the implicit assumption that ecological relationships do not vary within management areas (i.e. assuming spatially stationary processes). We question this assumption using a local modelling technique, geographically weighted regression (GWR), not previously used in fisheries analyses. Outputs of GWR are compared with those of global logistic regression and generalized additive models (GAMs) in predicting the distribution of northern cod off Newfoundland, Canada, based on environmental (temperature and distance from shore) and biological factors (snow crab and northern shrimp) from 2001. Results from the GWR models explained significantly more variability than the global logistic and GAM regressions, as shown by goodness-of-fit tests and a reduction in the spatial autocorrelation of model residuals. GWR results revealed spatial regions in the relationships between cod and explanatory variables and that the significance and direction of these relationships varied locally. A *k*-means cluster analysis based on GWR *t*-values was used to delineate distinct zones of species–environment relationships. The advantages and limitations of GWR are discussed in terms of potential application to fisheries ecology.

Keywords: Atlantic cod, fisheries ecology, generalized additive models, geographically weighted regression, logistic regression, non-stationarity, Northwest Atlantic, spatial modelling.

Received 6 April 2009; accepted 6 August 2009; advance access publication 4 September 2009.

M. J. S. Windle and G. A. Rose: Fisheries Conservation Group, Marine Institute, Memorial University of Newfoundland, St John's, NL, Canada A1C 5R3. R. Devillers: Department of Geography, Memorial University of Newfoundland, St John's, NL, Canada A1C 5R3. M-J. Fortin: Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada M5S 3G5. Correspondence to M. J. S. Windle: tel: +1 709 778 0504; fax +1 709 778 0669; e-mail: mattwindle@mi.mun.ca.

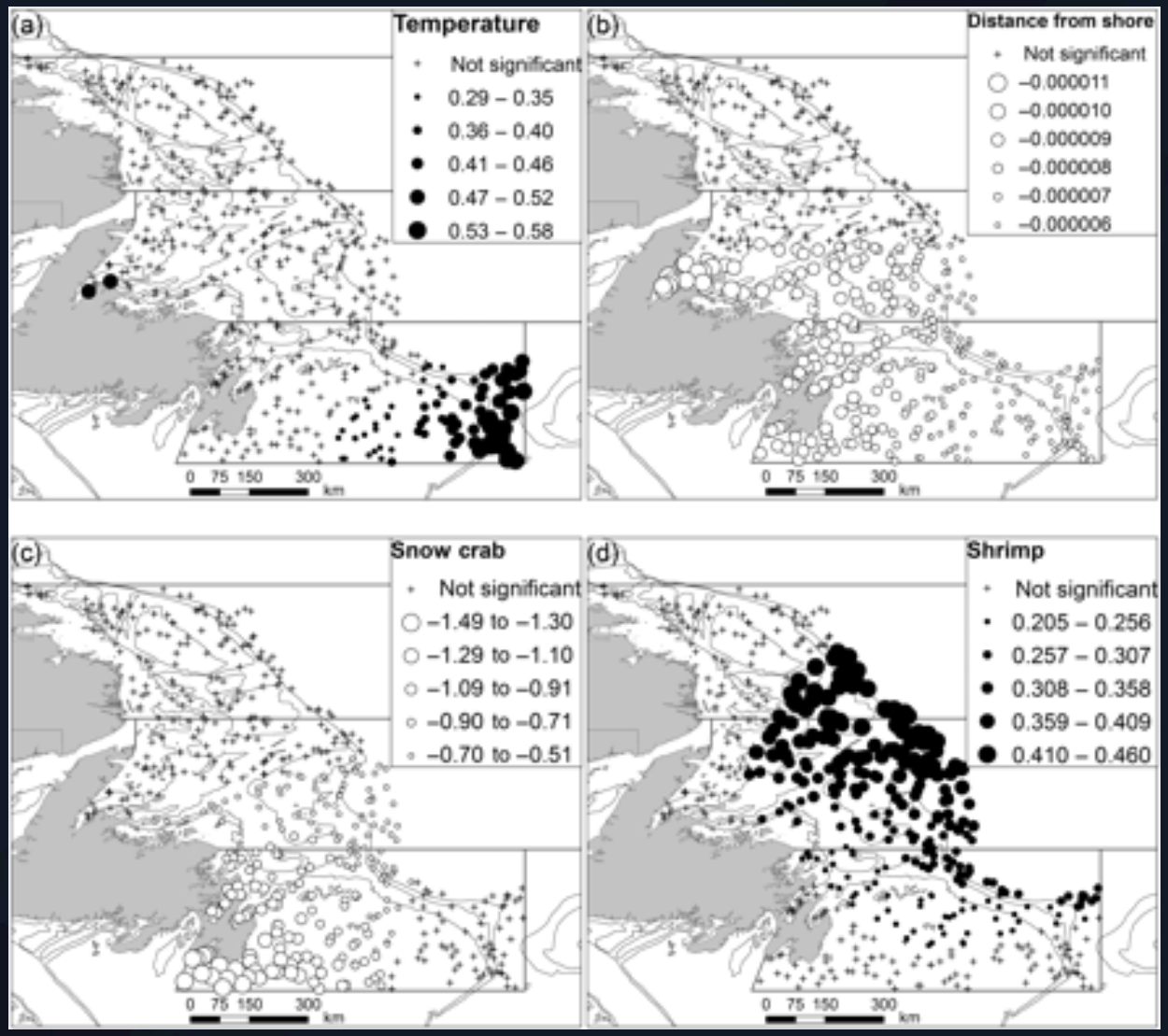
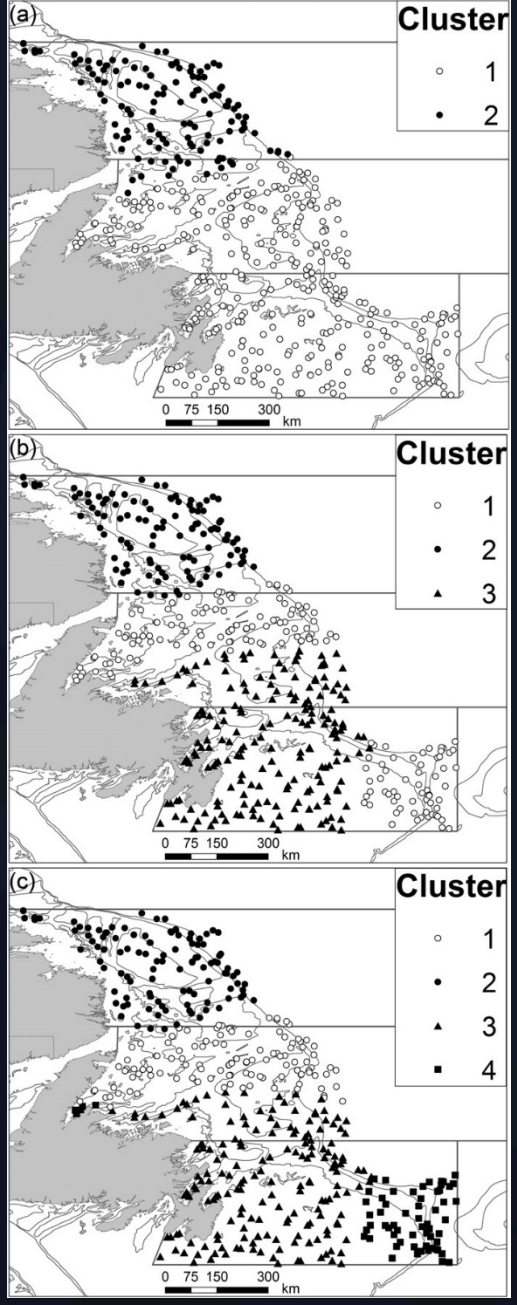


Figure 5: GWR-derived local coefficient estimates for (a) temperature, (b) distance from shore, and (c) snow crab and (d) shrimp as predictors of cod presence/absence in the 2J3KL region for autumn 2001. Positive values are shown as filled circles and negative values as unfilled circles. A significant threshold of 95% was used to mask out points where the relationship between cod and the predictor variable was not significant (plus signs).



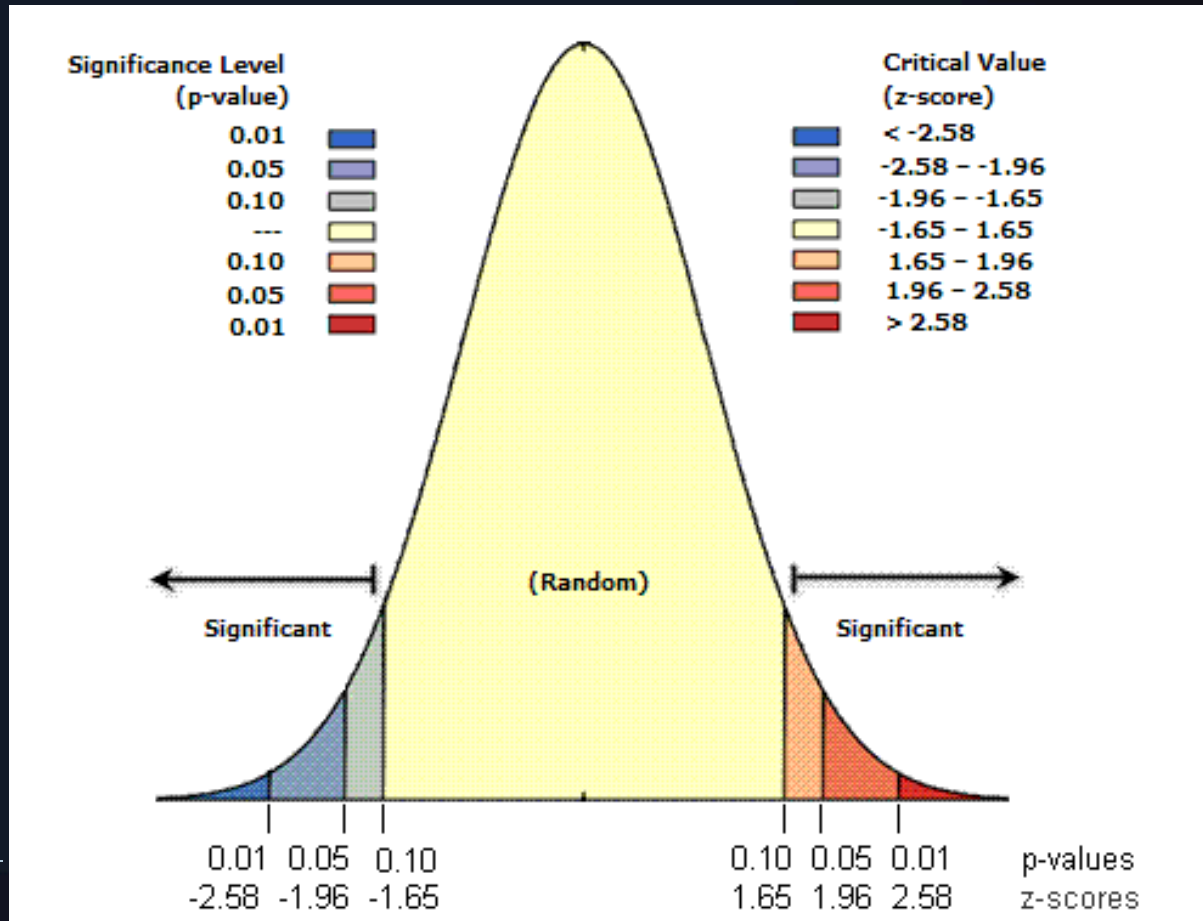
k	*c*	*n*	Temperature	Distance from shore	Snow crab	Shrimp
2	1	351	0.338 (0.086)	−0.000008 (0.000001)	−0.601 (0.329)	0.270 (0.069)
	2	130	0.321 (0.122)	0.000002 (0.000006)	0.247 (0.308)	0.436 (0.024)
3	1	172	0.408 (0.071)	−0.000008 (0.000001)	−0.316 (0.189)	0.317 (0.074)
	2	113	0.308 (0.126)	0.000003 (0.000005)	0.317 (0.265)	0.441 (0.020)
	3	196	0.283 (0.040)	−0.000008 (0.000001)	−0.818 (0.238)	0.241 (0.051)
4	1	117	0.369 (0.068)	−0.000008 (0.000001)	−0.418 (0.126)	0.367 (0.035)
	2	112	0.308 (0.126)	0.000003 (0.000005)	0.321 (0.263)	0.441 (0.020)
	3	186	0.285 (0.042)	−0.000008 (0.000001)	−0.829 (0.239)	0.234 (0.045)
	4	66	0.452 (0.067)	−0.000007 (0.000001)	−0.176 (0.201)	0.234 (0.026)

Table 4: Mean GWR parameter estimates from the model of 2J3KL cod distribution in autumn 2001 (212 km bandwidth) for each group identified by the k-means cluster analyses (s.d. in parentheses).

Windle M J S et al. ICES J. Mar. Sci. 2010;67:145-154

Inference with local statistics

- Random null hypothesis means z scores > 1.96 and < -1.96 are unusual



Inference with local statistics

- Random null hypothesis means z scores > 1.96 and < -1.96 are unusual

- Expected G_i is the proportion of the study area accounted for by the neighborhood of location *i*
  - *(under assumption of random distribution of attribute values…)*

How many “statistically significant” hot or cold spots are in this study area?

How many "statistically significant*" hot or cold spots are in this study area?

- 2034/5226 = 39%

* Compared to CSR/IRP, but we already determined that these data exhibit + SAC… (stat sig + Moran's *I*)

- **<u>Statistical tests of local statistics are inherently non-independent</u>**

How many "statistically significant*" hot or cold spots are in this study area?

- 2034/5226 = 39%

Statistical tests of local statistics are inherently non-independent

- Some solutions/strategies:
  - *Bonferroni correction*
    - $a' = a/n$; $a' = 0.05/3085 = 0.0000162$; $z = 4.15$
    - (http://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html)
  - Conditional Monte Carlo simulations (of attribute values) to determine *pseudosignificance* values

Simulation to get distribution of (random) values

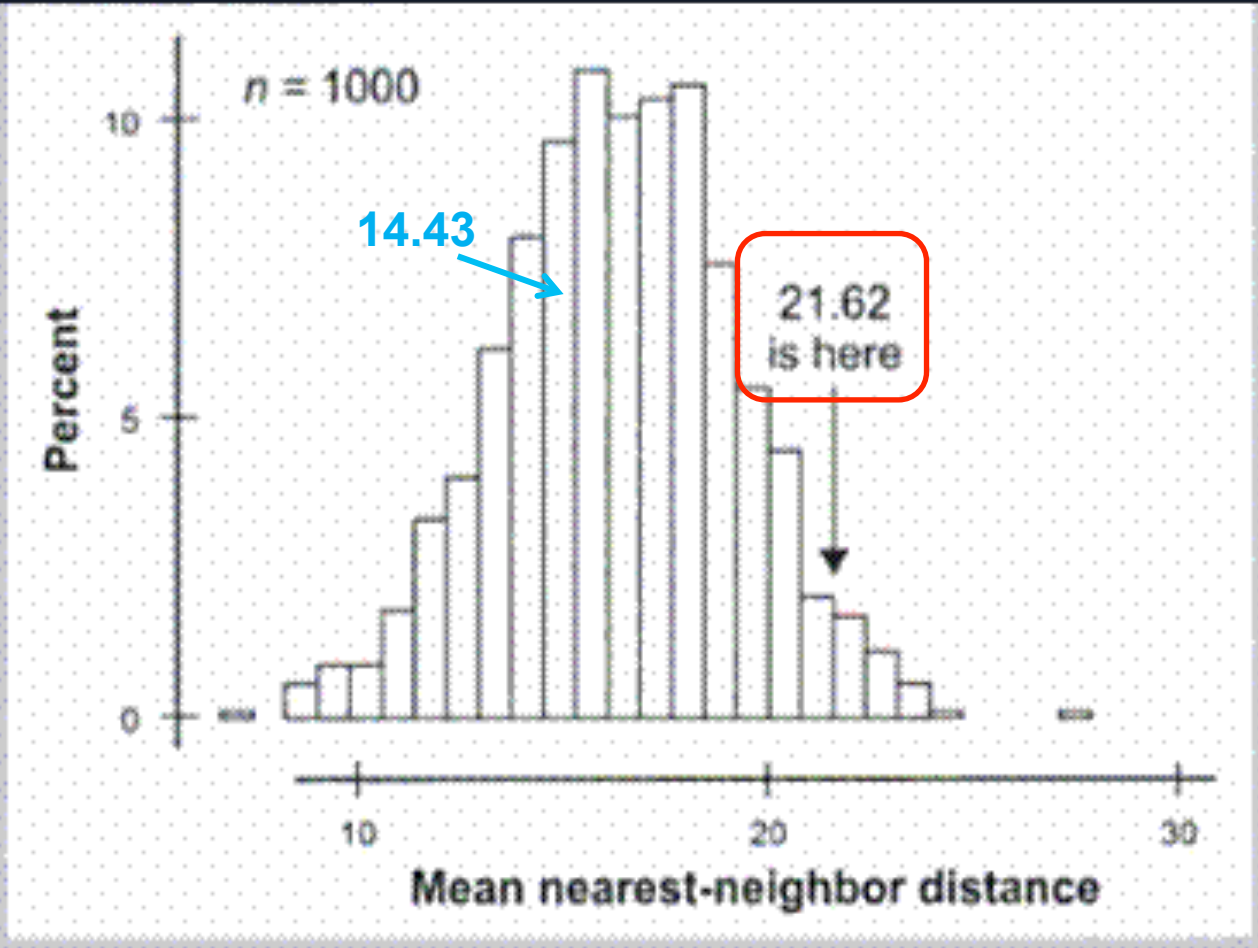


Figure 4.20 Results of a simulation of IRP/CSR for 12 events (compare Table 4.2).