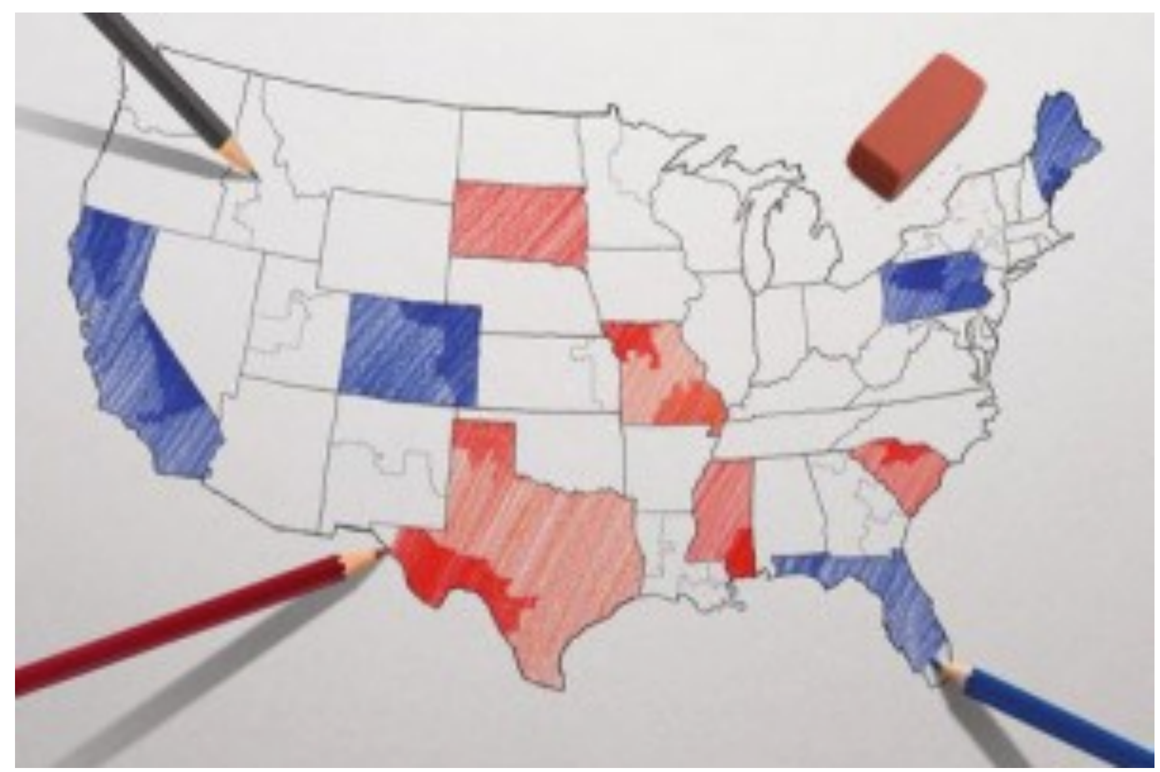
Choosing the Right Territory

Geospatial Data & Spatial Statistics in Insurance Analytics
Special Topic: Modifiable Areal Unit Problem (MAUP)

Satadru Sengupta

Liberty Mutual Group



Casualty Actuarial Society Annual Meeting
Chicago
November 2011

Antitrust Notice

- *The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.*

- *Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.*

- *It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.*

Next 45 Minutes

Spatial Statistics: A Statistical Framework for Geospatial Data in Insurance

- ***Motivation***
  - *Need of Strategic Growth, Targeted Products, Accurate Pricing and Efficient Operations*
  - *Availability of Geospatial Data*
  - *Availability of Robust Database Management System: Geographical Information System (GIS)*
  - *Availability of A Statistical Framework for Analyzing Geospatial Data*
- ***Geospatial Data Generating Process (DGP)***
  - *Stochastic Process, Random Fields and Analogy to Time Series Data*
  - *Elements of Geospatial Data - Spatial Index and Spatial Correlation*
- ***A Special Topic: Modifiable Areal Unit Problem (MAUP) - Aggregating Spatial Data***
  - *Gerrymandering*
  - *Elements of MAUP - Scale Effect and Zoning Effect*
- ***A Spatial Econometric Model***
  - *Spatial Simultaneous Autoregressive Error Model (SAR Error Model)*
  - *Comparison with GAM & Different Spatial Correlation Measurement*
- ***Conclusion***

Motivation

Tobler's First Law of Geography, Waldo R. Tobler, 1970

- **Statistical modeling and analysis starts with a perspective on the data to be analyzed**
- **A Model is built to predict the outcomes of a Process by mimicking the True Process**
  - **Physicists build Large Hadron Collider to find the law of nature**
  - **Flight Simulator**
  - **Actuaries build Mathematical Model to mimic a True Process**
- **Elements of a Spatial Data Generating Process (DGP)**

  I. Spatial Index - *Takes the data from the table and shows them on a Map*

  II. Spatial Correlation *- A relationship among the data points (as a function of the spatial index). Essentially, the observations are no more independent*

- **Location Matters**

  I. Observed value at one location is influenced by the observed values at other locations in a geographic area: ***There is an underlying correlation***

  II. Influence declines with distance: ***Decay in correlation with increasing distance***

  III. Influence can be positive as well as negative: ***Correlation can be positive or negative***

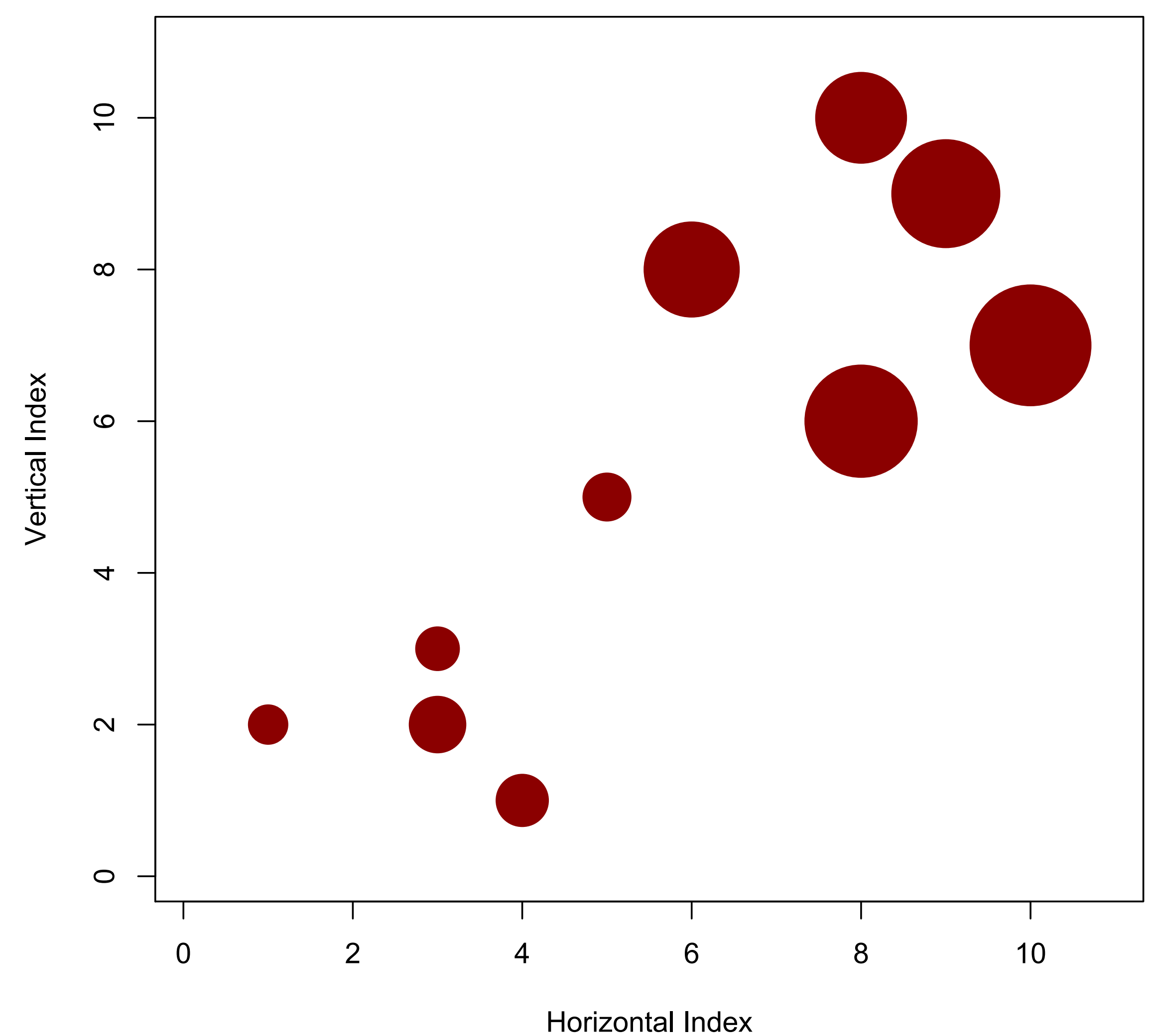
Index and Correlation in a Dataset

A Simple Illustration

How location indices increase the information in a dataset

Record	Location Index 1	Location Index 2	Observation
1	**9**	**9**	25
2	**5**	**5**	11
3	**3**	**3**	10
4	**6**	**8**	22
5	**1**	**2**	9
6	**8**	**10**	21
7	**4**	**1**	12
8	**3**	**2**	13
9	**8**	**6**	26
10	**10**	**7**	28

Table to Map - How Location Index Helps



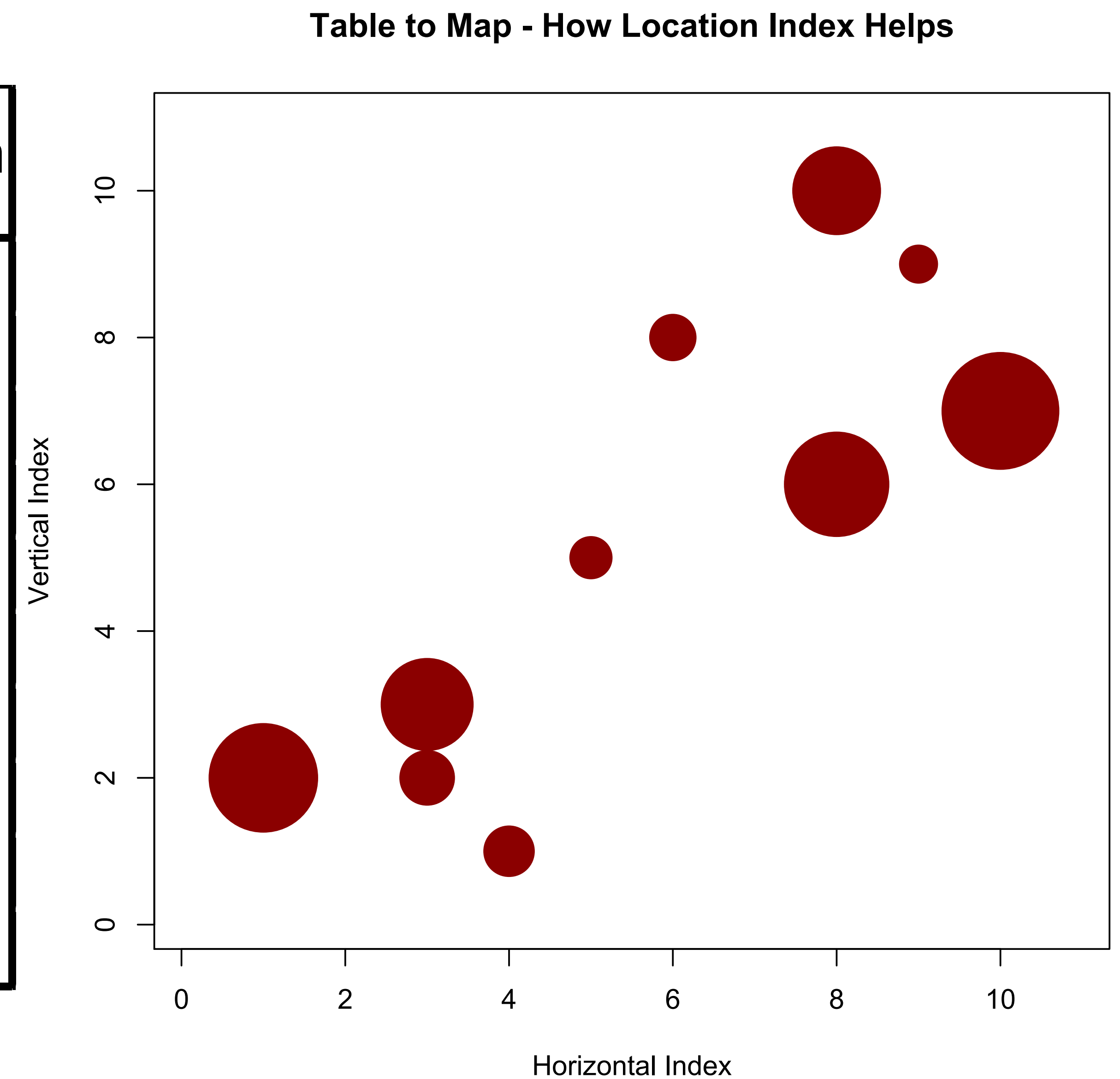
A map can show us what we can't see otherwise...

Index and Correlation in a Dataset

A Simple Illustration

Now let's shuffle the location indices of this data (rearranging the yellow columns)

Record	Location Index 1	Location Index 2	Observation
1	8	6	25
2	6	8	11
3	5	5	10
4	3	3	22
5	9	9	9
6	8	10	21
7	4	1	12
8	3	2	13
9	1	2	26
10	10	7	28



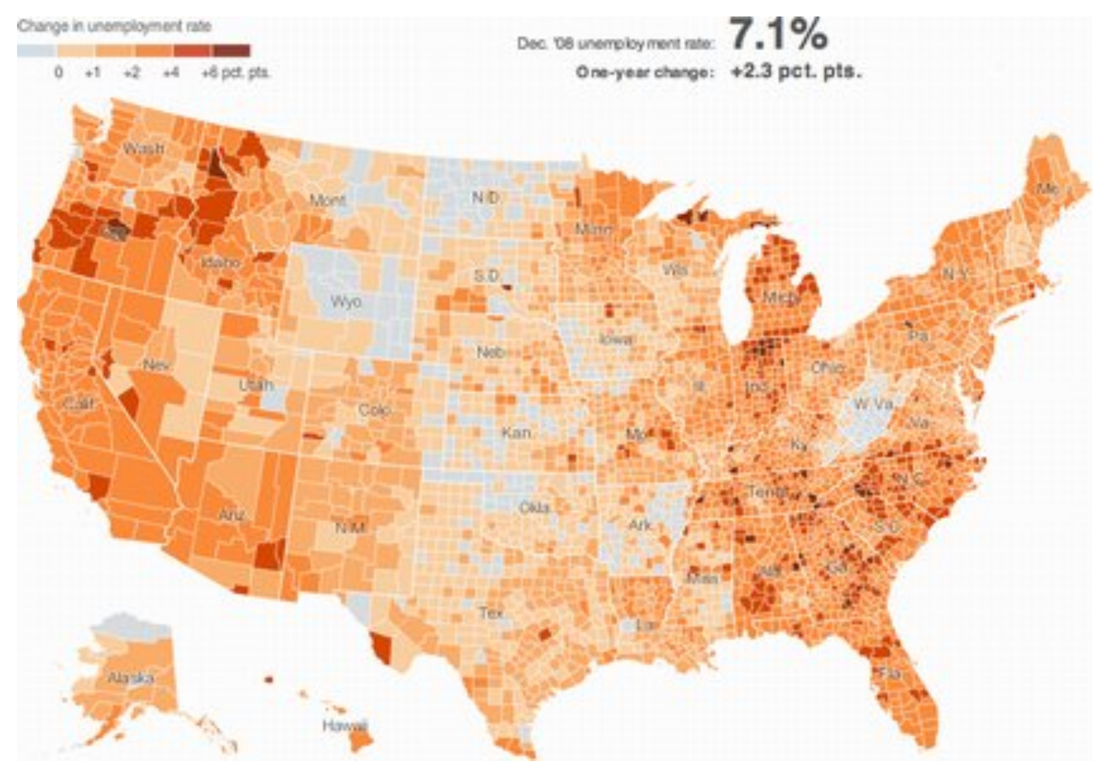
By changing the location index, we have lost the correlation in the data

Even for Non-Geographic Data

“Everything is related to everything else, but near things are more related than distant things

- **Concept of Near - Defining a “Cohort” - Spatial and Non-Spatial**
    - Euclidean distance, Territory with common boundaries, Transit distance (Manhattan distance)
    - Insured sharing the same Fire Station, Cars with Same Make and Similar Model (Car Symbols)
    - Friends in Facebook, Contacts in Linked-in, Contamination and Disease Propagation

- **Analyzing a Map or Network based Data Generating Process is asking two questions:**
    - How we can get those data and put them on a map
    - How we can quantify the interdependencies among the data points (nodes in the network, points in the map)



Mathematical Interpretation

Data Generating Process - Non-Spatial vs. Spatial

- ***Task - Regression in a Geographic Region:***
  - *Housing Prices in a State*
  - *Area with high crime rate in City - Crime Hotspot*
  - *Homeowners Insurance*
  - *Pollution Insurance*
  - *Primary Care Physician Availability in a Region*

- ***Assume A Non-Spatial Data Generating Process (DGP) : Good Old Regression Model***
  - *For location **i** and **k** in the region*

$$Y_i = X_i \beta + e_i$$

$$Y_k = X_k \beta + e_k$$

$$e_i, e_k \sim N(0,\sigma^2)$$

  - ***Conditional independence of the observed values*** *- observed value Y_i at location **i** is independent of observed value Y_k at location **k** (in a fully specified model)*

  - ***Independence of residuals*** *- e_i and e_k are independent*

Mathematical Interpretation

Data Generating Process - Non-Spatial vs. Spatial

- ***Spatial Data Generating Process*** *- For location **i** and **k** in the region*

$$Y_i = \alpha_k Y_k + X_i \beta + e_i$$

$$Y_k = \alpha_i Y_i + X_k \beta + e_k$$

$$e_i , e_k \sim N(0,\sigma^2)$$

- ***Spatial dependence of the observed values*** *- observed value* Y_i *at location **i** is influenced by the observed value* Y_k *at location **k***

- ***Motivation for an Spatial Econometric Model***

  1. ***A Time Dependence Motivation***

  2. ***Omitted Variable Motivation***

  3. ***A Spatial Heterogeneity Motivation - Panel Data***

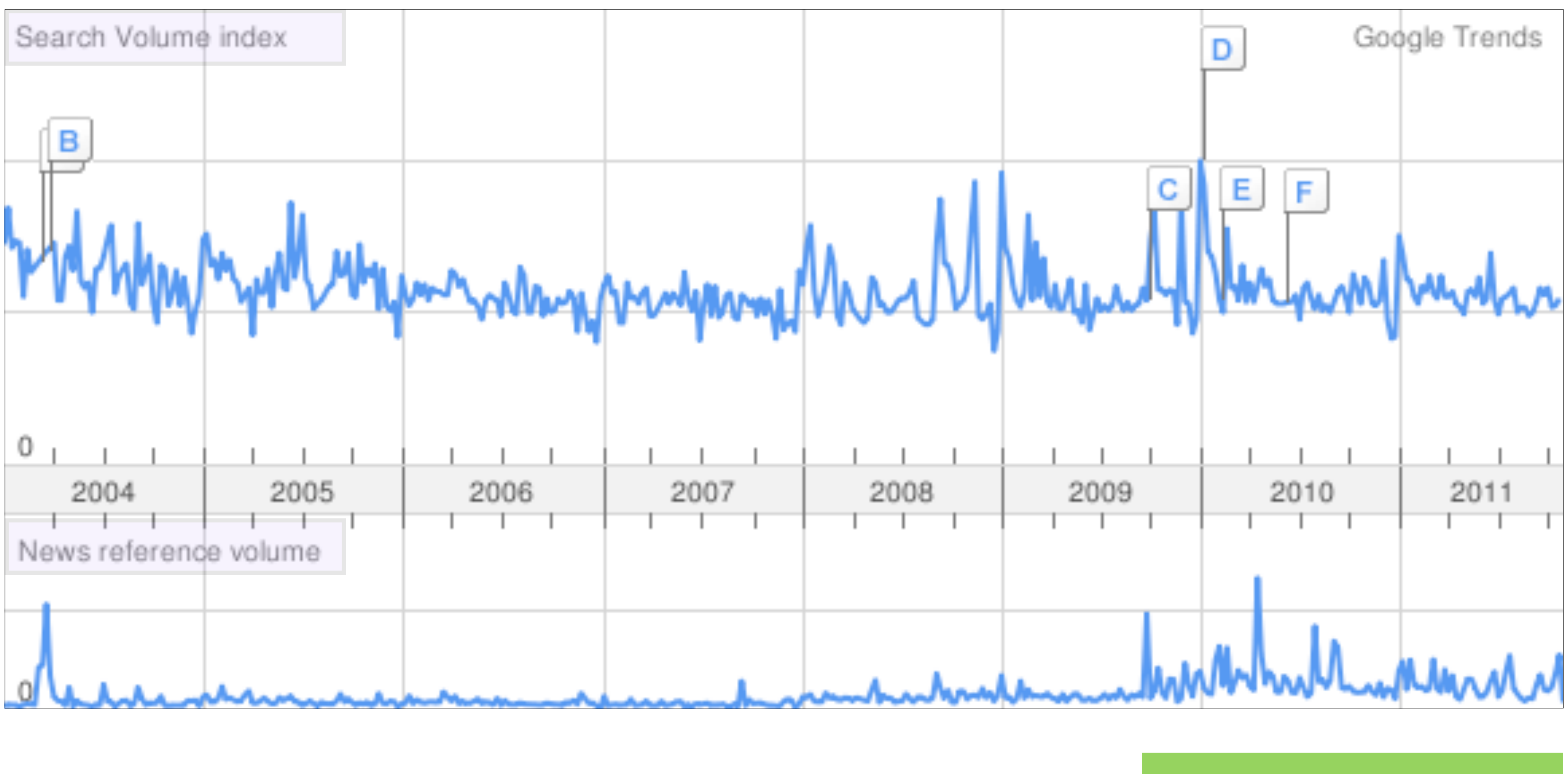
  4. ***An Externalities based Motivation - Positive or Negative Externalities***

- *For a detailed study refer to "Introduction to Spatial Econometrics" by James LeSage, R Kelley Pace (CRC Press)*

Spatial Data & Analogy to Time Series

Generic Stochastic Process and Random Field

- ***Stochastic Process : { Y(s) : s in D }*** *where* ***Y(s)*** *is* ***Random Observation, s*** *is an* ***Index*** *set from* ***D, a subset of*** R^r ***(r-dimensional Euclidean space)***
- ***Time Series -*** *Special case of stochastic process where index set s is 1-dimensional Euclidean space:* ***{*** Y_t ***: t in {1,2,3,4,...}}***
- ***How often the word "Actuary" appears in the online news and Google search?***



The word "Actuary" is appearing more often in the News starting mid-2009

source: http://www.google.com/trends

Three Types of Spatial Data

Stochastic Process, Random Field and Spatial Data

- ***Random Field -*** *When the Domain D is from a multi-dimensional Euclidean space (* ***r > 1*** *)*
- ***In simple words:*** *Random Field is a list of correlated random observations that can be mapped onto a r-dimensional space*
- ***Spatial Data Generating Process - The Process generates spatial data for r = 2***

{ Y(s) : s in D } where D is a subset of R^2

  - ***Coordinate Reference System (CRS)*** *- Latitude, Longitude, Northing, Easting, Different Projections*
  - ***Induced Covariance Structure -*** *Observations are spatially correlated based on a covariance function*

Three Types of Spatial Data

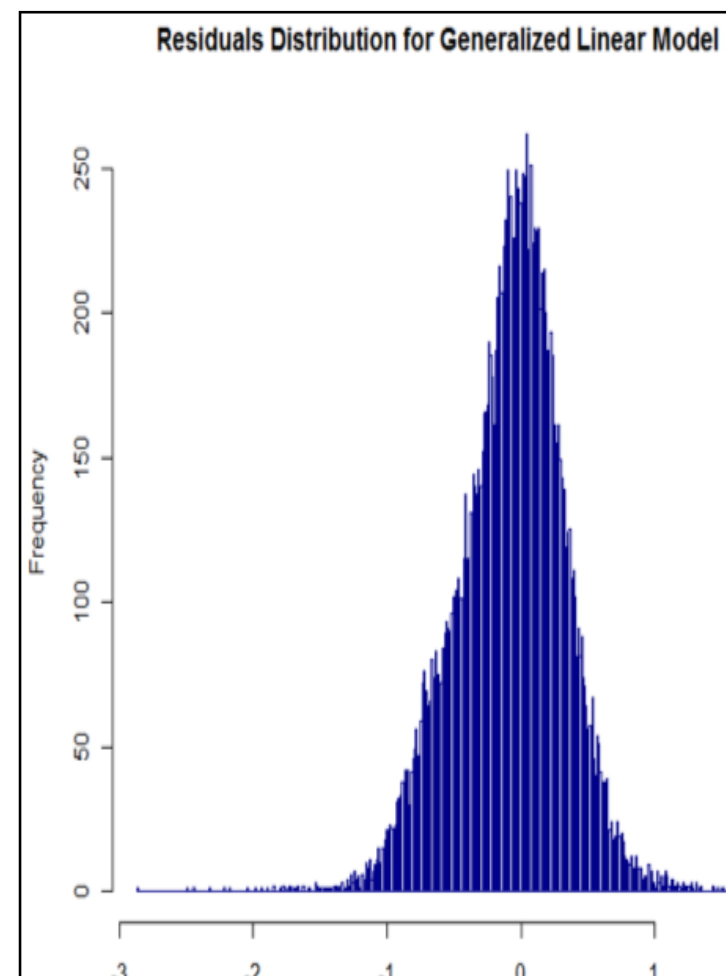
- *How s takes values in D (discrete/ continuous)?*
- *How D comes from R^2 (Fixed/ Random)?*
  - ***Point Referenced Data -*** *When s takes values in D continuously, D is a fixed subset of R^2*
  - *Temperature in Chicago (Possible to collect every point in Chicago)*
  - ***Lattice / Areal Data -*** *D is a fixed partitioned subset of R^2, $D = \{s_1, ..., s_n\}$, s assumes value from one of the partitions*
  - *Postal Zip Codes in Chicago - Non-overlapping Areal Unit*
  - ***Spatial Point Pattern Process -*** *The domain D itself is a random subset in R^2*
  - *Locations of Starbucks in Chicago - Are they more clustered in the Chicago Loop? Do their Cappuccinos taste better than the Starbucks at other places in the city?*

Point Referenced Data

Segmentation Pricing

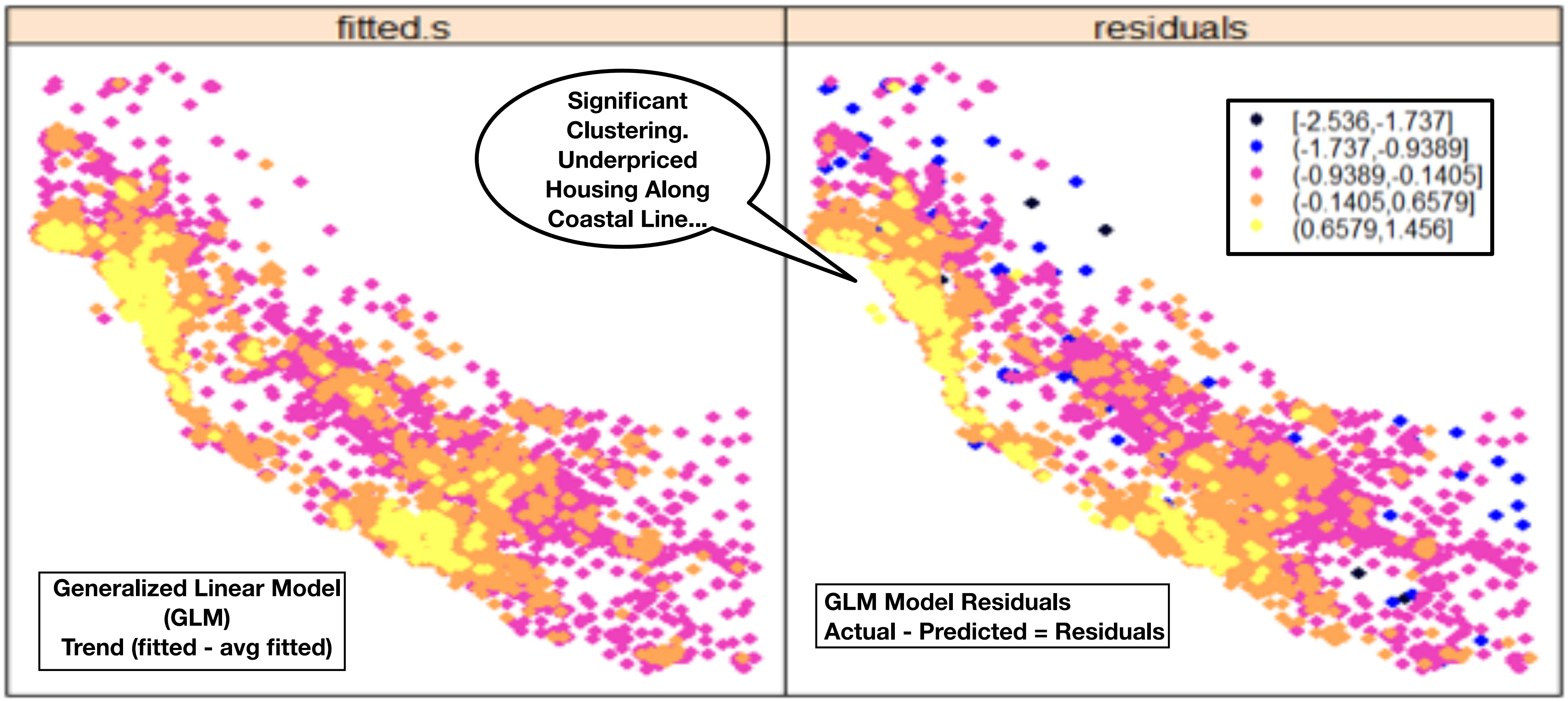
- ***Analysis and inference of Stochastic Process { Y(s) : s** runs **continuously** in **D } : D** is a **fixed subset** of **R^2***
- ***Common Practical Interest in Geostatistics***
    - *Given the observations in different location { $Y(s_1)$,,, $Y(s_n)$} : How to optimally predict Y(s) at a new location s*
    - *Estimation of spatial averages under spatially correlated data*
    - *Diagnostic of existing model: Spatial clustering of residuals in study region*
- ***A Simple Illustration -** California Housing Data (GAM example data) by Census Block*
    - *A typical example of Areal Data, but we will treat as Point Referenced Data*
    - *Assuming the data is a random selection of 20640 houses in California*
    - *Consider usual Generalized Linear Model (GLM) as in GAM Example*

```
glm(formula = value ~ income + I(income^2) + I(income^3) + log(age) +
     log(rooms) + log(bedrooms) + log(hh/pop) + log(hh), family = Gamma(link =
     log), data = ca.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.15154  -0.26238  -0.05152   0.15523   2.97976

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.3680015  0.0556922 204.122  < 2e-16 ***
income         0.6587304  0.0244209  26.974  < 2e-16 ***
I(income^2)   -0.0488121  0.0048773 -10.008  < 2e-16 ***
I(income^3)    0.0015019  0.0002929   5.127 2.97e-07 ***
log(age)       0.1924867  0.0060165  31.993  < 2e-16 ***
log(rooms)    -0.8568208  0.0171994 -49.817  < 2e-16 ***
log(bedrooms)  1.0472060  0.0261859  39.991  < 2e-16 ***
log(hh/pop)    0.2696699  0.0218861  12.322  < 2e-16 ***
log(hh)        0.0244465  0.0038982   6.271 3.65e-10 ***

    Null deviance: 6394.0  on 20639  degrees of freedom
Residual deviance: 2627.7  on 20631  degrees of freedom
AIC: 515354
```



GLM & Spatial Diagnostics

Independence of Residuals - Spatial Perspective

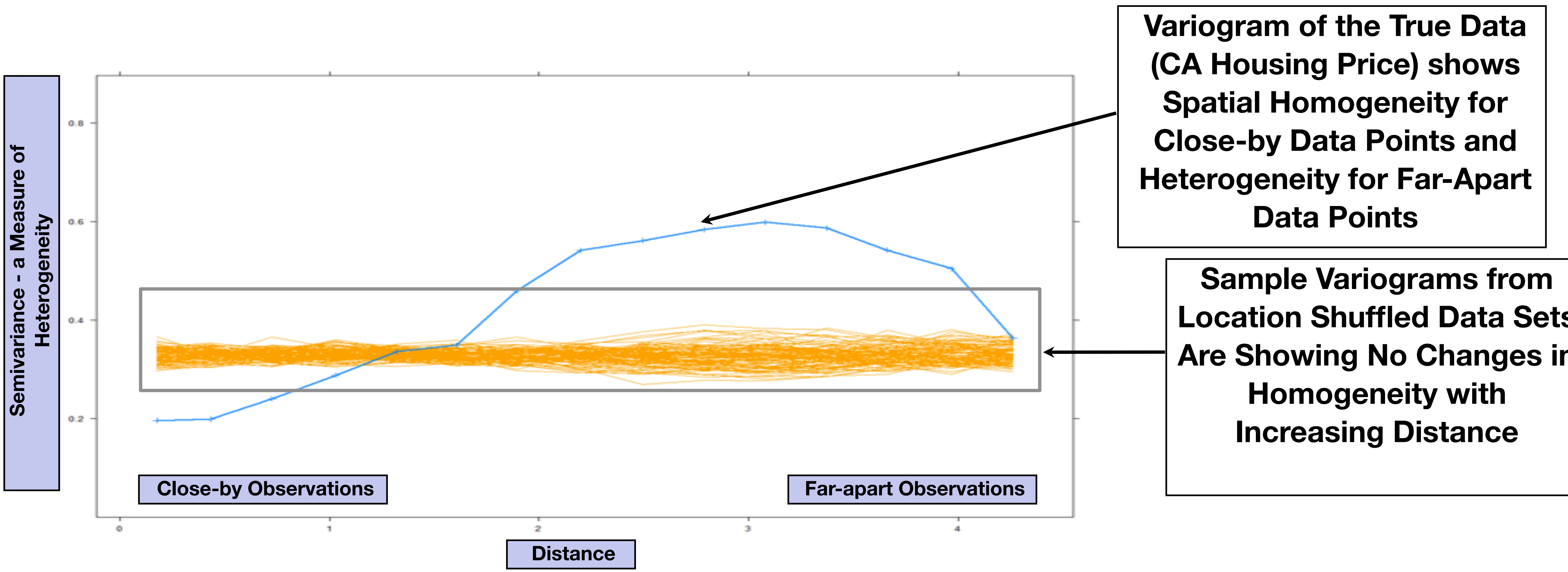


- *Residuals from the simple model are not distributed randomly over CA*
- *Model under-fits along coastline & Model over-fits in the locations away from coastline*
- *This example is an analogy to usual insurance adverse selection*
- *Can we show this Spatial Structure in a Quantitative Measure?*

Statistical Test for Spatial Correlation

Sample Variogram and Existence of Spatial Correlation

Variogram: $\gamma(h) = \frac{1}{2} * Var[Y(s+h) - Y(s)]$ ⟶ ***Higher the Semivariance Lower the Homogeneity Among Observations***



- *Calculate Variogram after re-assigning the observations (insured) randomly to different locations (street address) in the data (book of business) several times and obtain a 95% confidence interval*
- *Spatial Patterns become evident if the sample Variogram from true data falls outside the confidence interval*
- *Statistical packages can fit a parametric variogram to the sample variogram*
- *Some important parametric variogram: Linear, Exponential, Spherical, Gaussian, Matern*

Spatial Point Pattern Process

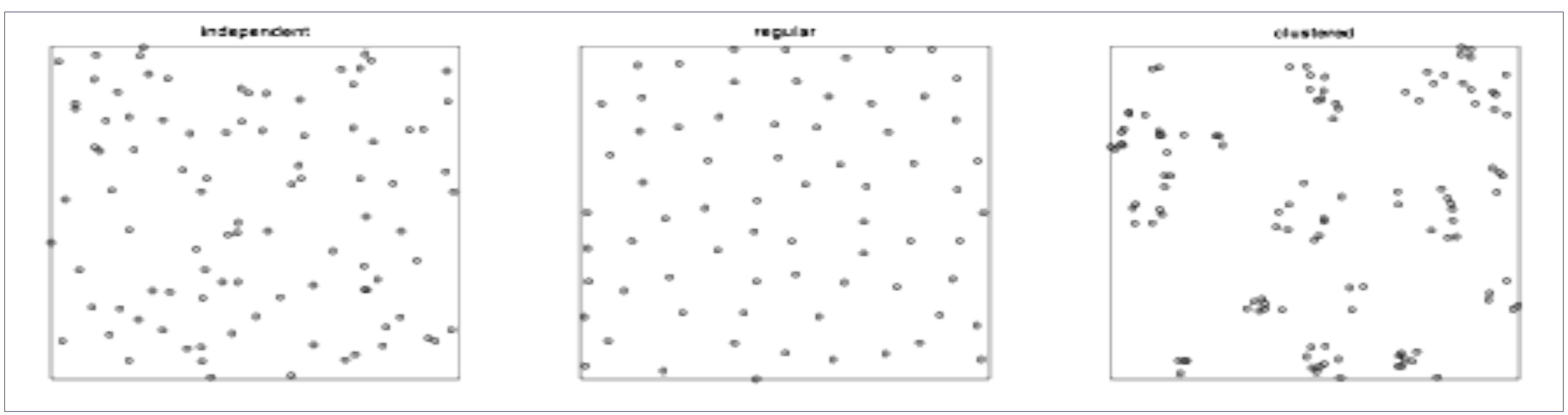
Exposure Concentration - Are we writing more business in the coastal area?

- ***Analysis and inference of Stochastic Process { Y(s) : s** in **D } : D** is a **random subset** of **R^2***

- ***Elements of Spatial Point Process:***
  I. ***First Order Properties - Distribution:** Spatial Distribution of Events - Intensity of Event Occurrence, Spatial Density*
  II. ***Second Order Properties - Interaction:** Clustering of Events, Independence*

- ***Complete Spatial Randomness (CSR) - Events occur independently and distributed uniformly over a geographic region***
  I. ***Clustering of Events** - Attraction between points over the region*
  II. ***Regularity of Events** - Presence of inhibition - Competition between points over the region*

- ***Spatial Poisson Process - Events occur independently and distributed according to a given intensity function λ(.) over a geographic region***
  I. ***Homogeneous Poisson Process (HPP)** - Intensity function is a constant : **$\lambda(x) = \lambda$***
  II. ***Inhomogeneous Poisson Process (IPP)** - Variable (often Parametric) Intensity function **$\lambda(x)$***

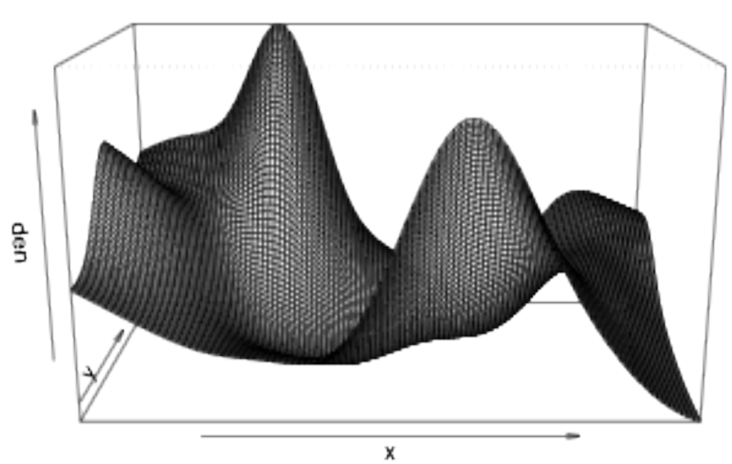


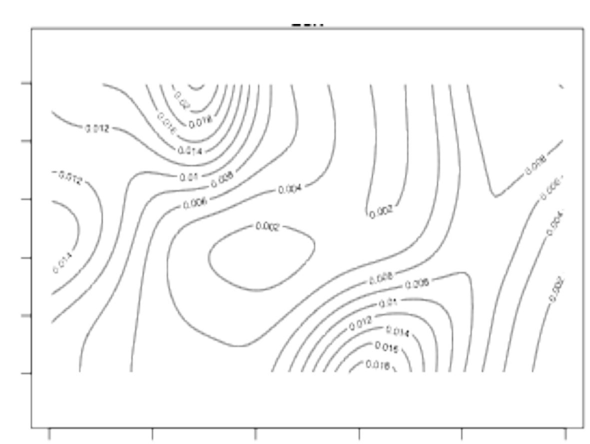
Spatial Poisson Process

Distribution of Events - A Similar Concept as Poisson Process in Waiting Time Problem

- ***HPP - Homogeneous Poisson Process -*** *A formalization of Complete Spatial Randomness (CSR)*
  - ***The number of events in a region W with area A is Poisson distributed with mean λA, where λ is the constant intensity of the process***
  - ***Given there is n number of events observed in the region W, they are uniformly distributed***

- ***Inference on the Poisson Process and Estimation of λ(x)***
  - *In Homogeneous Poisson Process Estimated Intensity is:* ***λ = (n / A)*** *: n(x) = # points in region W with area A*
  - ***Statistical Test for CSR: Quadrant based Chi-Square Test and Spatial Kolmogorov-Smirnov Test***
  - *In Inhomogeneous Poisson Process usual Kernel estimation is used to estimate the intensity function* ***λ(x)***
  - ***Perspective Plot*** *or* ***Contour Plot*** *are used as visual aid to understand intensity function*
  - *Maximum Likelihood Techniques are used to estimate a parametric intensity function in IPP*
  - *Estimated intensity function is used to fit Poisson Model and Residual Analysis takes place*

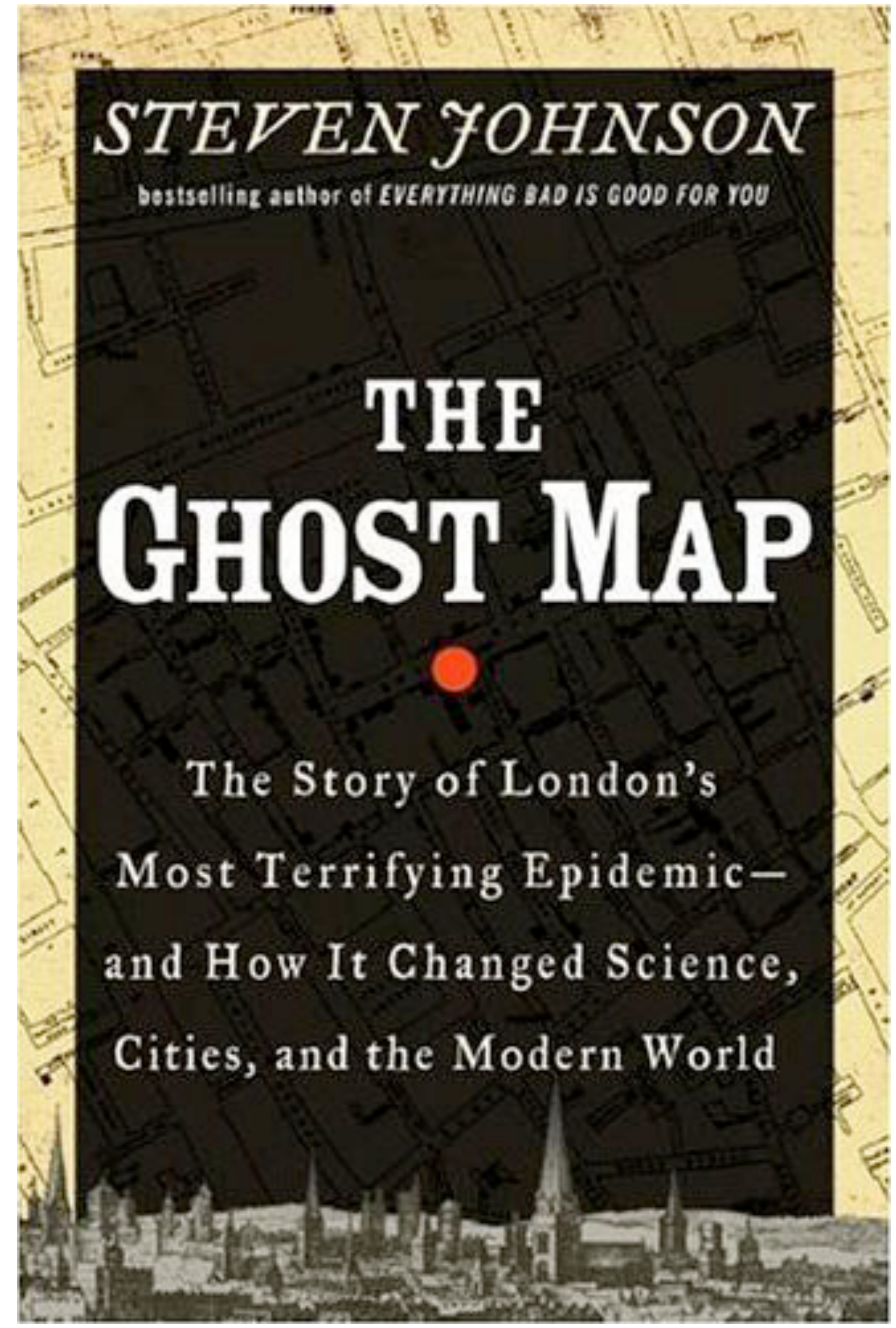




A Classic Illustration

1854 Broad Street Cholera - London

- ***The Story - John Snow Example***
- *Time: August, 1854*
- *Location: Soho District, London, UK*
- *Event: Cholera - Around 600 people died*
- ***Dr. John Snow's Study & Spatial Interpretation***
- *Miasma Theory - Disease such as Cholera/ Black Death were caused by noxious form of "bad air"*
- *Germ Theory - Disease is caused by Germs (micro-organisms)*
- *How Cholera deaths are distributed in Soho? Is there a Complete Spatial Randomness (CSR)?*
- *Snow draws a map to show that cholera deaths are clustered around Broad Street Pump and not Uniformly distributed*
- *Snow's visualization is considered to be the starting point of Modern Epidemiology and Disease Mapping*
- *Spatial Statistical Analysis can formally infer on the spatial distribution of cholera deaths*
- *For more Info: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak*



The Ghost Map

Map Reveals the Truth

Spatial Concentration of Deaths Around Broad Street Pump

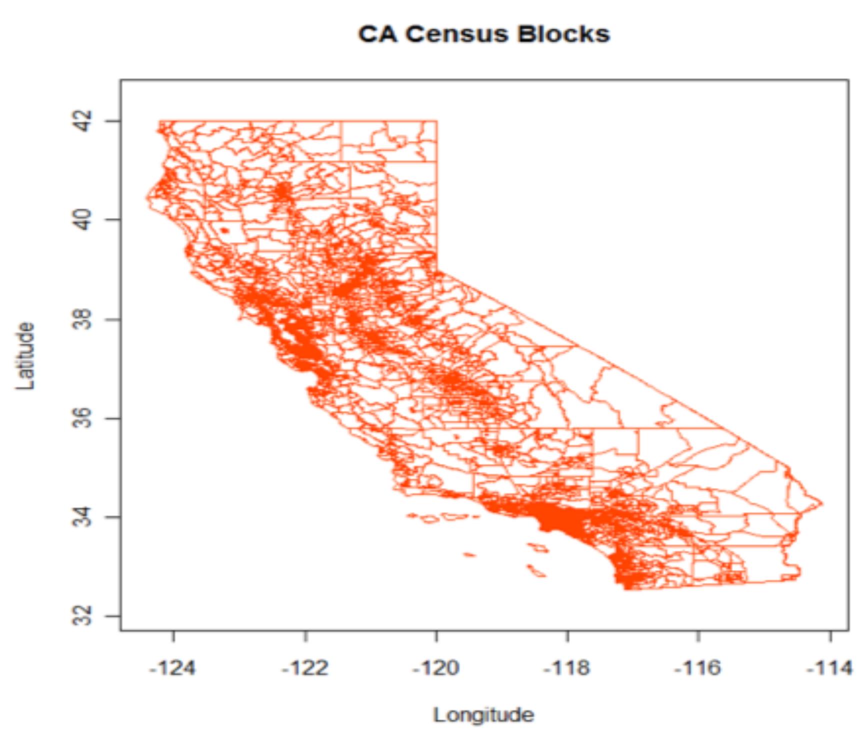
Areal Data

When grouping is a necessity - e.g. census data

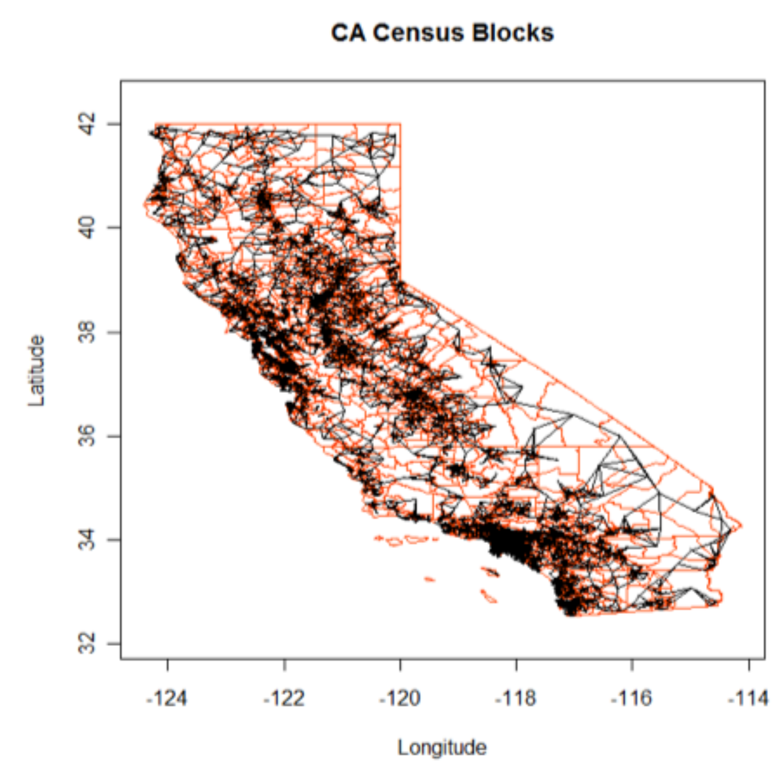
- ***Analysis and inference of Stochastic Process { Y(s) : s** in **D } : D = {s_1, ..., s_n}** is a **partitioned subset** of **R^2***

- ***Common Practical Interest:***
  - ***Spatial Correlation:** Spatial Correlations among territories/ areal units/ sub-regions and incorporating them into the model (mathematical notation : W)*
  - ***Model Based Smoothing:** Even out near-by Territories? How much smoothing should be done?*

- ***Correlation Quantification - Creation of Neighbors and Proximity Matrix W***
  - ***W - Proximity matrix - ((w_{ik})) - gets some value for each pair of locations (i,k)***
  - *Binary Proximity Matrix: **W = ((w_{ik})) = 1** if (i,k) has a common boundary; otherwise 0. Standardized for unit row sum.*
  - *Distance based neighbor criterion can be used (neighbors if within 50 miles of the Territory)*

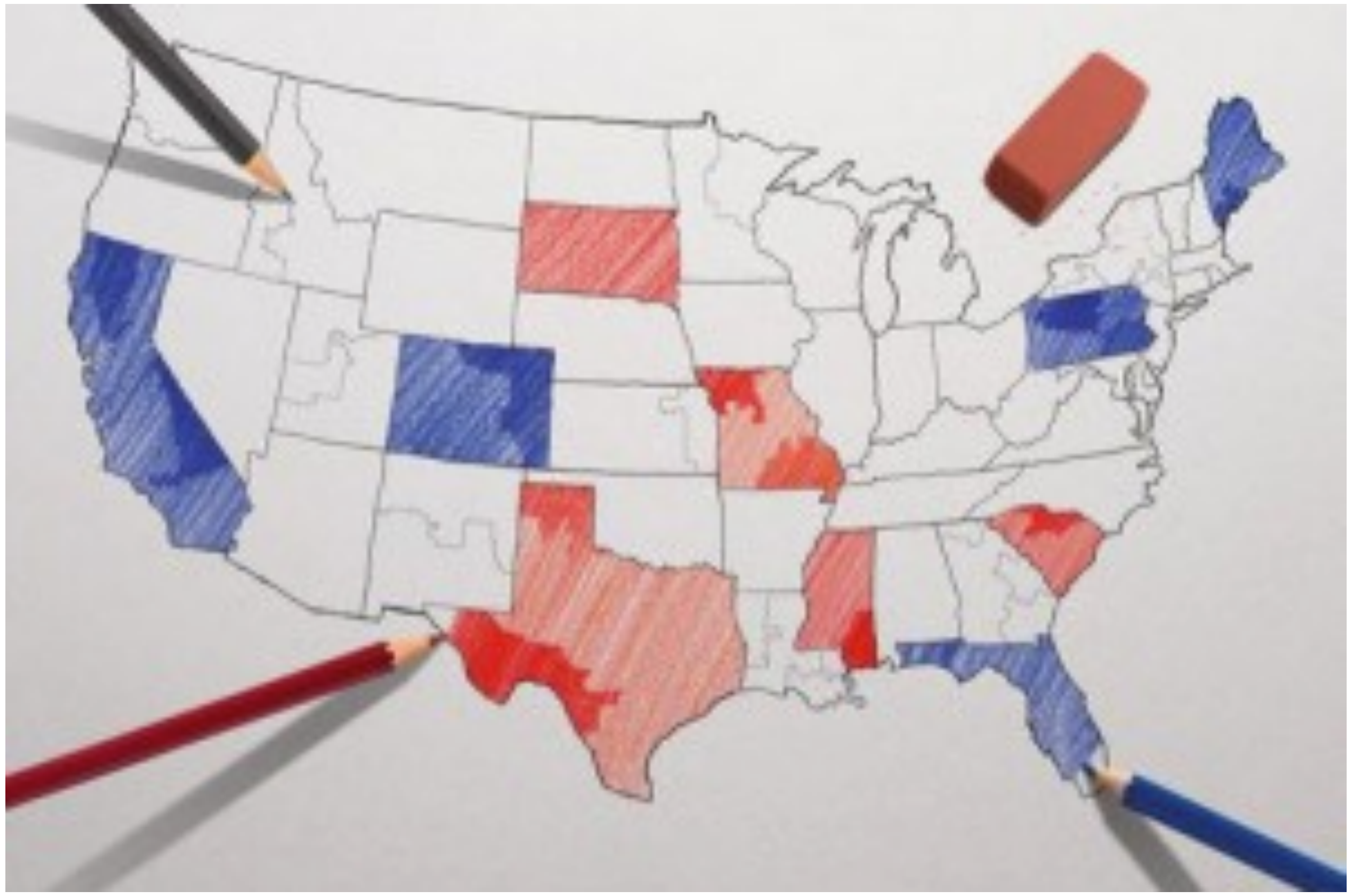


California Proximity Matrix for 4 Nearest Census Tract Neighbors



A Special Topic

Modifiable Areal Unit Problem (MAUP)



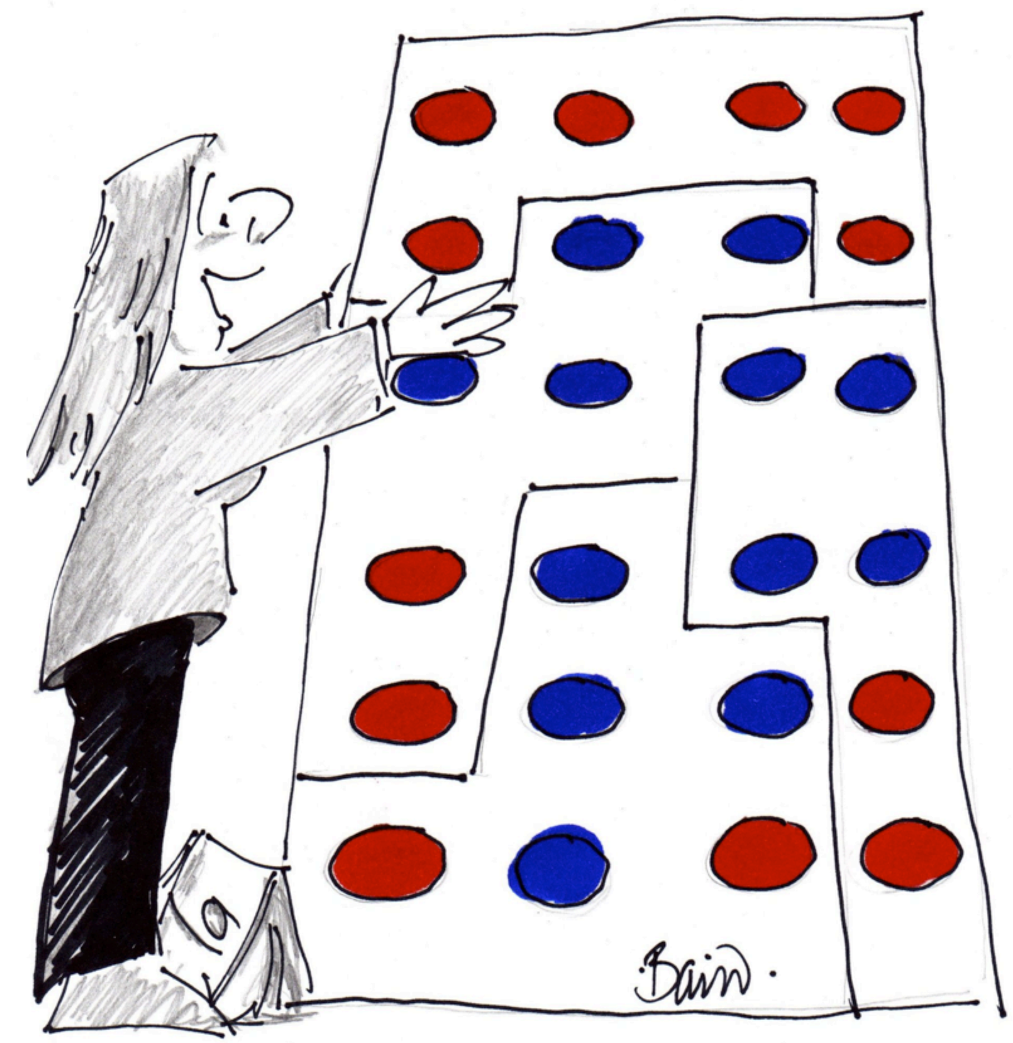
Gerrymandering

Why are people fighting over drawing a line?

Redistricting is process of drawing United States electoral district boundaries, often in response to population changes determined by the results of the decennial census.

Gerrymandering is the practice of drawing district lines to achieve political gain for legislators. The practice of gerrymandering involves the manipulation of district drawing in aims to leave out, or include, specific populations in a legislator's district to ensure his/her reelection.

The above cartoon-map first appeared in the Boston Gazette for March 26, 1812, to mimic that one of the parties forced through the Massachusetts legislature a bill rearranging district lines to assure them an advantage in the upcoming elections. Governor Gerry had only reluctantly signed the law, a Federalist editor is said to have exclaimed upon seeing the new odd looking district lines, "Salamander! Call it a Gerrymander."



source: wikipedia, gerrymandering

Modifiable Areal Unit Problem

Root of the Problem: Flexible Aggregation

- ***Modifiable Areal Unit Problem (MAUP):*** *"...the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating" S Openshaw in his 1984 paper*

Two Elements of MAUP

- ***Scale Effect / Aggregation Effect:*** *Variation in results obtained when data for one set of areal units are aggregated into fewer and larger unit of Analysis*
  - *Example: Using Zip Code or Census Track for an Analysis vs. Using County*
- ***Zoning / Grouping Effect:*** *Any variation in results due to the use of alternative units of analysis when the number of units is held constant*
  - *A positive spatial correlation (homogeneity) in the areal data will reduce the Zoning Effect*
  - *A negative spatial correlation (heterogeneity) in the areal data will increase the Zoning Effect*

- ***Territorial Ratemaking involves two fundamental step:***
  - *Defining Territorial Boundary*
  - *Obtaining Territorial Rate Relativity*
  - *MAUP can potentially affect both the steps based on the existing spatial correlation in the data*

source: MAUP

Choosing the Right Territory

An Illustration - Effect of Different Scaling and Zoning

- ***Actual Data:***
  - *9 Blocks : Base Population & Unemployed count*
  - *Measurement: Unemployment Ratio*

Different Spatial Aggregation Scheme:

(1)

- ***9 zones, unemployment rates are spatially heterogeneous***
- ***expected adverse effect of MAUP, with a change reduction in zone number and use of different zones***

(2) (3) (4)

- ***Use of 3 zones distorted the results from scheme 1. Different territorial or zone definition changed the final statistics/ measure significantly.***

image source: http://www.openmedicine.ca/article/view/286/359

Decision in the Presence of MAUP

How to Find an OptimalTerritorial Definition

- ***Territorial Ratemaking involves two fundamental step:***
  - *Defining Territorial Boundary*
  - *Obtaining Territorial Rate Relativity*
  - *MAUP can potentially affect both the steps based on the existing spatial correlation in the data*

- ***Decision Making in Territorial Boundary Definition:***
  - *How Many Territory/ Zone?*
  - *How to Draw the Line?*

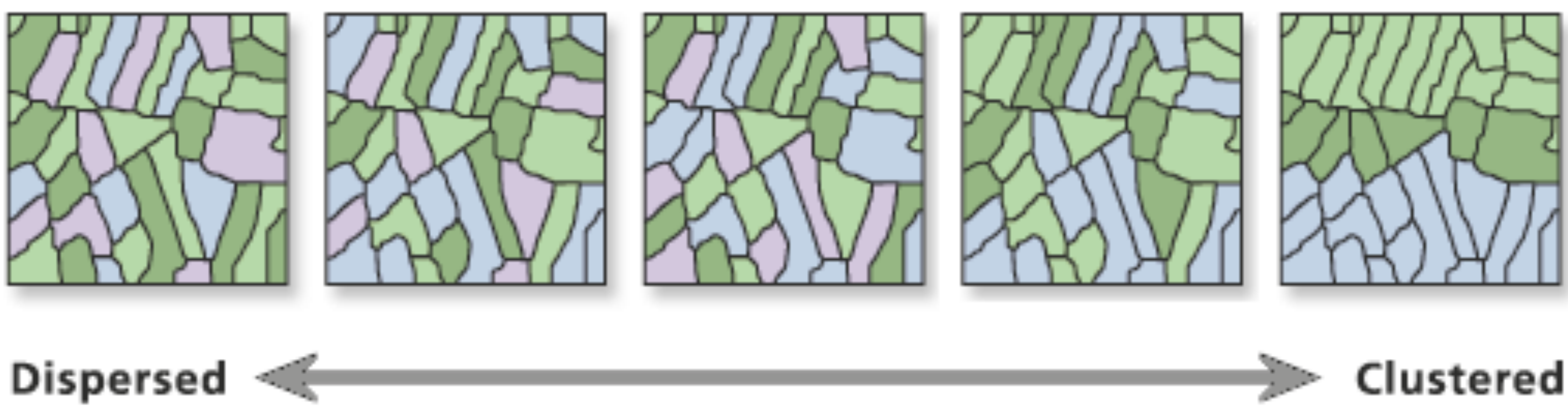
- ***Approach 1: Measuring Spatial Correlation***
  - *Using formal geospatial methods, e.g. Variogram*
  - *Using trial and error methods with different boundary definitions and consequently, calculating within territory and between territory correlations.*
  - *Optimal territorial definition is the one with high within territory correlation and low between territory correlation*

- ***Approach 2: Granular Territorial Definition***
  - *Increasing zone numbers can reduce the impact of both scale effect as well as zone effect*
  - *Using zip code, census track and small area polygonal grid is a easier way to tackle the problem*
  - *In future, we may be able to rate at street level address and won't need to use territories any more*

Moran I and Spatial Correlation

Exploring Different Territorial Boundary Definitions

- ***Spatial Homogeneity Measurement - Moran's I***
  - *Value Ranges from -1 (a perfect Heterogeneity) to +1 (a perfect Homogeneity) with 0 value means perfectly random arrangements*
  - *Moran I can be used to monitor changes in spatial correlations among territories as we change the boundary definitions*
  - *Moran I can be used to determine spatial associations in residual data after a modeling is done*
  - *Most of the GIS and Geospatial Tools have an inbuilt function for Moran I*



source: esri

A Spatial Econometric Model

Spatial Simultaneous Autoregressive Error Model

Spatial Simultaneous Autoregressive (SAR) Error Model For Spatial Process - { Y(s) : s in D } : D = {s_1, ..., s_n}

Y(s) = X(s) β + u(s) : Regression Model

↑

u(s) = λW u(s) + ε(s) : Latent Spatial Lag Model

X(s)β = Regression Covariate Structure (Mean)

u(s) = Spatial Error Structure

ε(s) = Pure Random Error

W = Proximity Matrix

λ = Spatial Lag Coefficient

λ = 0 leads to a purely non-spatial model

β = 0 leads to a purely spatial model

Actual Experience : *Y(s)*

Signal : *X(s)*β

Non-Geographic Predictors
(Example: Age of Insured)

Geographic Predictors
(Example: Avg. Snow Fall)

Geographic Residuals Variation
u(s)

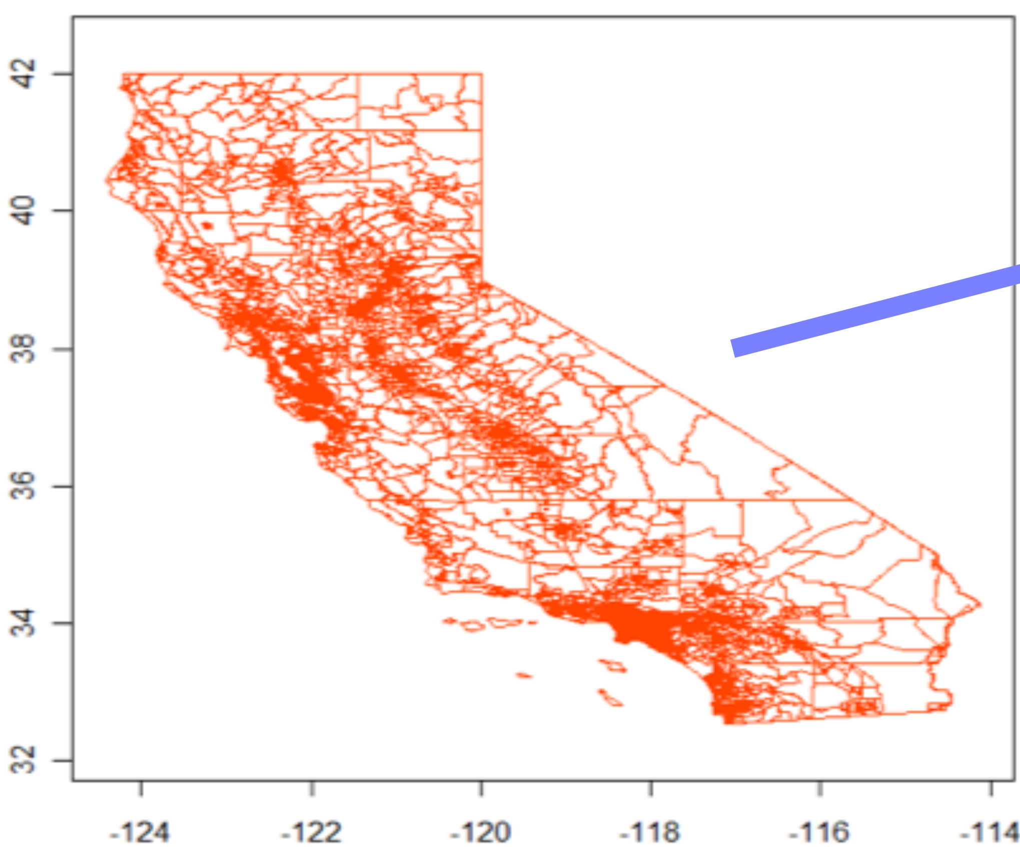
Noise
ε(s)

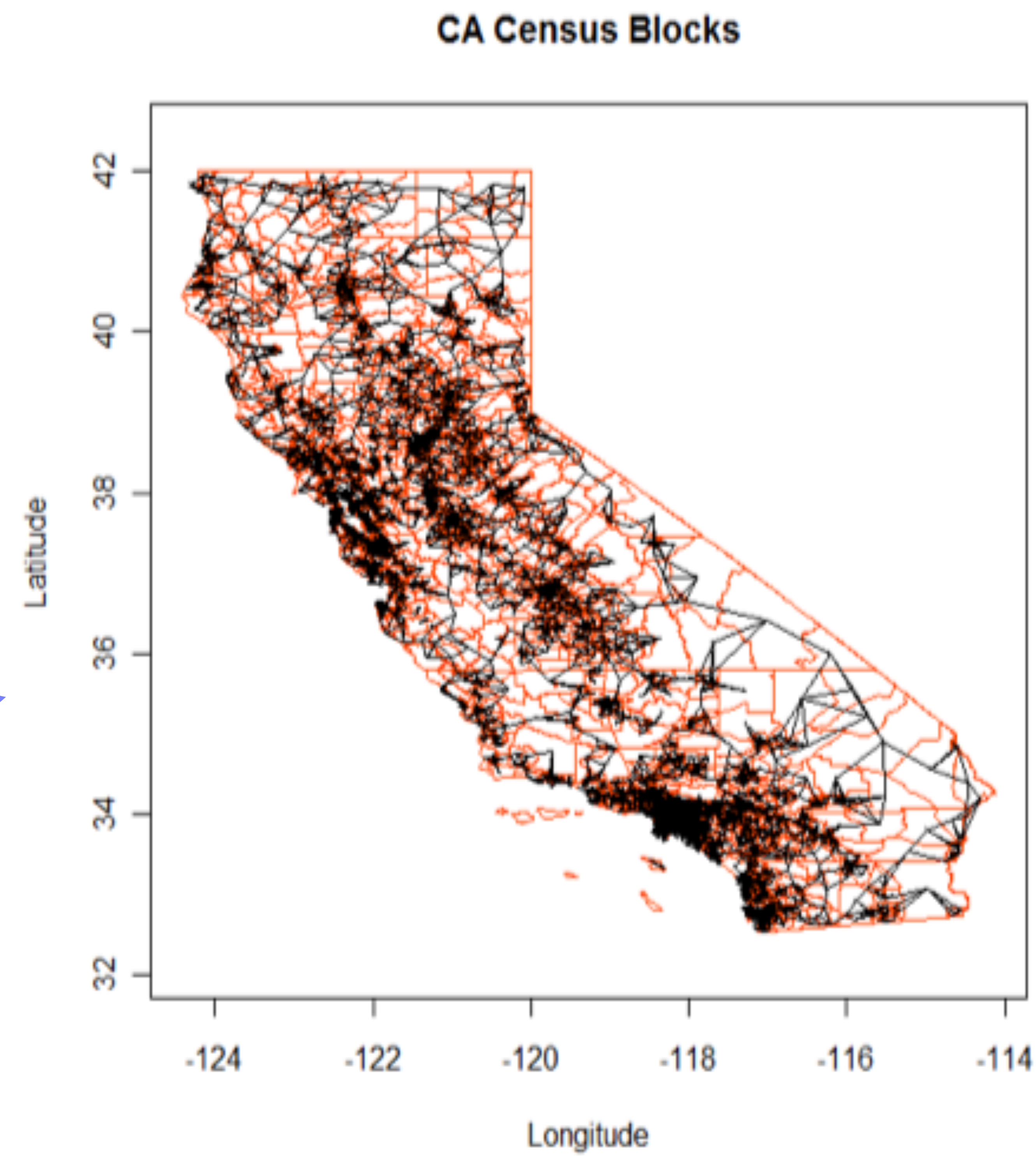
California Housing Price

Simultaneous Autoregressive Models - Neighborhood Creation

Liberty Mutual.

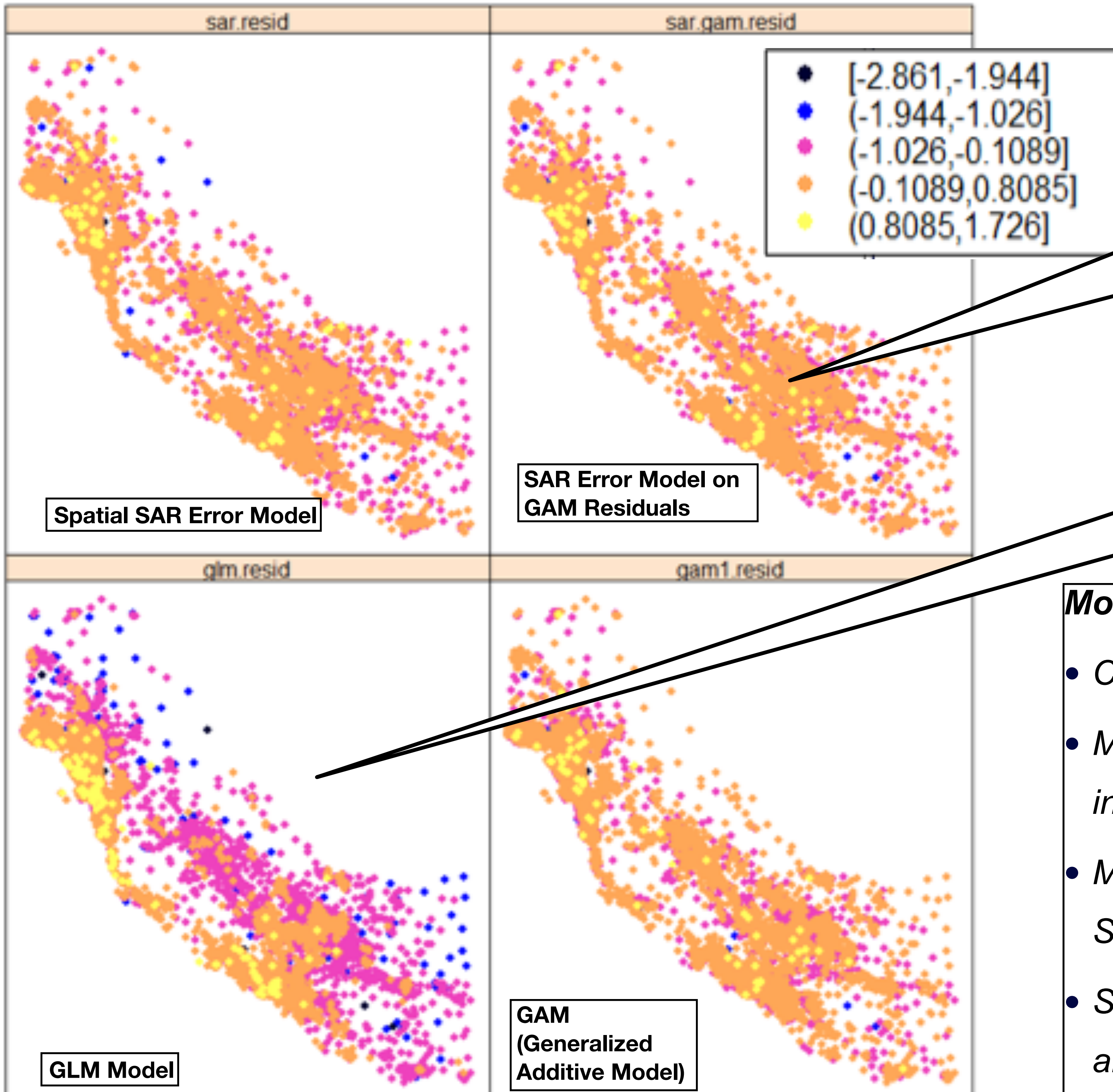
- A 4-Closest Neighbors Contiguity Matrix has been created for California 1990 Census Blocks
- Map (Census Block) data source - US Census
- R program has been used to create this map
- A Common Border Contiguity Matrix May be Tried





Diagnostics - Residual Mapping

Comparison of Different Models on California Housing Data



Lower Dispersion in Residual Distribution and Much Lesser Presence of Spatial Clusters

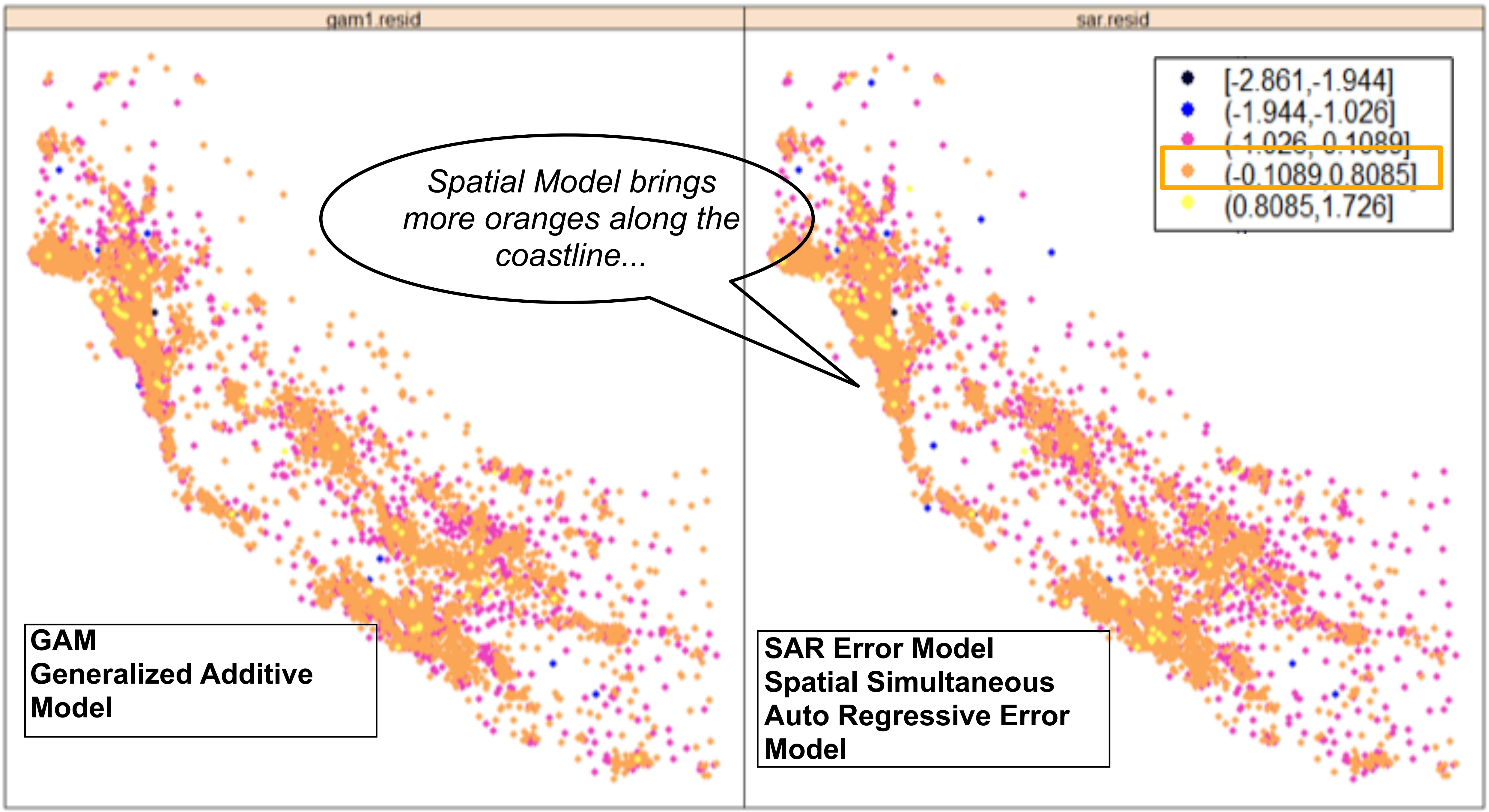
Higher Dispersion in Residual Distribution and Frequent Presence of Spatial Clusters

Model Fitting Steps

- *Creation of a Contiguity/ Proximity Matrix W*
- *Model I - Use the Contiguity Matrix as an Input in Spatial SAR Error Model*
- *Model II - Use the GAM as an offset for the Spatial SAR Error Model*
- *Statistical Software R is used or the entire analysis and graphics: http://cran.r-project.org*

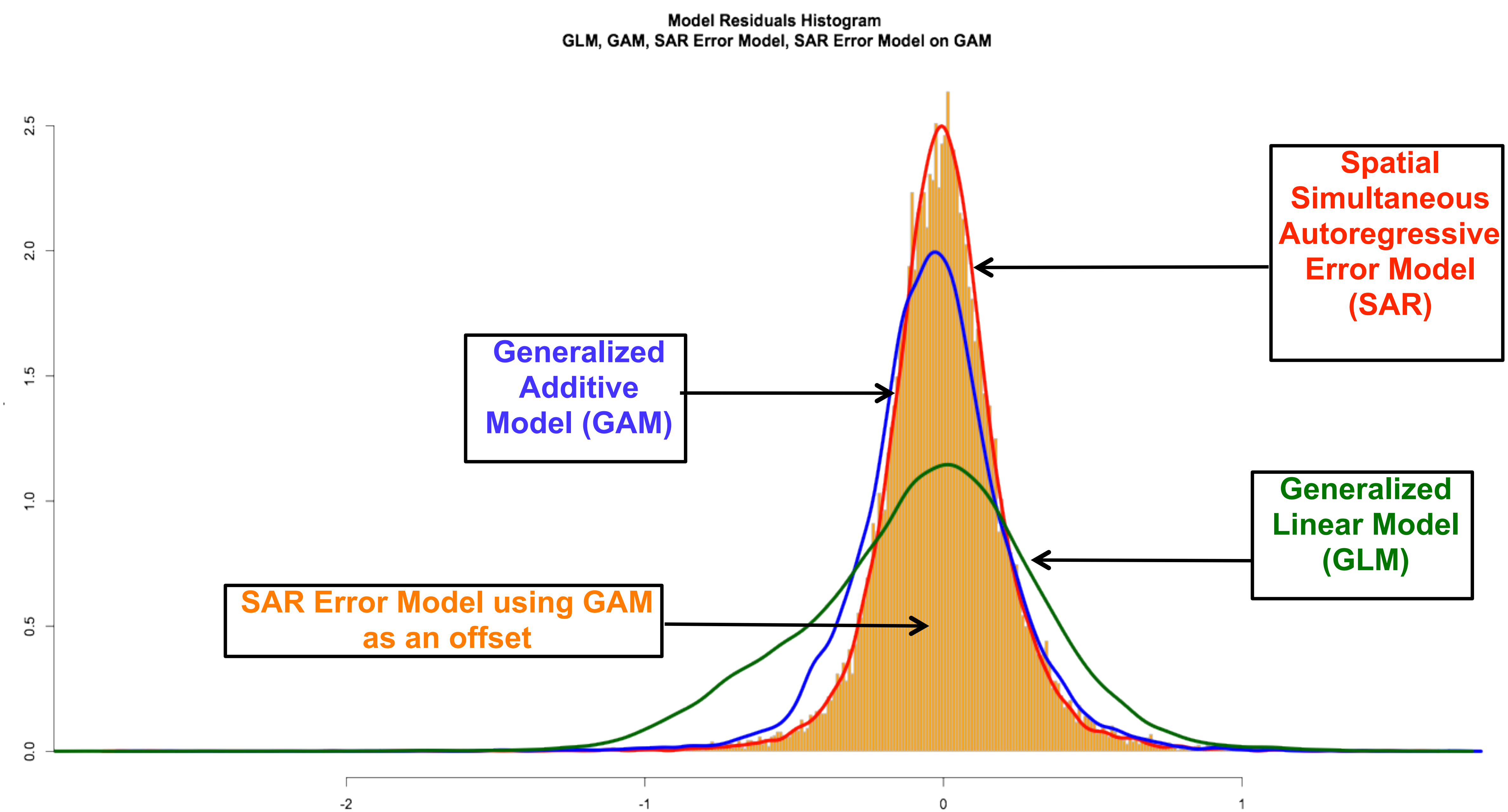
Diagnostics - Residual Mapping

Comparison GAM & SAR



Diagnostics - Residual Histograms

Comparison of GLM, GAM, Spatial SAR Error and GAM with SAR

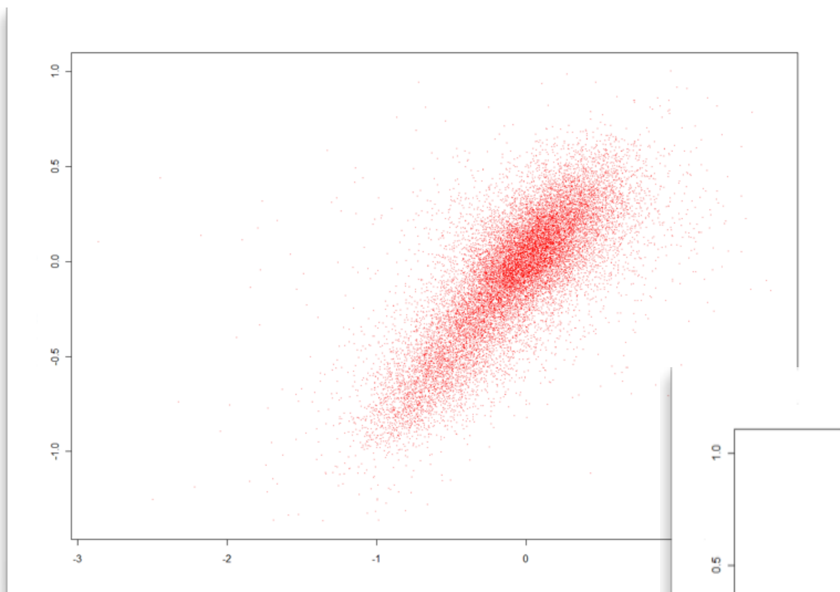


Spatial SAR Error Model shows lower dispersion and magnitude in model residuals distribution compare to GLM & GAM

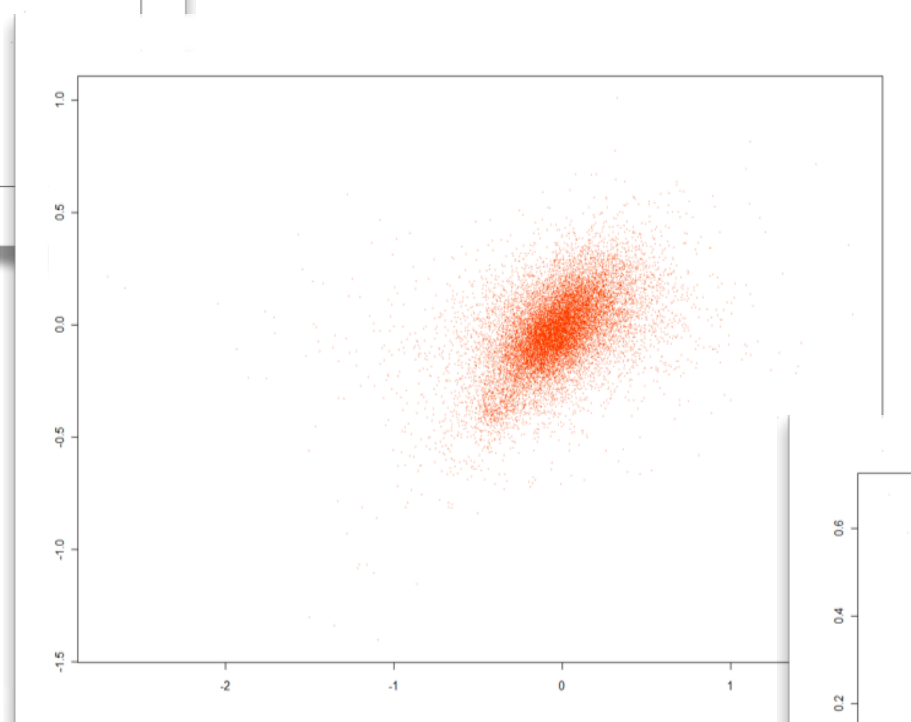
Further Diagnostics - Moran’s I

Correlations between Spatially Lagged Errors - Moran’s I Statistics

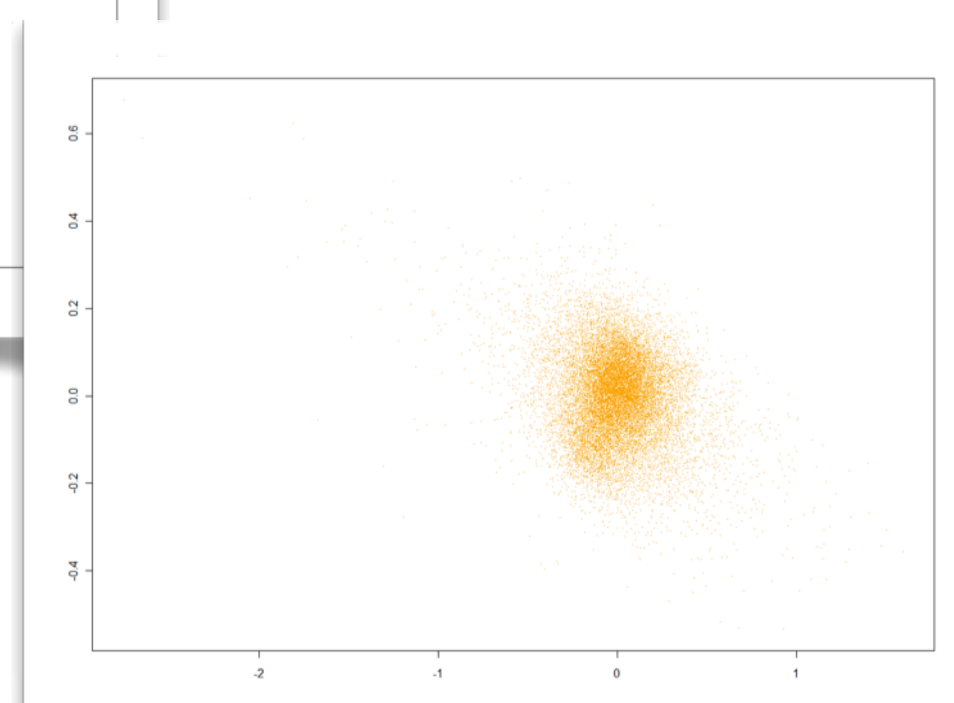
- ***Moran’s I, Measure of Strength Spatial Association among Areal Units***
- ***Time Series Analogous for Measuring Lagged Autocorrelation Coefficient***



Generalized Linear Model



Generalized Additive Model

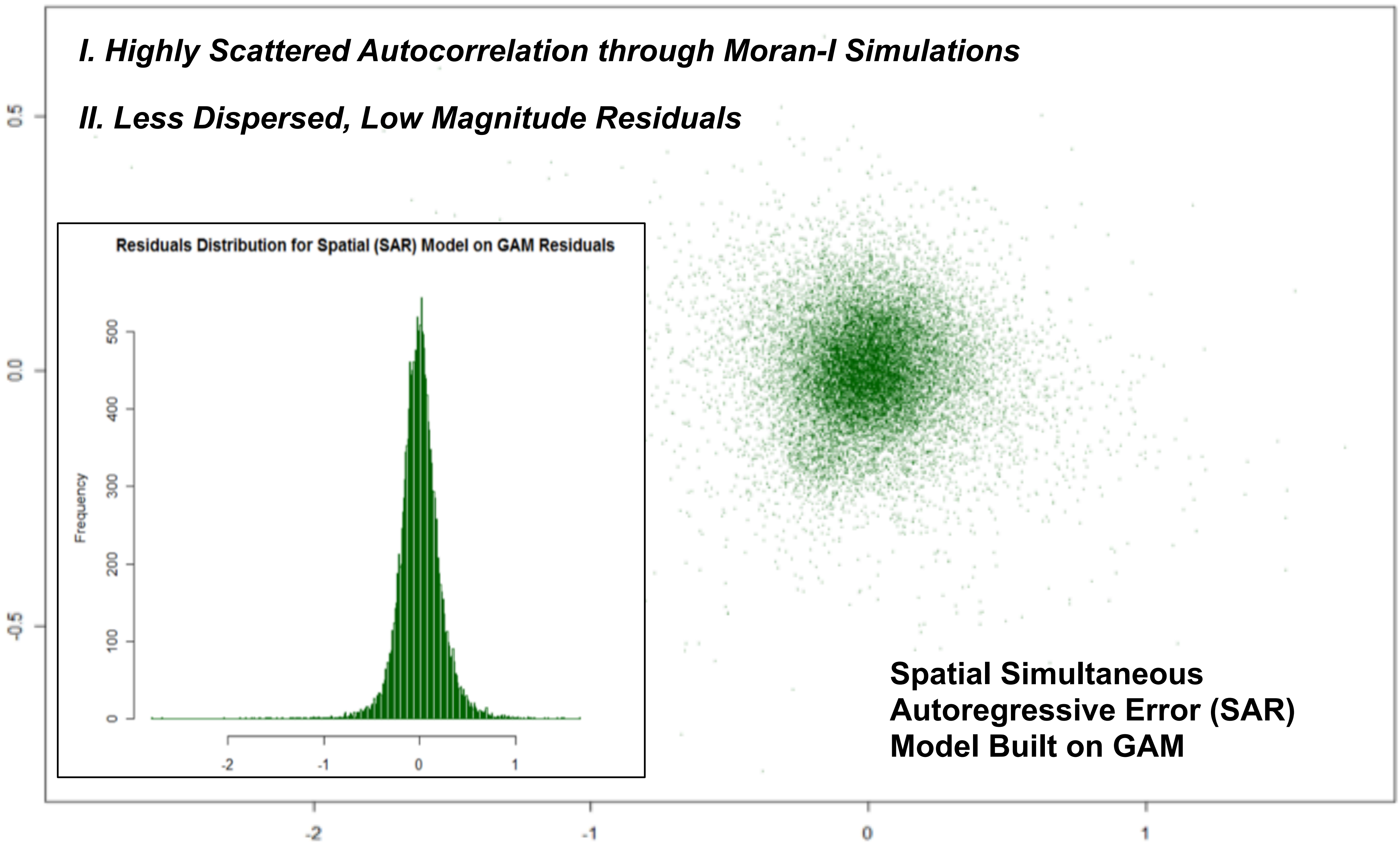


Spatial SAR Model

Filtering Spatial Dependence GLM - Highly Patterned > GAM - Moderately Patterned > Spatial SAR Error Model - Least Patterned

Further Diagnostics - Moran's I

Correlations between Spatially Lagged Errors - Moran's I Statistics



Conclusion

"We're drowning in information and starving for knowledge" - Rutherford Rogers

- ***Spatial Statistics - A Rigorous Statistical Framework For Analyzing Geographically Referenced Data***
  - *Map and Geospatial Explorative Analysis reveals more information from the same data*
  - *Extensive use in the other fields: Epidemiology & Public Health, Demographic Science, Marketing*
  - *Accuracy in results: Understanding MAUP or any other possible issue otherwise ignored*

- ***Computational Scope***
  - *Statistical software R (along with many well developed packages) offers extensive computational facilities and it has a high interaction capability with any standard GIS software*
  - *There are robust Geospatial Analysis Tools in the Market: e.g. ESRI*

- ***Communicating Model Results***
  - *Extensive Visualization Techniques*
  - *Model Results and Diagnostics: Add-on to the GLM based Rating Tool & consistent with GLM*
  - *Availability of interactive GIS software that can show results in a more revealing way*

- ***Text Book References:***

  - *"Hierarchical Modeling and Analysis of Spatial Data", by Banerjee, S., Carlin, B.P. and Gelfand, A.E. (Chapman and Hall/CRC Press, 2004)*
  - *"Applied Spatial Data Analysis with R" by Roger S. Bivand, Edzer J. Pebesma and V. Gómez-Rubio (UseR! Series, Springer 2008)*
  - *"Basic Ratemaking," Werner, G. and Modlin, C., Casualty Actuarial Society (January - 2010)*
  - *"Introduction to Spatial Econometrics" by James LeSage, R Kelley Pace (CRC Press - 2009)*

Question and Comments: *Satadru.Sengupta@LibertyMutual.Com*