Spatial Statistics

A Framework for Analyzing Geographically Referenced Data in Insurance Ratemaking

Satadru Sengupta

Personal Market
Liberty Mutual Group

CAS Ratemaking & Product Management Seminar
Chicago
March 2010

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 30 Minutes

An Introduction to Spatial Statistics For Territorial Ratemaking

---

- ***Motivation***

  Spatial Statistics - An Improvement to the Territorial Ratemaking

  Location Matters - Foundation of Spatial Statistics

  Standard Regression vs. Spatial Regression

- ***Spatial Statistics Theory & Connection to Insurance Ratemaking***

  Stochastic Process, Random Fields and Different Types of Spatial Data

  Spatial Structure in GLM Residuals & Measures of Spatial Dependence

  Why Loss Ratio is So High in North Atlantis?

  Are Theft Claims Coming More From South Atlantis?

  Territorial Boundary Definition - What Territories to be used?

- ***A Case Study - A Spatial Econometric Model***

  Housing Price in California - Simultaneous Autoregressive (SAR) Error Model

  Diagnostics & Model Comparison with GLM & GAM

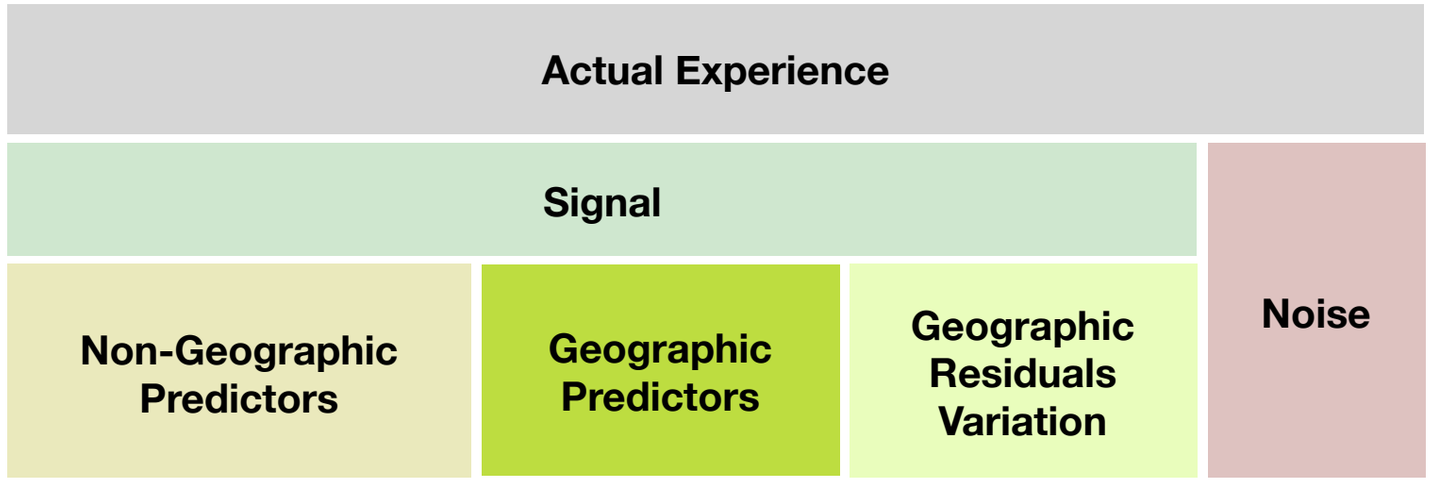
- ***An Evolution - Location in Insurance Ratemaking & Implementation***

  Three Different Assumptions, Three Different Framework and One Common Thread - Filtering

- ***Conclusion***

Territorial Ratemaking

Why We Want to Apply Spatial Statistics Methodologies?



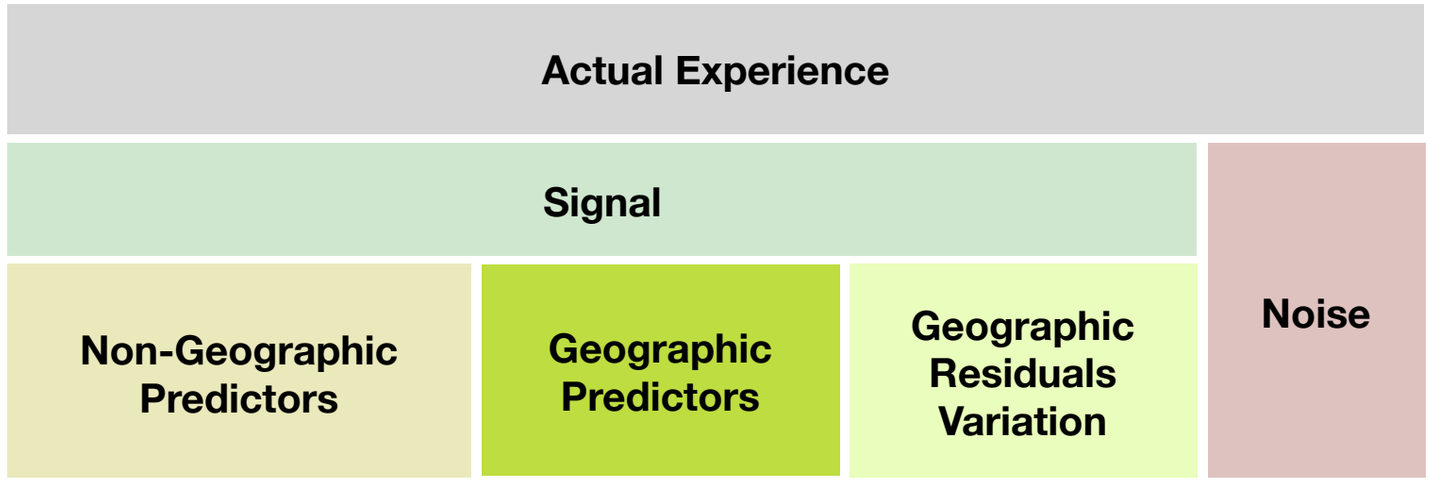
Elements of Territorial Ratemaking

1. ***Territorial Boundary Definition***
2. ***Setting up Territorial Relativities***

- ***Territorial Boundary Definition***
  - *Zip Code, Census Block, County*
  - *Territory acts as a proxy for many different variables that are hard to estimate*
  - ***Administrative Territories may not be optimal for insurance underwriting purpose***
  - *Same Territory may have inhomogeneous insured groups within; Different Territories may have homogeneous insured groups in between*
  - ***Spatial Models can "Filter-Out" this spatial overlap effect***

Territorial Ratemaking

Why We Want to Apply Spatial Statistics Methodologies?



- ***Setting Up Territorial Relativities***
  - ***Non-Geographic Predictors -*** *Age of Insured, Previous Loss History etc.*
  - ***Geographic Predictors -*** *Geo-demographic predictors (population density) as well as on Geo-physical predictors (average snow fall) etc.*
  - ***Geographic Residual Variation -*** *Accounts for possible left out Geographic Predictors*

Including Latitude-Longitude in the Model

- *Latitude-Longitude has a clear effect on Geographic Predictors. Generalized Additive Model (GAM) is the most intuitive way to include Latitude-Longitude in the Model that reduces Geographic Residual Variation.*

Including Spatial Correlation Structure in the Model

- *Practically, it is impossible to eliminate (Geographic) Residual Variation by including "all" possible predictors*
- *Spatial Statistics Methodologies have ability to include a Spatial Error Structure in the Model that accounts for the Geographic Residual Variation*

Motivation

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

- ***Idea*** *- "Everything is* ***related*** *to everything else, but* ***near*** *things are more related than* ***distant*** *things"*
    - **Location Matters -** Observed value at one location is influenced by the observed values at other locations in a geographic area
    - Influence declines with distance
    - Define "***near***" - Euclidean distance, Territory with common boundaries, Transit distance (Manhattan distance), Insured sharing the same fire station, Sphere of influence, other relationships e.g. Actuaries with a degree in Economics, Bostonians commuting in the green line T (subway)
- ***Theory and Computation***
    - Rapid theoretical development of Spatial Statistics in last few decades and widely available literature
    - Improved computation facility and advent of open source programming environment e.g. R, WinBugs
    - Application in the many fields - Epidemiology & Public Health, Political Science, Marketing, Real Estate, Economic Geography, Criminology
- ***Data*** *-* ***Cost effective and accurate geocoding process and easy availability of geocoded data***
    - Photos taken with most standard digital cameras, phones (e.g. iPhone) are geocoded
    - Different sources of Demographic and Geographic Data, Weather Data, Telematics data in coming days, Detailed and highly interactive GIS e.g. Google Earth

Mathematical Interpretation

Data Generating Process - Non-Spatial vs. Spatial

- ***Task - Regression in a Geographic Region -*** *Housing Prices in California, Area with high crime rate in Chicago (Crime Hotspot), Fire/ Water Insurance, Theft Insurance, Pollution Insurance, WC claims across a region*

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

$$Y_i = X_i \beta + e_i$$

$$Y_k = X_k \beta + e_k$$

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

  - ***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*

- ***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*

  - ***Omitted Variable Bias (OVB) -*** *Observations are influenced by a "latent" or "unobservable" factor (e.g. goodness of a good society/ neighborhood can increase demand of houses in that area)*

  - ***Spatial Heterogeneity -*** *Relationship between X and Y changes over Geographic Region (not a constant β)*

Spatial Data & Analogy to Time Series

A Generic Stochastic Process and Three Types of Spatial Data

- ***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,...}}***
- ***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

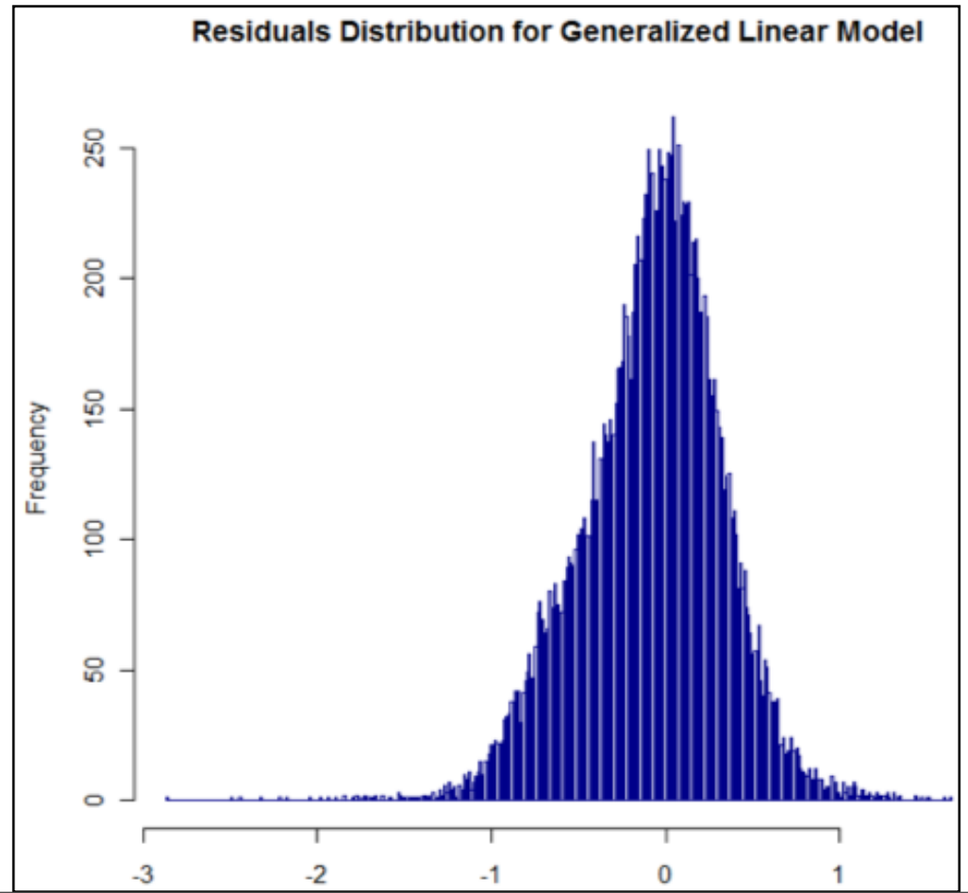
- *How s takes values in D (discrete/ continuous)?*
- *How D comes from R^2 (Fixed/ Random)?*
  - ***Point Referenced Data -*** *When s takes values in D continuously, D is a fixed subset of R^2*
  - *Temperature in Chicago (Possible to collect every point in Chicago)*
  - ***Lattice / Areal Data -*** *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?*

Why Loss Ratio Is So High In North Atlantis?

Point Referenced Data & Geostatistics

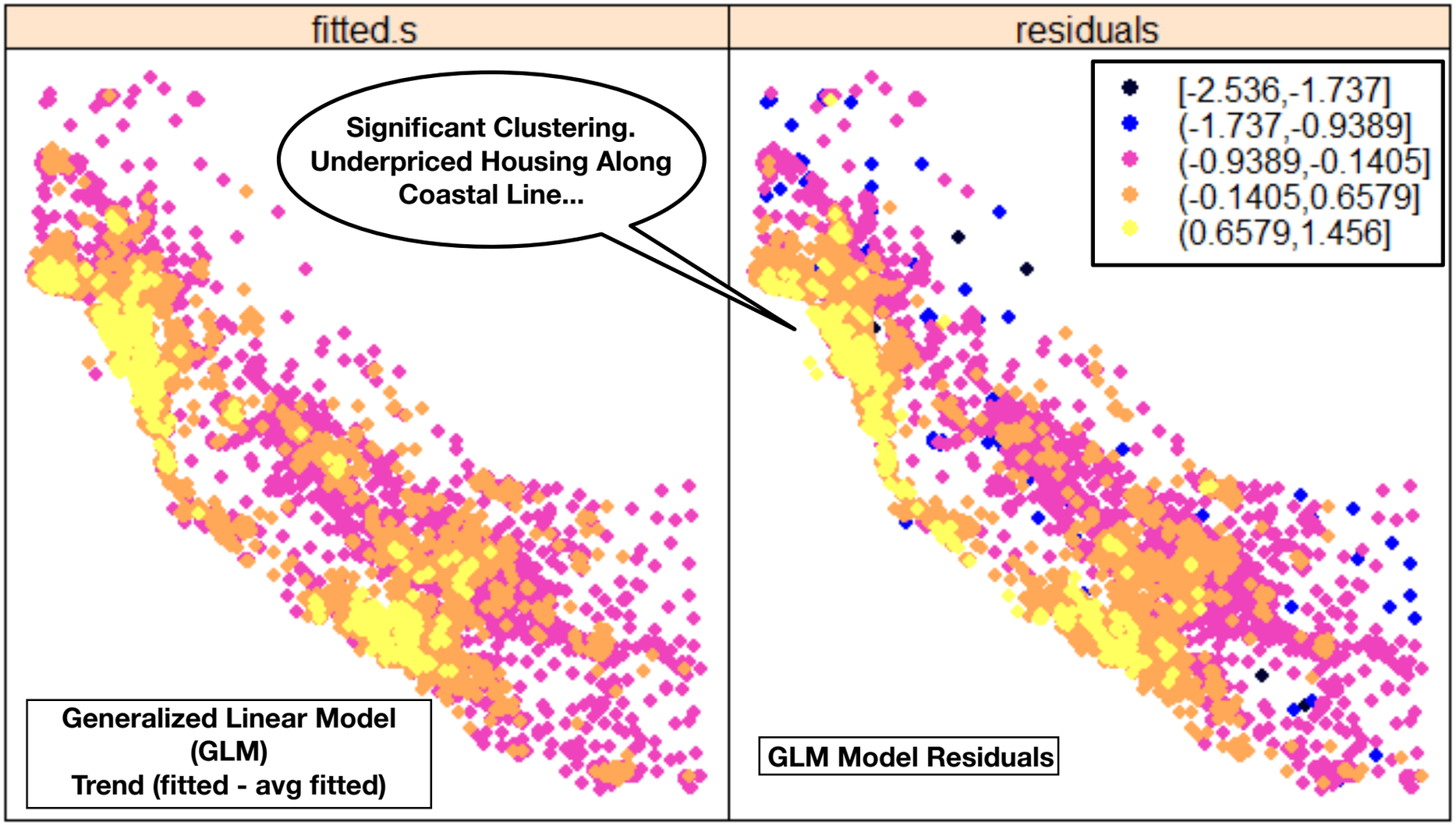
- ***Analysis and inference of Stochastic Process { Y(s) : s** runs **continuously** in **D } : D** is a **fixed subset** of **R^2***
- ***Common Practical Interest in Geostatistics***
  - *Given the observations in different location { $Y(s_1)$,,, $Y(s_n)$} : How to optimally predict Y(s) at a new location 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 & Geostatistics

Independence of Residuals - Spatial Perspective



- *Residuals from the simple model are not distributed randomly over CA*
- *Model underfits along coastline*
- *Model overfits 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?*

Spatial Correlation

Measure of Spatial Correlation

- ***Variogram / Semi-Variogram***
  - *Quantitative measure of Spatial Correlation between two near-by values (observations / errors)*
  - *Mathematical Formulation:*

    $$\textbf{Variogram} = 2\gamma(h) = Var[\,Y(s+h) - Y(s)\,] = 2\,[\,Cov(0) + Cov(h)\,]$$

    Assumes: $E[\,Y(s+h) - Y(s)\,] = 0$; $E[\,Y(s+h) - Y(s)\,]^2$ *depends only on the separation vector h*
  - *Statistical packages provide with the graph between different distances and corresponding* $\gamma(h)$
  - *Empirical graph of* $\gamma(h)$ *or sample variogram is then compared with different theoretical covariance function*
  - $\gamma(h)$ *plays an important role in the geostatistical prediction as the key to spatial correlation*

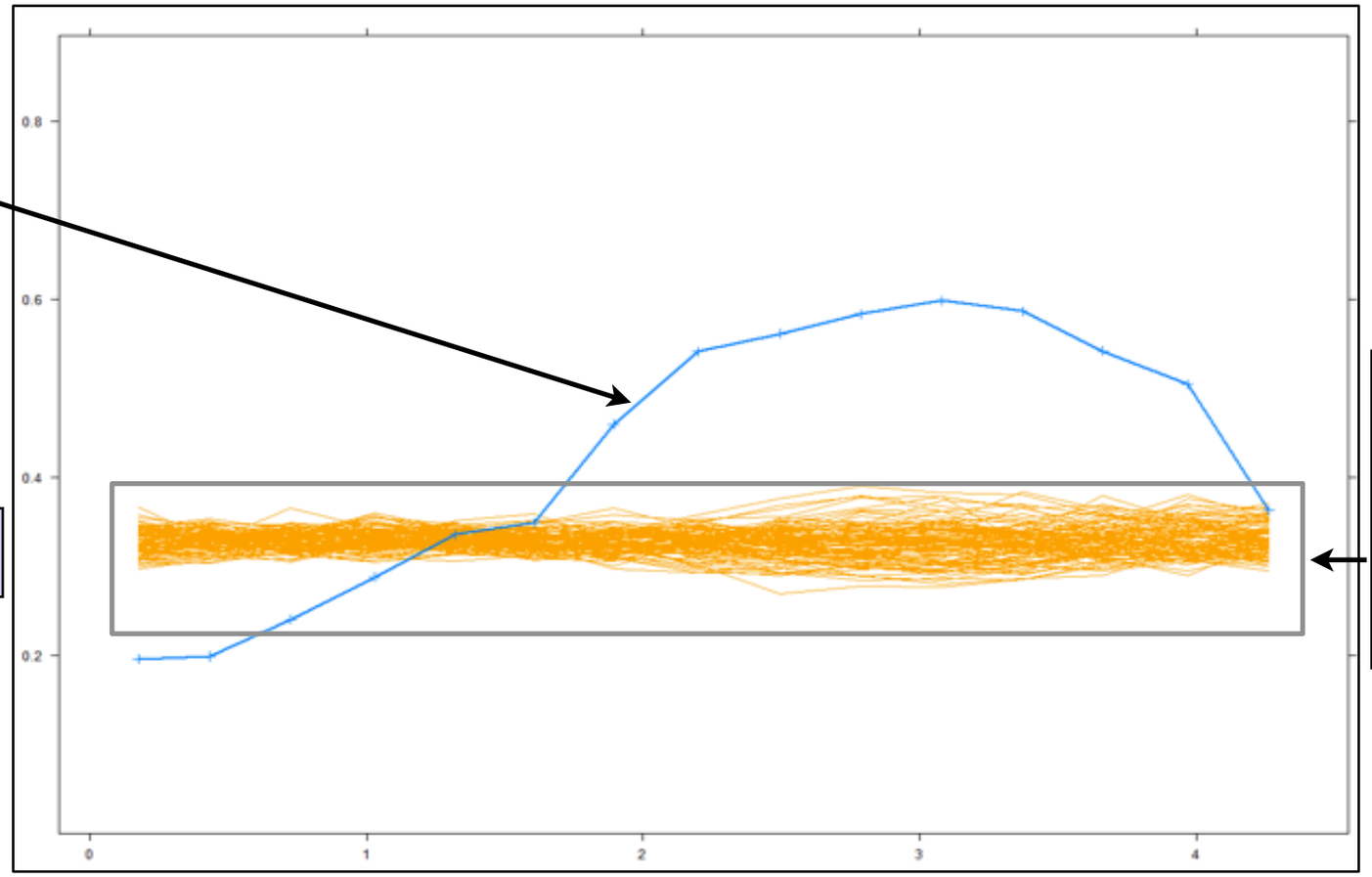
- ***Statistical Testing for Spatial Correlation***
  - ***Spatially Lagged Scatterplot:*** *A simple way to accept or reject spatial correlation is to check the scatter plots of pairs Y(s) and Y(s +h) for all possible separation vector h and grouped by the distances corresponding to h*
  - *In presence of spatial correlation Y(s) and Y(s+h) should show high correlation for lower degree of separation h and low correlation for higher degree of separation*
  - *In Simple words: In the scatter plot, observations in close proximity will show high pattern and observations at distant locations will show randomness.*

Spatial Homogeneity

Sample Variogram and Existence of Spatial Correlation

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

Variogram of the True Data (CA Housing Price) shows Spatial Homogeneity for Close-by Data Points and Heterogeneity for Far-Apart Data Points



Sample Variograms from Location Shuffled Data Sets Are Showing No Changes in Homogeneity with Increasing Distance

Semivariance - a Measure of Heterogeneity

Distance

Close-by Observations

Far-apart Observations

- ***Sample Variogram & Estimation of Spatial Correlation***
  - *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*

Are Theft Claims Coming More From South Atlantis?

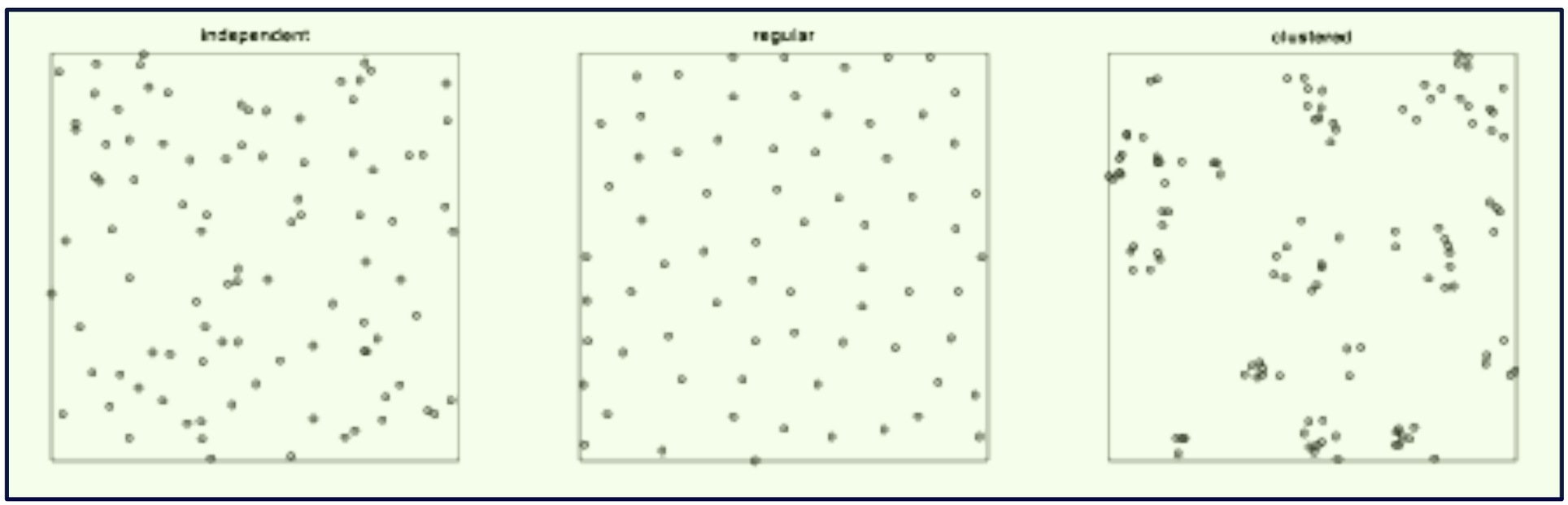
Spatial Point Pattern Process - Spatial Poisson Process

- ***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 Inhibitation - Competition between points over the region*

- ***Spatial Poisson Process - Events occur independently and distributed according to a given intensity function $\lambda(.)$ 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* **$\lambda(x)$**

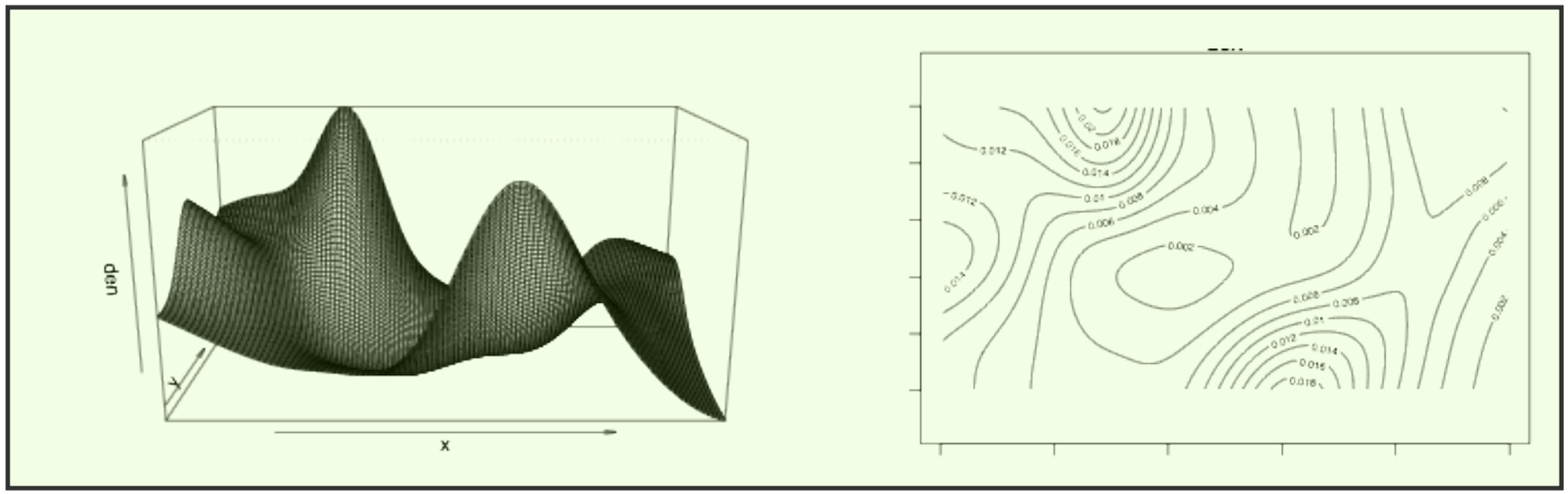


Spatial Point Pattern Process

Distribution of Events

- ***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

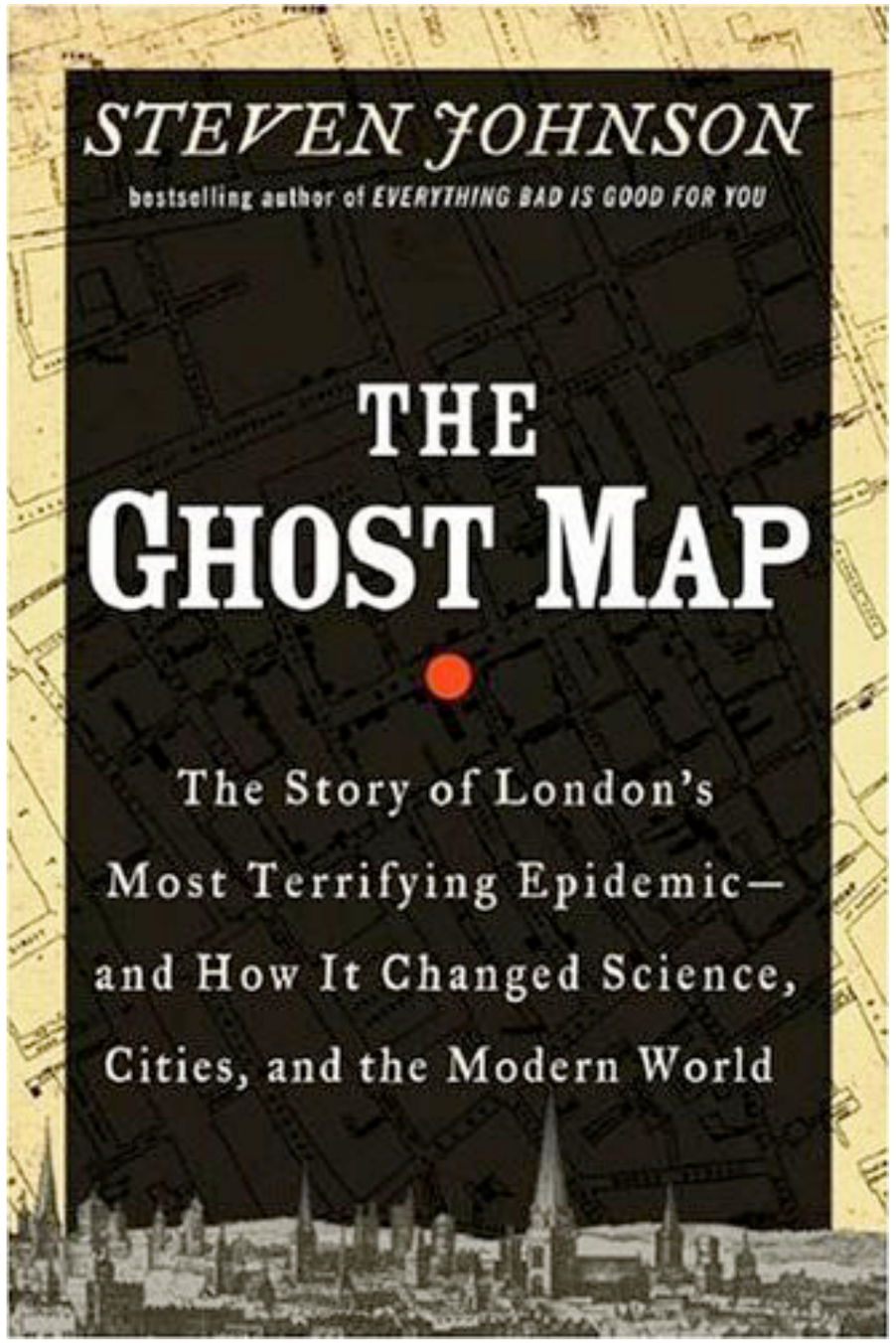
The Story - John Snow Example

- *Time: August, 1854*
- *Location: Soho District, London, UK*
- *Event: Cholera - Around 600 people died*

Dr. John Snow's Study & Spatial Interpretation

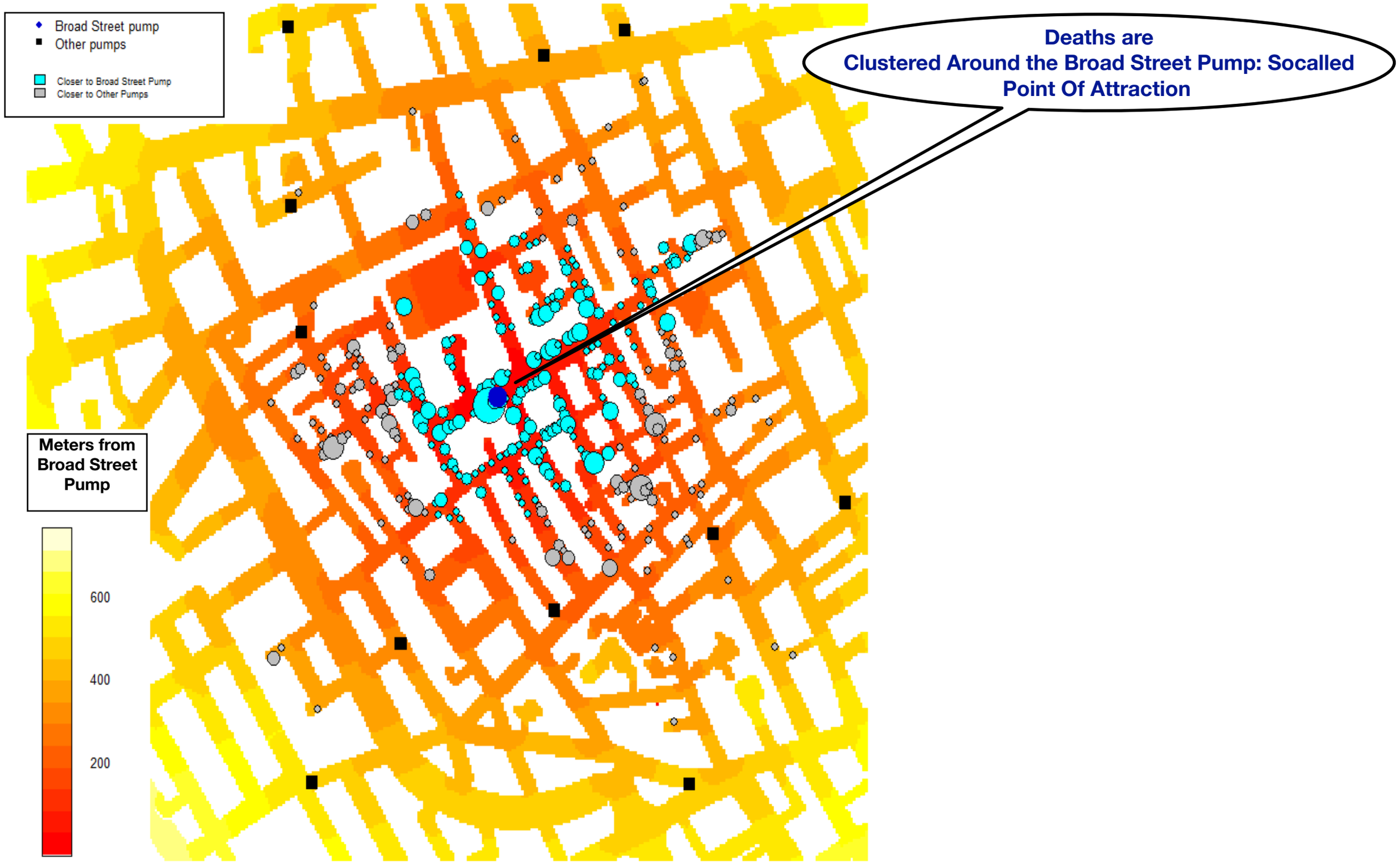
- *Miasma Theory - Disease such as Cholera/ Black Death were caused by noxious form of "bad air"*
- *Germ Theory - Disease is caused by Germs (micro-organisms)*
- *How Cholera deaths are distributed in Soho? Is there a Complete Spatial Randomness (CSR)?*
- *Snow draws a map to show that cholera deaths 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

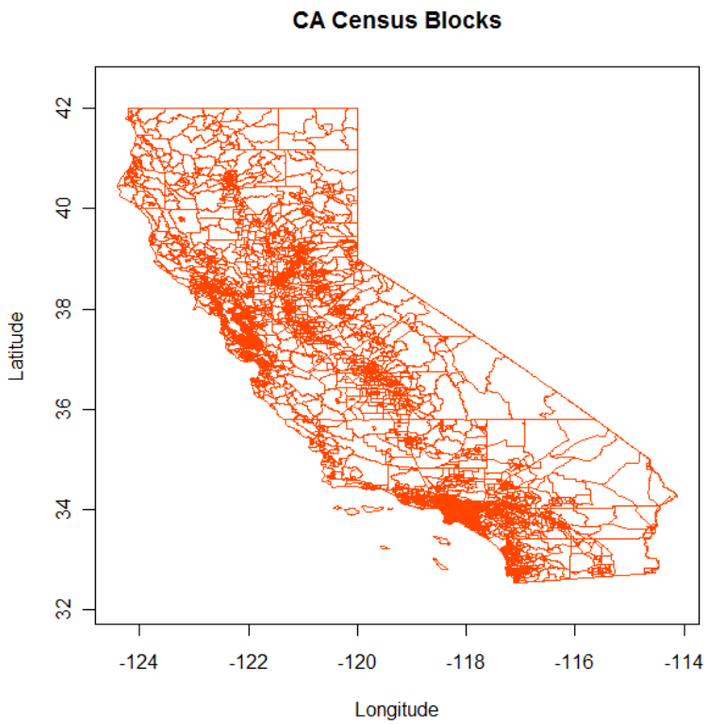
Spatial Concentration of Deaths Around Broad Street Pump



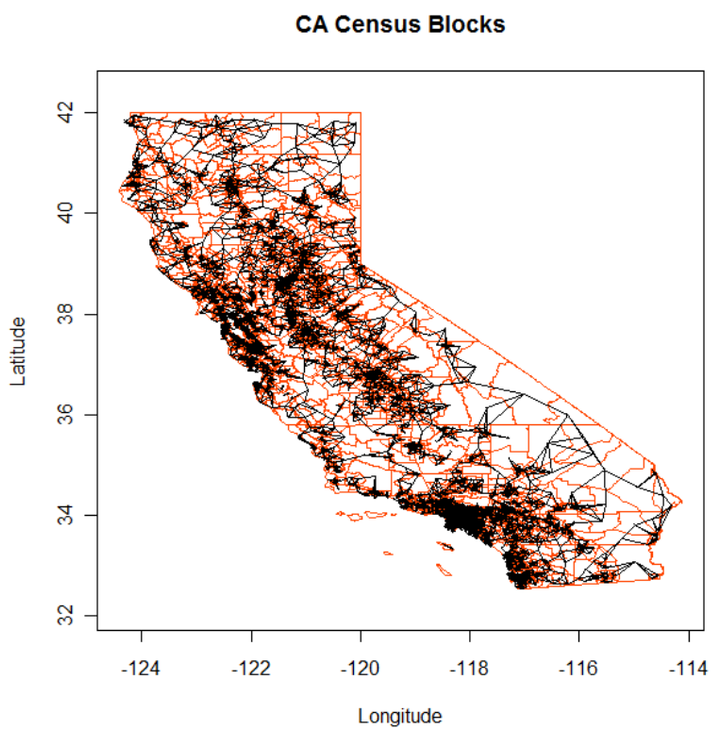
What Territories Should Be Used?

Lattice/ Areal 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*
  - ***Model Based Smoothing:** Even out near-by Territories? How much smoothing should be done?*
  - ***Modifiable Areal Unit Problem (MAUP) -** How to re-allocate observations when there is a change in territorial definition (a new set of territory to be used)?*
- ***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 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}***

$$Y(s) = X(s)\,\beta + u(s) : \text{Regression Model}$$

$$u(s) = \lambda W\,u(s) + \varepsilon(s) : \text{Latent Spatial Lag Model}$$

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

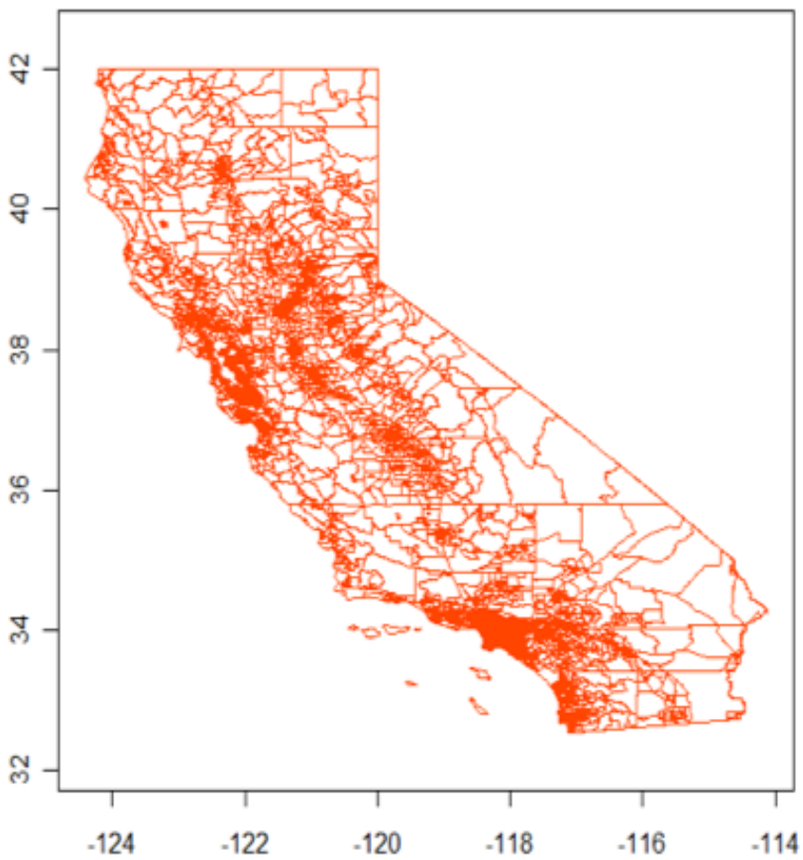
β = 0 leads to a purely spatial model

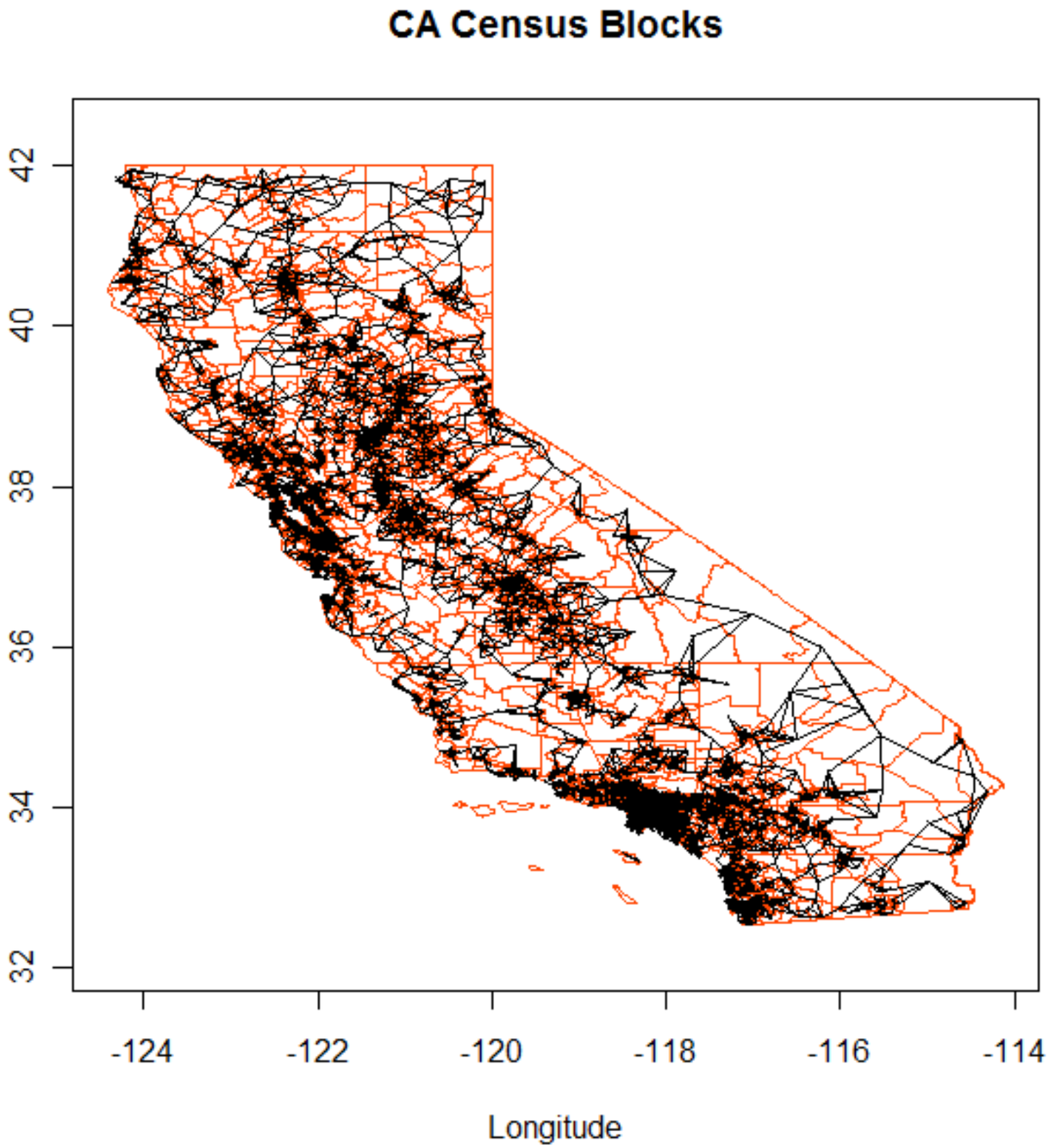
Actual Experience : *Y(s)*			
Signal : *X(s)β*			Noise *ε(s)*
Non-Geographic Predictors (Example: Age of Insured)	**Geographic Predictors** (Example: Avg. Snow Fall)	**Geographic Residuals Variation** ***u(s)***	

California Housing Price

Simultaneous Autoregressive Models - Neighborhood Creation

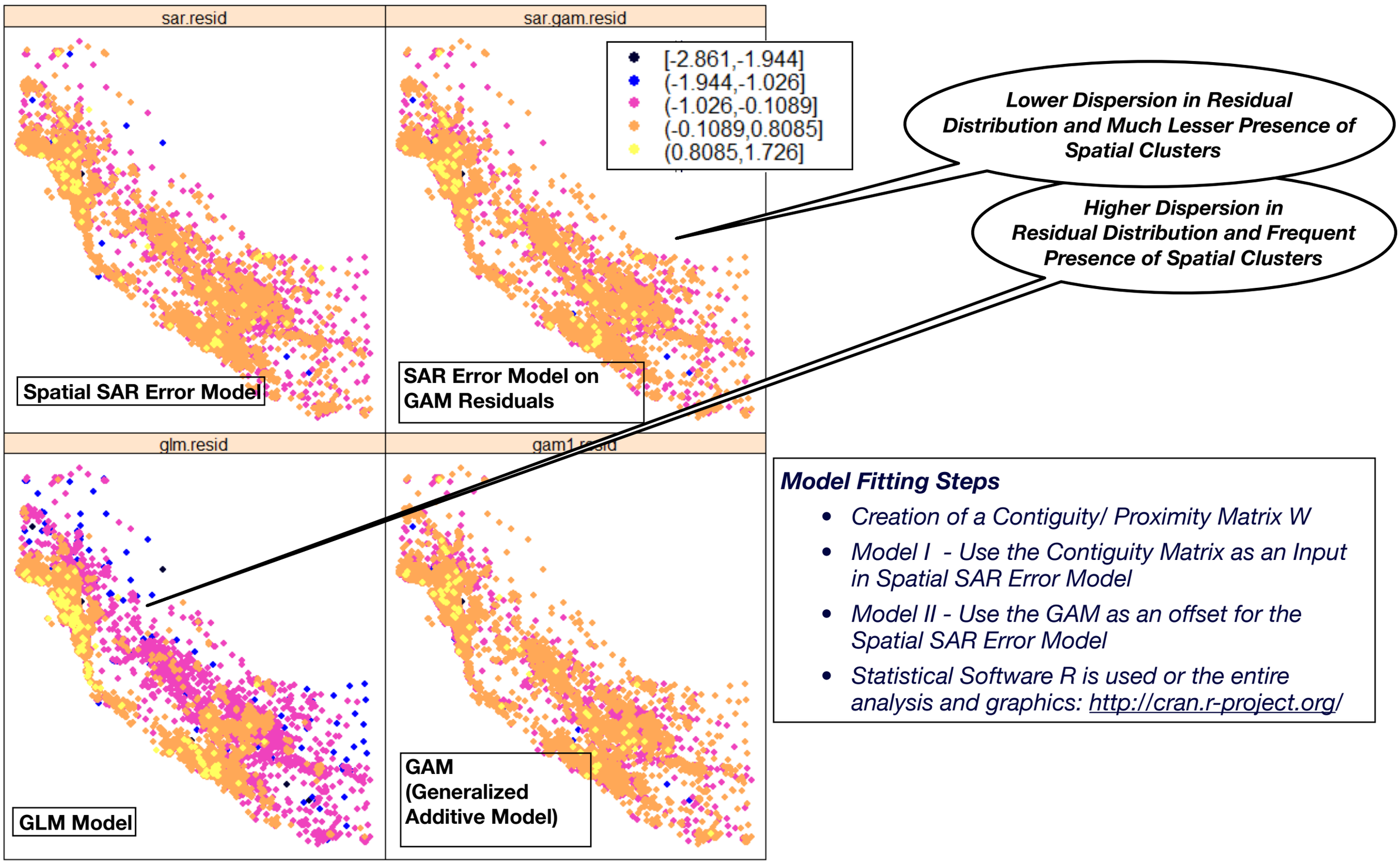
- **A 4-Closest Neighbors Contiguity Matrix has been created for California 1990 Census Blocks**
- **Map (Census Block) data source - US Census**
- **R program has been used to create this map**
- **A Common Border Contiguity Matrix May be Tried**





Diagnostics - Residual Mapping

Comparison of Different Models on California Housing Data

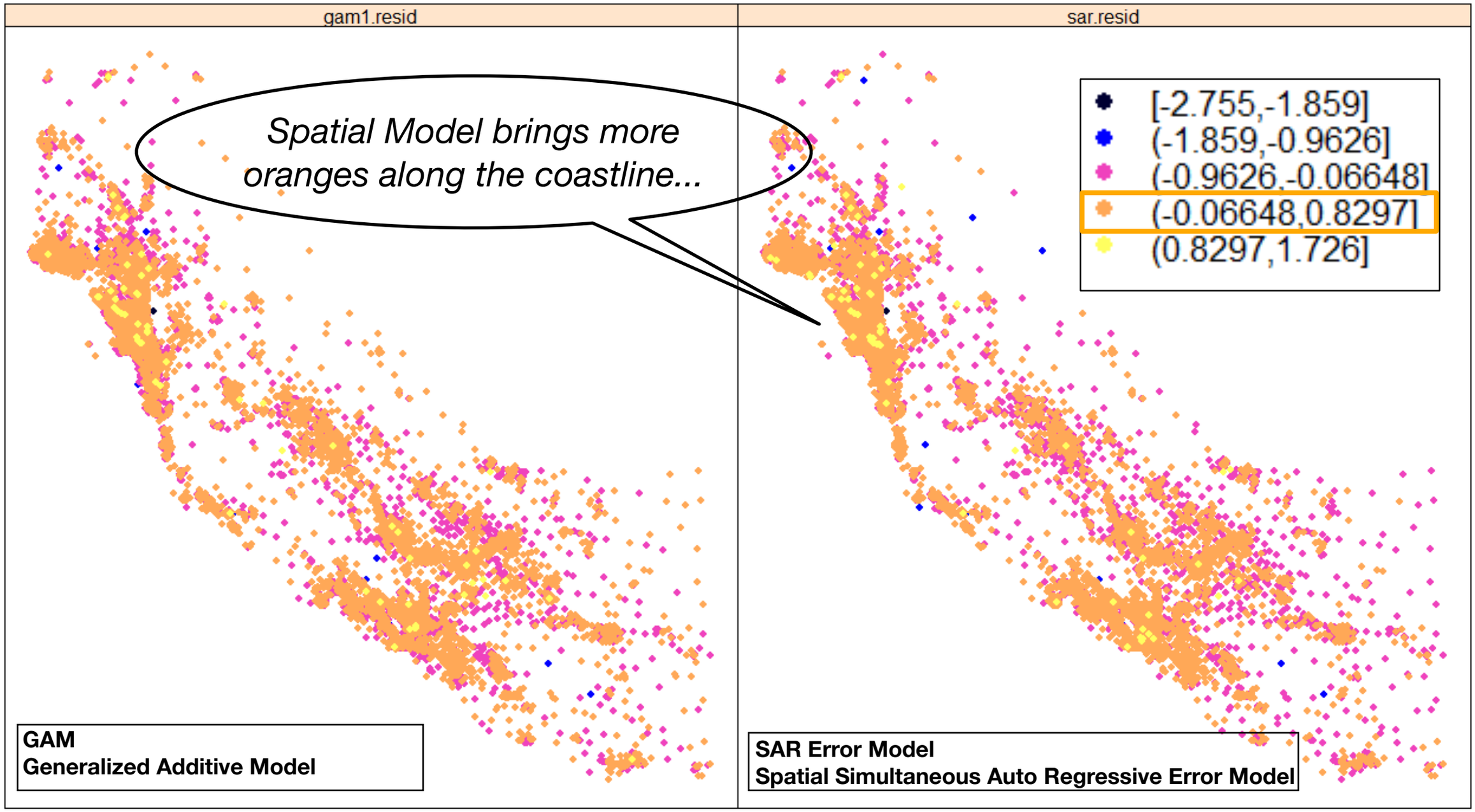


Model Fitting Steps

- *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/*

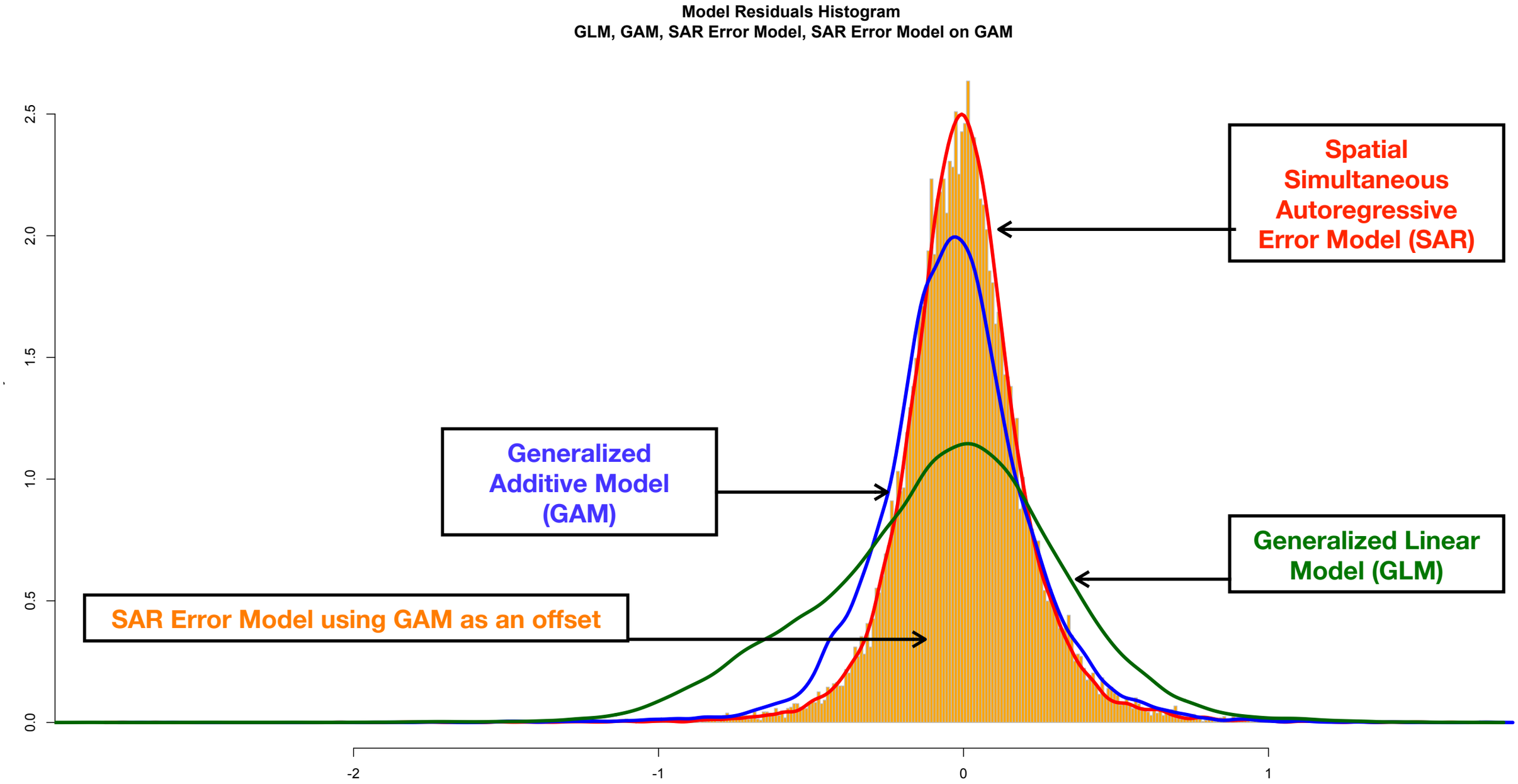
Diagnostics - Residual Mapping

Comparison GAM & SAR



Diagnostics - Residual Histograms

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



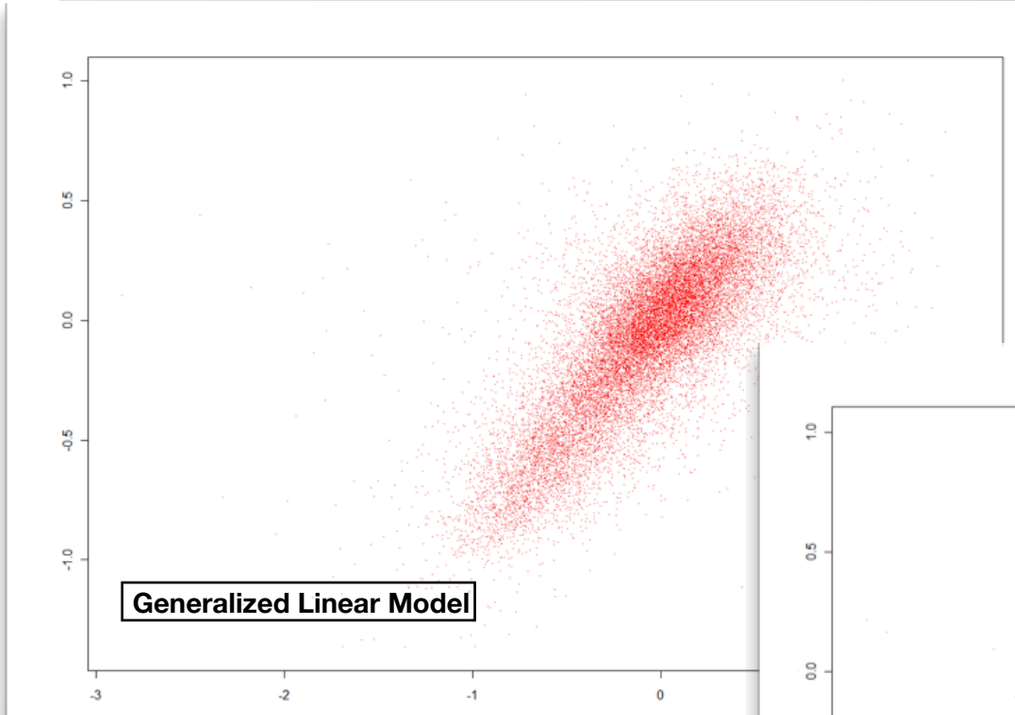
Spatial SAR Error Model shows lower dispersion and magnitude in model residuals distribution compare to GLM & GAM

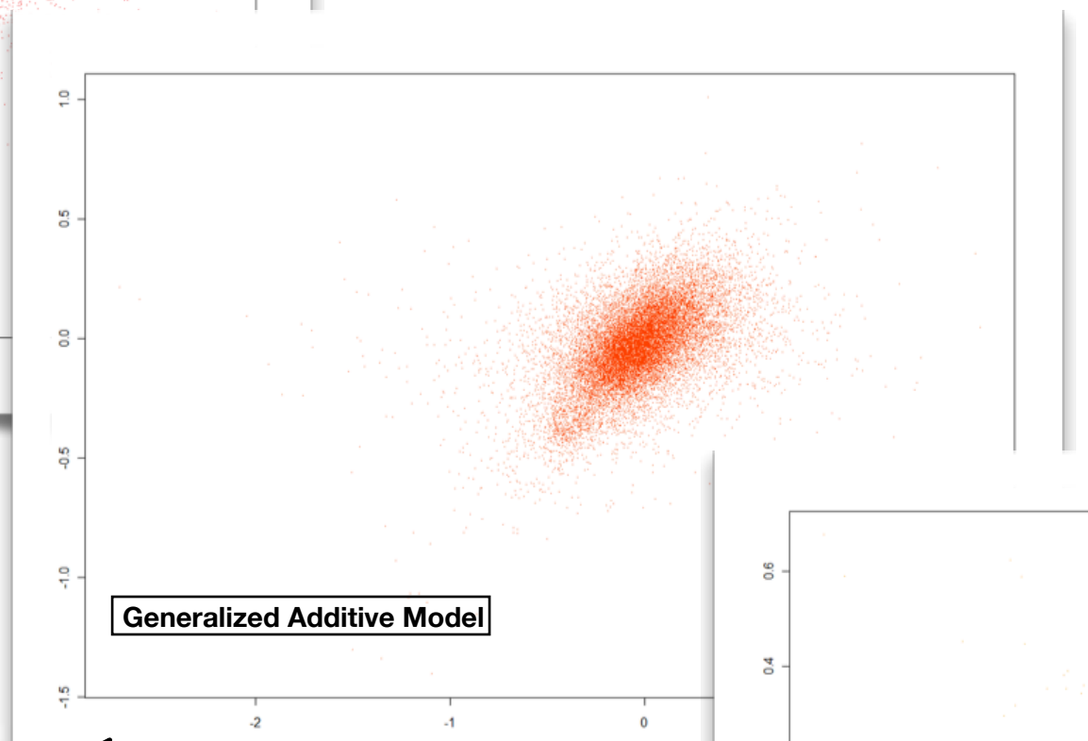
Further Diagnostics

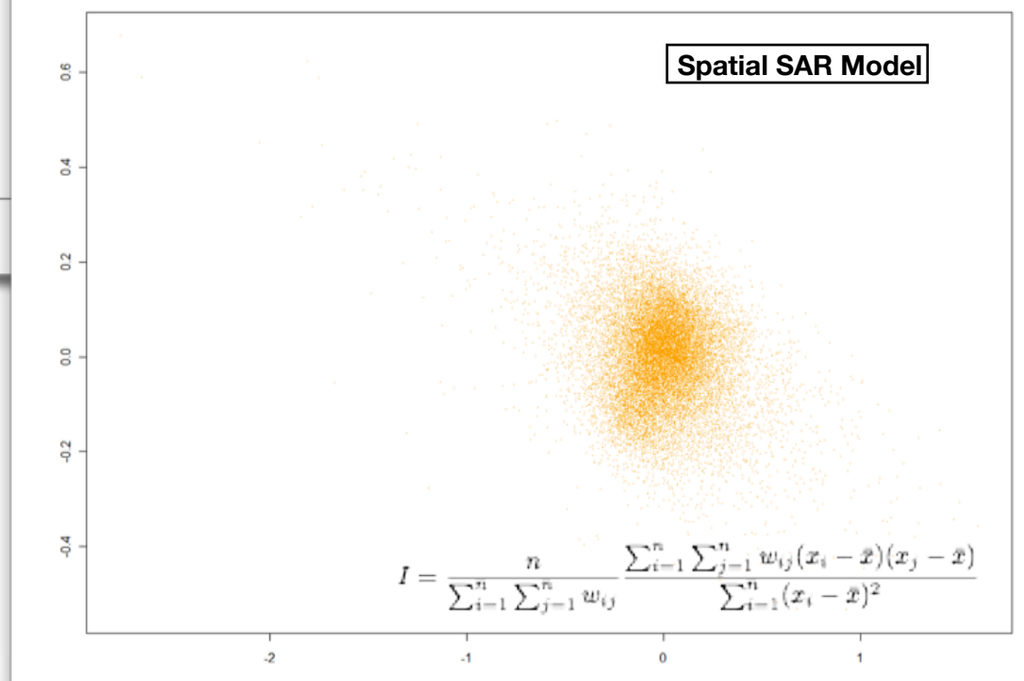
Correlations between Spatially Lagged Errors - Moran's I Statistics

I. Moran I, Measure of Strength Spatial Association among Areal Units

II. Time Series Analogous for Measuring Lagged Autocorrelation Coefficient







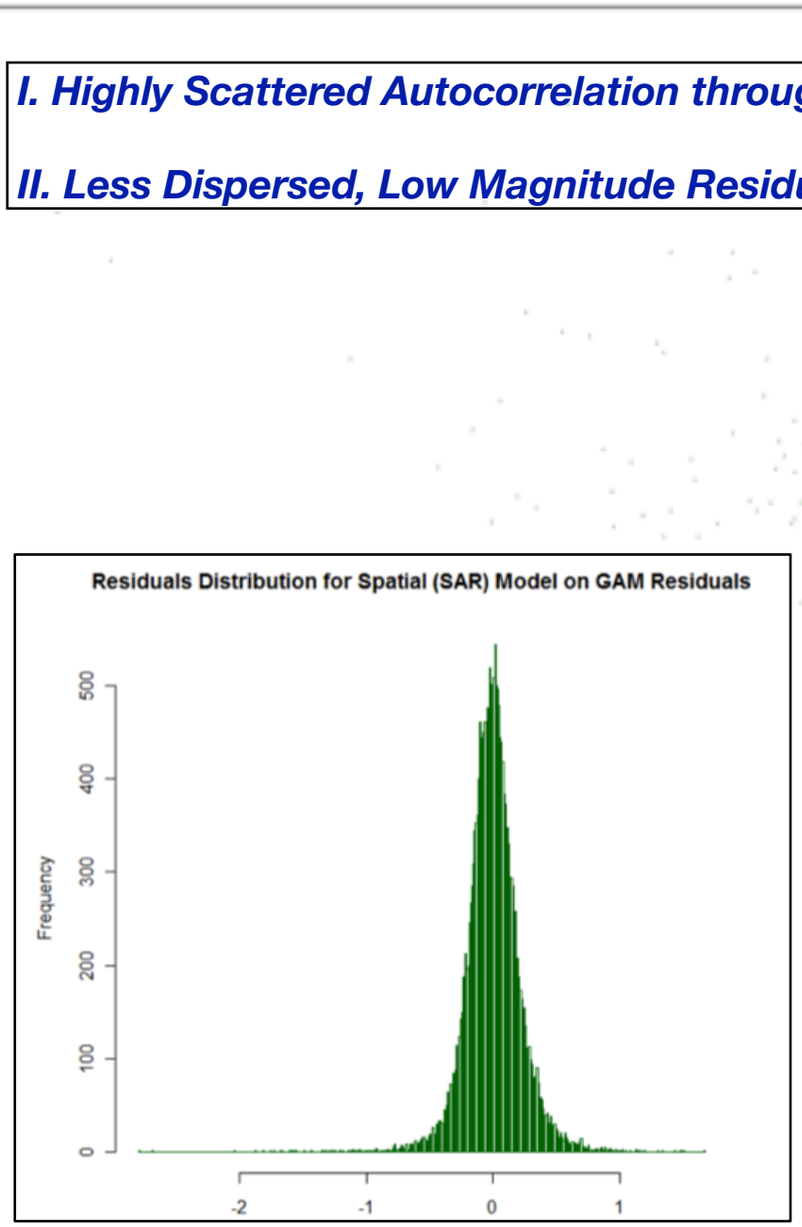
Filtering Spatial Dependence
GLM - Highly Patterned > GAM - Moderately Patterned > Spatial SAR Error Model - Least Patterned

Further Diagnostics

Correlations between Spatially Lagged Errors - Moran's I Statistics

I. Highly Scattered Autocorrelation through Moran-I Simulations

II. Less Dispersed, Low Magnitude Residuals



Spatial Simultaneous Autoregressive Error (SAR) Model Built on GAM

Residuals from SAR Error Model Built on GAM

Evolution

Location in Insurance Ratemaking & Implementation

- ***Generation I - Classical Territorial Ratemaking***
    - *Assumption: Complete Effect of location is captured in different location and demographic variables*
    - *Methods: Adding different proxy variables (Population Density, Other Geographic Variables, Different Location and Demographic Variables) in the GLM model, Credibility based approach (observed value, exposure, proximity), Kriging and Non-Geostatistical Smoothing (descriptive/ algorithmic opposed to model based)*
- ***Generation II - Latitude-Longitude in Predictive Models***
    - *Assumption: Latitude-Longitude holds significant predictive power*
    - *GLM – Use Latitude, Longitude as predictors (easting-northing effect – language, culture, food-habit). Not so promising in Insurance context.*
    - *GAM – Use a function of Latitude-Longitude as a predictor (location variables are function of latitude and longitude).*

Generation III - Spatially Correlated Observations (Insured) - Spatial Statistics Framework

- *Assumption: Unlike GLM or GAM set-up, underlying process has a spatial correlation structure that is only partially represented by GLM model*
- *Method: Filter the spatial effect to increase "correctness" in Model Estimation*
- *Consistent with GLM and GAM structure and can be built on existing GLM based Rating Tool*

Conclusion

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

---

- ***Spatial Statistics - A Rigorous Statistical Framework For Analyzing Geographically Referenced Data***
  - *Complete Distributional Inference*
  - *Captures Predictive Variation*
- ***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*
  - *Entire analysis (including all graphics) in this presentation are done in R*
- ***Communicating Model Results***
  - *Extensive Visualization Techniques*
  - *Add-on to the GLM based Rating Tool*
  - *Model Results and Diagnostics are consistent with GLM*
- ***Text Book References:***
  - *"Applied Spatial Data Analysis with R" by Roger S. Bivand, Edzer J. Pebesma and V. Gómez-Rubio (UseR! Series, Springer 2008)*
  - *"Hierarchical Modeling and Analysis of Spatial Data", by Banerjee, S., Carlin, B.P. and Gelfand, A.E. (Chapman and Hall/CRC Press, 2004)*
  - *"Basic Ratemaking," Werner, G. and Modlin, C., Casualty Actuarial Society (January - 2010)*

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