A Wavelet-Based Method for the Prospective Monitoring of Disease Occurrences in Space and Time

J. Brooke Marshall
Dan J. Spitzner
William H. Woodall
Department of Statistics
Virginia Tech

Joint Research Conference
8 June 2006

What is Public-Health Surveillance?

"Public health surveillance is the ongoing collection, analysis, interpretation, and dissemination of health data for the purpose of preventing and controlling disease, injury, and other health problems."

- Thacker, S. B. (2000) Principles and Practice of Public Health Surveillance, 2nd Ed.

Popularity of Public-Health Surveillance

- There has been greater interest in public-health surveillance lately.
  - More data available from the internet
  - Computers allow for quick assessment

- The amount of literature in this area is growing rapidly.
  - Link to methods used in Industrial Statistical Process Control (SPC)
  - Many methods incorporate control charts

Surveillance Method Concepts for Monitoring Disease Clusters

- Goal of these Methods:
  - To quickly detect a disease cluster in a geographic area as it is forming so that preventative measures can be taken.
- Retrospective vs. Prospective Analyses:
  - In retrospective analyses data is collected over time but assessment is only done once at the end of the study period. -- Delayed detection
  - In prospective analyses data is collected over time and assessment is done each time a new observation is collected. -- Quicker detection

Data used to Monitor Disease Clusters

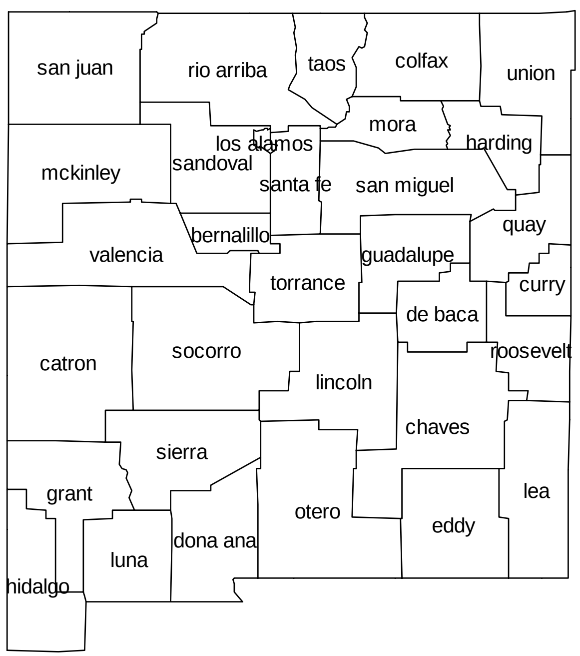
- Data for these Methods:
  - Space Component (Location in the geographic area where a disease incidence occurs.)
  - Time Component (Time at which a disease incidence occurs.)
- Forms of the Data:
  - Aggregation in Space and Time (Raubertas, 1989; Rogerson and Yamada, 2004)
  - Aggregation in Space only (Rogerson, 1997)
  - No Aggregation (Rogerson, 2001)

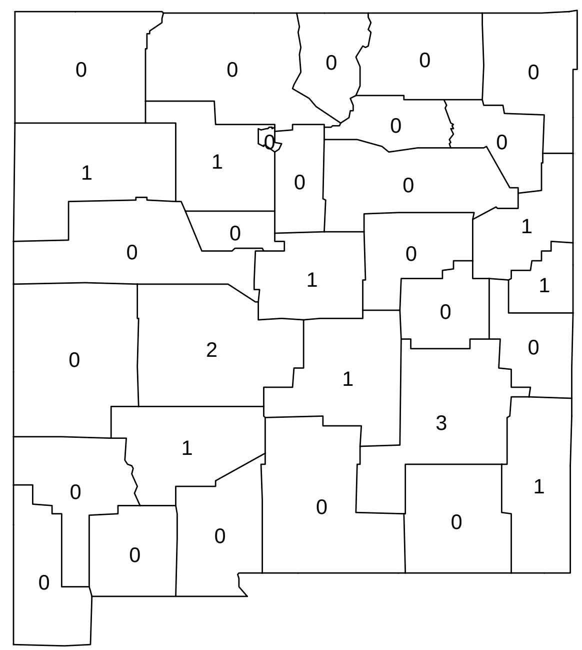
The Problem

- We want to detect clusters of disease in a geographical region.
- We have data available that is aggregated in space and time.
- In some cases, we may have information on population, age, gender, and baseline disease incidence.

Example

Yearly Male Thyroid Cancer Incidences in New Mexico (1973-1992)





Data Source: New Mexico Tumor Registry

Method Overview

- Prospectively monitor an incidence surface for the region over time
- Surface is estimated at each time point
- Estimate is obtained from a Poisson regression model with regressors from the Haar wavelet basis
- Current and past observations are used to estimate the surface
- Current observations are weighted more heavily

Outline

- Wavelet Introduction
- Monitoring Method
  - Case 1: Baseline Incidence Known
  - Case 2: Baseline Incidence Unknown
- Further Work

Wavelet Introduction
What are Wavelets?

- Wavelets are functions that have certain mathematical properties.
- A wavelet basis is a family of similar wavelet functions that is constructed from the 'mother wavelet'.
- Two common wavelet bases are the Haar basis and the Daubechies basis.

Wavelet Introduction

What do Wavelets do?

- Wavelets can be used to break down functions or signals into components of different scale or resolution.
- The lower resolutions represent the general shape of the function.
- The higher resolutions fill in more detail.

Wavelet Introduction
Traditional Uses of Wavelets

- Signal Processing
  - Radio
  - Cell Phones
- Image Processing
  - Computer Images - Internet
- Data Compression
  - Medical Images - X-rays
  - FBI fingerprints

Wavelet Introduction

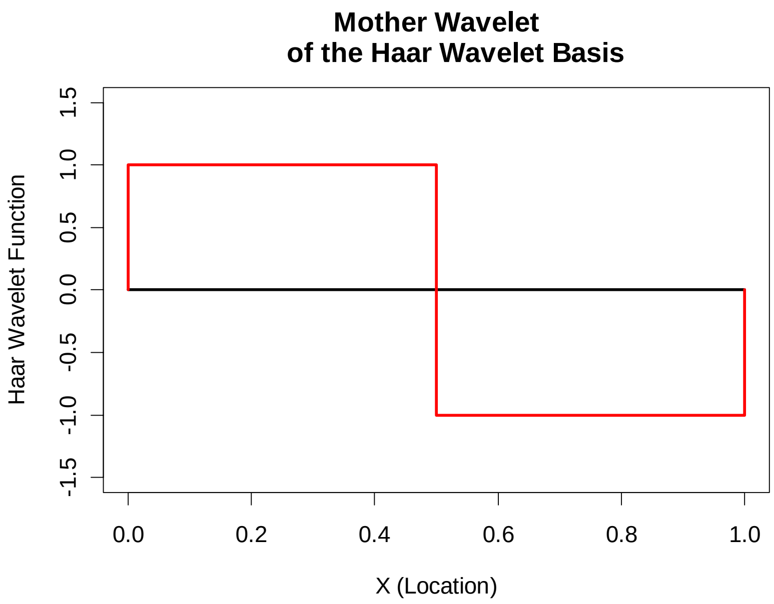
Similarity to Fourier Decomposition

- The wavelet decomposition of a function is analogous to the Fourier decomposition.
- Fourier decomposition breaks down functions into a sum of sine and cosine functions with different periods.
- A function can be approximated by a finite sum of these Fourier components.

$$F_J(x) \approx \frac{1}{2}a_0 + \sum_{j=1}^{J}[a_j \cos(jx) + b_j \sin(jx)]$$

Wavelet Introduction
The Haar Mother Wavelet

$$\psi(x) = \begin{cases} 1, & 0 \le x < \tfrac{1}{2} \\ -1, & \tfrac{1}{2} \le x < 1 \\ 0, & \text{otherwise} \end{cases}$$

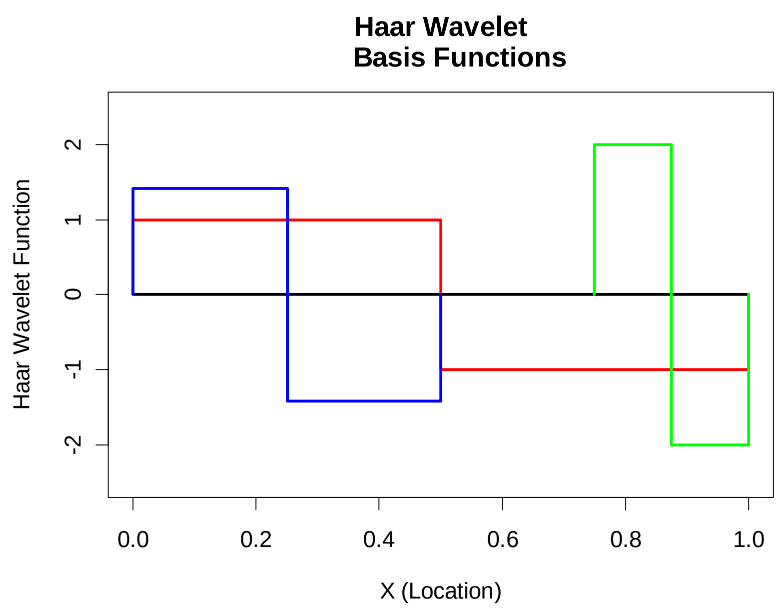


Wavelet Introduction

The Haar Wavelet Basis

- Haar family is produced by dilating and translating the mother wavelet
- $j = 0, 1, 2, \ldots$ is the dilation index
- $k = 0, 1, \ldots, 2^j-1$ is the translation index
- Basis Function:

$$\psi_{jk}(x) = 2^{j/2}\,\psi\left(2^j x - k\right)$$



Wavelet Introduction
Properties of Wavelet Bases

- A family of wavelets is a complete orthonormal system for $L^2(\Re)$.
  - Wavelet functions from the same family are orthogonal.
  - Any L^2-function, f, can be approximated by a finite linear combination of wavelet functions from the same family.

Wavelet Introduction
Aside: L^2-Functions

- An L^2-function is a function that is square-integrable and whose range is the set of real numbers.
- A square-intregable function is one where

$$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$$

- Therefore, the wavelet approximation of a function is only approximate in the L^2 meaning.

$$\int_{-\infty}^{\infty} |f(x) - \text{Approximation}|^2 dx \approx 0$$

Wavelet Introduction

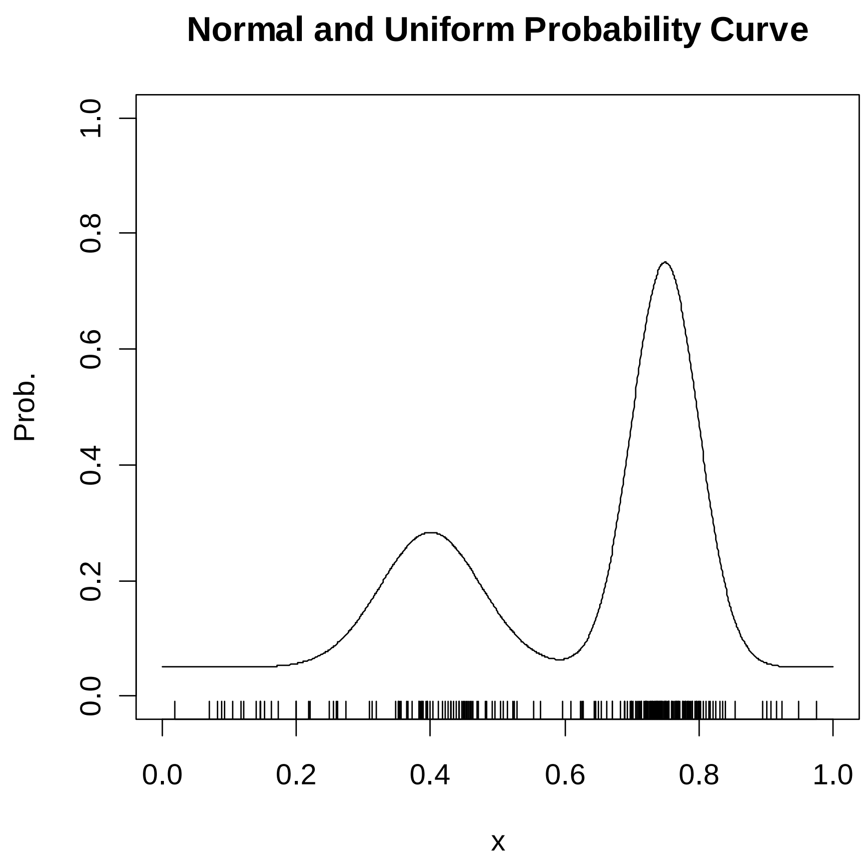
Haar Wavelet Function Approximation

- A function can be approximated with the following linear combination of Haar wavelets:

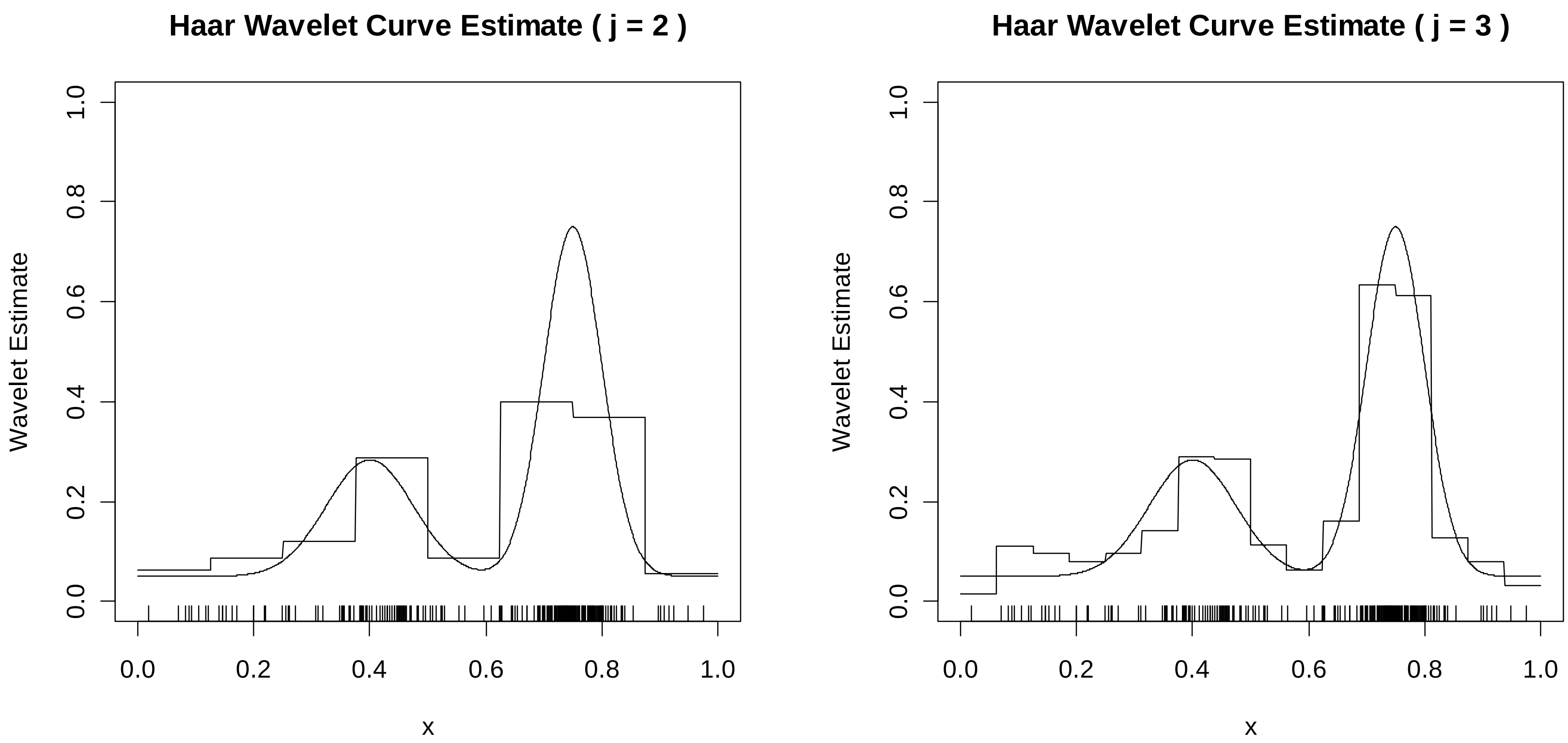
$$f(x) \approx a_0 + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} a_{jk} \psi_{jk}(x) \quad \text{where } x \in [0,1)$$

- The coefficients must be estimated.
- Regression can be used to do this easily.
  - β's represent the wavelet coefficients
  - regressors are the Haar wavelet functions

Wavelet Introduction
Example – Univariate Density Estimation

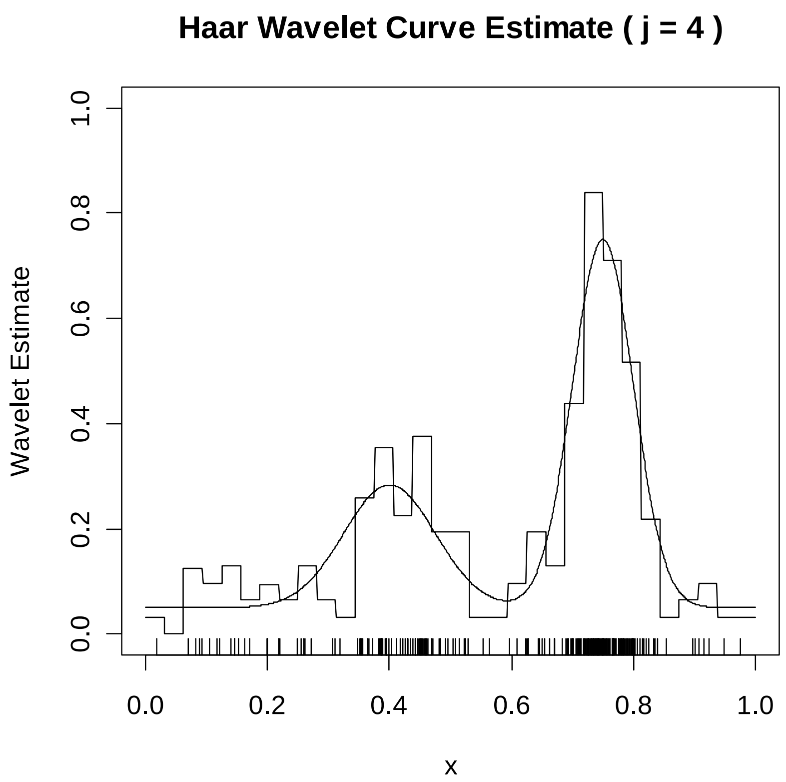


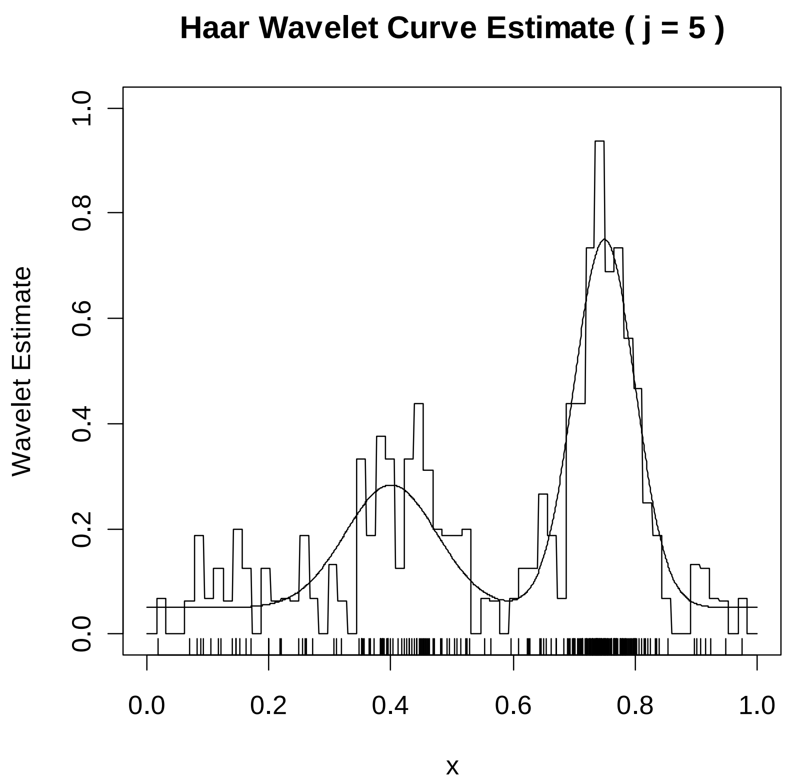
Wavelet Introduction
Example – Univariate Density Estimation



Wavelet Introduction

Example – Univariate Density Estimation





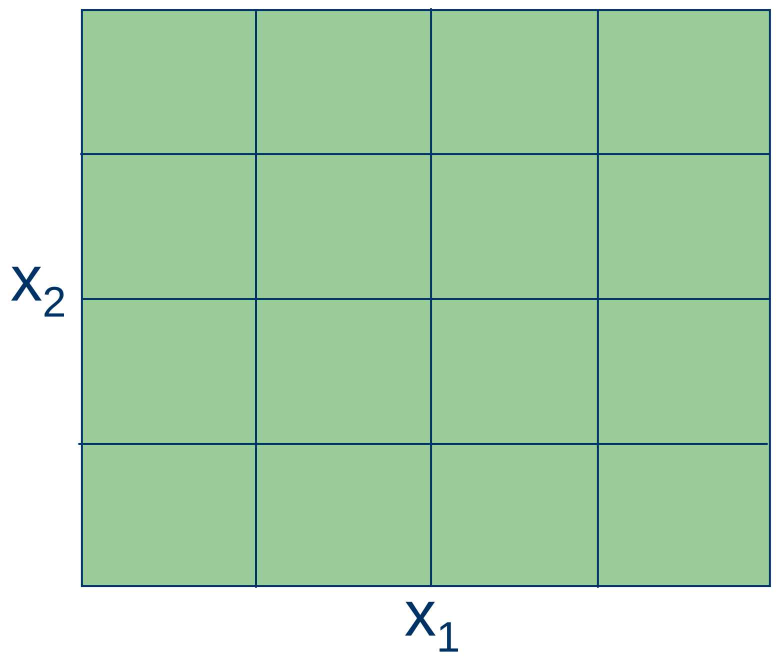
Wavelet Introduction
Haar Wavelet Scaling Function

- Notice that x must be between 0 inclusively and 1 exclusively.
- To approximate a function with different support, a scaling function, also called the father wavelet, is used.
- Haar scaling function for $x \in [0,1)$ is $\phi(x) = I_{[0,1)}(x)$
- Haar scaling function for other domains is $\phi_{jk}(x) = 2^{j/2}\,\phi(2^{j}\text{-}k)$

Wavelet Introduction

Two-Dimensional Haar Wavelet Basis

- The two-dimensional Haar wavelet basis can be constructed by taking cross products of the wavelet and scaling functions over $x_1 \in (0,1]$ and $x_2 \in (0,1]$:
  - $\psi(x_1) \times \phi(x_2)$
  - $\phi(x_1) \times \psi(x_2)$
  - $\psi(x_1) \times \psi(x_2)$



Wavelet Introduction
Haar Wavelet Surface Approximation

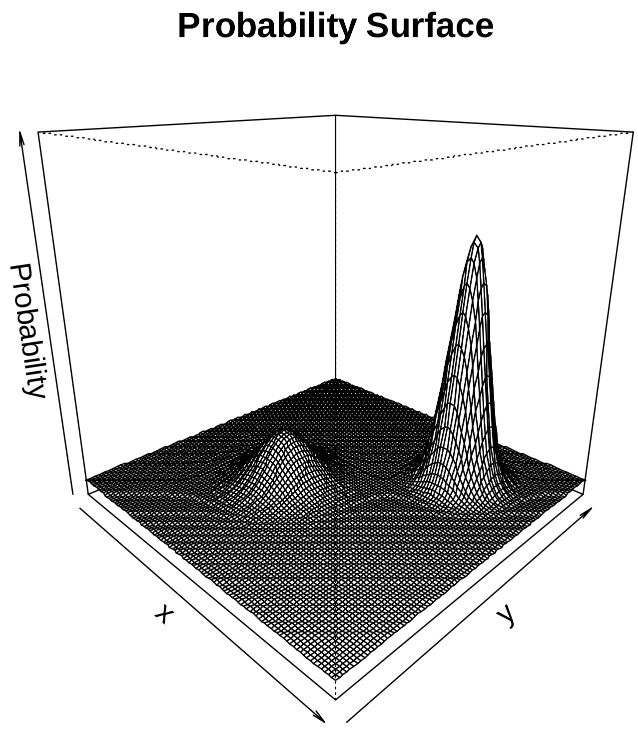
- A surface can be approximated with the following linear combination of Haar wavelets:

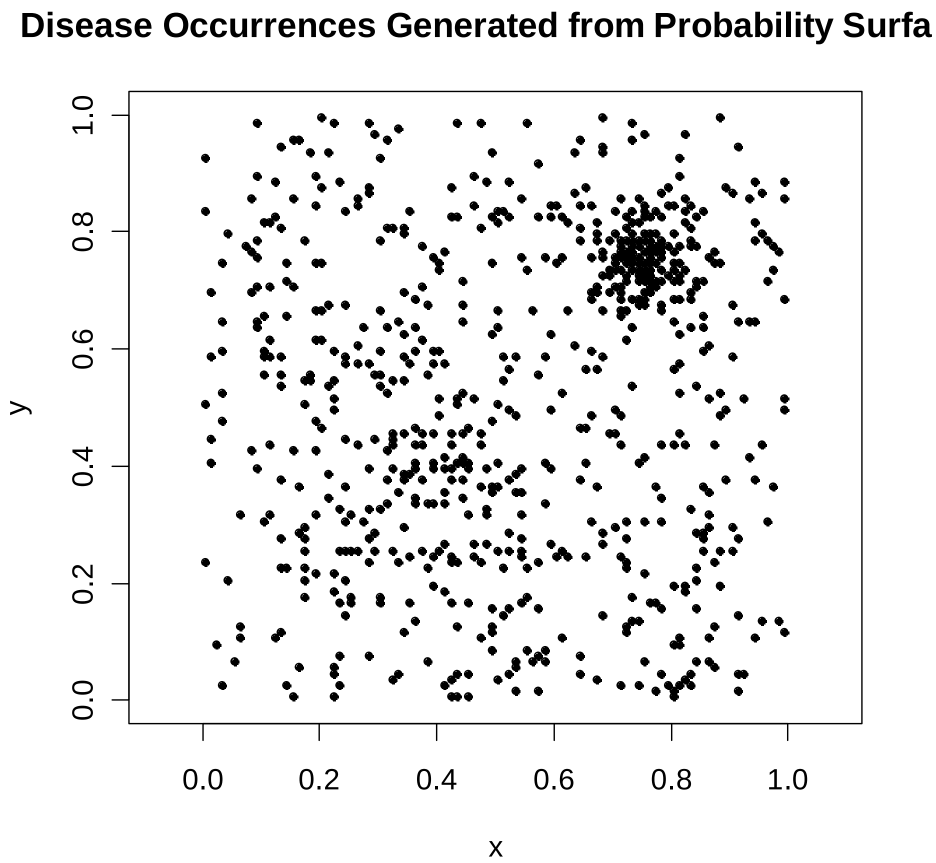
$$f(x_1,x_2) \approx a_0 + \sum_{j_1=0}^{J_1}\sum_{k_1=0}^{2^{j_1}-1} a_{j_1k_1}\psi_{j_1k_1}(x_1) + \sum_{j_2=0}^{J_2}\sum_{k_2=0}^{2^{j_2}-1} a_{j_2k_2}\psi_{j_2k_2}(x_2)$$

$$+ \sum_{j_1=0}^{J_1}\sum_{k_1=0}^{2^{j_1}-1}\sum_{j_2=0}^{J_2}\sum_{k_2=0}^{2^{j_2}-1} a_{j_1k_1j_2k_2}\psi_{j_1k_1}(x_1)\psi_{j_2k_2}(x_2)$$

Wavelet Introduction

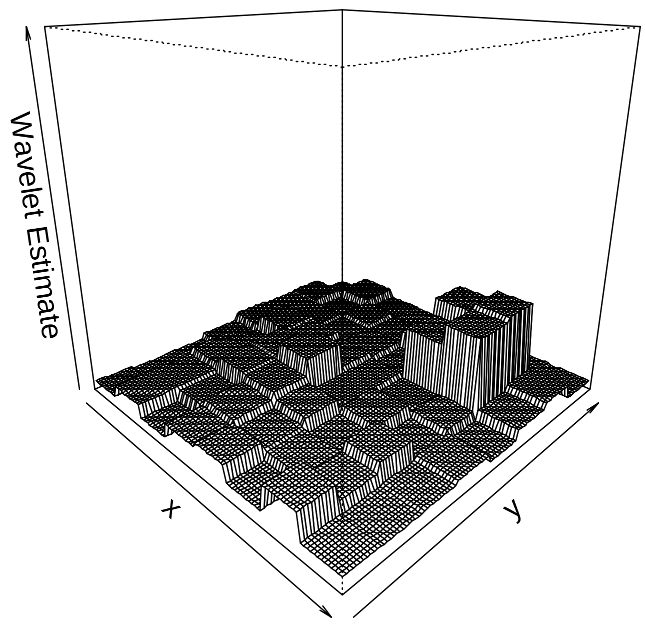
Example – Multivariate Density Estimation



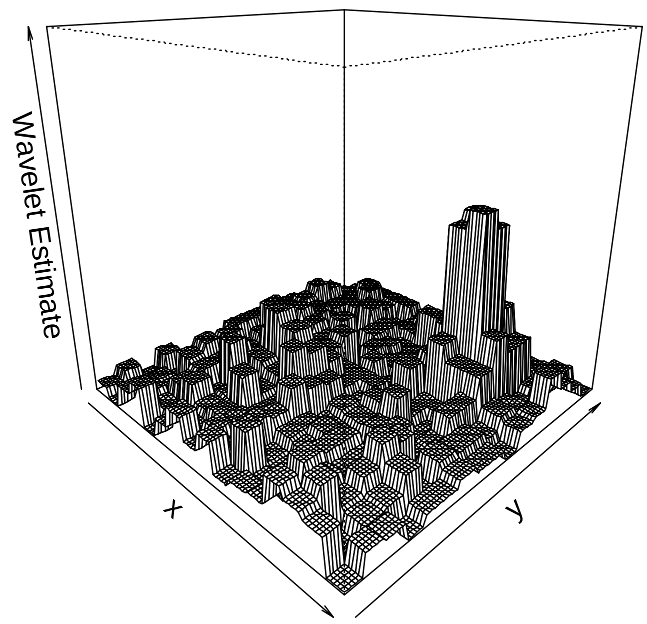


Wavelet Introduction
Example – Multivariate Density Estimation

Haar Wavelet Surface Estimate (j = 2)



Haar Wavelet Surface Estimate (j = 3)



Monitoring Method

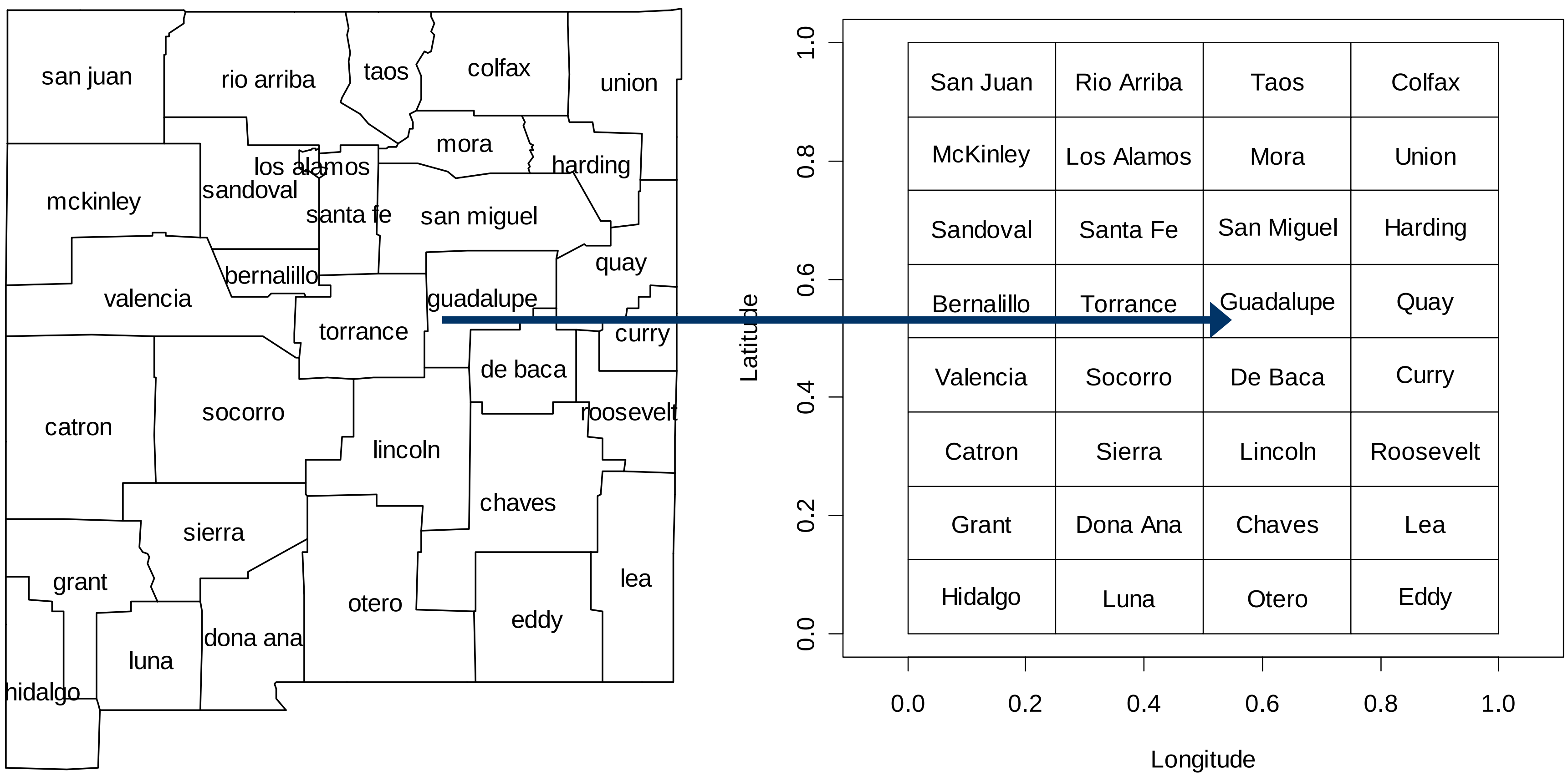
Why use Wavelets for Monitoring?

- Wavelets can easily model an incidence surface of any form and still give a parametric model for testing.
- Wavelet functions are orthogonal so the coefficients in the model are independent.
- The multiresolution of wavelet functions allows us to detect clusters of different size.
- Multiresolution can also give us more powerful global tests.

Monitoring Method

Mapping of Counties to Wavelet Domain

Allocation of Counties in New Mexico to Wavelet Domain



Monitoring Method
Case 1: Data

- Data is coming in at equal time intervals.
- At each time we get a count or rate of incidence for each county.
- Each observation is assumed to be independent.
- Only the current observation is used to estimate the incidence surface.

Monitoring Method
Case 1: Surface Estimation

Poisson Regression Model:

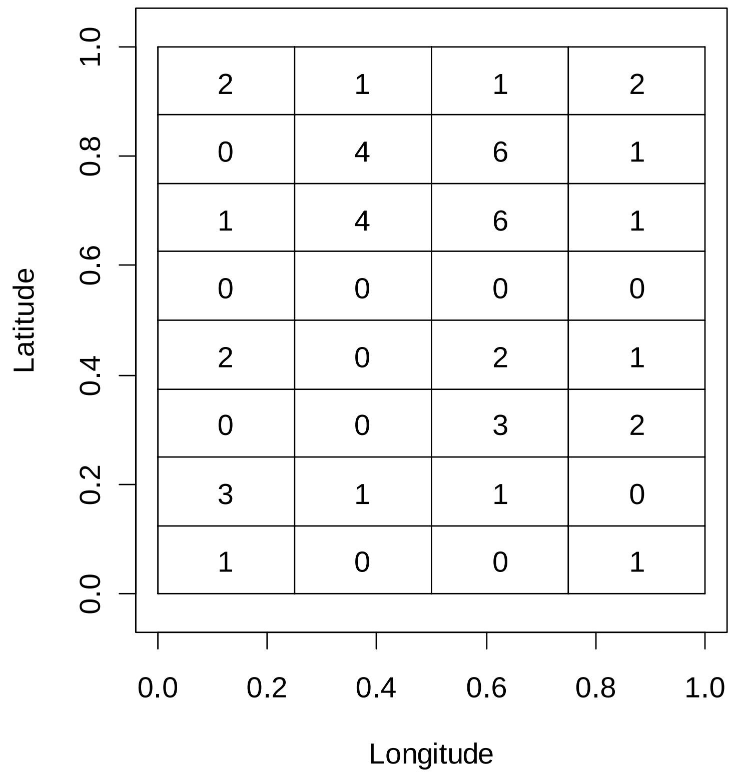
$$\underline{\mu} = e^{\underline{1}\beta_0 + \Psi_S \underline{\beta}_S}$$

Ψ_S are the wavelet function values over the region

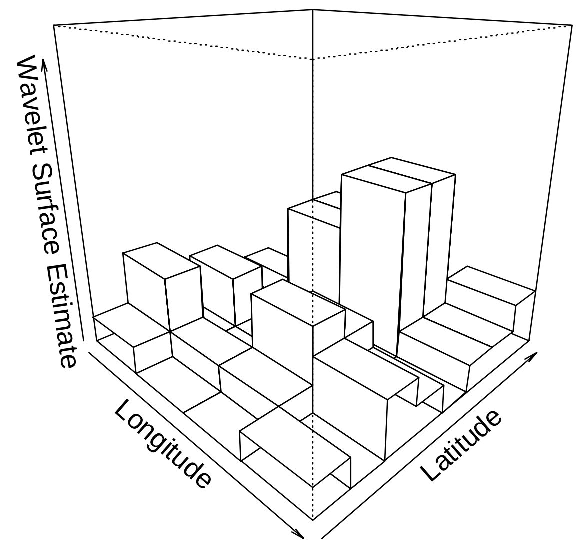
- A set of Haar wavelet functions for Longitude
- A set of Haar wavelet functions for Latitude
- All cross products of the wavelet functions for Longitude and Latitude

Monitoring Method
Case 1: Surface Estimation Example

Randomly Generated Incidence Counts at Time = 17

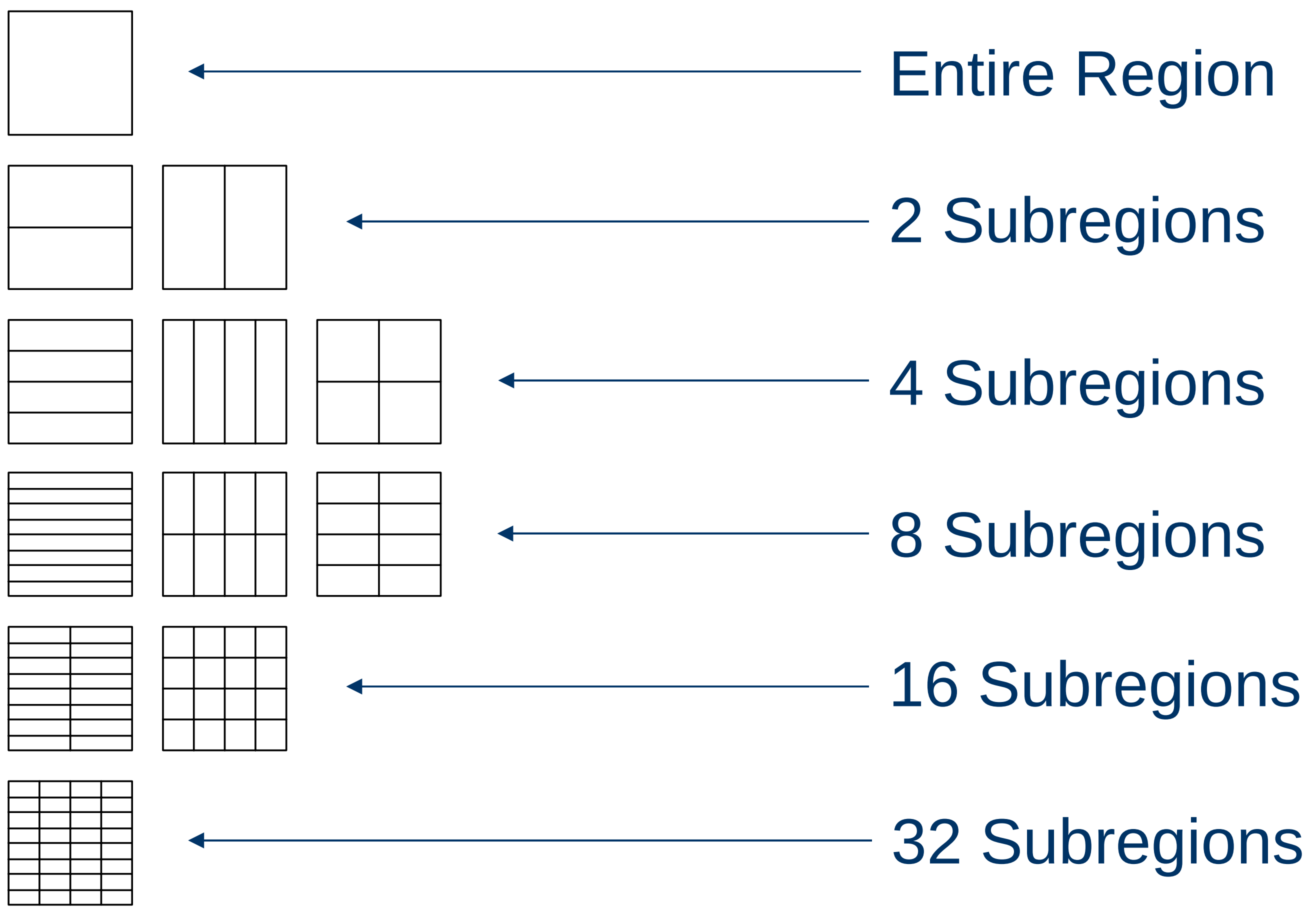


Haar Wavelet Surface Estimate at Time = 17



Monitoring Method

Possible Clusters



Monitoring Method

Case 1: Control Charts for Global Statistics

- Does the surface change from a baseline?

  H_0: $\underline{\beta}_S = \underline{\beta}_{BASELINE}$

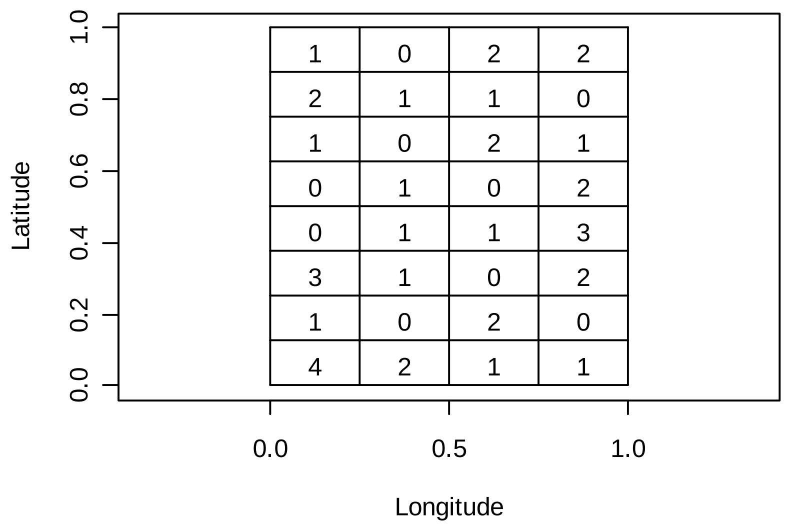
- Does the mean incidence increase over the entire region?

  H_0: $\beta_0 = \underline{\beta}_{0\ BASELINE}$

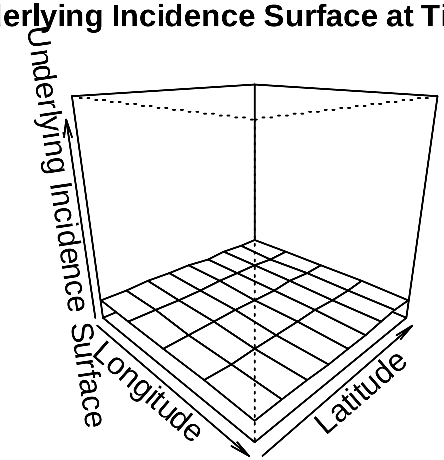
  *The Wald Test or GLRT can be used here.

- Monitor global statistics over time with Chi-square or Normal CUSUM control charts designed using standard ARL results.

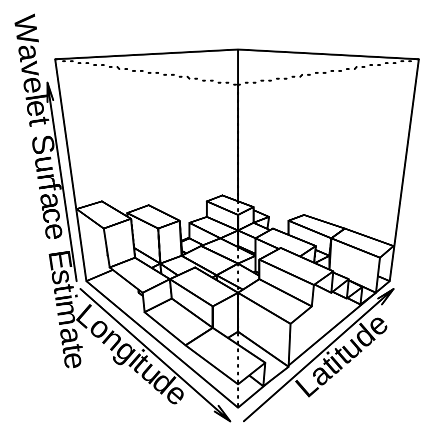
Randomly Generated Incidence Counts at Time = 10



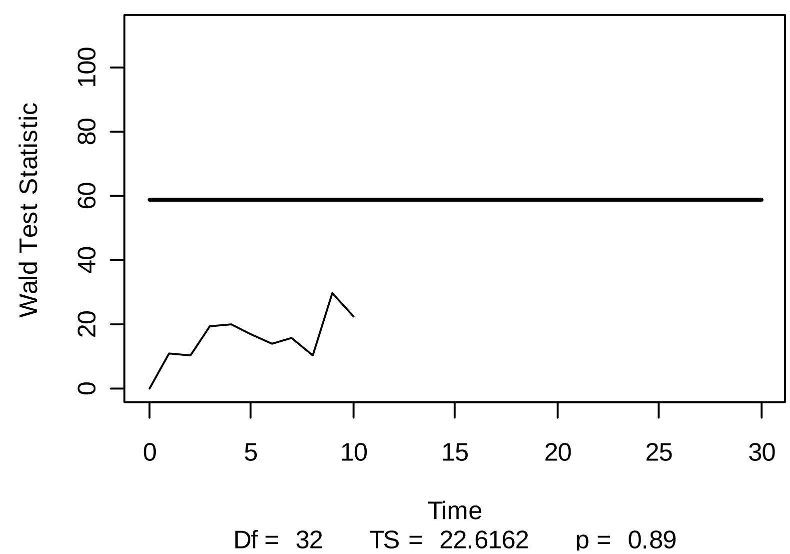
Underlying Incidence Surface at Time = 10



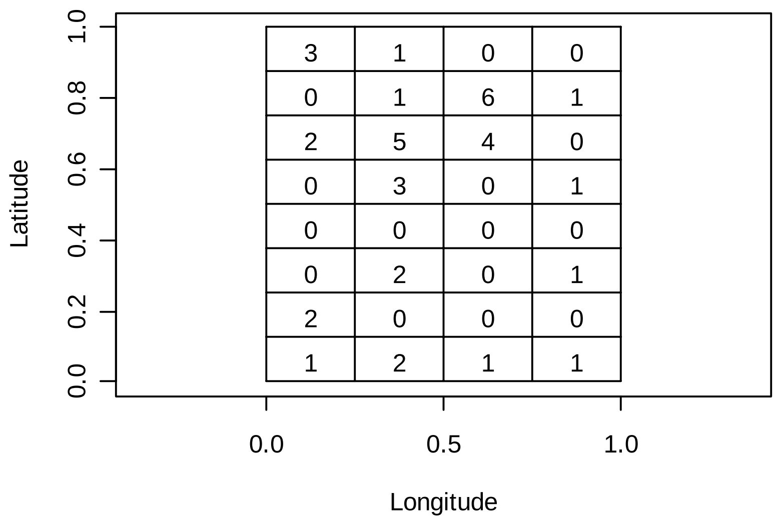
Haar Wavelet Surface Estimate at Time = 10



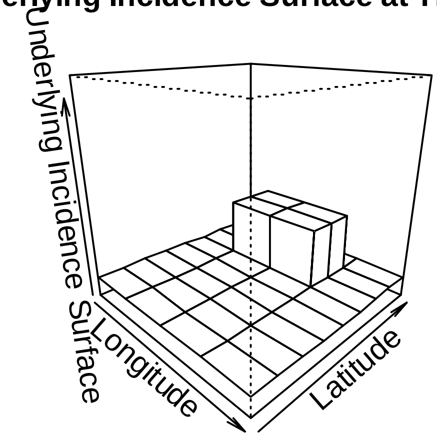
Global Baseline Control Chart at Time = 10



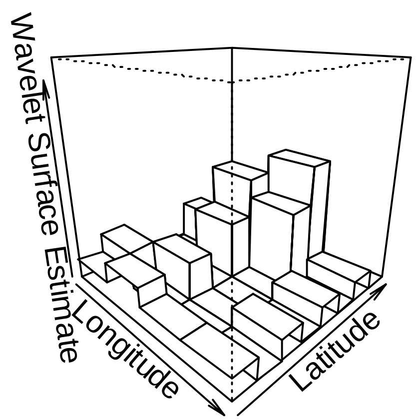
Randomly Generated Incidence Counts at Time = 16



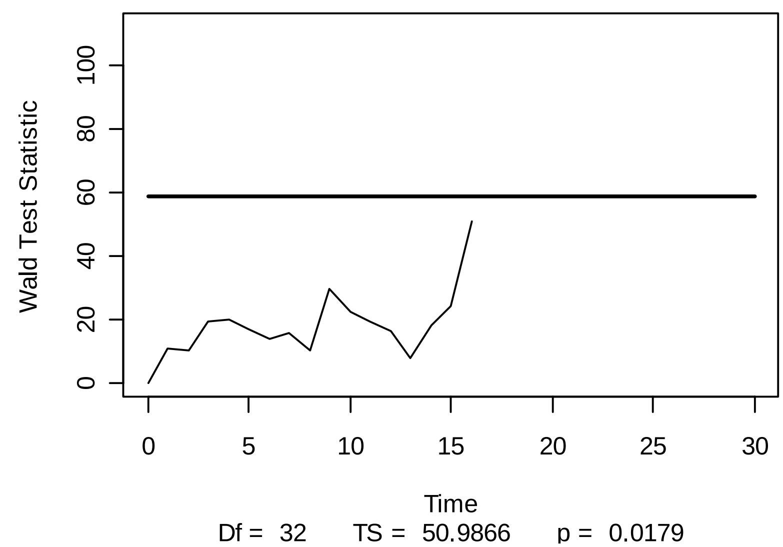
Underlying Incidence Surface at Time = 16



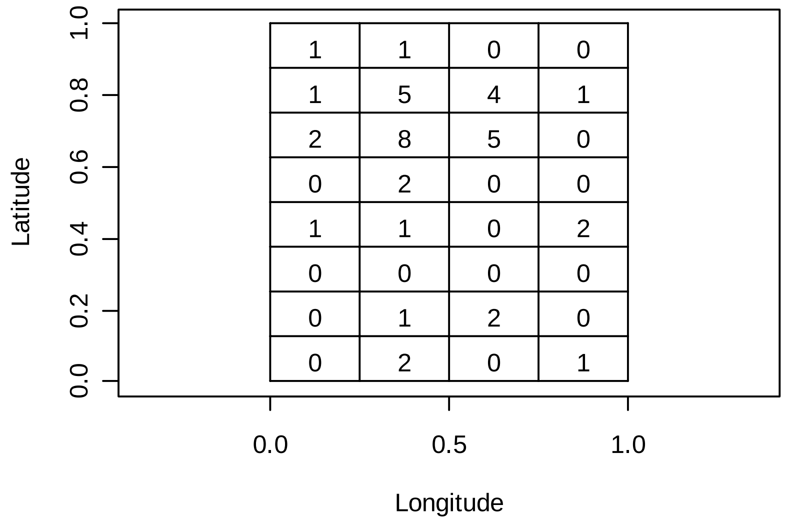
Haar Wavelet Surface Estimate at Time = 16



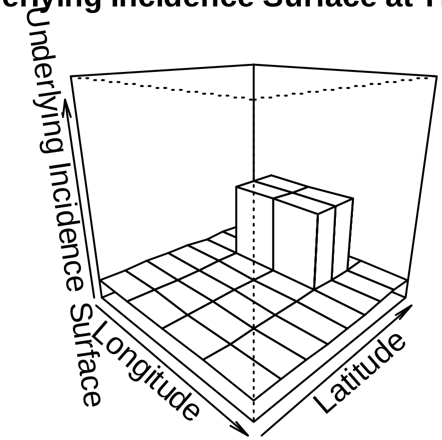
Global Baseline Control Chart at Time = 16



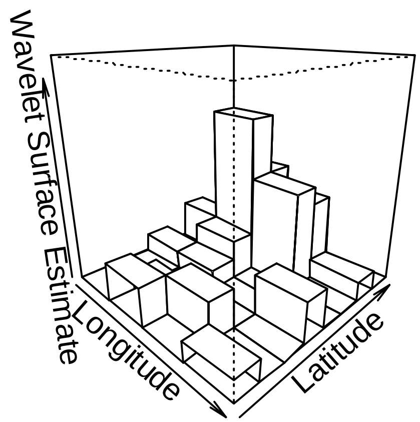
Randomly Generated Incidence Counts at Time = 30



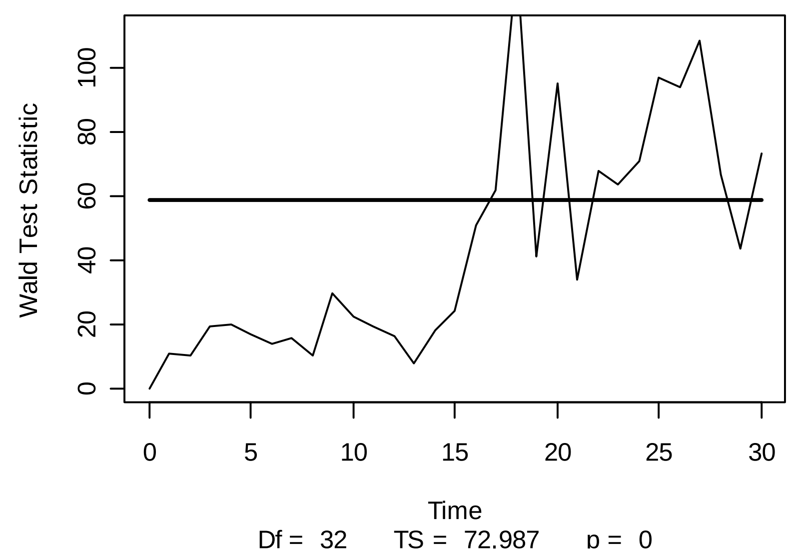
Underlying Incidence Surface at Time = 30



Haar Wavelet Surface Estimate at Time = 30



Global Baseline Control Chart at Time = 30



Monitoring Method

Case 1: An Alternative Global Statistic

- A weighted χ^2 test can be used in place of the Wald test.
- Wald = $\underline{\mathbf{a}}^T\underline{\mathbf{a}} \;\dot{\sim}\; \chi_\nu^2$
- Weighted $\chi^2 = \underline{\mathbf{a}}^T\mathbf{W}\underline{\mathbf{a}} \;\dot{\sim}\; \tau\chi_\nu^2 \Rightarrow \Gamma(\eta/2, 2\tau)$
- **W** is a weight matrix where the weights can be chosen to emphasize possible clusters of most importance.

Monitoring Method
Case 1: Control Charts for Local Statistics

- Detect multi-level clustering

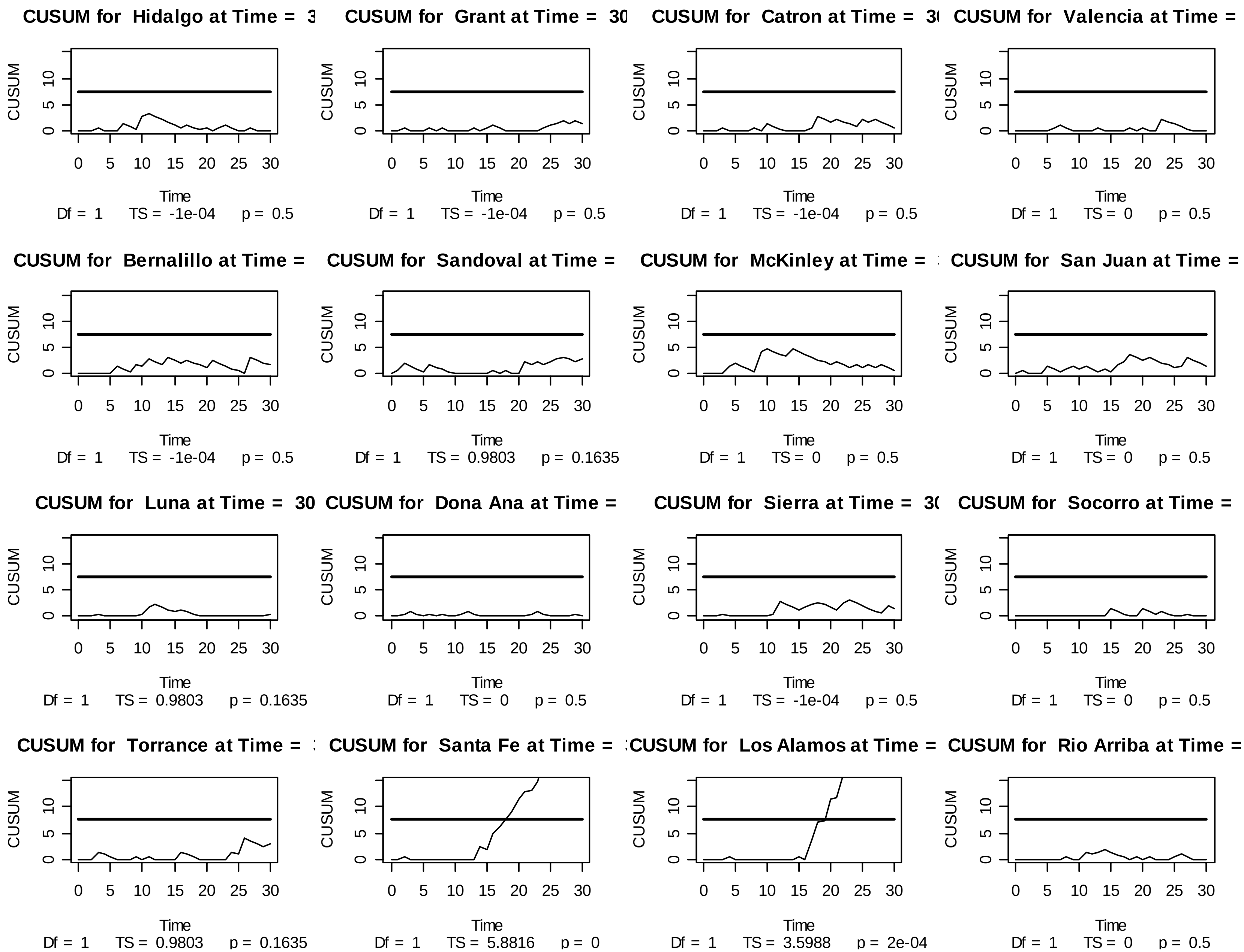
  H_0: $\beta_i = 0$ – test appropriate coefficients <u>OR</u>

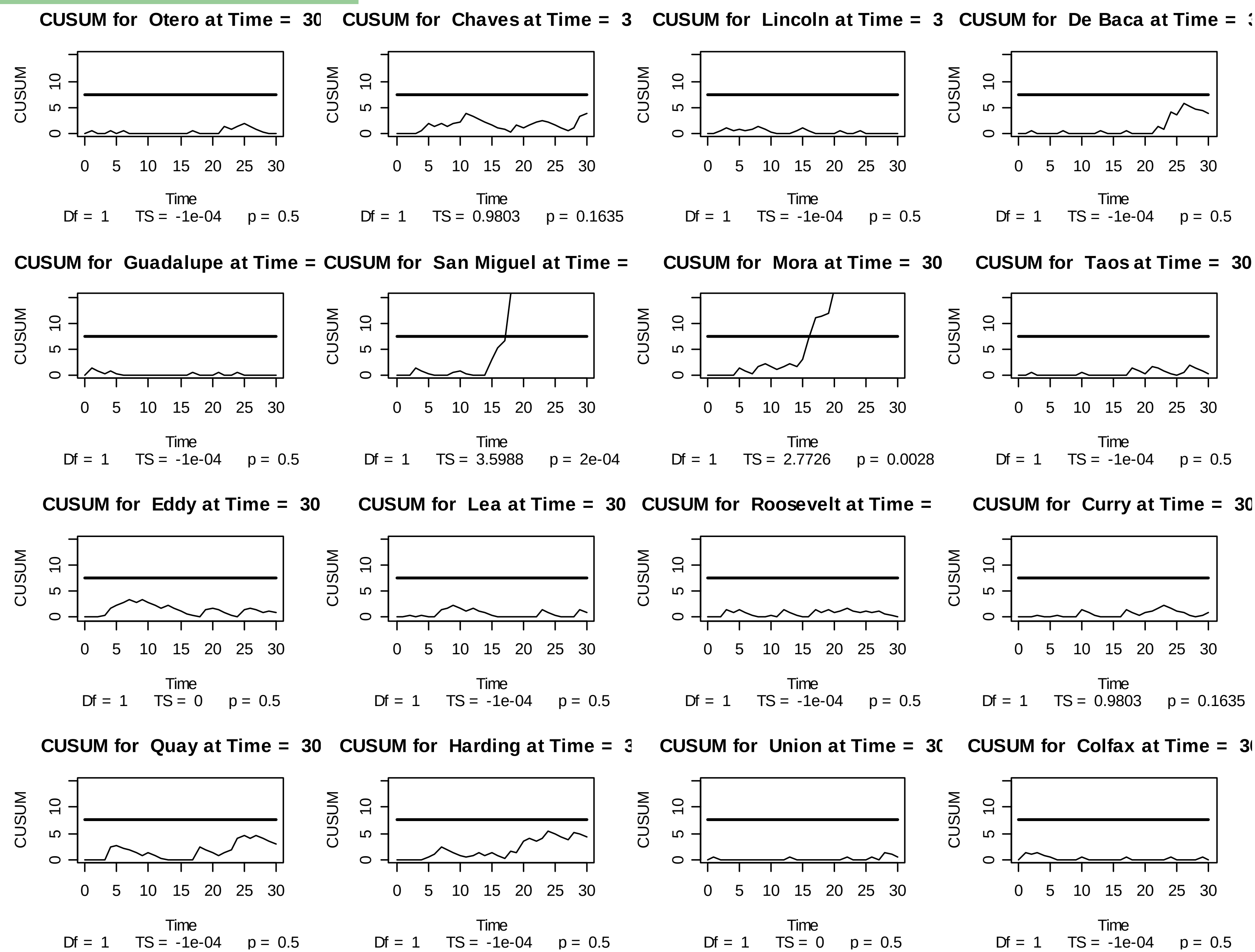
  H_0: predicted λ_i = baseline λ_i for each cluster i.

- Detect increases from baseline in individual areas

  H_0: predicted λ_i = baseline λ_i for each county i.

- Monitor local statistics over time with Chi-square or Normal CUSUM control charts designed using standard ARL results.


CUSUM for Hidalgo at Time = 3
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Grant at Time = 30
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Catron at Time = 3
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Valencia at Time =
Df = 1 TS = 0 p = 0.5
CUSUM for Bernalillo at Time =
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Sandoval at Time =
Df = 1 TS = 0.9803 p = 0.1635
CUSUM for McKinley at Time =
Df = 1 TS = 0 p = 0.5
CUSUM for San Juan at Time =
Df = 1 TS = 0 p = 0.5
CUSUM for Luna at Time = 30
Df = 1 TS = 0.9803 p = 0.1635
CUSUM for Dona Ana at Time =
Df = 1 TS = 0 p = 0.5
CUSUM for Sierra at Time = 3
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Socorro at Time = 3
Df = 1 TS = 0 p = 0.5
CUSUM for Torrance at Time =
Df = 1 TS = 0.9803 p = 0.1635
CUSUM for Santa Fe at Time =
Df = 1 TS = 5.8816 p = 0
CUSUM for Los Alamos at Time =
Df = 1 TS = 3.5988 p = 2e-04
CUSUM for Rio Arriba at Time =
Df = 1 TS = 0 p = 0.5
CUSUM
Time


CUSUM for Otero at Time = 30
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Chaves at Time = 3
Df = 1 TS = 0.9803 p = 0.1635
CUSUM for Lincoln at Time = 3
Df = 1 TS = -1e-04 p = 0.5
CUSUM for De Baca at Time =
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Guadalupe at Time =
Df = 1 TS = -1e-04 p = 0.5
CUSUM for San Miguel at Time =
Df = 1 TS = 3.5988 p = 2e-04
CUSUM for Mora at Time = 30
Df = 1 TS = 2.7726 p = 0.0028
CUSUM for Taos at Time = 30
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Eddy at Time = 30
Df = 1 TS = 0 p = 0.5
CUSUM for Lea at Time = 30
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Roosevelt at Time =
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Curry at Time = 30
Df = 1 TS = 0.9803 p = 0.1635
CUSUM for Quay at Time = 30
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Harding at Time = 3
Df = 1 TS = -1e-04 p = 0.5
CUSUM for Union at Time = 30
Df = 1 TS = 0 p = 0.5
CUSUM for Colfax at Time = 30
Df = 1 TS = -1e-04 p = 0.5
CUSUM
Time

Monitoring Method
Case 2: Data

- Data is coming in at equal time intervals. -- Same as Case 1
- At each time we get a count or rate of incidence for each county. -- Same as Case 1
- Each observation is assumed to be independent. -- Same as Case 1
- Current and past observations are used to estimate the incidence surface. Observations are weighted using the EWMA weighting scheme.

Monitoring Method
Case 2: Surface Estimation

- Regressors for Space
  - Same as in Case 1
- Regressor for Time
  - Change in time of each observation and the current observation
  - Motivation for this comes from the Taylor Series Expansion of a function -- $f(t) \approx f(T) + (t - T)f'(T)$
- Regressors for Space and Time
  - All space regressors are multiplied by the time regressor

Monitoring Method
Case 2: Surface Estimation

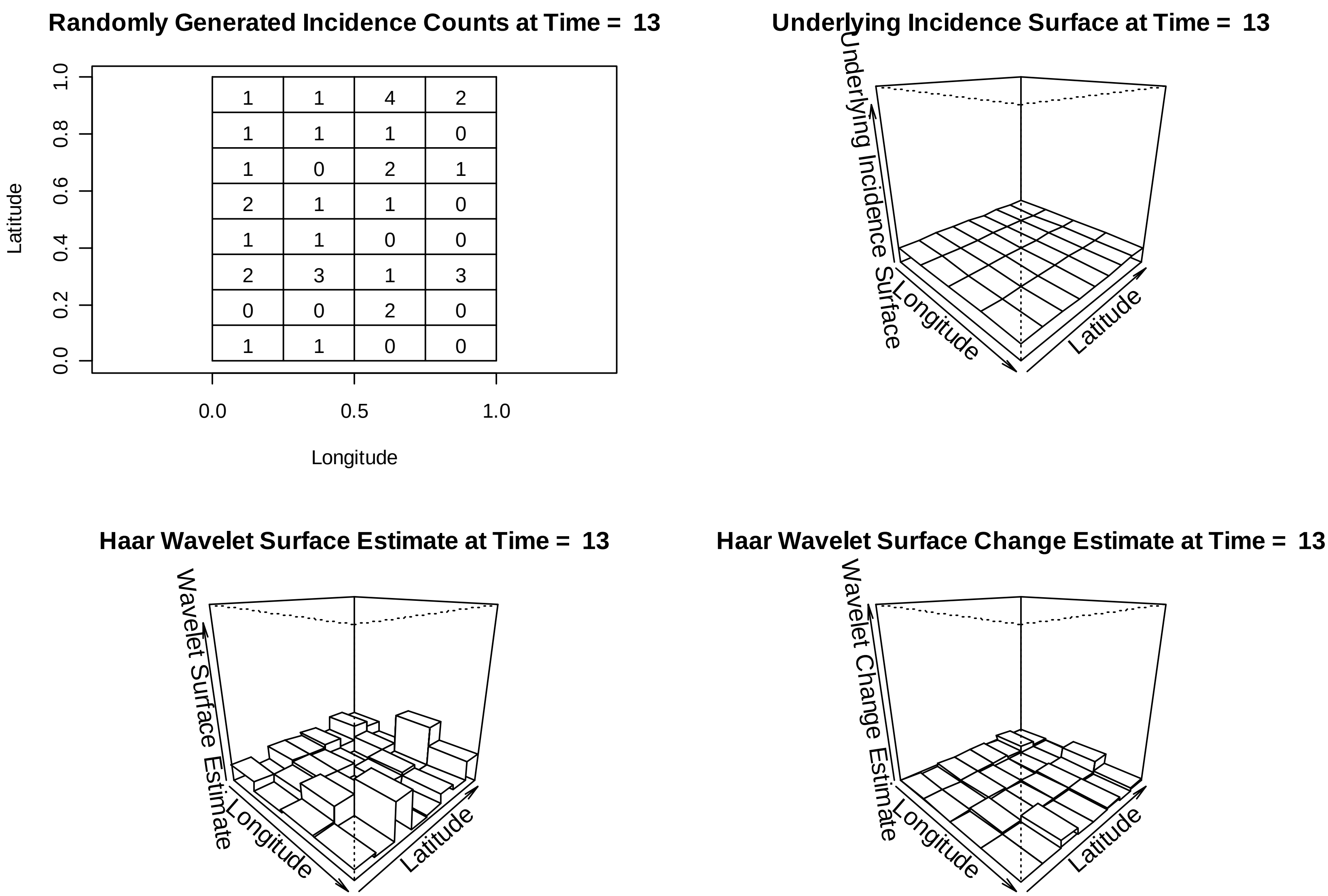
Poisson Regression Model:

$$\underline{\mu} = e^{\underline{1}\beta_0 + \Psi_S \underline{\beta}_S + \underline{\delta}_T \beta_T + [\Psi_S \times \underline{\delta}_T] \underline{\beta}_{ST}}$$

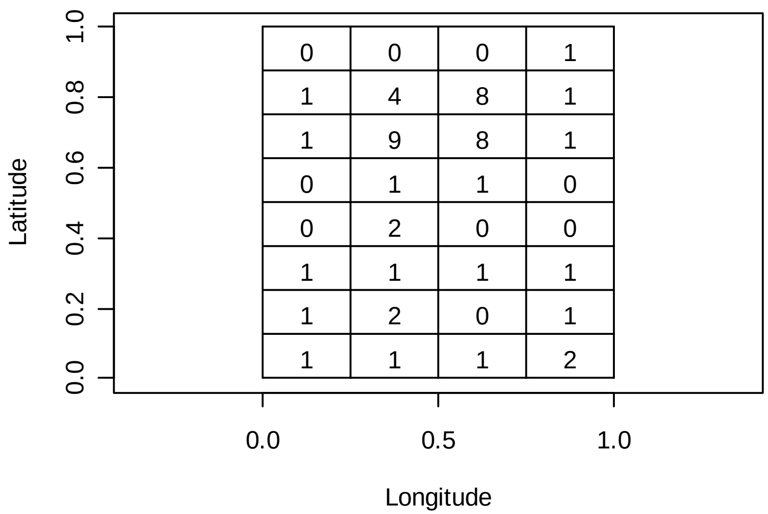
Ψ_S are the wavelet function values over the region (Space Regressor Matrix)

$\underline{\delta}_T$ are the changes in time from the current time (Time regressor)

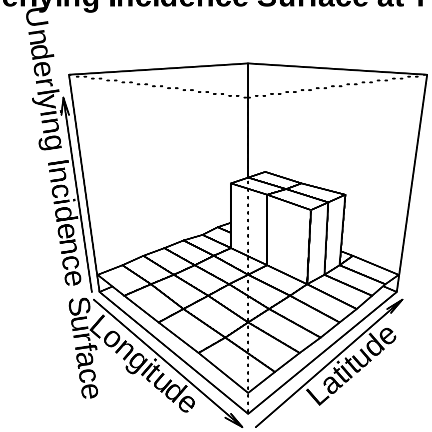
$[\Psi_S \times \underline{\delta}_T]$ are the products of the space and time regressors


Randomly Generated Incidence Counts at Time = 13
1 1 4 2
1 1 1 0
1 0 2 1
2 1 1 0
1 1 0 0
2 3 1 3
0 0 2 0
1 1 0 0
Latitude
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.5 1.0
Longitude
Underlying Incidence Surface at Time = 13
Underlying Incidence Surface
Longitude
Latitude
Haar Wavelet Surface Estimate at Time = 13
Wavelet Surface Estimate
Longitude
Latitude
Haar Wavelet Surface Change Estimate at Time = 13
Wavelet Change Estimate
Longitude
Latitude

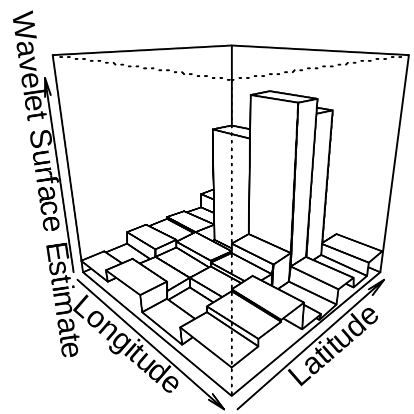
Randomly Generated Incidence Counts at Time = 18



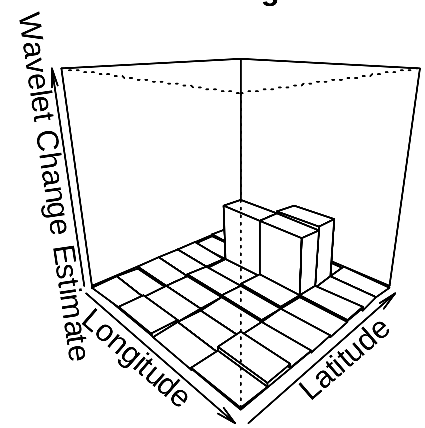
Underlying Incidence Surface at Time = 18



Haar Wavelet Surface Estimate at Time = 18



Haar Wavelet Surface Change Estimate at Time = 18



Monitoring Method
Case 2: Global Statistics

- Does the surface change over time?

  $H_0: \begin{bmatrix} \beta_T \\ \underline{\beta}_{ST} \end{bmatrix} = \underline{0}$

- Does the mean incidence increase over time?

  $H_0: \beta_T = 0$

- Is there space-time interaction?

  $H_0: \underline{\beta}_{ST} = \underline{0}$

*Same test statistics for Case 1 can be used here.

Monitoring Method
Case 2: Local Statistics

- Detect multi-level clustering

  H_0: $\beta_i = 0$ – test appropriate change coefficients

  OR

  H_0: predicted change in $\lambda_i = 0$ for each cluster i.

- Detect relative mean increases in individual counties

  H_0: predicted change in $\lambda_i = 0$ for each county i.

Monitoring Method

Case 2: Control Charts

- Similar control charts to those used in Case 1 are used to monitor the global and local statistics.
- There is autocorrelation present because past observations are used in the estimation process.
- ARL performance must be determined by simulation to design chart.

Monitoring Method
Further Work

- Covariate information, such as population size of the county, needs to be incorporated by changing the values of the wavelet functions.
- The wavelet values are changed so that the orthogonality is maintained with respect to the covariate.
- Covariate information on age and gender can easily be added to the model.

Monitoring Method
Further Work

- Problem with number of counties not equal to a power of two can be solved in two ways:
  - By assigning zero incidences to these squares at each time period
  - Or possibly by leaving out the observations associated with these cells
- Need to explore how weights of past observations influence performance

References

- Hawkins, D. M. and Olwell, D. H. (1998) *Cumulative Sum Charts and Charting for Quality Improvement*. New York, Springer.
- Ogden, T. (1997) *Essential Wavelets for Statistical Applications*. Birkhäuser: Boston.
- Spitzner, D. J. (2006) Use of goodness-of-fit procedures in high dimensional testing. *Journal of Statistical Computation and Simulation* **76**(5): 447-457.
- Spitzner, D. J. (2006) Testing in Functional Data Analysis using Quadratic Forms. Submitted to *Annals of Statistics*.
- Vidakovic, B. V. (1999) *Statistical Modeling by Wavelets*. Wiley: New York.
- Woodall, W. H. (2006) The use of control charts in health-care monitoring and public-health surveillance. *Journal of Quality Technology* **38**(2): 89-104.

Contact Information

J. Brooke Marshall

Department of Statistics

Virginia Tech

406-A Hutcheson Hall

Blacksburg, VA 24061-0439

email: jemarsh2@vt.edu