An Introduction to Pattern Statistics

- Nearest Neighbors
  - The CSR hypothesis
  - Clark/Evans and modification
  - Cuzick and Edwards and controls
- All events
  - k function
  - Weighted k function
  - Comparative k functions

Nearest Neighbors

- The CSR Assumptions
  1. All possible sites are equally likely to receive a point

  2. The placement of a point is independent of the placement of all other points

- Quadrats or distances
- The Poisson Distribution

$$P(x) = \lambda^x e^{-\lambda} / x! \quad \text{For } x=0,1,2,...$$

Clark/Evans and Modification

- Distance-based
- Finds expected distance to nearest neighbor in a CSR pattern: [E(d)]
- $E(d) = 0.5\ [(A/N)]^{0.5} + [0.0514 + 0.041/\ (N)^{0.5}]\ B/N$

  and

  $Var(mean\ d) = 0.070\ A/N^2 + 0.037\ B\ [A/(N^5)]^{0.5}$
- $Z = [(observed\ mean\ d) - E(d)]\ /\ [Var(mean\ d)]^{0.5}$

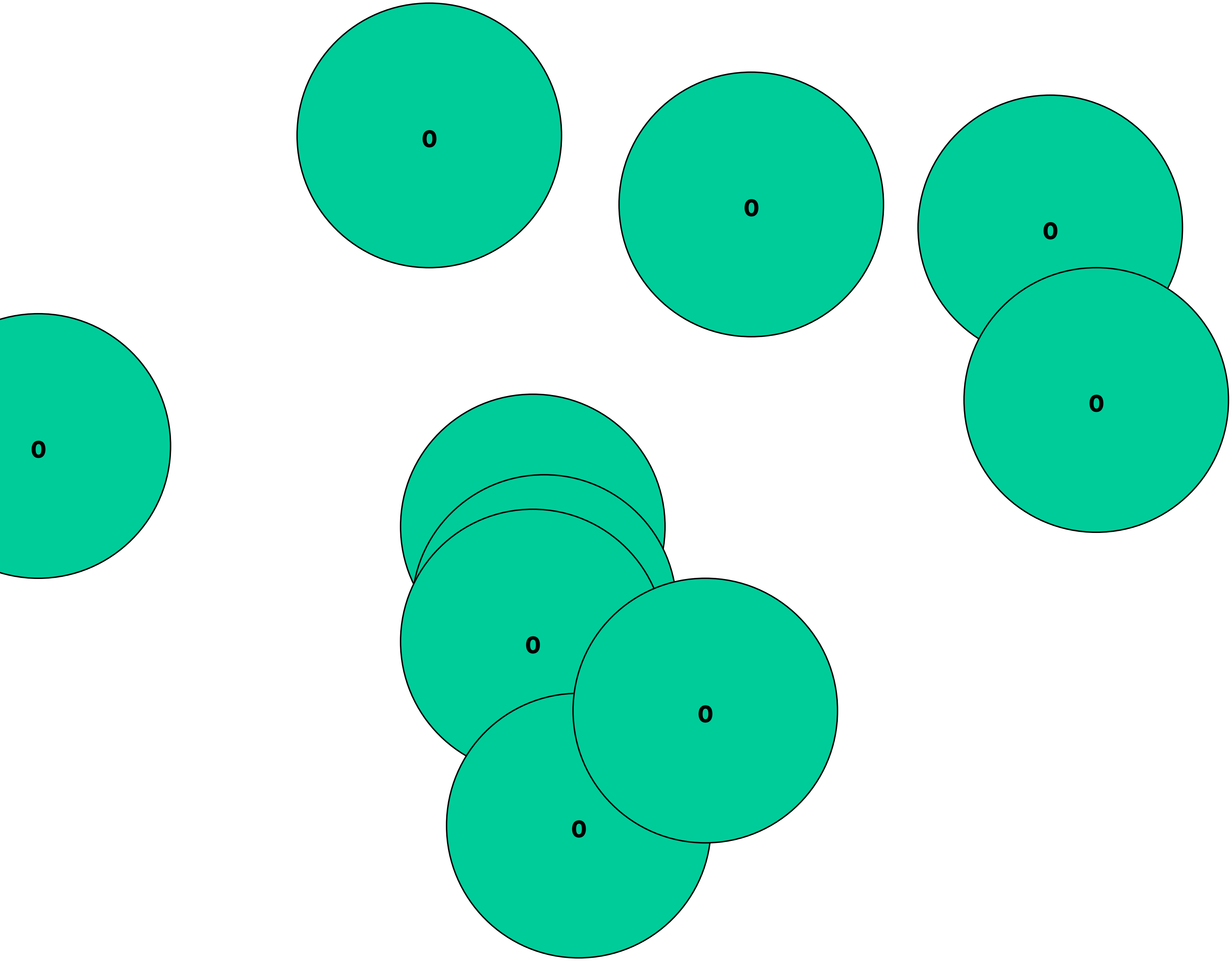
where A = area, N = total number of points, B = length of the perimeter

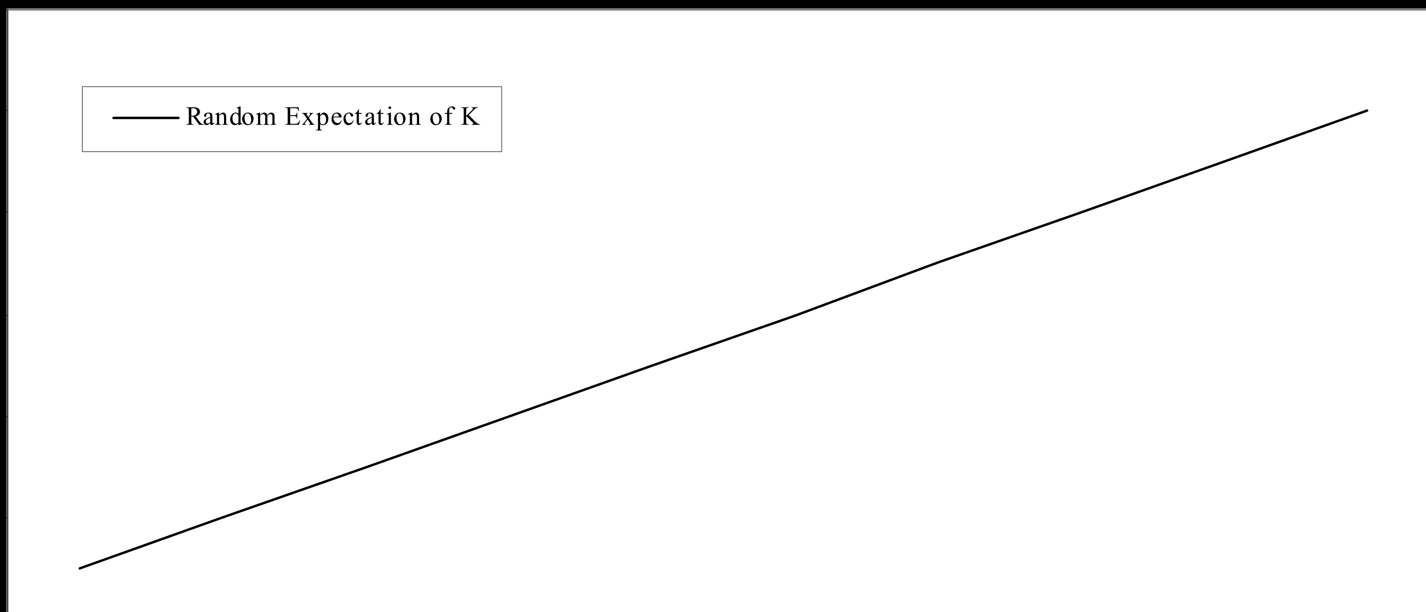
Cuzick and Edwards and Controls (k nearest neighbors)

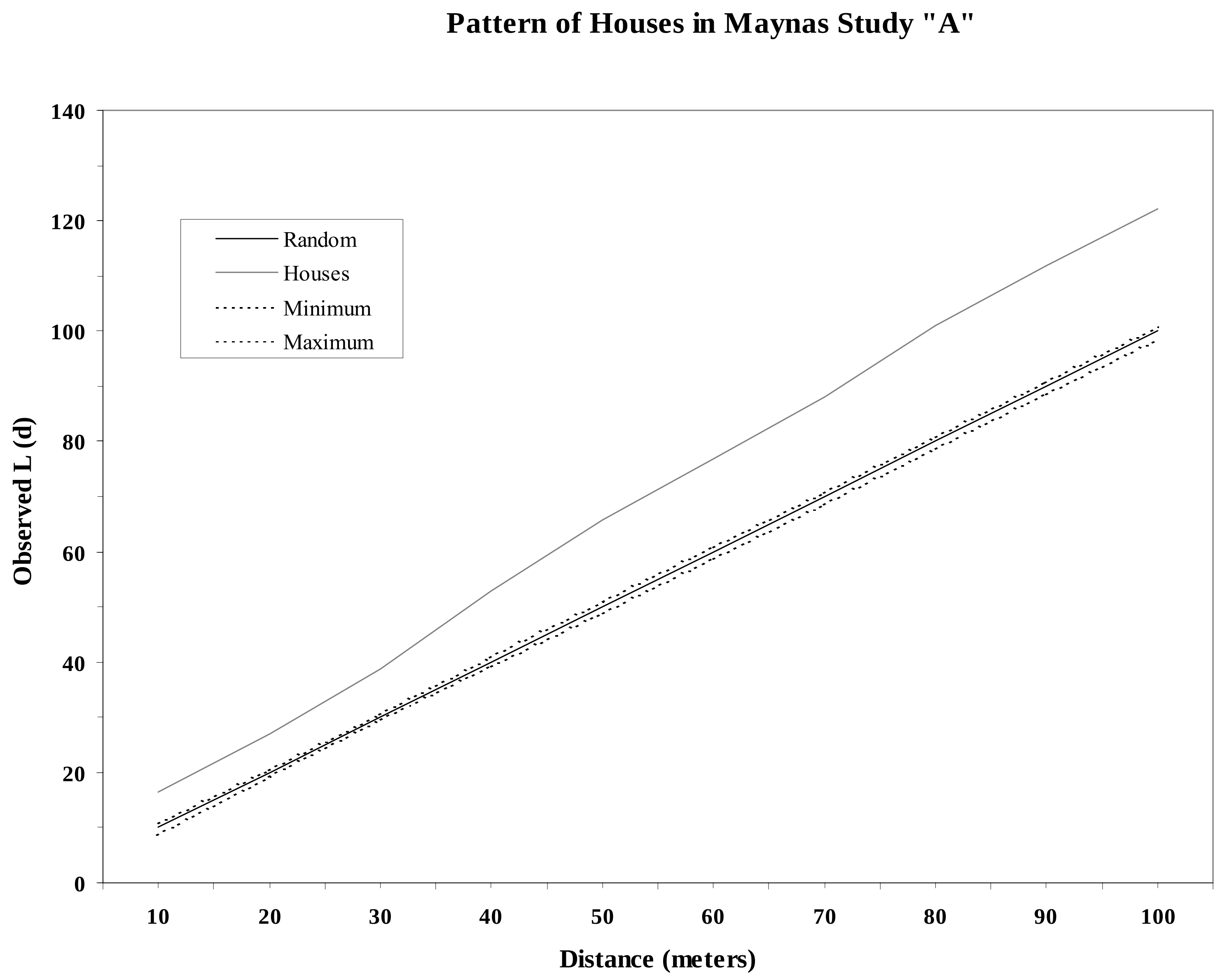
- A method for detecting spatial clustering for populations with non-uniform density.
- Label cases as x_i and controls as y_i
- Counts the number of cases (x_i) among the k nearest neighbors (x_i and y_i) to each case.
- Finds the theoretical distribution by permutation.
- Asymptotically normal. Provides test: the locations of the cases and controls follow a non-homogeneous Poisson process.

K Function Analysis: A Global Statistic

- $L(d) = \{(A[\Sigma\Sigma K\ (d_{ij})]/\ \pi N(N-1)\ \}^{½}$
- where K(d_{ij}) is the number of pairs of points within d of *i*, and *A* is the area of the region under study.
- Used to discern the clustering pattern of the specified variable within the entire study area.
- An output file gives a table showing *L* values for each distance *(d)* increment. *E[L(d)]* in a random distribution is *d*.
- Used to give access to controls.


0
0
0
0
0
0
0
0
0


Random Expectation of K


Pattern of Houses in Maynas Study "A"
Random
Houses
Minimum
Maximum
Observed L (d)
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
80
90
100
Distance (meters)

K-Function (second-order analysis) - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/4.cgi What's Related

Internet Lookup RealPlayer

K-Function (second-order analysis)

Your results:

```
The input data file: test.dat
The total number of points:  100
The minimum x coordinate: 1.000000
The maximum x coordinate: 99.000000
The minimum y coordinate: 0.000000
The maximum y coordinate: 98.000000
The total area: 9604.000000
The maximum search distance: 100.000000
The step size: 10.000000
The number of permutation for significance envelope:10
   Distance     Observed L(d)     Minimum L(d)  Maximum L(d)
   10.00000       10.13452        9.62932      10.74250
   20.00000       19.97004       19.41093      20.91819
   30.00000       29.04746       29.08554      31.51734
   40.00000       39.27837       39.09973      42.09946
   50.00000       50.12228       48.81093      52.01368
   60.00000       59.01874       58.35136      61.42986
   70.00000       66.99494       66.97278      69.04510
   80.00000       73.98763       73.91991      75.35072
   90.00000       79.12397       78.25648      80.07432
  100.00000       82.25774       79.72802      83.23661
```

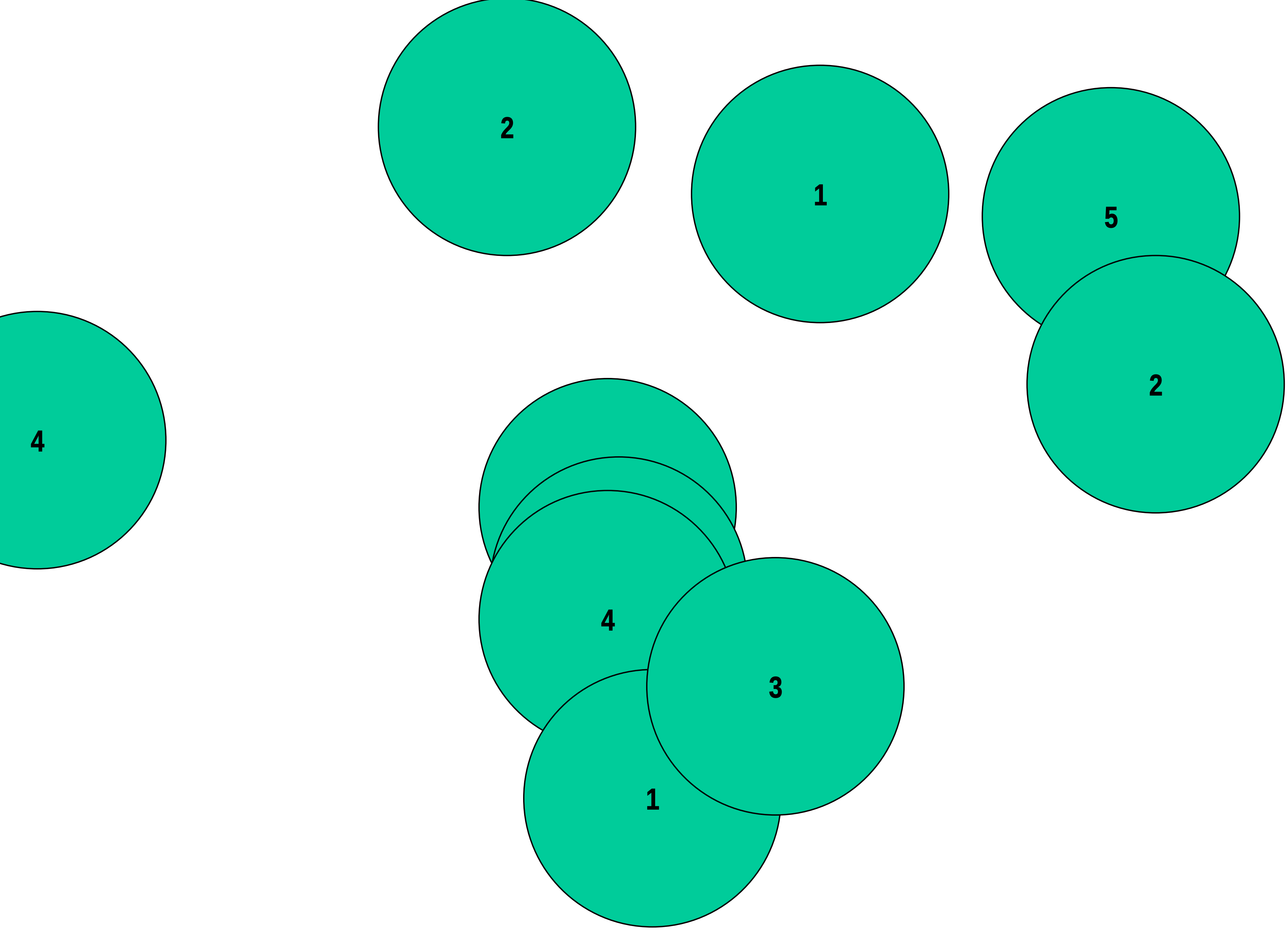
Report problems with this script to Andy Long.
cgi.tcl script creator: Don Libes, of NIST.
cgi.tcl script butcher: Andy Long, of BioMedware.

Powered by cgi.tcl

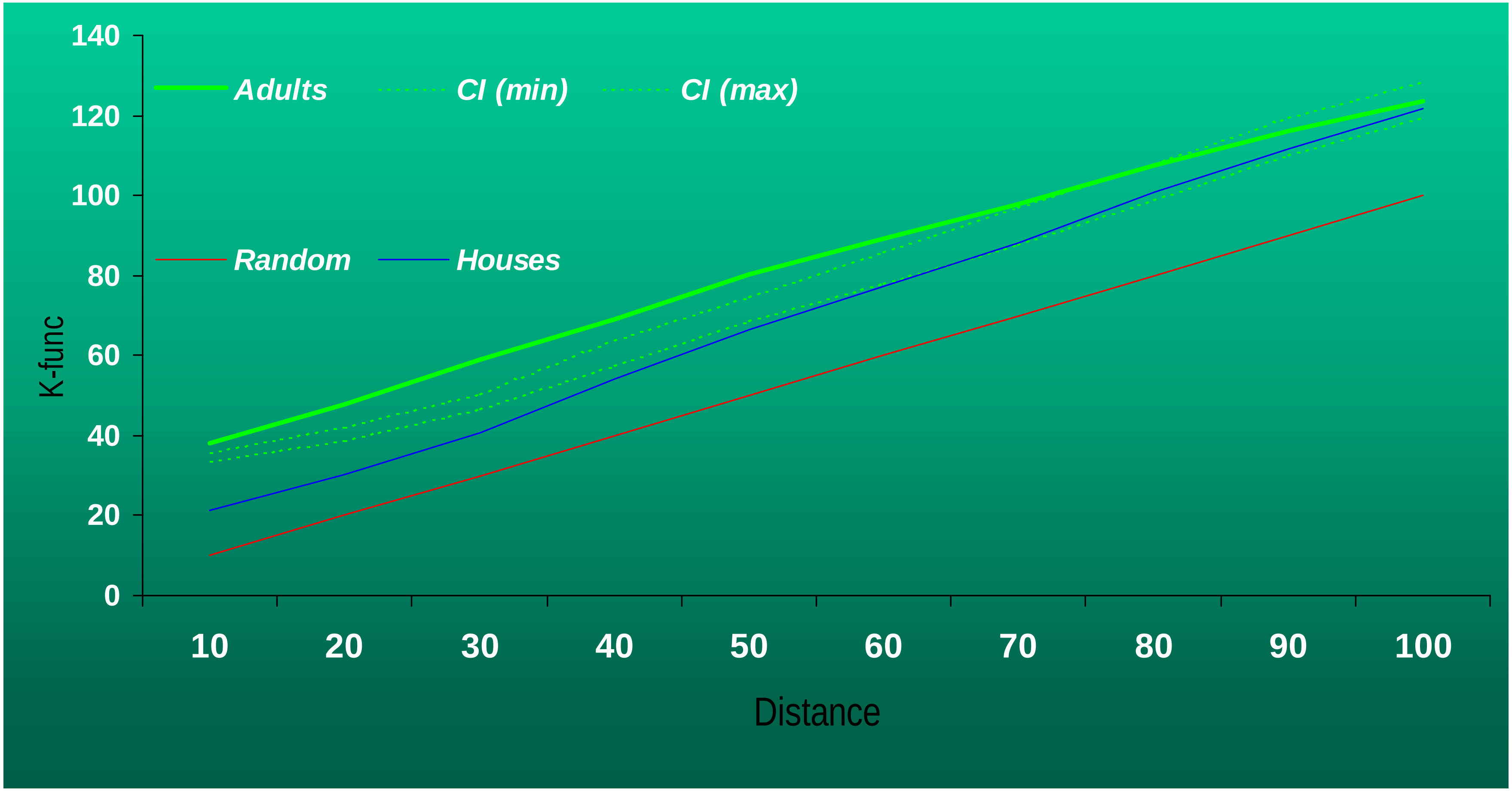
Document: Done

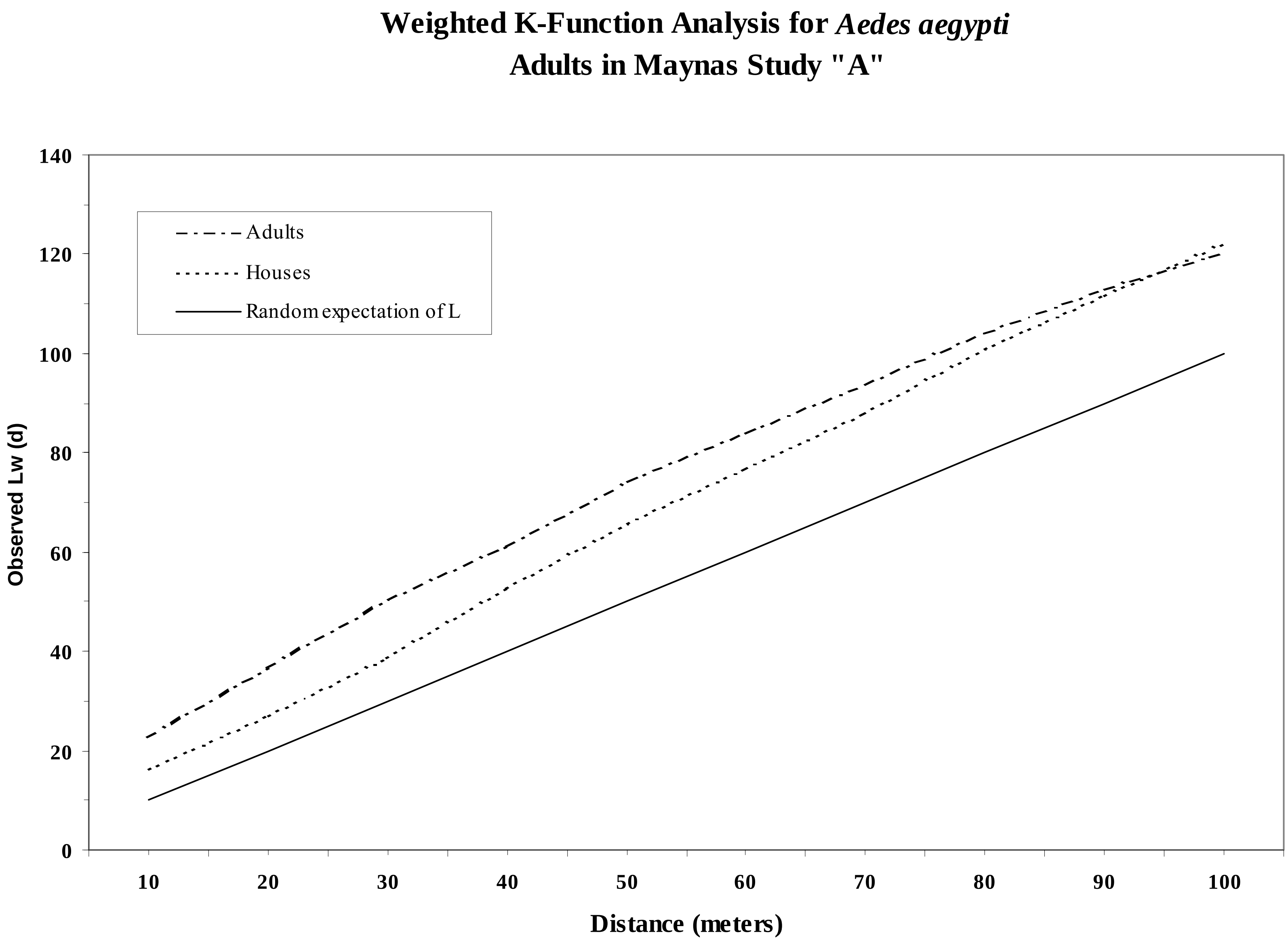
$$L_w(d)$$

$$L_w(d) = \left[\frac{\{ A \sum_i \sum_j u_{ij}^{-1} I_d(d_{ij} \leq d) x_i x_j \}}{\{ \pi [(\sum_i x_i)^2 - \sum_i x_i^2] \}} \right]^{1/2} \qquad i \neq j$$

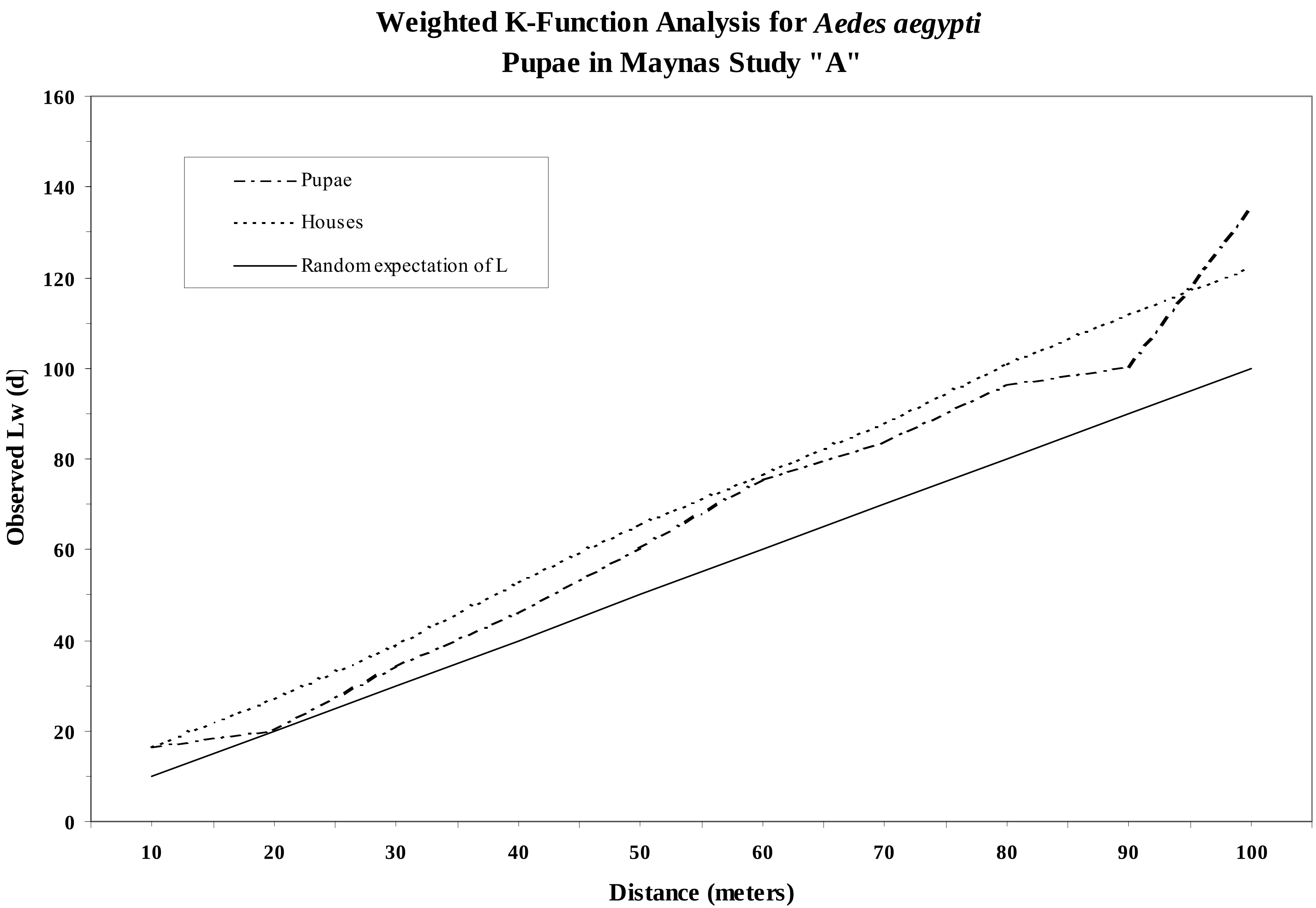

2
1
5
2
4
4
3
1

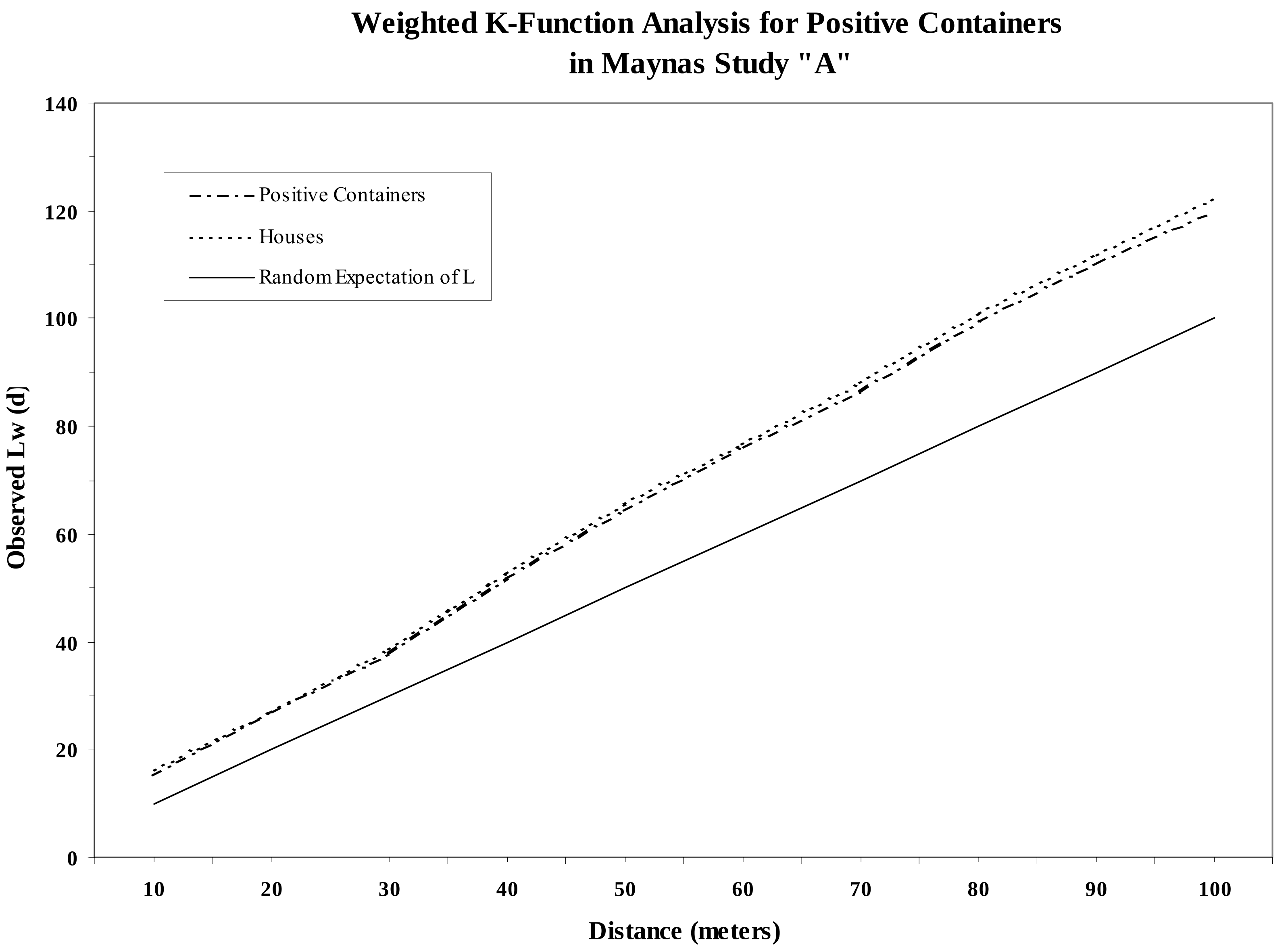
K-function for adult *Aedes aegypti*

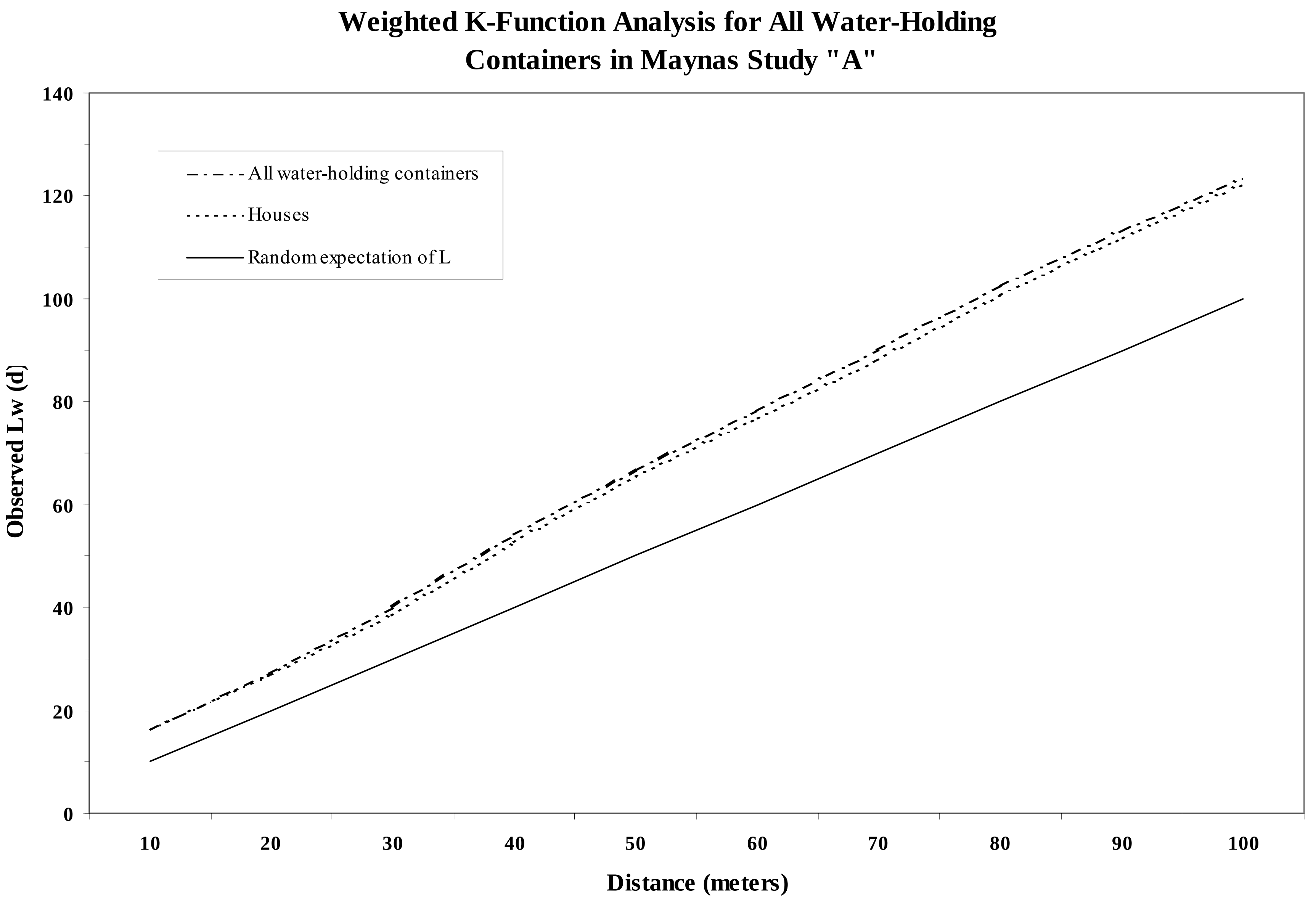




Weighted K-Function Analysis for *Aedes aegypti*
Adults in Maynas Study "A"
Adults
Houses
Random expectation of L
Observed Lw (d)
140
120
100
80
60
40
20
0
10
20
30
40
50
60
70
80
90
100
Distance (meters)


Weighted K-Function Analysis for *Aedes aegypti*
Pupae in Maynas Study "A"
Pupae
Houses
Random expectation of L
Observed Lw (d)
Distance (meters)
0
20
40
60
80
100
120
140
160
10
20
30
40
50
60
70
80
90
100


Weighted K-Function Analysis for Positive Containers
in Maynas Study "A"
Positive Containers
Houses
Random Expectation of L
Observed Lw (d)
Distance (meters)
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
80
90
100


Weighted K-Function Analysis for All Water-Holding Containers in Maynas Study "A"
All water-holding containers
Houses
Random expectation of L
Observed Lw (d)
Distance (meters)
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
80
90
100

POINT PATTERN ANALYSIS - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi What's Related

Internet Lookup RealPlayer

POINT PATTERN ANALYSIS

(Version 1.0a)

Developed by
Jared Aldstadt, DongMei Chen and Arthur Getis
San Diego State University

- Basic descriptive statistics
- Nearest neighbor analysis
- Refined nearest neighbor analysis
- K-Function (second-order analysis)
- Weighted K-Function analysis
- Cluster: Knox Time-space analysis
- Join-count statistic
- Global Moran's I
- Global Geary's c
- General Getis-Ord's G
- Local Ii statistic
- Local Gi statistic
- Local Gi* statistic
- Local K-Function

Submit Form

Stat-specific Documentation by Jared Aldstadt.
General PPA documentation by the authors.

Upload your own ASCII data for use in PPA

Please: no text in your files: just numbers, ASCII number!

- Text in the file (e.g. a word at the end of the file) causes ppa to choke, and leaves the process running on our machines. We then must kill these 'zombie' processes, as they bog down our system.
- At the moment, we kill ALL PPA jobs that are running as of 2:56 AM, Eastern time, to prevent these zombies from taking over the system and driving it into the ground. Try to avoid using the machine at this time!

Document: Done

POINT PATTERN ANALYSIS - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi What's Related

Internet Lookup RealPlayer

Stat-specific Documentation by Jared Aldstadt.
General PPA documentation by the authors.

Upload your own ASCII data for use in PPA

Please: no text in your files: just numbers, ASCII number!

- Text in the file (e.g. a word at the end of the file) causes ppa to choke, and leaves the process running on our machines. We then must kill these 'zombie' processes, as they bog down our system.
- At the moment, we kill ALL PPA jobs that are running as of 2:56 AM, Eastern time, to prevent these zombies from taking over the system and driving it into the ground. Try to avoid using the machine at this time!
- NB: you cannot submit Word documents, Excel documents, or any other binary format and expect useful results! Only ASCII (.txt) files, with the appropriate format, are allowed. You can check the file formats each test requires by selecting the test and examining the example files provided.

No Excel files!!! No Word files!!! E:\Crime\burglarydata [Browse...] These are BINARY formats, and WILL NOT WORK!	
Name for the file on the server (e.g. my.dat - do not use c:\...! - and a short (e.g. eight character) but *unique* name is better, as it is less likely to be overwritten by someone else):	trial.dat

Upload

(your file will be deleted from our system after one hour of inactivity).

You can submit multiple files by submitting them one-by-one (and giving them different names!).

Report problems with this script to Andy Long.
cgi.tcl script creator: Don Libes, of NIST.
cgi.tcl script butcher: Andy Long, of BioMedware.

Document: Done

POINT PATTERN ANALYSIS - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi What's Related

Internet Lookup RealPlayer

POINT PATTERN ANALYSIS

(Version 1.0a)

Developed by
Jared Aldstadt, DongMei Chen and Arthur Getis
San Diego State University

- ○ Basic descriptive statistics
- ○ Nearest neighbor analysis
- ○ Refined nearest neighbor analysis
- ○ K-Function (second-order analysis)
- ◉ Weighted K-Function analysis
- ○ Cluster: Knox Time-space analysis
- ○ Join-count statistic
- ○ Global Moran's I
- ○ Global Geary's c
- ○ General Getis-Ord's G
- ○ Local Ii statistic
- ○ Local Gi statistic
- ○ Local Gi* statistic
- ○ Local K-Function

Submit Form

Stat-specific Documentation by Jared Aldstadt.
General PPA documentation by the authors.

Upload your own ASCII data for use in PPA

Please: no text in your files: just numbers, ASCII number!

- Text in the file (e.g. a word at the end of the file) causes ppa to choke, and leaves the process running on our machines. We then must kill these 'zombie' processes, as they bog down our system.
- At the moment, we kill ALL PPA jobs that are running as of 2:56 AM, Eastern time, to prevent these zombies from taking over the system and driving it into the ground. Try to avoid using the machine at this time!

Document: Done

K-Function

K-function is also called second-order analysis to indicate that the focus is on the variance, or second moment, of pairs of interevent distances. It considers all combinations of pairs of points. It compares the number of observed pairs with the expectation at all distances based on a random spatial distribution of points. The density of points, the borders, and the size of the sample are taken into consideration.

Input

1. The input data file, which should contain N rows of X, Y coordinates, and W values (a column of 1s).
2. The maximum distance that you want to use. The statistically unbiased maximum distance is less than the circumradius of the study area, or one-half of the length of the shortest side of a rectangular study area.
3. The number of increments.
4. The number of permutations for creating the confidence envelope.
5. The output file.

Analysis

K-function analysis is a test of the hypothesis of CSR. The expected value of L(d) is d. The confidence interval in this analysis is generated by examining the specified number of permutations of randomly generated patterns of N points over the whole study area. If for any distance, the observed L(d) falls above or below the expected L(d) the null hypothesis of CSR can be rejected at an appropriate level of significance. The level of significance is determined by the confidence envelope. An observed L(d) below the envelope indicates that the points are dispersed at that distance, whereas an observed above the envelope indicates that clustering is present at that distance.

Formula

$$L(d) = \sqrt{\frac{A\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} k(i,j)}{\pi N(N-1)}}$$ [1]

where:

A is the study area,

N is the number of points

Formula

$$L(d) = \sqrt{\frac{A\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} k(i,j)}{\pi N(N-1)}} \quad [1]$$

where:

A is the study area,

N is the number of points

d is the distance

$\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} k(i,j)$ is the number of j points within distance d of all i points

$k(i,j)$ is the weight, which is estimated by

a) If no edge corrections,

$k(i,j) = 1$ in case $d(i,j) \leq d$

$k(i,j) = 0$ otherwise

b) If a point i is closer to one boundary than it is to point j, the border correction is employed

$$k(i,j) = \left[1 - \frac{\cos^{-1}\frac{e}{d(i,j)}}{\pi}\right]^{-1} \quad [2]$$

where e is the distance to the nearest edge.

Weighted K-Function analysis - Netscape

Location: http://xerxes.sph.umich.edu:2000/cgi-bin/cgi-tcl-examples/generic/ppa/ppa-result.cgi

Weighted K-Function analysis

Required Information:

Data Filename: trial.dat

maximum distance: 100

number of increments: 10

Number of permutations (for the confidence envelope): 99

Sample Data:	test.dat
Your data should have the same format!	three columns: x,y,z

Submit Form **This may take awhile, so be patient!**
The browser may appear to have stalled, but should eventually show the results (provided your input was valid).

Report problems with this script to Andy Long.
cgi.tcl script creator: Don Libes, of NIST.
cgi.tcl script butcher: Andy Long, of BioMedware.

Powered by cgi.tcl

Document: Done

Pattern Statistics

- GENERAL

  I, c, K, G, Knox, Mantel, Tango, Grimson, Cuzick and Edwards, Kernels, Scan

- FOCUSED

  I_i, c_i, G_i, G_i*, GWR, O_i

Global Statistics

- Nearest Neighbor
- K-Function
- Global Autocorrelation Statistics
  - Moran's I
  - Geary's c
  - Semivariance

WY :Covariance

- Set **W** to preferred spatial weights matrix
- Set **Y** to
- $(x_i - \mu)(y_i - \mu)$
- Set scale to $n/W\ \Sigma(x_i - \mu)^2$
- I = $n\ \Sigma\Sigma W_{ij} (x_i - \mu)(y_i - \mu) / W\ \Sigma(x_i - \mu)^2$

  where W is sum of all W_{ij}

WY : Difference

- Set **W** to preferred spatial weights matrix
- Set **Y** to
- $(x_i - y_i)^2$
- Set scale to $(n-1)/2W\Sigma(x_i - \mu)^2$
- c = $(n - 1) \Sigma \Sigma W_{ij} (x_i - y_i)^2 / 2W\Sigma(x_i - \mu)^2$

  where W is sum of all W_{ij}

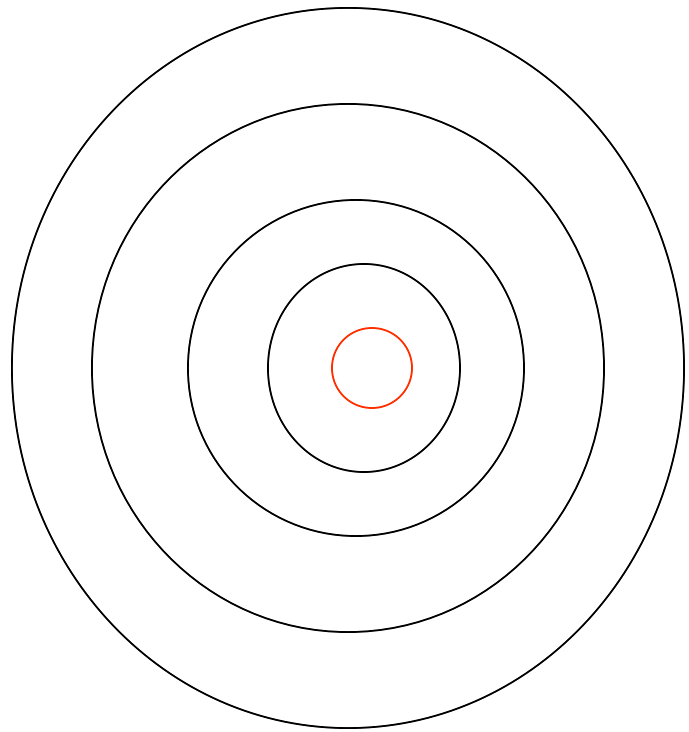
Local Statistic

$$G_i^*(d) = \frac{[\sum_j w_{ij}(d)x_j - W_j^* \bar{x}]}{s\{[NS_{1j}^* - W_j^{2*}]/(N-1)\}^{1/2}} \qquad all\ j$$

$w_{ij}(d)$ is element of 1/0 spatial weights matrix where 1 within d of i, 0 otherwise

$$W_j^* = \sum w_{ij}(d)$$

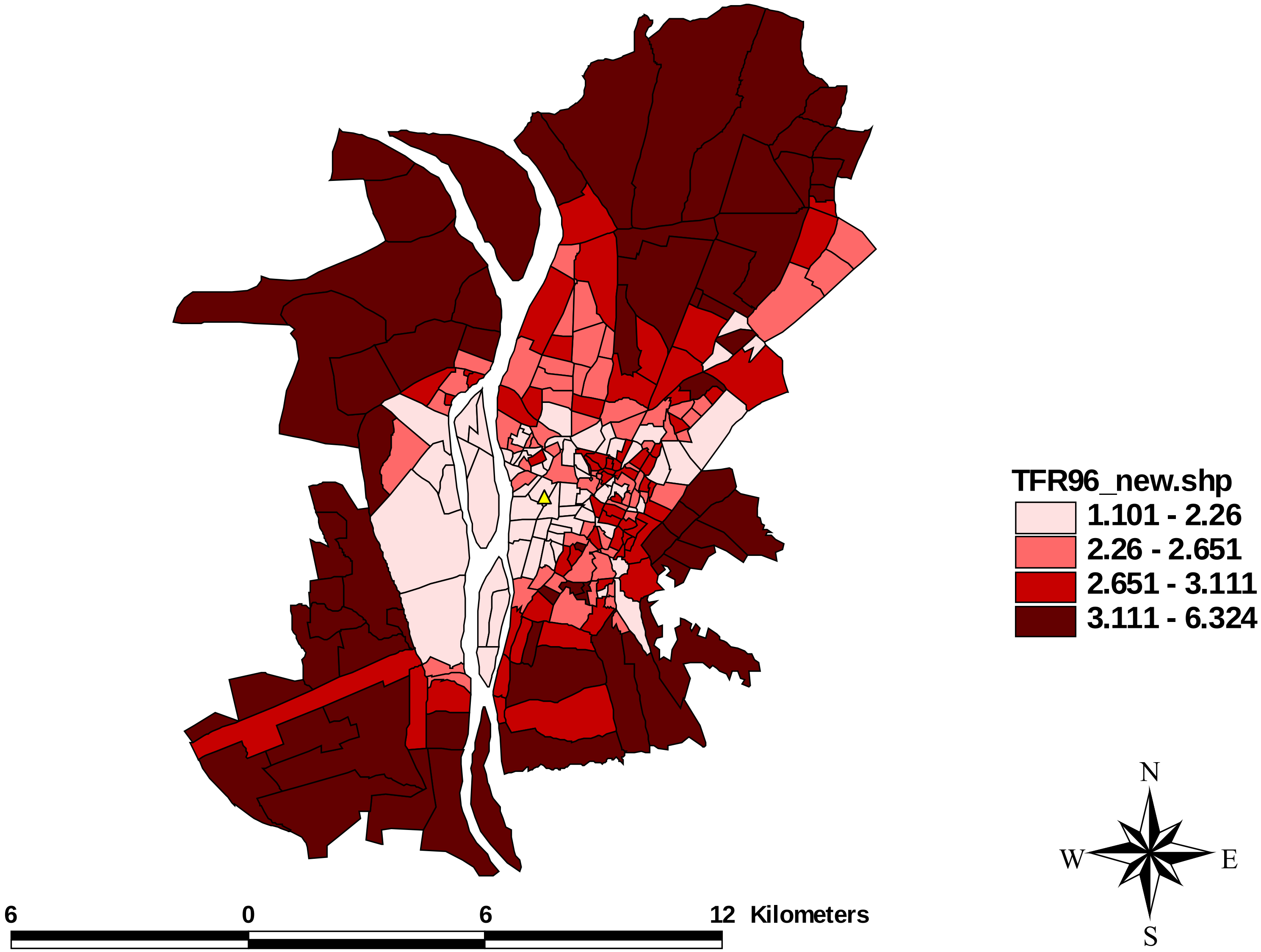
$$S_{1j}^* = \sum_j w_{ij}^2 \quad (all\ j)$$



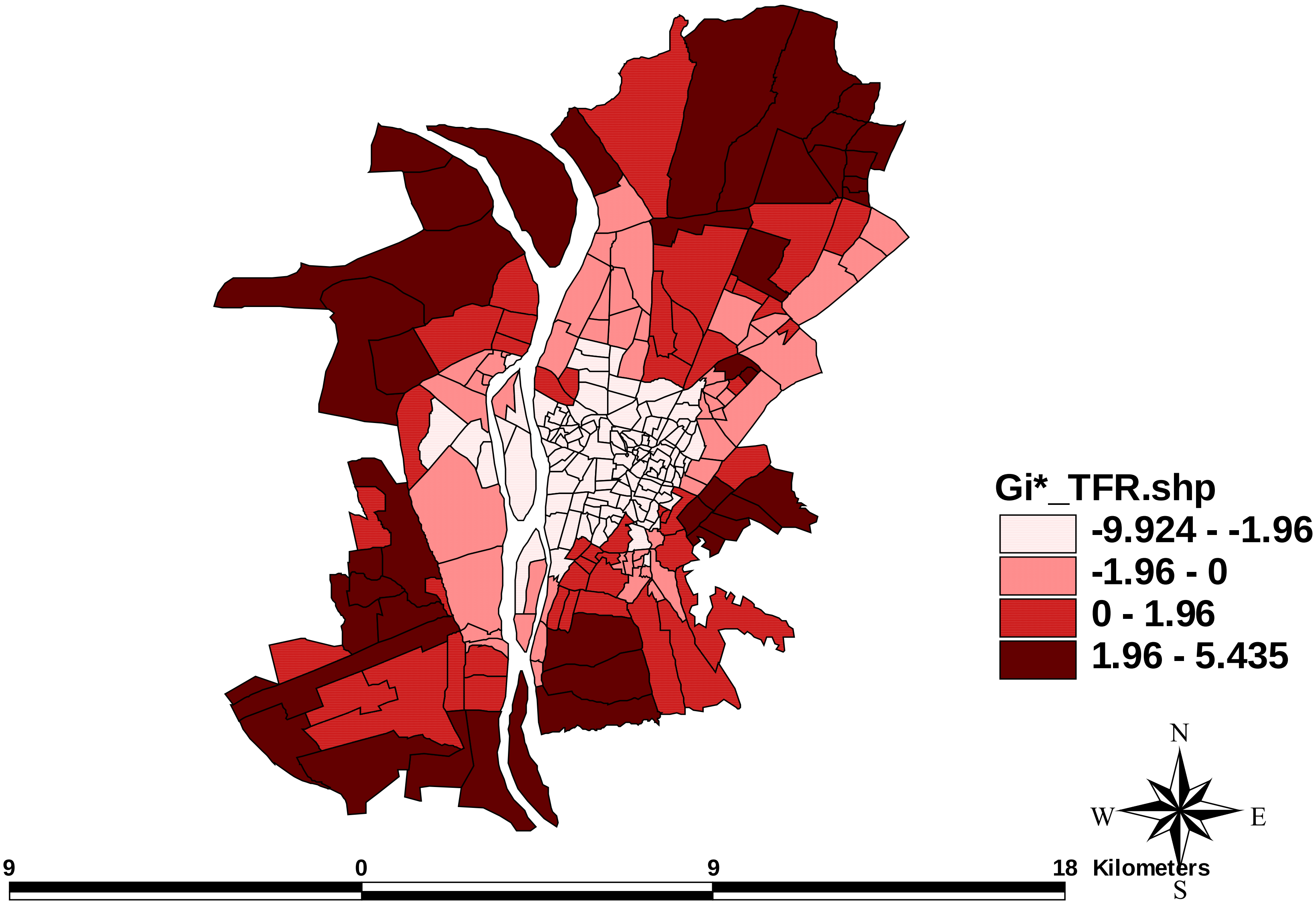
The G_i^* Statistic

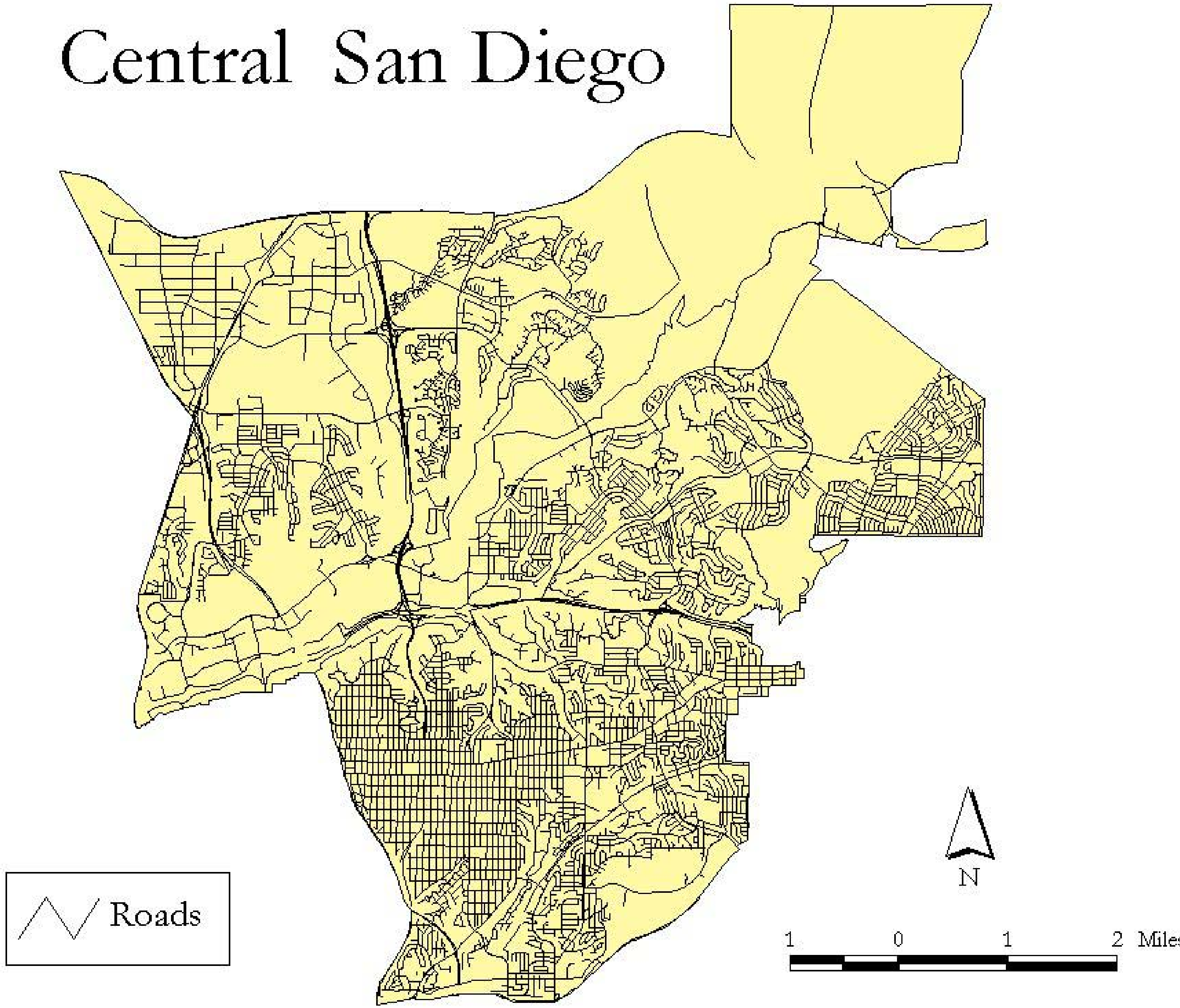
- The G_i^* statistic is local, that is, it is focused on sites and is normally distributed. It is designed to yield a measure of pattern in standard normal variates.
- Indicates the extent to which a location (site) is surrounded to a distance *d* by a cluster of high or low values (in this case, we focus on high values).
- The input is a file containing coordinates for each house and, for example, the number of adult *Aedes aegypti*. User specifies maximum search distance (100 meters in this case) and number of increments (10 10-meter increments).
- The output file contains a listing of the $G_i^*(d)$ value for each house at a specified distance *(d)*.

Talaat Harb Square

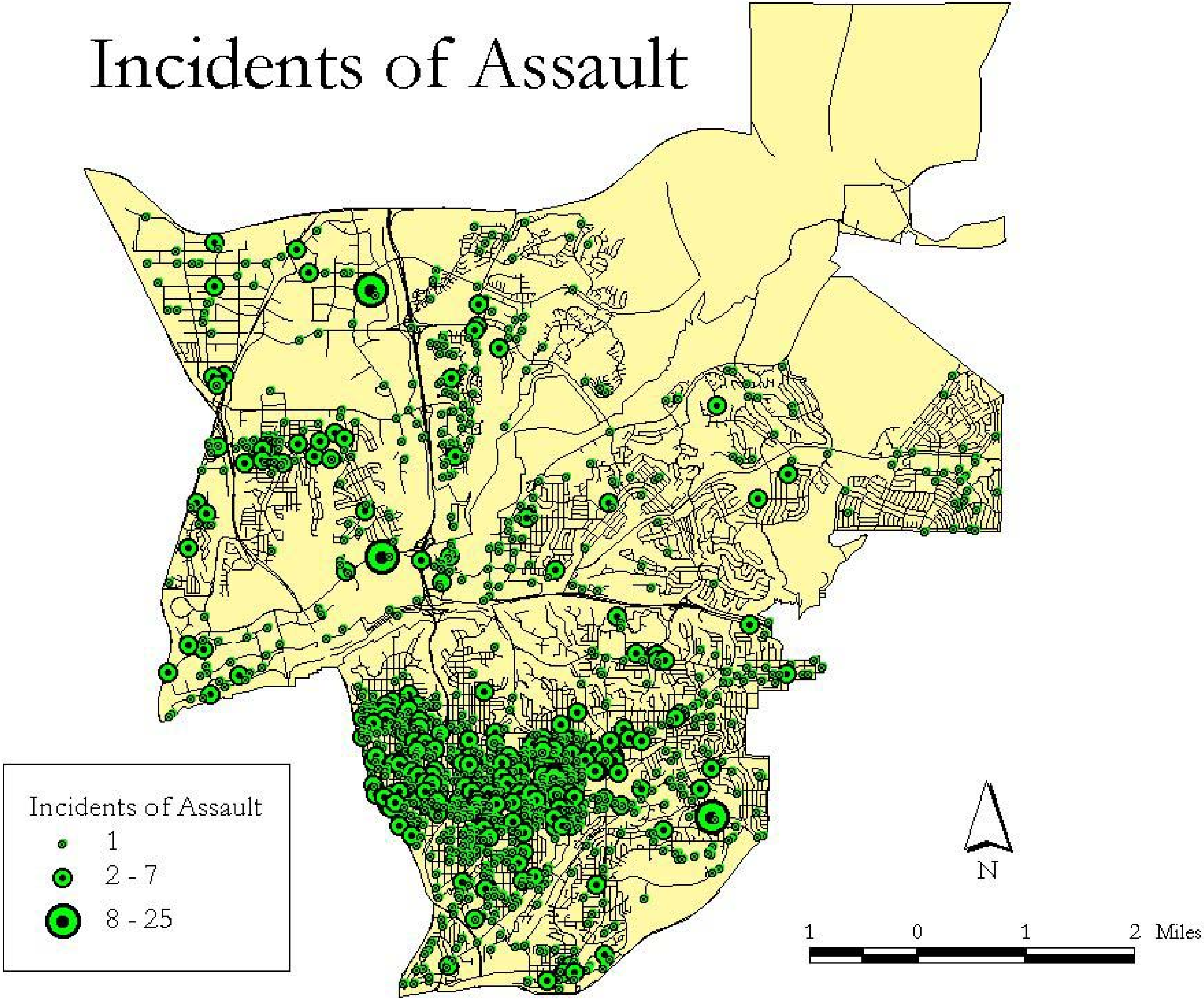


Gi* statistic

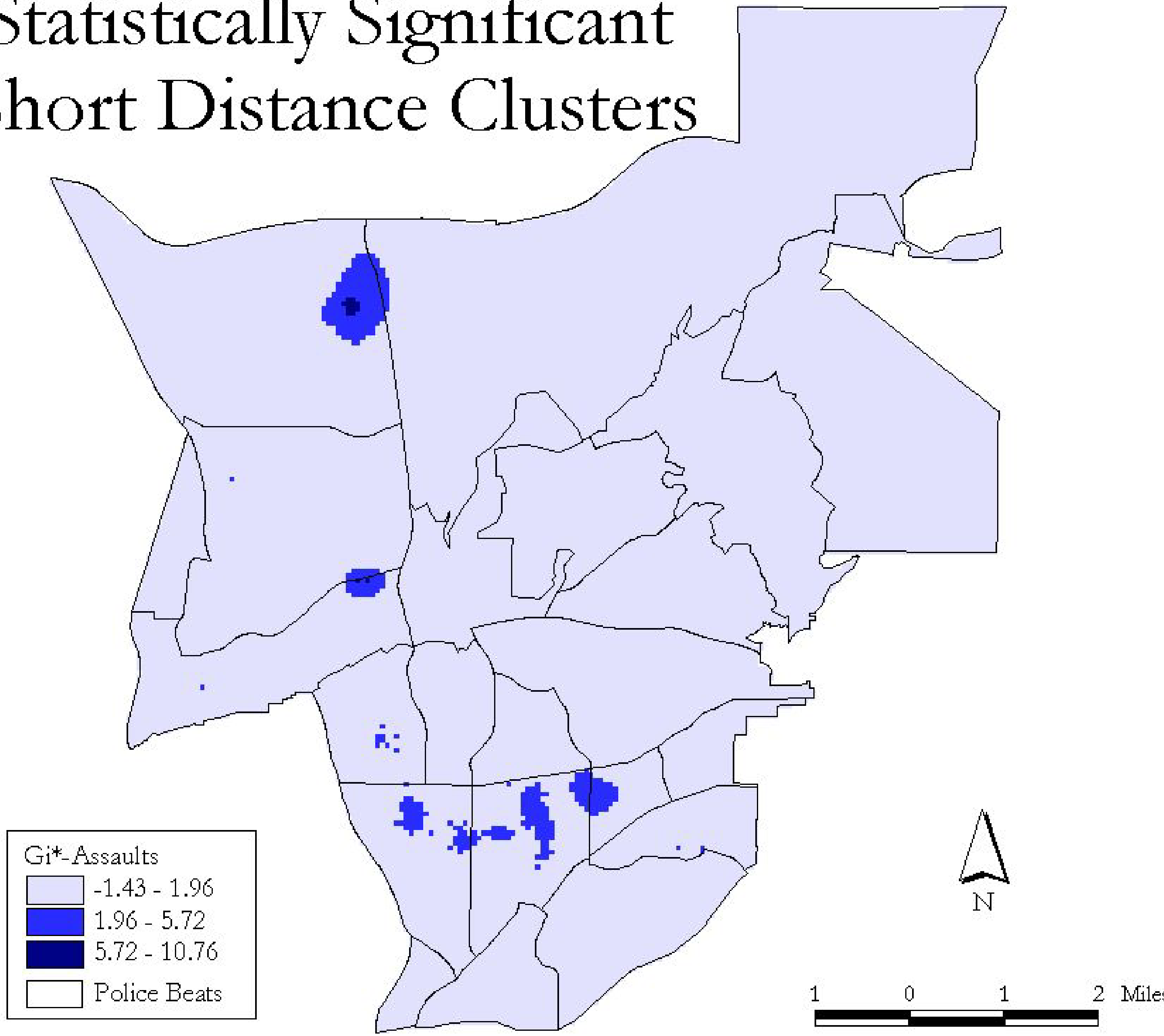



Central San Diego
Roads
N
1 0 1 2 Miles

Incidents of Assault



Statistically Significant Short Distance Clusters



Crime Clustering Packages

- STAC (Spatial and Temporal Analysis of Crime)
- GAM (Geographical Analysis Machine)
- SaTScan (Spatial and Space-Time Scan Statistic)
- CrimeStat (Crime Mapping Research Center)

Typology of Crime Mapping Applications

(after Craglia, Haining, and Wiles)

Application	Data	Scale	Function
Dispatching	Seconds/ Minutes	Site	Visualization
Community policing	Hours/days	Neighborhood	Mapping
Resource planning	Weeks/ months/years	City	Analysis/ modeling

Problems

- The problem of ellipses
- Smoothing effects
- Hierarchical scales
- Incidents and not risk

Problems with Local Stats

- Global heterogeneity
- Multiple Tests
- Dependent Tests

The O_i Statistic

- The null hypothesis of local autocorrelation
- A recommended partition

Multiple Dependent Tests

- Overlap
- Seemingly independent tests
- Virtual v

Seemingly Independent Tests:

Addressing the Problem of Multiple Simultaneous and Dependent Tests

Appropriate Inferential Bounds?

- Multiple Tests
- Simultaneous Tests
- Dependent Tests

Sidak: $1 - \alpha = (1 - p)^k$

Multiple tests but not dependent tests.

Bonferroni: α/k

Multiple tests but not dependent tests.

Virtual v

- $$K = mv$$

- where v is number of independent clusters,
- and m is number of observations within
- each cluster

Correlation Between Tests = r

- $v = k - r(k-1)$
- and $1 - \alpha = (1 - p)^v$
- when $r=1$, $v=1$; when $r=0$, $v=k$
- lower bound for r: $-1/(k-1)$
- highest possible v: $k+1$

Number of Tests with Dependence

- Possible Clustering of *aedes aegypti* in Mynas Section of Iquitos
- Houses = 543 = k
- Set d=10 meters
- Overlap (r estimated at 0.500)
- $v = k-r(k-1) = 271$
- for .95 level; $p^v = (.989007)^{271} = 0.0500$
- $Z = 2.290931$

Data Mining

- Extension of EDA
- Inductive methods
- Substance versus significance
- Selection biases
- Process

Geostatistics

- Semivariance and the Semivariogram
- Kriging

Semivariance

- A measure of the degree of spatial dependence between observations of a regionalized variable.
- Formulation

$$\gamma_h = \Sigma\ (x_i - x_{i+h})^2 / \ 2n$$

where h is the distance interval between points.

The plot for a number of h's is called the semivariogram.

Characteristics of Semivariogram

- Range
- Sill
- Nugget
- Autocorrelation
- Variance = Sill

Semivariograms

- OBSERVED
- THEORETICAL
  - Spherical
  - Exponential
  - Linear (with sill)
  - Gaussian

Intrinsic Stationarity

Variogram analysis cannot proceed without acceptable assumptions, chief of which is *intrinsic stationarity*.

Kriging

- The Idea of Kriging

- Models
    - Simple (punctual)
    - Ordinary (punctual)
    - Universal (punctual)
    - Block
    - Cokriging
    - Others

Simple Kriging

- $Z(x_0) = m + \mathbf{YW^{-1}B}$
- where m = assumed mean (known)
- **Y** = observations in the vicinity of x_0 (-m)
- **W** = correlation - semivariance (for all pairs of observations)
- **B** = correlation - semivariance (for all pairs between observations and x_0)

Ordinary Kriging

- $Z(x_0) = \mathbf{Y}\mathbf{W}^{-1}\mathbf{B}$

Universal Kriging

- Drift

Block Kriging

- Areas or volumes

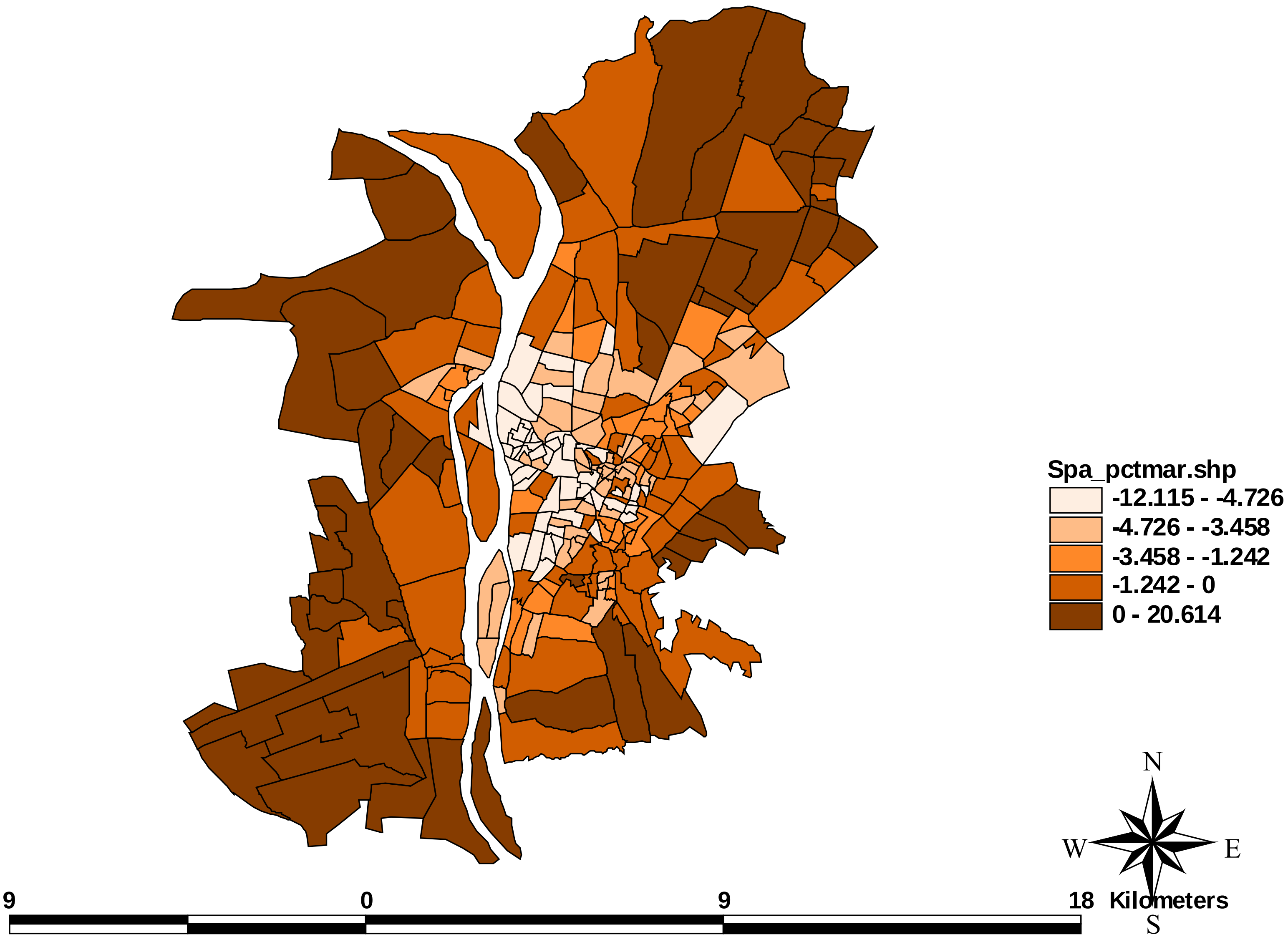
Cokriging

- More than one variable used to estimate value at a particular location.

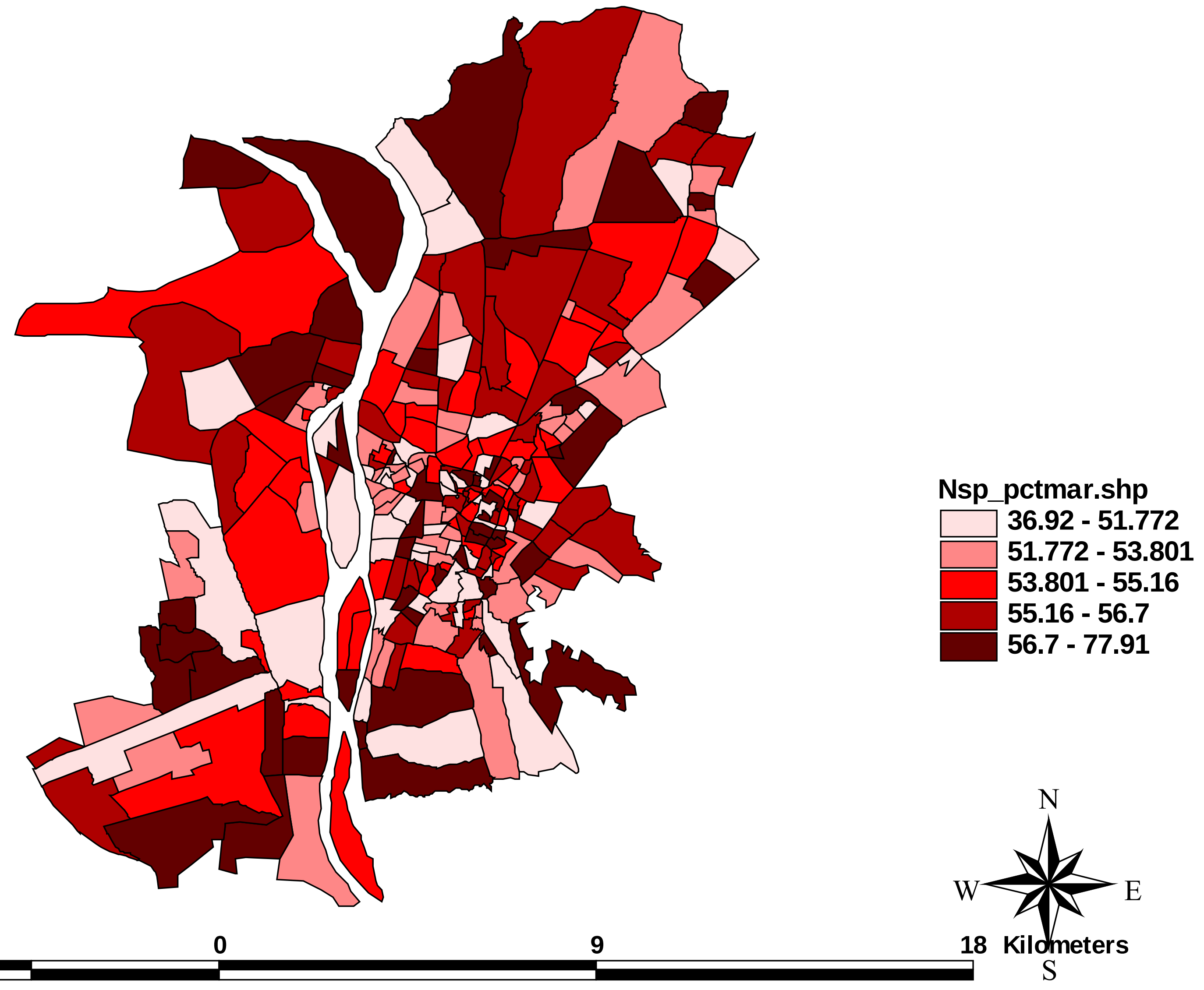
Spatial Filtering

- To find the degree to which each social variable is affected by spatial autocorrelation.
- Separate the spatial effects from the non-spatial effects.
- Develop spatial and non-spatial variables.
- Use Getis filtering approach (G_i* statistic).

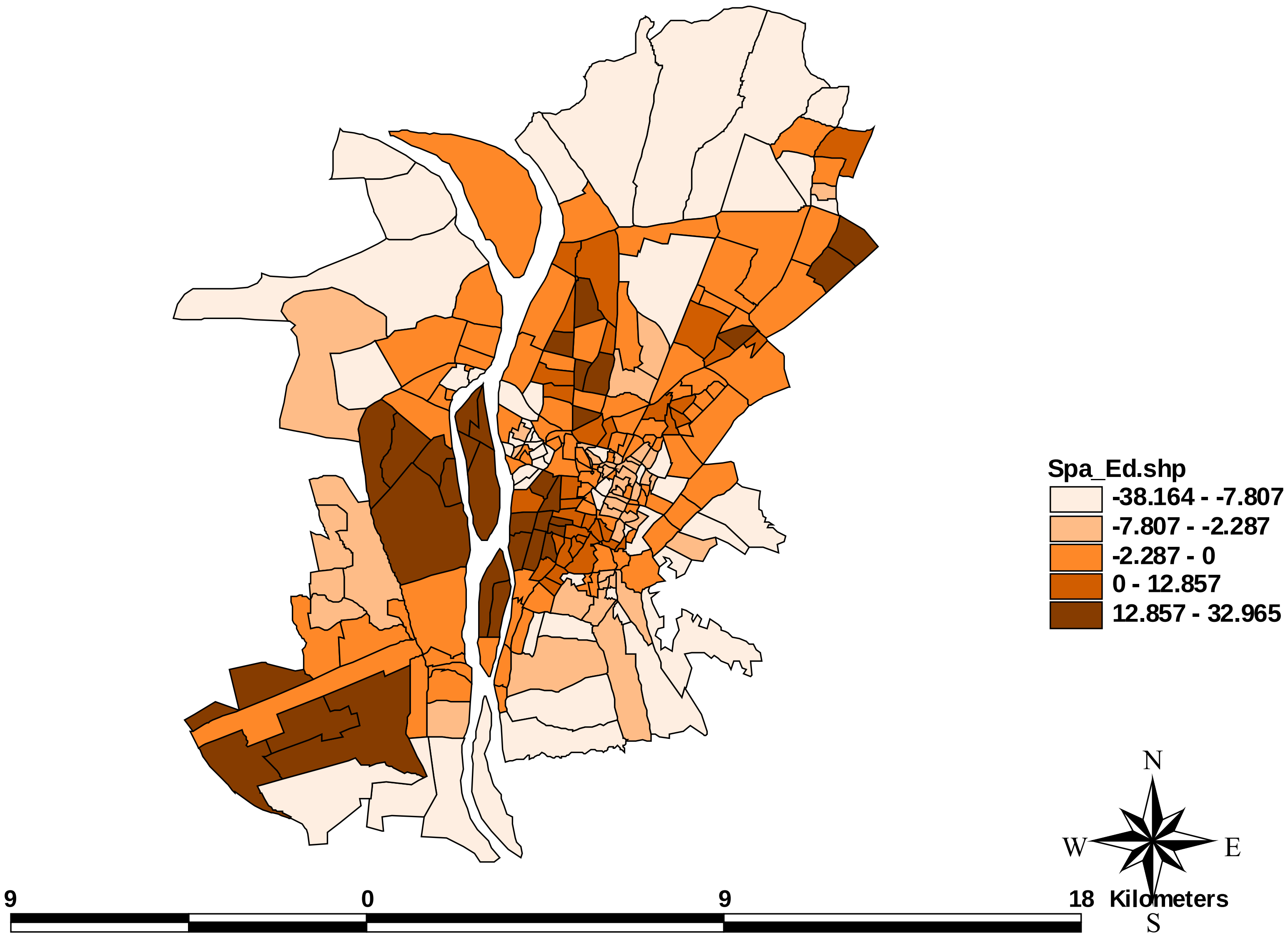
Spatial Component of Percent of women married



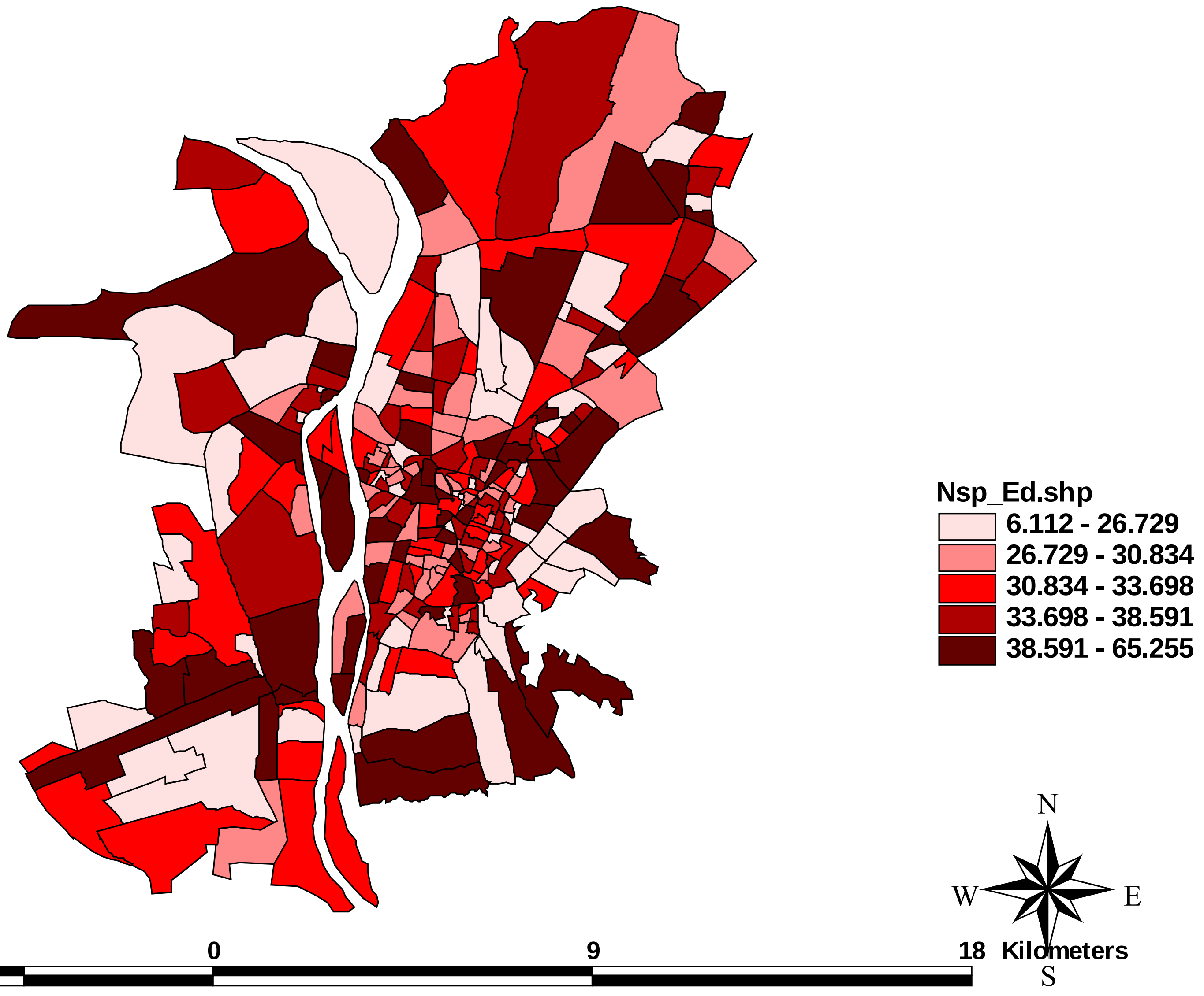
Non-Spatial Component of Percent of women married



Spatial Component of F96Ed_in



Non-Spatial Component of F96Ed_in



Large Dataset Problems

- heterogeneity
- partitions
- outliers
- missing data
- single fit
- multicollinearity
- data integration

A Local Variogram (1)

- We may define the effective range of the data as the distance d_r at which the variogram flattens out.

- At any distance less than d_r the correlation between any two pixels is greater than 0. The correlation between pixels d_r apart or greater is 0.

A Local Variogram (2)

- In this approach, we center on a pixel, i, and define the effective continuous region as that which contains:

1. pixels that are correlated with one another.
2. a discernible range, d_{ri} .

A Local Variogram (3)

- The number of pixels within the entire dataset is N and the number within the local variogram is M_i.
- M_i represents the partition.
- M may vary for each i^{th} pixel.
- If d_{ri} cannot be found for an i^{th} pixel, any analysis would have to be rethought for that partition.

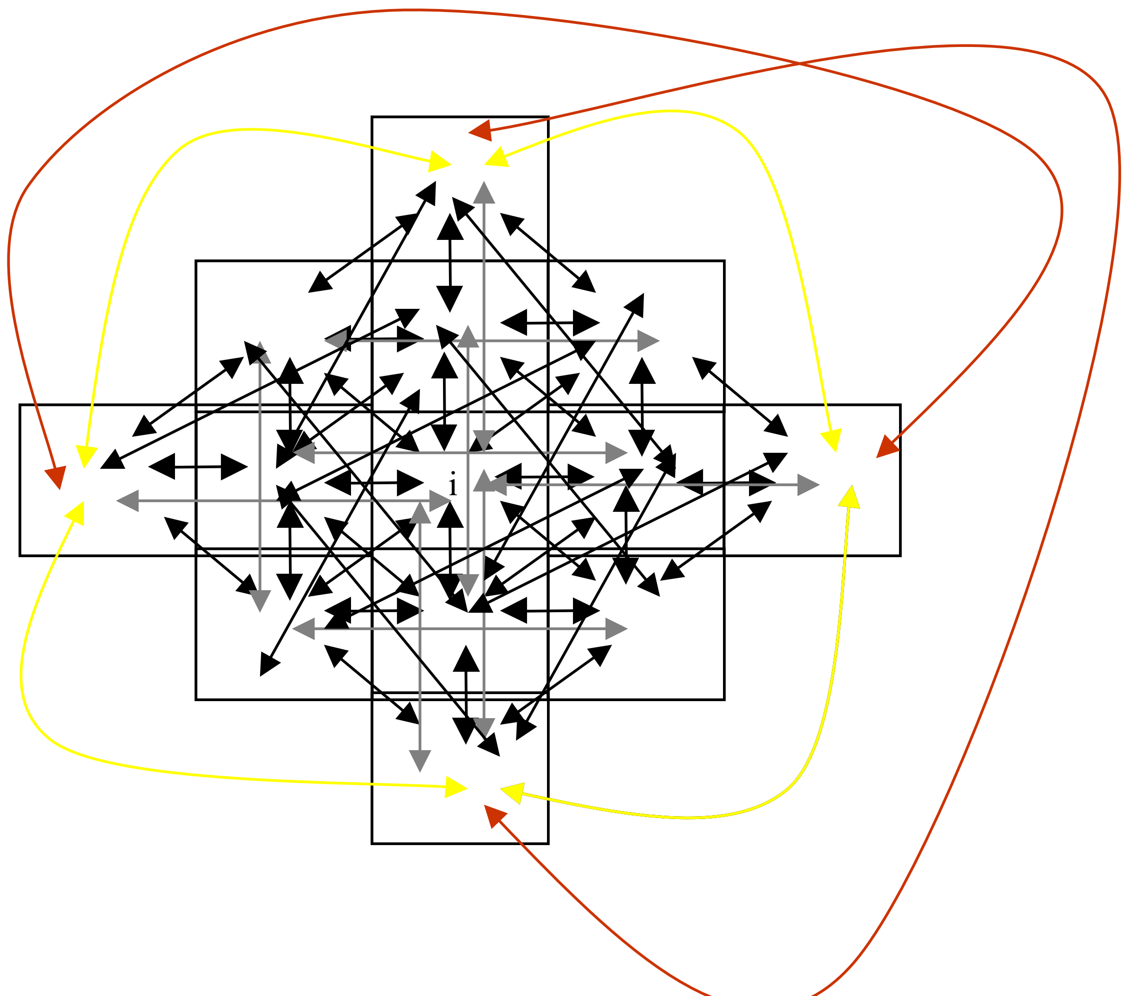
Finding d_{ri} and, therefore, M_i

$$Q_i = \Sigma\Sigma\, \rho(\mathbf{u}_j\text{-}\mathbf{v}_k) = \Sigma\Sigma\, C_{jk}\,(j,k) \qquad j,k \in M_i$$

- which sums the correlations between each pair of members within a trial partition starting with d = 1, then d = 2, and so on.
- when Q_i fails to increase after d is increased, d_{ri} and, therefore, M_i is reached.

Finding the Sum of the ρ's: $d = 1$

Finding the Sum of the ρ's: $d = 2$



Sum of the ρ's

- Autocorrelations (r) are estimated from the *M* observations as opposed to fitting a variogram model.

- Variogram models (spherical, exponential, linear, Gauss) may be poor descriptors of the spatial autocorrelation in the dataset.

The Number of Correlations for d=2

Distance	Number of correlations
1	16
$\sqrt{2}$	16
2	10
$\sqrt{5}$	16
$\sqrt{8}$	6
3	4
$\sqrt{10}$	8
4	2

Example

For the values of r, d^2 is in parentheses:

$r(1)$=0.236 $r(2)$=0.246 $r(4)$=0.215 $r(5)$=0.197 $r(8)$=0.184

$r(9)$=0.158 $r(10)$=0.166 $r(13)$=0.116 $r(16)$=0.106 $r(17)$=0.090

$r(18)$=0.053 $r(20)$=0.062 $r(25)$=0.058 $r(26)$=0.037 $r(29)$=0.036

$r(32)$=0.026 $r(34)$=0.021 $r(36)$=0.028 $r(37)$=0.016

$r(40\ to\ 169)$=0

Q = 1168.392

d_{ri} = 6.325

M_i = 129

Partition Routine