Lesson 7: Global and Local Measures of Spatial Autocorrelation

Dr. Kam Tin Seong
Assoc. Professor of Information Systems (Practice)

School of Computing and Information Systems,
Singapore Management University

2020-5-5 (updated: 2021-09-18)

Content

- What is Spatial Autocorrelation
  - Measures of Global Spatial Autocorrelation
  - Measures of Global High/Low Clustering
- Introducing Localised Geospatial Analysis
  - Local Indicators of Spatial Association (LISA)
- Cluster and Outlier Analysis
  - Local Moran and Local Geary
  - Moran scatterplot
  - LISA Cluster Map
- Hot Spot and Cold Spot Areas Analysis
  - Getis and Ord’s G-statistics
- Case Studies

What is Spatial Autocorrelation

- Toble's First Law of Geography
- Spatial Dependency
- Spatial Autocorrelation
  - Positive autocorrelation
  - Negative autocorrelation

Tobler's First law of Geography

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

The foundation of the fundamental concepts of:

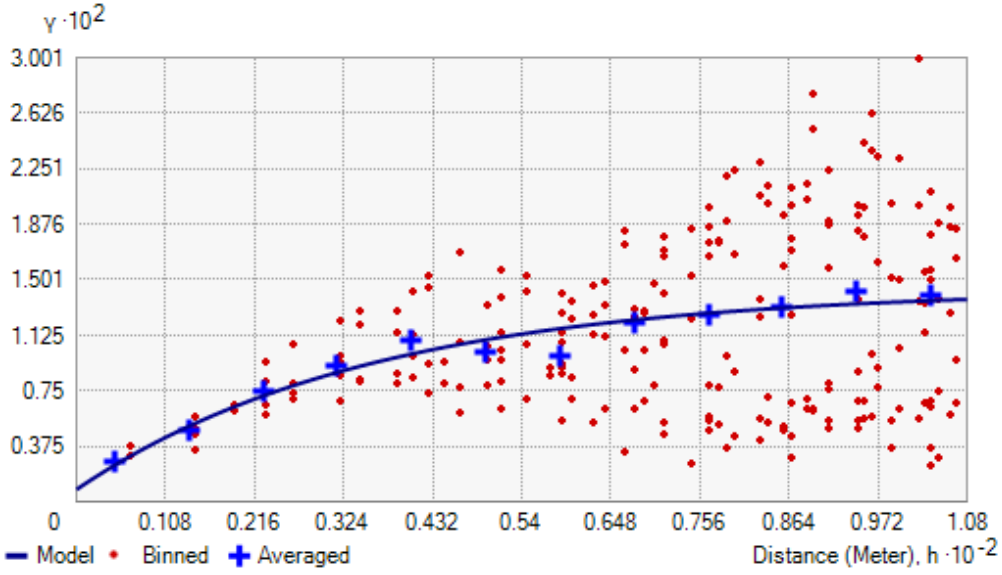
- spatial dependence, and
- spatial autocorrelation

Reference: A Computer Movie Simulating Urban Growth in the Detroit Region

Spatial Dependency

- Spatial dependence is the spatial relationship of variable values (for themes defined over space, such as rainfall) or locations (for themes defined as objects, such as cities).

- Spatial dependence is measured as the existence of statistical dependence in a collection of random variables, each of which is associated with a different geographical location.



Spatial Autocorrelation

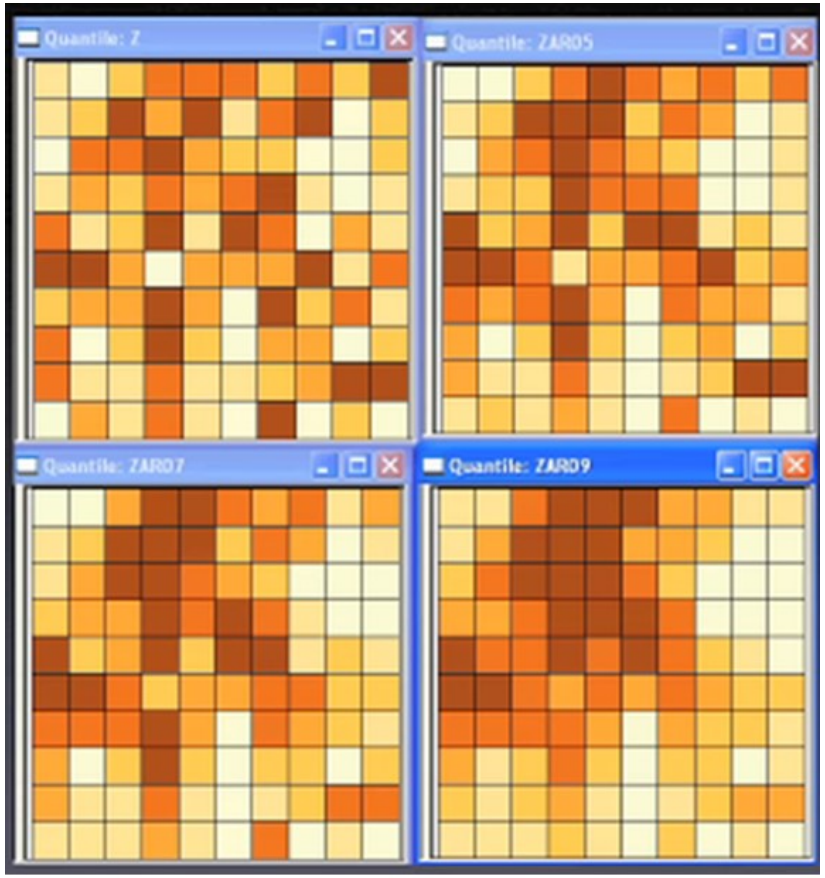
- Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable.

- The variable can assume values either:

  - at any point on a continuous surface (such as land use type or annual precipitation levels in a region);
  - at a set of fixed sites located within a region (such as prices at a set of retail outlets); or
  - across a set of areas that subdivide a region (such as the count or proportion of households with two or more cars in a set of Census tracts that divide an urban region).

Positive Spatial Autocorrelation

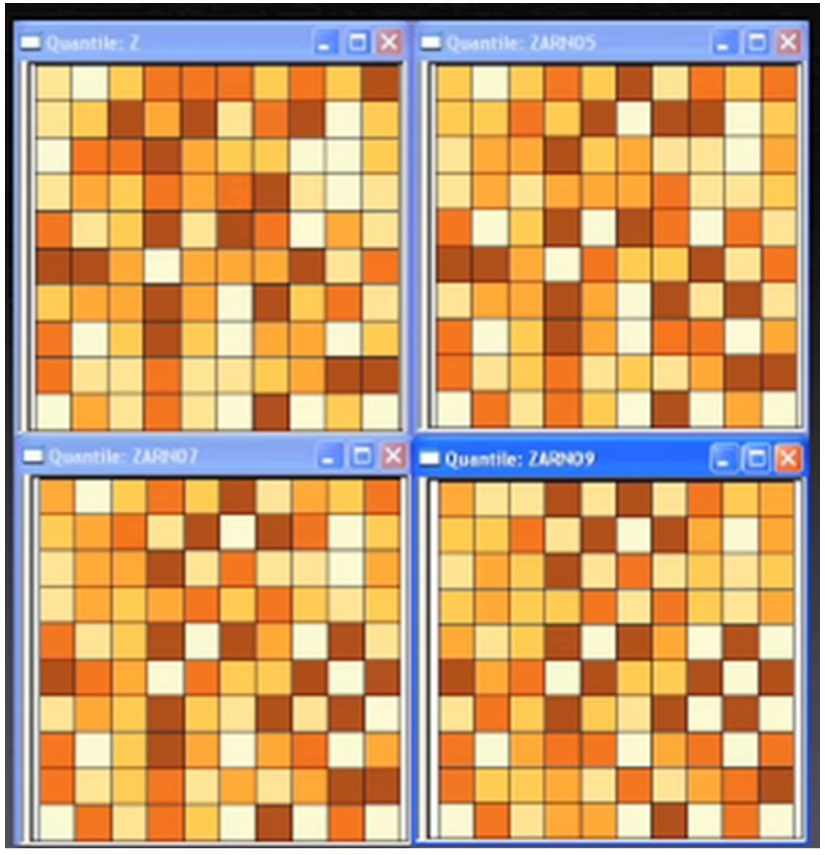
- Clustering
  - like values tend to be in similar locations.
- Neighbours are similar
  - more alike than they would be under spatial randomness.
- Compatible with diffusion
  - but not necessary caused by diffusion.



Legend: 0.1, 0.5, 0.7, 0.9

Negative Spatial Autocorrelation

- Checkerboard patterns
  - “opposite” of clustering
- Neighbours are dissimilar
  - more dissimilar than they would be under spatial randomness
- Compatible to competition
  - but not necessary competition



Legend: -0.1,-0.5, -0.7, -0.9

Measures of Global Spatial Autocorrelation

- Moran’s I
- Geary's c

Measures of Global Spatial Autocorrelation: Moran's I

Describe how features differ from the values in the study area as a whole

$$I(d) = \sum_{i}^{n}\sum_{j}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x}) \Big/ (S^2 \sum_{i}^{n}\sum_{j}^{n} w_{ij})$$

$$S^2 = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Moran I (Z value) is:
  - positive (I>0): Clustered, observations tend to be similar;
  - negative(I<0): Dispersed, observations tend to be dissimilar;
  - approximately zero: observations are arranged randomly over space.

Measures of Global Spatial Autocorrelation: Geary's c

- Describing how features differ from their immediate neighbours

$$C(d) = (n-1) \Big/ (2\sum_{i}^{n}\sum_{j}^{n} wij) \{\sum_{i}^{n}\sum_{j}^{n} wij(xi-xj)^2 \Big/ \sum_{i}^{n}(xi-\bar{x})^2\}$$

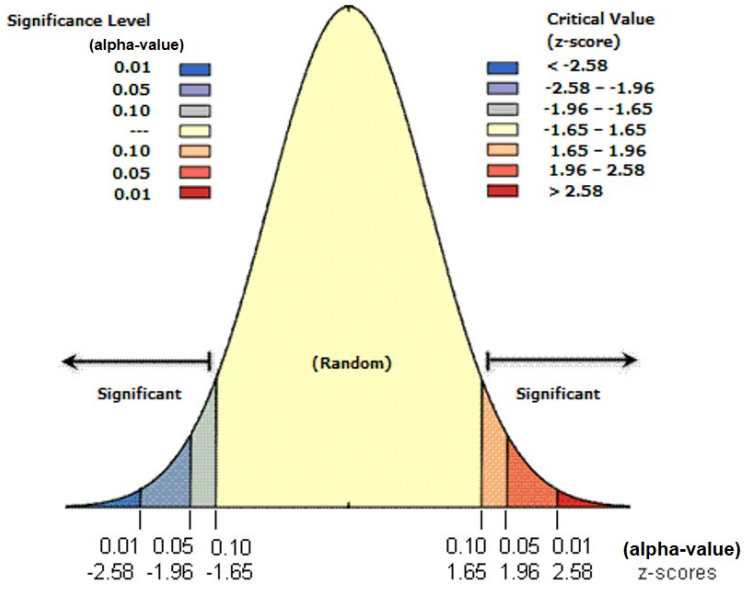
- Geary c (Z value) is:
  - Large c value (>1) : Dispersed, observations tend to be dissimilar;
  - Small c value (<1) : Clustered, observations tend to be similar;
  - c = 1: observations are arranged randomly over space.

Relationship of Moran's I and Geary's C

- C approaches 0 and I approaches 1 when similar values are clustered.
- C approaches 3 and I approaches -1 when dissimilar values tend to cluster. High values of C measures correspond to low values of I.
- So the two measures are inversely related.

z-score and p-value explained

- Statistically, we select the confident interval such as 95% => alpha value = 0.05.
- Reject the Null hypothesis (H0) if p-value is smaller than alpha value.
- Failed to reject the Null Hypothesis (H0) if p-value is greater than alpha value.



Reference: Confidence Interval or P-Value? https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689604/

Spatial Randomness

The Null Hypothesis:

- Observed spatial pattern of values is equally likely as any other spatial pattern.
- Values at one location do not depend on values at other (neighbouring) locations.
- Under spatial randomness, the location of values may be altered without affecting the information content of the data.

What if my data violate the assumptions?

- If you doubt that the assumptions of Moran's I are true (normality and randomization), we can use a Monte Carlo simulation.
  - Simulate Moran's I n times under the assumption of no spatial pattern,
  - Assigning all regions the mean value
  - Calculate Moran's I,
- Compare actual value of Moran's I to randomly simulated distribution to obtain p-value (pseudo significance).

Measures of Global High/Low Clustering

- Getis-Ord Global G

Measures of Global High/Low Clustering: Getis-Ord Global G

- The G(d)G(d)G(d) statistic is concerned with the overall concentration or lack of concentration in all pairs that are neighbours given the definition of neighbouring areas.
- The variable must contain only positive values to be used.

$$G(d) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d) X_i X_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j}$$

Source: Getis, A., & Ord, K. (1992). "The Analysis of Spatial Association by Use of Distance Statistics". *Geographical Analysis*, 24, 189–206.

Interpretation of Getis-Ord Global G

- The p-value is not statistically significant.
  - You cannot reject the null hypothesis. It is possible that the spatial distribution of feature attribute values is the result of random spatial processes. Said another way, the observed spatial pattern of values could be one of many possible versions of complete spatial randomness.
- The p-value is **statistically significant**, and the z-score is **positive**.
  - You can reject the null hypothesis. The spatial distribution of high values in the dataset is more spatially clustered than would be expected if underlying spatial processes were truly random.
- The p-value is **statistically significant**, and the z-score is **negative**.
  - You can reject the null hypothesis. The spatial distribution of low values in the dataset is more spatially clustered than would be expected if underlying spatial processes were truly random.

Localised Geospatial Statistics

- A collection of geospatial statistical analysis methods for analysing the **location related tendency** (clusters or outliers) in the attributes of **geographically referenced data** (points or areas).

- Can be indecies decomposited from their global measures such as local Moran's I, local Geary's c, and Getis-Ord Gi*.

- These spatial statistics are well suited for:

  - detecting clusters or outliers;
  - identifying hot spot or cold spot areas;
  - assessing the assumptions of stationarity; and
  - identifying distances beyond which no discernible association obtains.

Local Indicator of Spatial Association (LISA)

- A subset of localised geospatial statistics methods.
- Any spatial statistics that satisfies the following two requirements (Anselin, L. 1995):
  - the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;
  - the sum of LISAs for all observations is proportional to a global indicator of spatial association.

What is geographically referenced attributes?

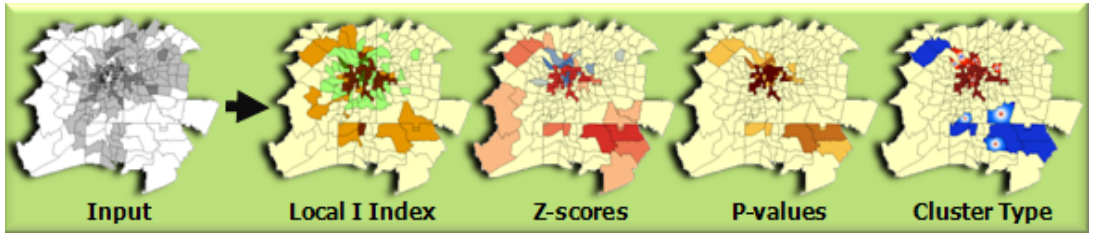
- Attributes that are recorded based on a geographical entity such as postal code, postal area, census block, district, state, province, and country
    - informal geographical entities includes regular grids or hexagons.
- These geographical entities can be in either point or polygon features.
- The attributes can be in absolute counts (i.e. number of people age 65 and above) or rates (i.e. proportion of population age 65 and above).
- It is univariate in nature.

Detecting Spatial Clusters and Outliers

- Given a set of geospatial features (i.e. points or polygons) and an analysis field, the spatial statistics identify **spatial clusters** of features with high or low values. The tool also identifies **spatial outliers**.

- local Moran's I is the most popular spatial statistical method used, other methods include local Geary's c.

- In general, the analysis will calculate a local statistic value, a z-score, a pseudo p-value, and a code representing the cluster type for each statistically significant feature. The z-scores and pseudo p-values represent the statistical significance of the computed index values.



Local Moran's I

Given a geographically referenced attribute field, X the formula of local Moran's I is:

The Local Moran's I statistic of spatial association is given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{X}) \qquad (1)$$

where x_i is an attribute for feature i, $\bar{X}$ is the mean of the corresponding attribute, $w_{i,j}$ is the spatial weight between feature i and j, and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{X})^2}{n - 1} - \bar{X}^2 \qquad (2)$$

with n equating to the total number of features.

Local Moran and Moran's I

- The summation of local Moran is

$$\sum_i I_i = \sum_i z_i \sum_j w_{ij} z_j$$

- Moran's I

$$I = (n/S_0) \sum_i \sum_j w_{ij} z_i z_j / \sum_i z_i^2$$

Test statistics of Local Moran

The z_{I_i}-score for the statistics are computed as:

$$z_{I_i} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{\mathrm{V}[I_i]}} \tag{3}$$

where:

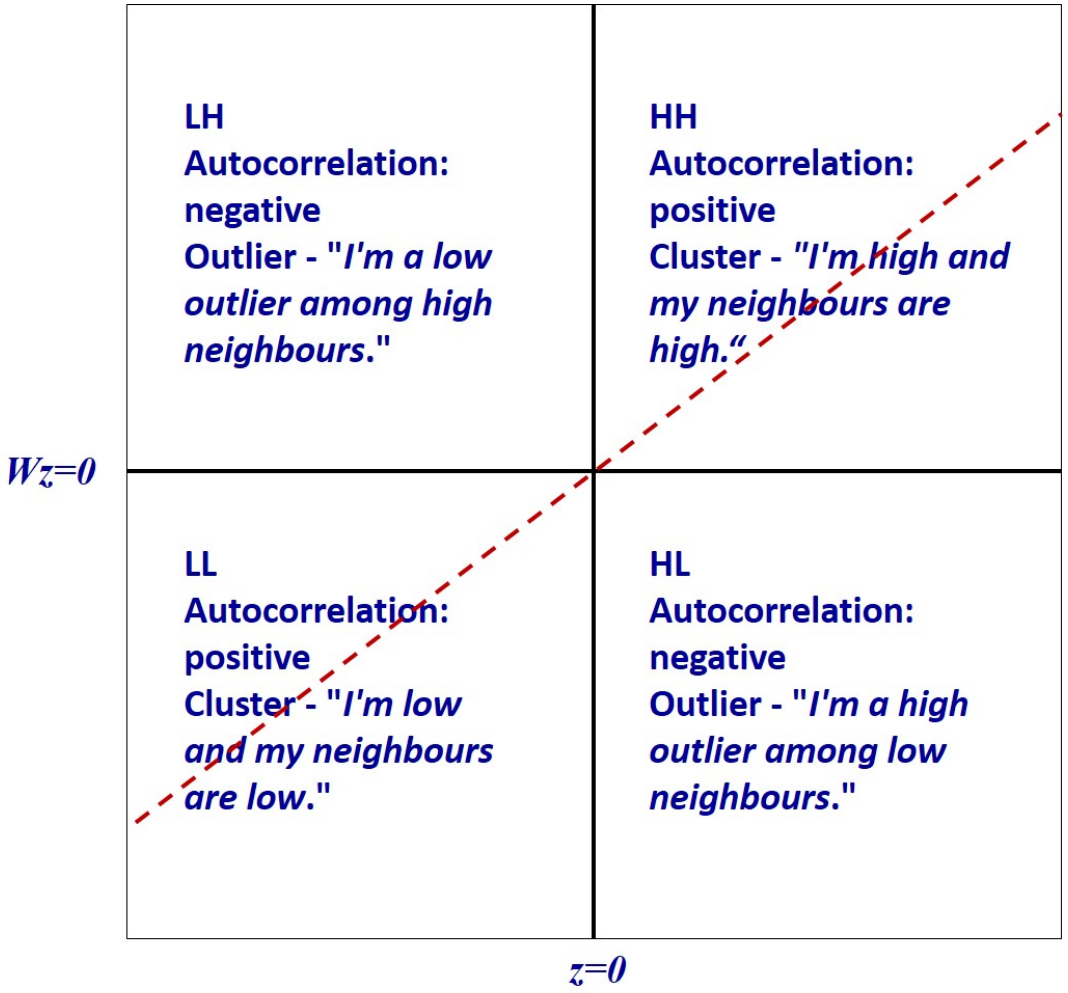
$$\mathrm{E}[I_i] = -\frac{\sum_{j=1, j\neq i}^{n} w_{ij}}{n-1} \tag{4}$$

$$\mathrm{V}[I_i] = \mathrm{E}[I_i^2] - \mathrm{E}[I_i]^2 \tag{5}$$

Interpretation of Local Moran

- An **outlier**: significant and negative if location i is associated with relatively low values in surrounding locations.
- A **cluster**: significant and positive if location i is associated with relatively high values of the surrounding locations.
- In either instance, the p-value for the feature must be small enough for the cluster or outlier to be considered statistically significant.
- The commonly used alpha-values are 0.1, 0.05, 0.01, 0.001 corresponding the 90%, 95, 99% and 99.9% confidence intervals respectively.

Interpretation of Local Moran and Scatterplot



Other forms of LISA

- Local Geary

$$LG_i = \sum_j w_{ij}(x_i - x_j)^2$$

- The observation x_i and x_j are in deviations from the mean, and the summation over j is such that only neighbouring values $j \in J_i$ are included.
- The weights w_{ij} may be in row-standardized form, though this is not necessary, and by convention, $w_{ii} = 0$.

Detecting hot and cold spot areas

- Given a set of geospatial features (i.e. points or polygons) and an analysis field, the spatial statistics tell you where features with either high (i.e. hot spots) or low values (cold spots) cluster spatially.

- The spatial statistic used is called Getis-Ord Gi* statistic (pronounced G-i-star).

Getis-Ord Gi*

Getis and Ord (1992) define the local G and G∗ statistics for region i (i=1,⋯,n) as

$$G_i(d) = \frac{\sum_{j \neq i}^{n} w_{ij}(d)\, x_j}{\sum_{j \neq i}^{n} x_j} \quad \text{when} \quad w_{ii}(d) = 0$$

$$G_i^*(d) = \frac{\sum_{j}^{n} w_{ij}(d)\, x_j}{\sum_{j}^{n} x_j} \text{when} \quad w_{ii}(d) \neq 0.$$

Getis-Ord Gi*

For variable x to be nonpositive and the weight matrix W(d) to be nonbinary the G and G* statistics are defined as

$$G_i(d) = \frac{\sum_{j \neq i}^{n} w_{ij}(d)\, x_j - W_i \bar{x}(i)}{s(i)\left\{\left[(n-1)S_{1i} - W_i^2\right]/(n-2)\right\}^{1/2}}$$

$$\text{when} \quad w_{ii}(d) = 0$$

$$\text{and} \quad G_i^*(d) = \frac{\sum_{j}^{n} w_{ij}(d)\, x_j - W_i^* \bar{x}}{s\left\{\left[(nS_{1i}^*) - (W_i^*)^2\right]/(n-1)\right\}^{1/2}}$$

$$\text{when} \quad w_{ii}(d) \neq 0, \tag{5}$$

where $W_i = \sum_{j \neq i}^{n} w_{ij}(d)$, $W_i^* = W_i + w_{ii}$, $\bar{x}(i) = \sum_{j \neq i}^{n} x_j/(n-1)$, $s^2(i) = \sum_{j \neq i}^{n} x_j^2/(n-1) - [\bar{x}(i)]^2$, $S_{1i} = \sum_{j \neq i}^{n} w_{ij}^2(d)$, $S_{1i}^* = \sum_{j}^{n} w_{ij}^2(d)$, and $\bar{x}$ and s^2 denote the usual sample mean and variance, respectively.

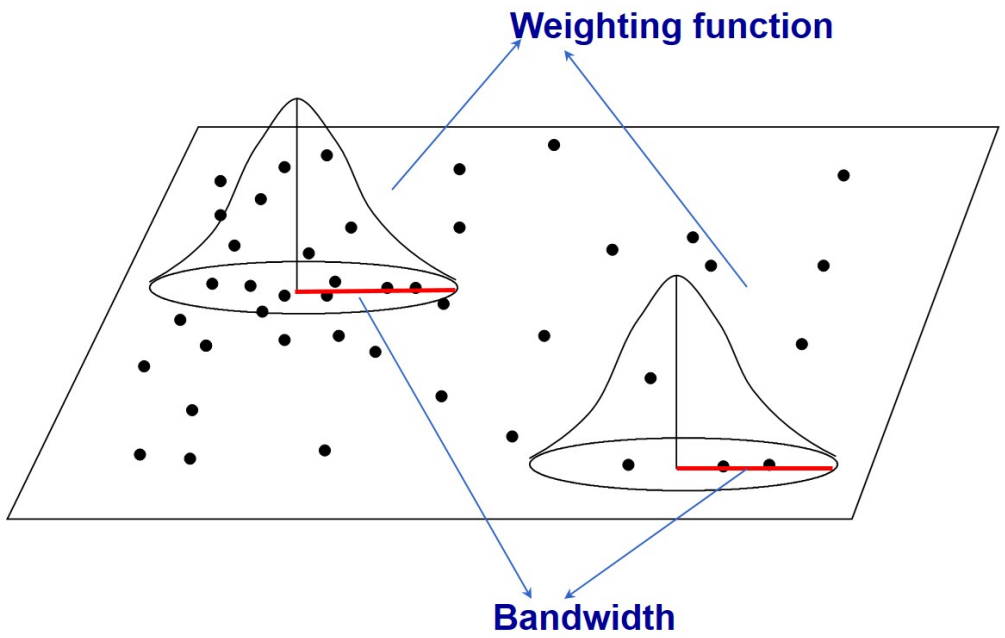
Interpretation of Getis-Ord Gi and Gi*

- A hot spot area: significant and positive if location i is associated with relatively high values of the surrounding locations.
- A cold spot area: significant and negative if location i is associated with relatively low values in surrounding locations.

Fixed weighting scheme

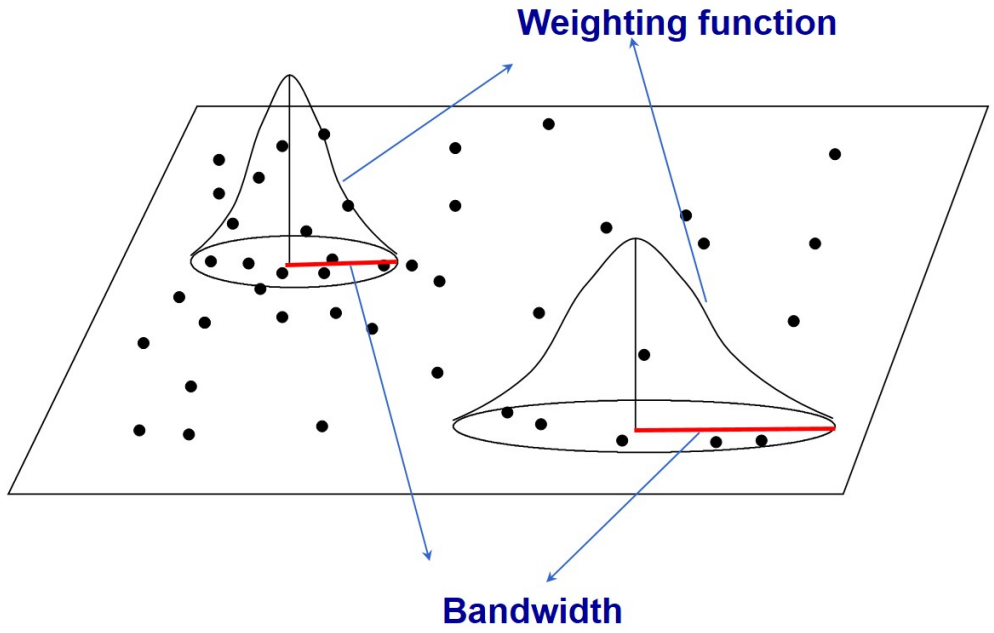
Things to consider if fixed distance is used:

- All features should have at least one neighbour.
- No feature should have all other features as neighbours.
- Especially if the values for the input field are skewed, you want features to have about eight neighbors each.
- Might produce large estimate variances where data are sparse, while mask subtle local variations where data are dense.
- In extreme condition, fixed schemes might not be able to calibrate in local areas where data are too sparse to satisfy the calibration requirements (observations must be more than parameters).



Adaptive weighting schemes

- Adaptive schemes adjust itself according to the density of data
  - Shorter bandwidths where data are dense and longer where sparse.
  - Finding nearest neighbors are one of the often used approaches.



Best practice guidelines

- Results are only reliable if the input feature class contains at least 30 features.
- The input field mst be in continuous data type such as a count, rate, or other numeric measurement, no categorical attribute field is allowed.

Best practice guidelines

Select an appropriate spatial weighting method

- The **polygon contiguity** method is effective when polygons are similar in size and distribution, and when spatial relationships are a function of polygon proximity (the idea that if two polygons share a boundary, spatial interaction between them increases).
    - When you select a polygon contiguity conceptualization, you will almost always want to select row standardization for tools that have the Row Standardization parameter.
- The **fixed distance** method works well for point data. It is often a good option for polygon data when there is a large variation in polygon size (very large polygons at the edge of the study area and very small polygons at the center of the study area, for example), and you want to ensure a consistent scale of analysis.

Best practice guidelines

Select an appropriate spatial weighting method

- The **inverse distance** method is most appropriate with continuous data or to model processes where the closer two features are in space, the more likely they are to interact/influence each other.
  - Be warned that with this method, every feature is potentially a neighbour of every other feature, and with large datasets, the number of computations involved will be enormous.

Best practice guidelines

Select an appropriate spatial weighting method

- The **k-nearest neighbours** method is effective when you want to ensure you have a minimum number of neighbors for your analysis.
    - Especially when the values associated with your features are skewed (are not normally distributed), it is important that each feature is evaluated within the context of at least eight or so neighbors (this is a rule of thumb only).
    - When the distribution of your data varies across your study area so that some features are far away from all other features, this method works well.
    - Note, however, that the spatial context of your analysis changes depending on variations in the sparsity/density of your features.
    - When fixing the scale of analysis is less important than fixing the number of neighbors, the k-nearest neighbours method is appropriate.

Futher guide on selecting a fixed-distance band value

- Select a distance based on what you know about the geographic extent of the spatial processes promoting clustering for the phenomena you are studying.
- Use a distance band that is large enough to ensure all features will have at least one neighbor, or results will not be valid.
- Try not to get stuck on the idea that there is only one correct distance band. Reality is never that simple. Most likely, there are multiple/interacting spatial processes promoting observed clustering.
- Select an appropriate distance band or threshold distance.
    - All features should have at least one neighbour.
    - No feature should have all other features as a neighbour.
    - Especially if the values for the input field are skewed, each feature should have about eight neighbours.

In colclusion

Spatial statistics methods are not a blackbox. Before performing the analysis, a geospatial analyst should consider the followings:

- What is the **geographical** question?
- What is the **geospatial feature**?
- What is the **analysis field**?
- Which **conceptualization of spatial relationships** is appropriate?

Case Study 1: Area-based

Is There Space for Violence?

A Data-driven Approach to the Exploration of Spatial-temporal Dimensions of Conflict

Vincent Z. W. Mack

&

Tin Seong KAM

Associate Professor of Information Systems (Practice)

School of Information Systems

Singapore Management University

Reference: Mack, Z.W.V. and Kam T.S. (2018) "Is There Space for Violence?: A Data-driven Approach to the Exploration of Spatial-Temporal Dimensions of Conflict" *Proceedings of 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities (ACM SIGSPATIAL'18)*. Seattle, Washington, USA, 10 pages.

Reference: https://ink.library.smu.edu.sg/sis_research/4331/

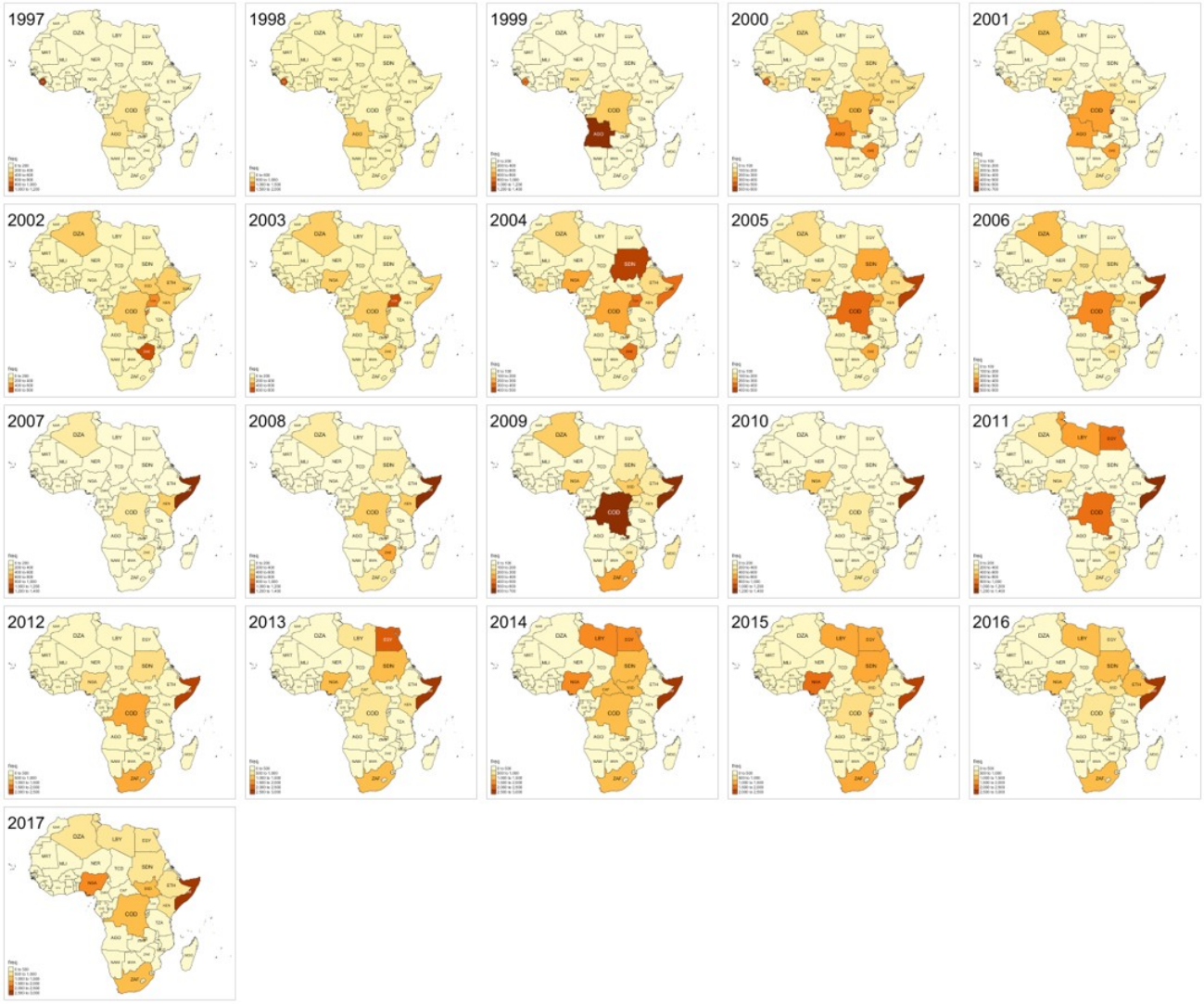
Objectives

- Using micro-level event data of armed conflicts in Africa, this study aims to show how a data-driven geospatial analytics approach can be used reveal useful spatio-temporal pattern of the conflict events,

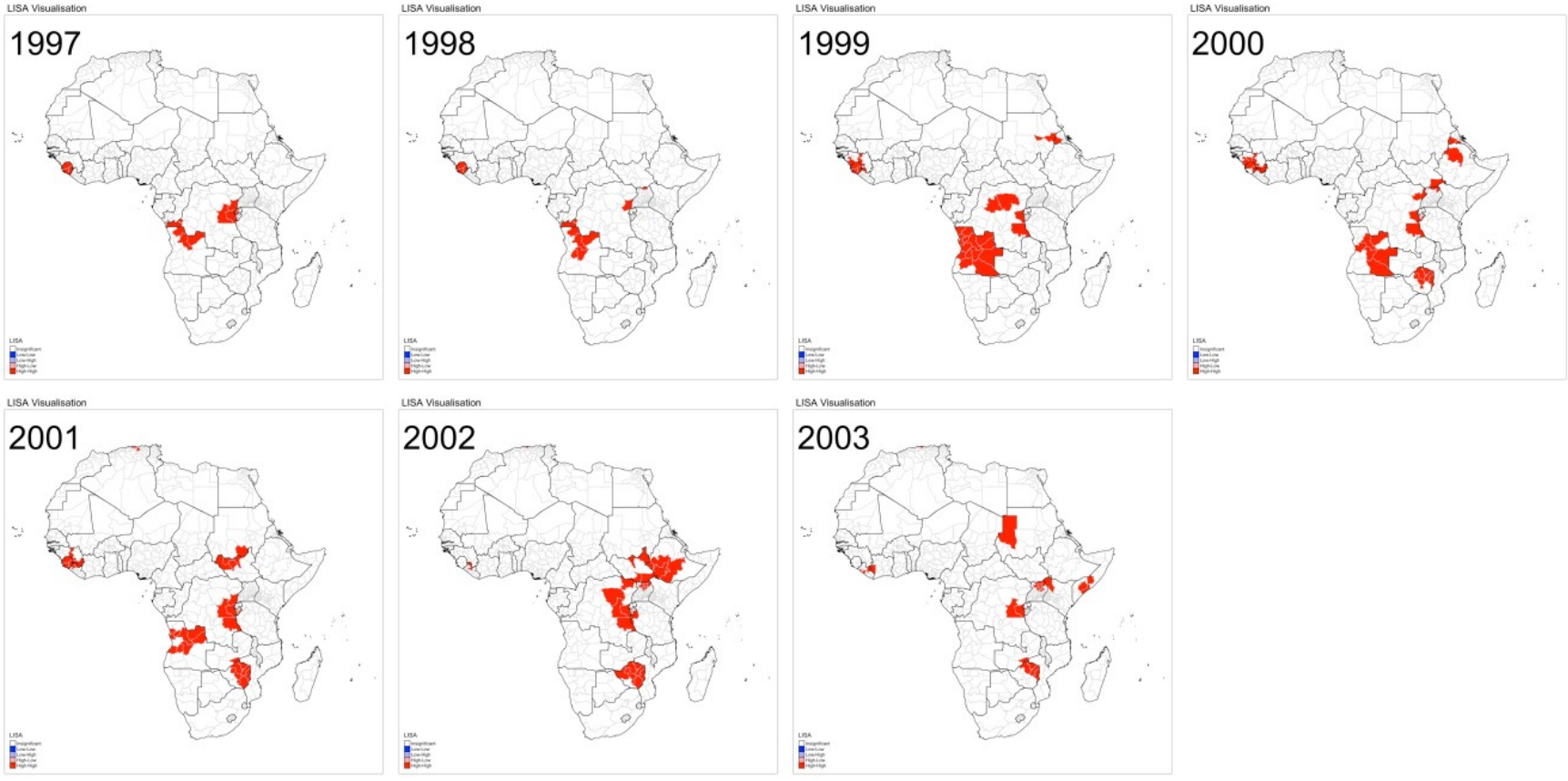
- Demonstrating how a reproducible research can be conducted by using R Markdown, Rstudio and other appropriate R packages, and

- Sharing the findings and more importantly, the approaches we used to the practice political researchers so that they are confident to conduct similar studies by themselves.

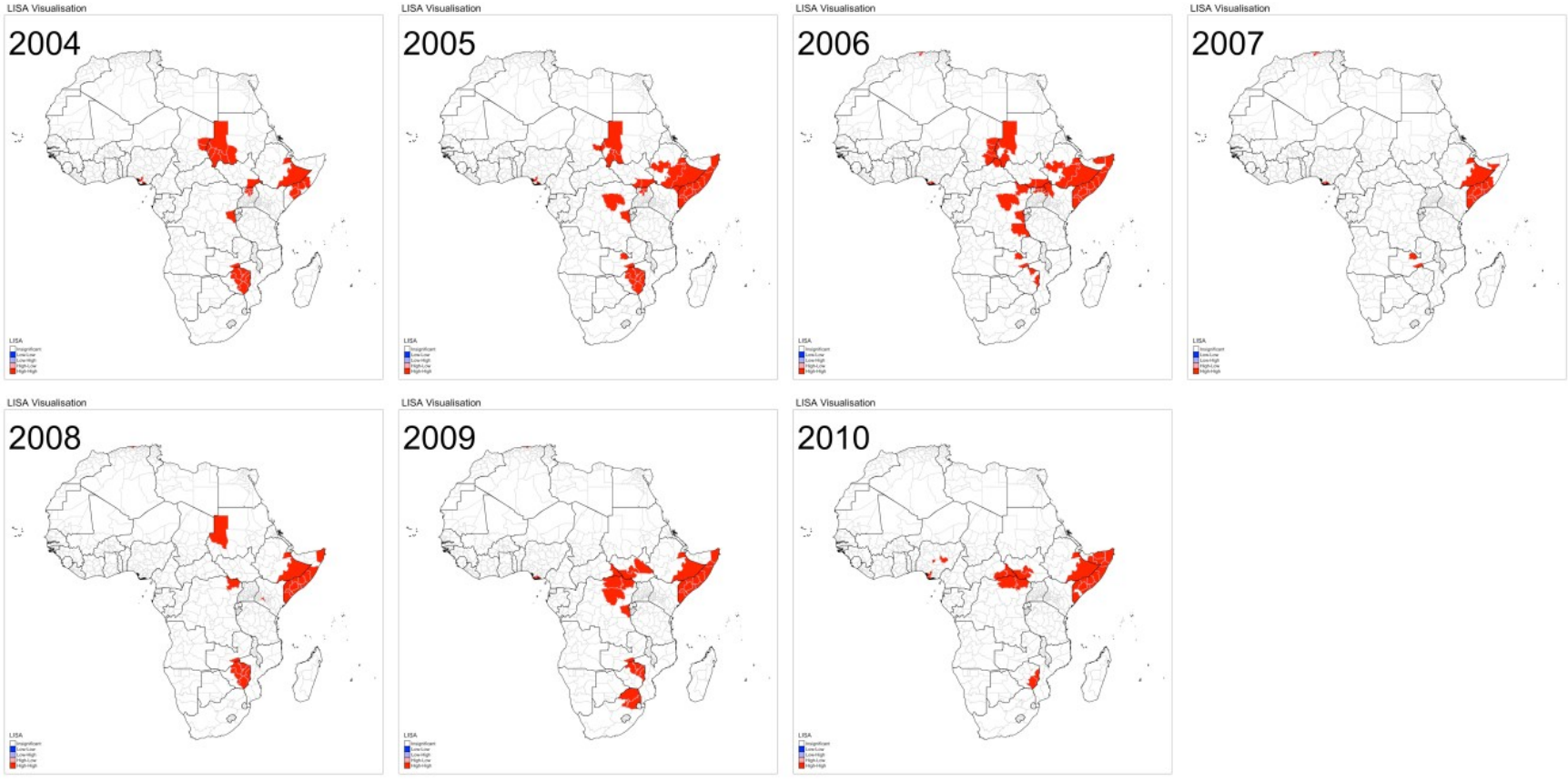
Violence against civilians in Africa, 1997-2017



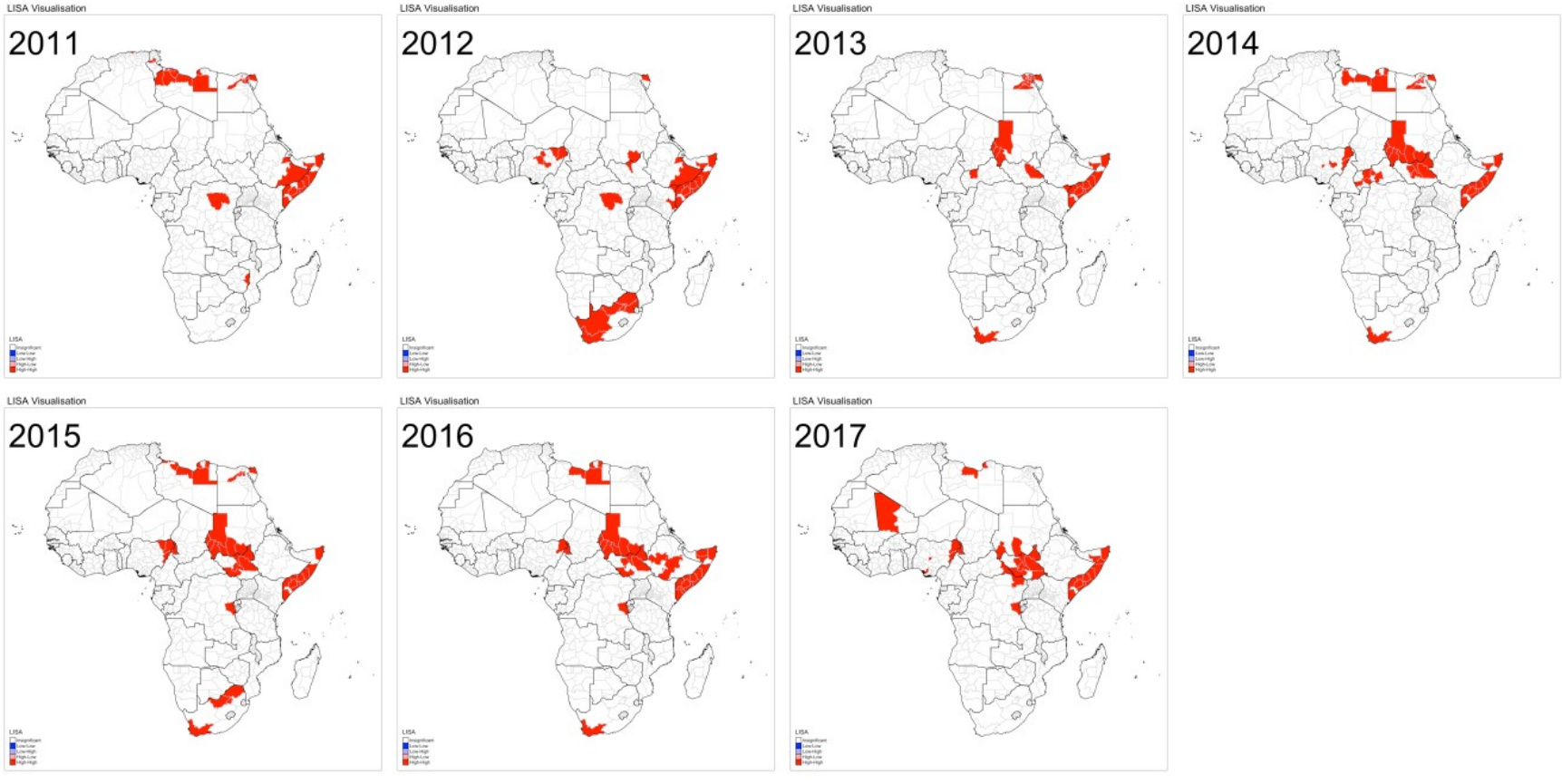
LISA Cluster Map for Phase I - from 1997 to 2003



LISA Cluster Map Phase II - from 2004 to 2010



LISA Cluster Map Phase III - from 2011 to 2017



Case Study 2: Point-based

Exploring and Visualizing Household Electricity Consumption Patterns in Singapore: A Geospatial Analytics Approach

Joanne Tan Yong Ying and Kam Tin Seong
Associate Professor of Information Systems (Practice)
School of Information Systems
Singapore Management University

Reference: TAN, Yong Ying and KAM, Tin Seong (2019) "Exploring and Visualizing Household Electricity Consumption Patterns in Singapore: A Geospatial Analytics Approach", by. Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings. Pp 785-796, 12 pages. https://doi.org/10.1007/978-3-030-15742-5_74.

Reference: https://ink.library.smu.edu.sg/sis_research/4376/

Objectives

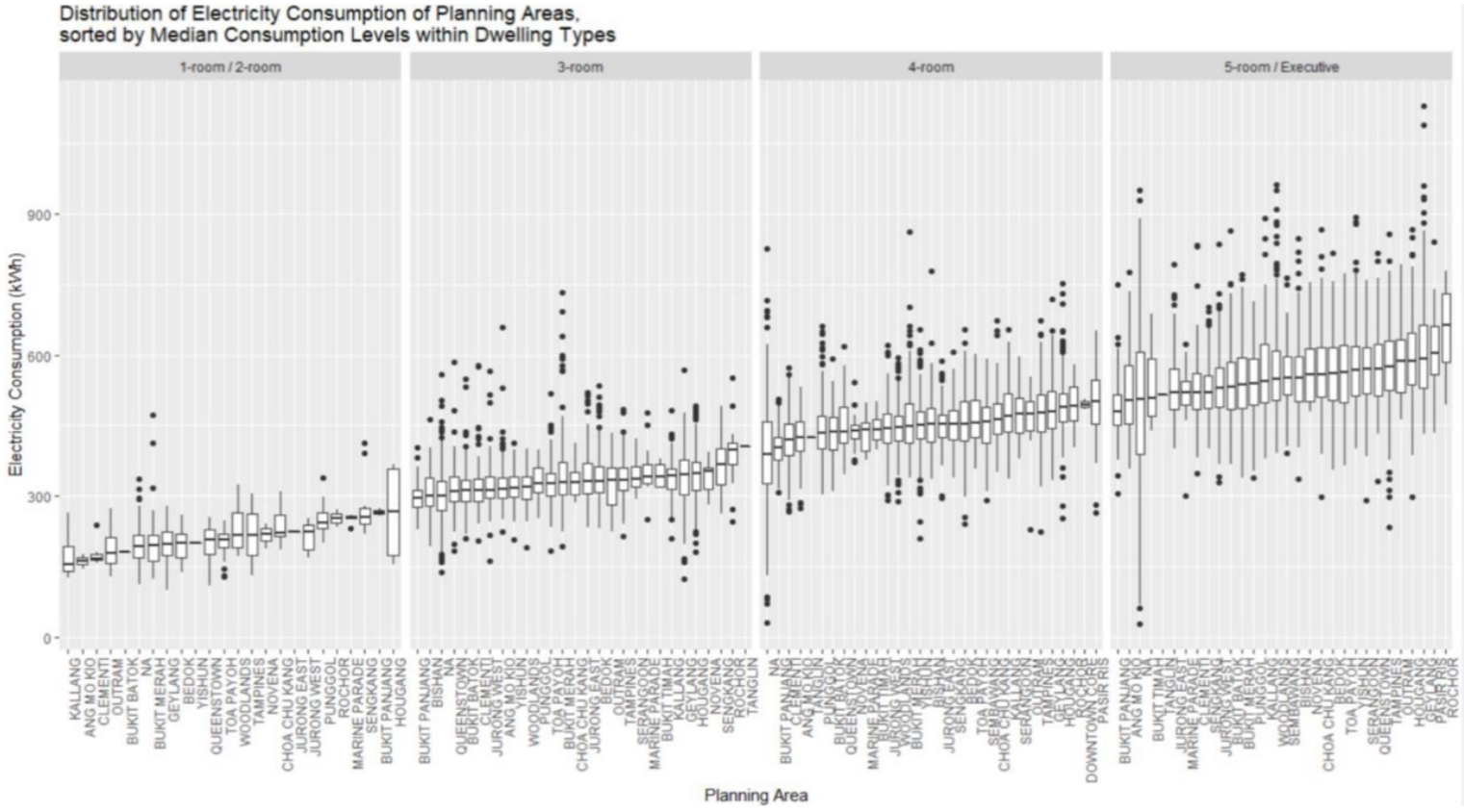
This project aims to contribute new knowledge towards the study of electricity consumption and its analyses in two ways.

- Firstly, we will analyze electricity consumption using mainly two spatial analysis methods, Local Indicators of Spatial Association (LISA) and Geographically-Weighted Principal Components Analysis (GWPCA) to discover spatial patterns on electricity consumption patterns.

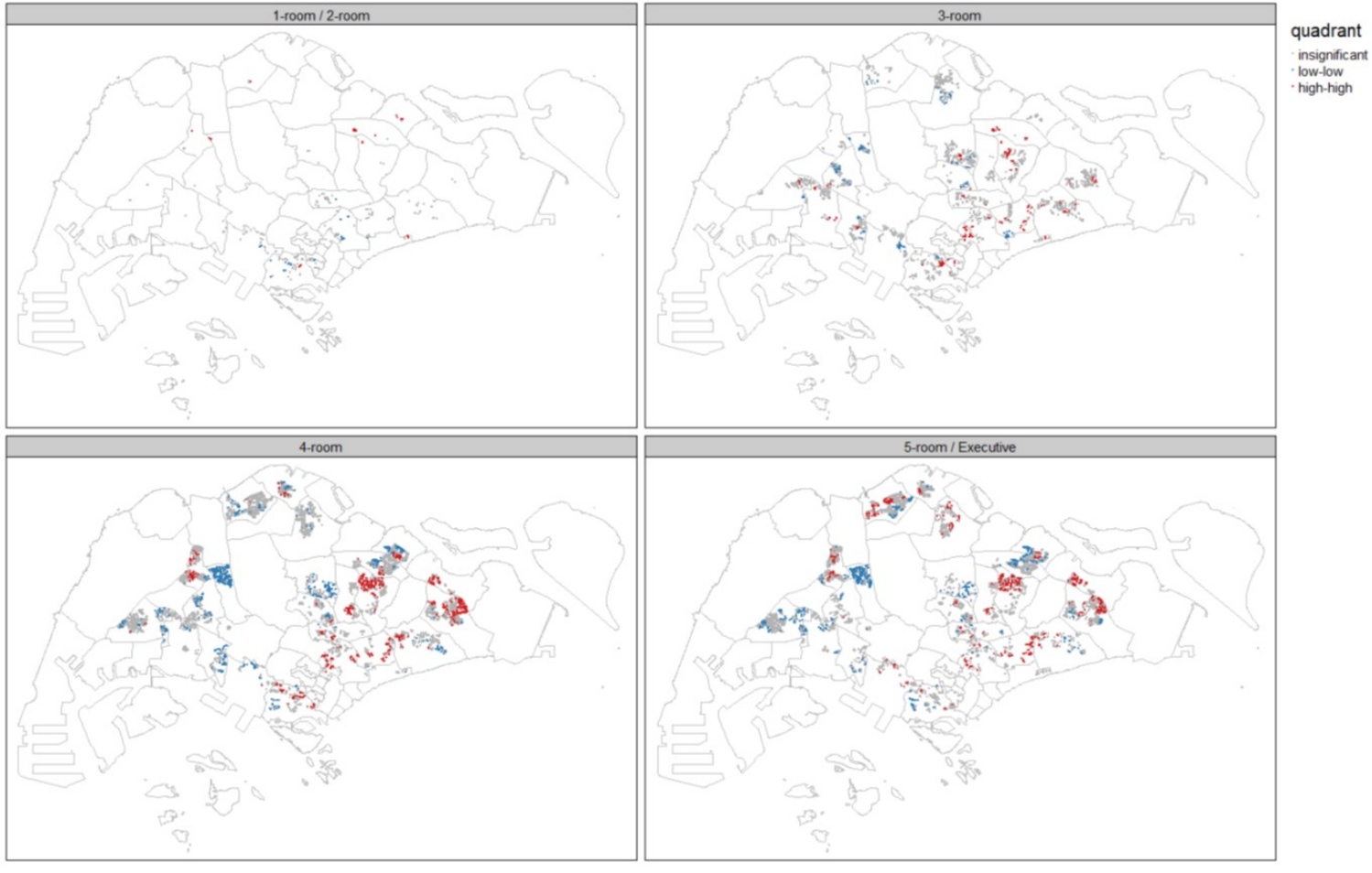
- Secondly, this project aims to be a proof-of-concept of using R to achieve an end-to-end analytics solution, from data cleaning and preparation all the way to visualization of results. The visualizations should not only present the analysis results in a clear manner but also prompt users to do their own exploration of electricity consumption in Singapore and derive insights that fit their purpose.

Exploratory Data Analysis

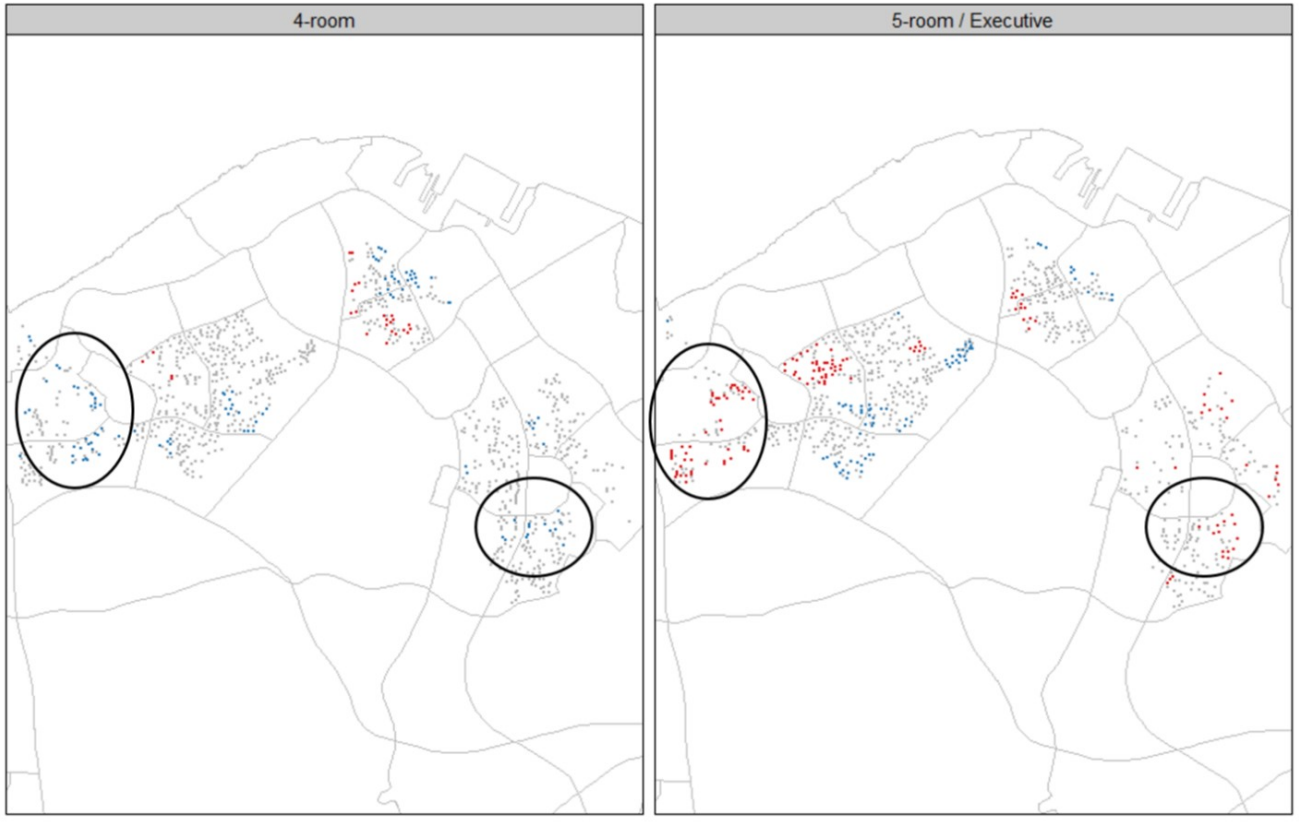
Boxplots are used to reveal the statistical distributions of electricity consumption by dwelling types.



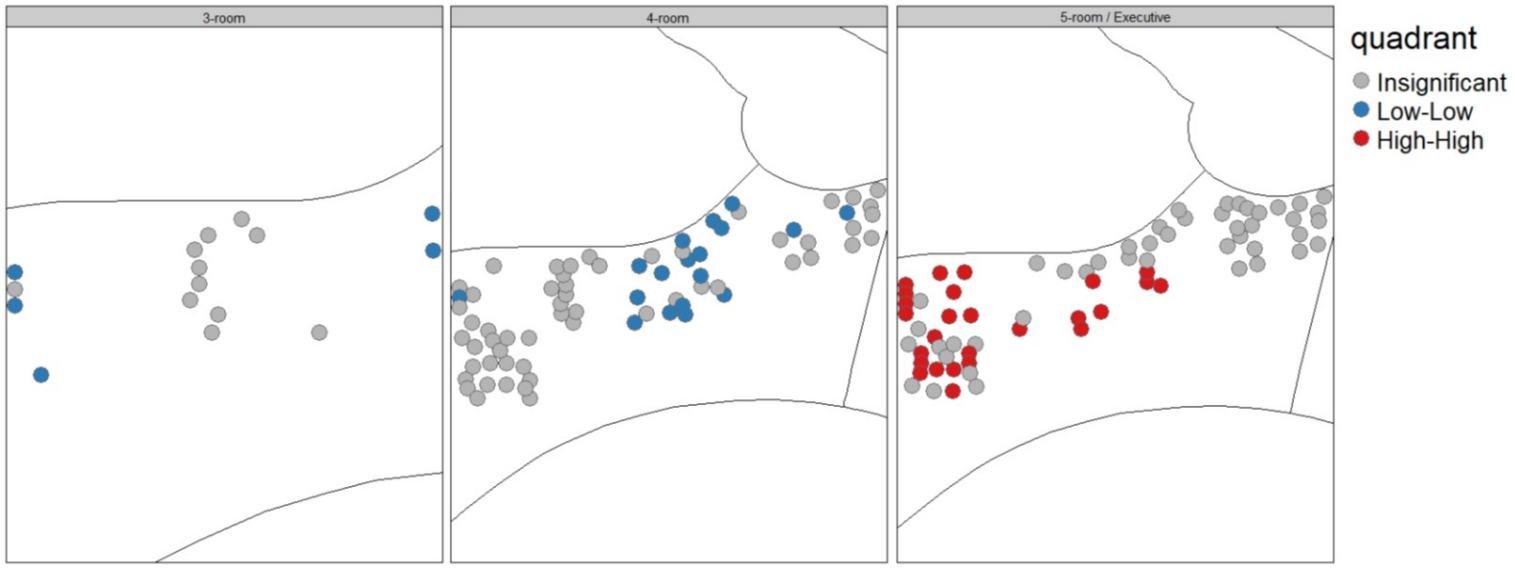
LISA Map of average electricity consumptions: At the national level



LISA Map of average electricity consumptions: At the planning area level



LISA Map of average electricity consumptions: At the local level



References

- Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena". Biometrika. 37 (1): 17–23.
- Geary, R.C. (1954) "The Contiguity Ratio and Statistical Mapping". *The Incorporated Statistician*, Vol. 5, No. 3, pp. 115-127.
- Moran's I
- Geary's c
- Getis, A., & Ord, K. (1992). "The Analysis of Spatial Association by Use of Distance Statistics". *Geographical Analysis*, 24, 189–206.
- Anselin, L. (1995). "Local indicators of spatial association – LISA". *Geographical Analysis*, 27(4): 93-115.
- Getis, A. and Ord, J.K. (1992) “The analysis of spatial association by use of distance statistics”. *Geographical Analysis*, 24(3): 189-206.
- Ord, J.K. and Getis, A. (2010) “Local spatial autocorrelation statistics: Distributional issues and an application”. *Geographical Analysis*, 27(4): 286-306.