Spatial Clustering

EE464: Urban Economics
EE562: Selected Topics in Development Economics 2
Semester 1 / 2020
Faculty of Economics, Thammasat University

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

Main Topics

1. Introduction
2. Global Clustering
3. Local Clustering

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

1.Introduction

- The ability to visualize spatial data allows for the quick **identification of any obvious patterns**, and in general, spatial pat terns can be classified as regular, random, or clustered.
- The term ‘clustering’ is used to describe the spatial aggregation of events, but as the observed spatial pattern may simply be a function of the distribution of the population at risk or of various risk factors.
- A more robust definition is the one proposed by Wakefield et al. (2000), that data are clustered if there is 'residual spatial variation in risk after known influences have been accounted for’.
- Besag and Newell (1991) classified the different methods for analyzing clusters as either specific or non-specific, although epidemiologists prefer to use the terms **'local'** and **'global'**.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

1.Introduction (Cont'd)

- **Global** (non-specific) clustering methods are used to assess **whether clustering is apparent throughout the study region** but ***do not identify the location of clusters***.
  - They provide a **single statistic** that measures **the degree of spatial clustering**, the statistical significance of which can then be assessed.
  - The **null hypothesis** for global clustering methods is simply that "no clustering exists' (i.e. random spatial dispersion).
- **Local** (specific) methods of **cluster detection define the locations** and extent of clusters, and can be further divided into focused and non-focused tests.
  - **Non-focused tests** identify the location of **all likely clusters** in the study region
  - **Focused tests** investigate whether there is an increased probability (e.g. risk of disease around a pre-determined point, such as a nuclear power plant or chemical factory.)
- Please be noted that the term clustering applies to global methods of cluster analysis, while cluster detect refers to local methods of analysis.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering

Statistical power of clustering methods

- This discusses some of the more commonly- used cluster analysis methods, and although certain tests are more appropriately used in specific situations, when there are competing methods a commonly asked question is ***'which is best?'***.
- In such instances, the **statistical power of the test** (i.e. the probability of detecting clustering or clusters when they actually occur) can provide an answer.
- Waller and Gotway (2004) discuss various issues affecting the ability of a test to correctly detect clustering, including the structure of the data and the alternative hypothesis.
- A comparison of the performance of different clustering methods is provided in Walter (1992; 1993), Kulldorff et al. (2003), Song and Kulldorff (2003) and Kulldorff et al. (2006).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- When compared with other global clustering tests, Tango's maximized excess events test (MEET) has been found to be the most consistent for evaluating cancer maps (Kulldorff et al. 2006a) and have the highest power of the tests under evaluation (Kulldorff et al. 2003; Song and Kulldorff 2003).
- Cuzick and Edwards' k-nearest neighbour test has also been shown to perform well, especially when the clustering occurs over small distances (Kulldorff et al. 2003).
- A detailed discussion of the statistical power of clustering and cluster detection methods can be found in Waller and Gotway (2004).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

Methods for aggregated data

- Autocorrelation statistics for aggregated data provide an estimate of **the degree of spatial similarity** observed **among neighbouring value** of an attribute over a study area.
- There are various autocorrelation statistics for aggregated data; two of which are discussed:
  - **Moran's I**
  - **Geary's c**
- Details of other autocorrelation tests can be found in Cliff and Ord (1973; 1981).
- **Fundamental** to all autocorrelation statistics is the **weights matrix**, used to define the spatial relationships of the regions so that those that **are close in space are given greater weight** in the calculation than those that are distant (Moran 1950).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- **Neighbours** can be defined based either on:
  - **adjacency**
  - **distance.**
- Methods based on **adjacency** (also known as contiguity) include:
  - **rook contiguity** (polygons are adjacent if they share a border)
  - **queen contiguity** (polygons are adjacent if they share a border or corner)
  - **higher order contiguities** (often called spatial lags) such as first-order (neighbours or second-order adjacency (neighbours of neighbours).
- When using **distance** to define neighbours, polygons with their **centroids** located within a **specified distance range** are **considered to be adjacent**.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- More complicated neighbour, definitions include **row-standardization**, length of common boundary or relationships that only act one way (e.g. large polygons may influence, but not be influence by, neighbouring, smaller polygons).

- Software for producing the weights matrix includes **GeoDa** and **ArcView**.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

Moran's I

- Moran's I coefficient of autocorrelation is **similar** to **Pearson's correlation coefficient**, and **quantifies the similarity** of an outcome variable among areas that are defined as spatially related (Moran 1950)
- **Moran's I statistic** is given by:

$$I = \frac{n \sum_i \sum_j w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_i \sum_j w_{ij} \ \sum_k (Z_k - \bar{Z})^2}$$

where Z_i could be the **data of an area**, and w_{ij} is a measure of the **closeness of areas i and j**.

- A **weights matrix** is used to define the spatial relationships so that regions **close in space** are **given greater weight** when calculating the statistic than those that are distant (Moran 1950).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- **Moran's I** is **approximately normally distributed** and has an expected value of

$$-1 / (N-1)$$

  where N equals the number of area units within a study region, when no correlation exists between neighbouring values.
- The expected value of the coefficient therefore approaches zero as N increases.
- Although Moran's I generally lies **between +1 and -1**, it is not bound by these limits (unlike Pearson's correlation coefficient, Waller and Gotway 2004).
- A Moran's I of **zero** indicates the null hypothesis of **no clustering**.
- A **positive** Moran's indicates **positive spatial autocorrelation** of areas of similar attribute relation.
- A **negative** Moran's I indicates negative spatial autocorrelation (i.e. that **neighbouring areas tend to have dissimilar attribute values**)
- The **significance of Moran's I** can be assessed using **Monte Carlo randomization** (Moran 1950).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- Moran's I can be calculated using various software packages, including ClusterSeer, R, and GeoDa.
- Disadvantages of the autocorrelation test are the assumptions that the population at risk is evenly distributed within the study area (Moran 1950) and that the correlation or covariance is the same in all directions (i.e. it is isotropic).
- The problems posed by an isotropic data can be overcome by manipulating the weights matrix to reflect directional inequalities.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- Although Moran's I is intended for use with continuous data it can also be used to analyse count data, even though, in such instances, any observed autocorrelation may simply be the result of variation in regional population sizes rather than a genuine spatial pattern in the disease counts (e.g. if regions with large populations are grouped together).
- Accounting for the population at risk by using regional incidence rates, instead of regional disease counts, increases the likelihood that an observed autocorrelation reflects a genuine spatial pattern rather than a heterogeneous population distribution.
- A **spatial correlogram** is a series of estimates of Moran's I evaluated at increasing distances. **Moran's I is plotted on the vertical axis and distance (spatial lag)** is plotted on the horizontal axis

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- The correlogram can therefore be used to determine where, on average, spatial autocorrelation is maximized.
- Correlograms can be calculated based on distance or adjacency.
- As Moran's I cannot adjust for a heterogeneous population density, Oden (1995) proposed the use of Oden's Ipop, a test statistic similar to Moran's I but in which rates are adjusted for population size.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

Geary's c

- Geary's contiguity ratio, or Geary's c, is **another weighted estimate** of **spatial autocorrelation** (Geary 1954) but whereas Moran's I considers similarity between neighbouring regions,
- Geary's c considers **similarity between pairs of regions**.
- Geary's c ranges from **zero to two**, with **zero** indicating **perfect positive** spatial autocorrelation and **two** indicating **perfect negative** spatial autocorrelation, for any pair of regions.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

2.Global Clustering (Cont'd)

- Geary's c is given by:

$$c = \frac{(n-1)\sum_i \sum_j w_{ij}(y_i - y_j)^2}{2(\sum_i (y_i - \bar{y})^2) \quad (\sum_i \sum_j w_{ij})}$$

where $\boldsymbol{n}$ is the number of polygons in the study area, $\boldsymbol{w}$ the values of the spatial proximity matrix, $\boldsymbol{y}$ the attribute under investigation, and $\bar{\boldsymbol{y}}$ the mean of the attribute under investigation.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering

- **Global measures** of spatial association assume that **the spatial process** under investigation is **stationary**.
- Under this assumption, which is **rarely met global tests** of association run the risk of **obscuring local effects**.
- As a result, **significant local clustering may not be detected** and conversely, large non-clustered areas within a study area may be ignored.
- The likelihood of including regions with inherently different local relationships increases as the size of the study area increases.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- With very large spatial datasets, and particularly with large raster datasets such as remotely sensed images, **global statistics** such as **Moran's I** (Moran 1948) run the **risk of losing information on spatial autocorrelation** since they summarize an enormous number of possibly dissimilar spatial relationships.
- **Local statistics** overcome this problem by **scanning the entire dataset**, but only measuring **dependence in limited portions of the study area**, the bounds of which have to be specified.
- Thus, for **clustering to be detected**, it need **not occur over the entire dataset**, nor need it have the same characteristics throughout the study area.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- In studying local area clustering we are concerned with **defining the characteristics** of clusters, such as their **location**, **size**, and **intensity**.
- Knox (1989) proposes some definitions of clustering, one of which defines a cluster as '**a geographically and/or temporally bounded group of occurrences of sufficient size and concentration to be unlikely have occurred by chance**.'

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- The detection of disease clusters, particularly around putative point sources, has been the subject of much debate.
- In 1989, the **Royal Statistical Society** reported on a meeting held specifically to **discuss the occurrence of cancer near nuclear installations** (Muirhead and Darby 1989).
- The meeting was triggered by the **highly popularized reports of excesses of childhood cancer**, in particular leukaemia, around the reprocessing plants at Sellafield in Cumbria and Dounreay in northern Scotland.
- Its objective was to bring together **epidemiologists (Doll 1989) and statisticians (Hills and Alexander 1989)** to try and **resolve whether excesses of cancer really were occurring in the vicinity of nuclear installations**.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- This triggered a spate of meetings, reported in Rothenberg et al. (1990), Lawson et al. (1995), Cliff (1995), Jacquez et al. (1996) and Smith (1996).
- These addressed the legal aspects of disease cluster detection (Henderson 1990; Jacquez et al. 1992, and provided some useful reviews of statistical tests including those by Walter (1993), Waller and Turnbill(1993), Cliff(1995) and Elliott et al.(1995)
- In 1999 the **Royal Statistical Society** hosted an **international conference on the analysis and interpretation of clusters** and ecological studies (Wakefield et al. 2001), during which the debate on whose method performs best was fuelled, old methodologies were adapted, and new methods introduced.
- Also addressed at this meeting were more philosophical questions such as **when and how to investigate clusters**, and indeed whether we should bother at all (Elliott and Wakefield 2001; Wartenberg 2001).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- The complexity of the issues involving the public, media, local health authorities, epidemiologists, and those responsible for putative sources of increased incidence of disease, is exemplified in a case where the media reacted to documents released by the Ministry of Defence in 1977 disclosing **simulation trials of germ warfare** along the south coast of England between 1961 and 1977 (Stein 2001).
- Families in East Lulworth, a coastal village in Weymouth Bay, believed that **exposure to the agents released** during these trials had caused high rates of **miscarriages, stillbirths, birth defects, and learning disabilities in the village**.
- Under pressure from the media, campaigning families, environmental activists, and Members of Parliament the local health authority was forced to conduct a **full scale investigation**.
- **No evidence of clustering could be found** and the 'cluster that never was' was eventually refuted (Spratt 1999).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- So 'cluster busting' has a long some difficult statistical, epidemiological, environmental, legal, and social territories.
- Cautious words on the interpretation of disease clustering studies (1992), Urquhart (1992), and Rothman (1990) who discuss some of the potential **pitfalls surrounding cluster detection**.
- Rothman (1990), for example, warns of the difficulties in defining the population base from which incidence rates can be calculated, and points out that due to high levels of publicity often surrounding such cases, **unbiased data are difficult to collect**.
- Rothman also warns that investigations into disease clusters are often **made for political**, rather than for scientific reason often, and, quoting from Oakes(1986), they are based on significance tests that lack meaningful descriptive statistics and are prone to abuse and faulty inference.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- Urquhart (1992) highlights the risks of drawing conclusions about clustering (or its absence) in a particular area **without knowledge of the underlying distribution**.
- Besag and Newell (1991) point to the assumptions behind cluster detection tests, emphasizing that the null hypothesis must be based on equal and independent risk to each individual, and that the **definition of the population at risk in terms of sex or age group, and of geographical and temporal boundaries** can have a **profound impact on the outcome and significance levels**.
- Gardner (1992) further emphasizes the risk of distorting estimates of disease excess, and of their significance levels by post hoc selection of controls.
- He aptly describes the problem as "moving the goalposts".
- Many texts have been written on the analysis of spatial point patterns. Examples include Cliff and Fingleton (1989), and Diggle (2003).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering(Cont'd)

- Reviews of some of the available methods are provided by Marshall (1991), Elliott et al. (1995), Morris and Wakefield (2000), Robinson (2000), Wakefield et al 2000), ward and Carpenter (2000) and Song and Kulldorff (2003).
- Some of the more commonly used methods are divided into those primarily designed for **aggregated data**, and those for **point data**, although this distinction is not rigid and most can be used with either type of data.
- Moreover, **point data can easily be aggregated** and **area data can be represented as points**, for example by using the centroids of the areas.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- The methods listed for point data are based on **variously defined circles that 'scan' the data for areas of elevated (or reduced) frequency**.
- Such statistics are **better suited to point than to area data, since points fall clearly inside or outside a scan circle**.
- Scan circles can be defined in terms of geographic distance (e.g. Openshaw's method), number of cases (e.g. Besag and Newell's method) or population size (e.g. Turnbull's method and Kulldorff's scan statistic).
- Cluster detection is demonstrated by applying Kulldorff's scan statistic to the British cattle TB data.
- Methods for investigating clusters around point sources are then reviewed, and finally methods that look for clusters in both space and time.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- **Unlike Moran's I** statistics, which **measures the correlation** between attribute values in adjacent areas, the ***Gi(d)*** local statistics (Getis and Ord (1992 , 1996)) is an indicator of local clustering that measures **the 'concentration'** of a spatially distributed attribute variables.
- Computational details of the of the ***Gi(d)*** statistics are provided by Ding and Fotheringham (1992), Getis and Ord (1996), and Kitron et al.(1997). The test statistic is calculated as:

$$Gi(d) = \frac{\sum_j w_{ij}(d)(x_j - \bar{x}_i)}{S\sqrt{\frac{w_i(n-1-w_i)}{n-2}}}, where\ j \neq i$$

where ***n*** is the number of areas within the region and interest

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

And X_i is the observed value of area i.

$$\bar{X}_i = \frac{1}{n-1}\sum_j x_j$$

And w_{ij} is a symmetric binary spatial weight matrix.

$$w_i = \sum_j w_{ij}$$

$$S^2{}_i = \frac{1}{n-1}\sum_j (x_j - x_i)^2$$

It has been shown that $E(G_i) = 0$ and $Var(G_i) = 1$, and that the distribution of G_i under the null hypothesis of no spatial association among x_i is approximately normal (Ord and Getis 1993).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- By comparing local estimates of spatial autocorrelation with global aver ages, the **$G_i(d)$ statistic** identifies '**hotspots**' in spatial data.
- The $G_i(d)$ statistic and a variety of other similar local statistics are discussed by Getis and Ord (1996), who also highlight some of their shortcomings, particularly those arising as a result of **small sample sizes** and the existence of high levels of global autocorrelation.
- Software to implement this statistic includes the **spdep** package in **R**, and **Clusterseer.**
- Getis and Ord (1996) use $G_i(d)$ the to explore the spatial pattern of sudden infant death syndrome (SIDS) by county in North Carolina between 1979 and 1984.
- In addition to a number of small clusters, they identify a major hotspot in the mid-south of the state they attribute to the distribution of health care facilities (Getis and ord 1996).

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- Kitron et al (1997) use the $G_i(d)$ statistic to identify significant clustering of **cases of La Crosse encephalitis** over distances of **5 and 10 km**, around particular towns in the Peoria area of Illinois.
- This analysis reveals a number of hotspots which they associate with homes on the edge of gullies and ravines where breeding and biting activities of the vector (*Aedes triseriatus*) are likely to occur.
- Dining and Fotheringham (1992) developed **the statistical analysis module (SAM) for ARC/I** based on the $G_i(d)$ local statistic.
- They use SAM to explore clustering of **population growth in China between 1982 and 1990** and show that **high rates of population growth are concentrated in the south**.
- They attribute this to less severe restrictions on family size and greater proportions of ethnic minorities with large families in this region.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- Moving towards exploring spatio-temporal clustering, Getis and Ord (1996) propose the use of local statistics to quantify the pattern and intensity of spread of a disease away from the core of a hotspot by estimating a series of local statistics at different time periods.
- Local statistics can be used to estimate the intensity of a disease at various distances from a core location, and the time dimension can then be used to estimate the rate of spread.
- Getis and Ord 01998) use the $G_i(d)$ statistic in this way to trace the spread of AIDS away from San Francisco.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York

3. Local Clustering (Cont'd)

Local Moran test

- The local Moran test (Anselin 1995) detects **local spatial autocorrelation** in aggregated data by **decomposing Moran's I statistic** into contributions for each area within a study region.
- Termed **Local Indicators of Spatial Association (LISA)**, the LISA statistic for each area is calculated as:

$$I = Z_i \sum_j w_{ij} Z_j$$

where Z_i and Z_j are the observed values in **standardized form**, and w_{ij} is a spatial weights matrix in row-standardized form.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York

3. Local Clustering (Cont'd)

- These indicators **detect clusters** of either **similar** or **dissimilar** frequency values around a given observation.
- The sum of the LISAs for all observations is proportional to the global Moran's I statistic.
- There are **two uses** of LISA statistics: either as indicators of **local autocorrelation** or as tests for **outliers in global spatial patterns** in the form of a Moran scatter plot (Anselin 1995; 1996).
- In a Moran scatter plot the horizontal axis represents the vector of observed values and the vertical axis specifies the weighted average of neighbouring values.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York

3. Local Clustering (Cont'd)

- The extent of the 'mix' of pairs among **the four types of association** in the quadrants of the plot defined by the axes (**low-high, high-high, high-low and low-low**) provides an indication of the stability of the spatial association throughout the data.
- It may also suggest the existence of different types of association in different subsets of the data; for example, positive association in one area and negative association in another.
- Software for implementing a local Moran test includes Clusterseer, GeoDa, SpaceStatlo and the spdep package in R.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- Anselin (1995) demonstrates the local Moran test using patterns of conflict in Africa and compares it with Getis and Ord's local $G_i(d)$ statistic (Getis and ord 1992).
- Anselin (1995) also developed LISAs for Mantel's test (Mantel 1967) and Geary's c (Geary 1954)
- The Clusterseer and GeoDa software can be used to implement the LISA for Moran's I.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York

3. Local Clustering (Cont'd)

- Burra et al. (2002) compare the ability of the local Moran test and Getis and Ord's local $G_i(d)$ statistic to detect clustering in mortality data from Hamilton, Ontario.
- These authors also evaluate the effect of geocoding errors on pattern detection, concluding that **small geocoding errors can significantly influence the results of, and therefore the conclusions drawn from, these statistical tests**.
- Their analysis confirms the clustering of breast cancer mortality previously identified by Kulldorff et al. (1997) but they find that the two methods identify slightly different cluster locations.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- Jacquez and Greiling (2003) use the **local Moran statistic** to analyse spatial clustering in diagnoses of **breast lung**, and **colorectal cancers** on Long Island, USA, identifying significant **spatial patterns** for all three diseases.
- As a result of these differences, Jacquez and Greiling (2003) recommend that **a combination of statistics be used when studying local clustering to ensure that different aspects** of spatial patterns are identified and to check whether the result from different analyses are consistent.

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- In the similar vein, Hanson and Wieczorek (2002) compare **the local Moran statistic** with **Kulldorf's spatial scan statistic** exploring alcohol-related mort in New York, USA.
- The clusters identified differ somewhat between the two methods and the conclude that the **scan statistic is the more sensitive** method.
- They advocate a **multi-method approach to cluster detection** since different methods tend to identify different characteristics of the clusters, the LISA identifying the core of a cluster and the scan statistic identifying its extent

Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

3. Local Clustering (Cont'd)

- A similar comparison is made by Nodtvedt et al. (2007) in an analysis of the spatial distribution of cases of atopic dermatitis in dogs in Sweden (1995-2000).
- They conclude that the tests produce similar results but that the **LISA statistic identifies more localized clusters** compared to the larger clusters identified by the spatial scan statistic.

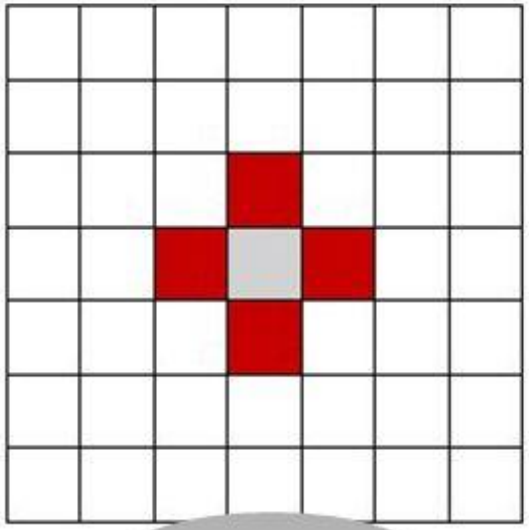
Source: Directly quoted from Dirk U.P., Timothy P.R., Mark S., Kim B.S., David J.R., and Clements A.(2008), *Spatial Analysis in Epidemiology*, Oxford University Press, New York.

Measuring Contiguity: *Lagged Contiguity*

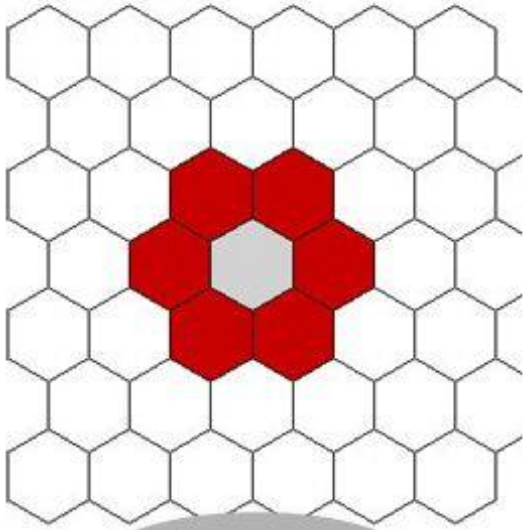
Should we include second order contiguity?

1st order

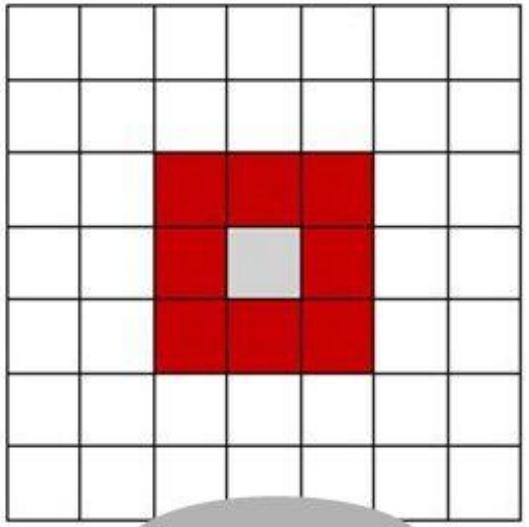
Nearest neighbor



rook



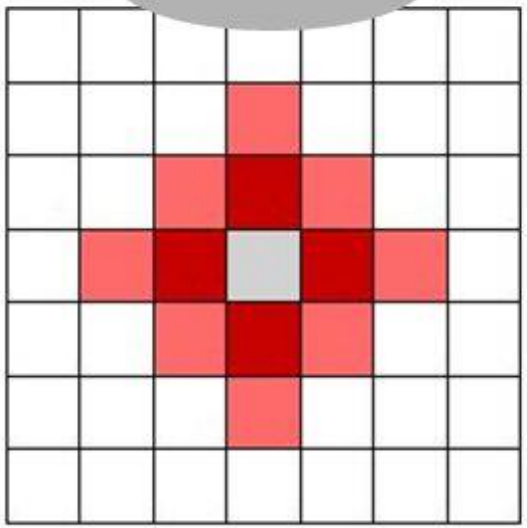
hexagon

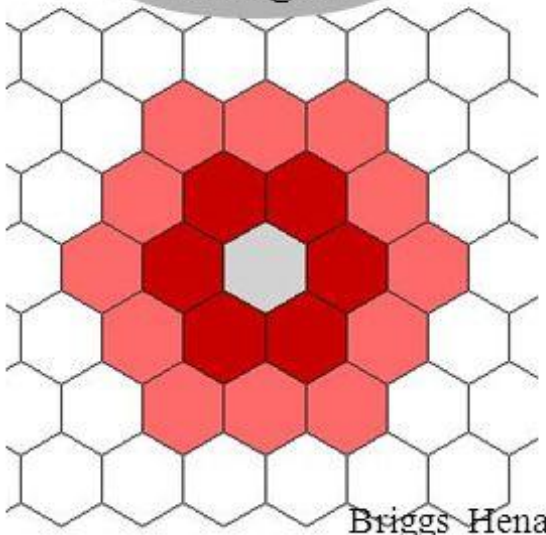


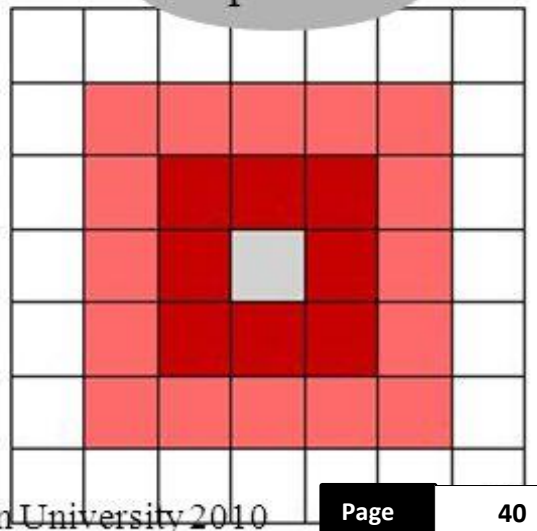
queen

2nd order

Next nearest neighbor



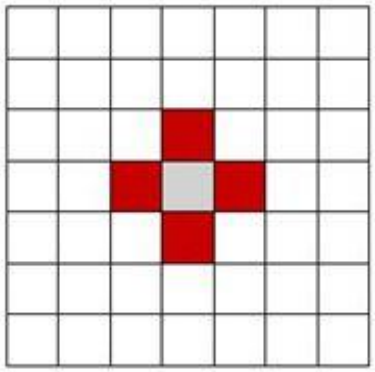




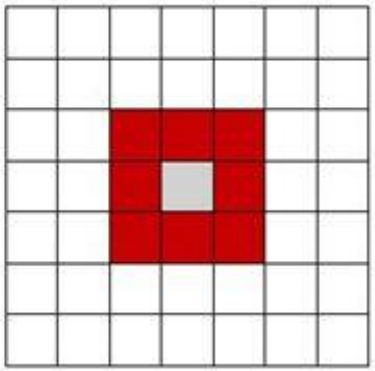
Spatial neighbors based on contiguity* (adjacency)

* Shares common border

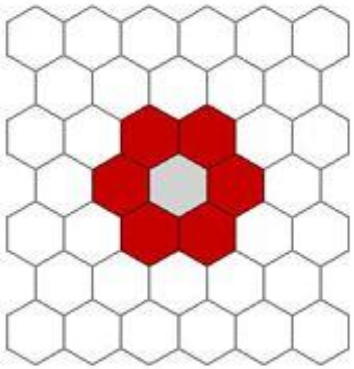
- Sharing a border or boundary
  - Rook: sharing a border
  - Queen: sharing a border or a point



rook



queen



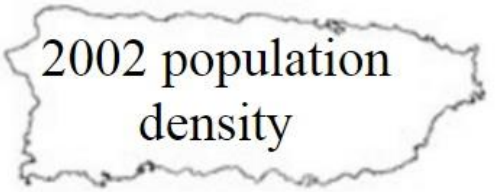
Hexagons

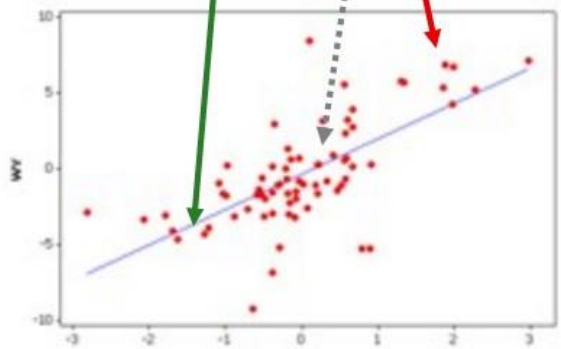
Irregular

Which use?

Positive spatial autocorrelation

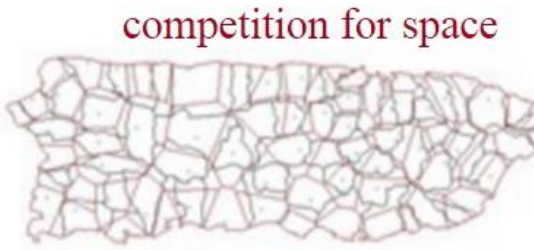
- high values surrounded by nearby high values
- intermediate values surrounded by nearby intermediate values
- low values surrounded by nearby low values



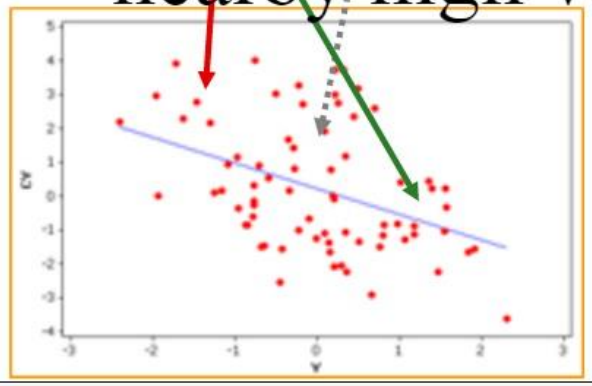


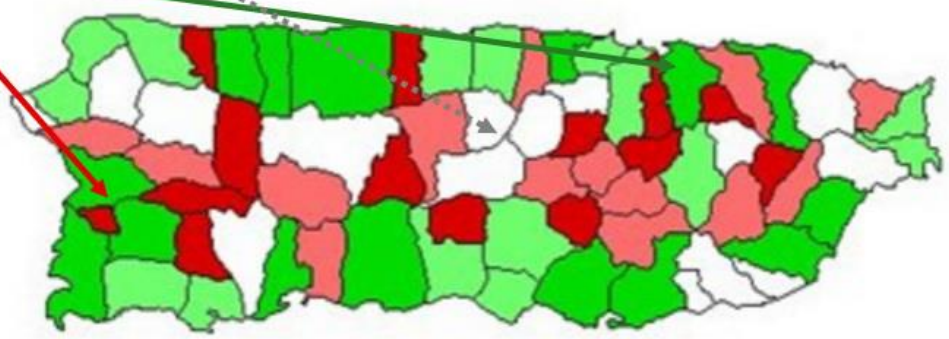
Negative spatial autocorrelation

- high values surrounded by nearby low values
- intermediate values surrounded by nearby intermediate values
- low values surrounded by nearby high values



Grocery store density

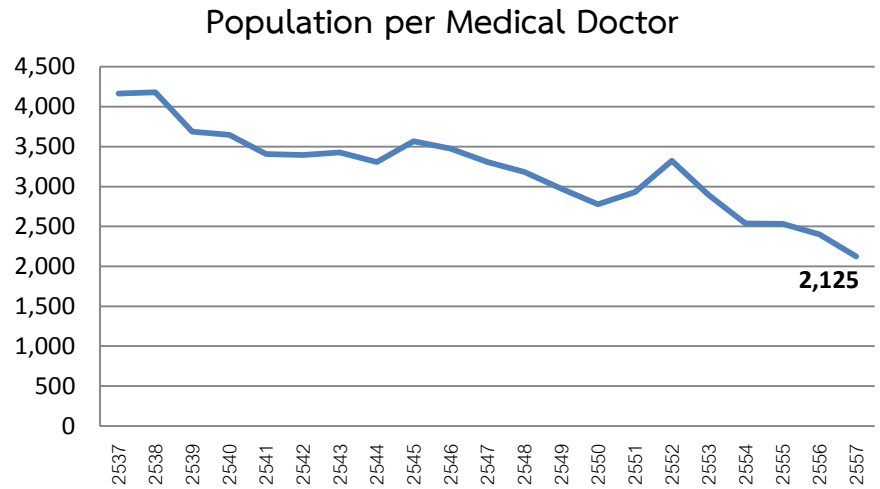


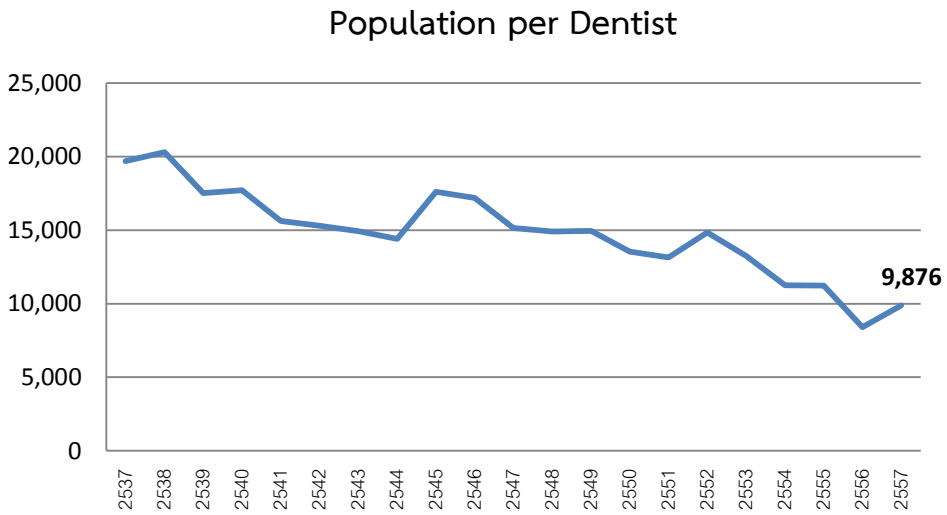


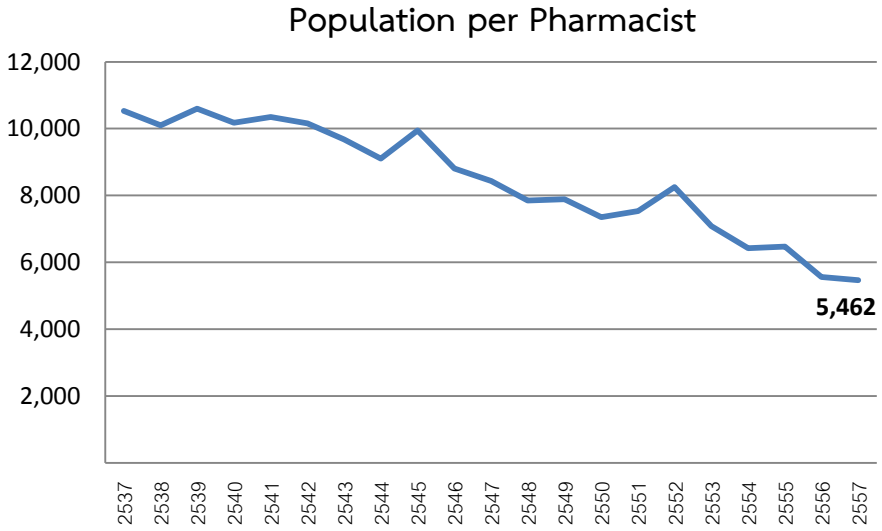
Example

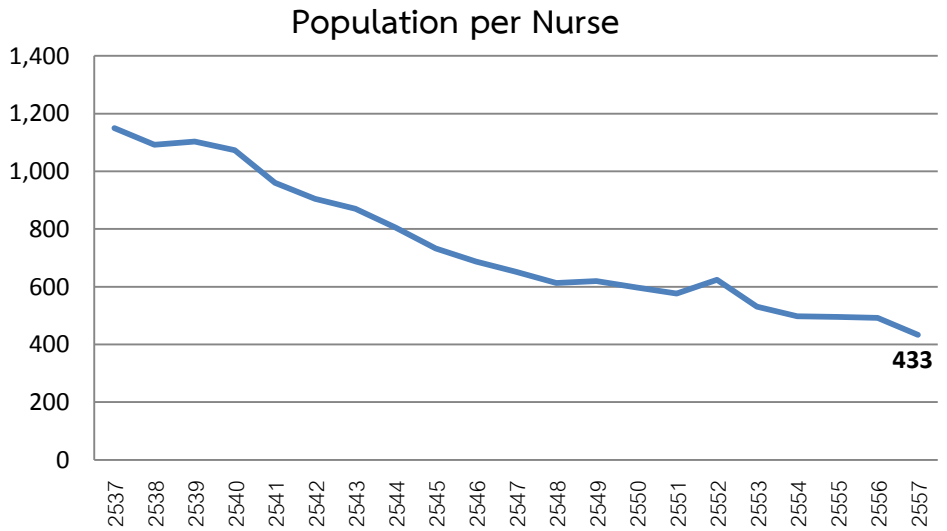
Public Health : Spatial Distribution of Human Resources in Thailand

Public Health – Human Resources in Thailand



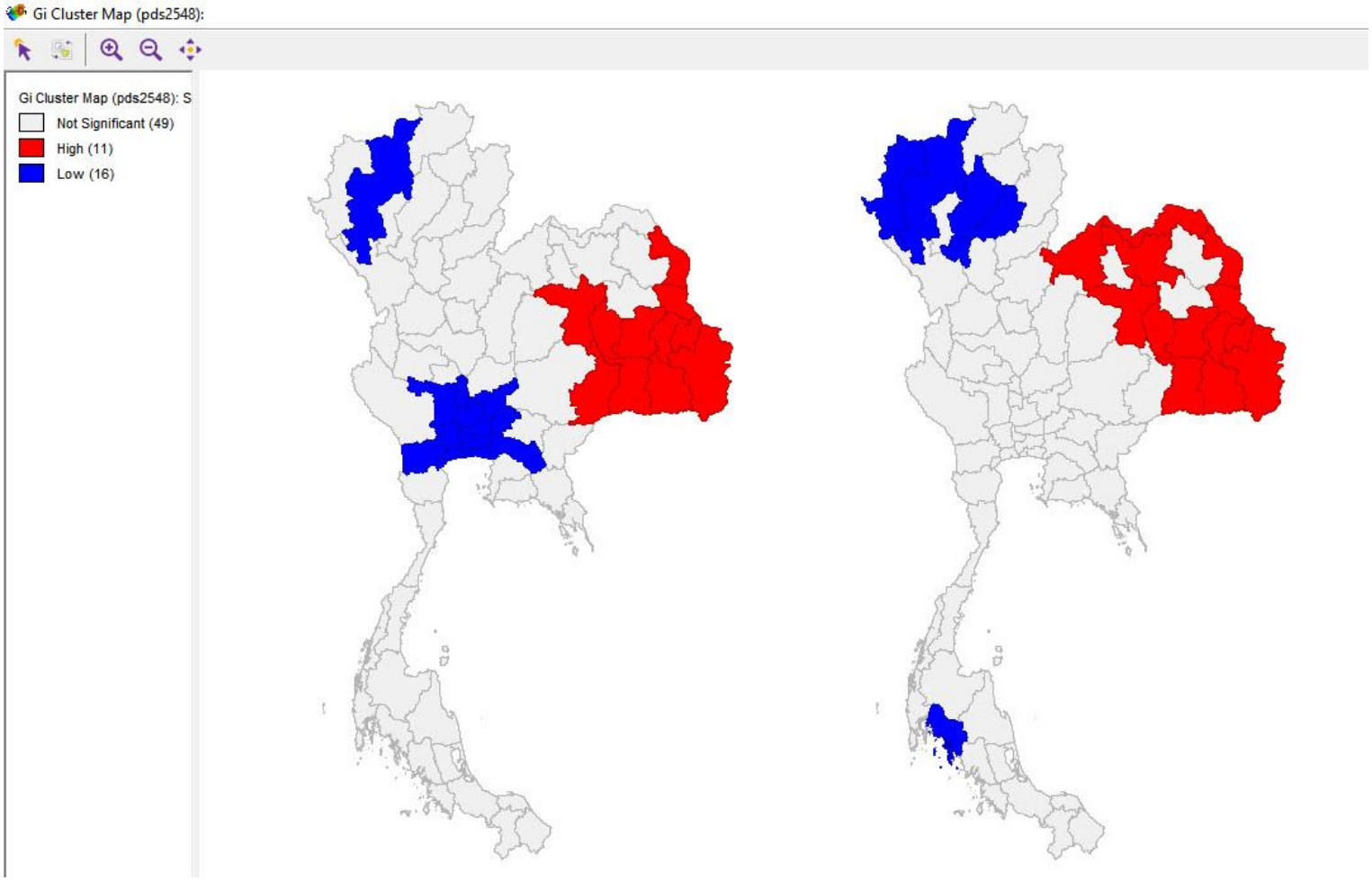




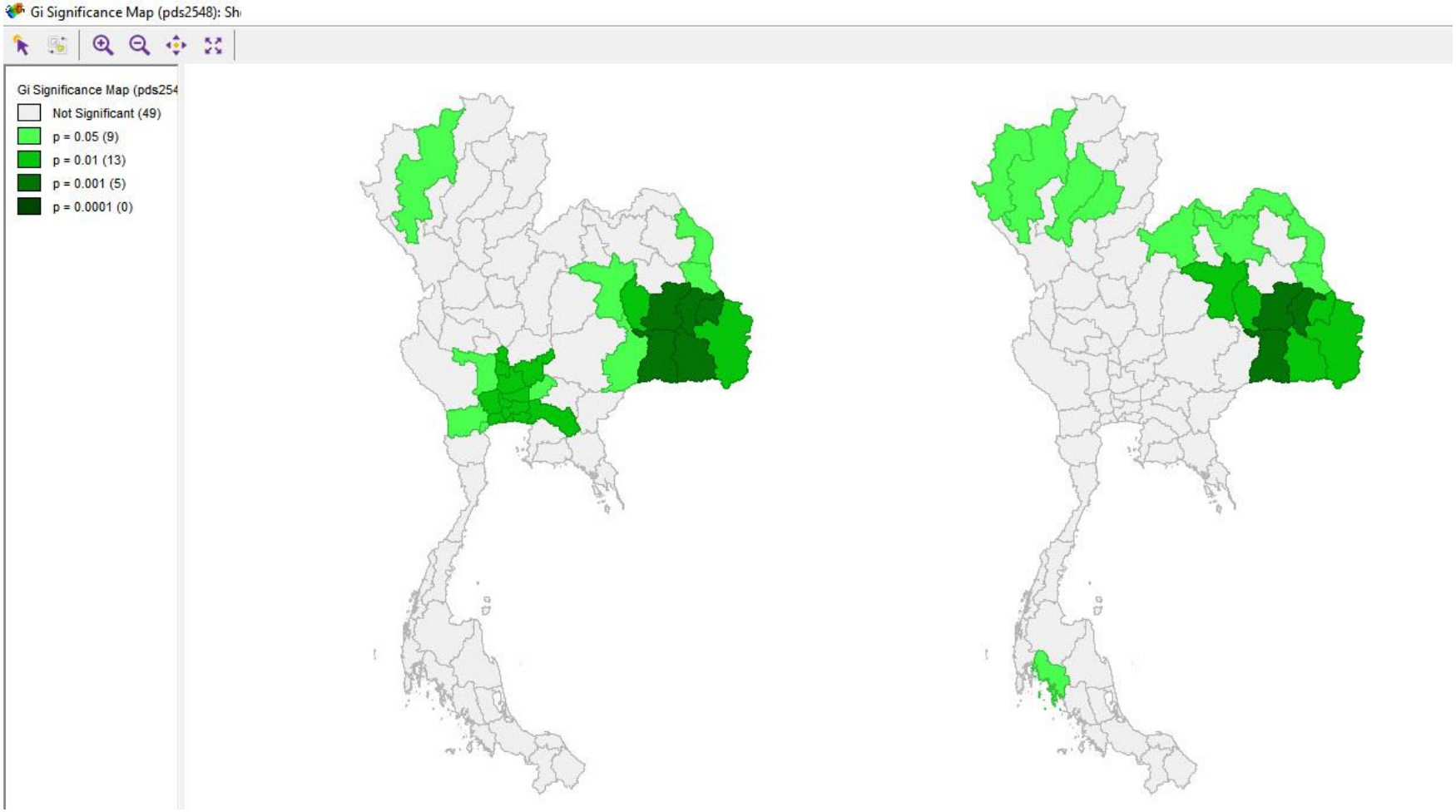


Univariate Analysis

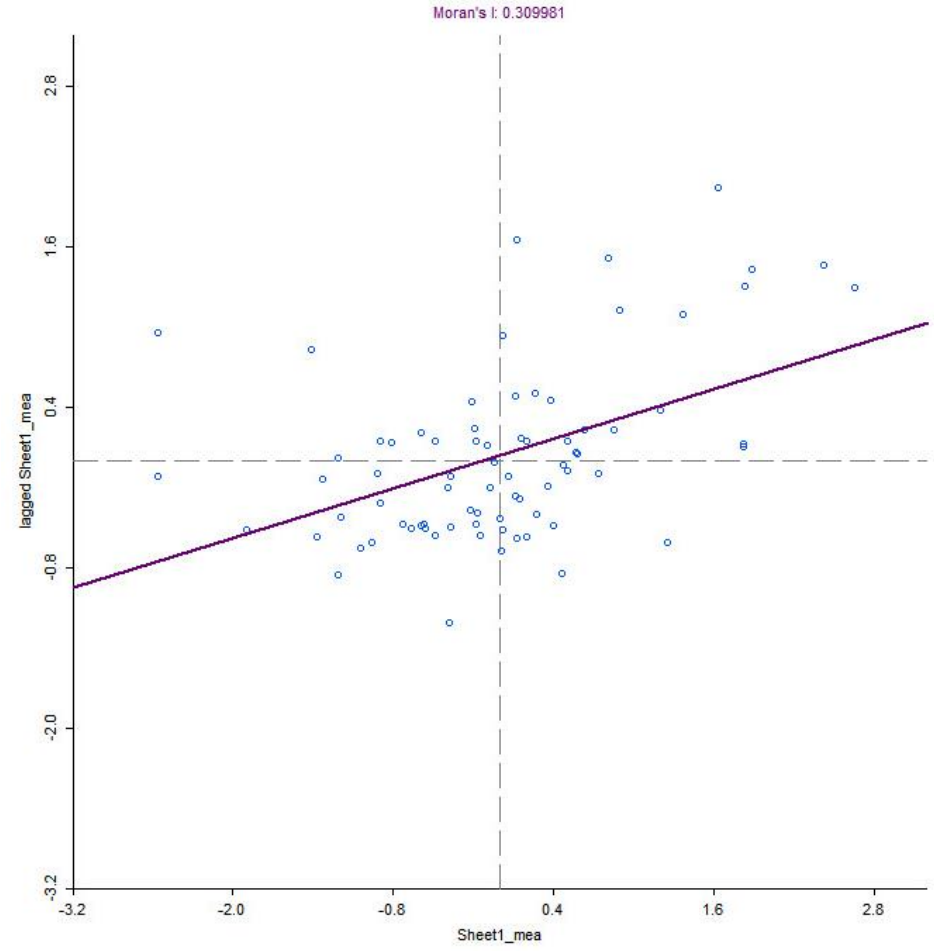
Gi cluster pds2548 VS pds2557

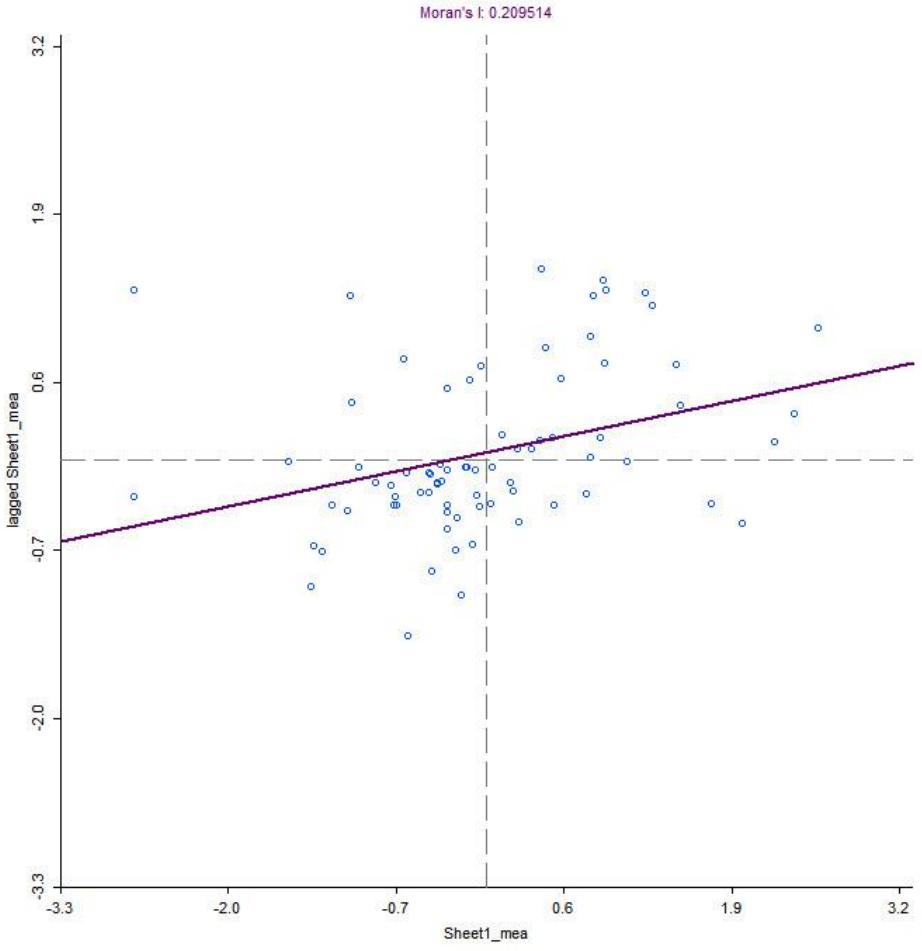


Gi cluster pds2548 VS pds2557

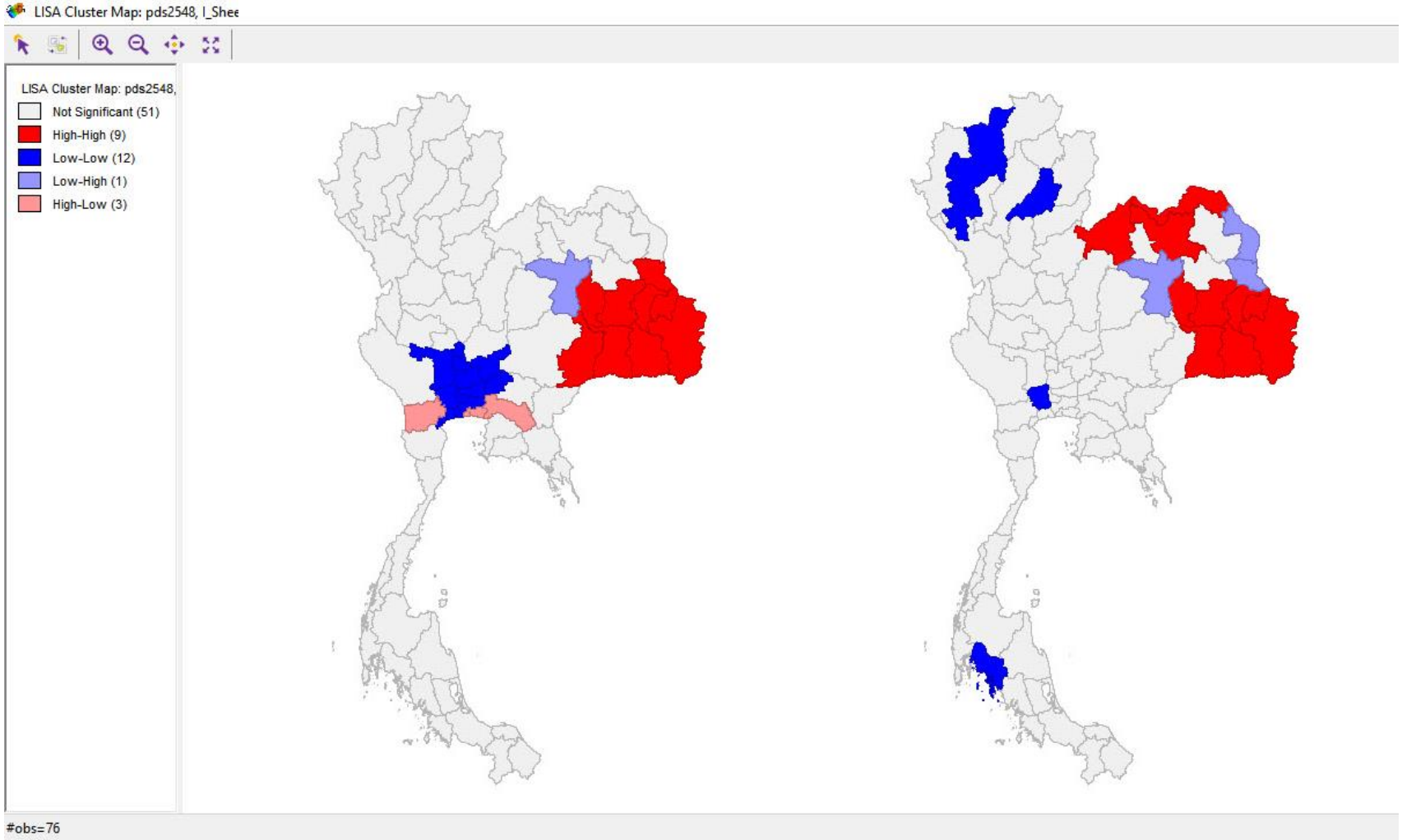


LISA pds2005 VS pds2013



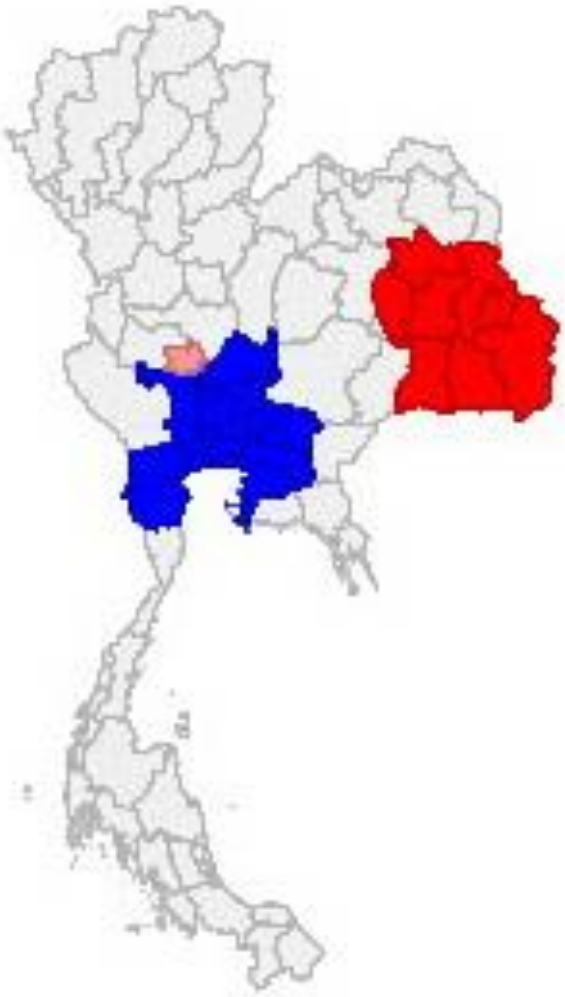


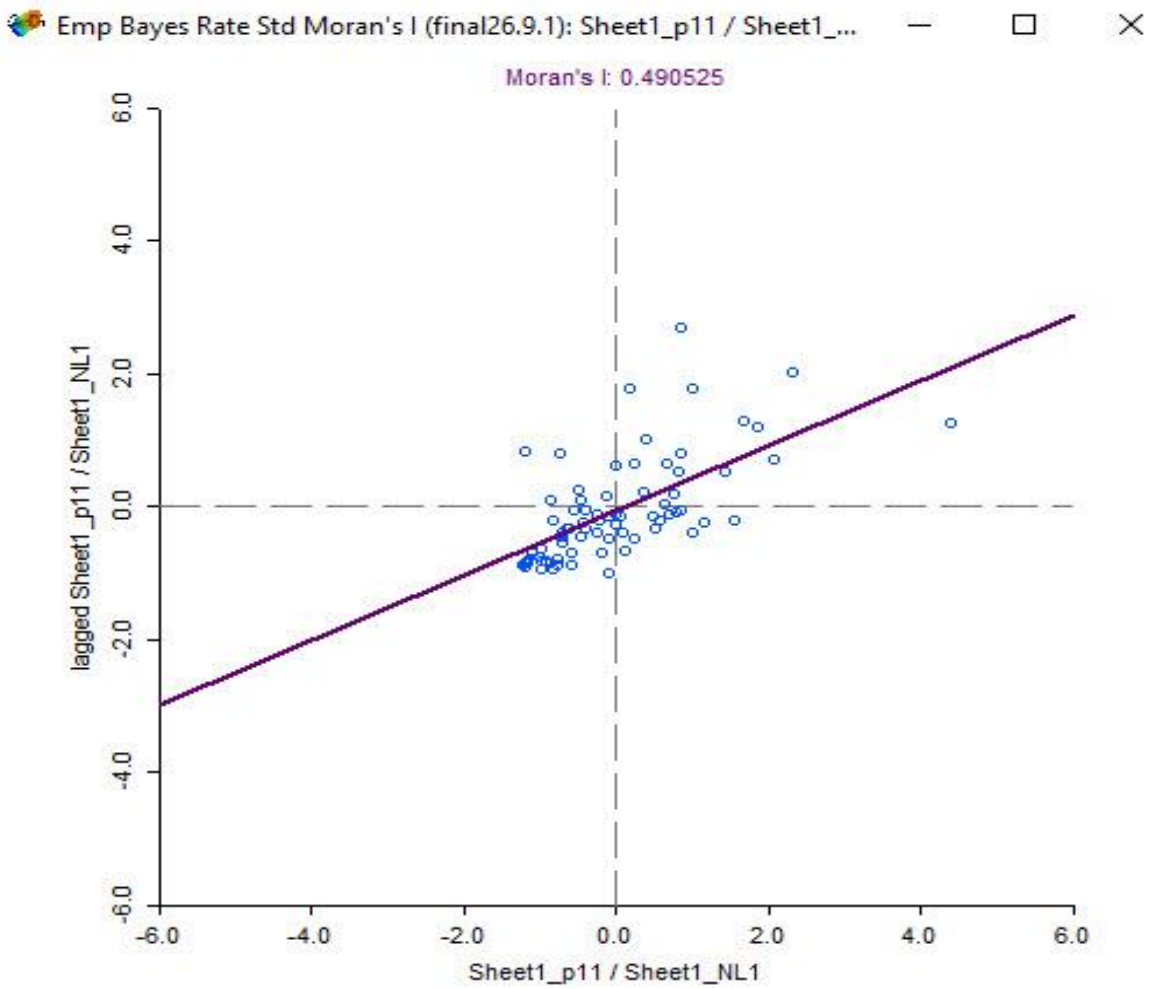
LISA pds2005 VS pds2013



Bivariate Analysis

Medical Doctor per Population, computed LISA with NightLight Index





Nurse per Population, computed LISA with NightLight Index