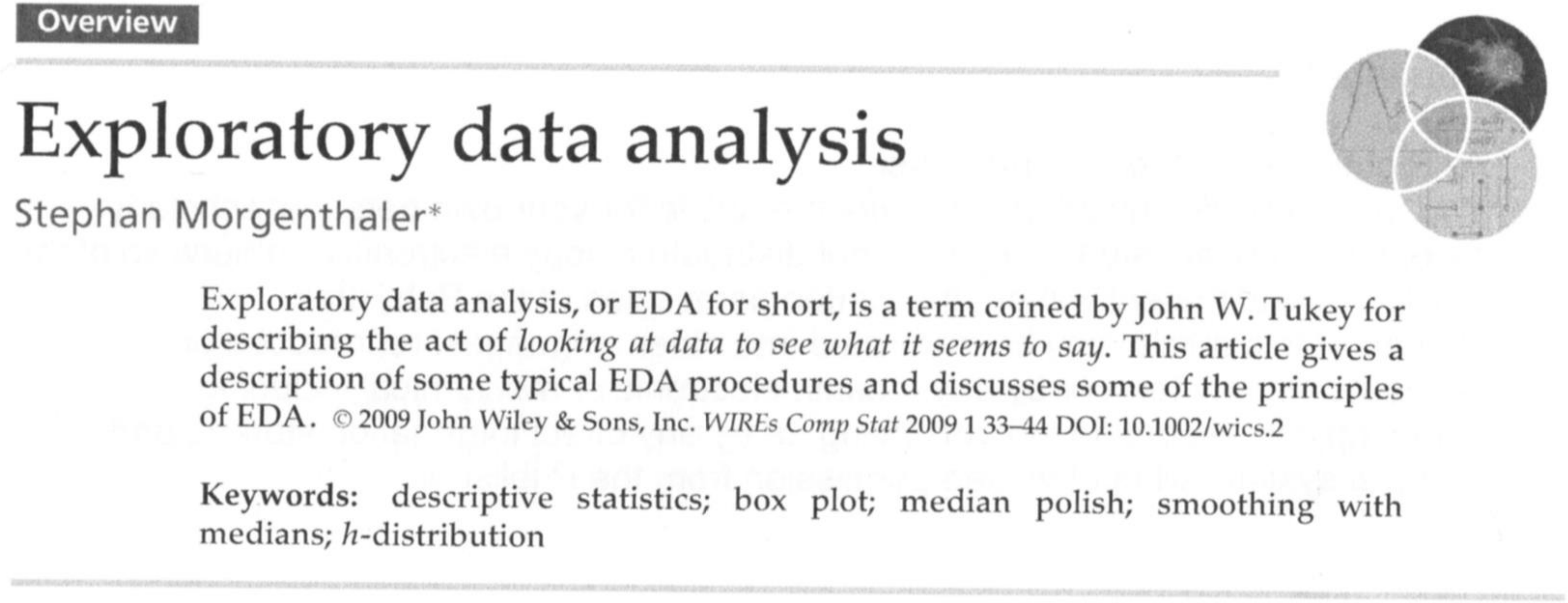
Exploratory data analysis in environmental health

Stéphane Joost & Mayssam Nehme

Solution Exercise 1 a&b

Two important papers on exploratory data analysis and Geoda

1. Exploratory data analysis, S.Morgenthaler



Overview

Exploratory data analysis

Stephan Morgenthaler*

Exploratory data analysis, or EDA for short, is a term coined by John W. Tukey for describing the act of *looking at data to see what it seems to say*. This article gives a description of some typical EDA procedures and discusses some of the principles of EDA. © 2009 John Wiley & Sons, Inc. *WIREs Comp Stat* 2009 1 33–44 DOI: 10.1002/wics.2

Keywords: descriptive statistics; box plot; median polish; smoothing with medians; *h*-distribution

A. How does EDA differ from the more traditional approaches to statistics (i.e. inferential or confirmatory methods) in terms of their tools, underlying assumptions/restrictions or objectives?

Classical statistical analysis

- Applies a specific model
- Derives procedures that are appropriate for the data
- Aim: confirm the presence of a particular effect

Exploratory data analysis

- No need to consider a model
- No restrictions on procedures, guided by the analyst’s intuitions and curiosity
- Based on graphical visualization
- Aim : find interesting indications about a dataset
- Multiple answers are possible!

A. More text

- Classic statistical analyses investigate data to elaborate a model and make several procedures available to compute values referring to the model like the standard deviation, and the confidence interval.
- Classic statistics are about randomness, stochastic models and population parameters, and this leads to questions related to the precision of estimates or to the significance of a finding. It is guided by the will to confirm the presence of a particular effect and it is supported by a statistical model translated by a mathematical expression.
- Exploratory Data Analysis has no need to consider a model. Each analyst is free to choose any procedure he likes to analyze the data.
- EDA puts more emphasis on data exploration using approaches which are deemed to be appropriate. Graphical visualization (histogram, normal quantile plot, box plot, etc.) is one of the most used process to analyze the data. Since EDA imposes no restriction on the methods to be applied, a model of reference is not required and no control of the validity of a finding is necessary.
- In a confirmatory analysis, a strict analytical procedure has to be applied. In fact there is no clear boundary between EDA and classic statistics, but the aim of EDA is to find interesting indications or relationships in the data. For any given dataset, multiple answers or hypotheses will be highlighted by EDA and then in turn confirmatory statistics will be applied to validate one or several of them.

B. In EDA, is it preferable to use the median or mean as a primary summary statistic? Why?

The median !

- Much more resistant to outliers
- More stable
- More representative of the variable

The median is preferable than the mean, because it is much more robust to abnormal, unusual values, or outlier values. The median is also more stable and more representative of the variable under study. Such exceptional values can rarely be avoided, even when the sampling is undertaken in a rigorous way.

C. In parametric statistics (e.g. t-test, ANOVA) data are sometimes transformed so that they approximate a normal distribution. What are the advantages of transforming (or re-expressing) the data prior to EDA?

- Most real-life data are not symmetric !
- Many advantages:
  - Handle skewed data -> makes it closer to a normal distribution (normality assumption required by many models)
  - Decreases the effect of the outliers (whiskers of more equal length)
  - We transform exponential trends into linear trends
  - Log transform is the most used but other transformations can be used: square root, square, inverse of the square root…
  - If 0 are present: one can use log(x+1)

C. More text

- Most of datasets are not symmetric, and thus it is frequent that a transformation has to be applied. This type of re-expression of the data may have many advantages. If the distribution of numbers is positive, one can transform it using a logarithmic scale. This will usually make the distribution normal and symmetric, centering the median and making the two whiskers of a boxplot of equal length – more or less.
- Other functions than the logarithm can be used to transform a distribution: the square, the square root or the inverse of the square root.

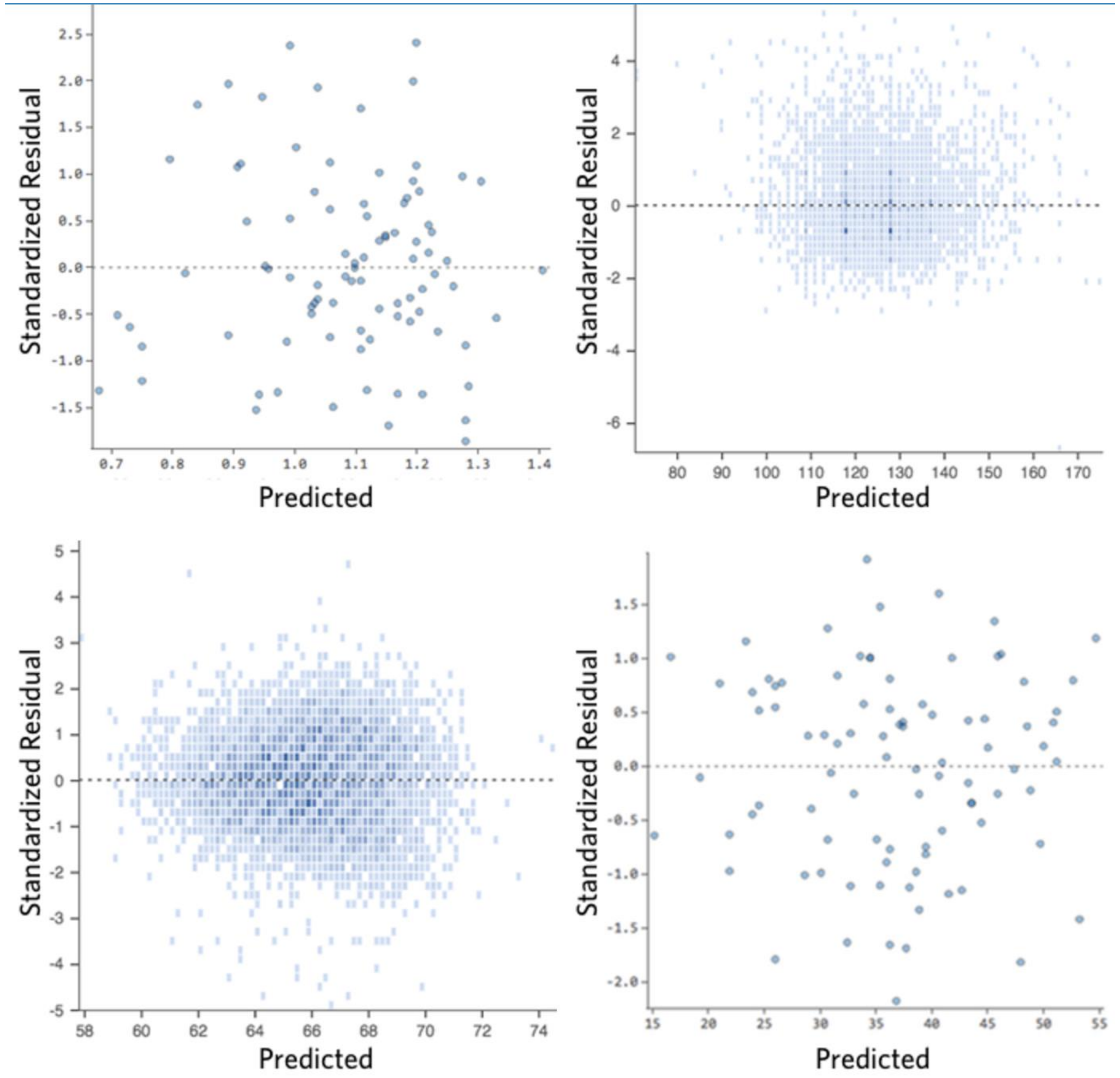
D. What sort of analysis can you perform to explore the residuals, what can this tell you?

- Boxplot of the data against the residuals.
  - If the model works well, the H-spread (Q3-Q1) of the residuals should be smaller than the one of the models.
  - The residuals should be symmetrically spread around 0, otherwise it indicates a tendency that the model cannot follow.
- Plot the residuals. It helps to understand the importance of each effect.

D. More text

Often we have to look at the residuals in order to find further indications and to improve the understanding of the data. A typical way of doing is to compare the boxplot of the data with the boxplot of the residuals. If the model works well, the interquartile range or H-spread of the residuals should be smaller.
When you plot the residuals, they should also be symmetrically spread around 0 (normal distribution of the residuals), otherwise it will indicate a tendency that the model is not able to follow.

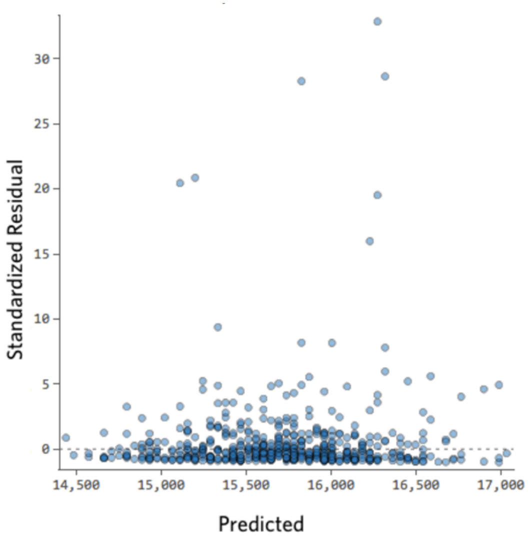
Ideally your plot of the residuals will have to look like one of these

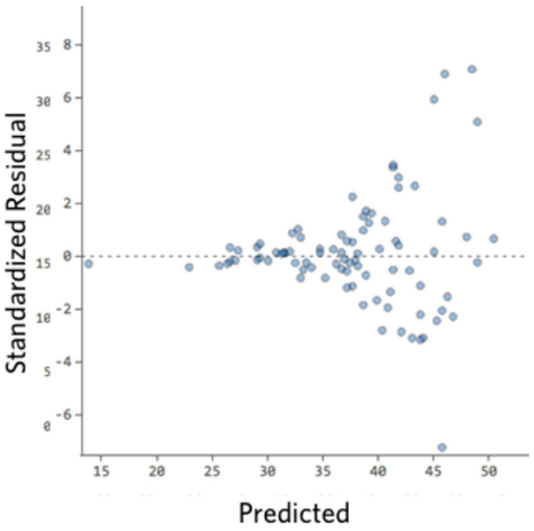


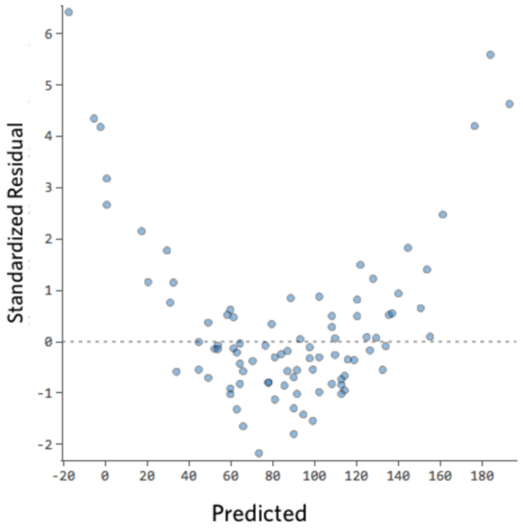
(1) they’re pretty symmetrically distributed, tending to cluster towards the middle of the plot.
(2) they’re clustered around the lower values of the y-axis (e.g., 0.5 or 1.5, not e.h. 200).
(3) In general, there is no clear pattern.

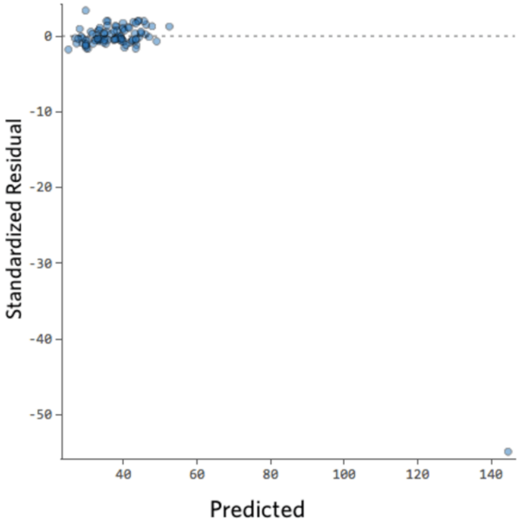
These analyses will improve understanding the data and can give clues to calculate a more suitable model.

The residual plots hereunder do not meet those requirements

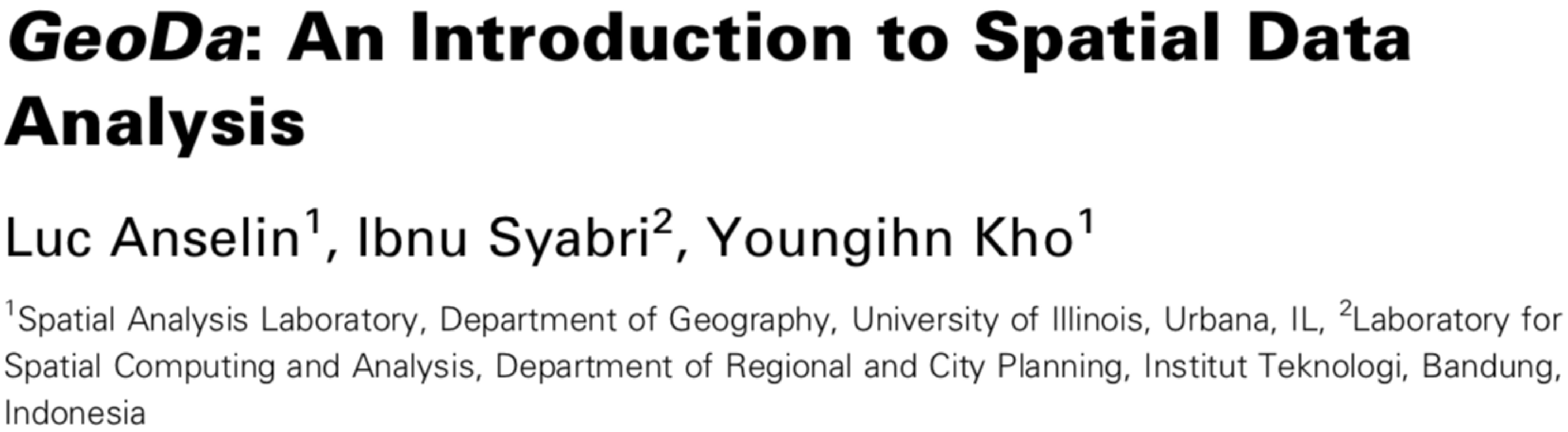








2. GeoDa: An introduction to spatial data analysis

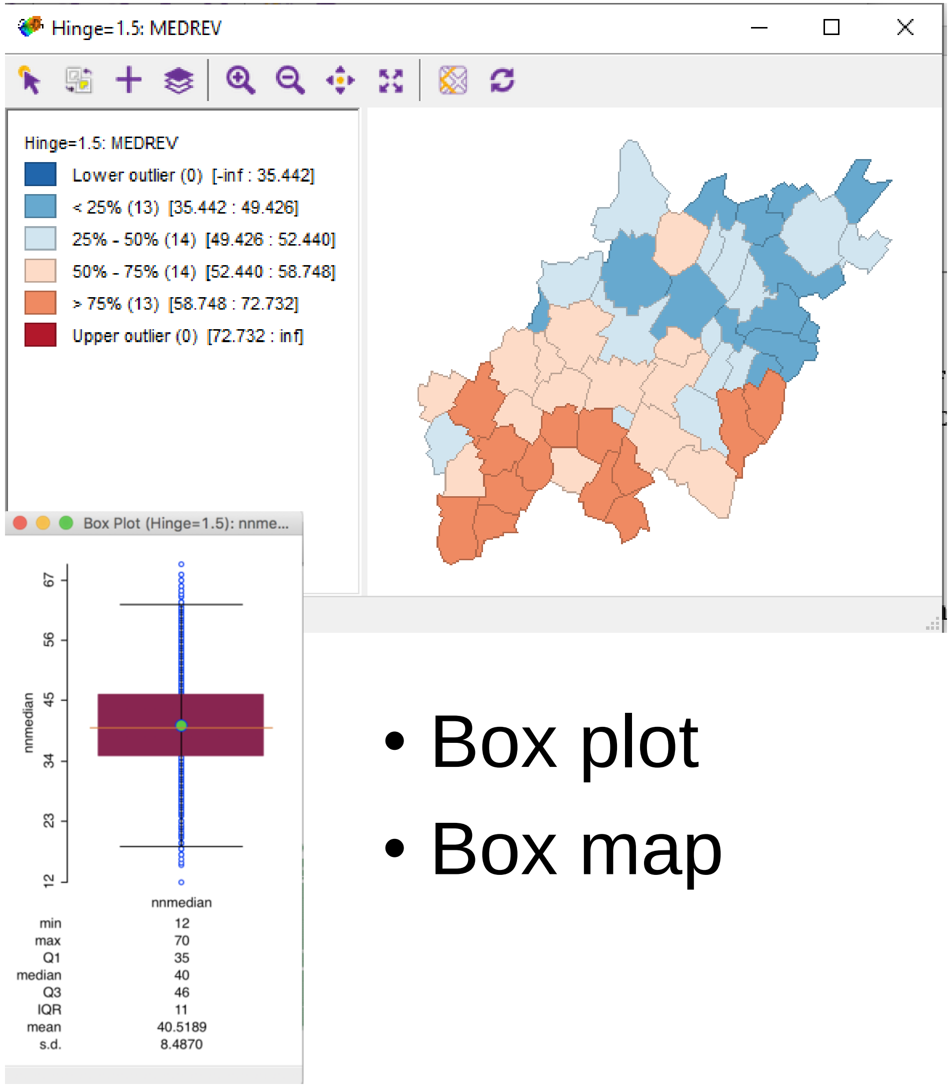


GeoDa: An Introduction to Spatial Data Analysis

Luc Anselin[1], Ibnu Syabri[2], Youngihn Kho[1]

[1]Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana, IL, [2]Laboratory for Spatial Computing and Analysis, Department of Regional and City Planning, Institut Teknologi, Bandung, Indonesia

A. After reading Anselin (2006), which statistical method(s) would you use for (spatial) outlier detection?



Hinge=1.5: MEDREV

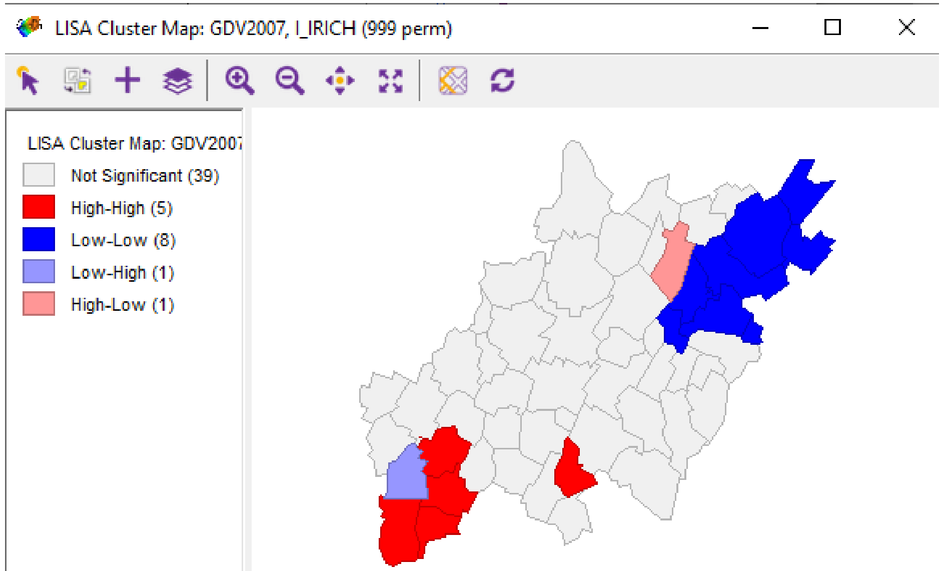
Hinge=1.5: MEDREV

🟦	Lower outlier (0) [-inf : 35.442]
🟦	< 25% (13) [35.442 : 49.426]
🟦	25% - 50% (14) [49.426 : 52.440]
🟧	50% - 75% (14) [52.440 : 58.748]
🟧	> 75% (13) [58.748 : 72.732]
🟥	Upper outlier (0) [72.732 : inf]

Box Plot (Hinge=1.5): nnme...

	nnmedian
min	12
max	70
Q1	35
median	40
Q3	46
IQR	11
mean	40.5189
s.d.	8.4870

- Box plot
- Box map



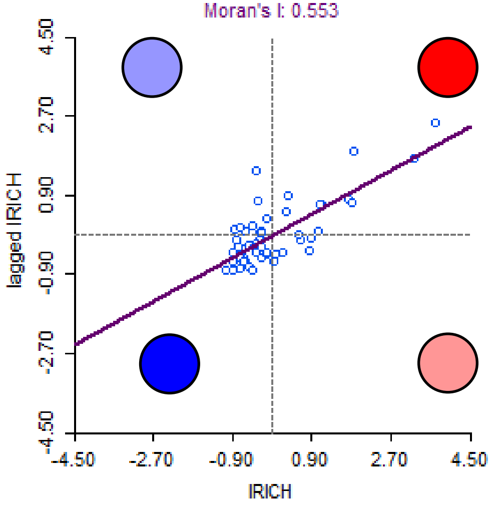
LISA Cluster Map: GDV2007, I_IRICH (999 perm)

LISA Cluster Map: GDV2007

⬜	Not Significant (39)
🟥	High-High (5)
🟦	Low-Low (8)
🟪	Low-High (1)
🟥	High-Low (1)

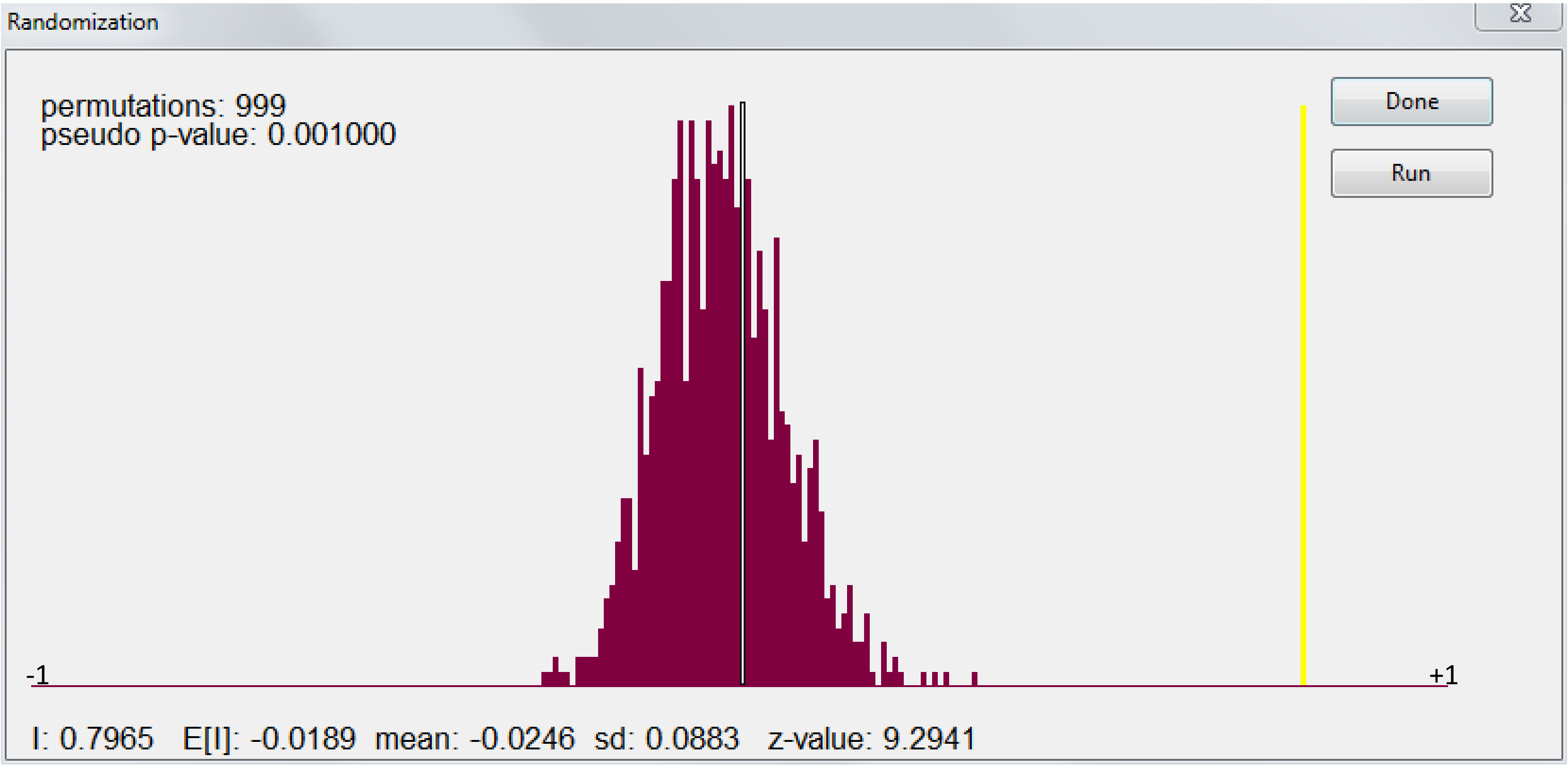
#obs=54 p <= 0.05

- Spatial statistics (local Moran’s I)



B. Which options does GeoDa perform to assess the sensitivity/significance of Local Moran’s I results?

GeoDa assesses the significance of the Moran’s I statistic through a permutation test (up to 9999 permutations).



permutations: 999
pseudo p-value: 0.001000

I: 0.7965 E[I]: -0.0189 mean: -0.0246 sd: 0.0883 z-value: 9.2941

C. "Everything is related to everything else, but near things are more related than distant things." (Tobler's First Law of Geography). Which method(s) presented in Anselin (2006) do take into account this observation? How?

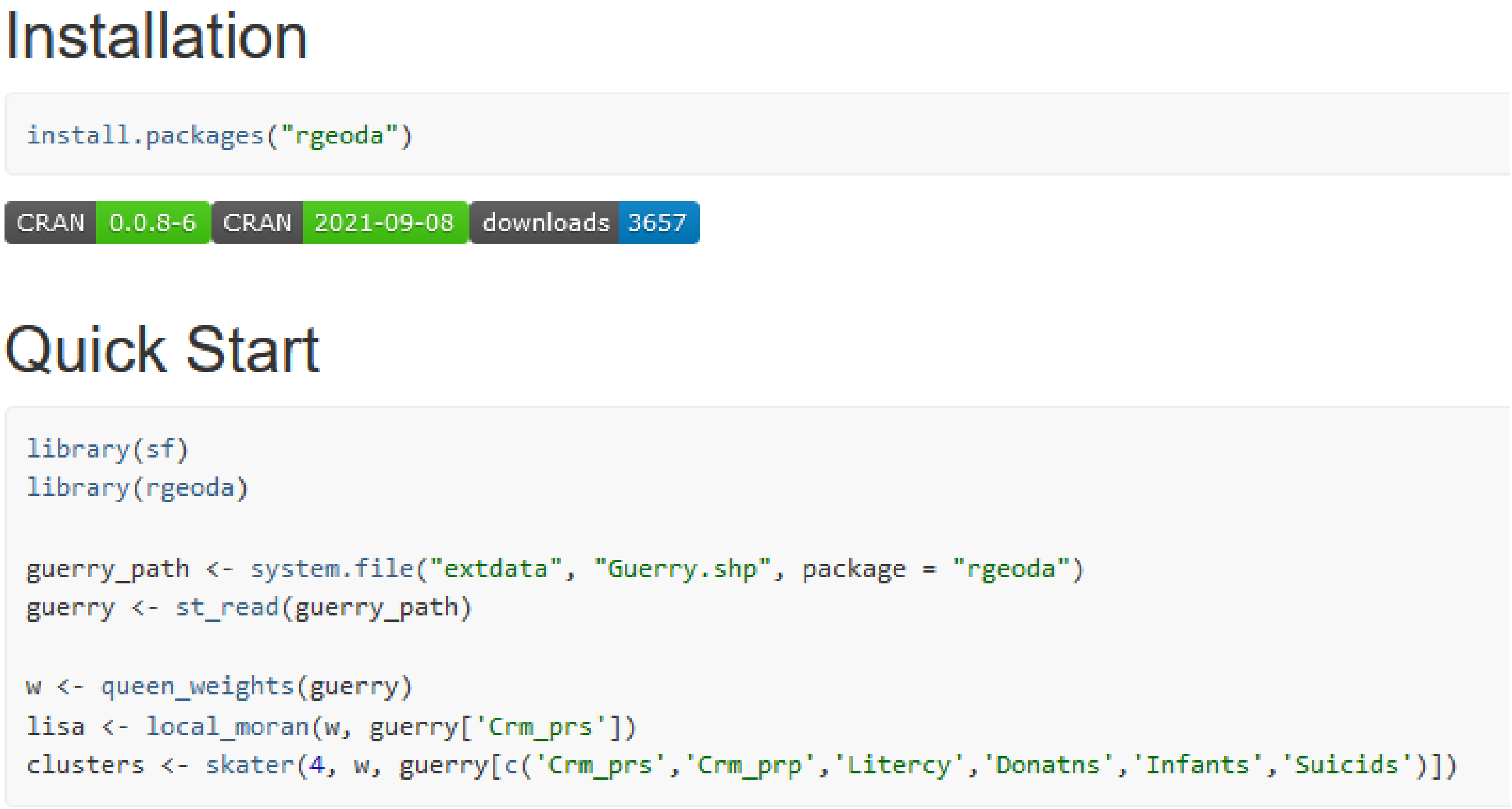
- Spatial autocorrelation analysis
  - Quantify the relatedness between neighboring objects.
- Spatial regression
  - Add a variable in the regression fit corresponding to the value of the neighboring units
- Spatial smoother
  - Takes advantage of this concept by incorporating the values of the neighboring objects in the measurement of the object’s value.

D. Why would you use GeoDa instead of R for spatial data analysis?

GeoDa vs R (or Python)

- Pros :
  - User friendly : Doesn’t require programming skills (Point and click interface)
  - Linking and brushing tools allows rapid identification/exploration of specific regions across all graphs
- Cons:
  - Not open source: no customization (users rely on updates)
  - Very large datasets (from N=30’000~) can cause slow behavior

Rgeoda



Installation

```
install.packages("rgeoda")
```

CRAN 0.0.8-6 CRAN 2021-09-08 downloads 3657

Quick Start

```
library(sf)
library(rgeoda)

guerry_path <- system.file("extdata", "Guerry.shp", package = "rgeoda")
guerry <- st_read(guerry_path)

w <- queen_weights(guerry)
lisa <- local_moran(w, guerry['Crm_prs'])
clusters <- skater(4, w, guerry[c('Crm_prs','Crm_prp','Litercy','Donatns','Infants','Suicids')])
```

https://geodacenter.github.io/rgeoda/articles/rgeoda_tutorial.html