UC Santa Barbara

Core Curriculum-Geographic Information Science (1997-2000)

Title

Unit 128 - Exploratory Spatial Data Analysis

Permalink

https://escholarship.org/uc/item/82w9m1rb

Authors

128, CC in GIScience
Haining, Robert
Wise, Stephen

Publication Date

2000

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Unit 128 - Exploratory Spatial Data Analysis

Written by : Robert Haining and Stephen Wise
The software used for the illustrations was written by Jingsheng Ma.
Department of Geography and Sheffield Centre for Geographic Information and Spatial Analysis
The University of Sheffield, S10 2TN, England.

This unit was edited by C. Peter Keller, Department of Geography, University of Victoria, Canada.

This unit is part of the *NCGIA Core Curriculum in Geographic Information Science*. These materials may be used for study, research, and education, but please credit the authors Robert Haining and Stephen Wise, and the project, *NCGIA Core Curriculum in GIScience*. All commercial rights reserved. Copyright 1997 by Haining and Wise.

Advanced Organizer

Topics and Intended Learning Outcomes

By the end of this lecture students can expect to

- Understand the purpose of EDA and ESDA
- Understand the model which underlies statistical ESDA
- Appreciate the importance of cartographical, graphical and tabular tools in processing geographical data for ESDA
- Understand a range of ESDA techniques and their use in analysing geographical data
- Appreciate the limitations of current GIS software for undertaking ESDA

Full Table of Contents

Metadata and Revision History

Exploratory Spatial Data Analysis

1. Introduction

What is exploratory data analysis (EDA)?

- aim is to identify data properties for purposes of
  - pattern detection in data
  - hypothesis formulation from data
  - some aspects of model assessment (e.g.goodness of fit, identifying data effects on model fit).
- based on the use of graphical and visual methods and the use of numerical techniques that are statistically robust i.e. not much affected by extreme or atypical data values. Hence use of median

  rather than mean for measure of 'location' - this is a term used in the EDA literature rather than
  'average' or 'central tendency'. Unfortunately it has the potential to cause confusion when dealing
  with spatial data.
- Emphasis on descriptive methods rather than formal hypothesis testing.
- Importance of "staying close to the original data" in the sense of using simple, intuitive methods.

What is exploratory spatial data analysis (ESDA)?:

- extension of EDA to detect **spatial** properties of data. Need additional techniques to thosefound in EDA for
  - detecting spatial patterns in data
  - formulating hypotheses based on the geography of the data
  - assessing spatial models.
- important to be able to link numerical and graphical procedures with the map - need to be able to

  answer the question "where are those cases on the map?". With modern graphical interfaces this is often done by 'brushing' - for example cases are identified by brushing the relevant part of a
  boxplot, and the related regions are identified on the map.
- This lecture only covers the case of attribute data attached to irregular areal units because we

  want to focus on the range of analytical tools required for ESDA and their provision using GIS.

ESDA and GIS

- GIS has not been developed with ESDA in mind but rather in terms of data management (eg

AM/FM applications), cartographic modelling (eg whole map operations such as sieve analysis to locate areas) and some selected forms of spatial analysis (eg network analysis).

- Nonetheless GIS does contain some ESDA-type facilities.
- Some have argued that the very large datasets now becoming available in GIS require new tools

to be developed which can detect patterns and anomalies automatically, and that traditionalESDA methods are not appropriate in a GIS context. (e,g,Openshaw in Fotheringham and Rogerson (1994)). Others feel there is a role for current statistical methods and that these can be given new power, and made more useful and widely available by linkage to GIS. This is the view taken here.

In the remainder of the lecture we outline some of the techniques of ESDA, and conclude with a summary of how many of these can currently be implemented using GIS

---

2. Data Model for ESDA

A set of data can be thought of as having general trends (e.g. average values, relationships) and local
variations from those trends. These are sometimes called the **smooth** and **rough** properties of the data repectively:

DATA = smooth PLUS rough

Any data value can be thought of as comprising two components : one deriving from some summary
measure (smooth) and the other a residual component (rough)

Data properties for a single variable identified through ESDA:

Non-spatial properties.

- smooth
  - location of the distribution (measured by the median)
  - spread of the distribution (measured by the inter-quartile range)
  - shape of the distribution (depicted by box plots, histograms and smoothed curves if the data takes the form of a sequence of values)
- rough
  - outliers e.g. values more than a certain distance above (below) the upper (lower) quartile of the distribution (Haining 1993,201).

Spatial properties

- smooth
  - spatial trends (or gradients)
  - spatial autocorrelation
- rough

- difference between data value and smooth value.
- spatial outliers. These are individual attribute values that are very different in magnitude from their neighbouring values.

Data properties for two variables identified through ESDA

In the case of two variables, the scatterplot is used to visualise the relationship between them. The best fit line through the scatter plot identifies the **smooth** element of the relationship, and the residuals from the best fit line the **rough** element. An outlier is a data value more than a certain vertical distance from the best fit line.

---

3. Classification of ESDA Methods

It is useful to distinguish between two classes of ESDA statistics:

- **global** or **"whole map"** statistics which process **all** the cases for one (or more) attributes.
- **focused** or **"local"** statistics which process **subsets** of the data one at time and which may involve a sweep through the data looking for evidence of smooth and rough elements of the mapped data.

This lecture only considers global statistics.

The application of ESDA might involve working with windowed subsets of the map (analyst defined
boxes, circles or polygons). Processing might involve:

- applying global or focused statistics only to cases in the window
- executing a spatial query: e.g. "identify all the areas within the window possessing attribute property x."

---

4. ESDA for Describing Non-Spatial Properties of Attribute

Below are some techniques for identifying non-spatial properties of a single attribute. All are standard
EDA techniques - the link to the map however makes them part of ESDA.

- Median.
  - The measure of the location (or centre) of the distribution of attribute values.
  - **ESDA query**: which are the areas with attribute values above (below) the median?
- Quartiles and Inter-quartile spread.
  - The measure of spread of values about the median.

ESDA query: which are the areas that lie in the upper (lower) quartile?

- Box plots.
  - Graphical summary of the distribution of attribute values.
  - **ESDA query**: where do cases that lie in specific parts of the boxplot occur on the map?; where are the outlier cases located on the map?

Figure 1.

Boxplot of incidence rates of a disease in Sheffield. Areas with values above the median are highlighted on map, showing a tendency for higher rates in the eastern part of the city.

---

5. ESDA for Describing Spatial Properties of an Attribute

The following are techniques which are only applicable to spatial data, although some are spatial equivalents of methods developed for non-spatial data (e.g. time series data). As above they apply to a single attribute at a time.

- Smoothing.

  Where the map consists of many small areas it is often helpful to apply simple smoothing methods which, depending on the scale of the smoother, may help to reveal the presence of general patterns that are unclear from the mosaic of values.
  ESDA technique: The simplest form is spatial averaging - simply take the attribute value of an area and its neighbours and average them. Repeat for each area. (The median could also be used in this way)

- Identifying trends and gradients on the map.

  Are there any general trends or gradients in the map distribution of values (eg the existence of a general increase in heart disease incidence from Southern to Northern England) **ESDA techniques**:
  - kernel estimation
  - taking transects through the data and plotting with attribute value on vertical axis and spatial location on horizontal axis
  - spatially lagged boxplots with lag order specified with respect to a particular area or zone (Haining 1993 p224)
  - two way median polish adapted to non-regular lattice data suggested by Cressie and described in Haining 1993 p215-220.

- Spatial autocorrelation.

  This is the propensity for attribute values in neighbouring areas to be similar.

  ESDA technique: use a scatterplot with area attribute value on the vertical axis plotted

against the average of the attribute values in the adjacent areas on the horizontal axis (Haining 1993 p205,206). A scatterplot where there is an upward sloping scatter to the right is indicative of positive spatial autocorrelation (adjacent values tend to be similar). Where the scatter slopes upward to the left this is indicative of negative spatial autocorrelation (adjacent values tend to be dissimilar).

- Detecting spatial outliers

This is the situation where an individual attribute value is not necessarily extreme in the distributional sense but is extreme in terms of the attribute values in adjacent areas.

ESDA technique: Use the scatterplot technique as for spatial autocorrelation and then run a least squares regression on the plot. Cases with standardized residuals greater than 3.0 or less than -3.0 might be flagged as possible outliers although for reasons beyond the level of this course this will tend to overstate the number of outliers (Haining 1993 p214). Cressie's test can be adapted for this purpose but has the same problem of tending to overstate the existenceof outliers.

Figure 2.

This is an example of detecting a spatial outlier. The plot shows the attribute values plotted against the average of values in neighbouring areas. One region has been selected since it is an outlier from the
regression line. As the histogram shows this region is not an outlier in the distributional sense - in fact
its value falls in the modal class.

---

6. ESDA for Model Assessment

- ESDA not used for model confirmation in the sense of significance testing.
- Techniques can be used to test model assumptions, as in the case of regression.

Testing for spatial autocorrelation

- Regression assumes model errors are independent. Spatial autocorrelation will invalidate this assumption - can be tested for by looking for spatial autocorrelation in the residuals.

ESDA technique: Map the residuals and look for evidence of positive residuals clustering together. The scatterplot method for spatial autocorrelation described above can also be run on the regression residuals. A strong gradient in the scatterplot is indicative of a failure to meet the assumption of independence

Figure 3.

Detecting autocorrelation in regression residuals. This study was looking for a relationship between the incidence rate of a disease (Y axis on scatter plot) and deprivation (X axis). The

map distinguishes
positive and negative residuals, showing evidence of spatial autocorrelation (clustering of similar
values).

---

7. GIS and ESDA

What currently can/cannot be done in standard GIS?

- CAN
  - Identify 'smooth' spatial properties but the techniques are statistically quite sophisticated (eg trend surface analysis, kriging) and have been developed for interpolation rather than ESDA (see unit xxx????).
  - Some non-spatial statistics such as the mean and standard deviation. But these are not robust measures of the location and spread of a distribution and median and quartiles are less commonly provided.
  - Presentation graphics, especially maps but also histograms.
- CAN'T
  - Identify 'rough' properties such as outliers or spatial outliers.
  - implement most ESDA techniques.
  - employ visualisation methods. GIS graphics tend to be strong on PRESENTATION of data rather than EXPLORATION. data.

  The problem of implementing these tools in GIS is not a technical one - GIS contain basic tools to do many of these things:
  - Contiguity information. This is needed for standard GIS operations such as polygon dissolve but is also used in the calculation of many spatial statistics.
  - DBMS for calculating statistics.
  - Windowing capability including hot links to allow linkage between graphs and maps (e.g. MapInfo)

There is considerable work currently exploring various mechanisms for providing ESDA software -
references to further reading are given in the references section

- Purpose written using a variety of methods
  - REGARD - Mac software (Haslett et al 1990)
  - cdv - Tcl/Tk Unix application building software (Dykes 1996)
  - XLisp-Stat - public domain programming language/stats package (Brunsdon and Charlton 1996)
- These packages are extremely flexible and quick but often require the re-writing of parts of GIS e.g. map drawing require movement of data between GIS and software
- Other packages linked to GIS
  - SAM - routine called from within GIS (Ding and Fotheringham 1992)
  - SAGE - Client/server link to GIS. (present authors)
- Linkage to GIS uses the existing strengths of GIS and saves data export but are specific to one GIS system and linked windows requires sophisticated software (e.g. SAGE) All the illustrations provided in these lecture notes have been taken from SAGE.

8. Conclusions

- ESDA provides a set of robust tools for exploring spatial data, which do not require a knowledge of advanced statistics for their use.
- GIS are currently only poorly equipped with many of these tools, despite containing the basic functionality to allow them to be implemented.

9. References

Much of the literature is concerned with spatial data analysis rather than ESDA specifically, and so includes techniques not covered in this lecture, such as spatial regression techniques.

For general introductions to spatial analysis see the following

- Bailey T.C. and Gatrell A.C. (1995) *Interactive spatial data analysis*. Longman, Harlow. An excellent introductory text, which contains a software package, INFOMAP, which is capable of carrying out some of the procedures described here.
- Haining R.P.(1993) *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press. More advanced than Bailey and Gatrell, but includes more on ESDA.
- Cressie N. (1991) *Statistics for Spatial data*. Wiley, New York.

The following volume contains numerous papers discussing aspects of linking spatial analysis with GIS, including reviews of work in the area, and a paper by Openshaw giving an alternative approach to the problem.

- Fotheringham A.S. and Rogerson P(1994) (eds) *Spatial Analysis and GIS*. Taylor and Francis, London.

The following papers all describe work to develop software which provides spatial analysis tools for use with spatial data.

The software (SAGE) used to produce the illustrations is described in the following two papers

- Haining R.P., Wise S.M. and Ma J (1996) The design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics* 11, 449-466.
- Haining R.P., Wise S.M. and Ma J. (1997) Exploratory spatial data analysis in a GIS environment (paper submitted).

Other important papers discussing the development of software in this area:

- Haslett J., Wills G., Unwin A. (1990) SPIDER - an interactive statistical tool for the analysis of spatially distributed data. *Int.J.Geographical Information Systems*. (Note: later versions of this software are called REGARD)
- Dykes J. (1996) Dynamic maps for spatial science: a unified approach to cartographic visualizaton. In Parker D. (ed) *Innovatons in GIS 3*, Taylor and Francis, London, 177-188.
- Brunsdon C. and Charlton M. (1996) Developing an exploratory spatial analysis system in XLisp-Stat. In Parker D. (ed) *Innovatons in GIS 3*, Taylor and Francis, London, 135-146.
- Ding Y. and Fotheringham S. (1992) The integration of spatial analysis and GIS. *Computers, Environment and Urban Systems* 16, 3-19

---

10. Questions and Discussion Points

1. Assess the value of ESDA techniques in analysing any geographical data with which you are familiar.
2. It is common to report crime statistics in terms of areal units, such as those covered by a single police officer or for which a police station is responsible. The data would normally consist of counts of the number of crimes committed
    - On a daily basis
    - Classified into different types of crime.

   How would you apply ESDA techniques to identify 'hot spots' i.e. areas with consistently high rates of crime?
3. Openshaw takes the view that in the 'data-rich' world of GIS, traditional techniques of ESDA are inappropriate, and that what is needed are methods which can identify patterns and hot spots automatically. Do you agree?
4. Discuss the strengths and weaknesses of current GIS software for undertaking ESDA.
5. Discuss how a map showing evidence of a linear spatial trend and a map showing evidence of spatial autocorrelation might differ.
6. Describe the difference between whole map and local statistics, and give examples where each would be appropriate.
7. In undertaking spatial smoothing, why might it be better in some cases to use the median rather than the average?

Citation

To reference this material use the appropriate variation of the following format:

Haining, Robert and Wise, Stephen (1997) Exploratory Spatial Data Analysis, *NCGIA Core Curriculum in GIScience*, http://www.ncgia.ucsb.edu/giscc/units/u128/u128.html, posted December 05, 1997.

Last revised: December 05, 1997.

Unit 128 - Exploratory Spatial Data Analysis

Table of Contents

Back to the Unit

Unit 128 - Exploratory Spatial Data Analysis

Metadata and Revision History

1. About the main contributors

- authors
  - Robert Haining and Stephen Wise
- software used for illustrations was written by
  - Jingsheng Ma

2. Details about the file

- unit title
  - Exploratory Spatial Data Analysis
- unit key number
  - 128

3. Key words

4. Index words

5. Prerequisite units

6. Subsequent units

7. Other contributors to this unit

8. Revision history

- 05 December 1997 - first draft

Back to the Unit.

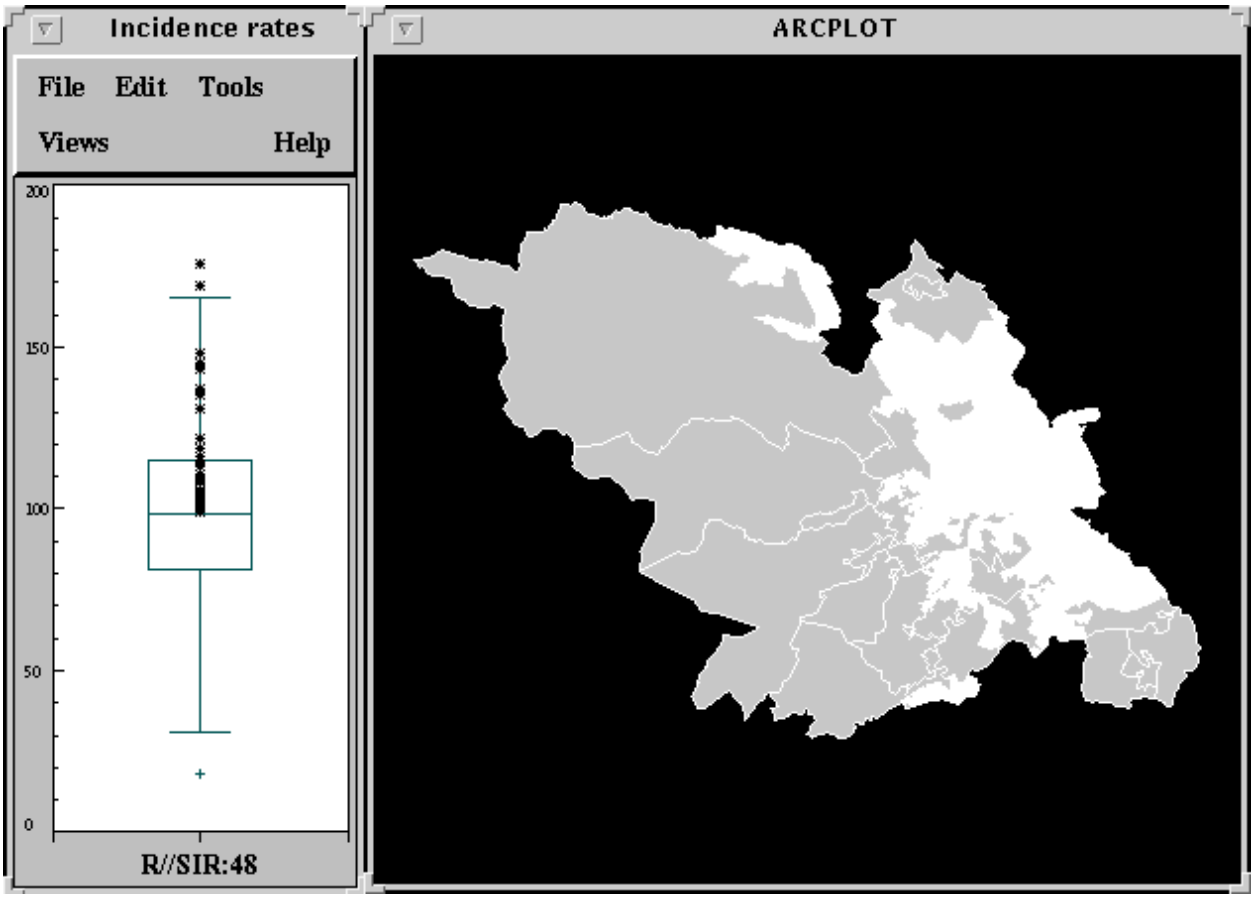

Figure 1. Boxplot of incidence rates of a disease in Sheffield. Areas with values above the median are highlighted on map, showing a tendency for higher rates in the eastern part of the city.

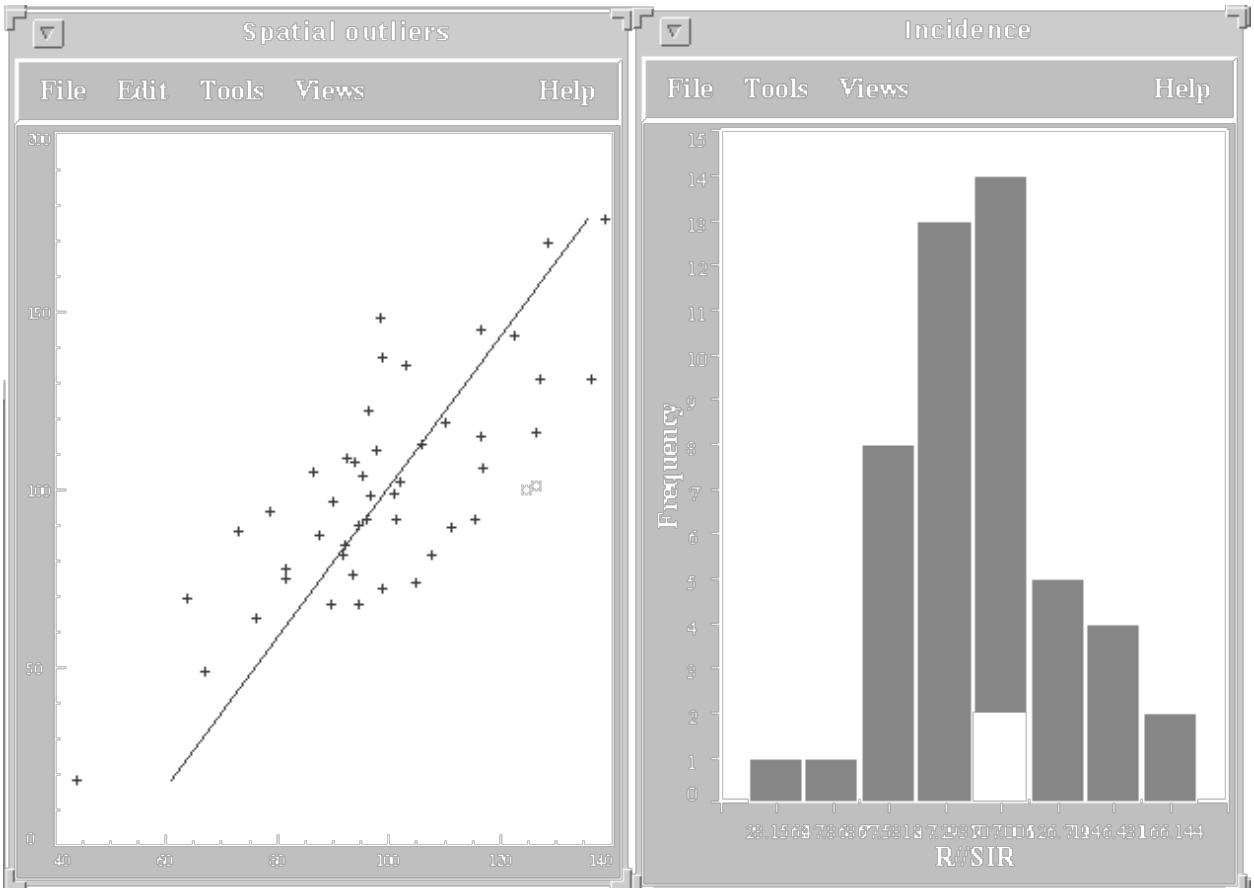

Figure 2. This is an example of detecting a spatial outlier. The plot shows the attribute values plotted against the average of values in neighbouring areas.

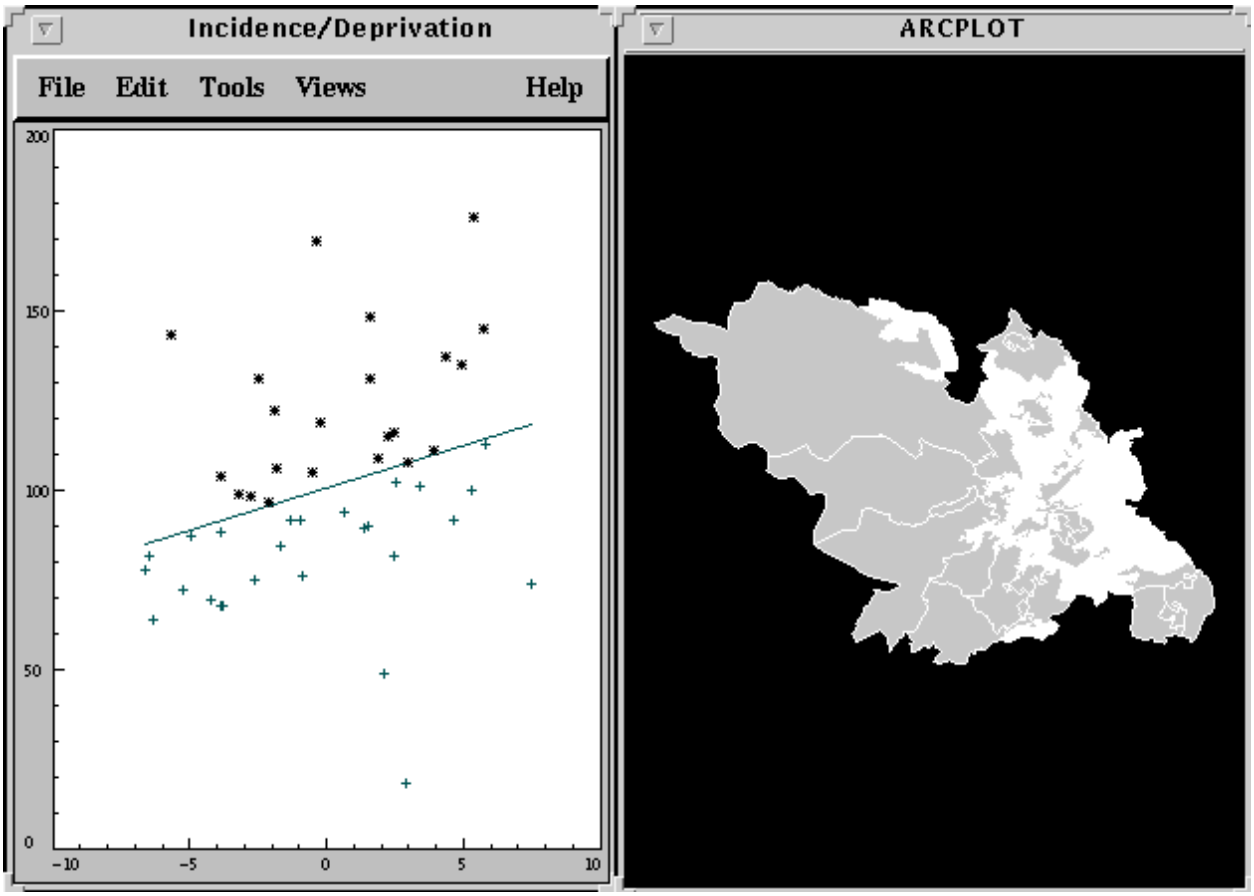

Figure 3. Detecting autocorrelation in regression residuals. This study was looking for a relationship between the incidence rate of a disease (Y axis on scatter plot) and deprivation (X axis).