# Open Big Data Quality:
# An exploration of Modifiable Areal Unit Problem

Brian Moran[*1], Egess Tiri[†1] and Chris Brunsdon[‡2]

[1,2]National Center for Geocomputation, Maynooth University

January 31, 2019

**Summary**

In the smart city era, open big data is taking the lead in city governance. Many city planners and government policies rely on census data, but the way these data are interpreted can lead to different city representations. In this paper, following the work of Gehlke and Biehl on Modifiable Areal Unit Problems (MAUP), we are looking into the problem of ecological fallacies by exploring census data for four different spatial units in Dublin.

**KEYWORDS:** smart cities, open big data, data quality, modifiable areal unit problem

## 1. Introduction

"Smart city is one that strategically uses networked infrastructure and associated big data and data analytics to produce a smart: economy, government, mobility, environment, living and smart people." (R, 2015) On one hand we have tons of real time data which are produced daily by thousands of sensors around the city and on the other hand there are non-real time data which are measured on a fixed timeline such as census data. City governors take their decisions by relying on these data. One of the purposes of a smart agenda is to increase citizen participation and governance transparency through open data initiative. Cities worldwide are using city dashboards as a tool of communicating these data to citizens.

> "*City dashboards use visual analytics – dynamic and/or interactive graphics (e.g., gauges, traffic lights, meters, arrows, bar charts, graphs), maps, 3D models and augmented landscapes – to display information about the performance, structure, pattern and trends of cities.*" (Kitchin, 2016)

In this framework, Building City Dashboards Project (http://dashboards.maynoothuniversity.ie/) at Maynooth University is trying to put in one platform all the open data for Ireland, by developing two city dashboards for Dublin and Cork, and addressing many issues which come along with the open data initiative, among them the data quality problems and how do we measure data quality?

[*] Brian.moran.2017@mumail.ie
[†] Egess.tiri.2017@mumail.ie
[‡] Christopher.Brunsdon@mu.ie

### 1.1. Data Quality

Dama UK Working Group (Group, n.d.) has identified six primary dimensions of data quality and related dimension for each core ones, as follows:
Figure 1 below shows six primary dimensions of data quality

- Completeness - The proportion of stored data against the potential of "100% complete"
  Related Dimension – Validity and Accuracy
- Accuracy - The degree to which data correctly describes the "real world" object or event being described.
  Related Dimension - Validity
- Timeliness - The degree to which data represent reality from the required point in time.
  Related Dimension - Accuracy
- Validity - Data are valid if it conforms to the syntax (format, type, range) of its definition
  Related Dimension - Accuracy, Completeness, Consistency and Uniqueness
- Consistency - The absence of difference, when comparing two or more representations of a thing against a definition.
  Related Dimension - Validity, Accuracy and Uniqueness
- Uniqueness - Nothing will be recorded more than once based upon how that thing is identified
  Related Dimension - Consistency

Completeness
Uniqueness
Accuracy
Quality
Consistency
Timeliness
Validity

**Figure 1** Six Primary Definitions of Data Quality

In this abstract we are trying to explore one of these dimensions which is Data Accuracy. The way data are interpreted and grouped for further analysis, for certain purposes, often gives biased results and as a consequence will affect city governance and misinform citizens.

## 2. Modifiable Areal Unit Problem

Following the work of Dr. Henry Sheldon, in 1934, Gehlke and Biehl did a detailed study of grouping effects in census tract data, where they concluded that "Variations in the size of the correlation coefficient is conditioned upon changes in the size of unit used, with a smaller value of r associated with the smallest unit rather than with the largest" (Gehlke & Biehl, 1934)

> "An ecological fallacy occurs when it is inferred that results based on aggregate zonal (or grouped) data can be applied to the individuals who form the zones or groups being studied." (S, 1984)

MAUP consists in two main issues:

a) *the Scale Problem*: when using the same analytical method for different spatial units, where the larger units are made up of smaller units, can lead to different results; and

b) *the Zoning Problem:* when small units are grouped together to make larger units and by using the same analytical method, the results can be different.

For further exploration of MAUP we are using data from Irish Census 2011, which was carried out Sunday 10th of April of the same year. These data are available to download from Central Statistics Office[§] website in comma separated variables (csv) format. In it we investigate the Index of Deprivation, known as the Pobal HP Deprivation Index in Ireland, by examining four different variables in four different spatial units.

The variables from the census we examine are

- "Very Bad Health / T12_3VBT",
- "Unemployed / T8_1_ULGUPJT",
- "Unskilled / T9_1_UST"
- "No Formal Education / T10_4_NFT"

The spatial areas we are using from the census are listed below along with the number of areas:

- Counties – 34 areas
- Local Electoral Areas – 171 areas
- Electoral Districts – 3,409 areas
- Small Areas – 18,488 areas

"Many ecological studies include the collection and use of data to investigate the relationship between a response variable and a set of explanatory factors (predictor variables). If the predictor variables are related to one another, then a situation commonly referred to as multicollinearity results." (B. Desta Fekedulegn, 2002)

Ridge Regression is a technique for analysing multiple regression data that suffer from multicollinearity. (Lambert M. Surhone, 2010) In ridge regression[**], we add a penalty by way of a tuning parameter called lambda which is chosen using cross validation. The idea is to make the fit small by making the residual sum or squares small plus adding a shrinkage penalty. The shrinkage penalty is lambda times the sum of squares of the coefficients so coefficients that get too large are penalized. As lambda gets larger, the bias is unchanged but the variance drops. The drawback of ridge is that it doesn't select variables. It includes all of the variables in the final model.

[§] https://www.cso.ie/en/census/

[**] http://wavedatalab.github.io/machinelearningwithr/post4.html

It seeks to minimise the equation 1[††]:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j}^{m} \beta_j^2 \qquad (1)$$

In **Figure 2 & 3** we can see strong correlations across all the variables. **Figure 4** shows there is a large change in magnitude of the eigenvalues; this is another result of high correlation. **Figure 5** shows less correlation for Local Electoral Area than for the Counties Level. Again, less correlation for the Electoral Districts as seen in **Figure 6,** where the areas are smaller. In **Figure 7-8,** which is for the small areas, there is *little* to *no* correlation in the variables.

The variables have high collinearity in the larger spatial areas and the figure falls as the amount of areas increase, i.e. as the areas get smaller, the correlation decreases. This can be seen in **Figure 9** which is a plot of the correlations of very bad health against unemployment.

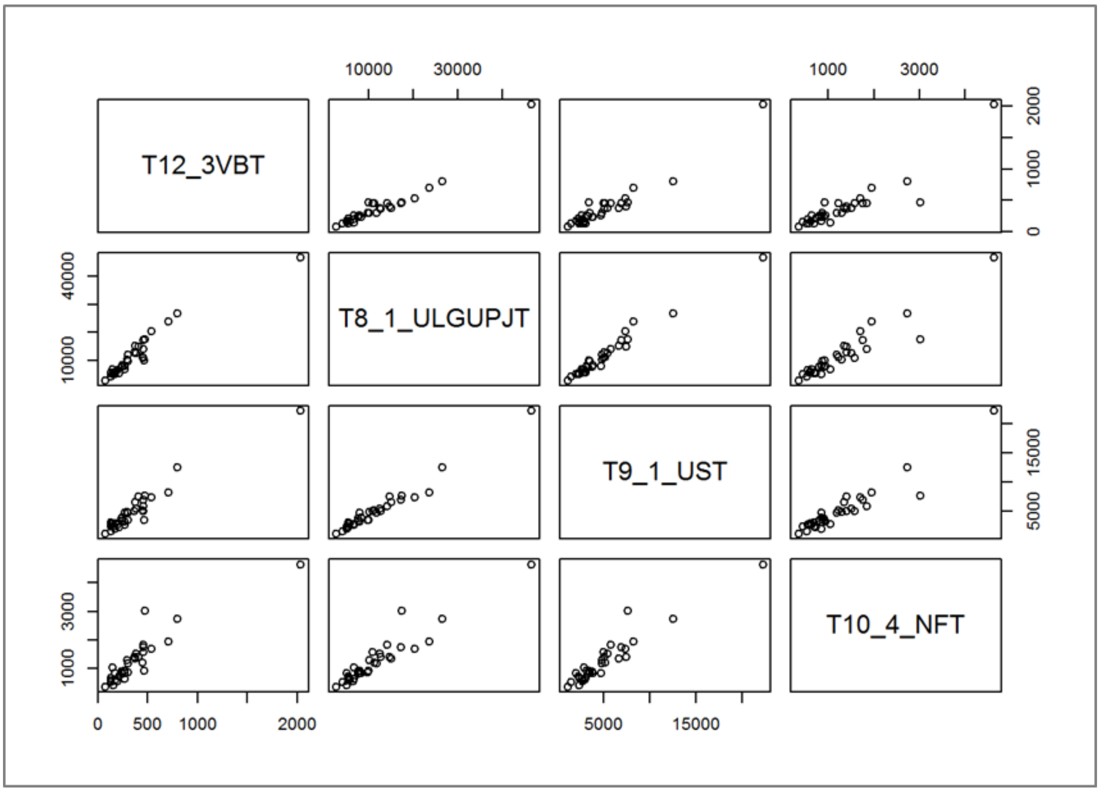
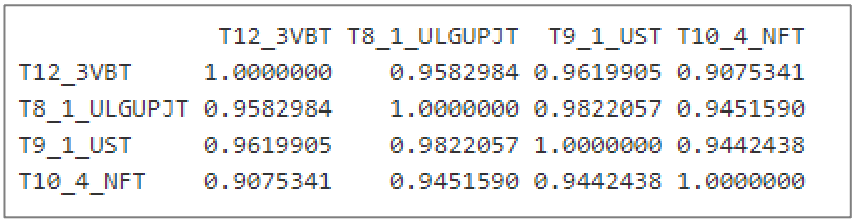


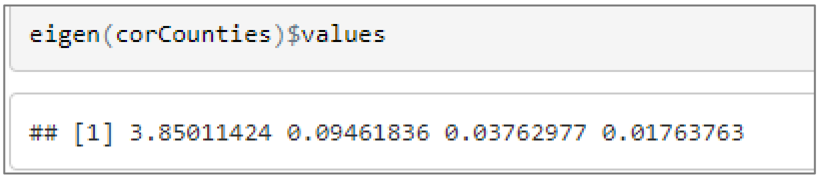
**Figure 2 -** Correlation Scatter Plot - Counties Level

[††] http://www.thefactmachine.com/ridge-regression/



```
              T12_3VBT T8_1_ULGUPJT  T9_1_UST T10_4_NFT
T12_3VBT     1.0000000    0.9582984 0.9619905 0.9075341
T8_1_ULGUPJT 0.9582984    1.0000000 0.9822057 0.9451590
T9_1_UST     0.9619905    0.9822057 1.0000000 0.9442438
T10_4_NFT    0.9075341    0.9451590 0.9442438 1.0000000
```

**Figure 3 -** Correlation Matrix - Counties Level



```
eigen(corCounties)$values
```

```
## [1] 3.85011424 0.09461836 0.03762977 0.01763763
```

**Figure 4 -** Eigenvalues - Counties Level

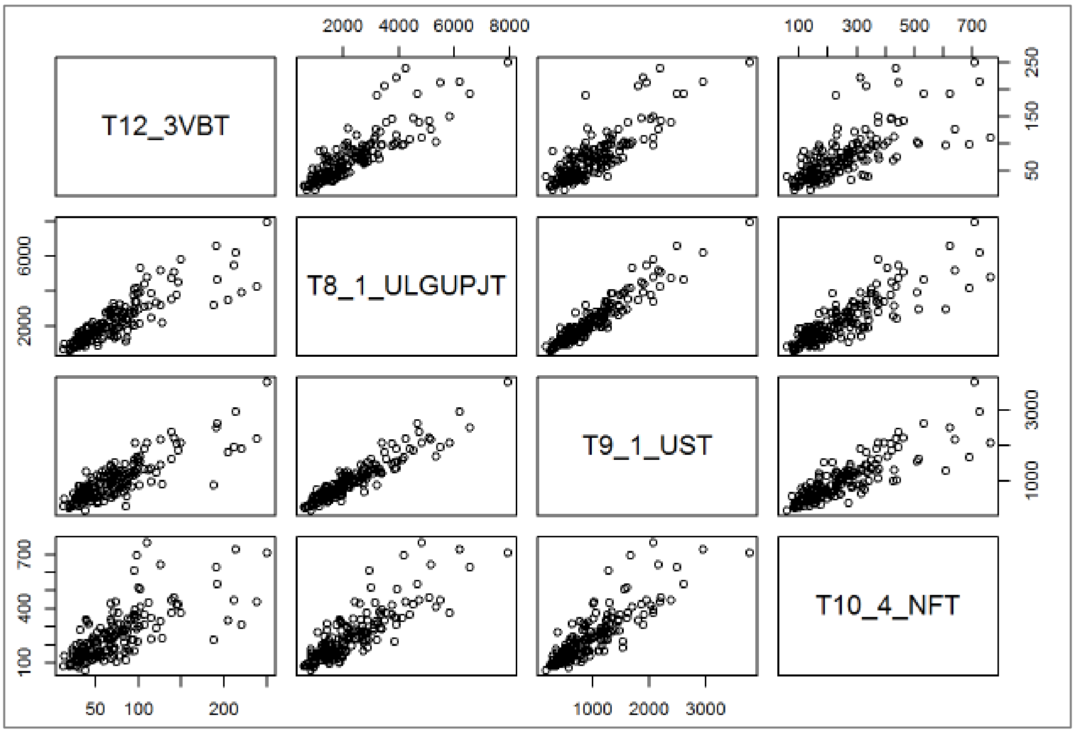


**Figure 5 -** Correlation Scatter Plot - Local Electoral Areas Level

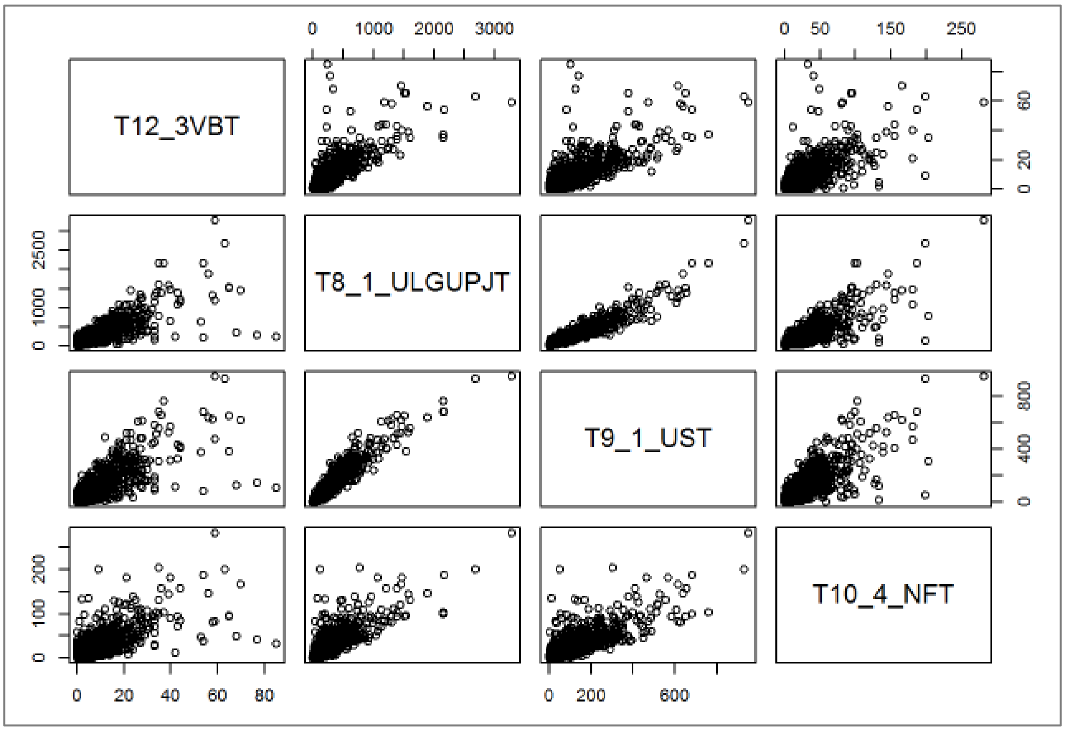


**Figure 6 -** Correlation Scatter Plot - Electoral Districts Level

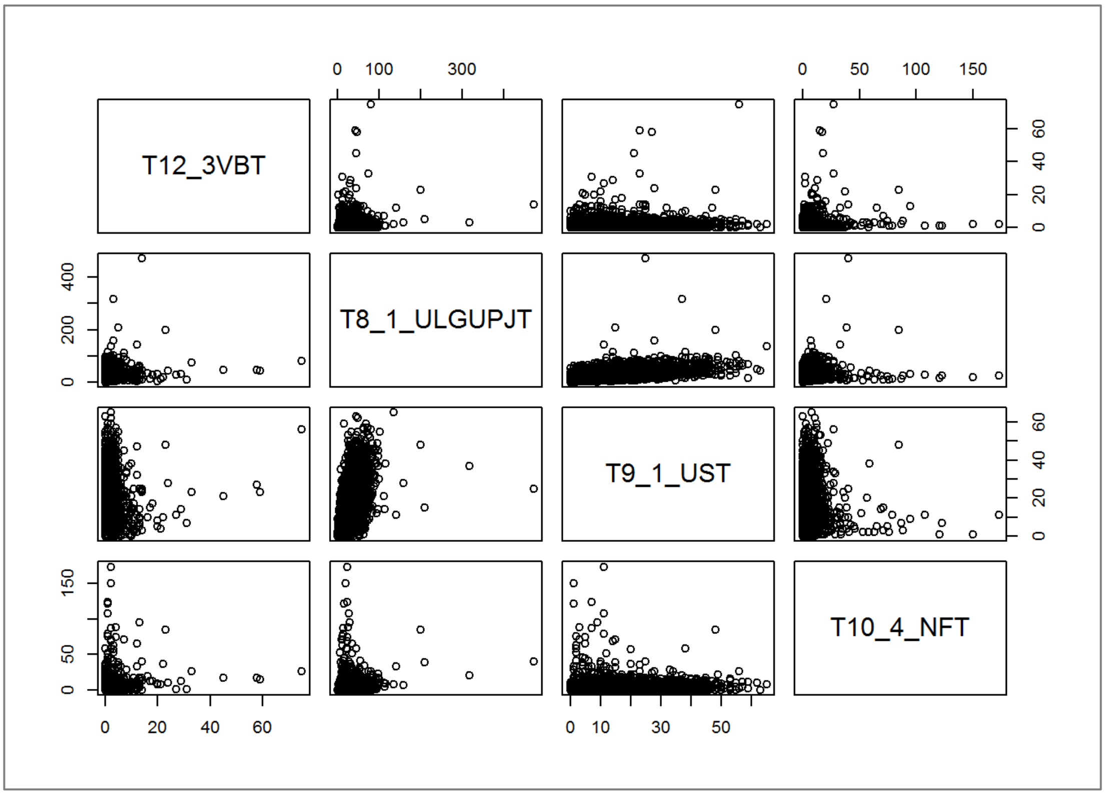
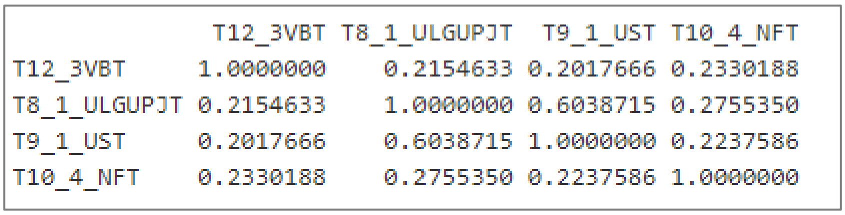


**Figure 7 -** Correlation Scatter Plot - Small Areas Level



```
               T12_3VBT T8_1_ULGUPJT  T9_1_UST T10_4_NFT
T12_3VBT      1.0000000    0.2154633 0.2017666 0.2330188
T8_1_ULGUPJT 0.2154633    1.0000000 0.6038715 0.2755350
T9_1_UST      0.2017666    0.6038715 1.0000000 0.2237586
T10_4_NFT     0.2330188    0.2755350 0.2237586 1.0000000
```

**Figure 8 -** Correlation Matrix - Small Areas Level

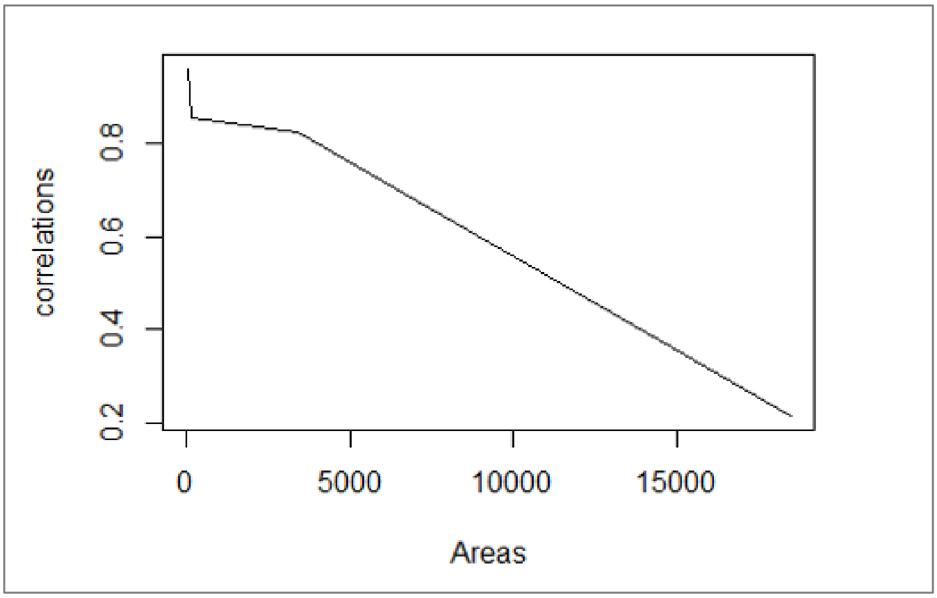


**Figure 9 -** Plot of the correlations

Another exploration method is leave one out cross validation. Leave-one out cross-validation (LOOCV) is a special case of K-fold cross validation where the number of folds is the same number of observations (i.e. K = N).[‡‡] We are exploring the data using this method and then comparing the results to ridge regression method.

Furthermore, we are going to present and discuss our findings during our conference presentation.

[‡‡] https://gerardnico.com/data_mining/cross_validation#loocv

## 3. Acknowledgements

"This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 15/IA/3090."

## 4. References

B. Desta Fekedulegn, J. C. R. H. J. M. E. S., 2002. *Coping with multicollinearity.* s.l.:s.n.

Gehlke, C. E. & Biehl, K., 1934. *Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material.* s.l., Taylor & Francis Ltd, pp. 196-170.

Group, D. U. W., n.d. *White Papers.* [Online]
Available at: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

Kitchin, R. M. G., 2016. *Urban data and city dashboards: Six key issues,* s.l.: SocArXiv Papers.

Lambert M. Surhone, M. T. T. S. F. M., 2010. *Ncss Statistical Software.* s.l.:s.n.

R, K., 2015. *Data-driven, networked urbanism,* Maynooth: SSRN.

S, O., 1984. *"The Modifiable Aeral Unit Problem".* Norwhich: Geoabstracts.

### Biographies

**Brian Moran** is a doctoral researcher, working on the Building City Dashboards Project at the National Centre for Geocomputation (NCG), Maynooth University, Ireland. He holds a MSc in “Data Science and Analytics” from Maynooth University, Ireland. His research focuses in analytics and modelling problems: moving beyond visual analytics to perform data analytics, statistical modelling, generate and evaluate predictions, simulations, and optimisations. He has previously worked in the private sector as an IT project manager and a service delivery manager

**Egess Tiri** has a BSc and MSc Degree in Geo Informatics from the Faculty of Geology and Mine, Polytechnic University of Tirana, Albania. After finishing her studies, she has been working as a Database and GIS Specialist for Public Offices and Large International Projects in Albania. Currently she is a Doctoral Researcher with National Centre for Geocomputation (Maynooth University, Ireland) on Building City Dashboards Project. Her research focuses on Data Problems which consists of deploying existing and developing new techniques for assessing data quality and veracity. Some of the issues include: (i) identification and correction of anomalous data; (ii) ecological fallacies; (iii) data standards; (iv) communication of metadata; and (v) calibration issues.

**Chris Brunsdon** is Professor of Geocomputation at the National Centre for Geocomputation, Maynooth University Ireland. Prior to this he was Professor of Human Geography at the University of Liverpool in the UK, and before this worked in the Universities of Leicester, Glamorgan and Newcastle, all in the UK. He has degrees from Durham University (BSc Mathematics) and Newcastle University (MSc Medical Statistics, PhD in Geography).