Interactive Visualization of Urban Data: Interactive Geodemographics

Burcin Yazgi Walsh & Chris Brunsdon

National Centre for Geocomputation (NCG)
Maynooth University (MU)

e-mail: burcin.yazgiwalsh@mu.ie

Abstract

Visualization of data is used to provide insight from information with the purpose of providing a better understanding of the process generating the data. It also aims to give the opportunity to users to see structure and pattern in a particular dataset in a relatively quick and intuitive way. Interaction as a tool carries these visual representations to a dynamic level. This way of presentation provides the advantages of simplicity and effectivity. Geodemographics on the other hand is another tool to represent the dynamics of spatial units based on some dataset, typically socio-economic. It is a tool that helps to classify the characteristics of areas based on their similarities. This paper aims to carry the two existing concepts of visualization of urban data and geodemographics to a new platform with a focus of interactive geodemographics. The purpose of this paper is not only limited to classifying the smallest scale of administrative units in Ireland based on the 2016 Census population data, but also to demonstrate these outputs with a newer approach like interactive visualizations. The data used in this study is obtained from Ireland's 2016 Census Population Data at small area level. The paper is also argues for the use of open data, open source software and reproducible research. Since some of the main disciplines like computer science, data science, geocomputation, information technology, geographical information sciences, geography, urban and regional planning, statistics are involved in the concepts of interest here, it is believed that the discussion will be useful and supportive for the formation of novel alternative platforms to represent urban data.

Introduction

Visualization of data is used to provide insight from data with the aim of providing a better understanding of the process that is generating the data. It also aims to give the opportunity to users to see the structure and pattern in a particular dataset in a relatively quick and intuitive way. Donolo et al. (2013) believe that "visualization is a means of making sense of data for both experts and non-experts". It is not any different when it comes to the visualization of urban data. The main aim of this type of representation is to provide easy ways to understand the relations in a dataset, especially in complex spatial patterns of urban life. While presenting these relations through visuals, the other purpose is to highlight differences in various processes which are related to urban structure.

Interaction is a way to carry the visual representation to a different level where the visualisations can act as dynamic elements rather than static ones. Simply, rather than trying to present different levels of same information or the same dataset in one static output, it is possible to add these spatial and temporal features as selectable layers to visualization outputs using interactive tools. This way of presentation provides the advantage of simplifying the exploration of patterns. In addition, this supports engagement with data and helps to view information in a form that a user is specifically looking to use. As Goodchild (2015) states "the new technologies allow maps to become dynamic and interactive, replacing a one-size-fits-all design with one that accommodates and supports

specific well-defined use cases". This is one of the other aspects of interactivity that is discussed – within its nature interactivity will help to user to find their own way of representation.

Geodemographics is a technique used to represent the social characteristics of spatial units based on some data set relating to those units. Since its early stages – around the 1970s – it has focused mostly on census data. This approach seeks to to classify the characteristics of areas based on the similarities found in this data. There are different approaches for selecting variables and selecting methods for the classification which will be discussed more in detail in the following section.

The purpose of this paper is not limited to classifying the smallest scale administrative units in Ireland based on the 2016 Census population data, but also discuss how to assess these outputs with approaches like interactive visualizations.

Within that frame, this paper is aiming to carry the concepts of visualization of urban data and geodemographics – to a new platform with a new perspective. With a focus on interactive geodemographics, the paper is structured around five different sections. After a brief introduction and background information, the data and method used in this study are introduced. The next section will be focusing on the outputs followed by a discussion around the visualization of spatial urban data in general and the paper is closed with some concluding remarks.

Background

Producing representations of urban data goes back centuries but visualization in the new form – interactivity - which is the focus of this paper is mostly discussed over the last five decades. Developments in information technology and computer systems are the underlying drivers for computer supported and georeferenced visualisations. Interactivity as an added attribute to these visualisations is a newer approach with the purpose of better, more efficient and intuitive representations of data.

There are varying representational forms of visualization and the map form is one of the more popular versions. It can be claimed that the first use of this form of representation is John Snow's (1854) cholera map which made the missing link (highly affected areas being close to the water supply) visually obvious on his map. Experiments with visualisation started in 19th century and at when considering their progress until today, there is a strong relationship between the technological limits and the things that can be shown on a map.

Following up, according to Friendly (2006), the second half of the 20th century saw a rebirth of data visualisation based on three major developments. First, Tukey's studies - starting with his paper in 1962 'The Future of Data Analysis' - introduced various effective graphical approaches and visual tools. Second, Bertin's study in France that visualized information based on the relations in multidimensional data. Third, the improvement in computing that helped to process data in different forms and carry out the visualisations more rapidly. It is around this time that interactive applications were also introduced. Limited versions of this kind of application based on the accessibility to technological infrastructures started to develop in the following decades. Improvements in computer technology then played a lead role in the development of visualization in the following decades.

Initial experiments with interactivity were based on graph forms rather than maps. Selecting, linking and brushing actions were applicable through these plots and such actions played a key role in visualization techniques developed. Friendly (2006) mentioned in his paper "it may be argued that

the greatest potential for recent growth in data visualization came from the development of dynamic graphic methods, allowing instantaneous and direct manipulation of graphical objects and related statistical properties". One of these examples is Becker and Cleveland's (1987) interactive graphics system called 'brushing scatterplots'. These early attempts at interactivity led the way for extensions of this subject including interactive maps. Around a decade later Dykes (1996) developed the Cartographic Data Visualiser (CDV) tool. Subsequently applications and tools using this method accelerated. It is important to mention that newly available hardware and software alternatives, and more easily obtained data also played a critical role in this acceleration. The research of Friendly and Denis (2001) on milestones in the history of data visualization is a great piece of work to follow the timeline under the categories of cartography, statistics and graphics, and technology.

Geodemographics, which is the other main focus of this paper, has its own unique development history, but in common with interactive visualization the method gained ground and grew in use with the rapid development of IT in the 1990s and 2000s. According to Singleton and Spielman (2014), the roots of geodemographic studies go back to human ecologists, social area analysis and factorial ecology. Later on it was used to deal with the census data and categorizations for geographies initially based on one city and then at national level. By the end of 1970s, geodemographics gained some more importance related to its commercial value.

By 1980s it was used for marketing purposes –consumer behaviour studies obtained through these demographic classifications (Brown, 1991; Singleton and Spielman, 2014). Webber in 1975 produced one of the first example of these classifications. In his study Liverpool census data was used at Enumeration District (ED) geographical level. There were different approaches to classification in different based on the availability of data and data type.

Some critiques of the method started to arise after the first attempts of geodemographics in the 1980s. Openshaw (1983) criticised the choices made during the classification process, such as the method for clustering or number of the clusters. He argued that there is no possible classification that would suit all purposes. Charlton et al. (1985) suggested an alternative to the proposed classification by Webber which was mostly for commercial use - for a better general purpose classification.

Even though the way it is applied and the purpose behind of it can clearly have an important effect on the outcomes, with the awareness of these facts, geodemographics are still believed to be a helpful tool. This can be supported by Charlton et al.'s study (1985) since they mentioned in their paper that "The problem is that there is no way of measuring 'bestness'. Instead, the results provide an exploratory spatial description of small-scale areal census data for an entire country." The other comment on this is by Singleton & Spielman (2014) who stated "It is entirely possible that more careful construction and broader use of geodemographic classification in the academy could support the development of a more robust theory of socio-spatial structure".

Computer science, data science, geocomputation, information technology, geographical information sciences, geography, urban and regional planning and statistics are some of the main disciplines that have had an important influence on both of the topics discussed above

Data & Method

In this section of the paper, after the data used for the geodemographics of Ireland is explained, the method behind the classifications will be discussed since it is important to know the derivatives behind the analysis.

Data

The list of the variables that are part of this study is represented in Table 1. The data used in this study are derived from Ireland's 2016 Census obtained through Central Statistics Office Ireland. The Census dataset includes more than 700 variables, but following the methodology applied by Brunsdon et al. (2011) for Ireland Census Data classifications 2011, 40 of these data subgroups were selected from the census dataset under the themes of Sex, Age and Marital Status; Migration, Ethnicity, Religion and Foreign Languages; Families; Housing; Education; Principal Status; Motor Car Availability, PC Ownership and Internet Access; Commuting; Disability, Carers and General Health; Industries. The level chosen to analyse cluster groups is the smallest possible scale, which is the Small Area (SA) level for Ireland. It is important to mention that all data used in this study are obtained through open governmental portals and processed using open source software.

Method

In the course of this paper, analysis performed includes three stages: variable selection, modelling process and visualisation of outcomes. First, variables are organised under the categories of demographic, household composition, housing, socio-economic, employment and internet access. These categories follow a study by Brunsdon et al. (2011) for 2011 census data for Ireland. Since variable organisation in that study maintains similarity with the work here, the same 40 different variables are selected which can be followed as a list form in Table 1.

Table 1- Variables used for classifications

Variable	**Census Theme**	**Description**	**Analysis Theme**
Age0_4	Sex, Age and Marital Status	Age 0 to 4	Demographic
Age5_14	Sex, Age and Marital Status	Age 5 to 14	Demographic
Age25_44	Sex, Age and Marital Status	Age 25 to 44	Demographic
Age45_64	Sex, Age and Marital Status	Age 45 to 64	Demographic
Age65over	Sex, Age and Marital Status	Age 65 and over	Demographic
EU_National	Migration, Ethnicity, Religion and Foreign Languages	EU Nationality	Demographic
ROW_National	Migration, Ethnicity, Religion and Foreign Languages	Nationality - Rest of world	Demographic
Born_outside_Ireland	Migration, Ethnicity, Religion and Foreign Languages	Birthplace out of Ireland	Demographic
Seperated	Sex, Age and Marital Status	Separated and Divorced	Household Composition
SinglePerson	Sex, Age and Marital Status	1 person households	Household Composition
Pensioner	Families	Retired households	Household Composition
LoneParent	Families	One parent family with children	Household Composition
DINK	Families	Pre-family	Household Composition
NonDependentKids	Families	Families with youngest child aged 20 and over	Household Composition
RentPublic	Housing	Rented from local authority	Housing
RentPrivate	Housing	Rented from private landlord	Housing
Flats	Housing	Flat/apartment	Housing
NoCenHeat	Housing	No central heating	Housing
RoomsHH	Housing	Average no. of rooms for household	Housing

PeopleRoom	Housing	Average no. of people for room	Housing
SepticTank	Housing	Individual septic tank	Housing
HEQual	Education	Higher education - including bachelor, postgraduate, doctorate	Socio Economic
Employed	Principal Status	At work population	Socio Economic
TwoCars	Motor Car Availability, PC Ownership and Internet Access	Having 2 or more cars	Socio Economic
JTWPublic	Commuting	Journey to work using public transport-including bus, minibus, coach, train, DART, Luas	Socio Economic
HomeWork	Commuting	Work mainly at or from home	Socio Economic
LLTI	Disability, Carers and General Health	General health condition - including bad and very bad	Socio Economic
UnpaidCare	Disability, Carers and General Health	Carers	Socio Economic
Students	Principal Status	Students	Employment
Unemployed	Principal Status	Unemployed - having lost or given up previous job	Employment
EconInactFam	Principal Status	Looking after home/family	Employment
Agric	Industries	Agriculture, forestry and fishing	Employment
Construction	Industries	Building and construction	Employment
Manufacturing	Industries	Manufacturing industries	Employment
Commerce	Industries	Commerce and trade	Employment
Transport	Industries	Transport and communications	Employment
Public	Industries	Public administration	Employment
Professional	Industries	Professional services	Employment
Internet	Motor Car Availability, PC Ownership and Internet Access	No. of households connected with broadband	Other
Broadband	Motor Car Availability, PC Ownership and Internet Access	No. of households with internet	Other

The next stage is the modelling process for geodemographic classification which includes the steps of running a Principal Component Analysis (PCA); computing k-means and Partitioning Around Medoids (PAM) algorithms – these are clustering algorithms and provide the classification of Eds into each group. Briefly framed, as a result of principal component analysis, the proportion of each component's variance explanation is obtained and cumulative effect is investigated. It was noticeable that first 14 components could explain the 81.3% of the variance in total in this case. Thus this number of principal components were used as an input for the k-means cluster analysis. The next step is to find the number of the clusters. In order to have a better understanding,the smallest cluster size versus total number of clusters is plotted. As a result of this plot, when the number of the clusters is more than 8, it looks like very small – and likely spurious -clusters are more common. Based on this information, using the cluster number of 8, the classification procedure is recalculated. The next goal in the process of classification is identifying the characteristics of each group.

Following Brunsdon et al. (2011) "Since the above computation applied the analysis to principal components it is helpful to characterise the clusters in terms of the original variables used in the PCA. To do this, firstly the z-scores of each variable for the cluster centroids are computed. If the initial variable is x, then the z-score is given by

$$z = \frac{x - \bar{x}}{s}$$

where $\bar{x}$ is the arithmetic mean of the *x* values in the data set, and *s* is the sample standard deviation defined by

$$s = +\sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

and this standardisation allows values of each variable to be compared on a consistent scale, the z-score in each case has a mean of zero and a standard deviation of one. Here, the x values are actually cluster-wise mean values, whereas $\bar{x}$ and *s* are computed for values for all Small Areas. This is useful for identifying which clusters have relatively high or low values of particular variables. The following code creates a set of z-scores for each cluster".

An alternative technique applied in this study for clustering is Kaufman and Rousseeuw's (1990) Partitioning Around Medoids (PAM) algorithm. Each step set out above for the modelling process was rerun with the new algorithm to compare the performance. "Difference between this and k-means is that the total absolute distances within clusters are minimised, rather than the sum of squared distances. The contribution to this sum from outliers is proportionally less, and so there is less tendency to form small clusters of outlying areas" (Brunsdon et al., 2011). The spread of the sizes of the clusters was less than with k-means and that is why the PAM algorithm results were used in the as input for later stages of the study.

The final step is the visualization of the classification results. The outputs will be presented in the following section but in this section mostly the approaches behind these visuals will be discussed. As mentioned before, the interactive visualization concept is applied through geodemographic classifications, in other words through cluster analysis, in this paper. There are several stages included in this process and these are related to the visualization. Querying and hovering events were the initial steps applied. Later some of the interactive visualization tools such as dynamic linking/brushing are illustrated through a spatial map of classification groups of Ireland. All the code for each step can be obtained through Yazgi Walsh et al. (2017).

Visualisations as well as all prior steps of this study are achieved using the open source R language and different packages/extensions of this programing language. During the analysis procedure for the classifications the specific R packages of 'plyr' and 'cluster' are used. For supporting the test stages, the graph production 'ggplot' package was helpful. For some of the outputs like heatmaps the 'RColorBrewer' package was preferred. For several static map alternatives the package 'tmap' was adapted. For interactive visual outputs 'ggplot', 'leaflet' and 'shiny' were integrated. Through all the stages, documentation was kept in R Markdown - this helped to publish a quick, online version of the outputs on the Rpubs website that gives the opportunity for this research to be reproducible.

Outputs & Discussion

Classification based on the 2016 census data for the whole country is one of the main outcomes of this study. Not to lose the main focus of this paper, the details of each group will not be discussed in detail here, but instead attention will be given to how to visualise the outputs of urban data to interact better with the audience or put another way how to make research more readable through these visual outcomes. Cluster analysis is one of the classic tools for data mining and provides an opportunity to explore the insights of urban data in any case, but adding a level of interactivity to it makes it even more powerful.

One of the first ways you can explore the classification output is by querying and investigating the different classifications. This gives you the opportunity to see the geographical distribution of a specific cluster group within the country as well as the distribution of all groups. With this approach (Figure 1), the plot shown on the side of the map contains the values of the variables based on the classification group will also update to give the user the chance to explore the varying values for each of these groups. The other tool that is provided by this interactive map is the hovering action over the groups that are presented in the map. With this you can easily read the location information of each group – which town and county they belong to (Figure 2).

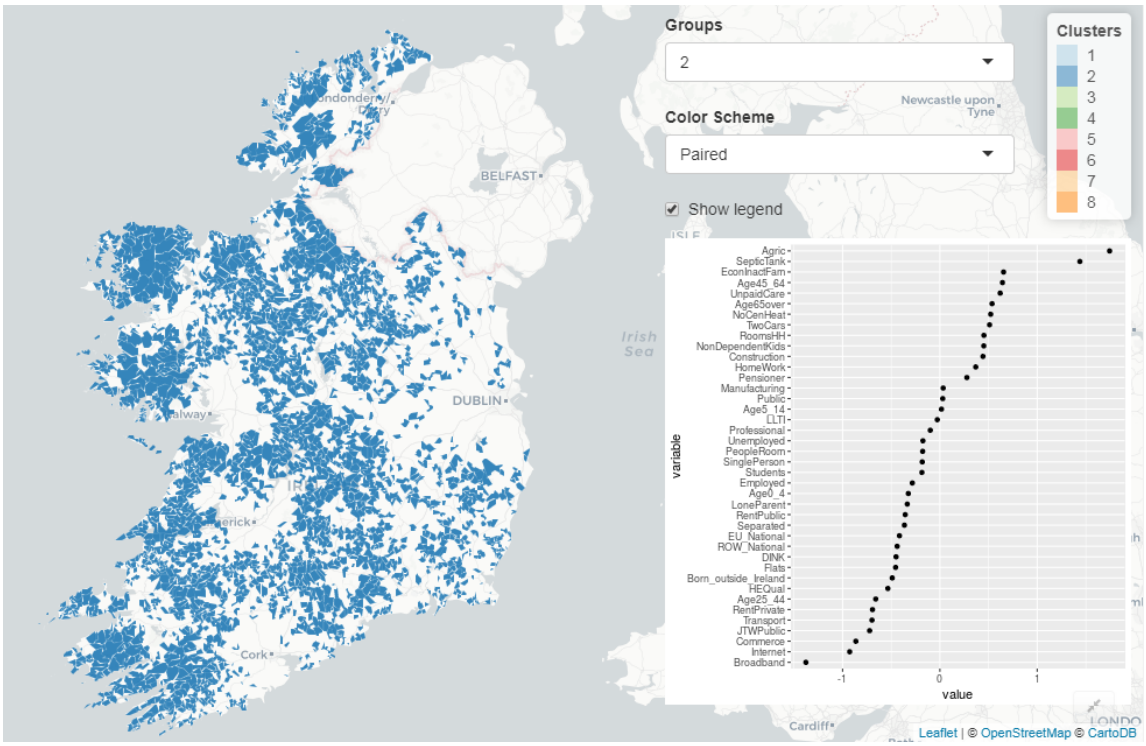

Figure 1 – Classification group 2 based on Ireland Census 2016

There is an extra layer of interactivity added as well through brushing and dynamic linking operations. Based on your selection, the average values of variables (e.g. population density, and some social and economic indicators) for your selection group is passed to another type of plot on a separate panel to display the linked information in text form. This gives the opportunity to user to learn more about any specific spatial units.

Since the main purpose here is to improve the interaction between the user and the information, it is important to consider how the data is presented. How to bring flexibility by avoiding complexity is one of the important questions in this context. Reaching out to more people and getting beyond the more typical purposes of the analysis are the useful targeted outcomes of these visualizations. Interacting with real urban dwellers through these visuals is one of the underlying intentions of this research.

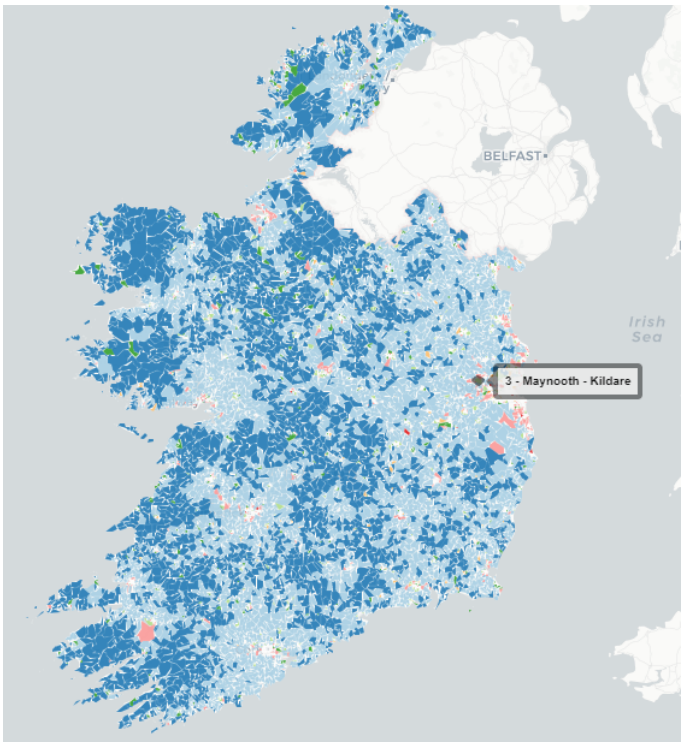

Figure 2 – All different groups based on Ireland Census 2016

There are a number of discussion topics on the agenda around visualization of data, especially urban data. As Swords and Liu (2015) mentioned “the confluence of big data and smart city agendas has seen cities visualised anew”. Big data and smart city movements in recent decades has accelerated the research in this topic and led into new types of data visualization discussions. At this point it is important to emphasize Kitchin and McArdle’s (2016) summary of six key issues;

- epistemology
- scope and access
- veracity and validity
- usability and literacy
- uses and utility
- ethics

These issues that have to be kept in mind by experts and non-experts alike when producing or using visualisations of urban data. One of the other popular discussion topics is design and interactivity. It is an understandable concern since it is a often considered best practice to follow simplicity but unfortunately it is not an easy task to achieve in visual representations of complex information.

Since many fields are included in these discussions, there are many alternative and critical approaches considered. While alternative approaches are trying to bring new perspectives to dealing with the data or new forms of representing the data, the critical approaches are mostly stressing the missing elements in urban visualisations, and the associated social issues. To have a comprehensive approach it is useful to consider each perspective. In order to achieve this goal, approaches that need to be followed through many disciplines should also be discussed. The fact that interaction is an efficient addition to visuals that are trying to represent data related to dynamic processes of daily

life can be enriched through these discussions. These debates are pertinent to many different disciplines.

The other important issue that has to be addressed is openness of the data as well as of the procedure followed to process the data. Reproducible steps will eventually increase the level of openness and help to uncover the unknown facts and procedures behind the representation of data. This might be useful as a bridging effect between the different understandings of each discipline and could bring more collaborations alive.

Conclusion

Data has always had a political aspect, but with the opportunity of reaching more people through different platforms and the ability of visualisation to provide insight, it has even more political power than before. As Swords and Liu (2015) commented, "tech companies are treating data like the new oil". In that sense, the way it is presented and the way it is interacted with have also equal importance.

There is no doubt that new developments in technology will lead to yet more alternative platforms to interact with data. There are already some movements in the mixed reality (virtual reality and augmented reality) environments in this aspect.

At this point, it is useful to emphasize the point that whatever a platform is used to represent the data, it has to be done with a critical approach since they are powerful highlighting tools.

The other important aspect of this paper is openness, supporting the use of open data and open source and promoting reproducible research. This study not only identifies the groups for the country, it also encourages updating these classifications in further census years by providing full background information of the analysis as well as providing the code for the interactive visualizations to build upon them.

Acknowledgement

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

References

Becker, R. A. and Cleveland, W. S., 1987. Brushing scatterplots, Technometrics, 29, 127–142.

Bertin, J., 1967. Semiologie Graphique: Les diagrammes, les reseaux, les cartes. Paris: Gauthier-Villars.

Brown, P.J.B., 1991. Exploring geodemographics, in I. Masser and M. Blakemore (eds.), Handling Geopgraphical Information: Methodology and Potential Applications, Longman, London, pp. 221-258.

Brunsdon, C., Rigby, J. and Charlton, M., 2011. Ireland Census of Population 2011: A classification of Small Areas, RPubs. https://rpubs.com/chrisbrunsdon/14998

Charlton, M., Openshaw, S. and Wymer, C., 1985. Some new classifications of census enumeration districts in Britain, A poor man's ACORN, Journal of Economic and Social Measurement, 13, 69-96.

Donolo, R. M., Favetta, F. and Laurini, R., 2013. A test to check the efficiency of visual representation of urban data, *Urban and Regional Data Management* – Ellul, Zlatanova, Rumor & Laurini (eds), Taylor& Francis Group, London.

Dykes, J., 1996. Cartographic Data Visualiser Toolkit.

Friendly, M., 2006. A brief history of data visualization. Handbook of computational statistics: data visualization.

Friendly, M. and Denis, D. J., 2001. *Milestones in the history of thematic cartography, statistical graphics, and data visualization.* Web document, http://www.datavis.ca/milestones/. Accessed: May 18, 2018.

Goodchild, M.F., 2015. Perspectives on the new cartography, Environment and Planning A, 47, 6, pp. 1341-1345.

Hartigan, J. A. and Wong, M. A., 1979. A K-means clustering algorithm. Applied Statistics, 28, pp. 100-108.

Snow, J., 1854. Cholera Map of London. In A Quick Illustrated History of Visualisation, DataArt. Web document, http://data-art.net/resources/history_of_vis.php. Accessed: May 18, 2018.

Kaufman, L. and Rousseeuw, P. J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons, New York.

Kitchin, R. and Mcardle, G., 2016. Urban data and city dashboards: Six key issues, The Programmable City Working Paper 21, Maynooth University.

Openshaw, S., 1983. Multivariate analysis of census data: The classification of areas. In D.W. Rhind (ed.). A Census Users Handbook. London: Methuen, pp. 243-264.

Singleton, A.D. & Spielman, S. E., 2014. The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom, The Professional Geographer, 66:4, 558-567.

Swords, J. and Liu, X., 2015. Visualizing urban and regional worlds: power, politics, and practices, Environment and Planning A, 47, 6, pp. 1235-1240.

Tukey, J. W., 1962. The future of data analysis. Annals of Mathematical Statistics, 33.

Webber, R. J. 1975. Liverpool social area study, 1971 Data. PRAG Technical Paper 14, Centre for Environmental Studies, London.

Yazgi Walsh, B., Brunsdon, C. and Charlton, M., 2017. Ireland Census of Population 2016: A classification of Small Areas, RPubs. https://rpubs.com/burcinwalsh/343141