Analysis of commuting patterns, changed by industry and location in England and Wales

May 2024

Abstract

The following project outlines the change in commuter patterns from 2011 to 2021, by industry and location. This is achieved through various dashboard and visualisations, which show the change in behaviours in a way that offers further insight as to commuting behaviours. Two dimensionality reduction techniques are employed to allow for comparison between locations or even locations and industries, to allow insight as to how different locations and industries share similar commuting patterns.

1 Introduction

Census data for England and Wales was recorded in 2011 and 2021. When the census was recorded in 2021, there was a national lockdown due to the coronavirus pandemic. This had a significant impact on the working population of England and Wales, as many people had to work from home. This report aims to investigate the shift from workers commuting, to working from home in this time period as well as analysing area-wise commuting patterns.

1.1 Task Setting According to Munzner's Taxonomy

Munzner's Taxonomy [1] outlines a clear structure for generating informative visualisations. This visualisation process is centred around: the domain, data/task abstraction, visual encoding and the algorithm.

The domain of this project is the analysis in commuting patterns between locations and years.

The data abstraction is multi-faceted and the exact details are outlined in more depth in the section "Data Description". The key structure of the data used contains information for each year, outlining the number of people in each census year who fit various combinations of distance travelled to work against method travelled to work and distance travelled to work against industry.

The task abstraction demands the use of analysis (for discovering patterns and trends, as well as enjoyment, for those who are curious as to what insights may be offered from the data); search (one might be interested to lookup their specific location to see how this location may be presented in this particular task) and querying (to summarise the overall behaviours for various industries or locations).

The visual encoding is displayed through three dashboards, each of which has various visualisations that reflect the task/data abstraction, with justifications to follow.

Algorithms are used to aid the visual encoding, for purposes of interactivity, to show higher volume of data without suffering from visual crowding, as well as dimension reduction, to reflect the important structure within the high dimensional data.

1.2 Ethical Considerations

Within the project ethical considerations were considered throughout.

All data is ethically sourced from the census data, and all computations were done in an unbiased man-

ner, as to not unfairly represent any Lower Tier Local Authority district.

The anonymity of census-takers was important. The selection and granularity of data had to be kept sufficiently low as to not make any census-taker identifiable within the data. Fortunately, the census website already has limitations in effect that preserve anonymity. If a granularity is too fine, then data is removed.

Another consideration is that some data could not be represented, due to the change in location boundaries. Consequentially Westminster and the City Of London could not be represented in this dataset. There was also difficulty surrounding the representation of the Isles of Scilly in the choropleth map for similar reasons.

Under-representing islands is a significant issue in the cartographic field. Reference [6] offers a comprehensive outline as to this issue; however the overriding message that applies to this report is that to omit the Isles of Scilly for arbitrary reasons would be unethical. The Isles of Scilly unfortunately could not be represented properly graphically (which is outlined further in the data preparation section) due to more complex reasons. The ethical implications of this are acknowledged, and it is essential to communicate this limitation of the presentations presented to the reader.

2 Data Preparation and Abstraction

2.1 Table Structure

The primary data of interest consists of four tables, each from the censuses of 2011 and 2021. Each table consisted of counts for how many people selected each unique combination of attributes, by location.

Granularity of location was an important factor. It was chosen that Lower Tier Local Authorities would best represent the regional diversity of locations in England due to the fact that it was neither too broad, losing specific regional insights, nor too specific, to violate the anonymity of census-takers. Henceforth Lower Tier Local Authorities are referred to as "Locations".

2.1.1 Method of Travel to Work

Two of the four tables had the following attributes: "distance travelled to work" and "method of travel to work". The distinction between the two tables is due to the measures for both 2011 and 2021. These two tables can be referred to as "Method of Travel to Work × Distance Travelled to Work 2011 (MTW×DTW11) from [2] and "Method of Travel to Work × Distance Travelled to Work 2021" (MTW×DTW21) from [3].

The "method of travel to work" attribute is unique to these two tables, and not present in the two tables containing information about industry. This is a categorical variable and contains 7 categories for example "Bicycle" and "On foot".

The second attribute in these two datasets, "distance travelled to work" has 10 distinct categories. These categories range from "less than 2km" to "More than 60km" - as well as including "working from home" and "other". When the "other" category is excluded, the data follows a clear ordinal structure, however, "Working from home" is somewhat outlierious - working from home doesn't necessarily have implication of travelling 0km to a place of work. For the purpose of this report this category is considered ordinally. The change measured of people "working from home" is expected to act as a proxy to the impact of the pandemic, as many people were forced to work from home, following lockdown.

2.1.2 Industry

The next two of the four tables had the following attributes: "distance travelled to work" and "industry". These two tables, again, each represented measures for 2011 and 2021 and can be referred to as: "Industry × Distance Travelled to Work 2011" (I×DTW11) from [4] and "Industry × Distance Travelled to Work 2021" (I×DTW21) from [5].

The structure of the I×DTW tables is similar to the MTW×DTW tables, but with an attribute for "industry", rather than "method of travel to work".

The industry attribute is categorical and contains distinct categories for each industry, including "C Manufacturing" and "F Construction" whereby the preceding character represents the industry section code.

2.2 Table Cleaning

It was first necessary to clean the tables, to enable comparison between data. Firstly, unique values in variables between the tables for two years were addressed.

The variable that first required addressing was location. Each location had a corresponding geocode. For consistency, this geocode was chosen (where possible) to match the entry of the 2021 dataset.

Between 2011 and 2021, many locations had either merged, renamed, or separated. Renaming locations proved a straightforward task, and the more recent name was chosen. An example is "Bristol, City of" was changed to "Bristol".

Amalgamating locations required more thought. The task of merging counts was simple, as the counts for each unique combination of attributes was added to create the new location. The location names were merged and the dictionary within the code reflects which locations were merged. An example is "Aylesbury Vale", "Chiltern", "South Bucks" and "Wycombe" all merging as "Buckinghamshire", with geocode preserved and chosen to match "Buckinghamshire"'s.

There were only two locations that split: "City of London and Westminster" and "Cornwall and Isles of Scilly". Consequentially there was no geocode for the individual locations in 2021 ("Cornwall", "Isles of Scilly", "Westminster" and "City of London"). For Cornwall and the Isles of Scilly, it is possible to attribute the counts to the geocode of Cornwall (due to the fact that the area of the Isles of Scilly is hard to see, because of the size and distance from Cornwall) (ethical considerations are considered, and sincere apologies are extended on behalf of the author to the population of the Isles of Scilly). When considering "Westminster" and "City of London" there is unfortunately no clear decision to be made when considering the geocode, consequentially both are removed. There are many surrounding areas of London that are included within this project, they have unique differences to the two omitted locations, however, London is still well represented within the datasets.

The methods of travel to work require renaming and aggregating. This process is similar to the process above. Examples of methods of travel to work that were aggregated include "Train" and "Underground, metro, light rail, tram", so that datasets had matching entries. "Not in employment or under 15" and "All" categories were removed too as they did not aid the towards the task abstraction.

The industry attribute required no further preparation.

When the tables were in sync in terms of denominations of the attributes, it was necessary to process the alignment of the tables for consistency and easier interpretation by Tableau. The tables of 2011 data contained location as a key, and each combination of previously outlined attributes was a unique column. It was consequentially necessary to use pivot tables in excel to adjust each previously mentioned attribute as a column.

For some dashboards the calculation of change and percent change was necessary. These were created from a combined table between tables with the same attributes, with an additional attribute for "year". "Change" and "Percent Change" were considered as unique entries in the "year" attribute section.

3 Dimension Reduction Techniques Outline

Before implementing dimension reduction techniques, it was necessary to have each location, or each unique combination of location and industry, as a key. This was again done using pivot tables in excel.

These techniques correspond to the algorithm facet of Munzner's Taxonomy, as these are computations used to assist in the learning from the data visualisations.

3.1 Principal Component Analysis

The first dimension reduction technique was used Principal Component Analysis (PCA). This algo-

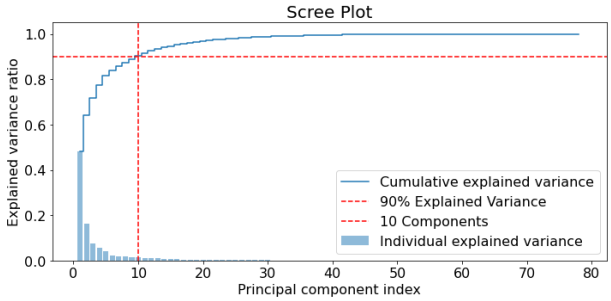

Figure 1: Scree Plot

rithm is used to take linear projections of the data to represent as much of the variance in as few components as possible. These components can be plotted against each other (visualising the projection space) to see how keyed data may cluster or repel.

PCA was applied to the MTW×DTW11 dataset, to analyse the similarities between each location's commuting behaviours. To do this it was first necessary to assign each location as a key in the table, and each unique combination of attributes (aside from year) were used for attributes. PCA was applied and the components that retained 90% of the variance of the data were kept.

The proportion of variance kept is shown in the scree plot in Figure 1. This figure depicts the cumulative proportion of explained variance that is preserved when using n principal components, (ordered by explained variance). It is possible to see that 10 principal components maintain 90% of the variance, as indicated by the red line, consequentially those are the principal components that are kept.

PCA assumes linear relationships between data. This is a shortcoming as it is not able to address more complex trends. PCA also focuses heavily on the global structure of the data, somewhat neglecting local structures and patterns.

3.2 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is considered the gold standard for dimensionality reduction. It preserves the local and global structure.

UMAP is applied to a version of the I×DTW datasets, which is structured to have each unique combination of location and industry in the key, and each cell represents the change in count between 2011 and 2021. This structuring allows for visualisation, showing which industries and which locations have similar and dissimilar trends in change in count in distance travelled to work.

4 Visualisation Justification

4.1 Dashboard 1

4.1.1 Introduction

Dashboard 1 displays an initial overview of the data, specifically the MTW×DTW datasets from 2011 and 2021. The chloropleth map has options for the user to select which "method of travel", and which "distance travelled to work". There is also a selection box for year, where the user can select which year, or change measure they would like to visualise. The stacked bar charts display the breakdown of the data. Visualising how people commuted, by how far they commuted. There is an obvious outlier in the "worked from home" as everyone who worked from home did not commute.

The Chloropleth map and selection choices exclusively specific to the chloropeth map have light blue boxes, for succinctness. The year selection applies to both plots so it is not necessary to highlight this with any colour. The colour key for the stacked bar chart is highlighted in the same colour as the stacked bar chart for similar reasons.

The chloropleth map can also be used to filter the stacked bar charts for even closer analysis.

It was important to avoid visual cluttering. The spatial considerations for both charts and all selection options was chosen to present the data in a way that isn't overwhelming to the eye, but still retains depth.

Due to the ordinality of the method travelled to work, discrete area charts were considered, however this visualisation fails when "change" is selected, as certain categories change from negative to positive

quite quickly, in a way that is visually challenging to interpret.

Munzner's Taxonomy focuses on asking why, what and how? for each visualisation, so these are answered as follows (in a different order than Munzner specifies however I believe the order I have chosen is best for representing my choices).

4.1.2 What, Why and How?

Chloropleth map:

- *What?*: Data is geo-spatial and contains discrete counts of various categories by year, method of travel to work.
- *Why?*: Chloropleth maps are an effective tool to visualise geo-spatial data. It is easy to draw comparative insights between areas. It is also possible to filter and draw insights for specific locations.
- *How?*: The chloropleth uses the spatial encoding of the data, and applies a colour embedding to the mark corresponding to the count, the interactivity allows for visualisations of more attributes.

Stacked bar charts: Reference [8] outlines the use of stacked bar charts, however, it is acknowledged that stacked bar charts may fall short in comparisons of heights of categories within bars (which visually encode the proportion of the measure accounted for by each subcategory, of the category in question). It can be challenging to compare heights against each other when they are presented at different locations. The user is likely to struggle to draw precise inferences, however, it is still possible to assume the overall summary or trend.

- *What?*: Data is categorical with multiple subdivisions within the data.
- *Why?*: The bar chart allows for comparison between various distances travelled to work and method travelled to work. The user can also identify specific counts or changes for specific areas. This visualisation is good for summaries.
- *How?*: The length of each segment of each bar corresponds to the count of each combination of attributes. Attributes are either shown on the x axis, through colour encoding or through a selection window.

4.1.3 Insights Drawn

Dashboard 1 offers many insights, an example is that by selecting "change", "works mainly from home" for all types of travel to work, one can infer that Leeds had the greatest change in number of people working from home. When clicking on Leeds, the tooltip shows that $96,588$ more people worked from home than in 2021, than in 2011. Referring to the stacked bar charts, the Leeds specific information shows that working from home was the only means of transport whereby the count increased between 2011 and 2021, across all subdivisions of distance travelled to work.

4.2 Dashboard 2

4.2.1 Introduction

Dashboard 2 shows another cloropleth map, a scatter plot and a bar chart.

The scatter plot shows the projection space for principal component analysis. This shows how different areas have similar or dissimilar commute patterns, according to the processed data from MTW×DTW11.

It is possible to select various principal components to plot against each other in selection boxes, as well as clustering algorithm.

Clustering algorithms that can be chosen include, K-means, DB Scan, Gaussian Mixture Model and adjusted K-means. The specific details of each clustering algorithm is not outlined, however the adjusted K-means requires justification.

When visualising in principal components 1 and 2, as well as 2 and 3: a central cluster, as well as 3 seemingly diverging clusters from the centre were apparent. All algorithms with various choices of hyperparameters failed to capture this well, besides the K-means algorithm with $K = 3$ (which is the value of K selected for the visualisation). The adjusted K

means incorporates a split within the cluster associated with Buckinghamshire. A new cluster label was created for those within this cluster, with values where PC1 < -4. This offers a much better split of the data and allows better insights to be drawn.

The chloropleth shows how each area is classified according to the clustering algorithm ran on the principal components, and can also be used to filter the bar chart.

The bar chart shows number of commuters for each method of travel in 2011. The emojis are used to indicate method of travel for simplicity, consequentially, the visual encoding is much more pleasing to the eye and interpretable.

4.2.2 What, Why and How?

Scatter Plot:

- *What?*: Data is multivariate and continuous, each point is assigned a specific pre-determined label. The choice to visualise two continuous variables through a scatter plot is justified in [7].
- *Why?*: Two dimensional spaces allows for the user to measure proximity in terms of euclidean distance, for comparison of locations. This is a reduction from a much higher dimension of data.
- *How?*: Shape and colour embeddings are applied to marks used to reflect different clusters. X and Y axis used to display each component of choice.

Chloropleth:

- *What?*: Geospatial data, each given a cluster label.
- *Why?*: To compare labels between areas that have geographical proximity, to identify overall trends.
- *How?*: Spatial and colour encoding, similar to Dashboard 1.

Bar Chart: Reference [8] outlines the choice for displaying continuous measures of categorical variables through a bar chart.

- *What?*: Categorical data with continuous measurements.
- *Why?*: To give insight as to commuter methods for a given area, allowing for some comparison of areas to the national behaviour. Use of lookup also possible.
- *How?*: Vertical height of bars corresponds to number of commuters that use each specific method of transport. A colour embedding is applied to the marks to indicate method of transport (to allow bars to stand out).

4.2.3 Insights Drawn

When selecting principle components 2 and 3, as well as the adjusted K means clustering, it is possible to see how different areas have similar commute patterns. This algorithm effectively identifies the London locations (without Barking and Dagenham) as a cluster which is intuitive, as London has a unique transport system with the London underground network.

Cluster 3 seems to capture the commuter patterns of major cities, for example Bristol, Birmingham, Cardiff and more. Interestingly East Suffolk, South Gloucestershire and a few other locations that aren't associated with being a major city are also encapsulated by this cluster.

Clusters 1 and 2 offer less distinct interpretations, however there are visual clusters of cluster 4, with Buckinghamshire, Central Bedfordshire, West Northhamptonshire and North Northhamptonshire each being spatially close and having similar commuting patterns. It is possible cluster 1 represents locations without distinct commute patterns.

Using the chloropleth map it is possible to select a London Borough, and draw insight from the bar chart. For example, Ealing (like many other areas in London) have the majority of people commuting by train, tram, underground, metro or light rail.

Comments on other clustering techniques:

- **K-means (with k=3)** effectively clusters London areas (without Barking and Dagenham).

Fails to effectively split between clusters so engineered cluster criteria is required.

- **DB-Scan** identifies clusters based on density of points. Hyper-parameters were selected with justification in code. This algorithm identifies the dense central cluster and a cluster that seems to correspond with some areas of London, and the rest as noise indicating this may be in inappropriate choice of clustering algorithm.

- **Gaussian Mixture Model** assumes the data is generated from normal distributions with centroids in the data and constant variances. This algorithm successfully identifies the London cluster, however includes Gosport and South Tyneside. This is unexpected, yet interesting. The remaining two clusters correspond similarly to the K-means clustering, but again fails to capture the fourth cluster, whose boundary requires engineering.

4.3 Dashboard 3

4.3.1 Introduction

The third and final dashboard, - Dashboard 3 - aims to present how different locations and industries have directly changed in distance travelled to work. Here each unique combination of location and industry is used as a key from the I×DTW, with cells representing the change between 2011 and 2021. UMAP is then applied to this dataset, to see which combinations of location and industry were affected similarly.

The scatter plot shows the UMAP embedding in the respective projection space. Marks have a colour embedding corresponding to industry.

The change in number of people working from home by location is shown in the chloropleth map, which can be used as a filter for the scatter plot and bubble plot, by location.

The bubble plot is used to show absolute percent change in people working from home by industry. This also works to filter the scatter plot by industry. The labels of the bubble plot show the industry. This plot does not support the use of emojis, consequentially the labels needed to be concise as possible to be displayed, but not all of the labels can be displayed, however it is still apparent with the tooltip and colour key.

A shortcoming of the bubble plot is that negative changes are not easily distinguishable from positive changes, as a result "absolute" change is plotted, and the reader is urged to use the tooltip and be careful before making dubious assumptions.

4.3.2 What, Why and How?

Chloropleth

- *What?*: Geo-spatial data with colour embedding corresponding to change in number of people working from home.

- *Why?*: Comparing behaviour in areas, of the most significant division of distance travelled to work.

- *How?*: Spacial and colour encoding, similar to Dashboard 1.

Bubble Plot

- *What?*: Standardised (by percentage) measures corresponding to categorical variables.

- *Why?*: To allow for comparison in absolute change between industries. Also offers a summary as to which industries displayed most change.

- *How?*: Radius of each bubble encodes absolute percent change values by industry. Colour encoding to relate to scatter plot, for which the bubble plot can act as a filter.

Scatter Plot

- *What?*: Two dimensional continuous UMAP encoding from higher dimensional space.

- *Why?*: Comparison between individual marks and between industries can offer insights as to which were impacted similarly. Option to filter between industries and location to further "zoom in" on data.

- *How?*: Positional representation of marks in the X and Y axes. Marks corresponding to industry, similar to the bubble plot. Circular marks chosen to further aid the similarity between the two.

4.3.3 Insights Drawn

The UMAP embedding display clear behaviours marks of various industries. For example it is apparent that marks associated with industries A and B (agriculture, forestry & fishing, and Mining and quarrying) are spatially close together. This indicates these two industries were similarly affected by the shift in commuting during the pandemic. Due to the fact these are very manual jobs that cannot be done from home, it is likely there was little impact of the pandemic on commuting habits. This is supported by the bubble plot which states that industry A had a -0.34% change in number of people working from home, and Industry B had a -8.84% change. These are fairly small percentages, especially for Industry A which has an even more distinct cluster. More similar analyses could be drawn but this is left as an exercise for the user.

If the bubble corresponding with the IT industry is selected, it is possible to visualise the UMAP embedding for exclusively, there is a distinct banana-shaped cluster, showing that commuting in the IT industry has been affected similarly across all locations. By highlighting the very top section of the projection space (as shown in 2) it is possible to see that there is a distinct continuous line of counties between Bristol and London. It is possible to deduce that there could be a commuter line that is significant to IT commuters, that significantly changed in time between 2011 and 2021. Inferences drawn here are spurious and speculative, however, this observation is interesting.

Among all industries, it is harder to draw comparisons between locations. When filtering on location in the chloropleth, points are scattered and do not retain any structure, this indicates that the change between 2011 and 2021 in number of people working from home, was more impactful industry-wise, than location-wise. However this is also possibly a consequence of the large number of locations and small number of industries, as it is harder to draw visual conclusions from a small sample size of marks.

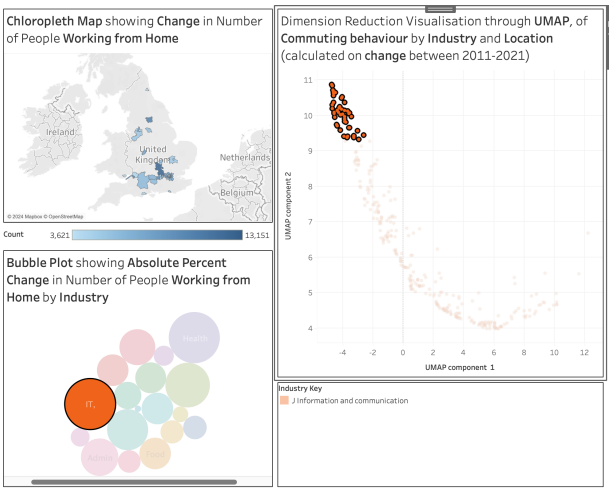

Figure 2: Behaviour of Change in Commuters in the IT Industry

5 Evaluation

In the evaluation session a survey was carried out to test how interpretable the visualisations were. Survey members were assured that their answers would be anonymous and their data would be stored securely.

Survey members were comprised of Data Science masters students, and were asked multiple choice questions to identify key variables and trends. While most of the answers indicated that key variables were identifiable (100% of people surveyed when asked, correctly identified that the number of people commuting by car or van has decreased across all distances travelled to work). However there were some trends that were less easy to identify.

One issue that was raised by the survey was that participants couldn't easily distinguish between attributes visualised (20% incorrectly answered when asked which category of distance travelled to work is outlierious). This is understandable as many of the labels and titles have a high density of words and

cause visual overcrowding. To remedy this, labels are simplified with emojis where possible (and appropriate) to reduce visual overcrowding. Another strategy used was the key words in titles being put in bold frame - this makes important information stand out without being lost in among less important words.

Another issue that was raised was the PCA scatter plot was difficult for colourblind people to interpret, consequentially different shapes were used to encode the cluster labels. This strategy is less effective on the UMAP scatter plot due to the large range of industries and colours displayed. Fortunately it is still possible to use the tooltip to comprehend what is being displayed (although much less effective).

6 Further Extensions

The pandemic had significant impacts on employment rates and (sadly) the population size. These are omitted in this report due to constrain the scope of this investigation to a suitable level.

Between 2011 and the beginning of the pandemic (before 2021) there was also likely many other factors that changed the face of employment in England and Wales. These factors are not accounted for in this project and are considered as an extension.

Both UMAP and PCA projections lose interpretability. The projection space is much less interpretable to those who do not have an understanding of what they represent. This can't be remedied easily as this is exactly what dimension reduction techniques are designed to achieve.

Reference [9] out lines how bubble plots might be used to visually encode three continuous variables, with two presented on the x and y axes, and the third through size. It may have been possible to leverage tableau to encode more information within the bubble plot provided on Dashboard 3, for example the bubbles could have been sorted into 2 clusters of positive and negative change. Tableau's software was particularly restrictive with the bubble plots and so this is considered an extension to improve upon.

7 Conclusion

It is clear to see how the pandemic has effected commuter patterns in England and Wales, through various visualisation techniques. These visualisations allow for the user to satisfy achieve various goals (such comparison, search, query) through the dashboard, which offer further insight and presentation of information, to further aid the user's inference.

References

[1] T. Munzner, *Visualization Analysis and Design*. A K Peters/CRC Press, 2014.

[2] Office for National Statistics. (2011). 2011 Census: Method of Travel to Work by Distance (CMLAD) [Data set]. Available from Nomis official labor market statistics.

[3] Office for National Statistics. (2021). 2021 Census: Method of Travel to Work by Distance (LTLA) [Data set]. Available from Nomis official labor market statistics.

[4] Office for National Statistics. (2011). 2011 Census: Industry by Distance Travelled to Work (CMLAD) [Data set]. Available from Nomis official labor market statistics.

[5] Office for National Statistics. (2021). 2021 Census: Industry by Distance Travelled to Work (LTLA) [Data set]. Available from Nomis official labor market statistics.

[6] M. Monmonier, *How to Lie with Maps*. University of Chicago Press, 1996.

[7] Tufte, E.R. *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[8] Few, S. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

[9] Wilkinson, L. *The Grammar of Graphics*. Springer-Verlag, 2005.