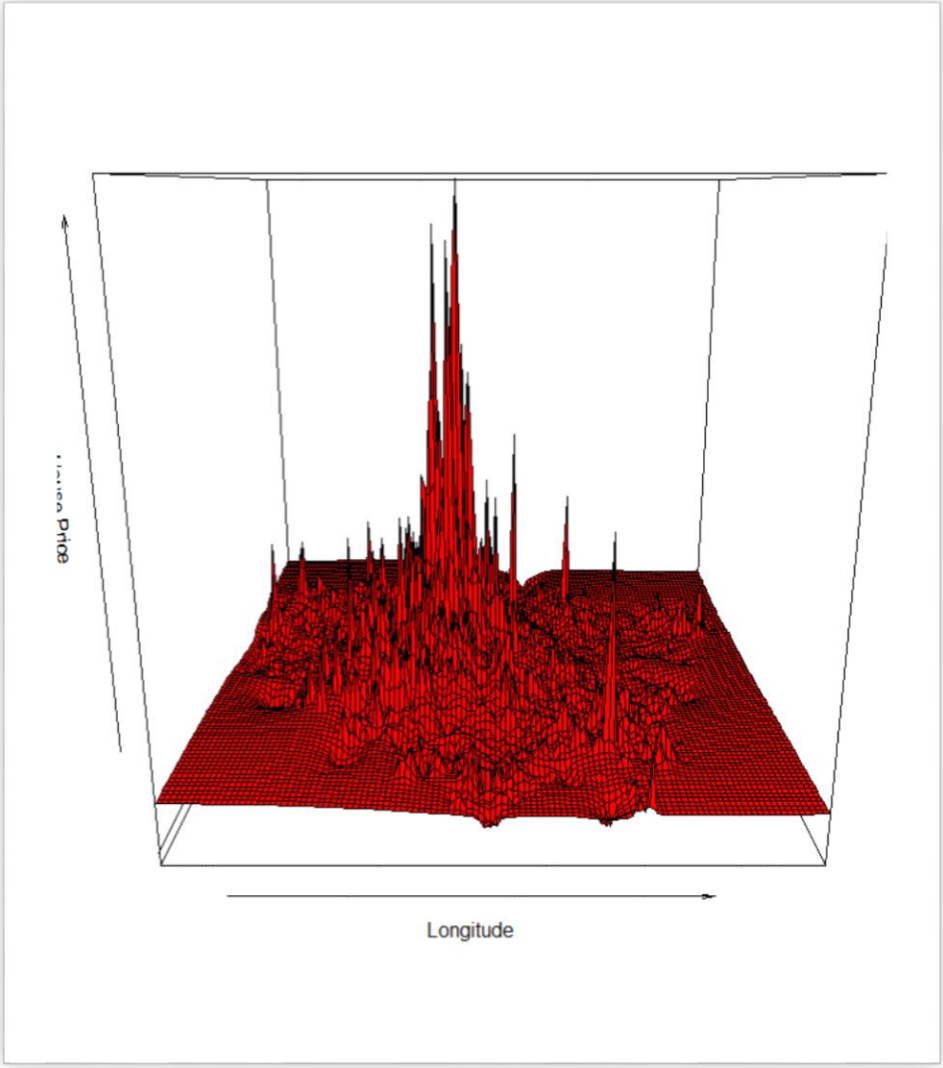
UCL DEPARTMENT OF GEOGRAPHY

YEAR 2016-17

| | |
|---|---|
| EXAM CANDIDATE ID: | RQJG1 |
| MODULE CODE: | GEOGG125 |
| MODULE NAME: | Principles of Spatial Analysis |
| COURSE PAPER TITLE: | Exploratory Spatial Data Analytics |
| WORD COUNT: | 1953 |

Are you registered as dyslexic with UCL Student Disability Services (SDS) and been given labels to 'flag' your written work NO *(please delete as applicable)*




# Exploratory Spatial Data Analysis

## A SPATIAL DISTRIBUTION OF HOUSE PRICES ACROSS LONDON

RQJG1 | GEOG125 Principles of Spatial Analysis | January 9, 2017

# Contents



Word Count: 1953

# 1 – Introduction

The purpose of the following report is to analyse variances in the spatial distribution of house prices in Greater London using open data published by the UK Land Registry. In order to analyse the data, commands in the programming language R were executed on the values of house prices sold within the year 2015 in relation to its geographical location detailed in British National Grid (BNG) eastings and northings. Univariate analysis was performed on the data in the form of a histogram of the house price data and spatial analysis of the data was employed in order to understand the dynamics of house prices over geographical location.

The research aim of this project is to analyse the mathematical and spatial elements of house prices in London using R programming language. In order to meet the aim, the following objectives will be reached:

- Univariate, mathematical analysis will be conducted on the house prices in Greater London for the year 2015;
- Visual analysis techniques will be employed on house price data plotted based on geographical location; and
- Spatial analysis techniques will be executed on the house price data in order to identify trends, connections or similarities within the data based on the location variable.

Contextually, the relevance of property values and, more specifically, the change in property values over space is crucial for many industries. For example, transport and infrastructure planning projects research property values in order to plan routes, station locations, roads etc. (Zhong & Li, 2016), however the connection is multi-directional as research suggests that property values are also affected by transport links in the area (Agostini & Palmucci, 2008). Moreover, a similar affect has been identified regarding the role of educational facilities and property values in an area (Wen et al. 2014). Therefore, much research has been conducted in the attempt to identify and understand the factors which affect house prices and their relative weighting with respect to each other (Wei & Cao, 2017, Atkinsomi et al., 2016, Nneji et al., 2013).

This study aims to critically review to what extent univariate and spatial analysis methods can contribute to understanding the fluctuations in property values in London. It is important to recognise in which stage this research fits in the contribution to knowledge; this report contributes

to the identification of trends within the data between the data collection process and the evaluation into reasons for the fluctuations in house prices in the study area. Visualised in figure 2.

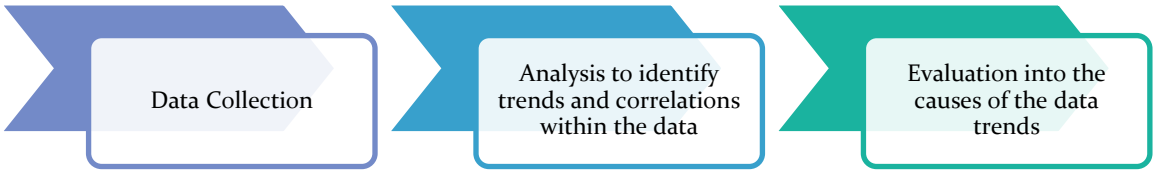



*Figure 1 - Flow diagram detailing data flow from raw data collection to fully interpreted data.*

## 2 – Data

Open source data used in the following report has been produced by the UK Land Registry as part of the *Price Paid Dataset* and includes information on "*all property sales in England and Wales that are sold for full market value and are lodged with us for registration*". (Land Registry, 2016a)

Accuracy of the data relies on the accuracy of data collection. The Land Registry publish on the official government website criteria under which the data would be excluded from the database. Reasons such as discount sales, the gifting of property and divorce settlements are included in the list and are excluded to maintain the 'quality of the Price Paid Data'. More information and the full list of exclusion criteria can be found on the government website (Land Registry, 2016b).

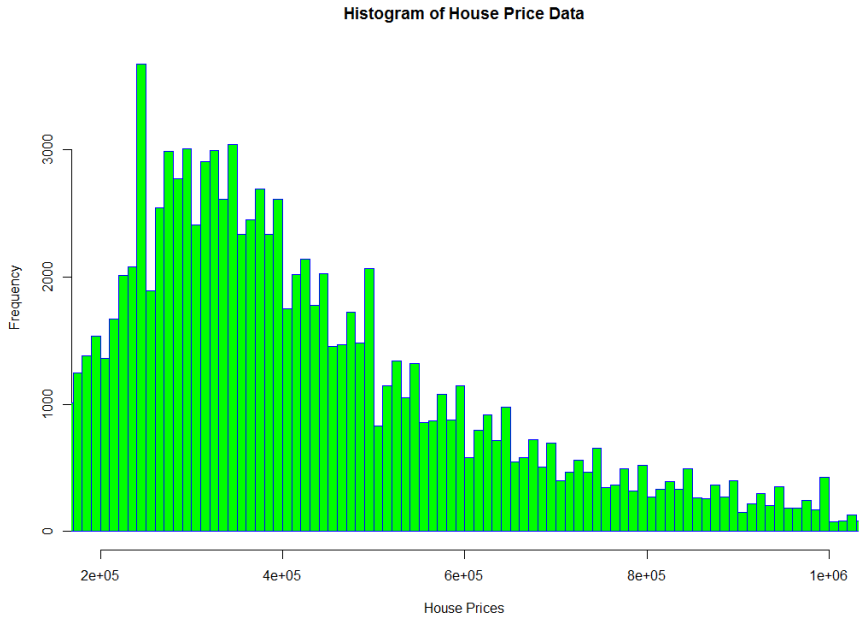



*Figure 2 - Histogram of House Prices sold in the year 2015. (Source UK Land Registry, 2016)*

Figure 2 visualises the first representation of the house price dataset. In the form of a histogram, a number of modifications have been made to the data in order to produce the histogram appropriately. The first adjustment is the exclusion of data regarding house prices more expensive than £1,000,000 in order to allow effective representation of the data. The reason for this is that when all the data is included, the tail of the histogram (data ranging from £1,000,000 to £36,500,000) requires such a high proportion of the x-axis, that the majority of the most frequent data values (between £200,000 and £500,000) were compressed into a single frequency column.

Therefore, a limitation of the histogram as a method of data visualisation and analysis is that, with this particular dataset, a section of the data must be excluded for the most appropriate graphic representation.

Visually, the data allows the viewer to understand some of the dynamics of house price fluctuation in London. Määttänen & Terviö (2014) produced a framework that analyses to what extent variances in income are reflected into house prices. The conclusion reached illustrates that the most influential factor is price gradient; thus, the difference in price of a property in comparison to nearby properties of a similar quality. This leads the researcher to believe that more can be understood from the data when it is analysed in a spatial context, rather than a univariate context and can therefore be considered a limitation of the histogram and univariate data analysis methodology. Spatial analysis will follow in section 3.

## 3 – Analysis

In order to compliment the mathematical, univariate analysis, spatial analysis will be performed on the data in order to visualise trends and connections within the data based on location. The basis on which the relevance of the spatial element is founded is Tobler's first law of Geography:

*"Everything is related to everything else, but near things are more related than distant things."*

Waldo Tobler (1970)

Figure 3 is a graphical representation of the house prices of Greater London using a scale of light blue to dark blue based on the location of each value. The image was produced in the programming language R and uses the Google Road Map image (Google, 2017) as a background reference to give locational context to the superimposed house price data.

A number of spatial themes and correlations become more apparent in the spatial analysis than the univariate analysis. For instance, it is possible to notice clusters of properties in similar price bands. For instance, the City of London Borough, in the very centre of London, contains a cluster of some the highest concentrations of property sales in the most expensive price category. This supports the conclusions reached by the research conducted by Agostini & Palmucci (2008) that stated that property prices increase with close proximity to major transport links as the borough such as the major London domestic and international train stations.

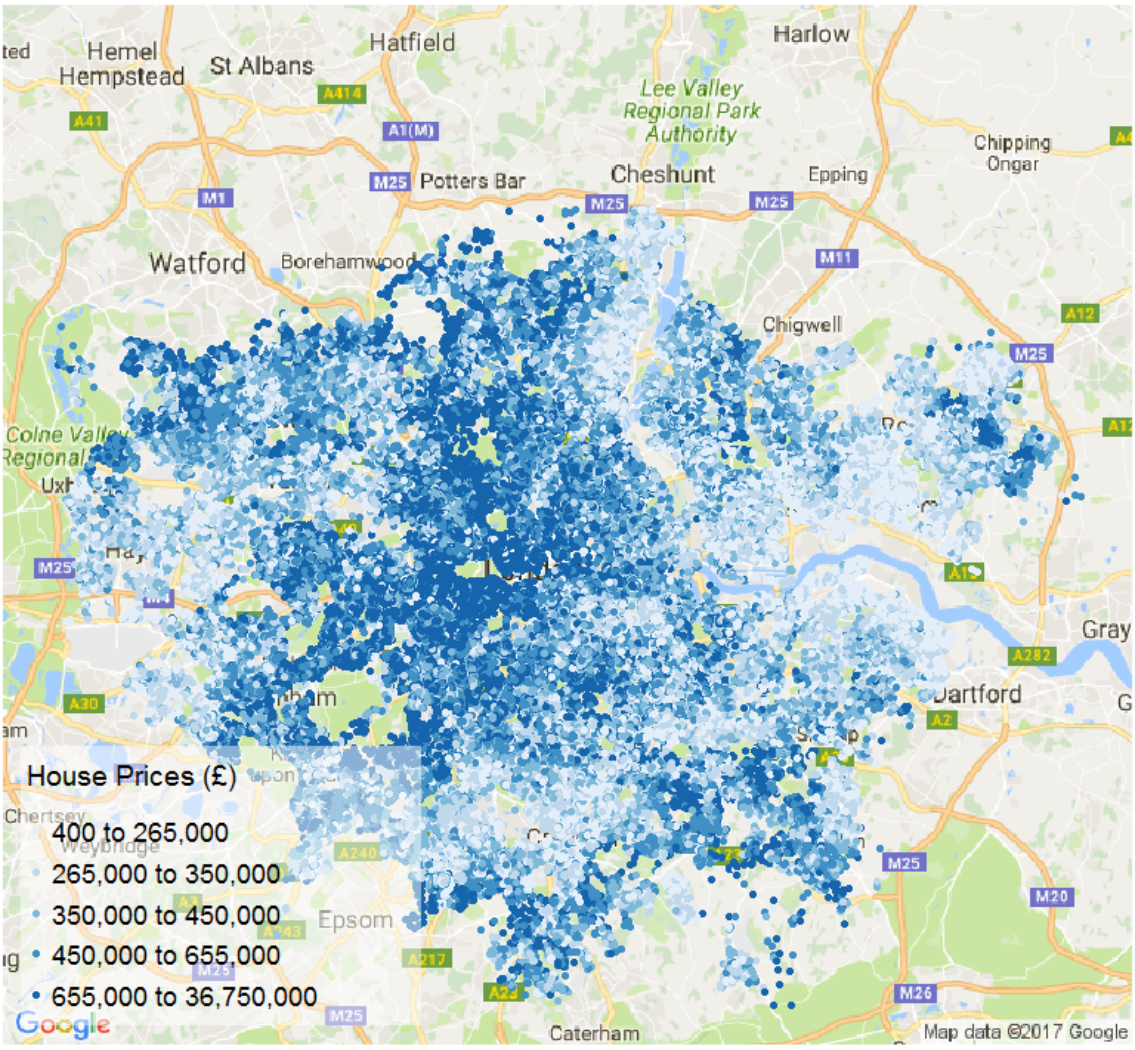



*Figure 3 - Visualisation of the property sale prices in the year 2015 plotted by location of property and colour based on sale price.*

However, the cluster of house price data in the Richmond borough near to Heathrow airport indicates a large proportion of property prices in the least expensive price category. This would contradict the conclusion made in the Agostini & Palmucci (2008) academic article as the residences would be in close proximity to both an international airport and to excellent transport links to Central London (such as the Heathrow Express train line). This is an example of a trend within the data which was only noticeable when spatial analysis was employed on the data. However, a limitation of the analysis is that the reasons for the variation in values requires further research and could also be coincidental rather than connected.

Figure 4 is a raster image displaying the interpolated house price data based on geographical location. The image was calculated based on two variables: the house price at which the property was sold; and the geographical location of the property that was sold. The main advantage of interpolating data is that the study areas becomes a continuous field rather than point cloud

dataset. Thus, any pixel could be selected and would be assigned a value based on the value of property sales nearby, inversely weighted to distance to the value. Therefore, the value of the pixel would be a reasonable estimate to the value of the property based on location and not property quality.

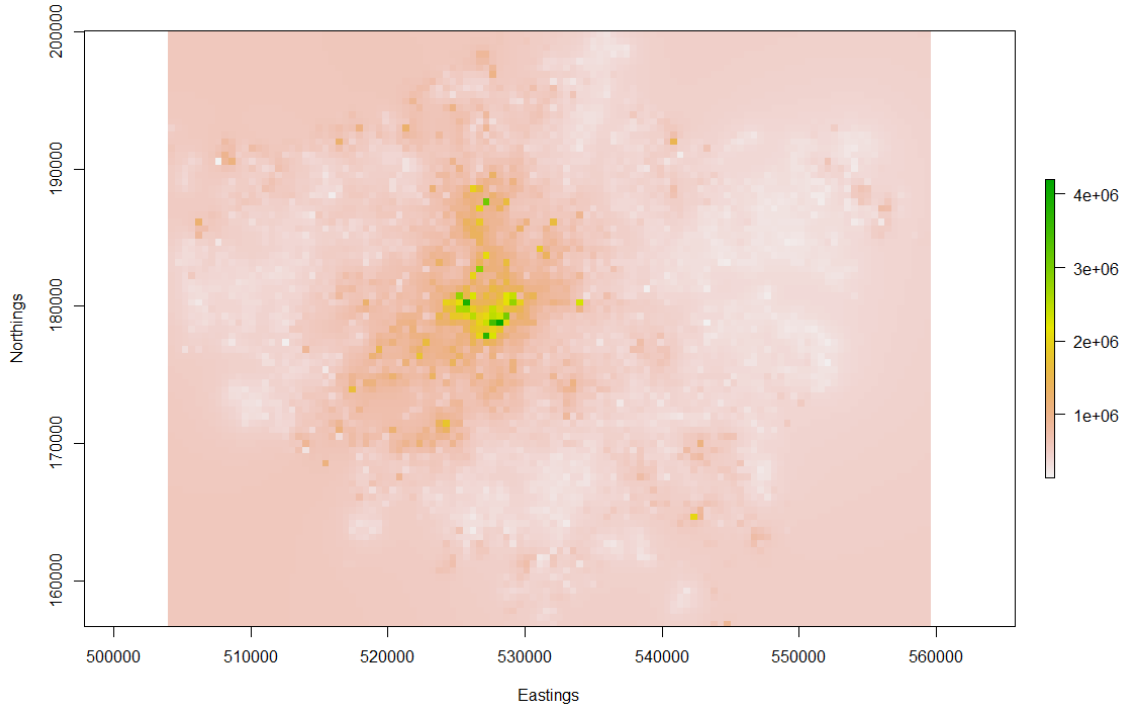



*Figure 4 - Interpolated raster image displaying the property prices of Greater London using an inverse distance weighting algorithm.*

Lu & Wong (2008) state that the inverse distance weighting algorithm is "*fast, easy to compute, and straightforward to interpret*". However, Lu and Wong (2008) argue that a limitation of the spatial autocorrelation method is that a constant distance-decay parameter homogenises the results. Thus, a variable distance-decay parameter would calculate a more accurate representation of reality. Therefore, it is suggested that a review into the relative merits of the two interpolation methods with the dataset for house price data in London is a possibility for further study.

Figure 5 is a visualisation of the house prices of Greater London using the *persp3d* command in the *rgl* RStudio package. The interpolated raster image displayed in figure 4 is rendered in 3-dimensions with the z-axis displaying the average house price for the area. The advantage of this data visualisation technique over the other 2-dimensional methods is that data clusters are more profound than the coloured clusters in figure 3. Moreover, the height of the spikes in the 3-dimensional figure 5 is on a continuous scale. Therefore, there is no issue of categorisation of data as the data is plotted on the axis based on its value. In figure 3, the boundaries at which the categories change dictate the data's visual impact on the viewer and is therefore a limitation of the spatial analysis employed in figure 3.

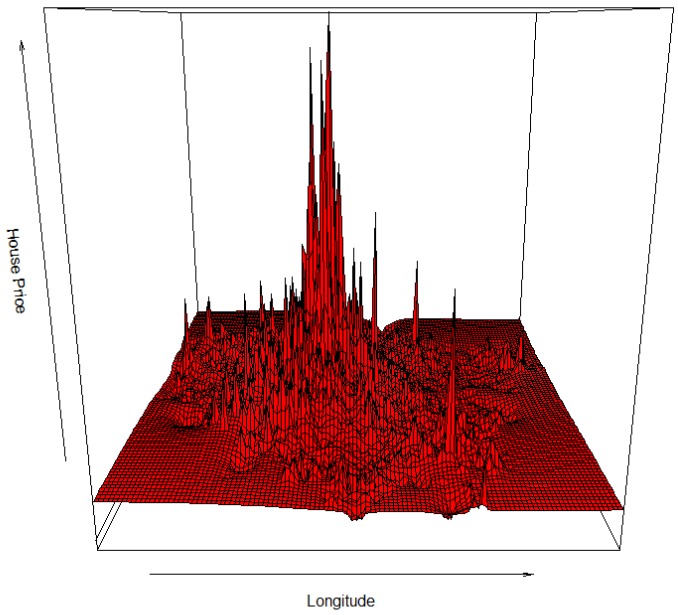



*Figure 5 - Visual representation of the house prices in Greater London in 3-dimensions.*

## 4 – Discussion

The aim of this research study was to review and critically assess the relative merits of univariate and spatial analysis methods to visualise house price data in Greater London. The univariate method reviewed was the histogram data representation graph which displayed the frequency of house price sales per £5,000 category. The graph indicated various attributes of the data and would be an excellent method of data visualisation if the relevant statistic of the data was the most common property sale price in Greater London. However, data had to be excluded from the histogram in order to appropriately display the graph. Moreover, correlations and similarities within the data based on spatial location could not be analysed using a univariate data display method alone.

Therefore, three spatial data depictions were reviewed and critically assessed in the study of the housing prices in Greater London. The first, figure 3, visualises the data plotted onto a base map of Greater London with the colour of the points representing the value of the property. This technique allows the viewer to understand the general trends of the data based on its spatial location, but is limited by the hiding of data values beneath other data values in densely populated areas.

Figure 4 is an example of using interpolation to eradicate the issue of overlap by calculating the value of each pixel based on the values of nearby pixels. However, a limitation of this

representation method is that the accuracy of the data to reality is subject to the algorithm that is executed.

The final representation of the house price data can be viewed in figure 5 which illustrates an example of how 3D data rendering can be used to visualise data. An advantage is that the data is displayed in a continuous field of all three variables at the same time but when published in 3D, data values are hidden behind high value data points. Thus, the data is distorted unless the viewer has access to rotate and view the data in 3-dimensions.

It is suggested that further study is completed into the relevant merits of constant and variable distance-decay parameters in the inverse-distance weighting (IDW) interpolation algorithm; and into the causes and connections between correlations and trends identified in the data analysis.

## 5 – References


Data produced by Land Registry © Crown copyright 2016.

Agostini, C. A. and Palmucci, G. A. (2008) 'The anticipated Capitalisation effect of a new metro line on housing prices'. *Fiscal Studies* 29 (2), 233–256

Akinsomi, O., Aye, G. C., Babalos, V., Economou, F., and Gupta, R. (2016) 'Erratum to: Real estate returns predictability revisited: Novel evidence from the US REITs market'. *Empirical Economics* 51 (3), 1191–1191

GOOGLE MAPS, 2017. Map of Greater London. [online]. Google. Available from: https://www.google.co.uk/maps/place/Greater+London/@51.487981,-0.6485948,9z/data=!3m1!4b1!4m5!3m4!1s0x47d8a00ba11ae26f:0x2ff173e384b8e98b!8m2!3d51.4309209!4d-0.0936496 [Accessed 05 January 2017].

Land Registry and UK Government (2016a) *How to access land registry price paid data* [online] available from <https://www.gov.uk/guidance/about-the-price-paid-data#data-excluded-from-price-paid-data> [5 January 2017]

Land Registry and UK Government (2016b) *'Price paid data'*. [online] GOV.UK. available from <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads> [5 January 2017]

Määttänen, N. and Terviö, M. (2014) 'Income distribution and housing prices: An assignment model approach'. Journal of Economic Theory 151, 381–410

Nneji, O., Brooks, C., and Ward, C. W. R. (2013) 'House price dynamics and their reaction to macroeconomic changes'. *Economic Modelling* 32, 172–178

Tobler, W. R. (1970) 'A computer movie Simulating urban growth in the Detroit region'. *Economic Geography* 46 (2), 234–240

Wei, Y. and Cao, Y. (2017) 'Forecasting house prices using dynamic model averaging approach: Evidence from china'. *Economic Modelling* 61, 147–155

Wen, H., Zhang, Y., and Zhang, L. (2014) 'Do educational facilities affect housing price? An empirical study in Hangzhou, china'. *Habitat International* 42, 155–163

Zhang, C., Jia, S., and Yang, R. (2016) 'Housing affordability and housing vacancy in china: The role of income inequality'. *Journal of Housing Economics* 33, 4–14

Zhong, H. and Li, W. (2016) 'Rail transit investment and property values: An old tale retold'. *Transport Policy* 51, 33–48