House Price Changes Across London and its Surrounding Areas

Mithileysh Sathiyanarayanan
Srishti i2i Biz Solutions
Email: Mithileysh.s@srishtibiz.in

1. Motivation, data, research questions

Since the year 2000, there has been a dramatic increase in house prices in London and its surrounding areas (henceforth: the London Region). This has led to an affordability crisis in which an ever-increasing proportion of the London Region is unaffordable to most of its citizens. It has been claimed that there has been a 'chain reaction' of price increases, whereby as central areas of London have become too expensive, demand has shifted to London's outer areas until they in turn have become too expensive, and demand has then shifted to the commuter belt [5].

This project provided an analysis of the patterns of growth in London Region's house prices. The project's scope was 51 local authorities for the period from 2000 to 2016.

Fig 1. Geographical map of the local authorities analysed in this project.

Analysis of house price growth could help to improve the regulation of London Region's housing. By better understanding house price growth, regulators would be better placed to ensure that the housing needs of their citizens are met e.g. by increasing the provision of housing stock, by adjusting stamp duty tax, by investing in transportation links to the commuter belt and so forth. This project focussed on the patterns of house price growth, including investigating whether there has been a 'price chain reaction'. It also briefly explored whether the patterns found could be explained using household income and dwelling stock data. The

analyses provided interesting insights into London Region's housing market, which could improve future analyses and modelling of London Region's housing. Such improvements could, in turn, lead to improvements in the regulation of London Region's housing.

This project's primary research question was:

What are the patterns of growth in house prices across the London Region?

In addressing this question, this project sought to determine if:

a. London Region's local authorities can be grouped into clusters that share similar patterns of house price growth.
b. there is any evidence for a 'chain reaction' of price increases

The Office for Budget Responsibility has successfully modelled national house prices, and has identified that key drivers of house prices are gross disposable household income (henceforth: household income) and housing supply [2]. A brief ancillary analysis was carried out for a secondary research question:

Can these patterns of growth be explained by differences in household income and housing supply?

The local authority house price data used in this project came from the Office of National Statistics' (henceforth: ONS) Monthly UK House Price Index data table[1]. This provided a reliable basis for measuring house prices as (i) its values were determined using a hedonic regression that takes account of the mix of property types sold in each period (ii) it used a geometric mean that reduces the weightings of the most extreme house prices in each local authority. The project also used ONS data for household income[2] and dwelling stock[3].

2. Tasks and approach

The analytic tasks employed in answering the research questions were:

2.1 Production of descriptive statistics and data maps. This included: plotting house price time-series and producing a heat map of the correlations between local authorities' house prices. Cartograms were also produced of house prices and house price growth.

2.2 Clustering. Clustering was used to identify those local authority house price time-series that shared similar shapes. It was established that the time-series could be grouped into a set of clusters where:

i.) each cluster's centroid was representative of the time-series of *all* of its members.
ii.) each centroid represented a significantly different pattern of house price growth.

The analysis then used the clusters' centroids as high-level summaries of their cluster members.

[1] www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-september-2017
[2] www.ons.gov.uk/economy/regionalaccounts/grossdisposablehouseholdincome
[3] www.gov.uk/government/statistical-data-sets/live-tables-on-dwelling-stock-including-vacants

The house price data was normalised, so that the time-series all had zero mean and unit standard deviation. If the data had not been normalised then clustering would have grouped time-series based on the magnitude of house prices rather than on similarity of shape.

A fuzzy c-means algorithm was chosen to perform the clustering as this:

i.) is an linear algorithm and hence differentiates between similar shaped time-series that are offset from each other [1]. This was important for analysing whether there were any price chain-reactions. The distinction between linear and nonlinear clustering algorithms will be further discussed in section 5.

ii.) measured the degree of membership of each time-series to its cluster. This enabled the identification of those local authorities that were borderline between two or more clusters.

A key part of the analysis was in setting the parameter *k* to the correct number of clusters to partition the dataset into. This involved partitioning the dataset for different values of *k*, and then:

a. creating geographical maps of the clusters and drilling down to compare the price data of individual local authorities.
b. visually comparing the shapes of cluster centroids to determine if they were sufficiently different, such that the *k*-partition was relevant to answering the research questions.
c. analysing the Euclidean distances between local authority time-series and their centroid.
d. analysing those local authorities with high degrees of membership to more than one cluster groups.

As well as establishing the correct value of *k,* the above analyses established that the centroids provided a good representation of their members' time-series.

The results of the clustering were mapped onto cartograms, enabling a visual analysis of the relationships between cluster membership and local authority house prices/ price growth.

2.3 Visual analysis of time-series shapes. The shapes of the three centroids were analysed to identify those time periods where there were significant differences in centroid growth. Box plots were then produced of local authority price growth rates for each of these time periods for each cluster. The box plots validated that the differences found between the centroids were also found with the local authority time-series. The box plots were inspected to see if they provided evidence of a 'chain reaction' in house price increases.

2.4 Visual analysis of 'key drivers' of house price growth. Scatter plots and correlation analysis were produced for each of the three periods comparing house price growth with household income. The volumes of new dwelling stock were also analysed.

The data processing in this project was carried out using SQL code generated in Microsoft Access. The clustering and statistical analysis were carried out in Matlab. Tableau was used to create the geographic maps, and the cartograms were produced in Microsoft Excel (adapted from a London Datastore template[4]).

[4] https://data.london.gov.uk/dataset/excel-mapping-template-for-london-boroughs-and-wards

3. Analytical steps

Fig 2 summarises the project's analytic steps:

3.1 Production of descriptive statistics and data maps.

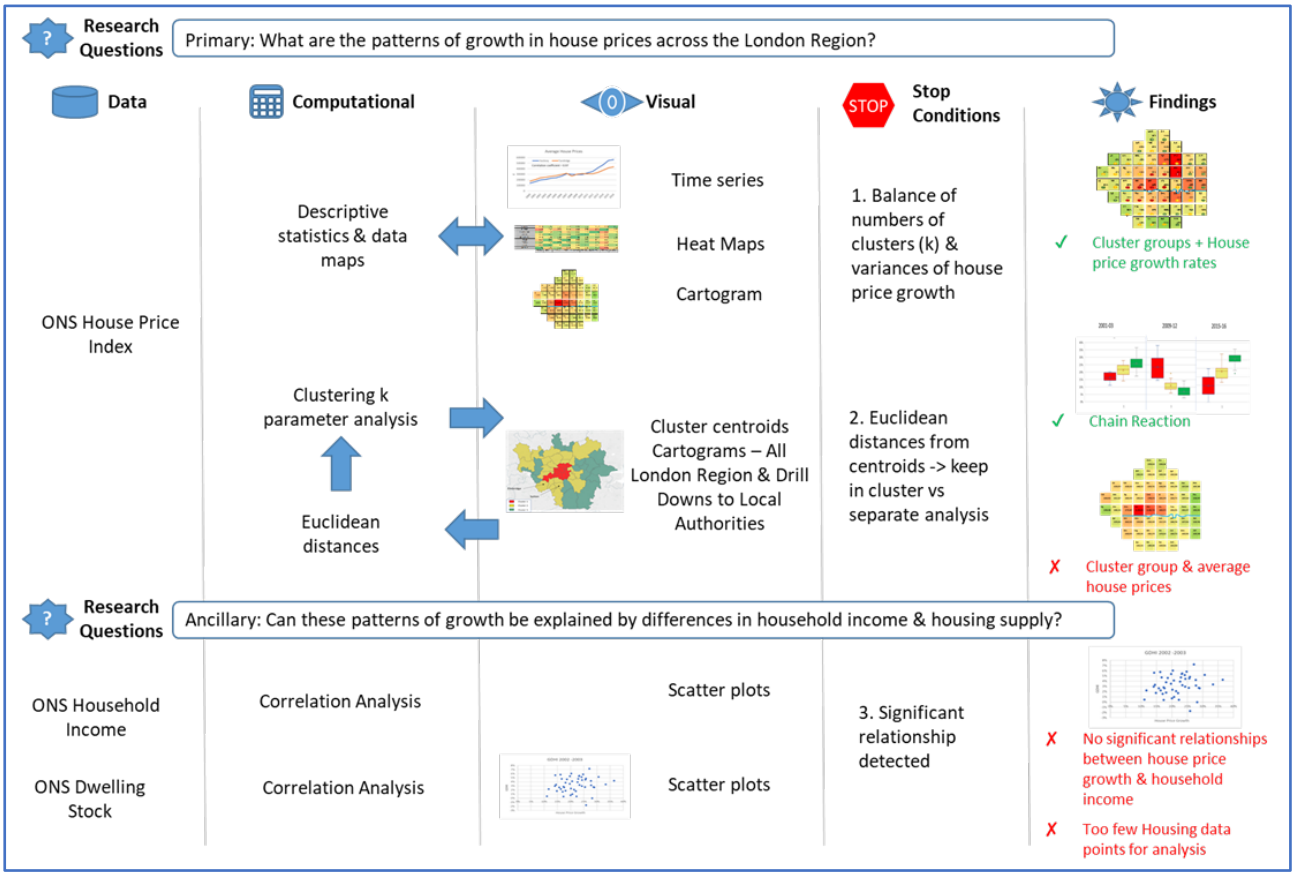

Fig 2. Overview of project steps.

The following are examples of the analyses that were carried out to gain an initial understanding of the trends and distribution of the house price data.

The time-series of house prices were plotted to assess their shapes. Figure 3 is for house prices across all 51 local authorities. This shows the steep rise in prices, with an exception in the year 2008, when the financial crash led to a sharp reduction in credit availability and consumer confidence[5].

[5] https://www.fca.org.uk/firms/mortgage-lending-statistics

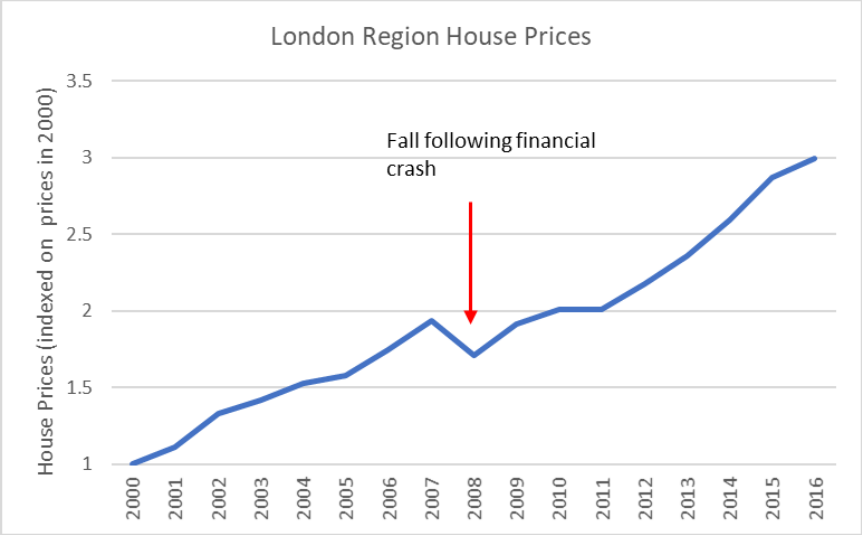

Fig 3. House prices across all 51 local authorities.

Correlations between the house prices of local authorities were investigated using a heat map. It was found that house prices were all highly correlated with each other.

	Barking	Barnet	Bexley	Brent	Brentwood	Bromley	Broxbourne	Camden	Chiltern
Barking	1.00	0.96	0.99	0.95	0.98	0.98	0.99	0.89	0.97
Barnet	0.96	1.00	0.98	1.00	0.99	1.00	0.98	0.98	0.99
Bexley	0.99	0.98	1.00	0.98	0.99	0.99	0.99	0.93	0.99
Brent	0.95	1.00	0.98	1.00	0.98	0.99	0.97	0.98	0.99
Brentwood	0.98	0.99	0.99	0.98	1.00	0.99	0.99	0.94	0.99
Bromley	0.98	1.00	0.99	0.99	0.99	1.00	0.99	0.96	0.99
Broxbourne	0.99	0.98	0.99	0.97	0.99	0.99	1.00	0.92	0.98
Camden	0.89	0.98	0.93	0.98	0.94	0.96	0.92	1.00	0.97
Chiltern	0.97	0.99	0.99	0.99	0.99	0.99	0.98	0.97	1.00

Fig 4. Extraction from the correlation coefficient heat map for London Region local authorities.

The least correlated local authorities were ‘Barking’ and ‘Kensington and Chelsea’, their correlation coefficient being 0.86. Even timeseries with very different growth rates over the 16 years (2000 to 2016) were highly correlated. The largest growth was for Hackney (341%) and the lowest growth was for Tandridge (176%).

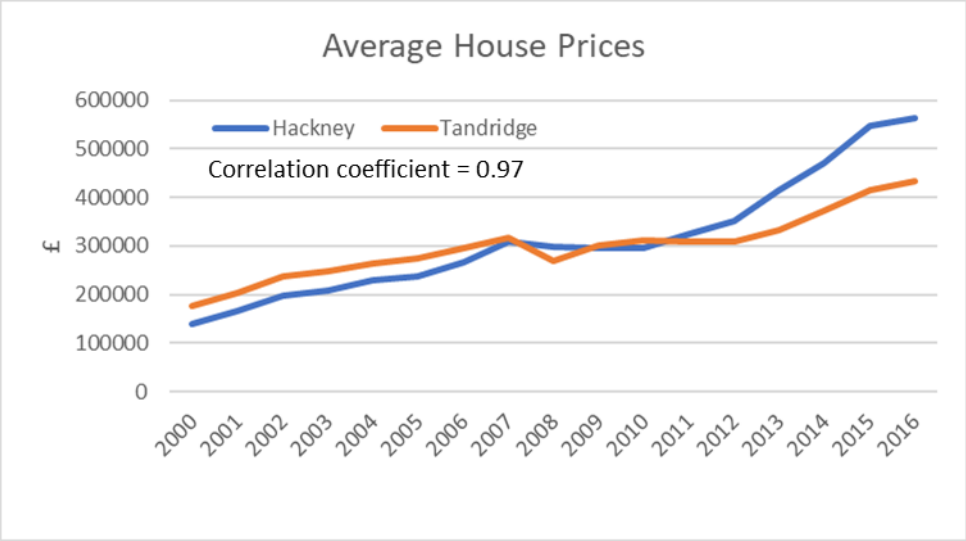

Fig 5. House price time-series for Hackney and Tandridge.

The distribution of house prices and house price growth was investigated using cartograms. The example below displays the average house prices at the end of 2016.

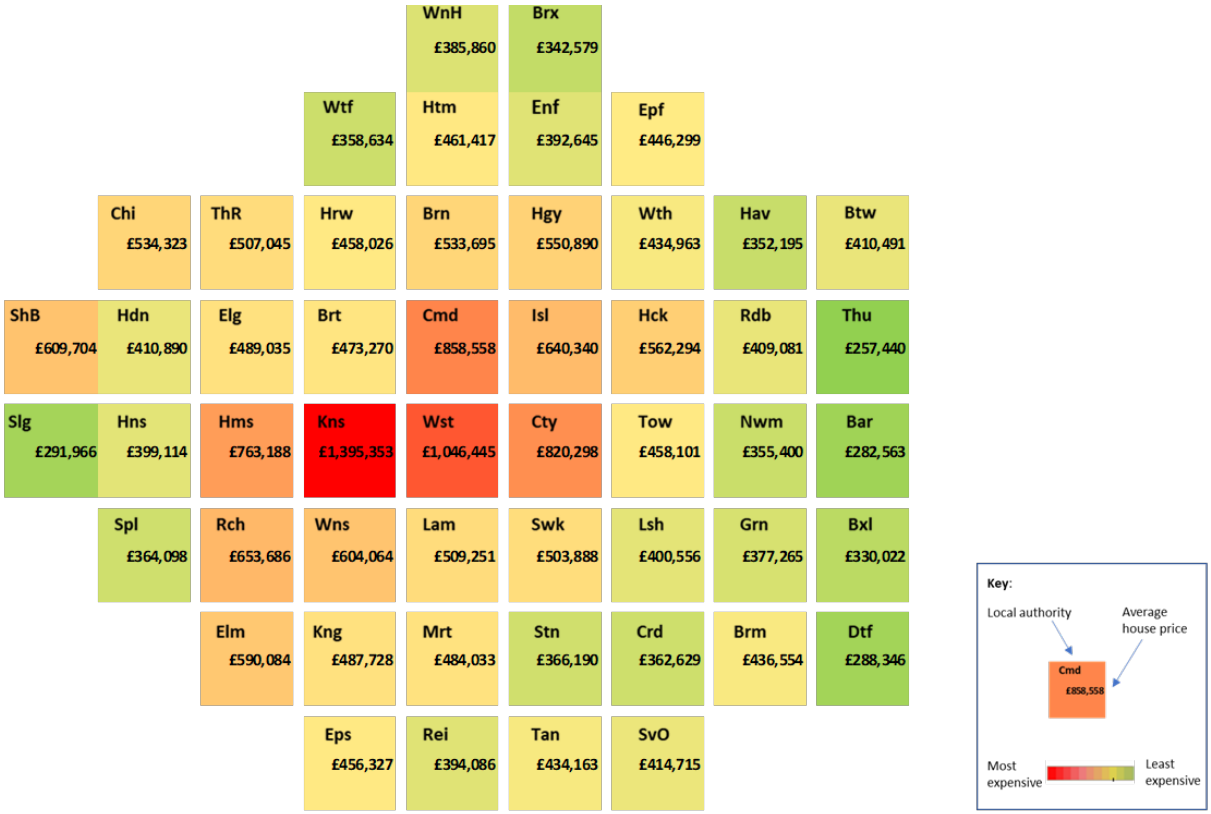

Fig 6. Cartogram of average house prices at the end of 2016.

3.2 Clustering

Analyses were carried out to determine the most appropriate number of clusters. The geographical map for five clusters is shown in figure 7. On the surface, the clustering appeared successful as there are some clear patterns to the clusters. However, the map also helped in identifying particular local authorities for further investigation. For example, Hounslow was classified as belonging to Cluster 3 even though it bordered a Cluster 1 local authority. This was also contrary to prior domain knowledge of West London that suggested that Hounslow was likely to have growth patterns similar to Ealing.

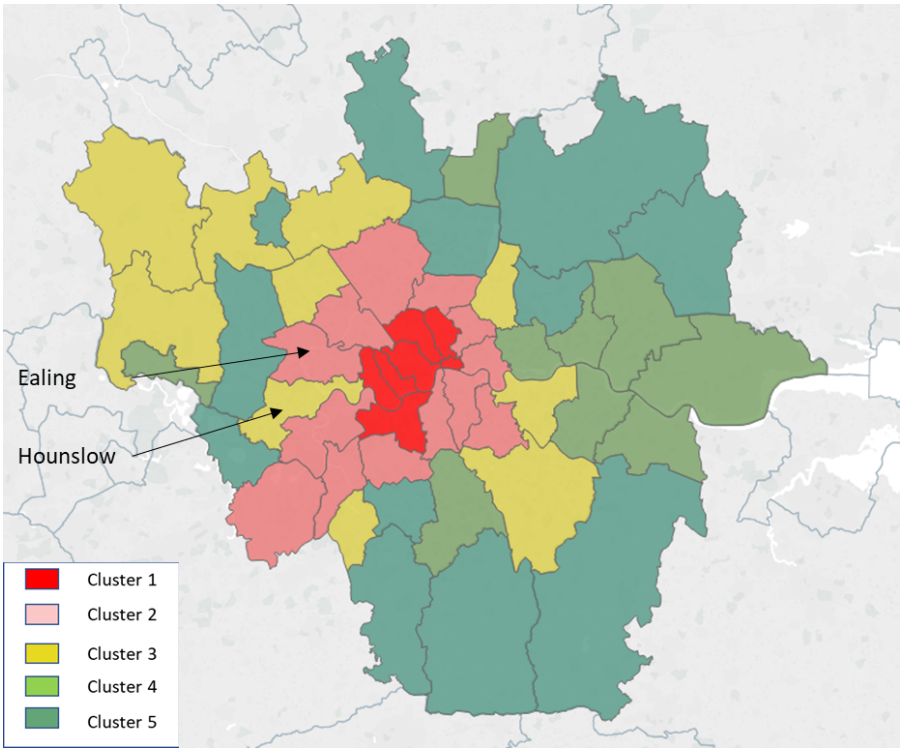

Fig 7. Map of London Region partitioned into five cluster groups.

The house price time-series for Ealing and Hounslow were plotted and found to be very similar. The centroids for the 5 clusters were then plotted and it was found that there was *no* significant difference between the centroids of clusters 2 and 3 that was useful in analysing differences in house price growth.

By completing the analytic tasks listed in section 2.3 for different k values, it was concluded that the most useful partition was achieved by having three clusters. The resulting Cluster 1 is contiguous and corresponds to central London. It is largely surround by Cluster 2, whilst Cluster 3 is mainly located in the south east of the London Region.

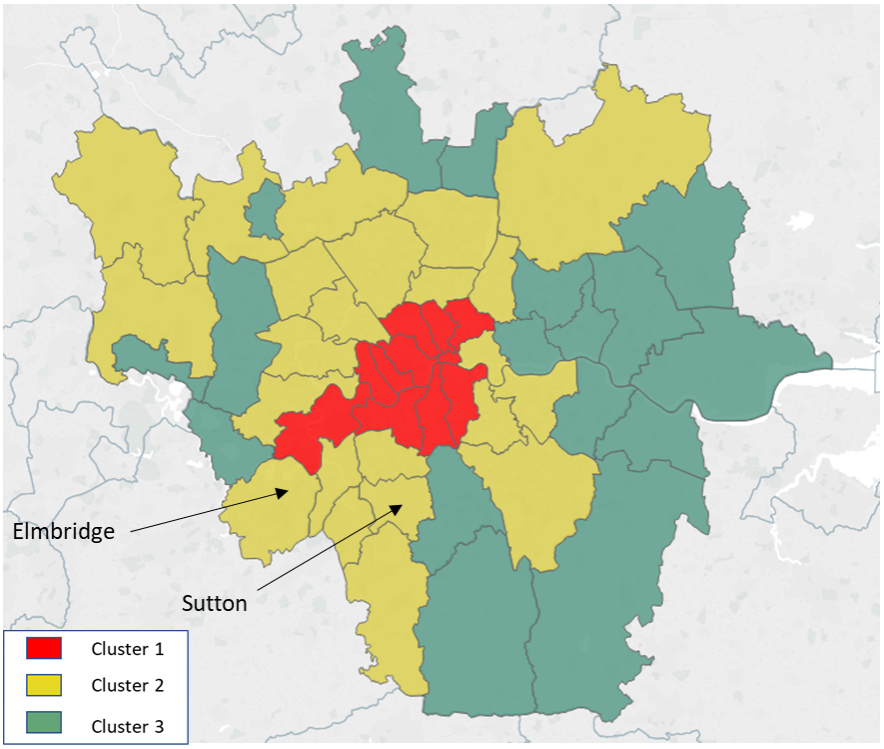

Fig 8. Map of London Region partitioned into three cluster groups

Overall, the clusters' centroids were found to be good representives of their cluster members' (i.e. the time-series of centroids were very similar in shape to that of their members). If a time-series had a significantly different shape to its centroid then this would have warranted excluding that time-series from its cluster and providing a separate analysis of its growth pattern. Euclidean distances were calculated to identify those time-series that were furthest from their centroid and their shapes were visually compared to their centroid. It should also be noted that there were a few time-series that were borderline between two clusters. For example: Sutton had degrees of membership of 0.51 for Cluster 2 and 0.45 for Cluster 3, whilst Elmbridge had degrees of membership of 0.42 for Cluster 1 and 0.46 for Cluster 2. However, overall it was judged that none of the time-series needed to be excluded from their cluster. Fig 9 is a plot of the centroid for Cluster 3 (the red line) together with each of the cluster's time-series, and helps to illustrate that the centroid is a good representative of its cluster's time-series.

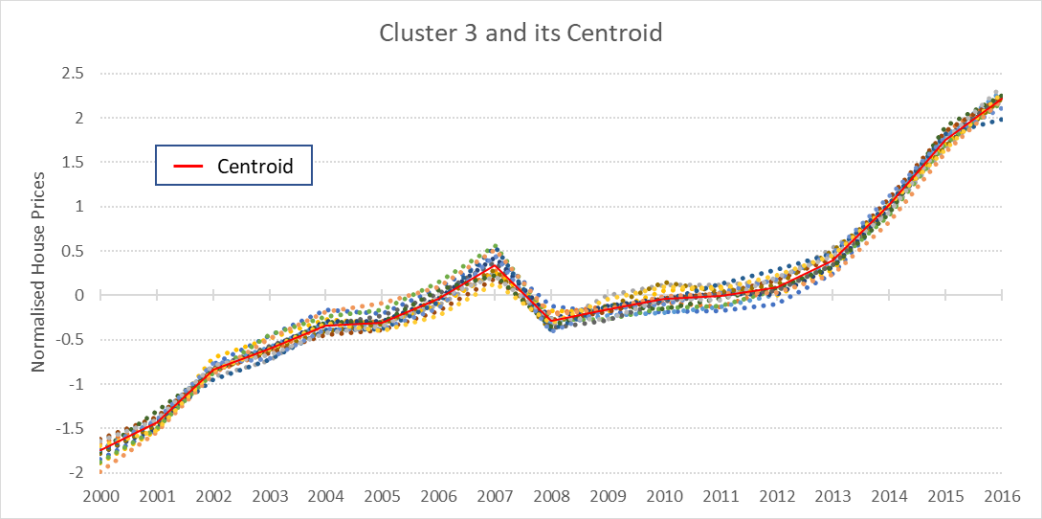

Fig 9. Time-series for the seventeen local authorities in Cluster 3, plus their centroid

Having established that the centroids were good representatives, visual analysis of their shapes identified three time periods where growth rates differed significantly between clusters. These were for 2001-2003, 2009-2012 and 2015-2016.

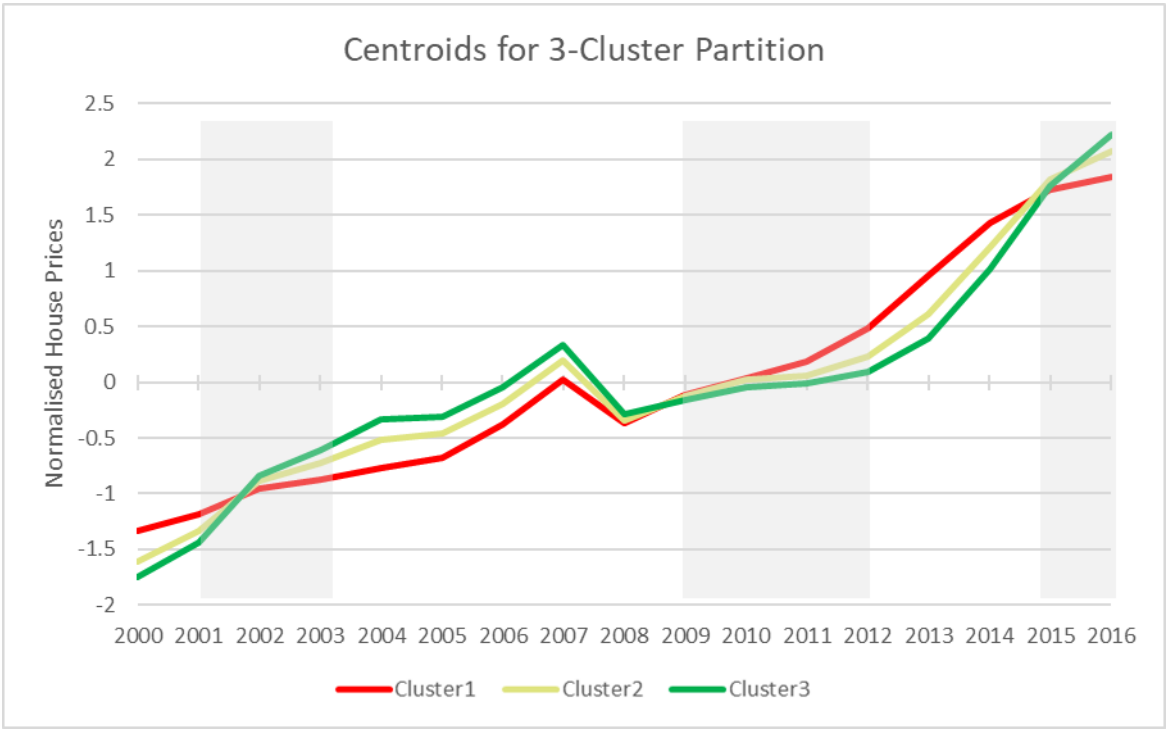

Fig 10. The three centroids- the shaded boxes show the three key time periods

3.3 Visual analysis of the growth rates.

The growth rates for the three time periods specified above were further investigated by:

i) producing boxplots of the local authority growth rates for each time period /cluster.

ii) analysing the time-series of particular local authorities.

The boxplots confirmed the patterns of growth rates found in the centroid graphs (see section 4). Cartograms were used to explore the relationship between cluster grouping and average house prices / growth rates.

3.4 Visual analysis of the key drivers of house price growth.

Scatterplots and correlation coefficients were produced in order to investigate the relationship between household income and growth rates. Separate scatterplots were produced for the three time periods for each of the cluster groups.

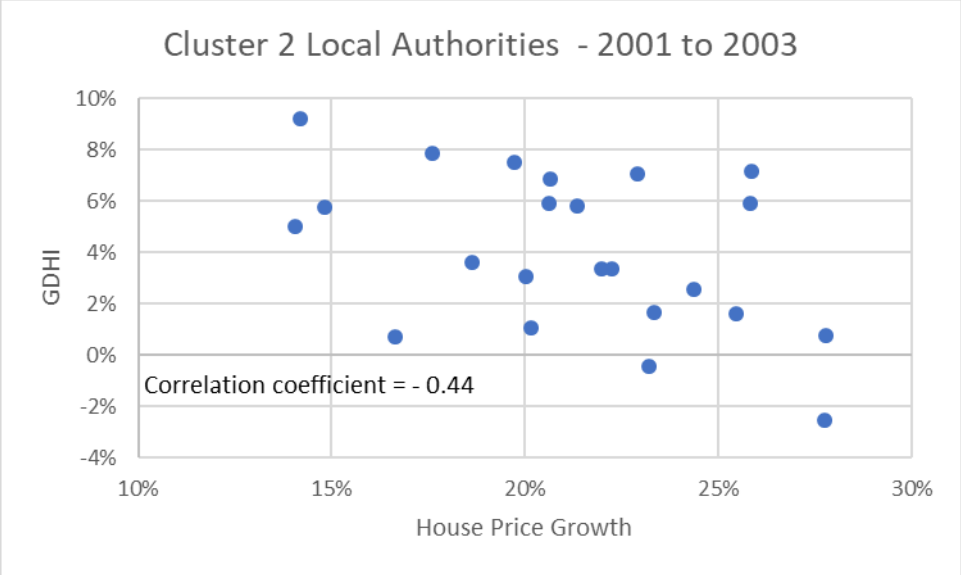

Fig 11. Example scatterplot of gross disposable household income vs house price growth.

Analysis of dwelling stock data showed that the volumes of new housing was too small to be relevant in explaining the differences in house price growth. For example, there was only a 0.006% increase in London borough housing stock between 2015 and 2016.

The stopping conditions for the analysis were when:

a) It was judged that the best balance had been struck between having a small number of clusters *k* that could be visually analysed and having clusters whose centroids captured the key shape differences between the house price time-series.
b) the time-series with the greatest Euclidean distances from their centroid had been inspected and a decision made either to keep them in their cluster or to analyse them separately.
c) either no significant relationships were found between house price growth and household income/ dwelling stock; or their relationships were quantified using regression modelling.

4. Findings

The project succeeded in identifying that the London Region could be partitioned into three cluster groups that had significantly different patterns of house price growth.

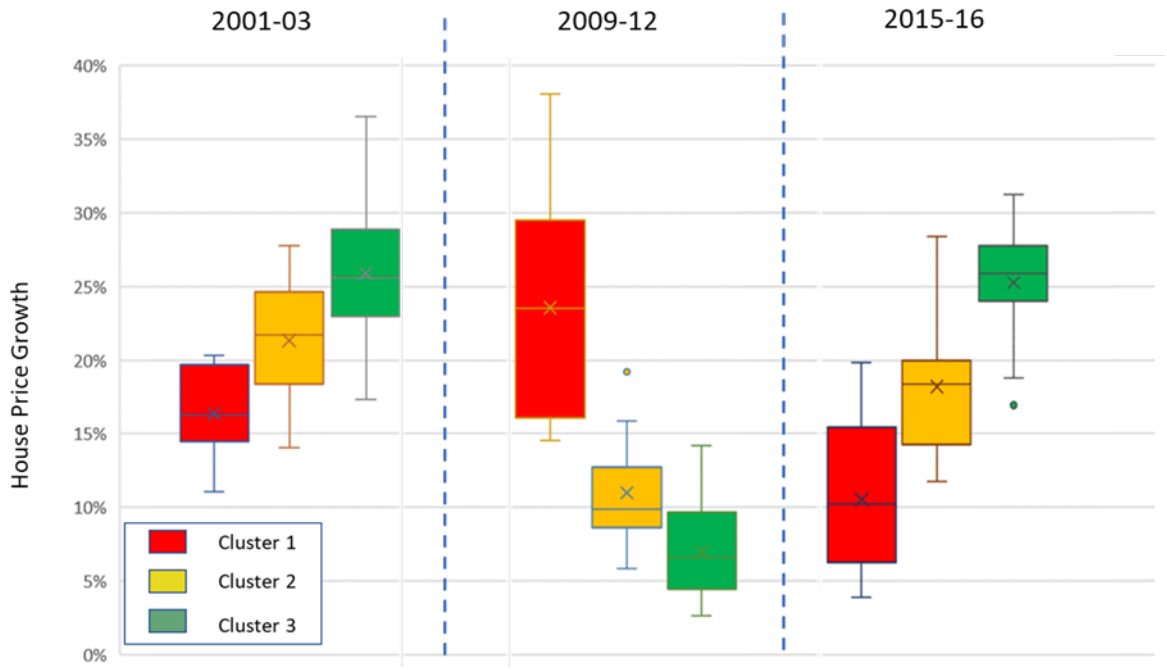

Fig 12. Boxplots of local authority house price growth by cluster group.

There was also found to be a strong relationship between a local authority's cluster group (determined by its pattern of growth) and that local authority's average house price. For example, the average house price in 2016 for Cluster 1 was £709k, in Cluster 2 was £455k and in Cluster 3 was £355k.

The above results provide some support for the 'price chain-reaction' hypothesis. The original hypothesis was that as central areas of London have become too expensive, demand has shifted to London's outer areas until they in turn have become too expensive, and demand has then shifted to the commuter belt. This project found that the geographical locations of the clusters differed to those described in the original hypothesis (see figure 8). Nevertheless, since the financial crash in 2008, house price growth has been consistent with there being a price chain-reaction. This is illustrated by figure 12, for example Cluster 1 (which had the highest average prices) experienced very high levels of growth between 2009-2012, whilst the highest levels of growth in 2015 -2016 was in Cluster 3 (which had the lowest average prices).

The different patterns of house price growth could not be explained with the household income or dwelling stock data. For example, a positive correlation was expected between house price growth and household income growth, with wealthier people moving into 'cheaper' local authorities and increasing those authorities' house prices. Instead the data failed to show a consistent relationship, in some periods even showing strong negative correlations (see for example figure 11).

Finally, there was only a weak relationship between cluster group and the overall growth rate from 2000 to 2016; the growth rate for Cluster 1 was 212%, for Cluster 2 was 201% and for Cluster 3 was 200%. The highest growth rates were in Hackney (Cluster 1) and Waltham Forest (Cluster 2). A cartogram showing the % growth rates and cluster group of each local authority is shown below.

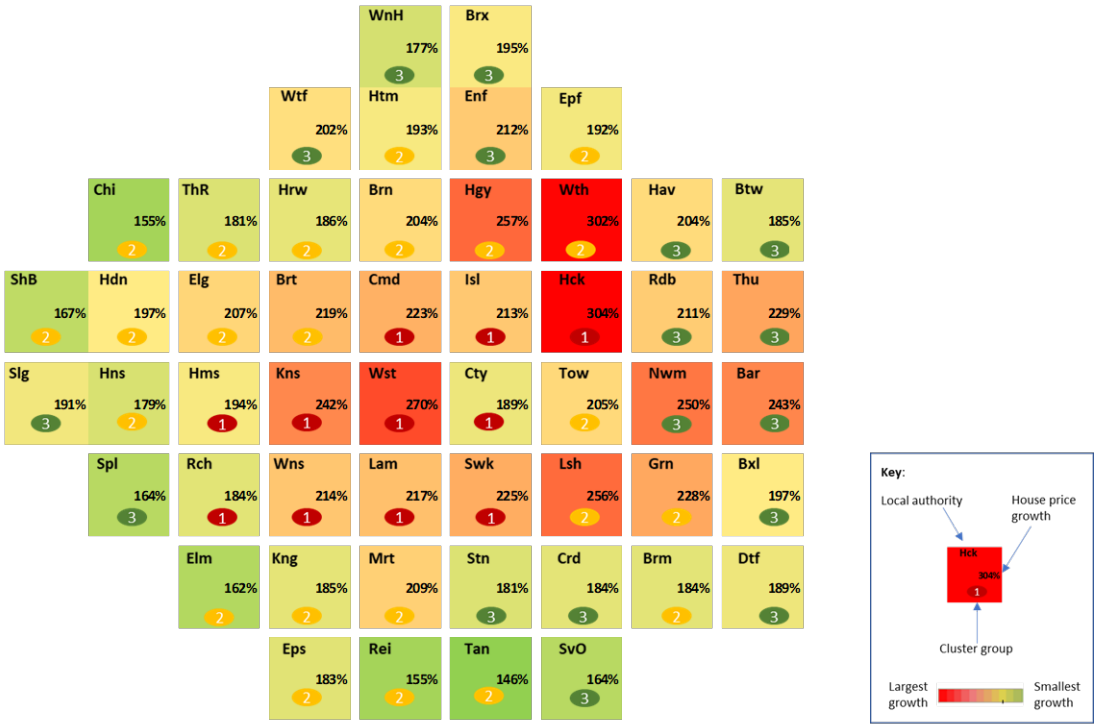

Fig 13. Cartogram of house price growth (2000 to 2016) and cluster group.

5. Critical reflection

The project has made substantial progress in answering its research questions. It established that the London Region could be meaningfully partitioned into three cluster groups, it provided some supporting evidence for a chain-reaction of house price growth, and showed that key data used in successfully explain national house prices growth, was less relevant in explaining the local authority growth rates.

The findings of this project have been beneficial in helping to understand the composition and dynamics of the London Region housing market. Identifying how the market should be partitioned, its correlations and its weak relationships to the household/dwelling data are important precursors to improving the analysis and modelling of London Region's housing. Such improvements should, in turn, lead to improvements in the regulation of London Region's housing.

The visual analytics methodology applied in this project has been successful at identifying meaningful clusters in London Region's housing market. The methodology combined using the computational technique of fuzzy c-means clustering with visualisation techniques.

Visualisation was central to the methodology, for example: in understanding the geographical locations of the clusters, in drilling down to investigate particular local authorities, in determining whether the centroids were sufficiently different to each other, in identifying the time periods for which there were significant differences in growth rates, and in detecting outliers.

The methodology was based on the assumptions that the house price dataset:

i) could be meaningfully partitioned into a small number of clusters using a fuzzy c-means algorithm.
ii) that the resulting centroids would be representative of their cluster members.

These assumptions turned out to be valid for the London Region. The box plots of growth rates (figure 12) demonstrate this validity. However, there will be many cases where these assumptions will fail.

Let us consider when the first assumption may fail. The fuzzy c-means algorithm is a 'soft' version of k-means clustering, that shares many of its advantages and disadvantages. The reason for going with a k-means type of algorithm was that, when applied to timeseries, they only compare feature values occurring at the very same times. For example, if two time-series were identical except that one time-series was offset from the other by a year, then they would be scored as being dissimilar. This differentiating between offset time-series was required as in this project, given that it was analysing price 'chain-reactions', hence offset timeseries needed to be assigned to different cluster groups. But there will be many cases where offset differences are not important, and non-linear clustering methods such as dynamic time warping will be more appropriate. There are also several well-known problems with *k*-means type clustering [4] including:

a) it is not robust to outliers, which can greatly impair its accuracy.
b) it can give too much 'weight' to large clusters and ignore small clusters.
c) it will 'identify' the user specified *k* clusters, even when there is no clustering in the actual data.
d) it can perform poorly with noisy data.

Visualising the time-series and clusters has an important role to play in assessing the performance of *k*-means type algorithms and in deciding whether a more sophisticated algorithm is needed.

The second assumption on 'centroids being good representatives' may also fail. For example, there can be a great deal of variance in the patterns of a particular cluster's time-series. Analysing Euclidean distances and visualising time-series help to assess the validity of this assumption.

This project provides an example where both assumptions were correct, and in such cases applying the methodology can provide valuable insights. There appears a strong case for testing the methodology within other domains where the time-series seem likely to have interesting regional variations e.g. unemployment, crime, personal wealth or pollution.

With respect to the ancillary analysis, the lack of correlation between local authority house prices and the household income data suggests that alternative data sources are needed. It seems uncontroversial that there would be a strong relationship between the changing wealth of individuals in a local authority and that local authority's house price growth. Although used in

forecasting national house prices [2], gross disposable household income may be a poor indicator of property owner wealth in the London Region. It is mainly derived from UK personal taxation and state benefits data of local authorities' residents. This misses the wealth of non-UK based owners, which is claimed to account for 20% of inner London sales [3]. It also misses the wealth of landlords, but includes the wealth of many residents who do not participate in house buying (e.g. those living in council estates). A successful analysis of the effects of property owner wealth may well need data on each of these factors. By contrast, there do not appear to be any problems with the housing stock data, however the volumes of new housing stock are too low to be relevant in explaining the patterns of house price growth.

6. Conclusion

The London Region housing market was analysed using a clustering/visual analytics methodology. There were found to be three cluster groups with significantly different patterns of house price growth. Visual analysis of centroids identified three key periods of house price growth. The project provided some supporting evidence for a price chain-reaction between the cluster groups. It appears that additional data sources are needed to assess how wealth affects local authority house price growth. The clustering/visual analytics methodology has worked well with the London Region housing data. Its applicability to other domains should be further explored.

7. Bibliography

[1] Aghabozorgi, S., Seyed Shirkhorshidi, A., Ying Wah, T. (2015). 'Time-series Clustering – A Decade Review.' *Information Systems* 53 (C): pp. 16–38.

[2] Auterson, T. (2014), 'Forecasting Housing Prices', Office for Budget Responsibility, Working Paper No. 6.

[3] Cassie, B., Wilson, W. (2017) 'Foreign Investment in UK Residential Property', House of Commons Library, Briefing Paper 07723

[4] Jain, A. (2010) Data clustering: 50 years beyond K-means, *Pattern Recognition Letters* 31 pp. 651–666.

[5] Price Waterhouse Coopers (2017) UK Economic Outlook