093

An empirical study on applying community detection methods in defining spatial housing submarkets in London

Stephen Law
Space Syntax Laboratory, The Bartlett School of Architecture, UCL
stephen.law@ucl.ac.uk

Kayvan Karimi
Space Syntax Laboratory, The Bartlett School of Architecture, UCL
k.karimi@ucl.ac.uk

Alan Penn
Faculty of the Built Environment, The Bartlett, UCL
a.penn@ucl.ac.uk

Abstract

Housing submarkets can be defined as a set of dwellings that are reasonably close substitutes with one another, but poor substitutes between other submarkets. This research argues similarities within submarkets are not only captured by its building and location characteristics but also in how each dwelling is inter-connected within its local area and embedded to the rest of the system. This research conjectures that spatial network local-areas as defined by community detection methods can be used to identify spatial housing submarkets. In order to test this conjecture, the hedonic approach will be used as an empirical strategy on the case study of London. The study found spatial network local areas correspond with planned known local area boundaries and that greater house price similarity is found within spatial network local-areas than between. The study also found that spatial network local area as defined by community detection technique can be used to identify spatial housing submarkets to explain house price. The contribution of this research is it represents a proof of concept in the use of community detection techniques in the definition of spatial housing submarket. Importantly it illustrates the significance in how spatial configuration influences housing market not just in terms of accessibility (Law et al. 2013) but also in terms of housing submarket. Further research will be carried out to study the spatial configuration of the spatial network local areas in understanding severances and connectivity between them. By understanding cities through multiple spatial representations will allow more informed policies at the local-area level.

Keywords

Housing submarket, hedonic model, house price, community detection, space syntax, London.

1.0 Background

Research examining intra-city house price variations often focuses on estimating the implicit price, buyers and sellers are willing to exchange at, for structural features, accessibility levels and local amenities from observed sold price using the hedonic approach (Rosen, 1974; Cheshire and Sheppard 1998). Following Alonso's (1964) monocentric model, measures such as "distance to CBD" and "gravitational potential" to employment were often applied to estimate the marginal willingness to pay for location differentials. Under similar vein, research in space syntax proposed the use of spatial configuration measures in estimating the implicit value of accessibility on the housing market without apriori data on employment location (Law et al., 2013). However, location differentials in house prices are not only captured by spatial configuration factors such as spatial accessibility or access to local amenity such as shops, schools and parks but it is argued in this research also by the distinctive housing submarket the property sits on. In simpler terms, the buyer is not only buying accessibility or amenities but also to live in specific housing submarket argued to be influenced by configuration. Below is an example describing two adjacent areas of London with similar travel time to Oxford Circus, similar global accessibility and similar number of shops and active uses but with significantly different house price. This suggests the two areas sit in different submarkets with different implicit values. It is thus unrealistic to assume a global housing market for the entire metropolitan region of London but rather a market of multiple submarkets with its own unique market conditions. An important question is how spatial housing submarkets can be defined.

	House Price 2013	Travel time to Oxford Circus	Integration	Retail units	Active Units
Crouch End	£595,000	30 minutes (+/- 5 mins)	10,400	92	494
Green Lanes	£373,000	30 minutes (+/- 2 mins)	10,950	96	573

Table 1 : An accessibility, amenity and house price comparison between two adjacent local areas in London.

Over the past decade, much research in regional studies had been conducted on the definition of housing submarkets. (Grisby et al, 1987; Bourassa et al., 1999; Dale-Johnson, 1982; Goodman and Thibodeau, 1998) Housing submarkets could be defined as a set of dwellings that are reasonably close substitutes for one another, but relatively poor substitutes for dwellings in other submarkets (Grisby et al., 1987) A simple example is defining a housing market through its dwelling type. A flat owner might value accessibility differently to a detached dwelling owner. A detached dwelling owner might value school quality differently to a flat owner. A greater understanding of housing submarkets can in turn improve the understanding on the value of different property characteristics which in turn would improve the prediction of a hedonic house price model.

Empirically, housing submarkets are defined by having similar dwelling characteristics using statistical clustering techniques at the postcode level or some administrative level. By resorting to administrative defined regions that are dependent on the past, properties are not being considered as part of a network of dwellings that embedded within the street network. These regions are insensitive to changes in spatial configuration over time. As illustrated in the figure below, the ward boundaries of Thamesmead do not accurately align with the spatial configuration of Thamesmead today.

This research aims to extend this line of thought by proposing a new type of spatial housing submarket based on spatial network local area using the topology of the street network rather than ward or postcode areas.

Figure 1: Ward boundaries of Thamesmead in London.

As both communities and neighbourhoods often have social meanings attached to it, the term sub-graph will be used in the methodology section and the term spatial network local-areas will be used for the rest of this research when community detection techniques are being applied to a spatial street network.

Two research fields have examined the definition of local-areas through its spatial network properties. One emerged from the field of space syntax where local areas are defined by its spatial network measures similarity, such as similarity in node count (Yang and Hillier, 2006) or similarity in intelligibility values within a local area. (Dalton, 2007) The second field, known as community detection, emerged under the field of network science, where subgraphs are detected through its topology. (Girvan and Newman 2002) Community detection techniques (Fortunato , 2010) have been widely adopted in network science; from uncovering organisations in social networks to uncovering pages with similar topics in the worldwide web to uncovering geographical regions on the commuting network. This research applies community detection methods on the spatial street networks in uncovering spatial network local areas.

1.1 Research Objective

This research conjectures that spatial network local-areas as defined by community detection methods is effective in identifying known local area and could be used to identify spatial housing submarkets. In order to test the conjecture, we first define the community detection method and apply on the case study of London street network. Second, we test the significance of the spatial local area boundary by testing it with known local area boundary. Third, we analyse the house price variations between the spatial local areas. Lastly, a new spatial housing submarket definition is defined by applying statistical clustering techniques on housing characteristics within each local area. The new spatial housing submarket will be compared with traditional submarket definition in explaining house price variation through the hedonic framework.

1.2 Datasets

Two open source datasets are used for the empirical section. The first key dataset is the Ordnance Survey Meridian Line dataset which covers the entire United Kingdom (Ordnance Survey, 2014). This network dataset is cropped up to the M25 and is used to construct spatial local areas for the Greater London Area. The second key dataset is the sold house price dataset for the same study area collected from Land Registry between the years 2009 – 2013 (Land Registry, 2014). Please see Appendix A for more details on the two datasets.

2.0 Spatial Network

In graph theory, a spatial network is a type of planar graph embedded in Euclidean space. As illustrated in figure below, two types of spatial network can be defined, the primal graph (PG) where vertices are junctions and edges are streets or the dual graph (DG) where the vertices are streets and edges are junctions (Porta et al., 2006). This study will employ community detection techniques on the dual graph commonly used in spatial configuration research (Hillier and Hanson, 1984). Specifically we will apply community detection on the dual graph produced from the road centre line segments (Turner, 2007).

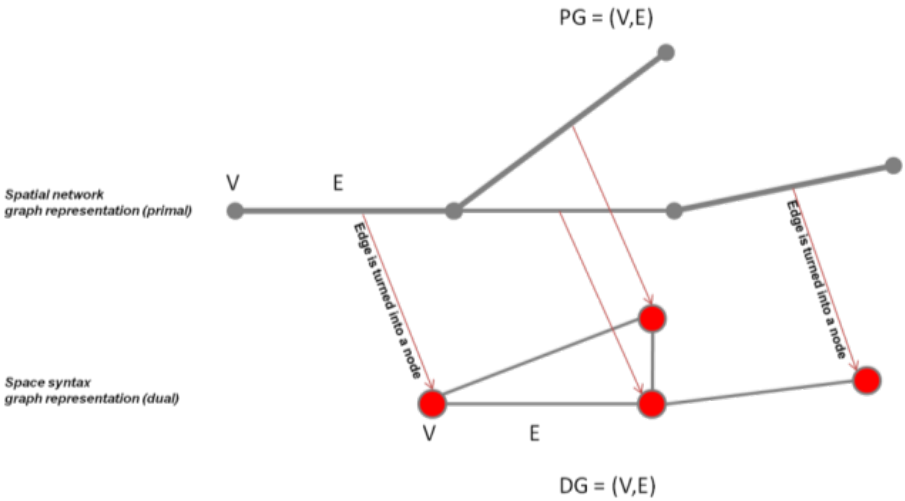


Figure 2 : Spatial Network Graph Definition. Primal representation at the top and Dual representation at the bottom.

2.1 Community Detection

A common topic in network science is community detection, whose objective is to define a set of subgraphs that maximises internal ties and minimises external ties using strictly the topology of the graph. (Girvan and Newman, 2002) Many methods exist in the definition of individual subgraph, such as divisive algorithms on high betweenness centrality edge (Girvan and Newman, 2002), dynamic algorithms (Reichardt and Bornholdt, 2004), vertex propagation algorithms (Raghavan et al., 2007) and optimisation algorithms (Newman and Girvan, 2004). This research will adopt optimisation algorithms to identify spatial local areas.

2.2 Modularity optimisation

A common method or criterion in defining subgraph is to optimise against a quality function. The most common quality function for community detection in network science is called Modularity (Q) (Girvan and Newman, 2002). Modularity (Q) calculates the difference between observed number of edges within a subgraph and the expected number of edges. The greater the observed number of edges relative to its expected the higher is its modularity. More formally, Modularity (Q) is defined where A is the adjacency matrix, m is the total number of edges in the graph, ki and kj are the degree for vertex i and vertex j. δ is 1 if i and j are in the same community and zero otherwise.

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - K_i K_j/2m)\delta(C_i, C_j)$$

A is the adjacency matrix
m is the total number of edges
Ki and Kj are the degree for the two subgraphs i,j
δ is a Kroneckar Delta function which equals 1 when its argument are the same and 0 otherwise.
***Equation 1**: Modularity(Q) equation (Girvan and Newman, 2002)*

Optimisation against the above function is currently impossible to solve for large datasets (class NP-hard problem). As a result, a number of heuristic algorithms have been implemented into finding the optimal sub-graph (Girvan and Newman, 2002). This study will apply one such type, the multi-level methods (Blondel et al., 2007) in optimising against this quality function.

2.3 Multi-level method

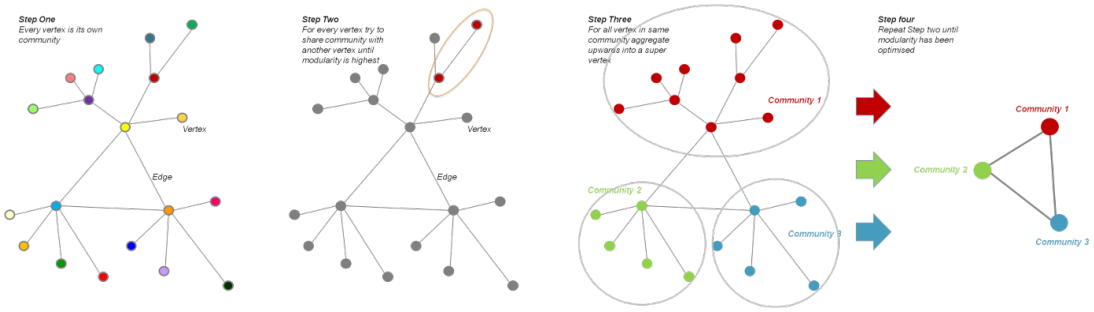


Figure 3 : The Multi-level method algorithm starts where every vertex is a community. Every vertex will then share community membership with one of its neighbours that attains the highest score. This continues for all vertices. Vertices within the same community will aggregate into a super vertex. These super vertices will again optimise its modularity sharing community membership until modularity can no longer be optimised. Diagram produced by the Author.

The multi-level algorithm starts where every vertex is a sub-graph. Every vertex will then share sub-graph membership with one of its neighbour that attains the highest modularity score. This continues for all vertices. After all vertices have been traversed, vertices within the same sub-graph will aggregate into a new super vertex and a new super-graph is formed. The super vertices of the new super-graph will again optimise its modularity sharing sub-graph membership with its neighbours. This aggregation continues until modularity can no longer be optimised. The method is hierarchical where each subgraph produced is part of a larger super-graph in the next iteration.

2.4 Multi-level method limitations

Despite being one of the most used algorithms, multi-level methods in modularity optimisation have some known technical limitations. The first is the resolution problem where the quality of the optimal aggregation might not necessarily have a more accurate partition than one with a smaller aggregation. (Lancichinetti and Fortunato, 2011) The second is the randomness of the starting node can produce potentially slightly different groupings where the network has multiple local maxima. Future research will respond to both of these limitations through a sensitivity test of the algorithm. Despite these known limitations, the multi-level methods have been evaluated against other community detection algorithm and have been found to be both computationally efficient and producing accurate memberships. (Lancichinetti and Fortunato, 2009) Secondly, multi-level methods have been applied previously to spatial networks such as the airplane network and commuting networks that found great similarity to geographical and functional regions.

2.4 Spatial network local-areas in London

Applying the multi-level method described in the last section on the OS Meridian line network, a total of 166 spatial network local-areas were detected for the Greater London area. Each spatial network local-area has on average 680 segments with a standard deviation of 269 segments.

Table 2 : 166 Spatial network local areas detected in Greater London

	Number of Community
	166
	Segments per community
Average	680
Std Dev	269
Min	102
Max	1572

The figure below shows the spatial network local-areas obtained from the multi-level method for the Greater London Area on the OS Meridian Line map. The figure shows distinct local areas mapped in GIS where the different colours correspond to different membership. The fifth and the last level of the multi-level aggregation was visualised. Visually the results shows clear distinction for local areas separated by River Thames such as the Isles of Dog and the Royal Docks area, spatial local areas separated by the railway tracks such as Crouch End and Harringay and spatial local areas separated by the Lea Valley. However the boundaries between spatial local-areas in Central London are not as apparent.

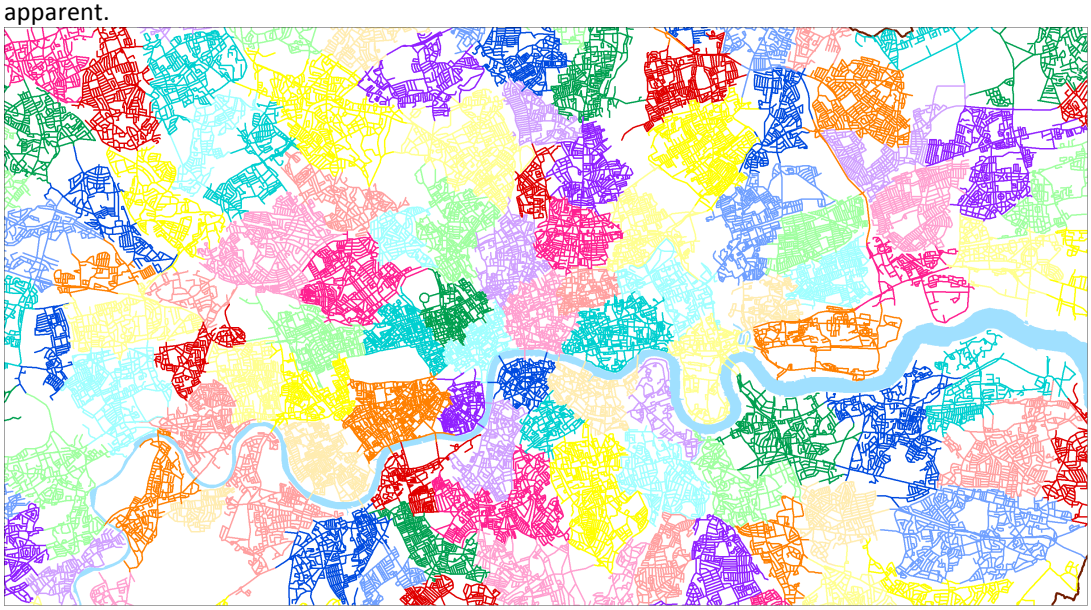


Figure 4 : Visualisation of Spatial network local areas for the Greater London Area

The figure below illustrates the spatial network local-areas overlaid on top of the Bedford Park planned local area boundary in London. The figure on the far left shows the smallest spatial network local-areas, the first level of aggregation, to the far right which shows the largest spatial network local-areas, the fifth level of the aggregation. Level 2 in this case shows the greatest similarity with the original Bedford Park local area historic boundary in black. Level 4 shows similarity to the larger Bedford Park area. The hierarchical nature of the algorithm allows each spatial network local-areas to be seen as embedded to a system of connected local-areas.

Figure 5 : Spatial network local areas for different aggregation level overlaid with the boundary of Bedford Park. The figure on the left shows the first level of aggregation, to the right which shows the fifth and last level of the aggregation. Level 2, 3 and 4 resembles different resolution of the Bedford Park suburb.

2.5 Known Local Area Test

To examine how spatial network local-areas associate with known local areas in London, an initial known local area test is conducted for five areas. Four of these are planned areas whilst one is unplanned. We will test both visually and statistically the similarity between the spatial network local-areas boundaries and the known local area boundaries. The figure below illustrates the OS Meridian line dataset (Ordnance Survey, 2014) for the Greater London Area overlaid for these five local areas namely Hampstead Garden Suburb, Brentham Garden Suburb, Bedford Park, Thamesmead and Soho in Central London. The following sources were used for the identification of the known local area boundary. (LB Barnet, NA; LB Ealing, 2007&2008; LB Hounslow, NA; Andrew Nunn Associates, NA; Thamesmead Trust, 2007; Sheppard, 1966; Walter, 1878; WIkitravel, 2011) The known local area boundary of Hampstead Garden Suburb, Brentham Garden Suburb and Bedford Suburb were based on historical sources from the councils. The known local area boundary of the Thamesmead development was based on the developer's masterplan. The known local area boundary for the Soho was based on the crowd source wiki-travel website. The objective here is to show how spatial network local-areas associate with known local areas in London using historic, developer or user-defined boundaries. A user-defined local area boundary will provide a stronger basis for the local area test but this is beyond the scope of the research.

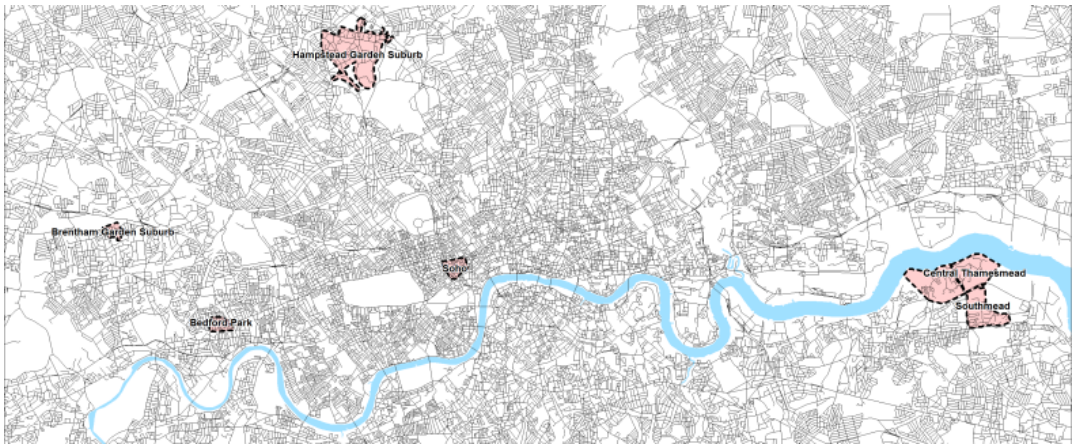


Figure 6 : Local Area Test Cases.

The figure below illustrates the overlay between the four planned known local area boundaries and the spatial network local areas defined by the multi-level method. The result shows, there is high levels of association between the known local area boundary in black with the spatial network local areas identified in pink for Hampstead Garden Suburb, in purple for Brentham Garden Suburb, in orange for Bedford Park and in green for Thamesmead Central and Thamesmead North. The spatial network local area boundaries do not align perfectly as the street network continue naturally beyond the borders of these known local area boundaries.

Figure 8 illustrates the overlay between the Soho known area boundary with the spatial network local areas defined by the multi-level method. The result shows that there is a poor association between known local area boundary and the spatial local area identified in red, blue and green. This is not surprising as this local area was developed organically overtime with a porous street network that continues across the edges of the district. The known area boundary is defined less by its topology but more by its width of the road and the centrality of the road.

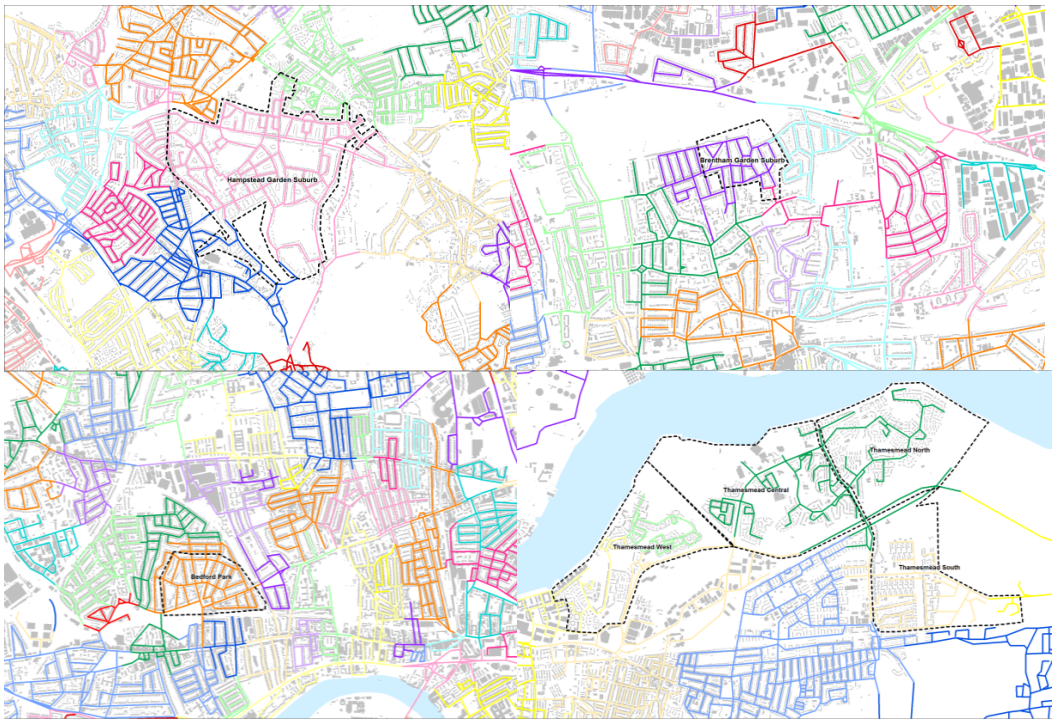


Figure 7 : The result shows strong similarity between the spatial network local area and four planned known local area boundary in black
Top Left Hampstead Garden Suburb boundary
Top Right Brentham Garden Suburb boundary
Bottom Left Bedford Park Suburb boundary
Bottom Right Thamesmead district boundary

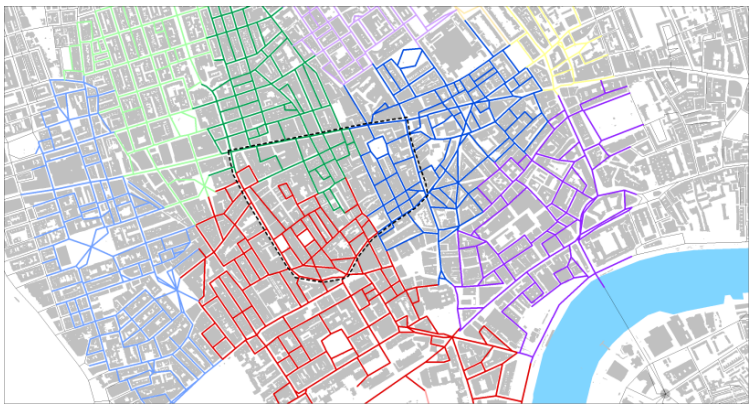


Figure 8 : Spatial network local area overlaid with Soho boundary. The result shows a poor association between the spatial network community definition and the unplanned area spatial boundary in black.

Table 3 summarises the local area tests between the five known local area boundaries and the spatial network local areas. Planned areas such as Hampstead Garden Suburb, Brentham Garden Suburb, Bradford Park and Thamesmead shows greatest similarity to the known local area boundary with a Cramér's V between 0.69 – 0.93. In contrast, the organically developed known local area of Soho achieved a Cramér's V of 0.34. This result reveals that community detection techniques are more effective in associating with planned local area than for unplanned ones. These results are not conclusive but points to association between community detection techniques and known local area that are spatially isolated. For details of the local area test please see Appendix B.

Table 3 : Local Area test statistics. The planned area achieves a significantly higher goodness of fit then the unplanned areas.

	Hampstead Garden Suburb	Brentham Garden Suburb	Bedford Park	Thames mead	Thames mead north	Soho
Pearson chi2	7.80E+04	5.40E+04	8.10E+04	4.50E+04	9.80E+04	1.30E+04
likelihood-ratio chi2	2.90E+03	532.021	924.943	1.60E+03	2.00E+03	753.188
Cramér's V	**0.828**	**0.690**	**0.845**	**0.632**	**0.930**	**0.342**

3.0 Exploratory Data Analysis Methodology

In order to study the significance of the spatial network local areas as a method in defining housing submarkets, we will explore the data and examine the house price variations between each spatial network local areas.

We will first describe the house price variations for the greater London area visually and measure the extent observed house price is spatially homogenous / heterogenous through the Global Moran's I. (Moran, 1950) Global Moran's I is an index of clustering which correlates a dwelling's sold price with its neighbouring sold price weighted by the distance between observation. It is calculated more formally where w is the weight matrix, x is the price of the observation, x bar is the mean price and N is the number of observations where the weight matrix assumes Euclidean distance up to 1200 metres. Similar to the Pearson correlation coefficient, the results range from -1 indicating perfect dispersion to +1 indicating perfect autocorrelation and 0 indicates insignificant spatial autocorrelation. The global statistic confirms the clustering nature of the London Housing market.

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

Equation 2 *: Global Moran's I equation (Moran, 1950)*

We will then examine the house price variations between the spatial network local areas. A one-way ANOVA (Analysis of variance) is adopted which will test whether the house price variations between the spatial network local areas differs comparing to the within variations. The null hypothesis is that the sample mean house price is the same for all spatial network local areas.

3.1 Housing Submarket Hedonic Model Methodology

Empirically, housing submarkets produce groupings that have a maximum degree of internal homogeniety and external heterogeneity. (Grisby et al. 1987) The hedonic approach (Rosen, 1974), which estimates the implicit price of a housing characteristic from observed sold price, is frequently adopted to test the significance of the housing submarket either by including it into an overall hedonic model or by estimating different housing submarket models. (Bourassa et al, 1999; Dale-Johnson, 1982; Goodman and Thibodeau, 1998) Statistical clustering techniques are often used to first delineate housing submarket before its application on the hedonic model. Bourassa et al. (1999) used principal component analysis and K-means clustering to delineate housing submarket cluster in Sydney and Melbourne. Dale-Johnson (1982) used factor analysis and cluster analysis in defining 10 housing submarkets. Goodman and Thibodeau (1998) used hierarchical clustering to define housing submarkets in Dallas. A report from The Greater London Authority used K-means clustering techniques to aggregate socio-economic and housing characteristics into defining 5-6 distinct housing submarkets for the metropolitan region. This research will follow similarly the definition of six housing submarkets in London.

Four models are defined for the empirical analysis. The first is a baseline hedonic model without the submarket variable. The second includes the postcode-attributes submarket variable. The third includes the ward attributes submarket variable. The fourth includes the spatial attributes submarket variable. The research procedure for each model is split into three stages. The first stage is to select the geographical unit for clustering. The second stage is to identify six housing submarket using K-means clustering on the averages of four property attributes namely dwelling type, dwelling tenure, new-built and global space syntax integration. The third stage would include the housing submarket variable into an overall hedonic model. Least Square is used for the estimation of the hedonic model, goodness of fit and test statistics are reported.

Model 1 is the standard empirical form of the hedonic approach which is to regress Log house price against a vector of dwelling specific and location specific variables through a simple Normal-Linear-Quadratic model (NLQ model) using cross section data. Dwelling specific variables include dwelling type, dwelling tenure and if the dwelling is new-built. Location specific variables include space syntax integration. Amenity specific variables include the number of shops at 800m and the number of offices at 800m. This is a reduced model compared to the previous research. (Law et al. 2013)

$$Log(HP_i) = \beta_0 + \beta_1 Log(Int) + \beta_2(Type) + \beta_3(Tenure) + \beta_4(New) + \beta_5 Log(Shop) + \beta_6 Log(Off) + \varepsilon$$

HP = house price
Int = space syntax integration
Type = the dwelling type (flat,terrace,semi,det)
Tenure = tenure type(leasehold, freehold)
New = if the dwelling is new built (new-built/not new built)
Shop = number of shops at 800m
Off = number of offices at 800m
Equation 3

Model 2 is the postcode-attributes housing submarket model where postcode area is the geographical unit. In the first stage, we take the averages of each attribute for each postcode area. In the second stage, we use K-means clustering to identify six housing submarkets minimising differences between global integration, type of house, tenure of house and if the house is new built or not. In the third stage, we include the postcode-attribute housing submarket variable into the global hedonic model as follows.

$$Log(HP_i) = \beta_0 + \beta_1 Log(Int) + \beta_2(Type) + \beta_3(Tenure) + \beta_4(New) + \beta_5 Log(Shop) + \beta_6 Log(Off) + \sum_{j=1}^{n=6} \gamma_j \, PostSub_j + \varepsilon$$

HP = house price
Int = space syntax integration
Type = the dwelling type (flat,terrace,semi,det)
Tenure = tenure type(leasehold, freehold)
New = if the dwelling is new build or not
Shop = number of shops at 800m
Off = number of offices at 800m
PostSub = housing submarket cluster
Equation 4

Model 3 is the ward-attribute housing submarket model where ward boundary is the geographical unit. In the first stage, we will take the averages of each attribute for each ward unit. In the second stage, we will apply K-means clustering to define 6 housing submarkets minimising differences between global integration, type of house, tenure of house and if the house is new build or not. In the third stage, we will include the ward-attribute housing submarket variable into the global hedonic model as follows.

$$Log(HP_i) = \beta_0 + \beta_1 Log(Int) + \beta_2(Type) + \beta_3(Tenure) + \beta_4(New) + \ +\beta_5 Log(Shop\ ) + \beta_6 Log(Off\ ) + \sum_{j=1}^{n=6} \gamma_j\ WardSub_j + \ \varepsilon$$

HP = house price
Int = space syntax integration
Type = the dwelling type (flat,terrace,semi,det)
Tenure = tenure type(leasehold, freehold)
New = if the dwelling is new build or not
Shop = number of shops at 800m
Off = number of offices at 800m
WardSub = Ward housing submarket cluster
Equation 5

Model 4 is the spatial-attribute housing submarket model where the spatial network local area is the geographical unit. In the first stage, we will take the averages of each attribute for each spatial local-area. In the second stage, we will apply K-means clustering to define 6 housing submarkets minimising differences between global integration, type of house, tenure of house and if the house is new build or not. In the third stage, we will include the spatial-attribute housing submarket variable into the global hedonic model as follows.

$$Log(HP_i) = \beta_0 + \beta_1 Log(Int) + \beta_2(Type) + \beta_3(Tenure) + \beta_4(New) + \ +\beta_5 Log(Shop\ R800) + \beta_6 Log(Off\ R800) + \sum_{j=1}^{n=6} \gamma_j\ SpatialSub_j + \ \varepsilon$$

HP = house price
Int = space syntax integration
Type = the dwelling type (flat,terrace,semi,det)
Tenure = tenure type(leasehold, freehold)
New = if the dwelling is new build or not
Shop = number of shops at 800m
Off = number of offices at 800m
SpatialSub = Spatial-attribute housing submarket cluster
Equation 6

Moran's I will be calculated on the residuals of the four models to test the extent spatial effect exists in the model. There are three key limitations to this research approach. First, only one community detection method is tested in defining housing submarket. Second, the hedonic regression model is only applied for one year and in one geographical region. Third, the research uses a simple least square specification in the estimation. Spatial temporal methods and empirical strategies are recommended for future research in responding to these limitations.

4.0 Descriptive Statistics

Below are the descriptive statistics for the house price in London between 2009 – 2013. The average house price in London for 2009 was approximately £380000 while its standard deviation was £400000. The mean rose to £500000 and the standard deviation to £660000 in 2013 where the mean and the deviation grew by 132% and 164% respectively. The results suggest house price in London have risen sharply the last five years while its distribution has become more dispersed and thus more unequal.

Year	Observations	average (GBP)	std dev (GBP)	Min (GBP)	Max (GBP)
2013	67859	500580	655542	10000	2.33E+07
2012	69311	459682	646612	50500	5.50E+07
2011	69653	437240	519080	50750	1.60E+07
2010	70955	425376	489518	51000	1.62E+07
2009	57684	377409	405487	50000	1.25E+07

Table 4 : London House Price statistics between 2009 - 2013 (Land Registry, 2014)

The figure below shows the house price in London for 2013 mapped in GIS where red indicates higher house price and blue indicates lower house price. The thematic distribution in GIS is calculated using the natural break method for 6 bands.

Figure 9 : Visualisation of London House Price in 2013 from red indicating high to green indicating low

The clustering nature of house price in London is obvious where the high house price cluster starts from the top of Hampstead Park passing through west London down to the south-western suburb of Richmond. The low house price cluster is concentrated to the east of Lea Valley, to the north and south of the Thames. To confirm the spatial clustering nature of house price in London, the global Moran's I statistics are calculated for the Log of house price for each year. The global Moran's I statistics ranges between 0.39 – 0.44 for the time period of 2009 – 2013. The exploratory result confirms significant levels of spatial clustering in London's house price over this time period. The next section will study how house price varies across the spatial network local areas.

Table 5 : London House Price Global Moran's I statistics from 2009 – 2013. The result confirms significant level of spatial clustering in London housing market. Please see appendix C for details. (Anselin et al. 2005)

Year	2009	2010	2011	2012	2013
Morans I	0.40	0.39	0.40	0.42	0.44

4.1 ANOVA Results

The table below illustrates the ANOVA results which tests whether the house price variations between the spatial local areas differs to within for the years 2009 - 2013. The P-value were statistically significant at the 0.01 level for all the years. The F-ratio are all above 100 and not significantly different between the five years. These initial results suggest house price in London are significantly similar within each spatial local areas but are different across each. The results are consistent between 2009 - 2013.

Table 6 : ANOVA Statistics. The result suggests house price are significantly different across spatial local areas.

2013	Sum of squares	Df	MS	**F**	**Prob > F**
Between Groups	7.00E+15	165	4.24E+13	**129.6**	**0**
Within Groups	2.22E+16	67693	3.27E+11		
Total	2.92E+16	67858	4.30E+11		

2012	Sum of squares	Df	MS	**F**	**Prob > F**
Between Groups	5.74E+15	165	3.48E+13	**103.49**	**0**
Within Groups	2.32E+16	69145	3.36E+11		
Total	2.90E+16	69310	4.18E+11		

2011	Sum of squares	Df	MS	**F**	**Prob > F**
Between Groups	4.20E+15	165	2.54E+13	**121.35**	**0**
Within Groups	1.46E+16	69487	2.10E+11		
Total	1.88E+16	69652	2.69E+11		

2010	Sum of squares	Df	MS	**F**	**Prob > F**
Between Groups	3.70E+15	165	2.24E+13	**119.18**	**0**
Within Groups	1.33E+16	7.08E+04	1.88E+11		
Total	1.70E+16	7.10E+04	2.40E+11		

2009	Sum of squares	Df	MS	**F**	**Prob > F**
Between Groups	2.26E+15	165	1.37E+13	**108.77**	**0**
Within Groups	7.23E+15	57518	1.26E+11		
Total	9.48E+15	57683	1.64E+11		

4.2 Regression Model Results

The table below illustrates the regression results for the four models. Model 1 is the basic hedonic model where the log of house price is regressed against a set of structural, accessibility and amenities variables. Model 2 includes the postcode-attribute housing submarket variable. Model 3 includes the ward-attribute housing submarket variable. Model 4 includes the spatial-attribute housing submarket variable.

Table 7 : Regression results comparing four hedonic models, a simple OLS model, a postcode-attribute housing submarket model, a ward-attribute housing submarket model and a spatial-attribute housing submarket model.

VARIABLES	(Whole) Model 1	(Postcode) Model 2	(Ward) Model 3	(Spatial) Model 4
intr20k	0.000327***	0.000199***	0.000242***	0.000221***
	(2.34e-06)	(8.72e-06)	(4.09e-06)	(3.07e-06)
2.type_id	-0.919***	-0.906***	-0.909***	-0.835***
	(0.0188)	(0.0186)	(0.0184)	(0.0169)
3.type_id	-0.499***	-0.487***	-0.486***	-0.441***
	(0.00980)	(0.00969)	(0.00962)	(0.00886)
4.type_id	-0.749***	-0.729***	-0.730***	-0.619***
	(0.00941)	(0.00935)	(0.00931)	(0.00871)
2.new_build_id	-0.0369***	-0.0304**	-0.0162	0.0527***
	(0.0122)	(0.0120)	(0.0120)	(0.0111)
2.hold_id	-0.347***	-0.355***	-0.351***	-0.376***
	(0.0166)	(0.0164)	(0.0162)	(0.0149)
shop_800	0.000456***	0.000388***	0.000295***	0.000359***
	(1.45e-05)	(1.45e-05)	(1.47e-05)	(1.32e-05)
off_800	3.31e-05***	5.84e-05***	4.60e-05***	0.000128***
	(5.47e-06)	(5.57e-06)	(5.53e-06)	(5.24e-06)
2.CL01_id		0.406***		
		(0.0270)		
3.CL01_id		0.405***		
		(0.0330)		
4.CL01_id		0.166***		
		(0.0209)		
5.CL01_id		0.0427***		
		(0.0108)		
6.CL01_id		0.00799		
		(0.0156)		
2.CL03_id			-0.000126	
			(0.00805)	
3.CL03_id			0.000150	
			(0.00850)	
4.CL03_id			-0.0185***	
			(0.00682)	
5.CL03_id			0.285***	
			(0.0101)	
6.CL03_id			0.395***	
			(0.0130)	
2.CL02_id				0.317***
				(0.0110)
3.CL02_id				-0.384***
				(0.00960)
4.CL02_id				-0.0731***
				(0.0122)
5.CL02_id				-0.0751***
				(0.00856)
6.CL02_id				-0.304***
				(0.0181)
Constant	12.40***	12.70***	12.64***	12.73***
	(0.0106)	(0.0184)	(0.0141)	(0.0114)
Observations	68,603	68,603	68,603	68,603
R-squared	0.408	0.422	0.432	0.523

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The P-value for all the variables are statistically significant at the 0.01 level for all four models. The R-square for Model 1 is 40.8% for the base case. The R-square for Model 2 is 42.2%. The R-square for Model 3 is 43.2%. Model 4 which uses the spatial-local-area as the geographical unit has the closest fit with an R-square of 52.3%. The estimates for all the variables have the expected sign where the estimates for the dwelling type, tenure, new built, shops density and office density all have similar estimates for all models. The estimates for integration are higher for Model 1 and lower for the other three models. The results suggest community detection methods can improve the definition of housing submarkets and are more effective than both the postcode-attributes and ward-attributes housing submarket model in explaining house price variations in London.

The last step of the analysis is to study the extent the four candidate models can explain spatial clustering of the housing market in London. Morans I have been computed for the residuals for all four model. Using the log-price of 2013 which achieved 0.44 as a base case, Model 1 achieved a 1% reduction. Model 2 and Model 3 achieved a 4% reduction in Morans I. Model 4 achieved a 16% reduction in Morans I. This signifies the spatial-attribute housing submarket model explains a significant amount of the spatial clustering effect of house price in London.

Table 8 : Model Residual Global Moran's I statistics in 2013. Please see Appendix D for details. (Anselin et al. 2005)

Model	1 Residual	2 Residual	3 Residual	4 Residual	Base case
Morans I	0.43	0.40	0.40	0.28	0.44

5.0 Discussion of Results

This research applied methods in community detection on defining spatial housing submarkets in London. The study first defined the spatial network local areas for the greater London area and found greater house price similarity within local areas than between. The study also found that these spatial local areas correspond with known local area boundaries. Importantly the study found that community detection technique can improve the definition of housing submarkets in explaining house price. This could help in the future for developing a more accurate predictive model. The goodness of fit is lower than previous research (Law, 2013) due to exclusion of some dwelling variables such as size and age with the reduced model. Future research would include these variables for validation to ensure the estimates are unbiased.

6.0 Conclusion

The contribution of this research is it represents a proof of concept in the use of community detection techniques in the definition of housing submarket. The spatial housing submarket model improves the explanation of house price in London. Notably it illustrates the significance in how spatial configuration influences house price not just at a global level in terms of accessibility (Law et al. 2013) but also at a meso level in terms of housing submarket. A more in depth analysis of each spatial housing submarket would be presented in the next paper.

Community detection technique not only improves the identification of housing submarket but it also relates to the identification of spatial network local areas. These spatial local areas were found to associate with known local area which differs between the planned and unplanned ones. This differences points to association between the fuzziness of spatial local areas and the fuzziness of unplanned areas. Further research is needed to study the association between spatial network local area and user-defined neighbourhood areas.

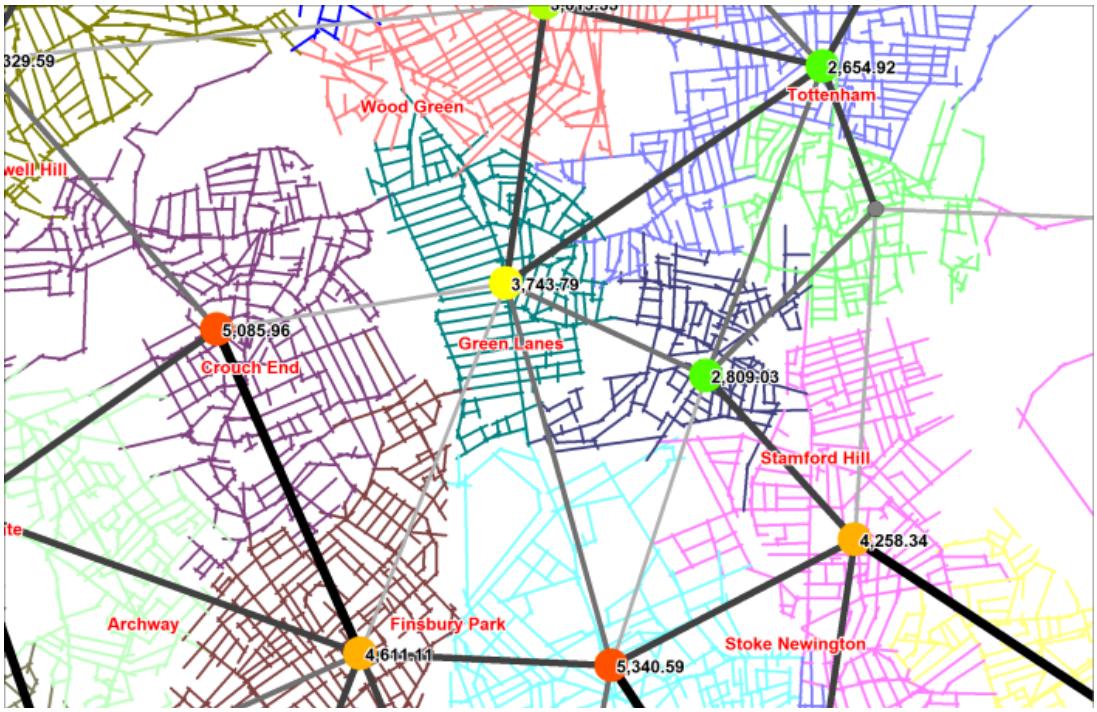


Figure 10 : Higher level representation of London supergraph. Thicker lines indicate high connectivity and thinner lines indicate low connectivity.

Lastly, further research will be carried out to study the spatial configuration of the subgraph. Interconnectivity and severances between local areas can thus be measured according to the connectivity of this super-graph. The figure above illustrates an example of a super-graph where the node represents the spatial network local area and the thickness of the lines represents the connectedness between them. By understanding cities as a system of connected spaces with multiple representations will allow a better spatial understanding and influences of the housing market.

Acknowledgement

We thank Dr. Sheep-Dalton for his inspiration, Dr. Dror Fidler for his kind advices, Jorge Gil and Shen Yao for their encouragements and discussions. We also thank the two reviewers for their thoughtful comments and Space Syntax Limited for their full support.

Appendix A

Data Item	Explanation
Transaction unique identifier	A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded.
Price	Sale price stated on the transfer deed.
Date of Transfer	Date when the sale was completed, as stated on the transfer deed.
Postcode	
Property Type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes
Old/New	Y = a newly built property, N = an established residential building
Duration	Relates to the tenure: F = Freehold, L-Leasehold etc.
PAON	Primary Addressable Object Name. If there is a sub-building for example the building is divided into flats, see Secondary Addressable Object Name (SAON).

SAON	Secondary Addressable Object Name. If there is a sub-building, for example the building is divided into flats, there will be a SAON.
Street	
Locality	
Town/City	
District	
County	
Record Status - monthly file only	Indicates additions, changes and deletions to the records.(please see guide

Appendix B Local Area Test Statistics

Hampstead_Garden_Suburb	FALSE	TRUE	Total
No	113222	74	113296
Yes	25	234	259
Total	113,247	308	113,555
Pearson chi2	7.80E+04	Pr = 0.000	
likelihood-ratio chi2	2.90E+03	Pr = 0.000	
Cramér's V	0.8281		

Hampstead Garden Suburb test statistics

Brentham Garden Suburb	FALSE	TRUE	Total
No	113483	32	113,515
Yes	4	36	40
Total	113,487	68	113,555
Pearson chi2	5.40E+04	Pr = 0.000	
likelihood-ratio chi2	532.0213	Pr = 0.000	
Cramér's V	0.6901		

Brentham Garden Suburb test statistics

Bedford Park	FALSE	TRUE	Total
No	113,471	24	113,495
Yes	0	60	60
Total	113,471	84	113,555
Pearson chi2	8.10E+04	Pr = 0.000	
likelihood-ratio chi2	924.9428	Pr = 0.000	
Cramér's V	0.8451		

Bedford Park Suburb test statistics

Thamesmead_Whole	FALSE	TRUE	Total
No	113,226	15	113,241
Yes	175	139	314
Total	113,401	154	113,555
Pearson chi2	4.50E+04	Pr = 0.000	
likelihood-ratio chi2	1.60E+03	Pr = 0.000	
Cramér's V	0.632		

Thamesmead District test statistics

Thamesmead_north_central	FALSE	TRUE	Total

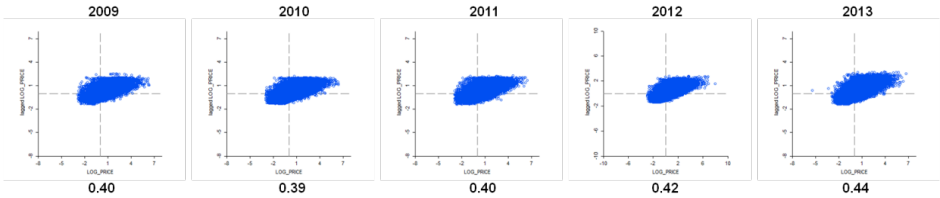
No	113,397	17	113,414
Yes	4	137	141
Total	113,401	154	113,555
Pearson chi2	9.80E+04	Pr = 0.000	
likelihood-ratio chi2	2.00E+03	Pr = 0.000	
Cramér's V	0.9296		

Thamesmead North and Central test statistics

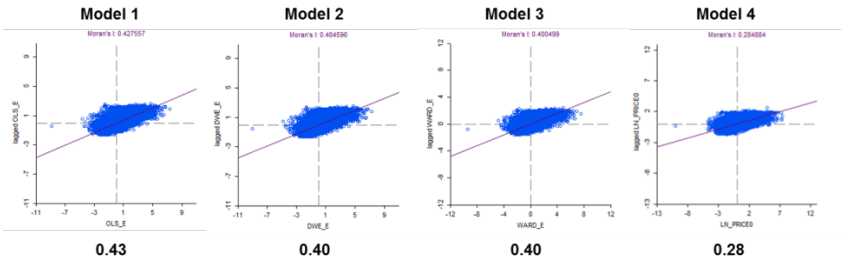
Soho	FALSE	TRUE	Total
No	113,131	237	113,368
Yes	103	84	187
Total	113,234	321	113,555
Pearson chi2	1.30E+04	Pr = 0.000	
likelihood-ratio chi2	753.188	Pr = 0.000	
Cramér's V	0.3415		

Soho Chi-square test statistics

Appendix C London House Price Morans I



Appendix D Model Residual Morans I



References

Rosen, S. (1974), Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34-55.7

Cheshire, P. and Sheppard, S. (1998), Estimating the demand for housing, land, and neighbourhood characteristics. Oxford Bulletin of Economics and Statistics 60: 357–382

Alonso, W. (1964), Location and Land Use: Toward a general Theory of Land Rent. Cambridge, Massachusetts: Harvard University Press.

Law, S., Karimi, K., Penn, A., Chiaradia, A. J. (2013), Measuring the influence of spatial configuration on the housing market in metropolitan London. In: Kim,Y.O.,Park, H.T.and Seo,K. W. (eds.),Proceedings of the Ninth International Space Syntax Symposium,Seoul: Sejong University, Article 121

Xiao, Y. (2012), Urban Morphology and Housing Market. A thesis submitted in partial fulfillment for the PhD. University of Cardiff

Grisby, W., Baratz, M., Galster, G. and Maclenna, N.D. (1987), The Dynamics of Neighbourhood Change and Decline. Oxford : Pergamon.

Bourassa, S. C., Hamelink, F., Hoesli, M. and MacGregor, B. D., (1999), "Defining Housing Submarkets," J. Housing Econ. 8, 160-183.

Bourassa, S.C., Cantoni, E., Hoesli, M. (2007), Spatial Dependence, Housing Submarkets, and House Price Prediction. Journal of Real Estate Finance and Economics, 2007,35:2, 143–60.

Dale-Johnson, D., (1982), "An Alternative Approach to Housing Market Segmentation Using Hedonic Pricing Data," J. Urban Econ. 11, 311-332.

Goodman, A.C. and Thibodeau, G. (1998), Housing market segmentation, Journal of Housing Economisc, 7(2) ,121-143.

Greater London Authority, (2004), London Housing Submarket Report. Greater London Authority. Accessed on 01/08/2014 via the world wide web. http://legacy.london.gov.uk/mayor/economic_unit/docs/housing_submarkets_report_for_web.pdf

Lynch, K. (1960), The image of the city, the MIT Press.

Galster, G. (2001), On the Nature of Neighbouhood. Urban studies 12:2111-2124. Publisher Full Text OpenURL

Kearns, A. and Parkinson, M., (2001). The significance of neighbouhood. Urban studies 2001, 38:2103-2110.

Bates, L.K. (2006), "Does Neighborhood Really Matter?: Comparing Historically Recognized Neighborhoods with Housing Submarkets." Journal of Planning Education and Research. 26(1):5-17.

Lebel, A., Pampalon, R., Villeneuve, P. Y. (2007), A multi-perspective approach for defining neighbourhood units in the context of a study on health inequalities in the Quebec City region. International Journal of Health Geographics 2007, 6:27 doi:10.1186/1476-072X-6-27

Hillier, B., Burdett, R., Peponis, J., Penn, A. (1987), "Creating Life: Or, Does Architecture Determine Anything?", Architecture & Comportment/ Architecture & Behaviour, 3 (3). pp. 233-250.

Peponis, J. (1988), Social structure of space and relations of community in six Greek cities, Secretariat of Research and Technology, Athens, Ministry of Energy Industry, Research and technology.

Read, S. (1999), Space syntax and the Dutch city. Environment and Planning B, 26, pp.251–264

Yang, T. and Hillier, B. (2007), The fuzzy boundary: the spatial definition of urban areas. In Proceedings 6th international space syntax symposium. Istanbul, Turkey, ed. A.S.Kubat, Ö Ertekin, and Y.Io. Güney, 091.01-22. Cenkler, Istanbul: Istanbul Technical University.

Dalton, N.S.C., (2006), Configuration and Neighborhood: Is Place Measurable. In Space Syntax and Spatial Cognition Workshop of the Spatial Cognition

Girvan, M. and Newman, M.E. (2002), Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), p.7821.

Reichardt, J. and Bornholdt, S., (2004), Phys. Rev. Lett. 93(21), 218701.

Raghavan, U. N., Albert, R., Kumara, S. (2007), Phys. Rev.E 76(3), 036106.

Newman, M.E.J. and Girvan. M. (2004). Finding and evaluating community structure in networks. Phys. Rev. E, 69:026113, 2004.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre. E.(2008), Fast unfolding of communities in large networks.J. Stat. Mech., 2008:P10008, 2008.

Guimer`a, R., Mossa, S., Turtschi, A., Amaral. L.A.N. (2005), The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proc. Natl. Acad. Sci. USA, 102:7794–7799, 2005.

Caschili, S., De Montis, A., Chessa, A., Deplano. G. (2009), Weighted networks and community detection: planning productive districts in sardinia. In G. Rabino and M. Caglioni, editors, Planning, complexity and new ICT, pages 27 – 36. Alinea Editrice s.r.l., 2009

Fortunato, S., (2010), "Community detection in graphs". Phys. Rep. 486 (3-5): 75–174. doi:10.1016/j.physrep.2009.11.002.

Porta,S., Crucitti, P., Latora,V. (2006), The network analysis of urban streets: a dual approach Physica A: Statistical Mechanics and its Applications 369 (2), 853-866

Hillier, B. and Hanson, J. (1984), The Social Logic of Space, Cambridge University Press: Cambridge.

Turner, A. (2007), "From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis" Environment and Planning B: Planning and Design 34(3) 539 – 555

Lancichinetti, A. and Fortunato, S. (2009), Community detection algorithms: a comparative analysis Physical Review E 80 (5), 056117

Lancichinetti, A. and Fortunato, S. (2011), Limits of modularity maximization in community detection Physical Review E 84 (6), 066122

Ordnance Survey (2014), Ordnance Survey Open Data Meridian 2 Dataset. Open source.Access on 01/08/2014 from the world wide web http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html. © Crown Copyright {2014}

Ordnance Survey (2015), Ordnance Survey Open Data Boundary Line Dataset. Open source.Access on 18/03/2015 from the world wide web http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html. © Crown Copyright {2015}

Ordnance Survey (2015), Ordnance Survey Open Data Open Local Map Dataset. Open source. Access on 18/03/2015 from the world wide web http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html. © Crown Copyright {2015}

London Borough of Barnet (Not available), Hampstead Garden Suburb Conservation Area Boundary Map. Accessed via the world wide web: http://www.barnet.gov.uk/downloads/file/2121/hampstead_garden_suburb

London Borough of Ealing (2008), Brentham Garden Estate Area Character Appraisal. Accessed via the world wide web: http://www.ealing.gov.uk/download/downloads/id/2091/brentham_garden_estate_area_appraisal

London Borough of Ealing (2007), Bedford Park Conservation Area Character Appraisal. Accessed via the world wide web: http://www2.ealing.gov.uk/ealing3/export/sites/ealingweb/services/council/committees/agendas_minutes_reports/regulatory_committees/planning_committee/15may2007-19may2008/_05_september_2007/Item_6_Bedford_Park_Conservation_Area_Report_Appendix.pdf

London Borough of Hounslow (Not Available), London Borough of Hounslow Conversation Areas. Accessed via the world wide web: http://www.hounslow.gov.uk/conservation_areas.pdf

Andrew Nunn Associates (Not Available**),** A guide to Bedford Park The first garden suburb. Accessed via the world wide web: http://www.andrewnunnassociates.co.uk/public/files/17981e29-c2a1-411b-ba1b-d6fff09decc5/client_files/09691700-ea9f-4f6b-8345-48bd3585642e/pdf/guide.pdf

Thamesmead Trust (2007), Thamesmead Street Map produced by Gallions Housing Association. Accessed via the world wide web: http://www.trust-thamesmead.co.uk/folder.cfm/id/461/folderName/Plans%20and%20Maps%20of%20Thamesmead

Sheppard, F.H.W. (1966), 'Map of the Parish of St. Anne, Soho', in Survey of London: Volumes 33 and 34, St Anne Soho, ed. F H W Sheppard (London, 1966), accessed via the world wide web. http://www.british-history.ac.uk/survey-london/vols33-4/map-of-st-anne-soho [accessed 12 January 2015].

Walter, T. (1878), 'Soho', in Old and New London: Volume 3 (London, 1878), pp. 173-184 Originally published by Cassell, Petter & Galpin, London, 1878. Accessed via the world wide web: http://www.british-history.ac.uk/old-new-london/vol3/pp173-184 [accessed 5 January 2015].

Soho boundary (Not Available), Soho. Wikipedia. Accessed via the world wide web: http://wikitravel.org/en/File:Areas_of_Central_London_I.png

Pearson, K., (1900), "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling".Philosophical Magazine Series 5 50 (302): 157–175.doi:10.1080/14786440009463897.

Cramér, H., (1999), Mathematical Methods of Statistics, Princeton University Press

Anselin, L., (1988), Spatial Econometrics: Methods and Models. Dordrecht: Kluwer Academic Publishers.

Anselin, L., Syabri, I., Kho, Y., (2005), GeoDa : An Introduction to Spatial Data Analysis. Geographical Analysis 38(1), 5-22.

Moran, P. A. (1950), "Notes on Continuous Stochastic Phenomena". Biometrika 37 (1): 17–23.

Wikipedia (2015), Analysis of Variance. Wikipedia. Accessed via the world wide web: http://en.wikipedia.org/wiki/Analysis_of_variance

Wikipedia (2015), F-test. Wikipedia. Accessed via the world wide web. http://en.wikipedia.org/wiki/F-test

Land Registry (2014), House Price Paid Data. Data produced by Land Registry © Crown copyright 2014.

Varoudis, T., (2012), 'depthmapX Multi-Platform Spatial Network Analysis Software', Version 0.30 OpenSource, http://varoudis.github.io/depthmapX/

Rossum, G. (2007), Python programming language. In USENIX Annual Technical Conference.