406

Estimating house price with spatial and land use accessibility components using a data science approach at the national scale

STEPHEN LAW, PO NIEN CHEN, YAO SHEN, KAYVAN KARIMI, & ALAN PENN,

UCL, LONDON, UK

ABSTRACT

Extensive research had been conducted studying the spatial configuration effects on house price using the hedonic price approach. Previous research has mostly focused on using econometric approaches in estimating house price. With the growing popularity of machine learning methods, there is an opportunity to study this problem from a data science perspective. Following Law et al (2017) which studied how economic value of closeness centrality (integration) differed across cities in England, we conduct here a similar experiment examining these differences using a data science approach. We leveraged on an integrated urban model, a large-scale geographic database to compute a series of land use accessibility and space syntax accessibility measures at the country scale (~120 measures). We then use a compressed set of spatial and land use accessibility components to estimate a set of hedonic price models in England; i. first for the entire country, then ii. for all 22 cities and then iii. for 22 cities individually. We found that spatial and land use accessibility features improve house price prediction accuracy jointly and the improvements are greater when using nonlinear methods. This research serves as a basis on the application of data science approaches in space syntax research for predicting real estate outcomes at the National-Scale.

KEYWORDS

Space syntax, land use, accessibility, data science , house price

1 INTRODUCTION

Estimating the economic value of geometric accessibility is an important and coherent line of research in the space syntax literature (Webster 2010). Earlier works suggest individuals are willing to spend more to live in more central areas with higher closeness centrality (integration), away from the main routes with lower betweenness centrality (choice) and are proximate to land use amenities. (Law et al., 2013; Xiao, 2015; Shen and Karimi, 2017). Many of these previous research focused on using econometric approaches to estimate house price at a city level. With the rise of data science and big data, there is a growing literature that focuses on using nonlinear methods to optimise out of sample predictive accuracy. A key reason to this shift is the emphasis on prediction accuracy as oppose to statistical inference. Both approaches have its strengths but in this research we will focus on the former.

In this research, we aim to combine space syntax and data science approach in trying to understand the extent spatial and land use accessibility measures can jointly improve house price prediction at the country-level. The results can improve our understanding on the joint effects of spatial and land use effect on house price but also how well can popular data science methods be used in space syntax research for house price prediction. The study will use a newly formed sold house price dataset and a novel Integrated Urban Model that calculates over ~120 spatial and land use accessibility variables in England.

The paper is organised as follows: the next section will introduce the related works on the use of space syntax and machine learning approaches on estimating house price; we will then describe the dataset, the street network and land use accessibility metrics and the empirical strategy for the house price prediction; we will end by reporting and discussing the results.

2 RELATED WORKS

Following Rosen’s economic framework (1974) , the hedonic price method is a popular approach to the valuation of intangible goods (Black 1999; Cheshire 1995; Goodman 1978; Gibbons and Machin 2005; Gibbons and Machin 2008; Ridker and Henning, 1967), and as inputs to land use and transportation models (Löchl & Axhausen, 2010). Theoretically the use of accessibility (Hansen 1959) stems from the concept of spatial equilibrium and the monocentric model (Alonso 1964; Muth; 1969 and Mills 1972) that explains how rent diminishes away from the central business district. The model operated through a bidding process whereby the people who capitalise the most from the assets acquire the right to the land. Based on the monocentric model, location differential is often estimated in the form of “Distance to the Central Business District (CBD)” in hedonic price modelling (Kain and Quigley, 1970). A limitation of this approach is the unrealistic definition of the CBD location (Ahlfeldt 2010; Heikkila et al. 1989). Responding to this limitation, space syntax research motivated the use of more complex forms of geometric accessibility in hedonic price model (Webster 2010; Xiao

2015; Law 2016). Holding all the property attributes constant, higher house prices can be found in more central areas with higher global integration (closeness), away from main routes with lower angular choice (betweenness) and more proximate to land use amenities. (Law et al., 2013; Xiao, 2015; Shen and Karimi, 2016). The strength of these approaches is it captures more complex accessibility potential than traditional measures of accessibility.

More recently, considerable successes have been found in the use of machine learning approaches in house price prediction. An early example is the works of Peterson and Flanagan (2009) who found a simple neural network model is able to outperform a linear regression model in predicting house price. Similarly, Peng et al. (2019) showed the extra gradient boosting regression (Chen and Guestrin 2016) is able to outperform other regression methods in predicting house price. These improvements are unsurprising as the more flexible functional form should be able to capture complex relations between the variables. For example the effects of being near a good school, a nice coffee shop and a neighbourhood park on house price are likely to be multiplicative. Despite these notable benefits, previous space syntax research have mostly focused on the use of linear econometric model on house price estimation. As a result, the aims of this research is to estimate house price using a combined space syntax and data science approach at the country scale. The following section will describe in greater details, the dataset, method and the experimental strategy.

3 DATASETS AND METHODS

3.1 Datasets

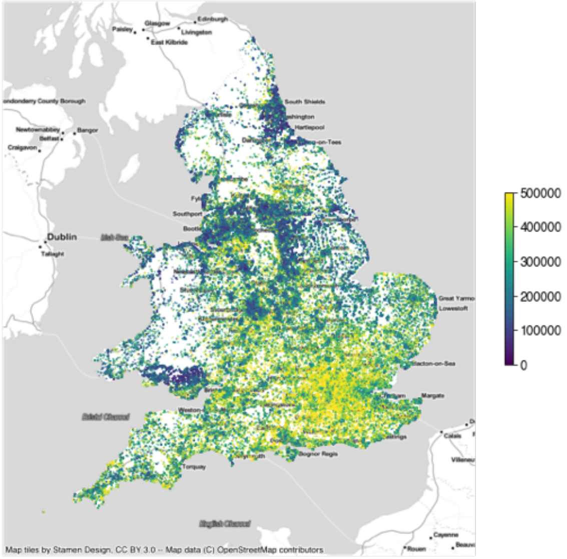
Following Law et al (2017), this study is framed within the area of England in United Kingdom. Three datasets have been compiled for the study which includes the property dataset, the spatial accessibility dataset and the land use accessibility dataset. The three datasets have been spatially joined at the postcode level.

3.1.1 Property dataset

The first dataset is a subset of the housing data from Chi et al. (2021) that contains the 2020 sold price data in England. This subset was constructed through fuzzy address-based matching between the Land Registry’s Price Paid Data (LR-PPD) and property information from the Ministry for Housing, Communities and Local Government (MHCLG)’s Domestic Energy Performance Certificates (EPCs). Table 1 shows a summary of the property dataset. A number of property attributes have been included in this dataset such as; the size of the property, the number of rooms in the property, the type of the property, the current epc score, the potential epc score, the tenure of the property and the building age of the property. A total of 512,587 transactions have been included in the experiment after data cleaning. The house price per sqm can be seen in figure 1 below.

names	description	source
price	price of property	land registry
size	size of property	mhclg
rooms	numbers of room	mhclg
cur_energy	current epc score	mhclg
pot_energy	potential epc score	mhclg
type	dwelling type {detached,semi,terrace,flat}	land registry
age	dwelling age {before 1960, post 1960}	UK census
tenure	dwelling tenure {freehold, leasehold}	land registry

Table 1: England house price per sqm



Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

Figure 1: England house price per sqm

3.1.2 Spatial accessibility dataset

The second dataset is the street network measures derived from the Space Syntax Database using the Ordnance Survey Open Roads and MasterMap Highways Network - Paths dataset. The spatial model has been further cleaned and segmented to reduce duplication and minimise the angular changes between streets. The spatial accessibility dataset has a total of 9,385,207 street segments in the mainland UK. Angular choice and angular integration have been calculated by the Place Syntax Tool (PST) for this research. The former measures the number overlaps at each street has for all shortest path pairs. While the latter measures the inverse mean angular distances from the root street to all other streets in the system. A summary of the measures can be found in (Freeman 1977; Hillier and Iida 2005; Hillier 2006; Hillier, Yang, Turner 2012). Figure 1 shows a visualisation of the spatial model and Table 2 shows all the variables from the spatial accessibility dataset.

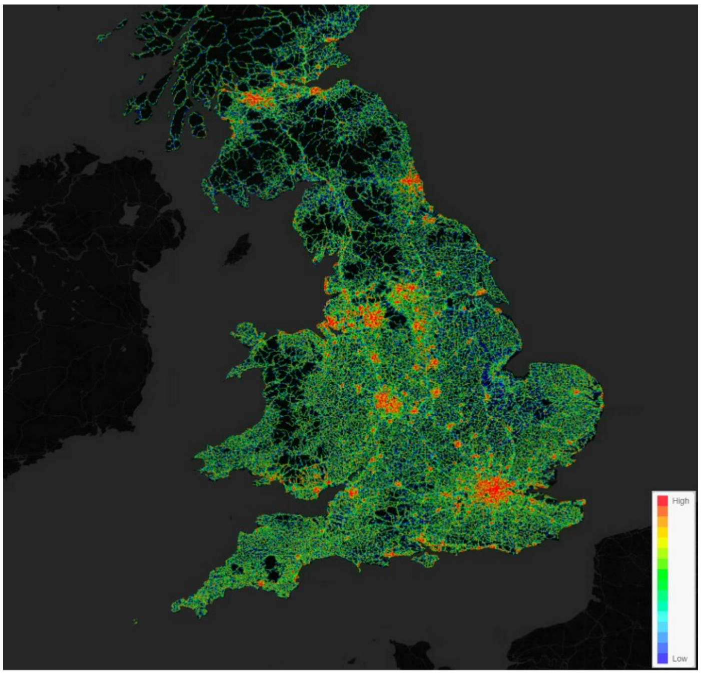


Figure 1: Angular Choice 2K

names	Description	source
ac_w400	choice at 400m	Place Syntax Tool (PST)
ac_w800	choice at 800m	Place Syntax Tool (PST)
ac_w1200	choice at 1200m	Place Syntax Tool (PST)
ac_w2000	choice at 2000m	Place Syntax Tool (PST)
ac_w5000	choice at 5000m	Place Syntax Tool (PST)
ac_w10000	choices at 10000m	Place Syntax Tool (PST)
ai_w400	integration at 400m	Place Syntax Tool (PST)
ai_w800	integration at 800m	Place Syntax Tool (PST)
ai_w1200	integration at 1200m	Place Syntax Tool (PST)
ai_w2000	integration at 2000m	Place Syntax Tool (PST)
ai_w5000	integration at 5000m	Place Syntax Tool (PST)
ai_w10000	integration at 10000m	Place Syntax Tool (PST)

Table 2: England house price per sqm

3.1.3 Land Use accessibility dataset

The third dataset is the land use accessibility measures derived from the Space Syntax Integrated Urban Model Database, using the urban land use data from 1. Ordinance Survey Addressbase Plus, 2. NHS Digital, 3. Ordinance Survey Open Greenspace, 4. School dataset from GOV.UK, and 5. NaPTAN from Department for Transport (DfT) and 6. the spatial model derived from the Ordnance Survey Open Roads and MasterMap Highways Network - Paths dataset in the previous section. The land use accessibility dataset has a total of 7,337,401 street segments and a total of 33,737,758 address points in the UK. A total of 117 variables have been calculated from this process as described in greater detail in Appendix A. Cumulative land use opportunities for every land use types, have been calculated from the Space Syntax IUM to create the land use accessibility index for 5mins (400m) and

15mins (1200m). The distance cost which has been assumed to process the catchment calculation is 5km/hour. Following this process, a reduced set of variables were calculated taking the sum of its individual categories. Table 3 shows the variables set from the Land use accessibility dataset.

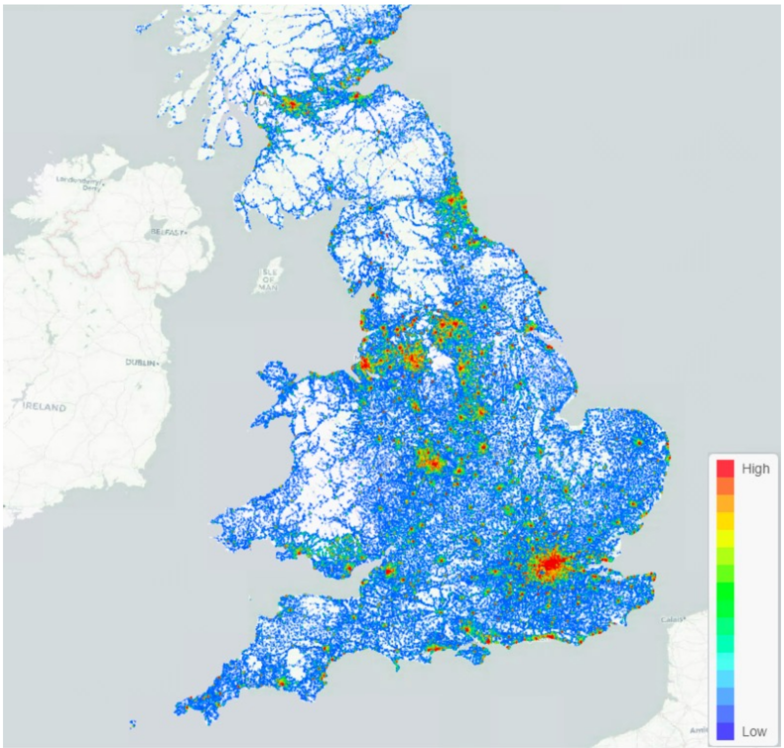


Figure 1: Land use mix accessibility 15mins

names	Description	source
nhs	No. nhs fac.	NHS Digital
greenspace	No. green space	Ordinance Survey
school	No. school fac.	GOV.UK
lesr	No. leisure fac.	Ordinance Survey
ofic	No. office	Ordinance Survey
rtl	No. retail	Ordinance Survey
srv	No. services	Ordinance Survey
cmrl	No. commercial	Ordinance Survey
naptan	No. transport fac.	DfT
nhs5	No. nhs fac. in 5mins	NHS Digital
greenspace5	No. green space in 5mins	Ordinance Survey
school5	No. school fac. in 5mins	GOV.UK
lesr5	No. of leisure fac. in 5mins	Ordinance Survey
ofic5	No. of office in 5mins	Ordinance Survey
rtl5	No. of retail in 5mins	Ordinance Survey
srv5	No. services in 5mins	Ordinance Survey
cmrl5	No. of commercial in 5mins	Ordinance Survey
naptan5	No. transport fac. in 5mins	DfT

Table 3: England house price per sqm

3.2 Exploratory Data Analysis and Principal Component Analysis

Figure 2 shows the correlation matrix between the land use and space syntax spatial accessibility variables. The correlation matrix shows high correlation between most of the space syntax accessibility variables in the upper left hand quadrant and some of the land use accessibility variables in the bottom right hand quadrant. Specifically, the choice and integration measures are highly correlated between similar radii reaching up to (r~0.93) for choice and (r~0.96) for integration. For the land use measures, a high correlation exists between retail and office use (r~0.76) and between retail and services (r~0.78).

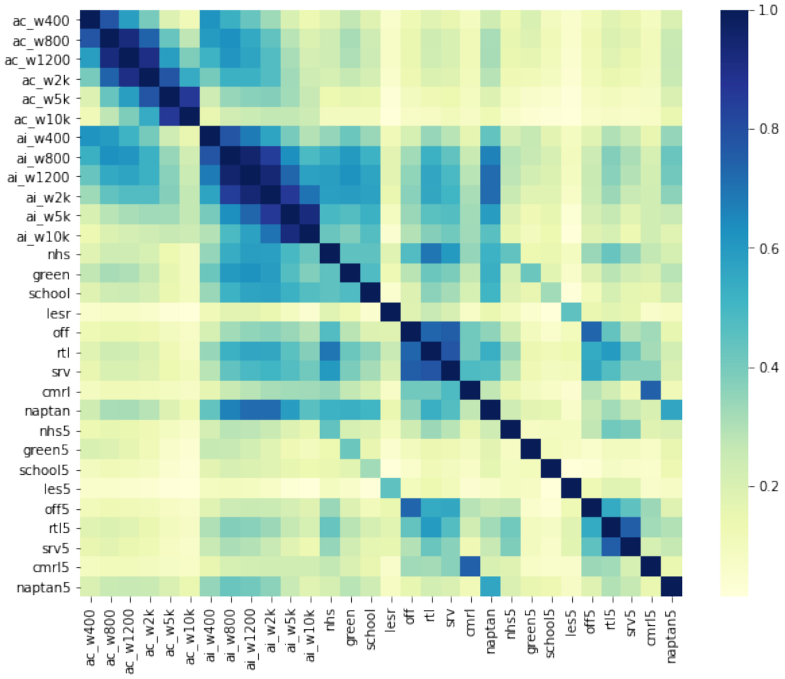


Figure 2: Correlation matrix between spatial accessibility and land use accessibility variables

Due to the high correlations among both sets of variables, we will apply a linear principal component analysis (Serra and Pinho 2013) on the spatial accessibility variables and the land use accessibility variables separately. We compress the 12 space syntax variables into 5 principal components that explains 97.47% of the data variance and the 18 land use accessibility variables into 5 principal components that explains 65.53% of the data variance. The spatial distribution of the two sets of principal components can be seen in figure 3 and figure 4 respectively. Descriptively, PC0 and PC1 of the spatial accessibility variables capture mostly a global structure while PC2 and PC3 captures mostly the local structure. On the other hand, PC0, PC2, PC3 of the land use accessibility variables capture mostly a global structure while PC1 and PC4 captures mostly a local structure. Please see Appendix B and C for details of all the variables that are used in the experiment.

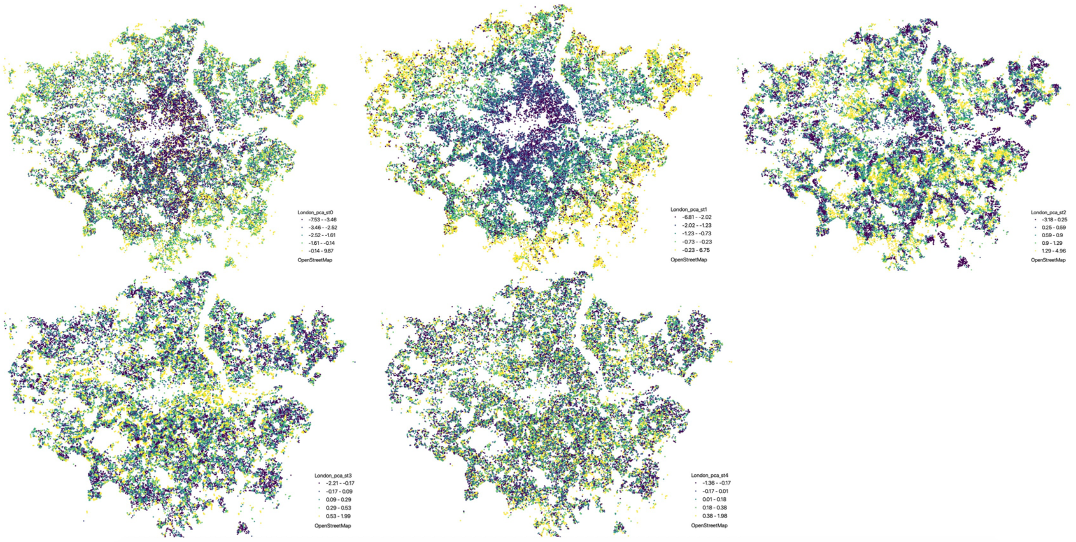


Figure 3: First five principal components of street network accessibility

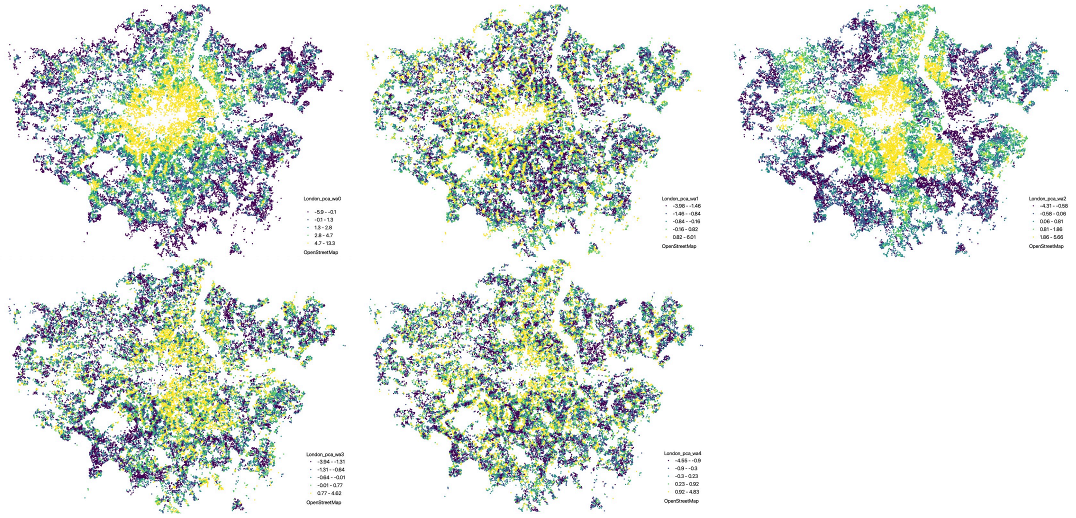


Figure 4: First five principal components of land use accessibility

4 EXPERIMENTAL DESIGN

4.1 Regression Details

In terms of experimental design, we will be running four regression models using three different types of regression methods for three spatial scales. More specifically we will compare a baseline model with only the housing features X which we call **H$_{\text{base}}$(X)** with a second model that includes both housing features X and space syntax features S **H$_{\text{st}}$(X,S)** a third model that includes the housing features X and land use features W **H$_{\text{wa}}$(X,W)** and finally a fourth model that includes housing, space syntax and land use features **H$_{\text{ws}}$(X,S,W)**. All four functions are parameterized respectively by set of weights θ (notation not shown explicitly). For each model, we will estimate three regression variants. The first is the baseline linear variant, which minimises the difference between the predicted $\hat{Y} = H(\cdot)$ and the observed Y with the loss as;

$$L(\cdot) = \frac{1}{n}\sum\left(Y - \hat{Y}\right)^2$$

We will then repeat this with the lasso variant which instead minimises the lasso objective[1] to reduce overfitting and to induce sparsity. We will finally repeat this with the extra gradient boosting regression algorithm otherwise known as *xgboost* (Chen and Guestrin 2016) which is a popular ensemble tree-base regression technique that sequentially fits new predictors to the residuals of the previous predictors that minimises the regularised differences between the observed and predicted while penalising tree complexity. The state of the art algorithm outperforms deep neural networks for tabular data (Shwartz-Siv and Armon 2022).

We will repeat this setup for three different scales. The country level scale where we will conduct this experiment for all the transactions in England at once, the pooled city scale which only includes major cities in the UK and lastly for the individual cities. The lists of cities can be found in table 4. In following a machine learning setup, model comparisons are made through the out of sample r2, mean absolute error and mean squared error where we split the dataset with a 70:30 ratio. Interpretability will be provided through feature importance plots. House price has been logged and all the input features have been standardised.

London	Middlesbrough	Portsmouth	Bristol
Manchester	Milton Keynes	Stoke-on-Trent	Southampton
Hudderfields	Liverpool	Brighton	Nottingham
Sheffield	Newcastle upon Tyne	Preston	Leeds
Birmingham	Leicester	Reading	
Bournemouth	Oxford	Cambridge	

Table 4: Major urban areas in England

5 RESULTS

This section summarises the empirical results for the three spatial scale namely the country level hedonic price model, the pooled-city hedonic price model and lastly the individual-city hedonic price model.

5.1 Country level

We first show the country-level results in table 4. The best model for all three regression variants is the baseline+street+landuse variables resulting in (r2=0.45,mae=0.38,mse=0.22) for the linear variant, (r2=0.44,mae=0.39,mse=0.23) for the lasso variant and (r2=0.61,mae=0.31,mse=0.16) for the xgboost variant. The results are substantially better with the nonlinear variant including both the space syntax and the land use accessibility variables. These results suggest at the country level that it is important

[1] The lasso ℓ-1 loss objective solves the original mean squared error loss subject to $||\theta||_1$ where θ are the parameters of the model tuned by a hyper-parameter α that is set to 0.001 for this study.

to include both space syntax and land use accessibility variables in a price prediction model. The difference are much greater when using a nonlinear regression approach (a difference in r2~0.6 for the linear approach vs a difference in r2~0.22 for the nonlinear approach). The feature importance in figure 5 shows structural features such as size, type of dwelling and age to be the most dominant features. Space Syntax PC2 and Land use PC4 which captures local structure are spatially the most important features for the three regression variants. The lasso variant interestingly shows fewer than half of the variables are necessary to achieve a regression with similar results to the linear model (difference of r2~0.01).

Country level

r2	linear	lasso	xgboost
base	0.39	0.38	0.43
base+st	0.42	0.41	0.53
base+wa	0.43	0.41	0.56
base+st+wa	**0.45**	**0.44**	**0.61**

mse	linear	lasso	xgboost
base	0.25	0.26	0.23
base+st	0.24	0.24	0.19
base+wa	0.24	0.24	0.18
base+st+wa	**0.22**	**0.23**	**0.16**

mae	linear	lasso	xgboost
base	0.4	0.4	0.38
base+st	0.39	0.39	0.35
base+wa	0.39	0.39	0.34
base+st+wa	**0.38**	**0.39**	**0.31**

Table 5: Regression results for country-level

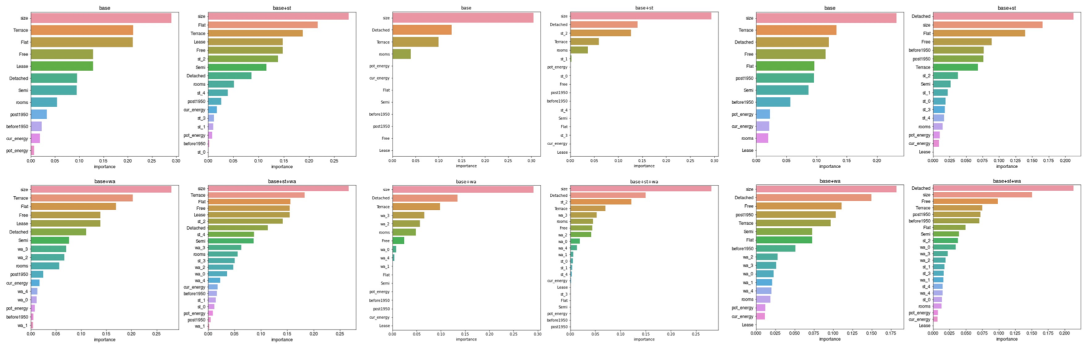


Figure 5: Feature importance for Linear Regressor (left) Lasso (centre) XGBoost (right)

5.2 Pooled-city level (Cities-only)

We then show the pooled-level cities only results in table 5. The best model for all three regression variants is similary the baseline+street+landuse variables resulting in r2~0.48, mae~0.42, mse~0.27 for the linear variant, r2~0.46, mae~0.43, mse~0.28 for the lasso variant and r2~0.68, mae~0.31, mse~0.16 for the nonlinear variant. The results are similar at the country level results with a slightly weaker loss only for the linear and lasso variant. Similarly results are better when including both space syntax and land use variables and substantially better with the nonlinear variant.
(diff of r2~0.10 vs r2~0.25). The feature importance in figure 6 shows similarly structural features such as size, type of dwelling and age and Space Syntax PC2 and Land use PC4 are the most important features for the three regression variants.

City Level

r2	linear	lasso	xgboost
base	0.38	0.37	0.43
base+st	0.43	0.42	0.6
base+wa	0.43	0.42	0.61
base+st+wa	**0.48**	**0.46**	**0.68**

mse	linear	lasso	xgboost
base	0.32	0.33	0.29
base+st	0.29	0.3	0.21
base+wa	0.29	0.3	0.2
base+st+wa	**0.27**	**0.28**	**0.16**

mae	linear	lasso	xgboost
base	0.46	0.47	0.43
base+st	0.44	0.45	0.36
base+wa	0.44	0.45	0.35
base+st+wa	**0.42**	**0.43**	**0.31**

Table 6: Regression results for the pooled-city level

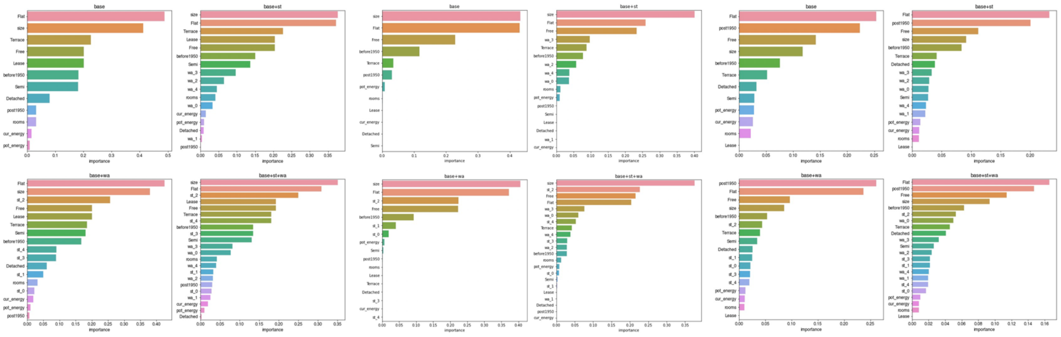


Figure 6: Feature importance for Linear Regressor (left) Lasso (centre) XGBoost (right)

5.3 Individual-City (cities-only)

Finally we show the results for individual cities in fig 7. The figure shows the results for the linear regression on the left, followed by lasso regression in the middle and xgboost regression on the right where the darker coloured cells indicate higher coefficient of determination (r2) and the lighter coloured cells indicate lower coefficient of determination (r2). The result shows generally the nonlinear variant r2~0.65-0.86 achieve a comparatively stronger predictive accuracy than the linear variant r2~0.6-0.82. The result shows substantial difference across cities.

The city that achieve the highest fit for the linear variant is Cambridge (r2~0.82), and Milton Keynes (r2~0.80) using all features. The model that achieve the highest fit for the nonlinear variant is Cambridge (r2~0.86), Bournemouth (r2~0.84). In all regression variants and for all cities, using only the housing features was worst than including both accessibility variables. Cities such as Greater London (diff in r2~0.18), Greater Manchester (diff in r2~0.22), Sheffield (diff in r2~0.23) achieved significant uplift when including the accessibility variables for the nonlinear variant. While cities such as Reading, Brighton, Oxford and Southampton only had a minor uplift when including the accessibility variables for the linear variant (diff in r2~0.01) but a larger uplift with the nonlinear variant (diff in r2~0.07-0.10).

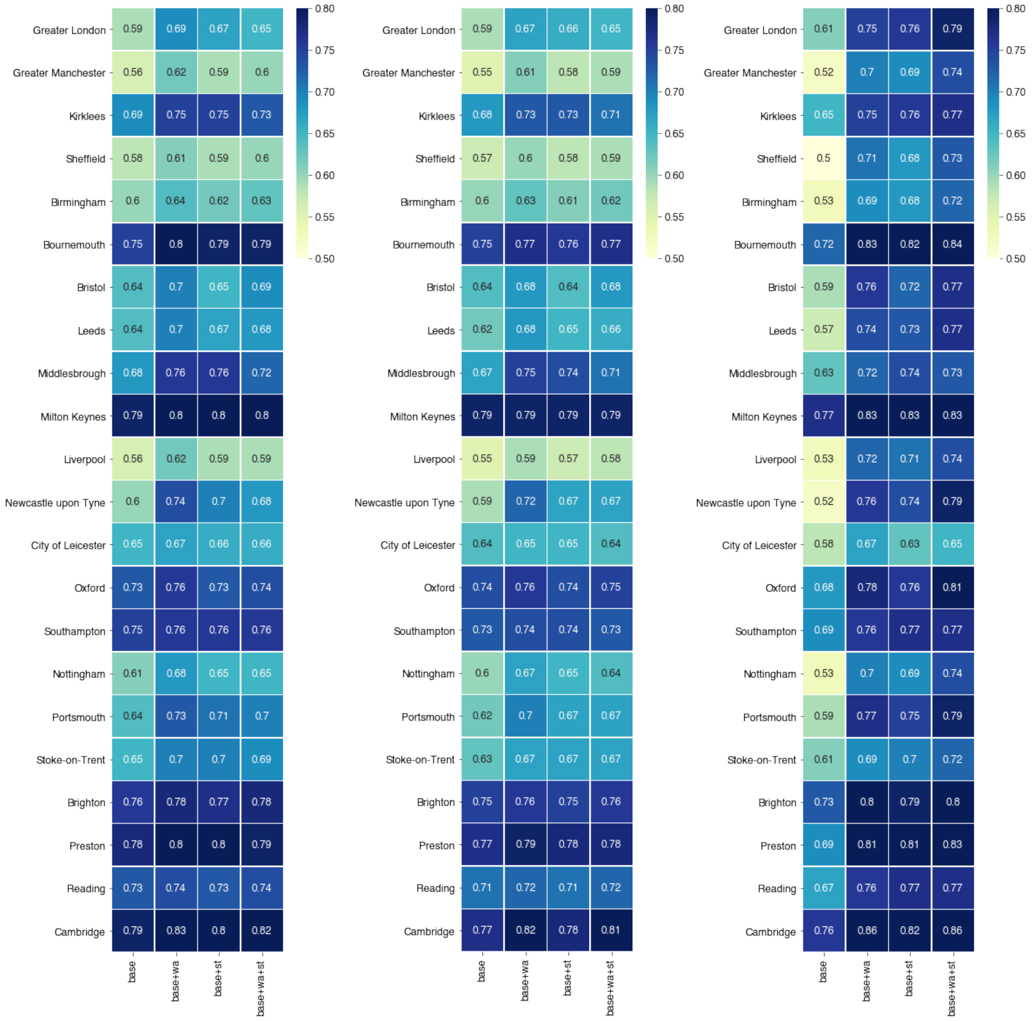


Figure 7: City-level goodness of fit r2 for Linear Regressor (left) Lasso (centre) XGBoost (right)

6 CONCLUSIONS

We estimated a series of regression models to predict house price in England comparing a baseline model with only housing attributes to a proposed model that includes housing and accessibility attributes using a data science approach. The result shows that out of sample prediction accuracy improves using both spatial and land use accessibility jointly. (from r2 ~43% to ~61% at the country level and from r2~43% to ~68% at the pooled city level and from r2~60-80% to ~72-86% at the individual city level) The implications here is that both spatial and land use accessibility components should be considered jointly when predicting house price at various scales, from the city up to the country level. The results were significantly better using a nonlinear variant as compared to the linear variant. These results show, interactions among variables should be considered when predicting house price. At the individual city level, the results show a similar trend but with notable differences between cities. As seen in previous research (Law et al 2017), results are difficult to generalise geographically. One interpretation is that different urban form induces different economic utilities from amenities. In summary, this research serves as a basis on applying data science approaches in space syntax research for predicting real estate outcomes at the National-Scale.

However, various limitations remain. Replicating the study in the future can potentially allow us to better understand the effect Covid-19 has on the housing market and how residents value accessibility differently before, during and after the pandemic. Figure 8 shows early result of such comparison with the feature importance map for 2019, 2020 and 2021(first months) at the country level (*xg.boost* variant). These early result do not show a significant absolute difference before and at the start of the pandemic in terms of feature importance for the accessibility components on house price. However it did have a minor increase in terms of ranking. Due to the smaller sample from the 2021 data, these early results are not conclusive. Further research using the 2022 data is necessary to help us better understand the effect the pandemic might have on the housing market.

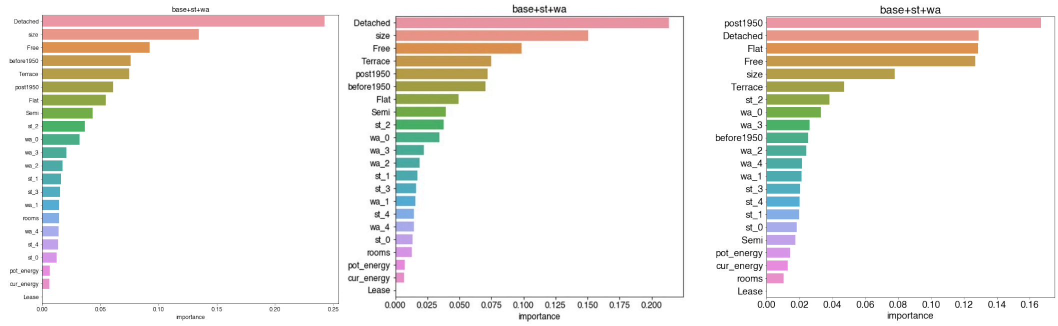


Figure 8: Feature importance for 2019 (left) 2020 (centre) and 2021 (right) at the country level

Furthermore, research to disentangle these differences are critical. One potential question is whether cities with different demographic profiles might have different value of spatial and land use accessibility following Law et al. (2017). Such investigations can help us better plan and allocate resources more efficiently. Repeating this investigation wasn't possible at the time of writing as the census data were not available. Another potential line of research is to use future versions of the Integrated Urban Model to estimate longitudinally the effect accessibility has on house price. Such research could estimate in a more robust manner (diff-in-diff) the effects spatial and land use accessibility has on house price, and would thus better bridge the gap between space syntax and real estate economics research.

REFERENCES

Ahlfeldt, G. (2010) If Alonso was right: Modeling Accessibility and Explaining the Residential Land Gradient. Journal of Regional Science

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

Black, S.E., (1999) "Do Better Schools Matter? Parental Valuation of Elementary Education," Quarterly Journal of Economics 114(2): 577-599.

Cheshire, P.C., and S. Sheppard, (1995) 'On the Price of Land and the Value of Amenities', Economica, 62, 247-267.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chi, B., Dennett, A., Oléron-Evans, T., & Morphet, R. (2021). Shedding new light on residential property price variation in England: A multi-scale exploration. *Environment and Planning B: Urban Analytics and City Science*, *48*(7), 1895-1911.

Freeman, L.C. (1977) A set of measures of centrality based on betweenness, Sociometry 40, 35-41. pp. 37. Lehigh University.

Gibbons, S. and Machin, S. (2005) Valuing rail access using transport innovations. Journal of UrbanEconomics, 57(1), 148- 169.

Gibbons, S. and Machin, S. (2008) Valuing school quality, better transport, and lower crime: evidence from house prices. Oxford review of economic policy, 24 (1): 99-119

Goodman A.C. (1978) 'Hedonic Price, Price Indices and Housing Markets', Journal of Urban Economics, 5, 471-484.

Hansen, W. G., (1959) How accessibility shapes land use. Journal of the American Institute of Planners, 25.

Heikkila, E., Gordon, P., Kim, J.I., Peiser, R.B., Richardson, H.W., Dale-Johnson, D., (1989). What happened to the CBD–distance gradient?: land values in a policentric city. Environment and Planning 21, 221–232.

Hillier, B. and Iida, S. (2005). Network and psychological effects in urban movement. In: Cohn, A.G. and Mark, D.M., (eds.) Proceedings of Spatial Information Theory: International Conference, COSIT 2005, Ellicottsville, N.Y., U.S.A.,September 14-18, 2005. Lecture Notes in Computer Science (Vol.

3693). Springer-Verlag, Berlin, Germany, pp. 475-490

Hillier, B., Yang, T., Turner, A., (2012). Normalising least angle choice in Depthmap - and how it opens up new perspectives on the global and local analysis of city space. JOSS 2012 P155-193

Hillier, B. (2006). Space is the machine. Cambridge, MA: Cambridge University Press.

Kain, J.F. , Quigley, J.M. (1970) Measuring the value of housing quality, Journal of the American Statistical Association 65 (440) (1970) 532–548.

Law, S., Karimi, K., Penn, A., Chiaradia, A. J. (2013)., Measuring the influence of spatial configuration on the housing market in metropolitan London. Published 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.

Law, S. (2016). Defining Street-based Local Area and measuring its effect on house price using the hedonic price approach: the case study of metropolitan London. Cities.

Law, S., Penn, A., Karimi, K., Shen, Y. (2017). The economic value of spatial network accessibility for uk cities: a comparative analysis using the hedonic price approach. In *Proceedings-11th*

International Space Syntax Symposium, SSS 2017 (Vol. 11, pp. 77-1). Instituto Superior Técnico, Portugal.

Löchl, M. , Axhausen, K.W. (2010) Modelling hedonic residential rents for land use and transport simulation while considering spatial effects The Journal of Transport and Land Use, Volume 3, 2010, pp. 39–63

Mills, S.E., (1972) Studies in the Structure of the Urban Economy. Baltimore: Johns Hopkins Press.

Muth, R., (1969) Cities and Housing. Chicago: University of Chicago Press.

Office of the Deputy Prime Minister (2006). State of the Cities Vol.1 – A research study. ODPM Publications.

Peng, Z., Huang, Q., & Han, Y. (2019, October). Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm. In *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)* (pp. 168-172). IEEE.

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of real estate research*, *31*(2), 147-164.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82(1), 34-55.

Serra, M., & Pinho, P. (2013). Tackling the structure of very large spatial systems-Space syntax and the analysis of metropolitan form. The Journal of Space Syntax, 4(2), 179-196.

Shen, Y., & Karimi, K. (2017). The economic value of streets: mix-scale spatio-functional interaction and housing price patterns. *Applied Geography*, *79*, 187-202.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84-90.

Webster C, (2010). "Pricing accessibility: urban morphology, design and missing markets" Progress in Planning 73 77–111

Von Thunen, J. (1826) Isolated State: An English Edition of Der isolierte Staat (translated by Wartenberg C and Hall P). Pergamon Press, Oxford, UK.

Xiao,Y., Webster, C., Orford, S. (2015) Urban Configuration, Accessibility, and Property Prices: A Case study of Cardiff, Wales. Environment and Planning B: Planning and Design. 2015. Vol.42, Pages 1-22

APPENDIX 1

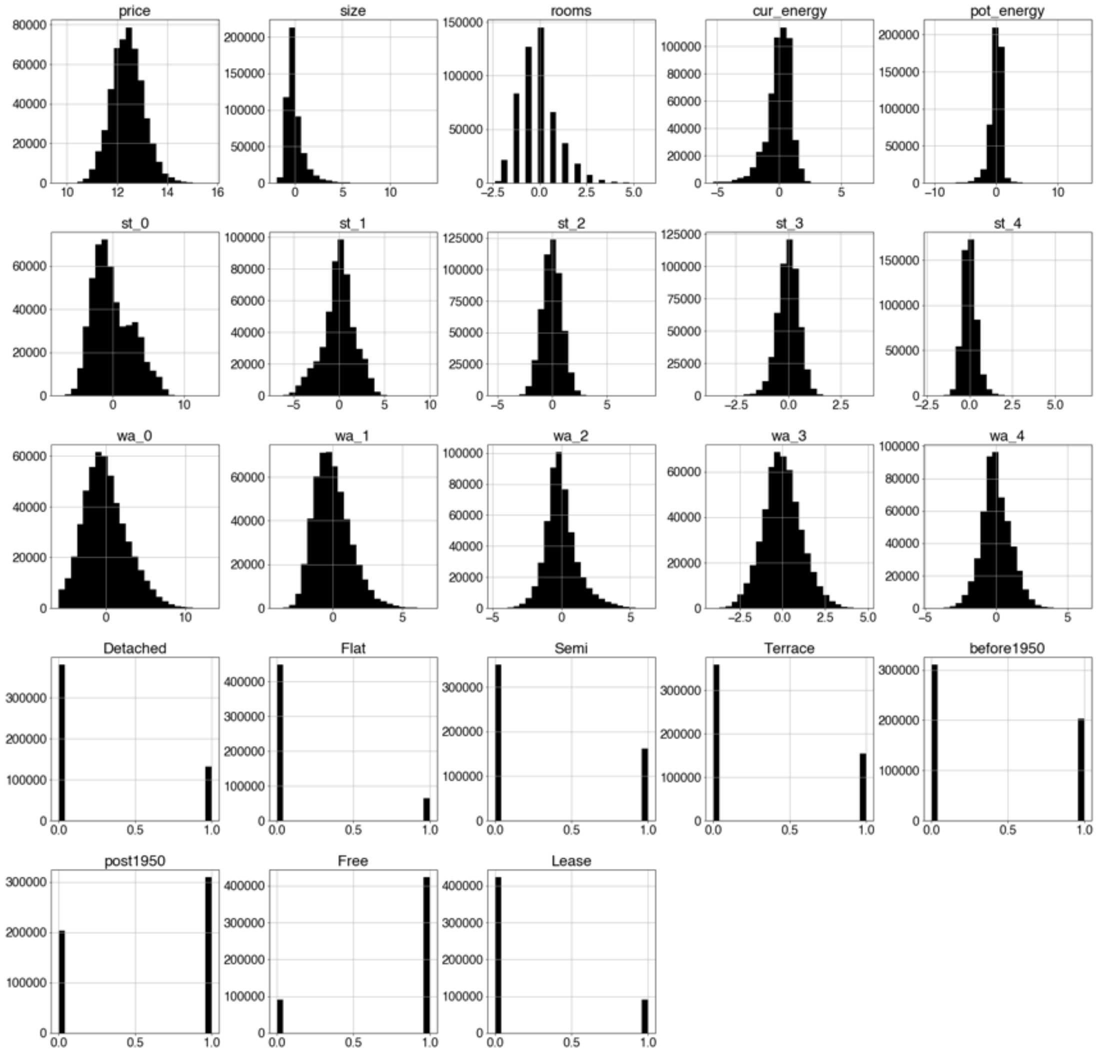
nhs	choices_gb_nhs_services_walk_gp,choices_gb_nhs_services_walk_nhs_nurhm,choices_gb_nhs_services_walk_nhs_den_all,choices_gb_nhs_services_walk_nhs_denpv,choices_gb_nhs_services_walk_nhs_den,choices_gb_nhs_services_walk_nhs_gp,choices_gb_nhs_services_walk_nhs_hospice,choices_gb_nhs_services_walk_nhs_hsptl,choices_gb_nhs_services_walk_nhs_hsptl_srv,choices_gb_nhs_services_walk_nhs_hsptl_ind,choices_gb_nhs_services_walk_nhs_ooh,choices_gb_nhs_services_walk_nhs_phmcy,choices_gb_nhs_services_walk_nhs_emerg,choices_gb_nhs_services_walk_nhs_wicooh,choices_gb_nhs_services_walk_nhs_wic
greenspace	choices_gb_greenspace_site_walk_grn_sprt,choices_gb_greenspace_site_walk_grn_plys,choices_gb_greenspace_site_walk_grn_plyf,choices_gb_greenspace_site_walk_grn_pb,choices_gb_greenspace_site_walk_grn_grow
school	choices_gb_school_walk_schl_pri,choices_gb_school_walk_schl_sec,choices_gb_school_walk_schl_prisec,choices_gb_school_walk_schl_nurs,choices_gb_school_walk_schl_midsec,choices_gb_school_walk_schl_spec,choices_gb_school_walk_schl_16,choices_gb_school_walk_schl_midpri,
lesr	choices_gb_lesr_gnl_walk,choices_gb_lesr_amus_walk,choices_gb_lesr_camp_walk,choices_gb_lesr_clt_walk,choices_gb_lesr_thmp_walk,choices_gb_lesr_bch_walk,choices_gb_lesr_soci_walk,choices_gb_lesr_stdm_walk,choices_gb_lesr_sprt_walk
ofic	choices_gb_ofic_gnl_walk,choices_gb_ofic_stdo_walk,choices_gb_ofic_bcst_walk
rtl	choices_gb_rtl_gnl_walk,choices_gb_rtl_mkt_walk,choices_gb_rtl_fd_walk,choices_gb_rtl_shp_walk,
srv	choices_gb_srv_anml_walk,choices_gb_srv_comn_walk,choices_gb_srv_tolt_walk,choices_gb_srv_ofic_walk,choices_gb_srv_bnk_walk,choices_gb_srv_info_walk,choices_gb_srv_mail_walk,choices_gb_srv_post_walk,choices_gb_srv_atm_walk,choices_gb_srv_rtl_walk,
cmrl	choices_gb_cmrl_gnl_walk
naptan	choices_gb_naptan_stops_walk_bct,choices_gb_naptan_stops_walk_rse,choices_gb_naptan_stops_walk_tmu

nhs5	choices_gb_nhs_services_walk_nhs_gp_5,choices_gb_nhs_services_walk_nhs_hsptl_5,choices_gb_nhs_services_walk_nhs_hsptl_srv_5,choices_gb_nhs_services_walk_nhs_hsptl_ind_5,choices_gb_nhs_services_walk_nhs_ooh_5,choices_gb_nhs_services_walk_nhs_phmcy_5,choices_gb_nhs_services_walk_nhs_emerg_5,choices_gb_nhs_services_walk_nhs_wicooh_5,choices_gb_nhs_services_walk_nhs_wic_5,choices_gb_nhs_services_walk_nhs_nurhm_5,choices_gb_nhs_services_walk_nhs_comnsrv_5,choices_gb_nhs_services_walk_nhs_denpv_5,choices_gb_nhs_services_walk_nhs_den_5,choices_gb_nhs_services_walk_nhs_hospice_5,choices_gb_nhs_services_walk_nhs_den_all_5
greenspace5	choices_gb_greenspace_site_walk_grn_grow_5,choices_gb_greenspace_site_walk_grn_sprt_5,choices_gb_greenspace_site_walk_grn_plys_5,choices_gb_greenspace_site_walk_grn_plyf_5,choices_gb_greenspace_site_walk_grn_pb_5
school5	choices_gb_school_walk_sch_pri_5,choices_gb_school_walk_schl_sec_5,choices_gb_school_walk_schl_prisec_5,choices_gb_school_walk_schl_nurs_5,choices_gb_school_walk_schl_midsec_5,choices_gb_school_walk_schl_spec_5,choices_gb_school_walk_schl_16_5,choices_gb_school_walk_schl_midpri_5,choices_gb_school_walk_schl_all_5,choices_gb_school_walk_schl_pri_5,choices_gb_school_walk_schl_all
lesr5	choices_gb_lesr_clt_walk_5,choices_gb_lesr_sprt_walk_5,choices_gb_lesr_gnl_walk_5,choices_gb_lesr_amus_walk_5,choices_gb_lesr_camp_walk_5,choices_gb_lesr_bch_walk_5,choices_gb_lesr_soci_walk_5,choices_gb_lesr_stdm_walk_5,choices_gb_lesr_thmp_walk_5
ofic5	choices_gb_ofic_gnl_walk_5,choices_gb_ofic_stdo_walk_5,choices_gb_ofic_bcst_walk_5
rtl5	choices_gb_rtl_gnl_walk_5,choices_gb_rtl_mkt_walk_5,choices_gb_rtl_fd_walk_5,choices_gb_rtl_shp_walk_5,
srv5	choices_gb_srv_bnk_walk_5,choices_gb_srv_post_walk_5,choices_gb_srv_atm_walk_5,choices_gb_srv_comn_walk_5,choices_gb_srv_tolt_walk_5,choices_gb_srv_ofic_walk_5,choices_gb_srv_rtl_walk_5
cmrl5	choices_gb_cmrl_gnl_walk_5
naptan5	choices_gb_naptan_stops_walk_rse_5,choices_gb_naptan_stops_walk_tmu_5,choices_gb_naptan_stops_walk_bct_5

APPENDIX 2

	count	mean	std	min	25%	50%	75%	max
price	512587	12.41	0.64	9.68	11.98	12.41	12.83	15.76
size	512587	0.00	1.00	-1.88	-0.59	-0.24	0.30	14.17
rooms	512587	0.00	1.00	-2.36	-0.49	0.14	0.76	5.74
cur_energy	512587	0.00	1.00	-5.31	-0.51	0.19	0.63	7.00
pot_energy	512587	0.00	1.00	-10.47	-0.47	0.19	0.59	14.28
st_0	512587	0.00	2.77	-7.53	-2.08	-0.55	1.99	13.84
st_1	512587	0.00	1.73	-6.81	-0.93	0.09	1.04	9.94
st_2	512587	0.00	0.88	-5.19	-0.61	0.00	0.60	8.74
st_3	512587	0.00	0.51	-3.57	-0.32	0.02	0.34	3.70
st_4	512587	0.00	0.41	-2.23	-0.27	-0.03	0.24	6.68
wa_0	512587	0.00	2.74	-5.91	-1.93	-0.28	1.66	13.30
wa_1	512587	0.00	1.29	-3.98	-0.93	-0.15	0.77	6.93
wa_2	512587	0.00	1.17	-4.80	-0.71	-0.13	0.56	6.27
wa_3	512587	0.00	1.11	-4.04	-0.72	-0.04	0.70	4.89
wa_4	512587	0.00	1.02	-4.55	-0.65	-0.08	0.65	6.18
Detached	512587	0.26	0.44	0.00	0.00	0.00	1.00	1.00
Flat	512587	0.13	0.33	0.00	0.00	0.00	0.00	1.00
Semi	512587	0.32	0.47	0.00	0.00	0.00	1.00	1.00
Terrace	512587	0.30	0.46	0.00	0.00	0.00	1.00	1.00
before1950	512587	0.39	0.49	0.00	0.00	0.00	1.00	1.00
post1950	512587	0.60	0.49	0.00	0.00	1.00	1.00	1.00
Free	512587	0.82	0.38	0.00	1.00	1.00	1.00	1.00
Lease	512587	0.18	0.38	0.00	0.00	0.00	0.00	1.00

APPENDIX 3


price
0
20000
40000
60000
80000
10
12
14
16
size
0
50000
100000
150000
200000
0
5
10
rooms
0
50000
100000
150000
−2.5
0.0
2.5
5.0
cur_energy
0
20000
40000
60000
80000
100000
−5
0
5
pot_energy
0
50000
100000
150000
200000
−10
0
10
st_0
0
20000
40000
60000
0
10
st_1
0
20000
40000
60000
80000
100000
−5
0
5
10
st_2
0
25000
50000
75000
100000
125000
−5
0
5
st_3
0
25000
50000
75000
100000
125000
−2.5
0.0
2.5
st_4
0
50000
100000
150000
−2.5
0.0
2.5
5.0
wa_0
0
20000
40000
60000
0
10
wa_1
0
20000
40000
60000
0
5
wa_2
0
20000
40000
60000
80000
100000
−5
0
5
wa_3
0
20000
40000
60000
−2.5
0.0
2.5
5.0
wa_4
0
20000
40000
60000
80000
100000
−5
0
5
Detached
0
100000
200000
300000
0.0
0.5
1.0
Flat
0
100000
200000
300000
400000
0.0
0.5
1.0
Semi
0
100000
200000
300000
0.0
0.5
1.0
Terrace
0
100000
200000
300000
0.0
0.5
1.0
before1950
0
100000
200000
300000
0.0
0.5
1.0
post1950
0
100000
200000
300000
0.0
0.5
1.0
Free
0
100000
200000
300000
400000
0.0
0.5
1.0
Lease
0
100000
200000
300000
400000
0.0
0.5
1.0