Multivariate Analysis of Spatial Inequality in the United Kingdom

April 27, 2025

1 Introduction

I have been interested in spatial economics, i.e. how economic agents interact with one another with respect to distance. Particularly, I find spatial inequality to be a particularly pressing and persistent problem, where some regions prosper while others are left behind. One key driver of spatial inequality is real estate and housing affordability, some areas are just more expensive than others, and so blocks off opportunities to those who can't afford to live there. To put this into perspective, Figure 1 shows that in the United States, the average home price has been steadily rising since the 1960s to the point that owning a home becomes less likely for the average person. This is prevalent in the rest of the world as housing prices grow much faster than do people's incomes and rent (Piketty, 2014). Such regional disparities are even apparent in the richer and developed parts of the world, these trends make the United Kingdom a compelling case to examine spatial inequality more closely.

In this project, I examine how spatial inequality manifests across towns in the United Kingdom. Specifically, I am interested in the following questions: What latent factors capture the structural variation in housing, economic, and demographic characteristics across UK towns (Factor Analysis)? Can towns be grouped into distinct clusters based on their observed socioeconomic, housing, and demographic characteristics, and do these clusters reflect meaningful differences in wealth, affordability, and vulnerability (Hierarchical Clustering)? Do differences in towns' socioeconomic characteristics predict significant multivariate differences in housing outcomes, confirming that spatial inequality is meaningfully structured (Generalized Linear Model)?

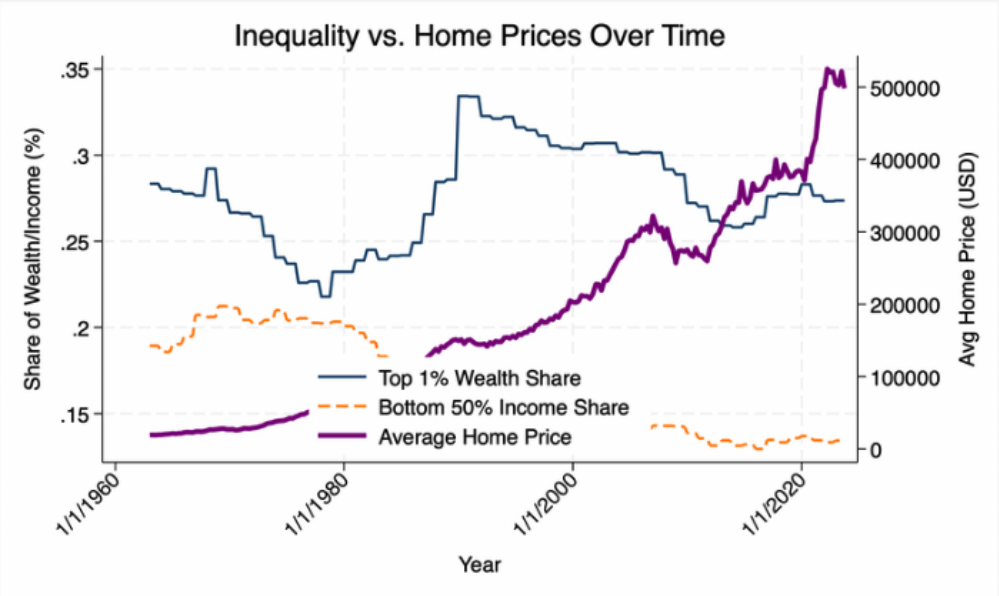


Figure 1: Rising Housing Prices in the US

Sources: World Inequality Database and Federal Reserve Economic Data, Federal Reserve Bank of St. Louis.

2 Data Overview

To best study the nature of spatial inequality in the UK, I employ the 2016 Housing and Commuting dataset compiled by Prothero (2016) from the Office for National Statistics, which operates under the UK Statistics Authority. This dataset is drawn from several official sources, including the 2011 Census, the 2015 Index of Multiple Deprivation (IMD), and Land Registry housing statistics dating back to 2004. It uses a geographical framework that defines towns and cities based on built-up areas rather than administrative boundaries, allowing for more consistent comparisons across urban regions. Key indicators are sourced at the Lower Super Output Area (LSOA) level and then aggregated to the town level.

However, the dataset poses several limitations. First, it spans multiple points in time—such as the 2011 Census and housing data up to 2015 (which may affect the temporal coherence of the analysis). Second, aggregating LSOA-level data to the town level may mask internal variation within towns, obscuring pockets of deprivation or concentrated affluence. Finally, the IMD is a composite measure that integrates several domains of deprivation, each with varying degrees of data quality and consistency; as a relative ranking, it indicates a town's position compared to others but does not convey the absolute level of deprivation. The 10 key variables span housing, demographic, and vulnerability indicators:

- Median house price in 2015 (*all_price_2015*)

- Percentage change in house prices since 2004 (*housepricechange*)
- Percentage of town population that are homeowners (*households_owned*)
- Percentage of town population that privately rents (*households_rent_private*)
- Relative level of recorded crime (*crime_rank*)
- Proportion of residents with health vulnerabilities (*prop_health_vuln*)
- Percentage of working-age adults with no qualifications (*noqual*)
- Proportion of residents aged 16–74 who are students (*student_prop_16_74*)
- Net commuter flow (*net_commuting*)
- Proportion of workers in manufacturing sector (*prop_manu*)

Later, in the multivariate Generalized Linear Model (GLM) analysis, I construct three binary variables based on categorical predictors: whether a town has experienced significant house price growth (greater than the median 10% increase) since 2004 (*high_growth*), whether it ranks above the median in income deprivation (*high_icd*), and whether it has commuter flows exceeding the median (*high_commuting*). These variables' summary statistics are detailed in Table 1, while their pairwise correlations are visualized in Figure 2. Figure 3 displays the chi-square quantile plot of the dataset and it suggests that the data is somewhat but not perfectly multivariate normal (especially with that one notable outlier). Since the variables are measured on different scales (e.g. house prices, percentages, commuting flows), all variables are standardized prior to analysis to ensure comparability for factor analysis and hierarchical clustering.

3 Factor Analysis

Going back to Figure 2, the correlation matrix reveals several strong and interesting correlations among the variables. Most notably, median house price in 2015 is highly positively correlated with house price growth ($r \approx 0.86$), suggesting that wealthier towns have also experienced greater housing market acceleration, reinforcing patterns of spatial inequality. Additionally, house prices show strong negative correlations with indicators of socioeconomic vulnerability, including the proportion of residents with no qualifications ($r \approx -0.84$) and those with health vulnerabilities ($r \approx -0.79$). These patterns suggest that more affluent towns tend to have better-educated and healthier populations. The matrix also shows expected tenure dynamics: private renting and owner-occupation are negatively correlated

Variable	Type	Min	Median	Mean	Max
all_price_2015	Continuous	78,000	145,000	170,363	390,000
housepricechange	Continuous	-3.85	10.74	13.30	46.94
households_owned	Continuous	33.55	60.71	60.30	81.01
households_rent_private	Continuous	9.93	17.69	17.74	32.45
crime_rank	Categorical	1	55	55	109
prop_health_vuln	Continuous	2.76	6.05	6.21	10.96
noqual (in % of town population)	Continuous	12.18	23.92	24.13	37.94
student_prop_16_74	Continuous	5.61	7.82	9.99	26.69
net_commuting	Continuous	-18,189	4,676	12,765	498,946
prop_manu	Continuous	3.13	9.27	9.69	23.80
high_growth	Binary	0	0	0.50	1
high_icd	Binary	0	0	0.50	1
high_commuting	Binary	0	0	0.50	1
region	Categorical				

Table 1: Type and summary statistics for variables used in the analysis

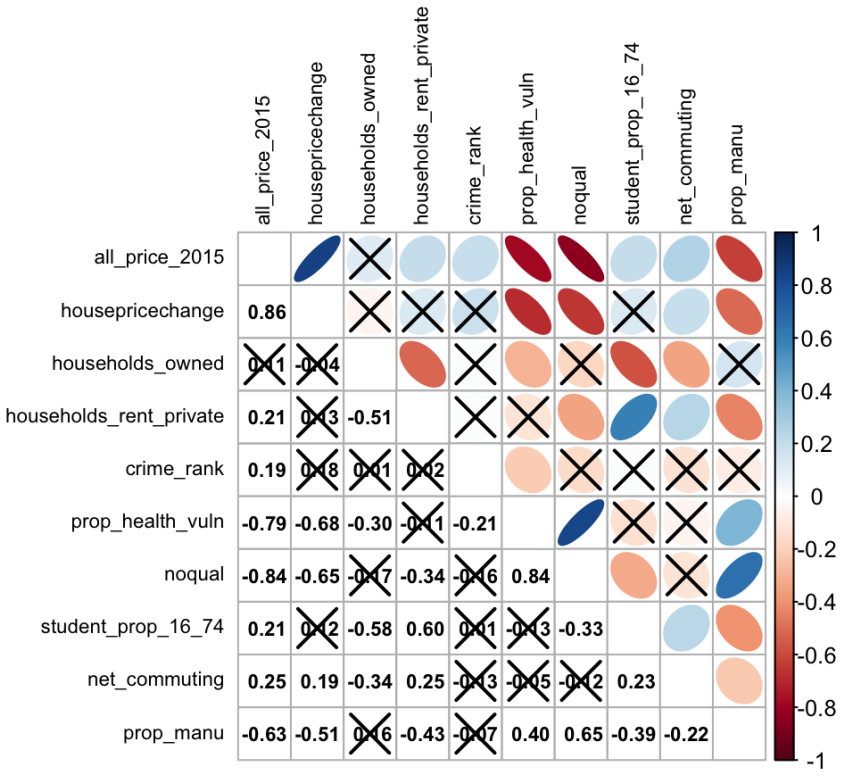


Figure 2: Correlation Matrix of Key Variables with 95% Confidence Intervals

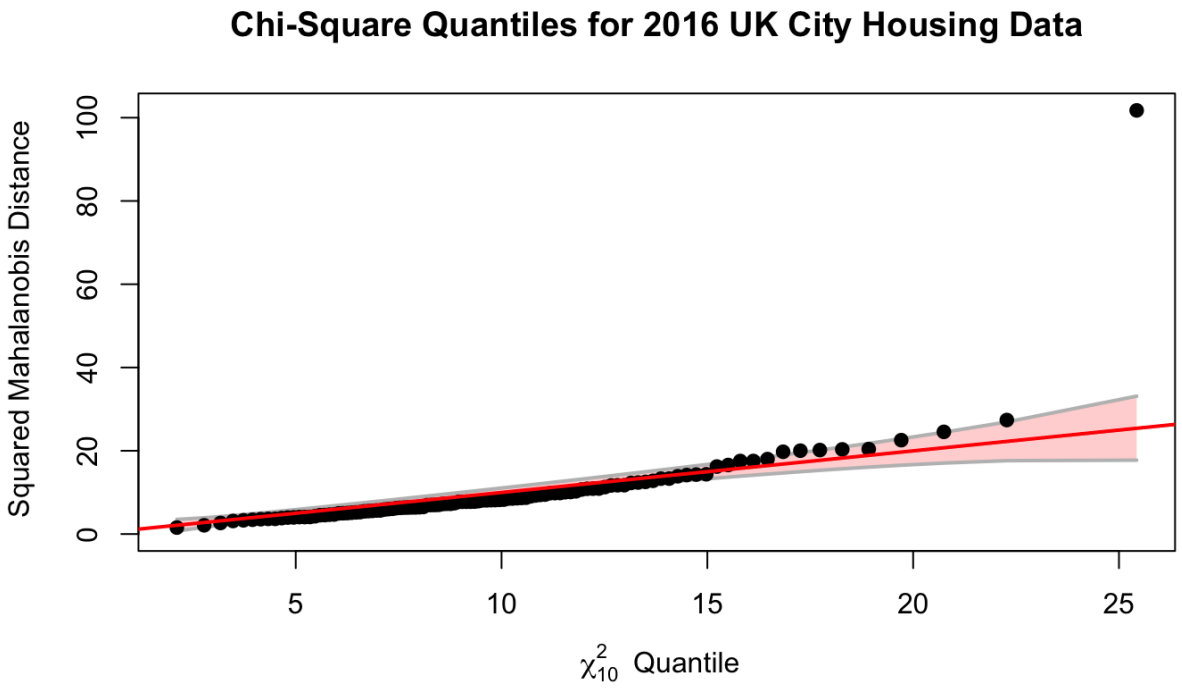


Figure 3: Chi-Square Quantile Plot for 2015 UK Town Data

($r \approx -0.51$), and towns with larger student populations tend to have lower homeownership ($r \approx -0.58$). Meanwhile, the manufacturing employment share is positively associated with socioeconomic disadvantage, showing moderate-to-strong positive correlations with health vulnerability and lack of qualifications. Net commuting seems to show relatively weak correlations with other variables, suggesting that commuting patterns are more structurally independent. Overall, the observed correlations reinforce the existence of two broad dimensions: one capturing a wealth–vulnerability gradient, and another reflecting housing tenure and transience.

With these correlations in mind, factor analysis on the variables above should identify latent factors. But before proceeding, I should disclose that the Kaiser-Meyer-Olkin (KMO) measure for this data is 0.711, which implies that factor analysis is likely appropriate though not ideal with my data. I run principal components to find the number of latent factors and the scree plot in Figure 4 shows that the point of diminishing returns (or the elbow) happens at 3 principal components, so we pick the number above that elbow and are guided that there are likely 2 latent factors.

I compare three different orthogonal models (all with the varimax rotation) to see which best minimizes residual error. First, the examination of the residual correlation matrix

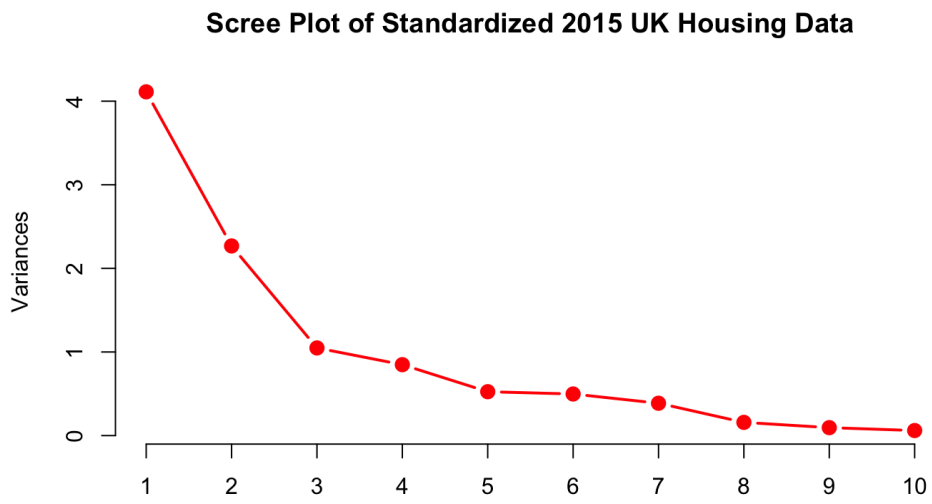


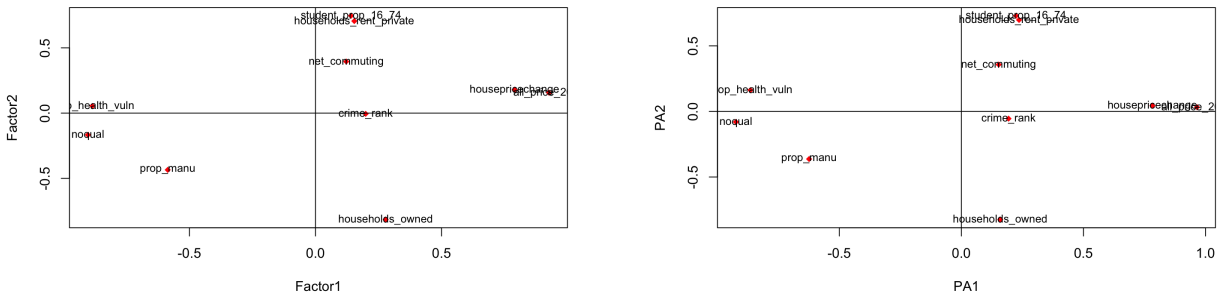
Figure 4: Scree Plot of Standardized 2015 UK Housing Data

indicated that a significant portion of off-diagonal residuals were small for most models, suggesting a reasonable fit of these factor models to the data. By comparing the root mean squared and the number of residual correlations greater than 0.05, each criterion suggests a different extraction method. The former is least with iterative PCA (0.05723) while the latter suggests to use maximum likelihood estimation (36%). So I discuss both of these methods. Table 2 shows their factor loadings which are visualized in loading plots in Figure 5. Both of these extraction methods seem to find two distinct factors based on their factor loadings. The first factor highlights the socioeconomic gradient since it's characterized by strong positive loadings on variables such as *all_price_2015*, *housepricechange*, and *crime_rank*, and strong negative loadings on indicators of marginalization including *noqual*, *prop_health_vuln*, and *prop_manu*. This factor delineates affluent towns from more disadvantaged areas as the former faces high housing costs, low deprivation, and more service-based economies whereas the latter has lower human capital, poorer health outcomes (on average), and a greater reliance on manufacturing industries (which are traditionally less profitable than services) (Yang and Pan, 2020). The second factor reflects the transiency of housing with its high positive loadings on *student_prop_16_74* and *households_rent_private*, and a strong negative loading on *households_owned*. This pattern likely captures towns with more transient or rental-heavy populations, such as college towns or urban centers with high concentrations of students and young professionals, as opposed to more stable, owner-occupied communities.

So these two factors imply that towns are distinguished by how affluent its residents are and for how long they tend to stick around. The dichotomy from the first factor captures the

Variable	PA1 (Iterative PCA)	PA2 (Iterative PCA)	Factor 1 (MLE)	Factor 2 (MLE)
all_price_2015	0.97	0.03	0.93	0.16
housepricechange	0.78	0.04	0.79	0.18
households_owned	0.16	-0.82	0.28	-0.82
households_rent_private	0.24	0.70	0.15	0.71
crime_rank	0.19	-0.05	0.20	
prop_health_vuln	-0.86	0.16	-0.89	
noqual	-0.92	-0.08	-0.90	-0.17
student_prop_16_74	0.22	0.73	0.14	0.75
net_commuting	0.15	0.36	0.12	0.40
prop_manu	-0.62	-0.36	-0.59	-0.44

Table 2: Comparison of factor loadings from Iterative PCA and Maximum Likelihood Estimation (MLE)



(a) Loading Plot for MLE

(b) Loading Plot for Iterative PCA

Figure 5: Two Loading Plots

long-term consequences of economic restructuring and deindustrialization as areas transition into strong knowledge-based economies / financial hubs that diverge from post-industrial areas that face persistent deprivation and health challenges. Meanwhile, the second dimension of transiency arises from differences in towns' urban functions and demographic dynamics. University towns and large cities tend to have more transient, rental-heavy populations linked to education and early-career migration, whereas mid-sized commuter towns and more residential areas exhibit greater housing stability through higher rates of homeownership.

4 Hierarchical Clustering

To explore the grouping of towns in a robust manner, I apply hierarchical clustering under four distance-agglomeration combinations to prevent yielding findings that were spurious or a result of the choice of distance/agglomeration method. Regarding distance measures, I use Euclidean distance since most of my variables are continuous and that it gives equal attention/weight to every standardized variable. Then I compare it with maximum distance that highlights the greatest distance between towns and magnifies sharp socioeconomic differences. Meanwhile for agglomeration methods, complete linkage's resilience against outliers and the fact that it gets the maximum distance between clusters which allows outliers to be singled out which is to be expected among towns with some super star cities (I expect London to stand out). Contrastingly, Ward's method fuses clusters by minimizing the total within-cluster variance, and thus results in compact and similar groupings which are ideal for the profile of towns I want to identify. Based on diagnostic criteria (root-mean-square standard deviation, R-squared, semi-partial R-squared, and cluster distance plots which is consistent for most clusters), the analysis suggests a six-cluster solution, I present the dendrograms for these four different clusterings in Figure 6.

Across all four dendrograms, the same broad pattern emerges. First, London consistently severs from the rest of the tree, standing alone as an outlier—its exceptional house-price level, rapid appreciation, heavy in-commuting, and distinctive rental market sharply distinguish it from every other town. Beneath this top-level split, five robust clusters appear across all specifications:

1. A high-affluence university cluster (Cambridge, Oxford, Bath, Brighton & Hove, Guildford, and under Ward's rule, St Albans), combining very high prices and price growth with large student shares, extensive private renting, and minimal manufacturing employment.

2. A prosperous commuter-belt group (e.g., Reading, Woking, Bracknell, Crawley, Bas-

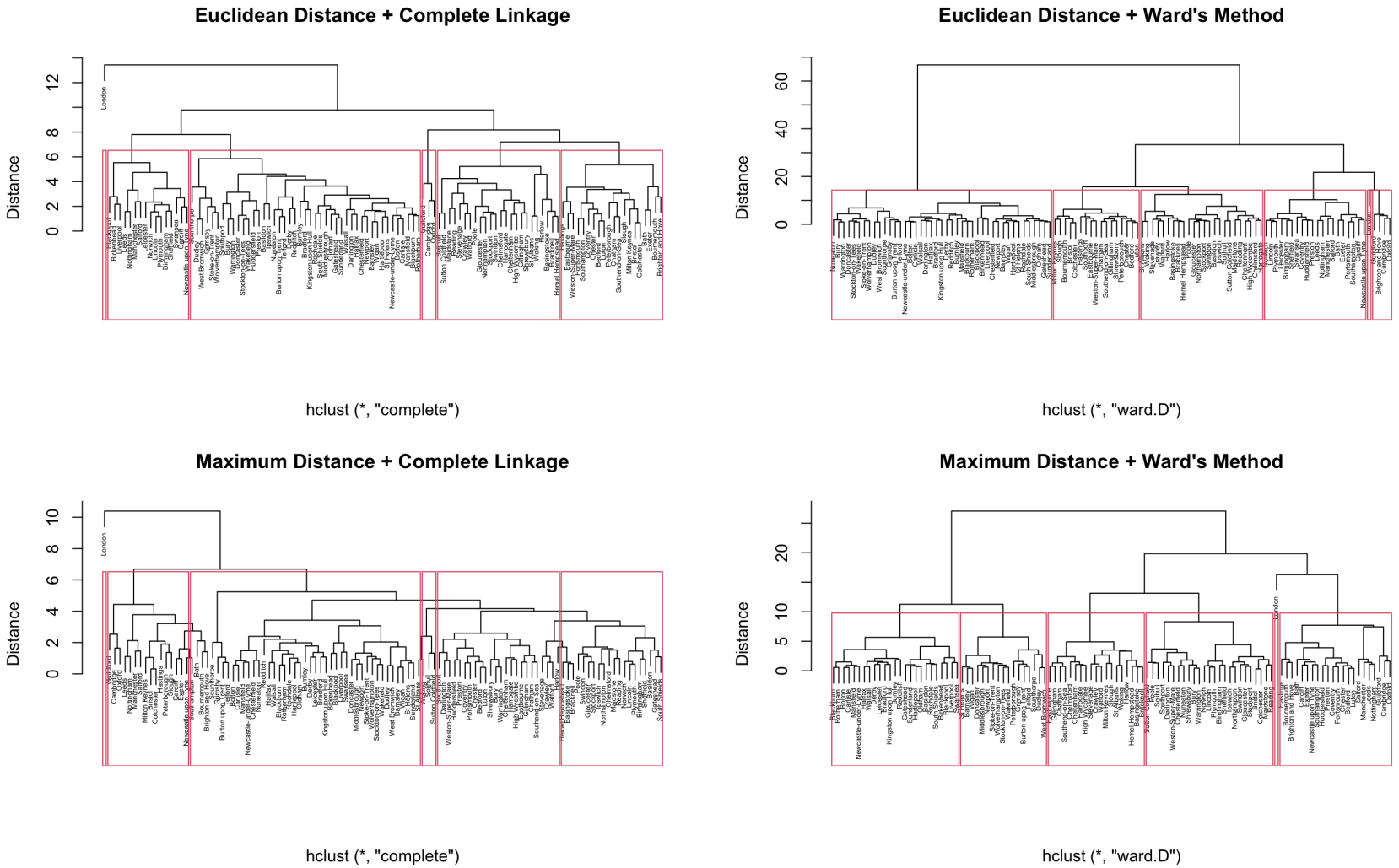


Figure 6: Dendrograms under Different Specifications

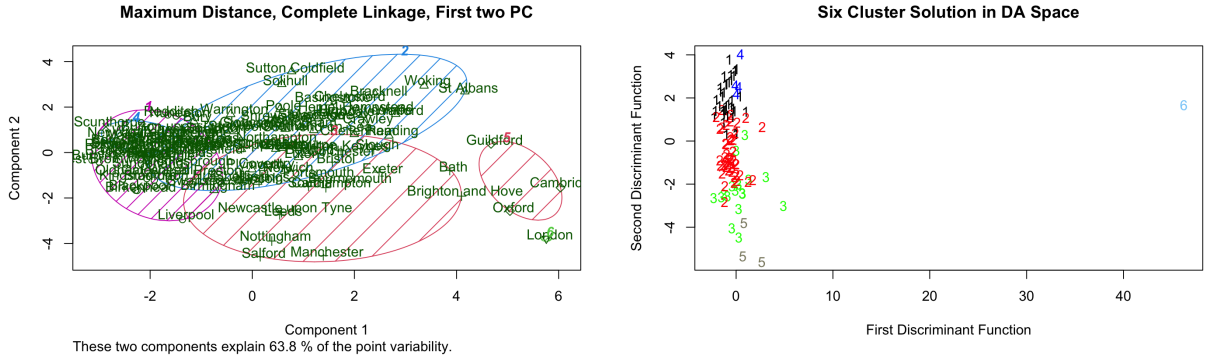
ingstoke), characterized by elevated but not extreme house prices, high owner-occupation, and strong net in-commuting.

3. Major regional service hubs (Manchester, Birmingham, Liverpool, Leeds, Newcastle, Sheffield, and Bristol) forming a cluster with mid-range house prices, sizable student populations, and mixed deprivation indicators.

4. A post-industrial branch (Blackpool, Burnley, Stoke-on-Trent, Grimsby, Sunderland, and similar towns) combining low house prices and weak price growth with high health vulnerability, low qualifications, and a heavy reliance on manufacturing.

5. A mid-tier stable group (e.g., Ipswich, Colchester, Swindon), generally lying close to the sample median across most indicators.

To better visualize these groupings, I project the maximum-distance complete-linkage clustering into principal component space (Figure 7a) and discriminant analysis space (Figure 7b). The consistency of the same five substantive clusters—elite knowledge centers, commuter belts, regional service hubs, post-industrial towns, and average mid-sized settlements—across all distance–linkage pairings suggests that this typology reflects genuine structural differences rather than artifacts of a particular clustering method.



(a) Principal-component space

(b) Discriminant-analysis space

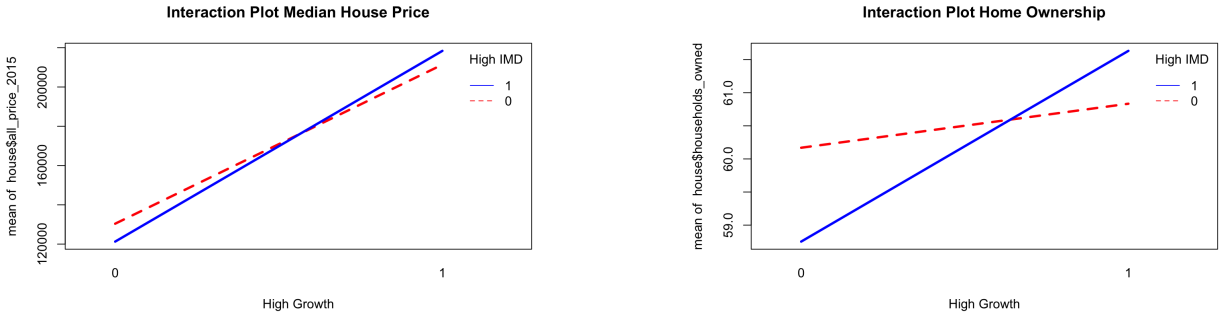
Figure 7: Six–Cluster Solution in 2D Projections

5 Multivariate Generalized Linear Models

I use a multivariate GLM to investigate whether socioeconomic conditions (commuting patterns, health vulnerability, manufacturing employment, and geographic region) predicts significant multivariate differences across towns' median house prices and rates of home ownership which serve as the analysis' response variables. I include categorical predictors to

model spatial variation (*region*), commuting patterns (*high_commute*), and the interaction between income deprivation and price growth (*high_icd* × *high_growth*). On top of that, I incorporate covariates that capture broader demographic characteristics (*prop_health_vuln*, *noqual*, and *prop_manu*). In doing so, this approach would allow for a more comprehensive understanding of how structural and spatial socioeconomic factors jointly shape housing outcomes, highlighting the complexity of inequalities across different local contexts.

Before proceeding, I first examine if there is evidence to suspect an interaction between high income deprivation and price growth on the response variables. I check for this interaction since I believe that unaffordability of housing could possibly be both a demand and supply side issue where people are receiving lower incomes while houses are also growing at a faster rate. The interaction plots in Figures 8a and 8b show that they seem to intersect which strongly suggests some interaction between these variables, justifying my approach to include their interaction in the model.



(a) Median House Price

(b) Proportion of Homeowners

Figure 8: Interaction Plots on Median House Price and Homeownership

Table 3 summarizes the multivariate significance tests based on Pillai's trace, Wilks Lambda, Hotelling-Lawley Trace, and Roy's Largest Root. The multivariate GLM results show that commuting patterns, regional location, health vulnerability, and proportion of no qualification residents are significant multivariate predictors of housing outcomes. The proportion of manufacturing is significant only at the 10% significance level but is worth pointing out nonetheless. Meanwhile, deprivation alone (*high_imd*), house price growth (*high_growth*), and their interaction (*high_icd:high_growth*) are not significant predictors in the multivariate model.

The multivariate regression coefficients for statistically significant predictors, presented in Table 4, suggest that towns with higher commuting inflows tend to have lower median house prices and a homeownership rate approximately 6 percentage points lower than other towns. This pattern is intuitive: towns with high commuting flows often have more transient

Table 3: Type III Multivariate GLM Results

Predictor	**Approx. F**	**p-value**
high_imd	0.106	0.8993
high_growth	1.360	0.2617
high_commuting	11.567	<0.001***
region	2.898	<0.001***
prop_health_vuln	6.821	0.0017**
noqual	11.740	<0.001***
prop_manu	2.741	0.0698*
high_imd:high_growth	0.499	0.6089

populations and a higher proportion of renters, while lower local demand can also contribute to suppressed house prices. Regional differences are also pronounced. The regional comparisons are noticeable where London stands out with average house prices approximately £132,380 higher than those in the base region of Yorkshire and the Humber. In contrast, Wales has significantly cheaper homes, with prices £32,557 lower than Yorkshire and an 8 percentage point higher homeownership rate compared to London, where homeownership is nearly 9 percentage points lower.

Turning to demographic characteristics, towns with higher proportions of health-vulnerable populations tend to have lower house prices, with each unit increase associated with a reduction of approximately £1,000. This could reflect broader issues in these towns, such as poorer access to healthcare services or broader socioeconomic disadvantage, factors that may make them less attractive to potential homeowners. Similarly, towns with higher proportions of residents without qualifications see significantly lower median house prices (by about £7,100), which may signal a weaker local labor market, fewer productive firms, and less economic dynamism overall. Finally, a higher share of manufacturing employment is also associated with modest declines in house prices, while its relationship with homeownership appears mixed and relatively weak.

To ensure the validity of the multivariate results, I first check whether the assumption of multivariate normal residuals is satisfied using a chi-square quantile plot. Figure 9 shows that the residuals are approximately multivariate normal, supporting the reliability of the findings. Additionally, I inspect the univariate Type III ANOVA results in Table 5. These results largely align with the multivariate conclusions: commuting patterns, regional location, and health vulnerability emerge as significant predictors, particularly for the proportion of homeowners. Educational attainment (no qualifications) is strongly associated with median house prices but not with homeownership rates. In contrast, deprivation, house price growth, and their interaction remain non-significant across both housing outcomes.

Table 4: Multivariate GLM Coefficients for Housing Outcomes

Predictor	**all_price_2015 (Estimate)**	**households_owned (Estimate)**
high_commuting	-4,209.29	-5.92
base region: Yorkshire and The Humber		
region1: East Midlands	-32,565.80	0.39
region2: East of England	14,727.34	-5.72
region3: London	132,380.44	-8.78
region4: North East	-20,534.30	2.75
region5: North West	-29,592.55	6.10
region6: South East	5,021.10	-4.30
region7: South West	-16,259.11	-1.14
region8: Wales	-32,556.96	8.10
region9: West Midlands	-1,700.88	1.84
prop_health_vuln	-1,000.25	-3.47
noqual	-7,100.13	0.29
prop_manu	-1,712.56	0.45

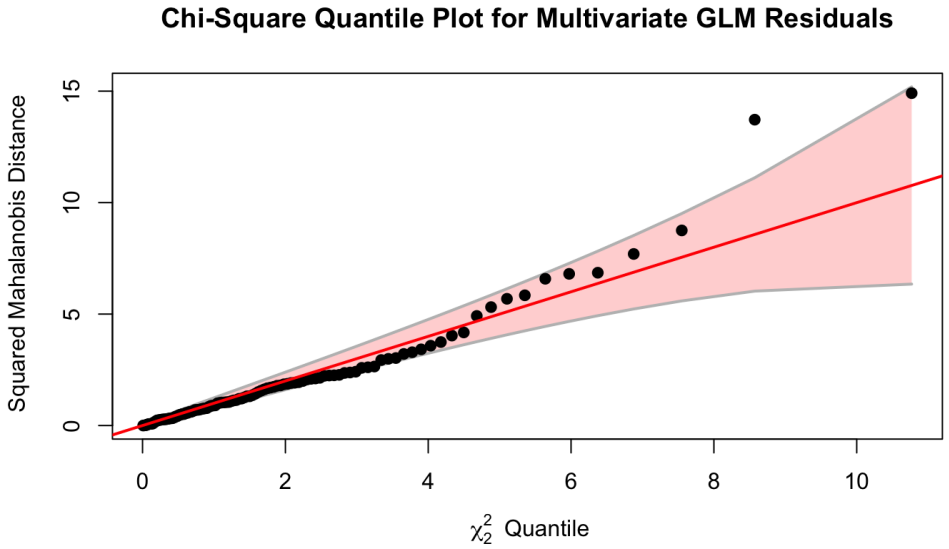


Figure 9: Chi-Square Quantile Plot for Multivariate GLM Residuals

The multivariate GLM assesses the joint predictive power of the socioeconomic explanatory variables on both housing outcomes simultaneously. To further explore these effects individually, I turn to the univariate ANOVA results. Table 5 presents these findings. For homeownership, the results show that commuting patterns, regional location, and health vulnerability remain significant predictors, consistent with the multivariate results. However, the proportion of residents without qualifications, which was significant in the multivariate model, appears to have no meaningful impact on homeownership rates, as indicated by its very large p-value. This suggests that while education levels matter for broader housing patterns, they do not seem to directly influence the decision or ability to own a home by itself.

In contrast, the determinants of median house prices differ. Here, region and the proportion of residents without qualifications emerge as the primary significant factors. Regional differences align with broader geographic disparities in the housing market, while lower educational attainment may reflect weaker local economies that struggle to sustain higher property values. Other variables such as commuting patterns and health vulnerability, though important for ownership, appear less critical in explaining variations in median house prices when considered independently.

Table 5: Follow-up Type III ANOVAs by Dependent Variable

Predictor	*all_price_2015*		*households_owned*	
	F	**p-value**	**F**	**p-value**
high_imd	0.05	0.8225	0.16	0.6902
high_growth	0.20	0.9931	0.97	0.3265
high_commuting	0.49	0.4837	23.04	<0.001***
region	37.45	<0.001***	19.97	<0.001***
prop_health_vuln	0.05	0.8261	13.77	0.00035***
noqual	22.64	<0.001***	0.10	0.9996
prop_manu	2.05	0.1561	3.37	0.0697*
high_imd:high_growth	0.23	0.6358	0.76	0.3849

However the regional disparities shown in the multivariate model is especially striking, so I use a multivariate contrast to compare London to the rest of the regions since it really stands out and see how everywhere else stands relative to it. A formal multivariate contrast (Table 6) comparing London against all other towns reveals a highly significant difference in housing outcomes, based on Pillai's trace ($F(2, 99) = 11.27$, $p < 0.001$). London exhibits markedly higher median house prices and lower homeownership rates compared to the rest of the United Kingdom, reinforcing its distinct role in shaping spatial inequality patterns.

Table 6: Multivariate Contrast: London vs Other Towns

Test Statistic	**Value**	**Approx. F (2, 99)**	**p-value**
Pillai's Trace	0.1855	11.27	<0.001***
Wilks' Lambda	0.8145	11.27	<0.001***
Hotelling-Lawley Trace	0.2277	11.27	<0.001***
Roy's Largest Root	0.2277	11.27	<0.001***

6 Conclusion

Taken together, the three strands of analysis paint a coherent picture of how and why spatial inequality persists across the United Kingdom.

First, the factor analysis reduced the variation of ten place-based indicators into two clear, orthogonal dimensions/factors with the affluence-vulnerability gradient that captures the long-run legacy of economic restructuring. Prosperous towns combine high house prices, rapid capital gains, and well-educated, healthy populations while post-industrial areas show almost the opposite. The second factor which captures the transiency of housing situations distinguishes stable, homeowning communities from rental-heavy, student-oriented, or highly mobile labor markets. These latent factors hint at a map divided not just by wealth but the nature of people's day to day lives.

Second, hierarchical clustering was used to group towns using the same variables and ended up with six distinct town profiles using different distance-agglomeration specifications:

1. London (Unique enough to be its own category)
2. Elite university cities Knowledge industries and intense rental demand)
3. Affluent commuter belts (Great access to metropolitan jobs at the cost of housing stability)
4. Major regional service hubs (Large cities that balance growth with some pockets of deprivation)
5. Post-industrial towns (Low skills, poor health, and sluggish housing markets)
6. Mid-tier stable settlements (Variables hover around the national median)

These distinct profiles somewhat confirm that spatial inequality is not random noise, but a patterned hierarchy reproduced across England and Wales.

Finally, the multivariate GLM showed that these patterns matter for concrete housing outcomes. Commuting intensity, region, health vulnerability, and human capital jointly explain significant variation in both median home prices and ownership rates. However, the expected interaction between income deprivation and price growth since 2004 were insignificant alongside these other variables. After controlling for local demographics, London's house prices remain more that £130,000 more expensive (with lower ownership rates) than the national baseline (Yorkshire and the Humber). In sum, mobility and place intersect: towns that export workers, import students, or inherit poor health adn skills see this encoded directly in their housing markets.

Implications The results from this paper suggest that policies that do not take into consideration the socioeconomic fabric of the town will fail to meet its goals. Bridging the UK's regional disparities requires interventions that takes into consideration the profile of the town, their history and their potential.

- Post-industrial cities require investment in the human capital of residents for them to transition into more lucrative industries that may attract non-manufacturing firms and hopefully revitalize the area.
- University cities and commuter belts need policies that increases the housing supply and tame the high-rent dynamics that disproportionately harm poorer households.
- Regional hubs need policies that promote growth while protecting their comparatively affordable housing situation.
- London needs metropolitan-scale solutions that recognize its national gravitational pull in terms of goods and people.

Points for Further Analysis Future studies can continue studying these spatial dynamics by improving the kind of data being analyzed. By updating the analysis using more recent data like the 2021 census and the 2024 land registry, one may capture the resilience/persistence of these town profiles amidst Brexit and COVID-19. Another point to consider is to go further down in granularity by using LSOA as the geographic unit to see whether average towns mask sharp neighborhood disparities (pockets of gentrification or deprivation). Also, it would be good to include variables that capture the local amenities and infrastructure of the towns that may contribute to one's mobility or productivity, providing greater depth to the affluence-vulnerability gradient.

References

Piketty, T. (2014). *Capital in the twenty-first century.* Harvard University Press, London, England.

Prothero, R. (2016). Towns and cities analysis. Dataset. Statistics on towns and cities in England and Wales with a focus on housing and deprivation.

Yang, Z. and Pan, Y. (2020). Human capital, housing prices, and regional economic development: Will "vying for talent" through policy succeed? *Cities*, 98:102577.