A micro-geographic house price index for England and Wales*

Gabriel M. Ahlfeldt Felipe Carozzi Lukas Makovsky

August 28, 2022

Abstract

We generate a mix-adjusted house price index for England and Wales from 2010 to 2020 at the level of *lower-layer super output areas*. To this end, we blend parametric and non-parametric estimation techniques and leverage on a matched *Land Registry-Energy Performance Certificate data set*. The key advantage of our index is that it combines full spatial coverage with high spatial detail.

Key words: Index, real estate, price, property.
JEL: R10

*Ahlfeldt: London School of Economics and Political Sciences (LSE) and CEPR, CESifo, CEP; g.ahlfeldt@lse.ac.uk. Carozzi: London School of Economics and Political Sciences. Makovsky: London School of Economics and Political Sciences. The usual disclaimer applies.

1 Introduction

Following the distribution of amenities and disamenties, house prices can vary significantly significantly over short distances. Since property transactions are relatively rare events, the creation of mix-adjusted house price indices for micro-geographic areas is not straightforward. Standard problem is that in thin markets there will be few observations within small geographic units like output areas.

To generate a micro-geographic house price index for England and Wales, we apply the algorithmic approach developed by Ahlfeldt et al. (2021). This approach uses spatial methods to overcome the limitations of sparse property data. The input is a conventional data set containing pooled cross sections of real estate transactions with information on prices or rents along with geographic coordinates, transaction dates, and property characteristics. The output is a balanced panel data set of mix-adjusted purchase or rental prices for arbitrary spatial units. The algorithm automatically adjusts to spatially varying densities of observations using a combination of parametric and non-parametric estimation techniques. It treats the computation of the indices for *any* spatial unit as a separate problem that is addressed in a *separate* iteration of of a *locally weighted regression* (Cleveland and Devlin, 1988).

In each iteration, the algorithm considers the density of observations in the vicinity of the targeted location and flexibly defines the size of a spatial window that provides a sufficient amount of observations. Inside this spatial window, observed prices are adjusted for structural and location characteristics using the conventional hedonic regression technique (Rosen, 1974). To predict the price and rent indices right at the target location, the method controls for a first-order polynomial of distance from the center. In addition, there is a spatial fixed effect, whose diameter also depends on the density of observations. The strength of the algorithmic approach is that it loads the predictive power on non-parametric components where many observations are available, such as in high-density urban neighborhoods, whereas the predictive model becomes more parametric if observations are sparse, e.g. in rural regions.

We apply this data set to a matched *Land Registry-Energy Performance Certificate data set*. The advantage of this data set over the readily accessible *Land Registry Price Paid* data set is that we can adjust property prices for various observable property price characteristics such as, floors space, type of the property (detached house, semidetached house, terrace house or apartment), energy consumption of the property, type of load bearing walls, whether the property is freehold (or leasehold), new, or equipped with a fireplace.

The result is a balanced-panel mix-adjusted house price index at the lower-layer super output area (LSOA) level that combines comprehensive coverage with high spatial detail. It is straightforward to aggregate this micro-geographic price index to higher spatial units. Since output areas vary significantly in geographic area to reduce the variation in the population living within, even an unweighted average across output areas within larger spatial units will deliver a decent approximation of the population-weighted average.

Several sources of local level housing price estimates are available for the United Kingdom at relatively fine levels of spatial disaggregation. We see the index provided here as dominating the existing sources in either the level of spatial detail, the methodology used in its construction or both.

Perhaps the most established of these existing sources of spatial price data is the hedonic price index released by the Land Registry (LR). One important limitation of this index is that it is only available at relatively coarse levels of aggregation, with the finest granularity being at the level of local authorities. Given that there are a total of 355 local authorities and roughly 35,000 LSOAs in England and Wales, our price index is two orders of magnitude more granular than the one provided by this source. A second important difference between our index and the one released by LR relates to the variables used to conduct the hedonic adjustment for property characteristics. In our index, we use a richer set of property characteristics including—crucially—the floor area of units. This means that our index is more successful at adjusting for differences between units even if working at comparable levels of aggregation.

More disaggregated data on prices is released yearly by the Office of National Statistics as the House Price Statistics for Small Areas in England and Wales (HPSSA). This dataset provides median prices and transaction volumes at the level of Middle Layer Super Output Areas. These spatial units are coarser than the LSOAs in our index release, but still reasonably fine relative to most other available products. One important shortcoming of the (HPSSA) is that it simply reports median prices. While this allows for a methodology that can easily be applied consistently over time, it also means that observed changes in prices across geographies can result from changes in the types of units sold in this period, an issue that we deal with our hedonic adjustment.

A series of independent efforts have produced finely disaggregated price maps that have some characteristics in common with our own index. Anna Powell-Smith, from the Centre for Public Data, produced a of average prices per square metre for map of 2,280 postcode districts England and Wales in 2007 based on the same combination of LR and EPC data we use here. Our advantage is that we provide a

dataset covering a decade of price developments at a finer level of spatial disaggregation that also ensures that prices are calculated using a sufficient number of sales by virtue of our estimation algorithm.

Finally, online real estate platforms such as Zoopla and Rightmove, as well as lenders such as Nationwide and Halifax also release spatially disaggregated indexes. In most cases, however, there are released at the regional level only which severely limits their use for spatial analysis of local cross-sectional and longitudinal trends. Zoopla is perhaps an exception, with postcode district level estimates and property price heatmaps available on the website. The disadvantage of this source is that neither the method nor the data sources used to construct its indices are disclosed for evaluation.

The rest of the documentation is organized as follows. Section 2 introduces our algorithm. Section 3 discussed our parametrization and data. Section 4 present some descriptive statistics.

2 Algorithm

The below description is borrowed from Ahlfeldt et al. (2021) from which we adopt the methodology. We use the method to create a mix-adjusted property price index for an arbitrary set of *target* spatial units indexed by $j \in J$. For each j, we run a locally weighted regression (LWR) of the following type:

$$\begin{aligned}\ln \mathcal{P}_{i,t} = \; & a_t^j + \bar{S}_i b^j + \sum_z d_z^j (D_i^j \times I(z=t)) + e^j I(D_i^j > T^j)_i \\ & + f^j (X_i - X^j) + g^j (Y_i - Y^j) + \epsilon_{i,t}^j,\end{aligned}$$

where $\mathcal{P}_{i,t}$ is the purchase or rental price of a property i transacted in year t. $\bar{S}_i$ is a vector of covariates stripped off the national average (we subtract the national mean from the observed value of S_i), and b^j are the LWR-j-specific hedonic implicit prices. D_i^j is the distance from a transacted property i to the target unit j with d_z^j being the LWR j-specific gradient in year z. $I(.)$ is an indicator function that returns a value of one if a condition is true and zero otherwise and T^j is a threshold distance. Hence, $e^j I(D_i^j > T^j)_i$ is a fixed effect for all transacted properties i that are outside the vicinity of the catchment area. X_i and Y_i are the coordinates of transacted properties, X^j and Y^j are the coordinates of the target unit, and f^j and g^j are spatial gradients. $\epsilon_{i,t}^j$ is the residual term.

The threshold T^j is chosen using the following rule:

$$T^j = \begin{cases} T^1, & \text{if } N^{(D_i^j \leq T^1)} \geq N^T \\ T^2, & \text{if } N^{(D_i^j \leq T^1)} < N^T \leq N^{(D_i^j \leq T^2)} \\ T^3, & \text{if } N^{(D_i^j \leq T^2)} < N^T \leq N^{(D_i^j \leq T^3)} \\ T^4, & \text{if } N^{(D_i^j \leq T^3)} < N^T, \end{cases}$$

where $N^{(D_i^j \leq T^{s \in \{1,2,3,4\}})}$ gives the number of transacted units from a target unit within distance threshold $T^{s \in \{1,2,3,4\}}$ and N^T is a minimum-number-of-transactions threshold, all to be chosen by the user in the program implementation of this algorithm.

In each LWR j, all transacted properties i are weighted using the following kernel weight:

$$W_i^j = \frac{w_i^j}{\sum_i w_i^j}$$

$$w_i^j = \begin{cases} I(D_i^j \leq A^1), & \text{if } N^{(D_i^j \leq A^1)} \geq N^A \\ I(D_i^j \leq A^2), & \text{if } N^{(D_i^j \leq A^1)} < N^A \leq N^{(D_i^j \leq A^2)} \\ I(D_i^j \leq A^3), & \text{if } N^{(D_i^j \leq A^2)} < N^A \leq N^{(D_i^j \leq A^3)} \\ I(D_i^j \leq A^4), & \text{if } N^{(D_i^j \leq A^3)} < N^A, \end{cases}$$

where $\{A^1, A^2, A^3, A^4\}$ are distance thresholds and N^A is a minimum-number-of-transactions threshold, all to be defined by the user in the program implementation of this algorithm.

The price index for a target unit is then simply defined as:

$$\hat{\mathcal{P}}_t^j = \exp(\hat{\alpha}_t^j),$$

which we recover from the LWR-j-specific estimates of time-fixed effects α_t^j. To facilitate the computation of confidence bands, we also report standard errors

$$\hat{\sigma}_{\mathcal{P}_t^j} = \exp(\hat{\sigma}_{\alpha_t^j}) \times \hat{\mathcal{P}}_t^j,$$

where ($\hat{\sigma}_{\alpha_t^j}$ are estimated allowing for clustering within the areas inside and outside the spatial fixed effect ($I(D_i^j > T^j)_i$). Intuitively, the price index for a target unit is a year-specific local conditional mean that is adjusted for property characteristics (deviations from the national average), location (time-varying distance from j effects, and time-invariant spatial trends in X and Y coordinates), and a spatial fixed effect. Since $\{w_i^j, T^j\}$ are endogenously chosen by the algorithm, the precision of the index

automatically increases as the density of observations increases.

Via the parameters $\{A^1, A^2, A^3, A^4, N^A, T^1, T^2, T^3, T^4, N^T\}$, the user has flexible control over the *bias-variance trade-off*. Smaller values in all parameters will generally lead to greater spatial variation, at the cost of an increasing sensitivity to outliers in the underlying micro-data. In choosing N^A, it is worth recalling that N^A describes the number of observations that occur over multiple years, but estimates of conditional means and distance gradients are year-specific. Thus, as a rule of thumb, N^A should increase proportionately to the number of years over which an index is predicted.

3 Application

In this section, we describe the data that we input into the algorithmic approach introduced in Section 2 and discuss our parametrization.

3.1 Data

To facilitate our analysis. we match the *Land Registry Price Paid* data to the *Land Registry Energy Performance Certificate* data set at the property level. This way, we observe ,along with price of the property, its floor area, type of the property (detached house, semidetached house, terrace house or apartment), energy consumption of the property, the type of load-bearing walls, and whether the property is freehold (or leasehold), now, or equipped with a fireplace. The data are geo-referenced at the postcode level, which many of the denser areas corresponds to address-level. Our data set covers years from 2010 to 2020 and contains almost 8.9 million observations in England and Wales.

To remove outliers, we drop observations with (i) a property price bellow $250£/m^2$ or above $25000£/m^2$; (ii) a floor area below $30m^2$ or above 500 m^2; (iii) observations with missing attributes. We also remove properties for which the per-m^2 price is bellow 20% or above 500% of the median price within the Local Authority district. This data cleaning procedure removes shrinks the sample by about 2%. As a further input into the property algorithm, we use the the geocraphic centroids of lower-layer super output areas (LSOA) as of year 2011.

3.2 Parametrisation

The spatial unit of our house price index are LLSOAs. These are built from group of contiguous output areas. They are designed to be consistent in terms of population,

with an average of about 1500. To achieve this similarity in terms of population size, LLSOAs vary significantly in geographic size. The mean area of LLSOAs is 4.35 km^2, with a standard deviation of 14.78 km^2. While the 10% of the smallest LLSOAs are smaller than 0.17 km^2, the largest 10% are larger than 10.25 km^2 and the largest 1% exceeds 64.77 km^2. This shows the size distribution of LLSOAs is skewed to the right. In the 2011 LLSOA definition, there are 34,753 units in England and Wales. LLSOAs are somewhat smaller than postcodes in Germany. Hence, we build on the parametrisation Ahlfeldt et al. (2021) recommend for postcodes, but make some adjustments. Specifically, we allow for the following choices for thresholds: $\{A^1 = 2.5, A^2 = 5, A^3 = 25, A^4 = 100, T^1 = 2, T^2 = 4, T^3 = 10, T^4 = 20\}$ (all in km) and we require a minimum of $\{N^A = 10,000, N^T = 1,000\}$ transactions. These choices allow for a tight local fit in areas where the density of transactions is high while ensuring that the LWR are run on a sufficiently large sample in areas that are more sparsely populated.

4 Index

Figure 1 illustrates the spatial distribution of hour house price index for 2020. Evidently, there is striking variation in house prices within and between cities in the UK. Mix-adjusted house prices within the London functional urban area are 2.5 times as high as in the rest of the England and Wales. At the same time, our index

Table 1: Summary statistics (8,876,986 observations)

	Mean	Std.dev.
Price (£/m^2)	2853	1946
Transaction year	2015.1	2.89
Property size (m^2)	94.0	43.4
Energy consumption (kWh/m^2 per annum)	256.6	125.3
Fireplace (0,1)	0.16	0.37
New-built (0,1)	0.10	0.30
Leasehold (0,1)	0.24	0.42
Detached houses (0,1)	0.24	0.43
Semi-detached houses (0,1)	0.28	0.45
Terrace houses (0,1)	0.29	0.46
Apartments (0,1)	0.19	0.39
Solid walls (0,1)	0.21	0.41
Cavity walls (0,1)	0.56	0.50
Unknown walls (0,1)	0.23	0.42

varies from 2,600 £/m^2 to 24,507 £/m^2 within the London functional urban area making the most expensive LLSOA almost 10 times more expensive than the least expensive one.

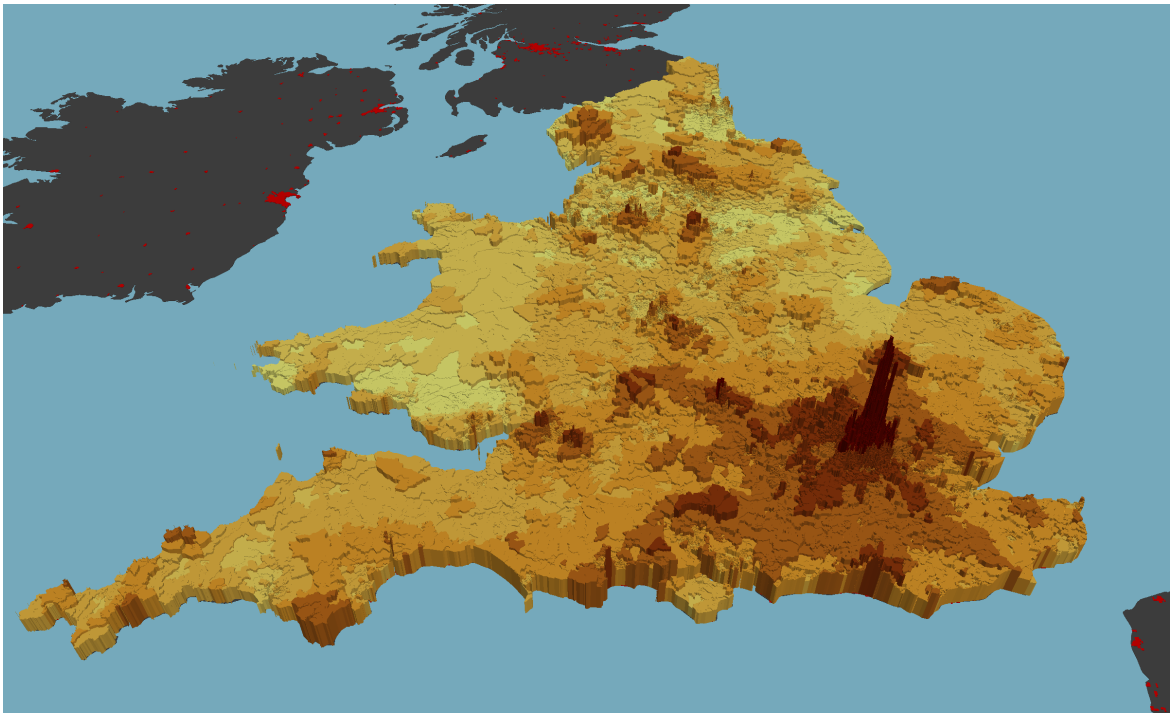
Similarly, there is significant spatial heterogeneity in changes in house prices over the past decade, both between (see Figure (2) and within (see 3) cities. Figure 4 shows that the London housing marked recovered quickly from the financial crisis as prices rose throughout the decade. In contrast, it took until 2014 for prices in Manchester and West Midlands to return to 2010 levels. In Leeds, it took even one year longer. At the same time, steep growth in the first half of the decade in London was followed by price stagnation from 2017 until the end of analyzed time series in 2020. Price growth in the central area of the London functional urban area outpaced growth in the outer area and the difference was highest during the period of fastest growth. In Manchester and West Midlands, house prices underperformed relative to other areas during the first half of the decade, but caught up during the second. While Leeds lags the trend in the other areas, house price growth has been steady since 2013.

In Figures 5 and 6, we correlate initial price levels and growth rates over the decade across LLSOAs in England and Wales and selected travel-to-work areas. There is a strong mean-reversion tendency within the London travel-to-work area, consistent with gentrification of formerly affordable areas. We do not find a similar trend in any other major cities. To the contrary, across LLSOAs in the other parts of England and Wales, levels and trends are positively correlated, pointing to spatial divergence.

5 Conclusion

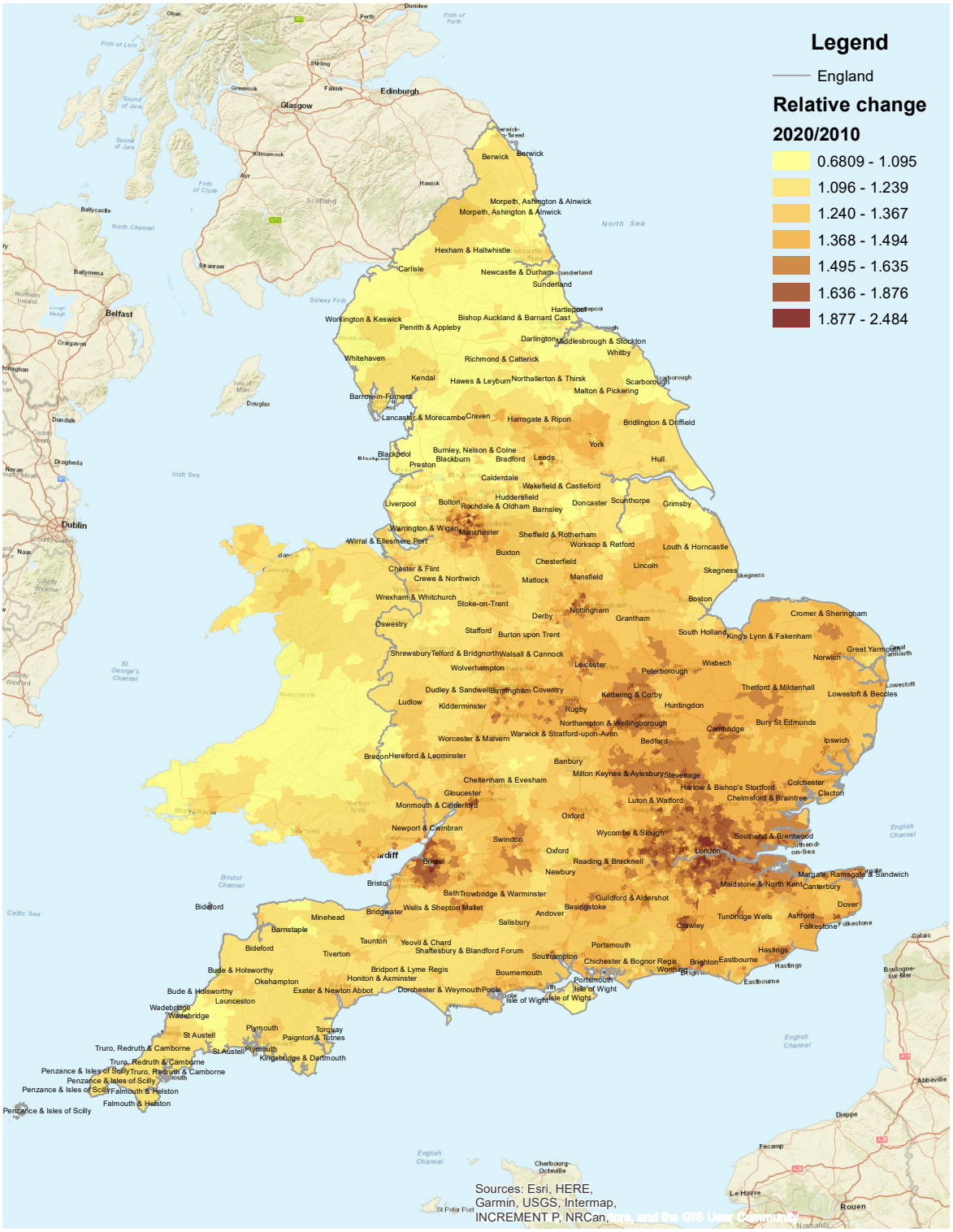
We document the generation of a new mix-adjusted house price index at the level of LLSOAs for England and Wales. The index combines full spatial coverage with great spatial detail. As such it is an ideal input into quantitative spatial models that often require the price of a homogenous housing service for the inversion of structural fundamentals.

Figure 1: 2020 mix-adjusted house prices by output areas



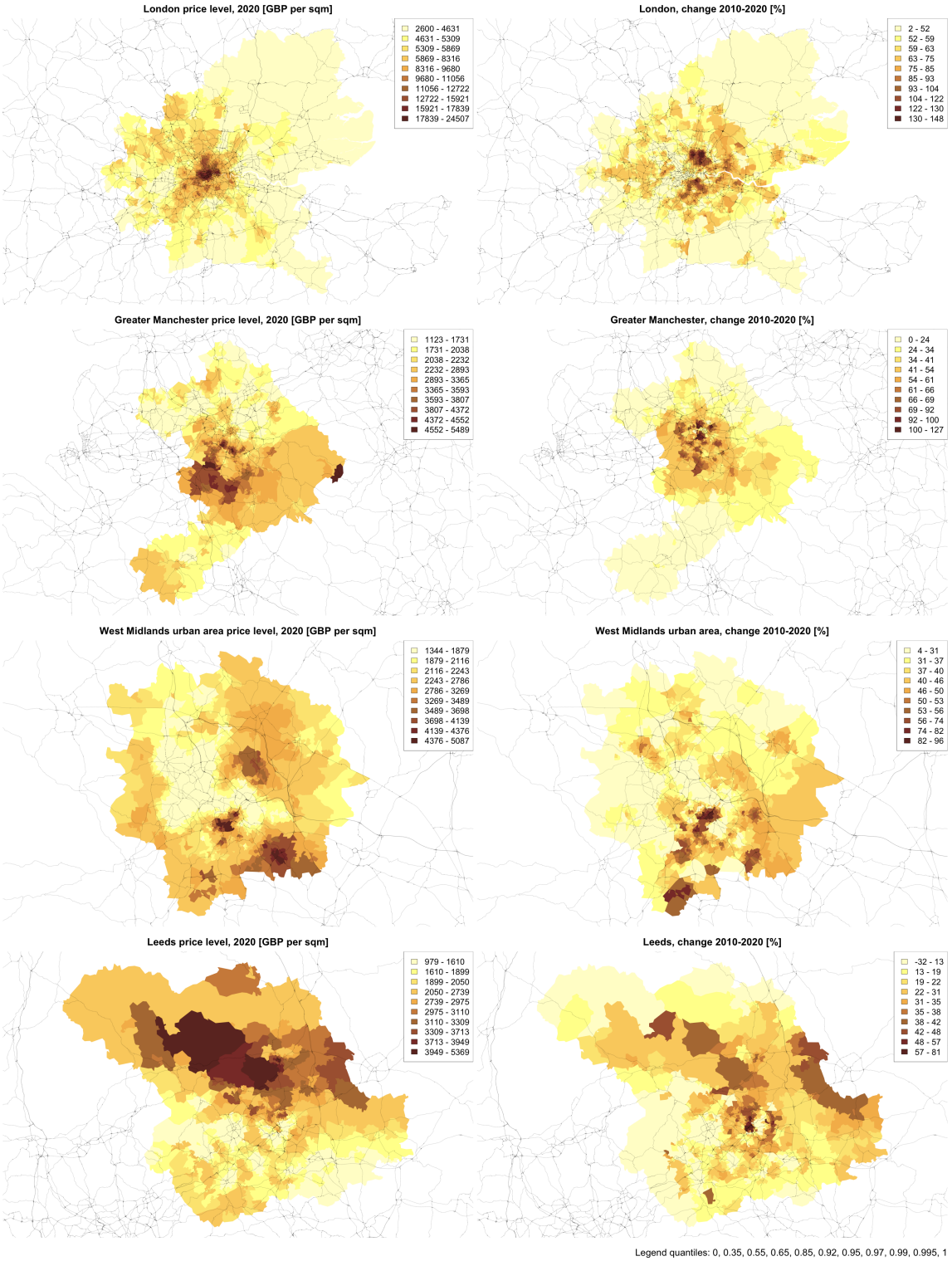
Notes: Bar height is proportionate to mix-adjusted per square meter purchase prices.

Figure 2: 2010-2020 price change



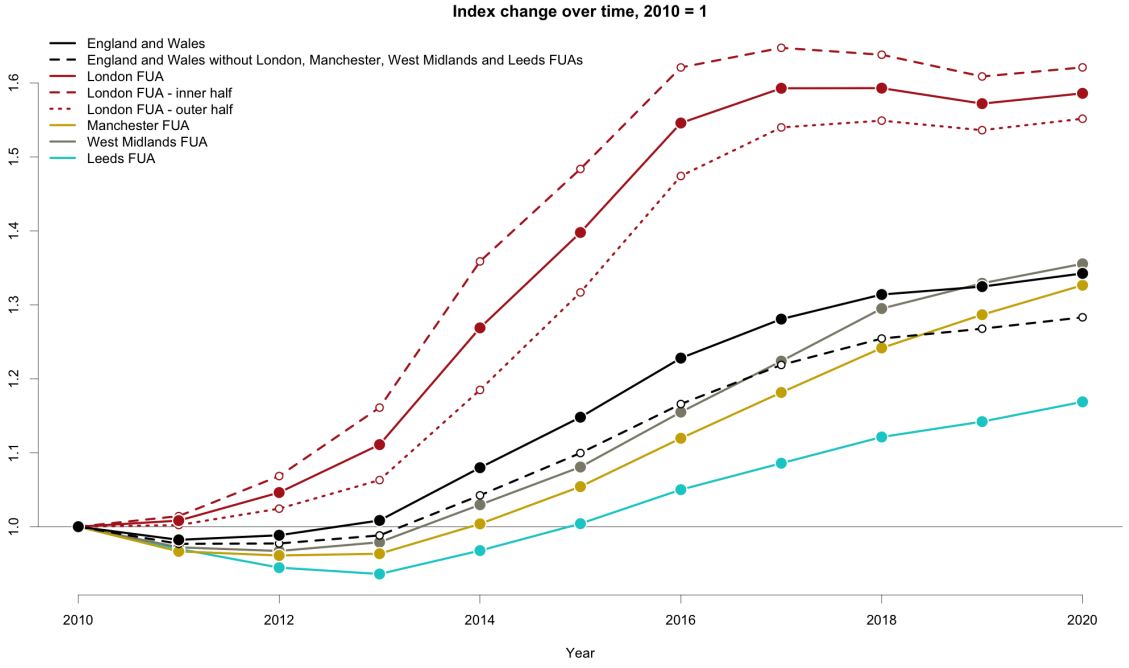
Notes: Unit of observation is lower-layer super output area. Comparison is made based on mix-adjusted per-square meter house prices.

Figure 3: Detail of 4 largest functional urban areas



Notes: Unit of observation is lower-layer super output area. Comparison is made based on mix-adjusted per-square meter house prices. OECD definition of functional urban areas is used (?)

Figure 4: Index development in England and Wales and 4 largest functional urban areas

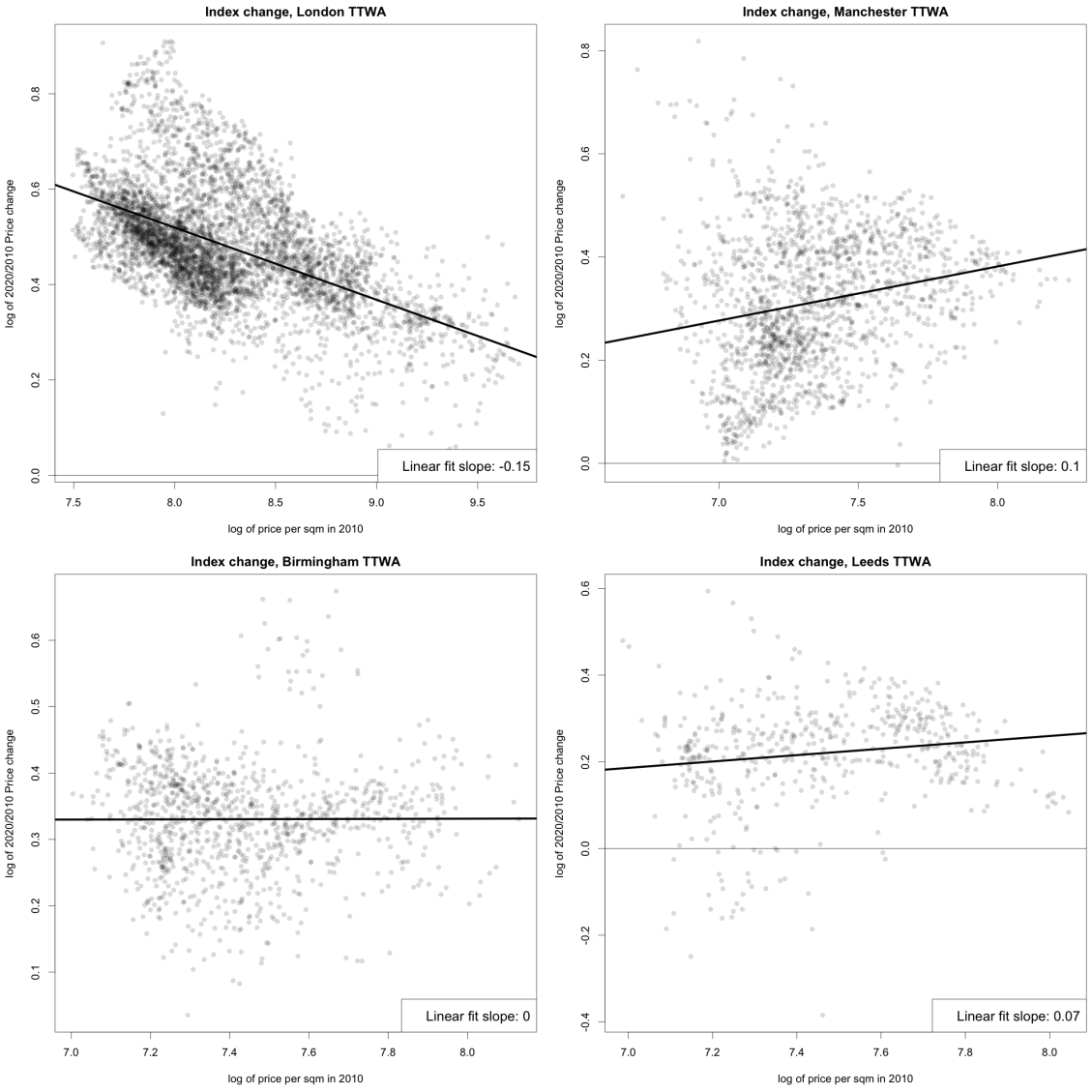


Notes: Region-specific house price trends are based on the year fixed effects (with 2010 being the base year) recovered from region-specific regressions of the log of house prices against location and year effects. Inner half of the London FUA refers to one half of all LSOAs within London FUA located closest to the Holborn area. The outer half are the remaining LSOAs. OECD definition of functional urban areas is used (OECD, 2012)

Figure 5: Initial price and price change

Notes: Each observation represents one LLSOA

Figure 6: Initial price and price change, TTWA of 4 largest cities



Notes: Each observation represents one LLSOA.

References

Ahlfeldt, Gabriel M., Stephan Heblich, and Tobias Seidel, "Micro-geographic property price and rent indices," *CEP Discussion Paper*, 2021, *1782.*

Cleveland, William S and Susan J Devlin, "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 1988, *83* (403), 596–610.

OECD, *Redefining Urban: A New Way to Measure Metropolitan Areas*, Organisation for Economic Cooperation and Development, 2012.

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

ONLINE APPENDIX—not for publication

1 Additional figures and tables