UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

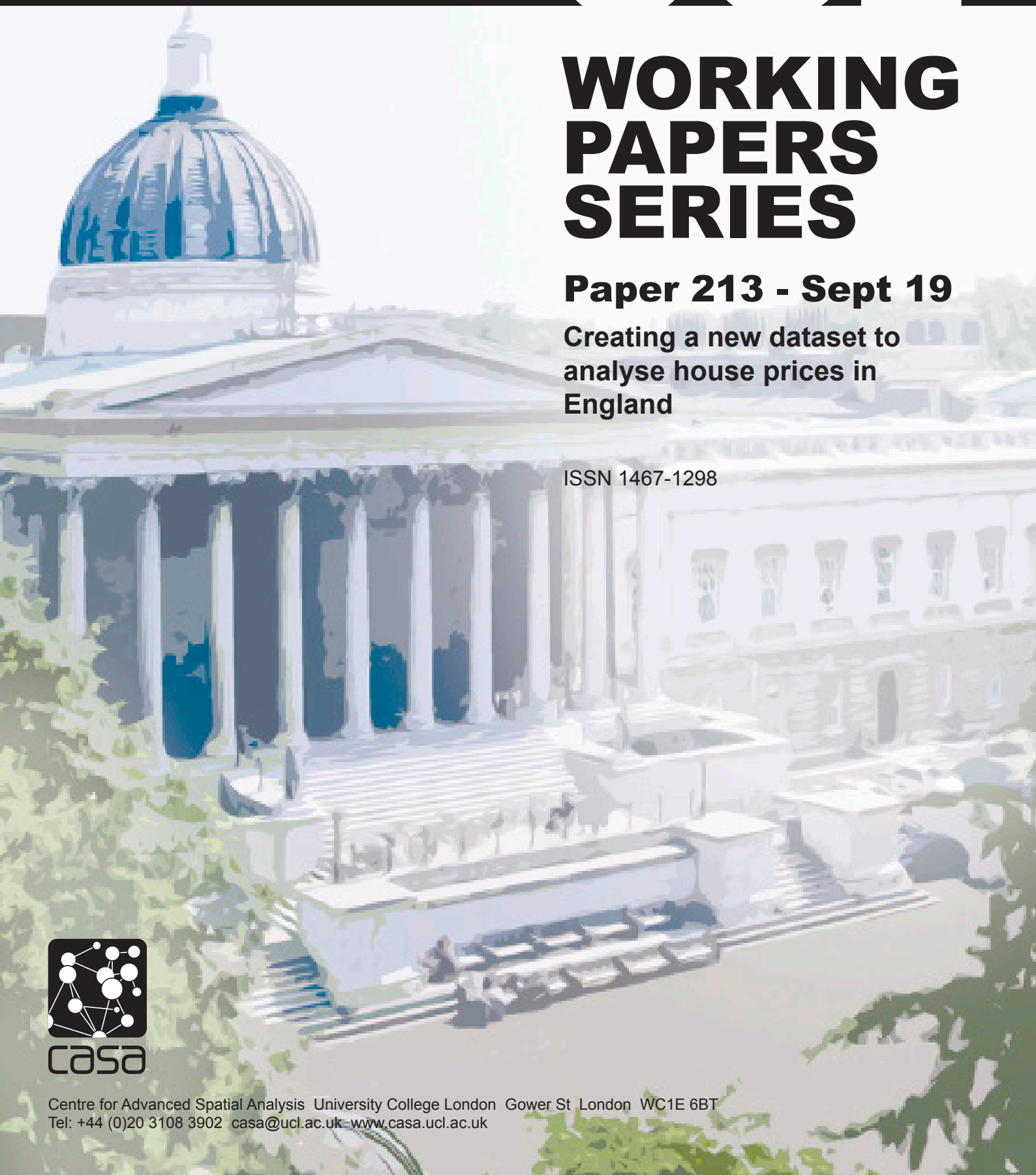
UCL

WORKING PAPERS SERIES

Paper 213 - Sept 19

Creating a new dataset to analyse house prices in England

ISSN 1467-1298



casa

Centre for Advanced Spatial Analysis University College London Gower St London WC1E 6BT
Tel: +44 (0)20 3108 3902 casa@ucl.ac.uk www.casa.ucl.ac.uk

Creating a new dataset to analyse house prices in England

Bin Chi, Adam Dennett, Thomas Oléron-Evans, Robin Morphet

Abstract

House price data deficiencies hinder UK housing market research. House price research in the UK is limited by lack of an open and comprehensive house price database that contains transaction price alongside individual property characteristics. This research outlines one approach which addresses this deficiency in England. Land Registry Price Paid Data (PPD) is the official house price dataset in England covering all residential transactions in the housing market. It has two main disadvantages: first it is not geo-referenced and second, it lacks accurate information on housing size. We create two data linkage methods to overcome these two shortcomings, first by linking the Land Registry PPD with Ordnance Survey (OS) MasterMap and OS AddressBase Plus, second by linking the resulting data with total floor area information from Domestic Energy Performance Certificates (EPCs). This new linked dataset offers greater flexibility for the exploration of house price variation in England over different scales. A strong positive relationship is observed between house price and total floor area. This relationship varies at different geographic scales and over different property types across England.

Keywords: Land Registry Price Paid Data, data linkage, England.

1. Introduction

Housing is a major source of inequality in the UK, particularly in England (Dorling, 2014). In some areas of England, buying a house is becoming prohibitively expensive (Inman, 2017). The current UK government has recognised that the UK housing market is broken and they want to fix it (DCLG, 2017). A more nuanced understanding of residential house prices in England will support better decision making to facilitate this goal. However, data deficiencies are an obstruction to a comprehensive analysis. House price data in England is imperfect (Gibb and Bailey, 2016; Wood, 2015) and this poses significant practical problems in exploring house price variation across England, especially over small geographical areas.

Current house price statistics are normally presented at a macro-geographic scale (i.e. region or local authority), while house prices actually show spatially heterogeneous patterns at smaller geographical scales (ONS, 2016, 2017). It is necessary to explore house price patterns at smaller geographic levels to gain a better understanding of the UK housing market. To support this, the choice of the dataset is regarded as critically important, but there has been little discussion of this in the literature (Gibb and Bailey, 2016; Whitehead et al., 2008; Wood, 2015). Meanwhile, the current official house price dataset (Land Registry PPD) covers all residential transactions in England and Wales since 1995, and includes information on a number of housing characteristics, but does not contain accurate housing size information, such

as total floor area. House price data linked with information on individual property characteristics are difficult to obtain within the UK (Gibbons and Machin, 2003; Orford, 2010), but dwelling size is regarded as one of the most important determinants of house price variation in house price modelling (Orford, 2010). Building a comprehensive housing price database will produce an advanced understanding of house price variation.

Presently, there is no comprehensive database which contains transaction price along with property characteristics in England (Wood, 2015). This research aims to overcome this limitation by developing a composite research dataset[1] comprising Land Registry PPD ,OS data and EPCs to better support house price analysis over small geographical areas in England. In this paper, we create two methodologies to enrich Land Registry PPD starting with an overview of Land Registry PPD along with a basic descriptive analysis in Section 2. Section 3 introduces the two data linkages which are created to overcome two deficiencies of the Land Registry PPD. This allows an exploration of the relationship between transaction price and total floor area in Section 4. Finally, the conclusions and implications for future study are discussed in Section 5.

2. Land Registry Price Paid Data

Land Registry PPD is the administrative dataset from the Her Majesty's Land Registry, which has been published as open data since 2013 (HM Land Registry, 2015). It comprehensively records all the actual residential transactions since 1995 at address level. The Office for National Statistics (ONS) uses this data to calculate certain house price statistics, such as House Price Statistics for Small Areas (South and Henretty, 2017) and the Official House Price Index (Office for National Statistics et al., 2016). Table 1 displays an explanation of data items in the Land Registry PPD; it not only contains the property sales price, transaction date and property address information, but also shows house type (detached, semi-detached, terraced houses or flats/maisonettes), tenure (freehold/leasehold), and whether a property is newly built or whether it was sold at full market value.

Table 1 Explanations of information fields in Land Registry PPD[2]

Data Item	Explanation
Transaction unique identifier	A reference unique number which is recording each published sale. e.g. {955B1020-9223-4981-AFF1-72C47E6CC60E}
Price	Sale price (transfer deed) .
Date of transfer	Date when the sale was completed. e.g. 2006-10-13
Property type	Indicates the type of house: D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other
Old/New	Indicates the age of the property and applies to all price paid transactions, residential and non-residential. There are two categories: a newly built property, an established residential building. If the property is firstly sold since 1995 it will identify as 'a newly built property'.

[1] The composite research dataset comprises data with a range of licenses conditions which make it available for academic research. This research includes one open data under the Open Government Licence v3.0 (Land Registry PPD: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads), two OS datasets under individual academic licenses(OS MasterMap and OS AddressBase Plus) and one Domestic EPCs data (https://epc.opendatacommunities.org/) under a restrictive licence and copyright. These datasets are available for academic research upon application to the relevant authorities.

[2] Resource: https://www.gov.uk/guidance/about-the-price-paid-data .

Data Item	Explanation
	Y = a newly built property, N = an established residential building
Duration	The tenure of property: freehold, leasehold
PPD category type	Indicates the type of Price Paid transaction. A = Standard Price Paid entry, includes single residential property sold for full market value. B = Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage) and transfers to non-private individuals. Category B is identified from October 2013.
Postcode	e.g. WC1H 9QH
PAON	Primary Addressable Object Name. such as the house number or name. e.g. 36
SAON	Secondary Addressable Object Name. Where a property has been divided into separate units (for example, flats), the PAON (above) will identify the building and a SAON will be specified that identifies the separate unit/flat. e.g. Flat 302
Street	e.g. Tottenham Street
locality	e.g. London
towncity	e.g. London
district	e.g. Camden
county	e.g. Greater London
Record status	Indicates additions, changes and deletions to the records A = Addition; C = Change; D = Delete.

The Land Registry PPD records 22,578,068 transactions in England and Wales between 1/1/1995 and 31/7/2017. Figure 1 shows the house price distribution from 1995 to 2016. Over this period, house price distributions in each year are seen to be positively skewed. It means prices are mainly clustered around a relatively low value together with a few extreme high values. Meanwhile, house prices have become more and more dispersed over time as the overall range of house price has dramatically widened during the last 22 years. The two local peaks (at £125,000 and £250,000) that may be observed in the graphs since 1998 reflect the Stamp Duty Land Tax (SDLT) thresholds. Moreover, house prices after 2006 exhibit a new peak at £500,000, which is also SDLT related.

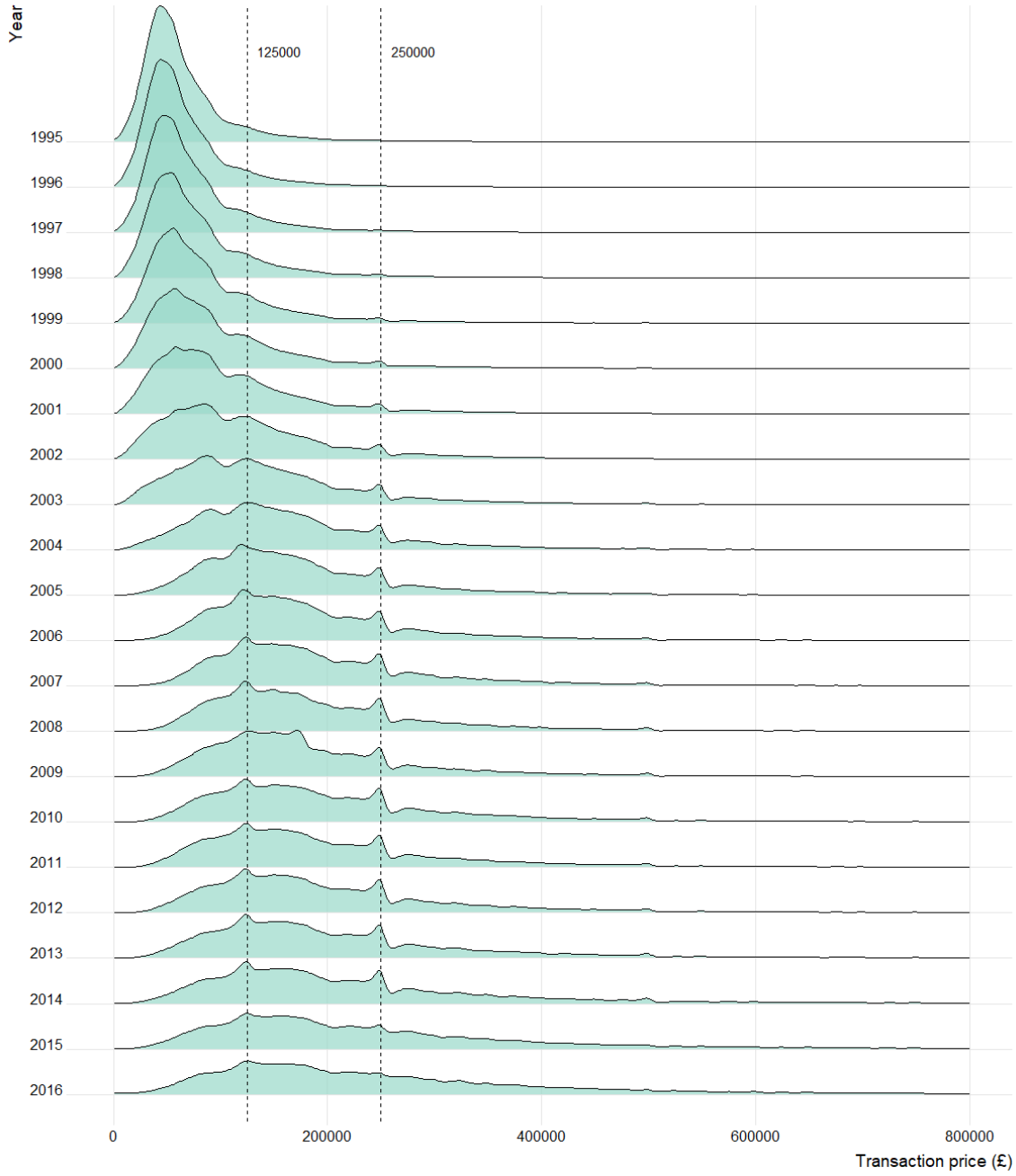


Figure 1 A Joyplot version of transaction price density plots in England and Wales,1995-2016[3]

The average number of annual transactions in England and Wales from 1995 to 2016 is around 1 million. Figure 2 shows the change of transaction volume from 1995 to 2016. There is a significant turning point when the global financial crisis erupted in 2007. Transaction numbers show a generally increasing trend from 1995 to 2007, but suddenly decrease by about a half in 2008. The number of residential property sales continues to recover after 2009, with an increase to over 1 million after 2015.

[3] The Land Registry Price Paid Data covers the period from 1/1/1995 to 31/7/2017. It does not cover the whole transactions occur in 2017. Thus all the description analysis within this section below not include the transactions in 2017. As the house price distribution shows a long tail and this graph only plot the distribution below £800,000.

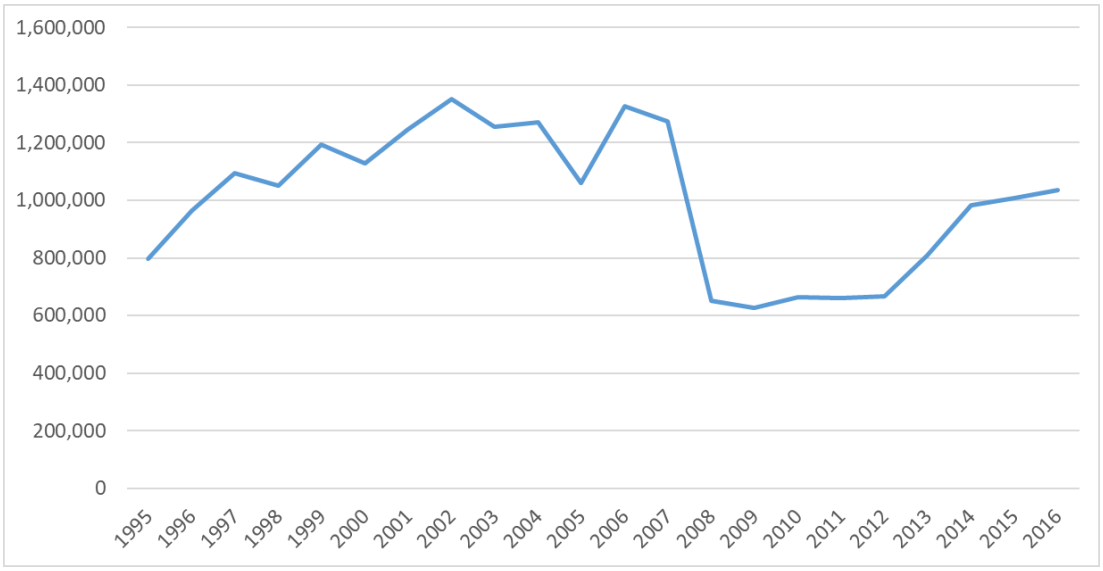


Figure 2 Transactions sales change in England and Wales, 1995-2016

3. Creating a new dataset by enriching the Land Registry Price Paid Data

Dwellings have heterogeneous characteristics and therefore the house prices will differ, even within the same neighborhood. Moreover, house prices show spatial sensitivity (Halket et al., 2015; Palm, 1978), meaning they vary across locations. That is why house price is normally presented at a certain location. Given this, the Land Registry PPD has two potential limitations for understanding house price variation. One is that it is not geocoded, the other is that it does not include property size information. Two methods are outlined below to overcome these limitations. One method aims to geo-reference transactions at the building level, whilst the other aims to add in property size (i.e. total floor area and number of habitable rooms) to the geo-referenced transactions by linking with Domestic EPCs. A brief flowchart of these two methods are shown in Figure 3.

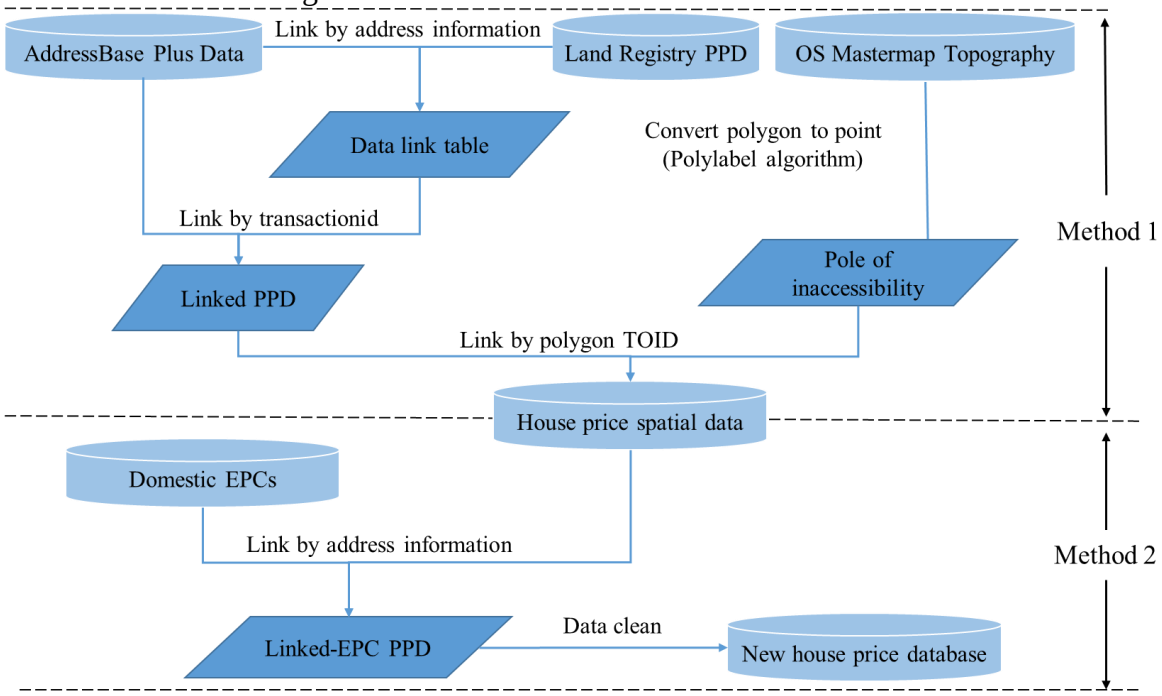


Figure 3 A brief flowchart for enhancing the Land Registry PPD

3.1 Geotagging the price paid data at building level

Geographic information exists in the form of the address string in the Land Registry PPD. The National Statistics Postcode Lookup (NSPL) is frequently used to link geographic information (i.e. latitude and longitude) to the Land Registry PPD through matching the postcode (South and Henretty, 2017). This method cannot accurately pinpoint the dwelling's real location, since it only locates the postcode's centroid point. OS MasterMap Topography Layer is a spatial dataset which represents individual buildings as geolocated polygons along with a unique geocode (TOID, Topographical Identifier). OS AddressBase Plus contains geocodes (TOID) and dwellings' postal delivery addresses from the Royal Mail across England and Wales. Linking these two datasets through geocodes (TOID, Topographical Identifier) creates a database that is interchangeable between the building's postal delivery address and its geographic information. Therefore, geocoding the Land Registry PPD can be achieved at the building location by integrating Land Registry PPD with AddressBase plus and OS MasterMap data.

Land Registry PPD, OS AddressBase Plus data and OS MasterMap Topography Layer build a foundation for the geo-referencing of transaction prices. As shown in Figure 3, Land Registry PPD and AddressBase Plus data can be linked by address information (postcode along with address strings), then link back to the OS MasterMap matching through the TOID. On the other side, an iterative grid algorithm called Polylabel (Garcia-Castellanos and Lombardo, 2007; Hügel, 2017) is used to calculate the pole of inaccessibility[4] of each polygon as proxy of geolocation of the building. The last step is to link these three datasets with TOID to build a house price spatial database.

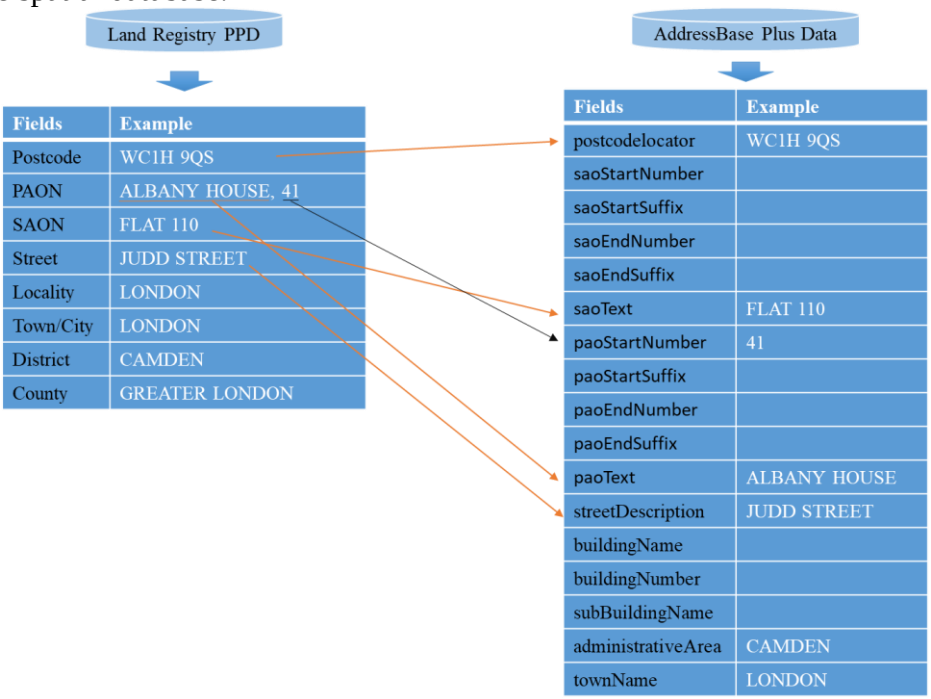


Figure 4 Address components difference in Land Registry PPD and AddressBase Plus data

Linking Land Registry PPD with AddressBase Plus by address information presents difficulties as the address records between these two datasets are structured differently (Figure 4). The full postal delivery addresses in the Land Registry PPD are categorized into four

[4] Pole of inaccessibility is a geographical point that represents the most remote place reach in a given area. The definition of pole of inaccessibility is the point within a polygon that is farthest form an edge. In cartographic visualization, it is use to label the text label on the centre of polygon.

address information items (i.e., postcode, paon, saon and street). TheAddressBase Plus data not only contains the same postcode and street records, but also includes building name, building number and sub-building name. Moreover, it divides PAO (Primary Addressable Object) information as 'paostartnumber', 'paostartsuffix', 'paoendnumber', 'paoendsuffix' and 'paotext'. Moreover, SAO (Secondary Addressable Object) information divides in the same way, named as 'saostartnumber', 'saostartsuffix', 'saoendnumber', 'saoendsuffix', 'saotext' respectively. These differences mean that matching is not straightforward and a multi-stage process is required to achieve successively more matches.

Some basic data cleaning and standardization are implemented to support the data linkage. As shown in Table 2, 33 new address variables are created to in Land Registry PPD and AddressBase Plus data, eight of them are created in the Land Registry PPD and the rest of 25 new variables are created in the AddressBase Plus data.

Table 2 New address variables created from existing address field[5]

Type	New variable	Create method
Combine	SAONPAON	Combine SAON and PAON with a blank space
	PAONSTREET	Combine PAON and street with a blank space
	bb	Combine buildingname and buildingnumber, using a comma
	pp	Combine paostartnumber and paostartsunffix
	pp1	Combing paotext and paostartnumber fields using a comma
	pp2	Combing paotext and pp fields using a comma
	pp3	Combing buildlingname and pp fields using a comma
	pp4	Combine paostartnumber and paostartsunffix using hyphens
	ppp	Combine paotext and pp4 with a blank space
	ss	Combine saostartnumber and saostartsuffix
	ss1	Combine saostartsuffix and saostartnumber
	subss	Combine subbuildingname and ss with a blank space
	saopp	Combine saotext and pp with a comma and a blank space
	sp	Combine ss and paotext fields using a blank space
	ssp	Combine saotext and sp with a comma and a blank space
	saobui	Combine fields saotext and buildingname using a blank space
	psao	Combine the paostartnumber and saotext1
	paosao	Combine the paostartnumber and saotext1
Stripping	PAON1	Stripping surrounding whitespace from hyphens and the comma in PAON field.
	PAON2	Stripping surrounding whitespace from hyphens in PAON field
	saotext1	Deleting the 'FLAT ' leading string in saotext
Prepend string	FLATSAON	Prepend the SAON with 'FLAT ' string
	FLATPAON	Prepend the PAON with 'FLAT ' string
	UNITPAON	Prepend the PAON with 'UNIT ' string
	flatsao	Prepend the saostartnumber with 'FLAT ' string
	flatss	Prepend the ss with 'FLAT ' string
	flatsub	Prepend the subbuildingname with 'FLAT ' string
	unitss	Prepend the ss with 'UNIT ' string
	flatpao	Prepend the paostartsuffix with 'FLAT ' string
	paostartnumber1	Prepend the paostartnumber with 'FLAT ' string
Replace	subbuildingnamenew	Replace 'UNIT' and 'APARTMENT' string in subbuildingname to 'FLAT ' string
	saotext2	Replace the 'APARTMENT' ,'SUITE' sting in saotext to 'FLAT ' string and delete '.' string in saotext
	SAON1	Replace the 'APARTMENT' ,' STORE' sting in SAON to 'FLAT ' string

[5] Variables written as capital stands for the new variable in Land Registry PPD.

The linkage between Land Registry PPD and AddressBase Plus data is designed to match within each unique postcode unit belonging to Land Registry PPD. However, some postcodes included in the PPD are not covered by the AddressBase Plus data. The transactions which have these postcodes are deleted first. A data linkage is created using a 13 stage process that has 84 matching rules; it is based on the address string fields shown in Figure 4 and Table 2. Details of the 13 stage process and matching rules are shown in Appendix A. The matching rate for each stage is shown in Table 3.

Table 3 Match rate for different stages

Stage	Match rate	Cumulative match rate
Stage 1	0.002%	0.002%
Stage 2	91.49%	91.50%
Stage 3	2.17%	93.67%
Stage 4	0.23%	93.90%
Stage 5	0.74%	94.64%
Stage 6	0.11%	94.75%
Stage 7	0.31%	95.07%
Stage 8	1.83%	96.90%
Stage 9	0.32%	97.22%
Stage 10	0.46%	97.67%
Stage 11	0.01%	97.68%
Stage 12	0.17%	97.86%
Stage 13	0.04%	97.89%

Land Registry PPD used here covers all transactions before 31/7/2017 in England and Wales. Using the 13 stage/84 rules model, 97.89% of transactions (22,102,551) are successfully matched. This data linkage result is designated as the data link table as shown in Figure A1 and also Figure 3. Stage 2 and stage 3 achieve a 93.67% match rate without additional stages being performed. These two stages therefore constitute the main matching process. Given the differences in address string format between the Land Registry PPD and AddressBase Plus datasets, a more complete data linkage was achieved by processing the newly created address variables through other 11 stages. These 11 stages are termed the match cleaning up process.

Following the work flow in Figure 3, the data link table obtained from the 13-stage matching linkage contains a unique transaction identifier (transactionid) from the Land Registry PPD and Topographical Identifier (ostopotoid) from OS AddressBase Plus data. Then using the Land Registry PPD with the data link table we can successfully add geocodes (Topographical Identifier, TOID) to the transaction price to give the linked PPD. After that, the linked PPD can be geo-referenced by linking the building's centre point (Pole of inaccessibility) by TOID. The method 1 process (Figure 3), successfully geo-referenced 22,019,341 records at building level and this new dataset is designated the house price spatial data set.

A sample of house price spatial data is shown in Figure 5. Following this linkage procedure, combining the linked PPD and Pole of inaccessibility derived from OS MasterMap Topography, confers two major advantages on the newly created data set. First, unlike the original PPD data, house price spatial data can now be aggregated at the level of any geographical unit (e.g. street level, OA level, regional level, etc.). Second, fully georeferenced house price data is more analytically flexible than data represented at postcode unit by linking NSPL. This flexibility allows for a much wider range of spatial analyses to be conducted, such as exploratory spatial data analysis and spatial interpolation.

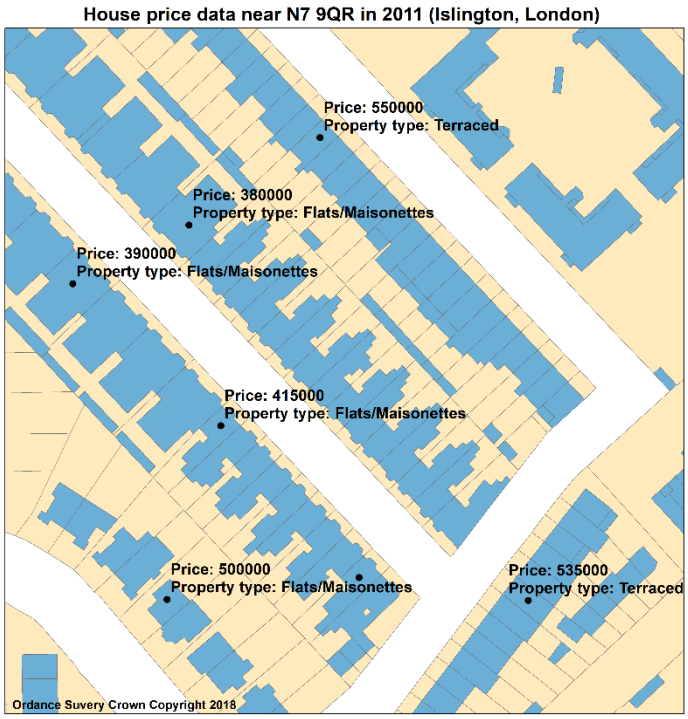


Figure 5 Sample of matched building polygons from Master Map with Land Registry data,1995

A 100% match rate is not to be expected mainly because in both datasets the addresses are structured differently. Additionally, there are three other reasons. First, 0.12% of the Land Registry records lack the postcode information in the price paid dataset. Second, some transactions do not possess matching address information in the AddressBase Plus Dataset; this may be because some properties no longer exist. Third, some transaction address records are insufficiently detailed to identify the unique building TOID in which they are situated. This issue caused one-to-many relationship problems with one (transaction) being related to many buildings during the matching process.

3.2 Enrichment house price data spatial data with property size information

Modelling suggests floor area is the most important determinant of house price (De Nadai and Lepri, 2018; Morancho, 2003; Orford, 2010; Sirmans et al., 2006; Thwaites and Wood, 2005). Thus enriching Land Registry PPD with floor area information will be highly valuable in supporting house price analysis, especially for house price variation analysis. Domestic Energy Performance Certificates (Domestic EPCs) released by the Department for Communities and Local Government (DCLG) describe a property's energy performance and its building stock information, such as its total floor area and its number of habitable rooms. EPCs are legally required when a building or building unit is offered for sale or rent in the UK and remain valid for 10 years. Some researchers have started to use the combination of Land Registry and EPC data to undertake house price research. The first house price per square metre map in England and Wales was based on data links between the Land Registry PPD and EPCs (Powell-Smith, 2017). This map offered a new insight into house price per square metre patterns at postcode district level. Moreover, Fuerst et al (2013) combined Land Registry data and EPC data to explore the relationship between energy performance and house prices across the UK in the period from 1995 to 2011. This section describes method 2 in Figure 3, which aims to enrich the house price spatial data with the total floor area information from Domestic EPCs.

(1) Data linkage

The current EPC dataset available to the public contains 85 items with 15,623,536 Domestic EPCs from 1/1/2008 to 1/10/2016. Table 4 shows the description of the key property characteristics recorded in Domestic EPCs.

Table 4 Explanations of address string and key property characteristics in EPC data[6]

Item	Explanation
Address1	First line of the address.
Address2	Second line of the address..
Address3	Third line of the address.
Postcode	The postcode of the property
Property type	Describes the type of property. e.g. Maisonette, Flat, House, Bungalow, Park home.
Built form	The building type of the Property e.g. Enclosed End-Terrace, Detached , End-Terrace, Semi-Detached, Mid-Terrace, Enclosed Mid-Terrace.
Inspection date	The date that the inspection was actually carried out by the energy assessor.
Lodgement date	Date lodged on the Energy Performance of Buildings Register.
Total floor area	The total useful floor area is the total of all enclosed spaces measured to the internal face of the external walls, the gross floor area as measured in accordance with the guidance issued from time to time by the Royal Institute of Chartered Surveyors or by a body replacing that institution.
Floor level	Flats and maisonettes only. Floor level relative to the lowest level of the property (0 for ground floor). If there is a basement, the basement is level 0 and the other floors are from 1 upwards.
Number of habitable rooms	Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar; and also a non-separated conservatory. A kitchen/diner having a discrete seating area (with space for a table and four chairs) also counts as a habitable room. A non-separated conservatory adds to the habitable room count if it has an internal quality door between it and the dwelling. Excluded from the room count are any room used solely as a kitchen, utility room, bathroom, cloakroom, en-suite accommodation and similar; any hallway, stairs or landing; and also any room not having a window.
Floor height	Average height of the storey.
Address	Field containing the concatenation of address1, address2 and address3.

Figure 6 shows the process of data linkage between house price spatial data and Domestic EPCs. These two datasets offer the property information at address level but their address structures are different, thus basic data standardization is needed before linking house price spatial data and Domestic EPCs. First, all the address string in the Domestic EPCs were capitalised and then new address variables were created separately in the house price spatial data and Domestic EPC data sets. Finally, the newly created address variables were used to achieve the data linkage. Following this process, 175 new variables were created in the house price spatial data and 94 new variables were created in the EPC data to assist the data linkage. Details of the new variable creation methods are shown in Table B1.

[6] Resources: https://epc.opendatacommunities.org/docs/guidance

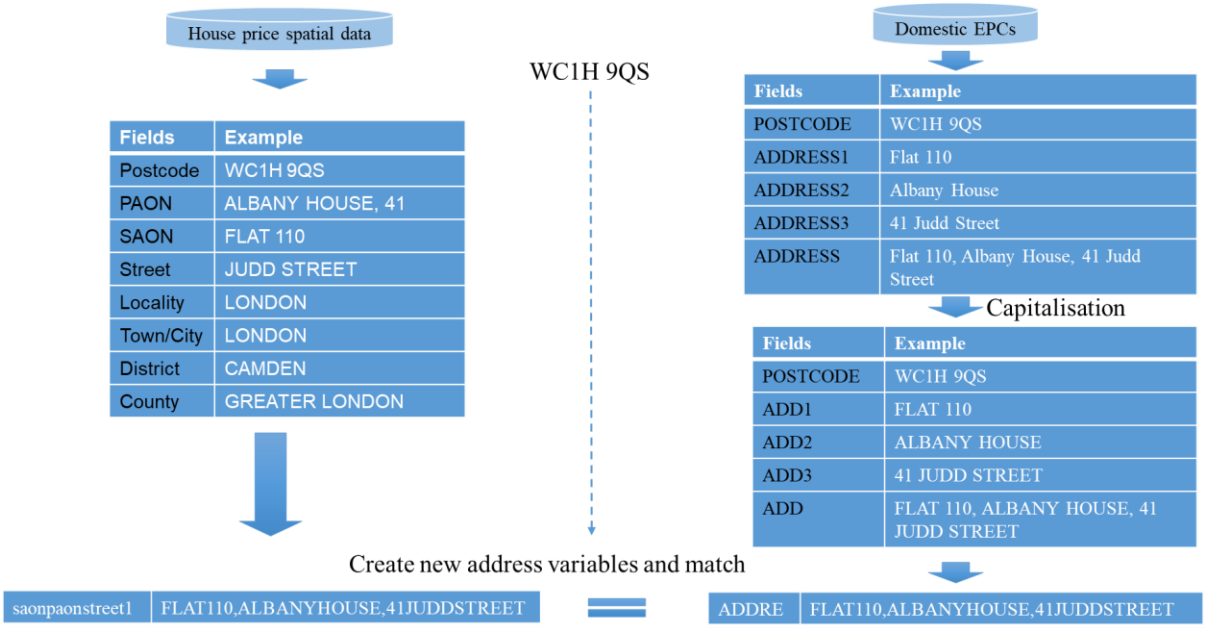


Figure 6 An example of data linkage process

Before the matching, transactions without postcodes in the Domestic EPCs dataset were excluded. A total of 0.63% of the data was deleted after applying this rule. Then with the newly created address variable in Table B1, a matching method containing a 4 stage (160 matching rules) matching process was designed to combine the house price spatial data and Domestic EPCs. Details of the matching process and matching rules are shown in Appendix C. Following the combination of house price spatial data and Domestic EPCs, 14,509,783 geo-referenced transaction records were successfully linked with EPC.

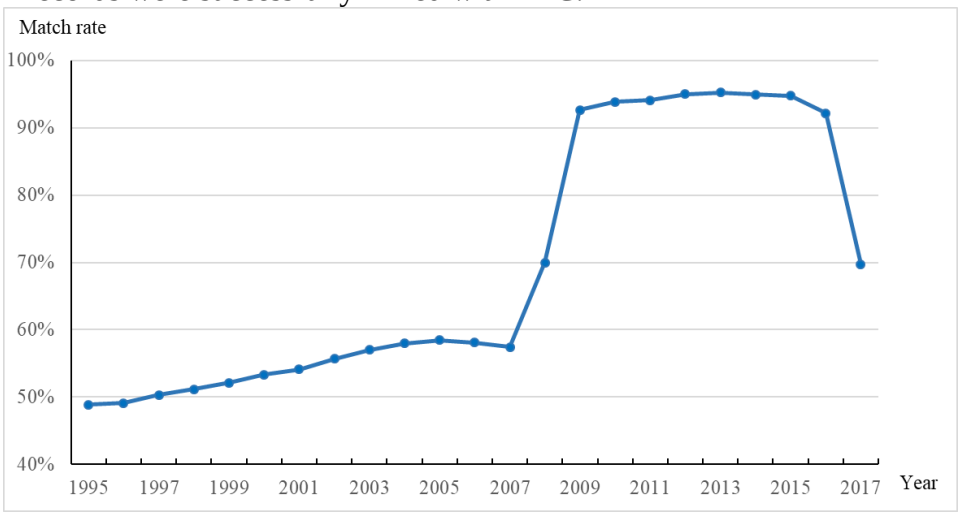


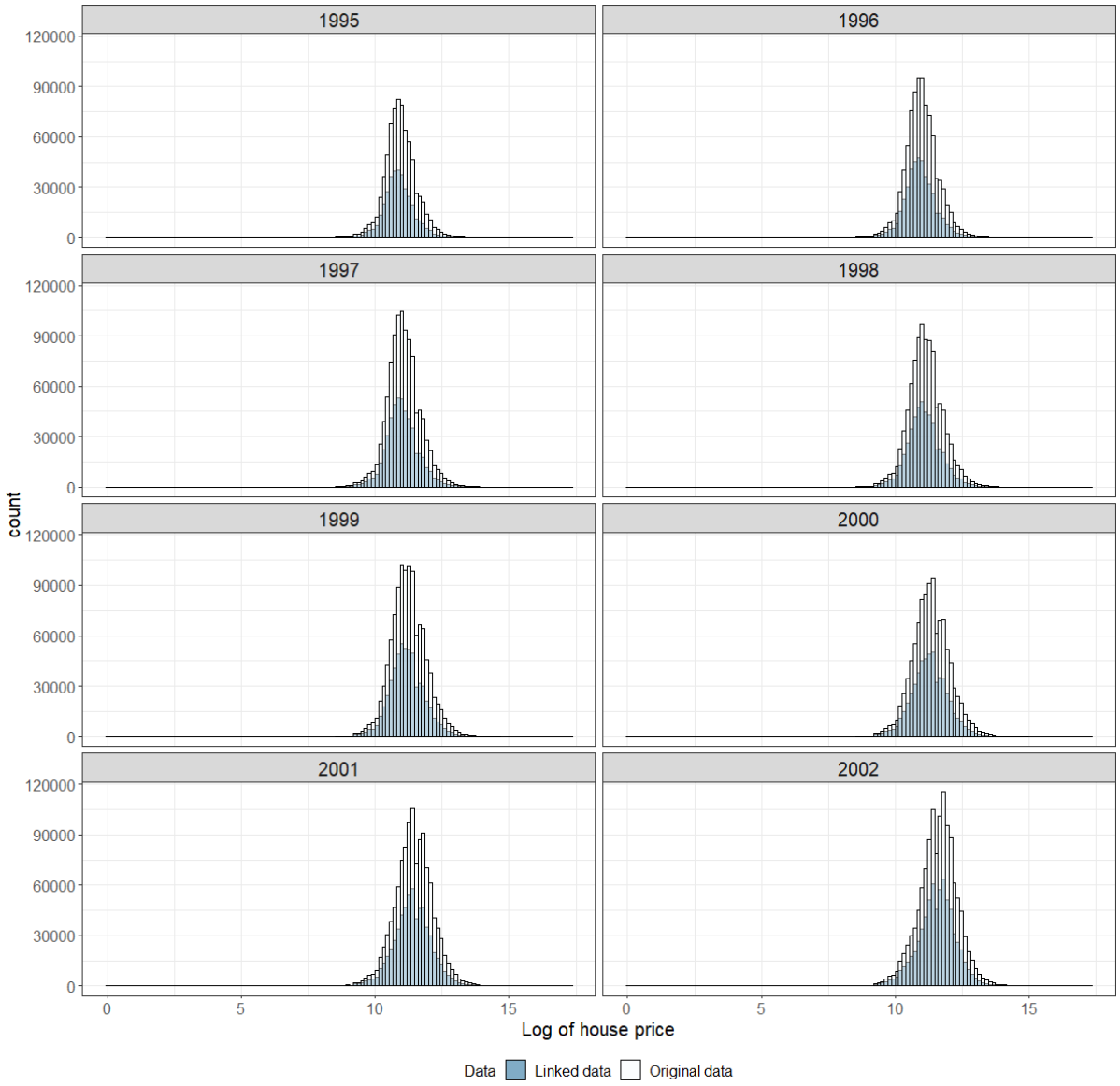
Figure 7 Match rate of house price spatial data in England,1995-2017

Within the linked EPC data, 13,878,139 of the entries are transactions in England. The match rate of transactions in England is shown in Figure 7. The matching rate between 2009 and 2016 is higher than 92%, while the matching rate of the rest of the period is lower than 70%. As the publicly available EPC data only covers the period between 2008-1-1 and 2016-10-1, the match rate is relatively high (over 90%) for the same year period (2008-2016). After checking the transactions (2008-2016) which failed to link, we found there are some sold dwellings which were not recorded in the publicly available EPC data. This makes 100% matching unachievable. The matching rate of the period before 2008 and after 2016 is in the range of 50% to 70%. This is mainly due to the dwellings sold before 2008 or after 2017 having

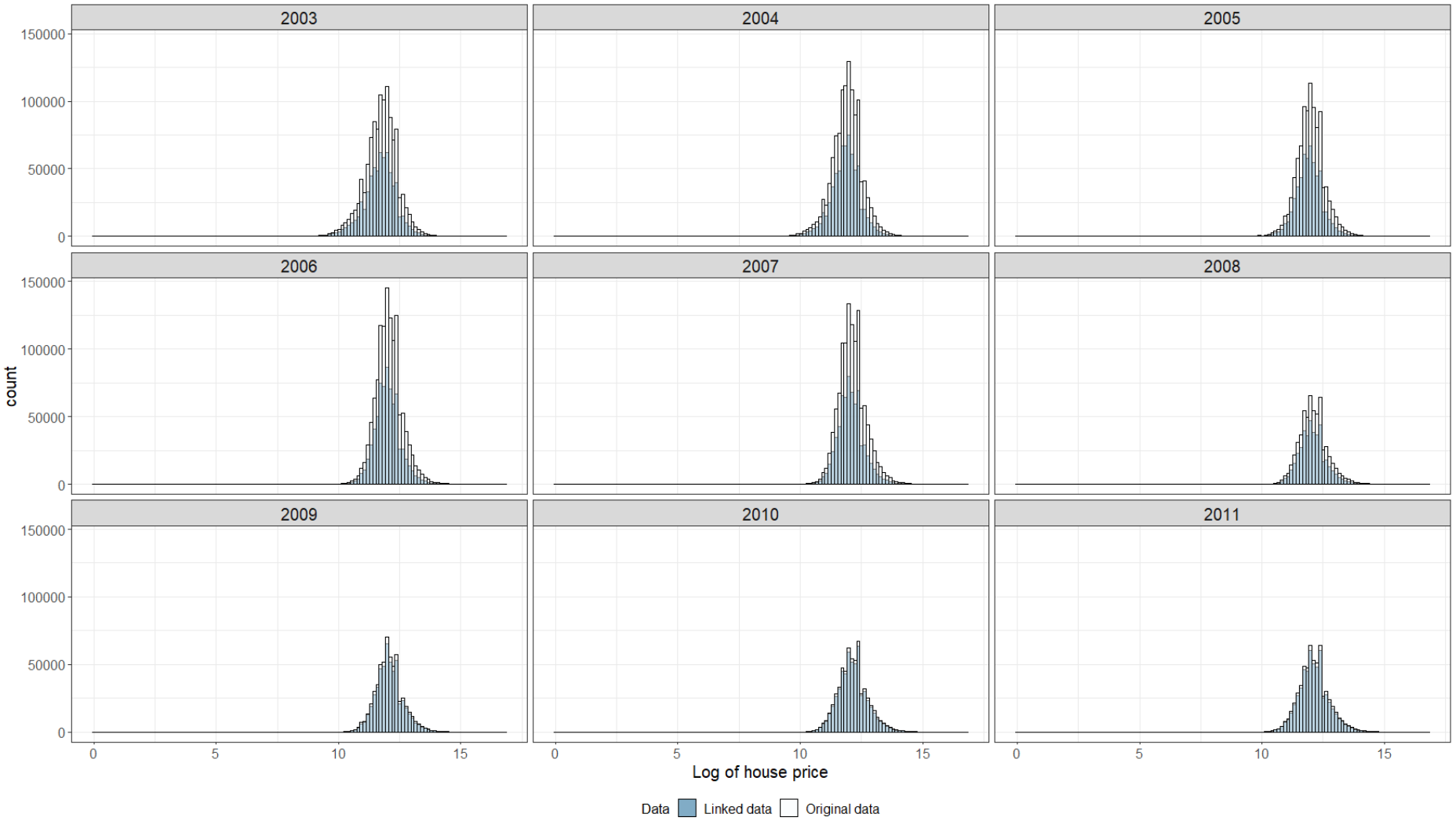
also been sold again or rented during 2008 to 2016, permitting them to be matched in the Domestic EPC.

(2) Evaluation of house price information lost after data linkage

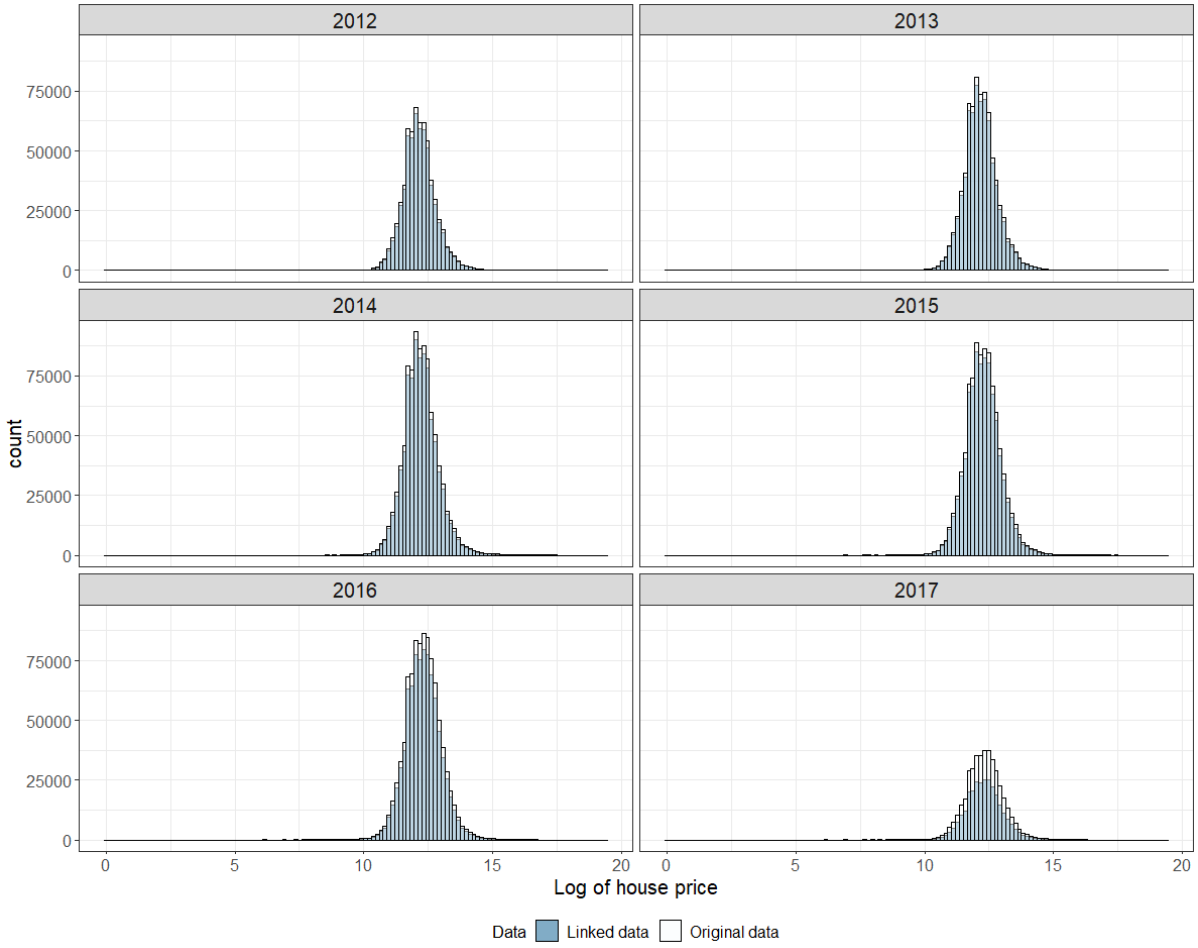
Match rates offer a crude way to quantify the matching performance, but visualization of the house price difference before and after linkage displays a clear picture of the matching performance by considering all the available house price values in the dataset. As the house price distribution follows a positive skew distribution with a long tail (Figure 1), the logarithm of house price is used to rescale the house price range. Histograms of the logarithm of house price from the transaction data in house price spatial data (geo-referenced PPD) and linked-EPC PPD in a certain given year is chosen to visualise the house price distribution change (Figure 8). In each graph, the distribution of the linked-EPC PPD (linked data) is overlaid onto the distribution of the house price spatial data (original data). The histogram of linked data is colored in white and the histogram of original data is colored in blue. Therefore, the area between the white bar and green bar represents the extent of the transactions which failed to match. After linking to the EPC data, more data was lost during the period of 2008 and 2017. However, no certain value of house price was significant lost after the data linkage.



(8A)


2003
2004
2005
2006
2007
2008
2009
2010
2011
count
150000
100000
50000
0
0
5
10
15
Log of house price
Data
Linked data
Original data

(8B)



(8C)

Figure 8 House price distribution of original data and linked-EPC PPD, 1995-2017[7]

The Kolmogorov–Smirnov test (K-S test) and the Jeffreys divergence (J-divergence) are used to quantify the extent of house price information lost. The Kolmogorov–Smirnov (K-S) test is a nonparametric test that examines the differences in the shape of a distribution. The K-S test, statistic D, is based on the maximum absolute difference between two cumulative distribution functions. Here, the test will be used to quantify the difference of two house price distributions (original data versus linked data). The Jeffreys divergences (J-divergence), derived from information theory, is a function used to establish the distance of one probability distribution to another (Jeffreys, 1946; Nielsen, 2010; Rohde, 2016). To calculate the J-divergence, the data from two different samples must first be assigned to *k* different categories. In the case of this research, these categories are a simple subdivision of the log house price into bins. The J-divergence is then defined as

$$J = \sum_{j=1}^{k} p^j \ln\left(\frac{p^j}{q^i}\right) + \sum_{j=1}^{k} q^j \ln\left(\frac{q^j}{p^i}\right) \quad (1)$$

where k is the number of categories, p^j is the proportion of data points in category j in the original house price data, and q^j is the proportion of data points in category j in the linked house price data. The final divergence measure, J, ranges from 0 to 1. If the distribution of both data samples across all the categories is the same, J will be 0. Larger values of J indicate greater

[7] Note: Original data in the graph above means house price spatial data in figure 3. Linked data means the Linked-EPC PPD data in figure 3.

differences between the two distributions.

To compute the J-divergence, the original data and linked data are divided into 150 bins, the 150 bins are created based on the 150 equal intervals of log house price in the original data in a given year. The results of J-divergence and K-S tests are shown in Figure 9. P-values of all the K-S tests are less than 0.05, which means there is a statistically significant difference between the original house price data and the linked house price data. The D statistic drops markedly after 2009, remaining at a low level thereafter. This demonstrates the distribution of house price before and after linkage are highly similar between 2009 to 2017. The J-divergence results also show that the final linked data exhibits relatively low information loss between 2009 and 2017. Considering the time period between 2009 to 2017, the information loss is slightly higher after 2016 than that shown by K-S. The information loss situation after 2015 is not as bad as for the period before 2008. Both K-S test and J-divergence test shows that the newly created house price data between 2009 to 2017 is representative of the pre-linked data and can offer a more reliable data set to represent the housing market than that for other years. As the house price data does not contain the whole of 2017 the following analyses will focus on the time period 2009 to 2016.

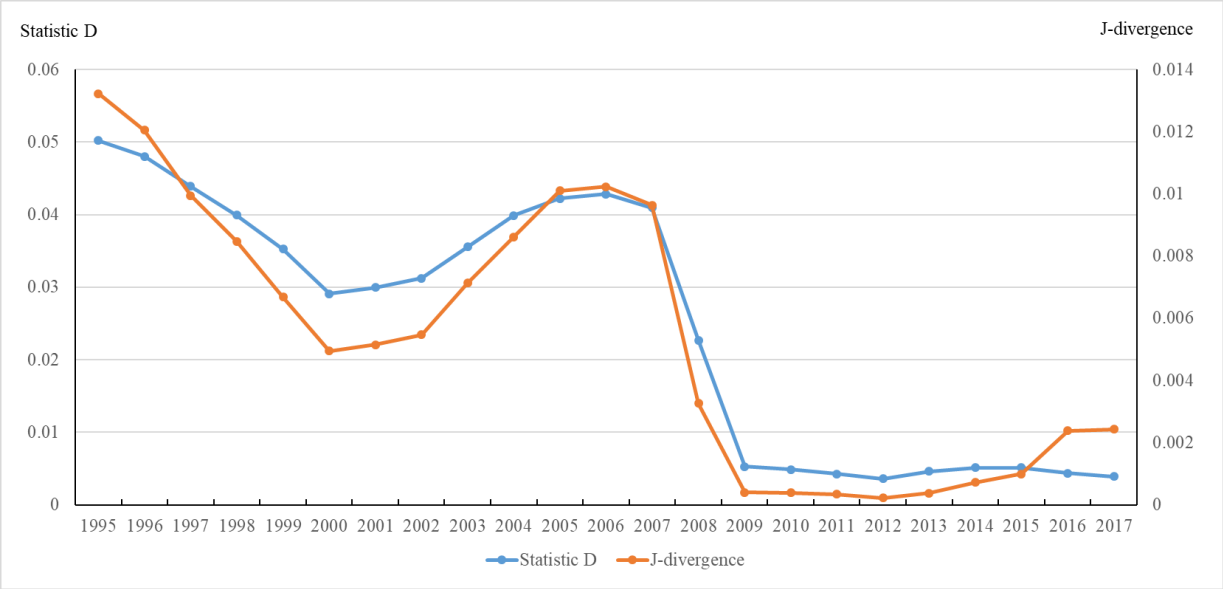


Figure 9 Results of K-S test and J-divergence method

(3) Data cleaning

There are 5,595,837 linked-EPC PPD in England for the period between 2009 to 2016. The overall matching rate of this period is 93.56%. A set of outliers are excluded from the linked-EPC PPD prior to analysis. All the excluded transactions are listed in Table 5, which accounts for 16.3% of the linked-EPC PPD. After removing these transactions, 4,681,253 transactions are left to support the house price analysis. This is the "new house price database" shown in Figure 3.

Table 5 List of transactions exclude from the linked-EPC PPD

No.	Details	Transaction numbers	Proportion
1	Transaction's property type is other .	11,580	1.27%
2	Transaction's categorytype is B.	179,757	19.65%
3	Transaction's total floor area or number of habitable rooms are NA value or 0.	719,392	78.66%
4	Transaction's total floor area is small than 9 m^2 or bigger than	578	0.06%

No.	Details	Transaction numbers	Proportion
	974 m^2.[8]		
5	Transaction's price per total floor area is bigger than 50000 £/m^2 or transaction price per total floor area is small than 200 £/m^2 .	768	0.08%
6	Transaction's floor size per habitable room is bigger than 100m^2.	704	0.08%
7	Transaction's number of habitable rooms are bigger than 20.	374	0.04%
8	Transaction's floor size per habitable room is smaller than 6.51m^2.[9]	1431	0.16%
Overall		914584	100%

4. Relationship between transaction price and total floor area in England

Using the newly created house price database, a strong positive linear association between transaction price and total floor area (as measured by the Pearson correlation coefficient) can be observed within individual local authorities. Figure 10 displays examples of this relationship for two sample local authorities in England.

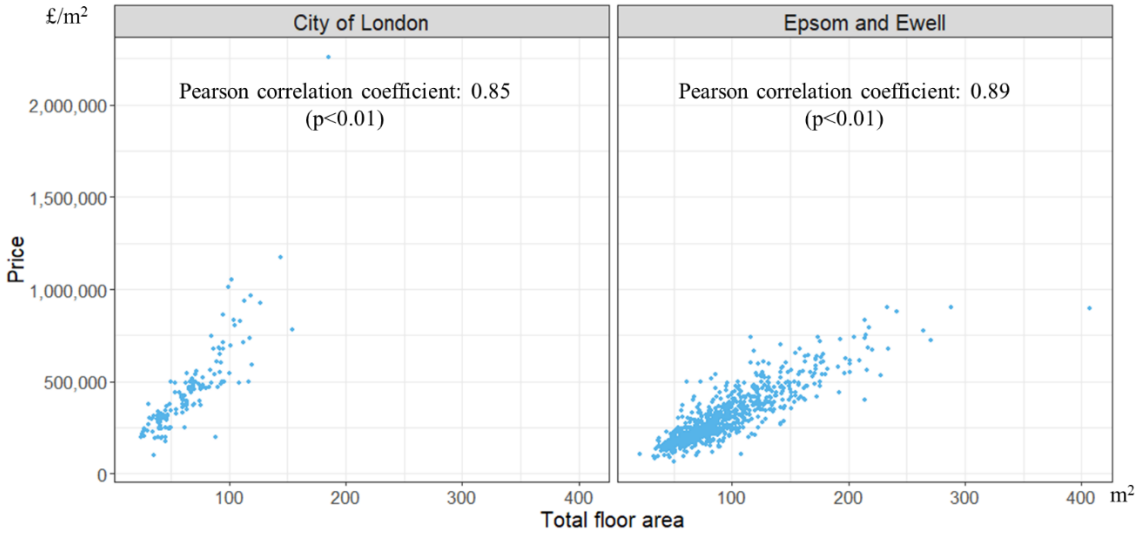


Figure 10 Transaction price against total floor area in local authorities, 2009

We are also able to observe the geography of this relationship. Figure 11 shows the extent of linear association between transaction price and total floor area in each local authority across England in 2009. For 99% of local authorities, the correlation coefficient between price and total floor area (ρ) is larger than 0.5. 79% of local authorities have ρ larger than 0.7; using the total floor area distribution in one of these local authorities, 70% of the residential house price variation can be estimated. Lower correlations reveal areas where other contextual factors are having an increased influence on house prices and these can be observed in parts of London, Manchester, Liverpool and South Yorkshire.

[8] According to the total floor area from the English housing survey (2008-2016), the range of total floor area is from 9 square metres to 974 square metres (statistics by author). All total floor area data that is not inside the range of the English housing survey is classified as outliers

[9] According to the min room size for one person aged over 10 years in The Licensing of Houses in Multiple Occupation (Mandatory Conditions of Licences) (England) Regulations 2018.Resources: http://www.legislation.gov.uk/ukdsi/2018/9780111167359/regulation/2

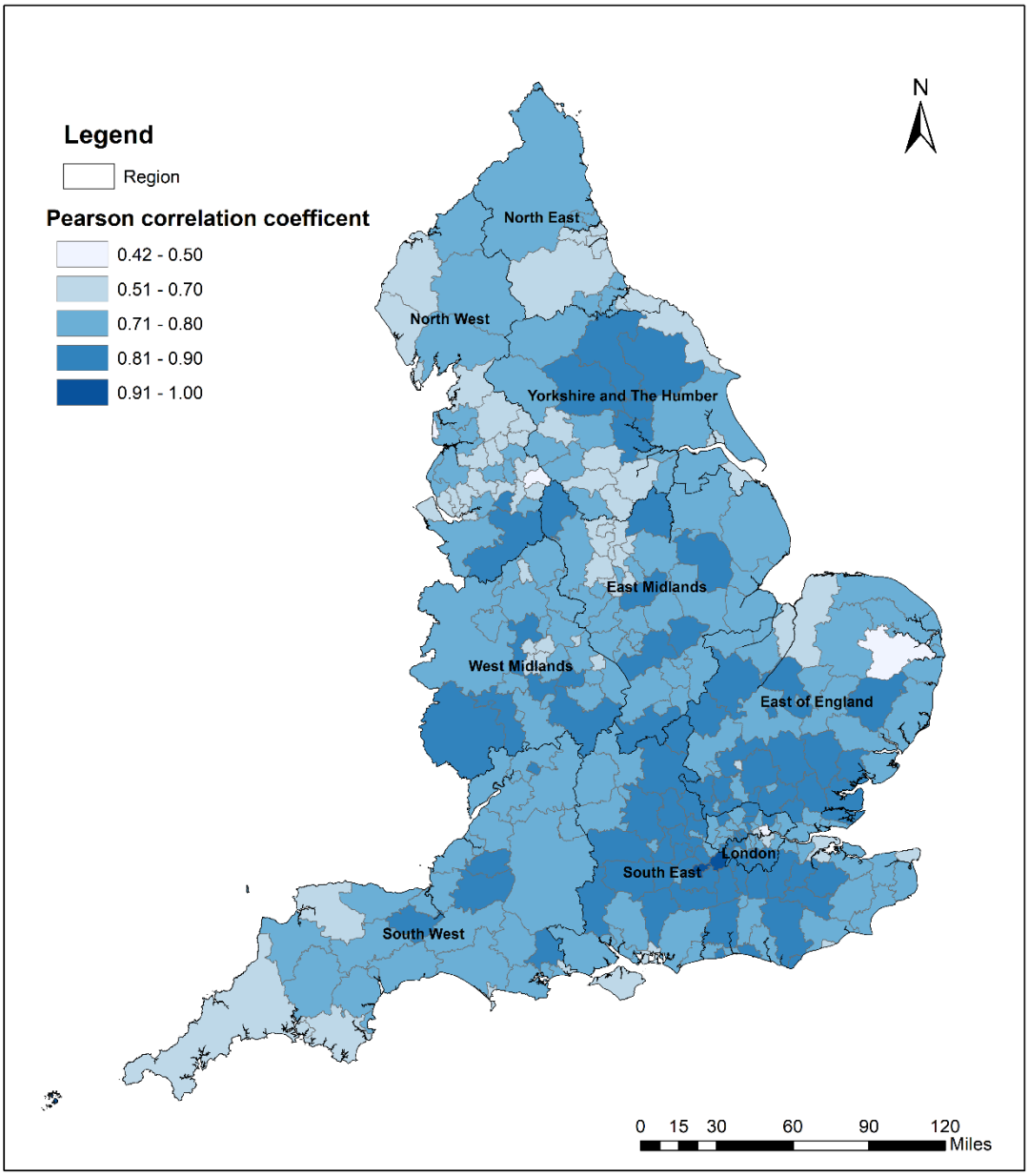


Figure 11 Pearson correlation coefficient at local authority level in England, 2009

We are able to unpick these relationships further by altering the scale of analysis. In some local authorities, house price and total floor area show a stronger linear relationship when moved to a smaller area of analysis, such as Middle Layer Super Output Area (MSOA) level and property type is controlled for. One sample is shown in Figure 12, where in Richmond upon Thames, local variations in floor area are particularly important for the price of semi-detached houses.

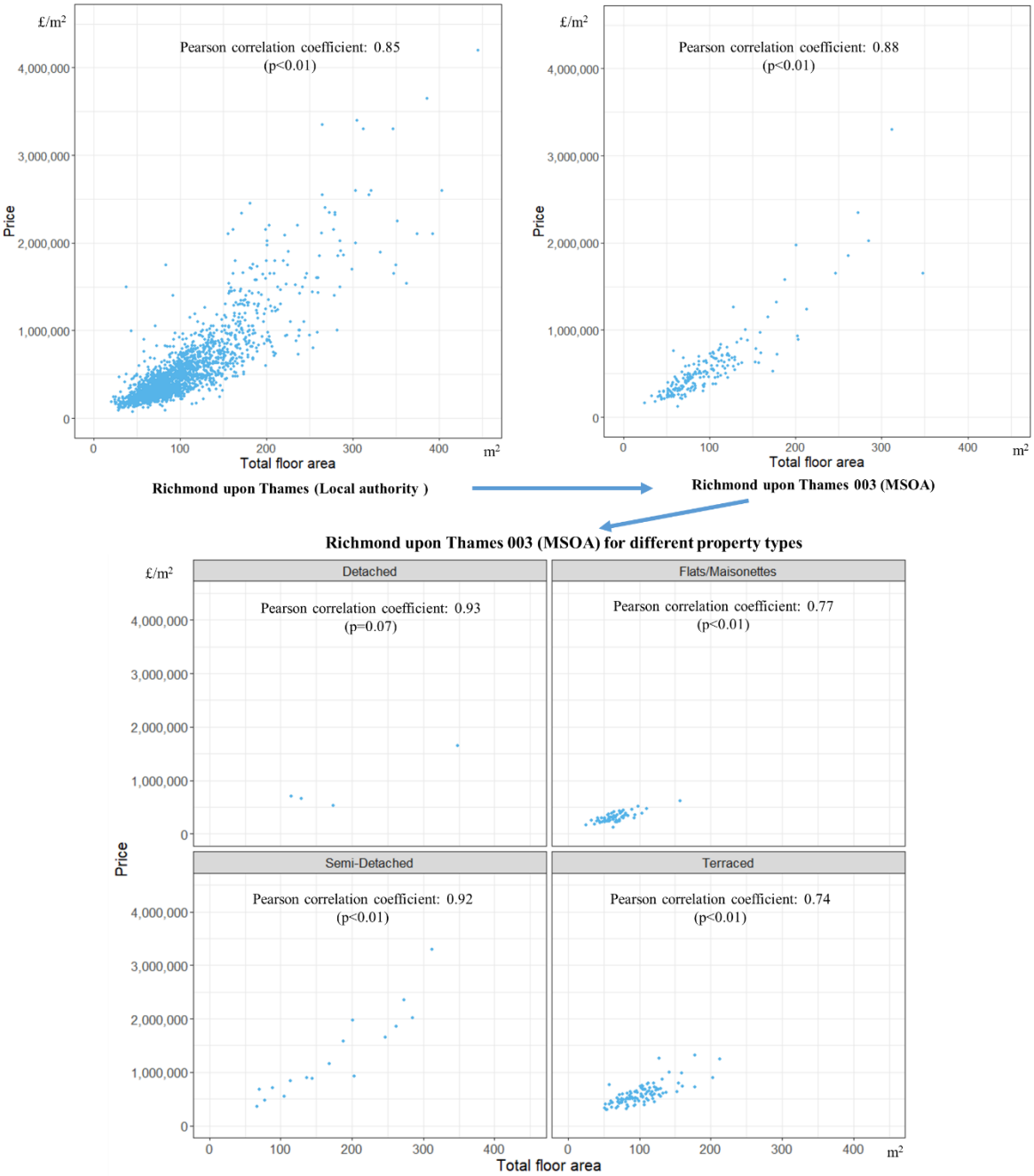


Figure 12 Transaction price against total floor area in Richmond upon Thames, 2009

5. Conclusion and Future study

This research has created a new house price data set to address the issue of incomplete house price data in England, for which there is no comprehensive database integrating both transaction price and property characteristics. We outlined one approach to address this data deficiency in England, based on the most comprehensive transaction data available (Land Registry PPD). Two data linkage methods were proposed to overcome two specific limitations of the Land Registry PPD: the lack of transaction's geo-location and of accurate property size. For the first data linkage method, a new spatial house price data can be achieved. This newly created spatial house price contains 98% of transactions in England and Wales (1995 -2016). For the second data linkage method, the newly created spatial house price has been added in the total floor area and number of habitable rooms information. According to the results of a

K-S test and J-divergence measurement, the time period from year 2009 to 2016 demonstrates a relatively high matching performance. The overall matching rate within the 2009 to 2016 period in England is 94%, which is higher than those given in previous research (Powell-Smith, 2017; Simpson et al., 2018). This valuable new dataset advances explorations in house price variation, and offers new insights into the housing market across England.

The new house price database contains transaction price and total floor area. It advances the understanding of house price variation through exploring the relationship between transaction price and total floor area. House price and total floor area show a moderate or strong linear relationship in local authorities across England. This relationship varies between different geographic scales and by different property types across England. For some areas, a stronger linear relationship was observed at MSOAs and for individual property types within individual MSOAs. The strong relationship between transaction price and total floor area shows total floor area is an importation factor that impacts house price variation. Total floor area is one measure of property size, but others, such as building volume and plot size, are also worthy of investigation since they also impact house price variation. More descriptive and statistical analysis between house price and different property sizes will be conducted in the future.

References

De Nadai M and Lepri B (2018) The economic value of neighborhoods: Predicting real estate prices from the urban environment. In: *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Italy: IEEE, pp. 323–330. IEEE. Available at: http://arxiv.org/abs/1808.02547 (accessed 20 August 2018)

Fuerst F, McAllister PM, Nanda A, et al. (2013) *Is Energy Efficiency Priced in the Housing Market? Some Evidence from the United Kingdom*. Rochester, NY: Social Science Research Network. Available at SSRN: https://papers.ssrn.com/abstract=2225270 (accessed 31 May 2018).

Garcia-Castellanos D and Lombardo U (2007) Poles of inaccessibility: A calculation algorithm for the remotest places on earth. *Scottish Geographical Journal* 123(3): 227–233. DOI: 10.1080/14702540801897809.

Gibb K and Bailey N (2016) *Data Scoping Study for a UK Housing Evidence Centre*. Available at: https://esrc.ukri.org/files/funding/funding-opportunities/uk-housing/data-scoping-study-for-a-uk-housing-evidence-centre/(accessed 13 October 2018)

Gibbons S and Machin S (2003) Valuing English primary schools. *Journal of Urban Economics* 53(2): 197–219. DOI: 10.1016/S0094-1190(02)00516-8.

Halket J, Nesheim L and Oswald F (2015) *The housing stock, housing prices, and user costs: The roles of location, structure and unobserved quality*. London, cemmap. Available at: https://www.ifs.org.uk/publications/8091(accessed 3 September 2018)

HM Land Registry (2015) Additional Price Paid Data release improves market insight. Available at: https://www.gov.uk/government/news/additional-price-paid-data-release-improves-market-insight (accessed 22 November 2018).

Hügel S (2017) *Polylabel_cmd: A Command-Line Utility for Generating Optimum Polygon Label Coordinates*. Rust. Available at: https://github.com/urschrei/polylabel_cmd (accessed 29 January 2018)

Jeffreys H (1946) An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 186(1007): 453–461.

Morancho AB (2003) A hedonic valuation of urban green areas. *Landscape and Urban Planning* 66(1): 35–41. DOI: 10.1016/S0169-2046(03)00093-8.

Nielsen F (2010) A family of statistical symmetric divergences based on Jensen's inequality. Available at: https://arxiv.org/abs/1009.4004 (accessed 8 October 2018).

Office for National Statistics, Land Registry, Registers of Scotland and Land & Property, et al. (2016) *Development of a single Official House Price Index*. Available at: https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/developmentofasingleofficialhousepriceindex (accessed 10 October 2018).

ONS (2016) *House price statistics for small areas in England and Wales: year ending September 2015*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/bulletins/housepricestatisticsforsmallareas/yearendingdec1995toyearendingsept2015 (accessed 18 November 2018).

ONS (2017) *House price statistics for small areas in England and Wales: year ending June 2017*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/bulletins/housepricestatisticsforsmallareas/yearendingjune2017 (accessed 4 November 2018).

Orford S (2010) Towards a Data-Rich Infrastructure for Housing-Market Research: Deriving Floor-Area Estimates for Individual Properties from Secondary Data Sources. *Environment and Planning B: Planning and Design* 37(2): 248–264. DOI: 10.1068/b35082

Palm R (1978) Spatial Segmentation of the Urban Housing Market. *Economic Geography* 54(3): 210–221. DOI: 10.2307/142835.

Powell-Smith A (2017) House prices by square metre in England & Wales. Available at: https://houseprices.anna.ps (accessed 19 November 2018).

Rohde N (2016) J-divergence measurements of economic inequality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(3): 847–870. DOI: 10.1111/rssa.12153.

Simpson P, Nesheim L, Halket J, et al. (2018) Estimating the benefits of transport investment. Available at: https://www.ifs.org.uk/publications/13241 (accessed 24 January 2019).

Sirmans GS, MacDonald L, Macpherson DA, et al. (2006) The Value of Housing Characteristics: A Meta Analysis. *The Journal of Real Estate Finance and Economics* 33(3): 215–240. DOI: 10.1007/s11146-006-9983-5.

South B and Henretty N (2017) House price statistics for small areas: Using administrative data to give new insights. *Statistical Journal of the IAOS* 33(3): 609–614. DOI: 10.3233/SJI-160340

Thwaites G and Wood R (2005) *The Measurement of House Prices*. Bank of England Quarterly Bulletin, Spring 2003. Available at SSRN: https://ssrn.com/abstract=707043(accessed 17 October 2018).

Whitehead C, Monk S, Clarke A, et al. (2008) Measuring Housing Affordability: A Review of Data Sources. Cambridge: Cambridge Centre for Housing and Planning Research.

Wood R (2015) A comparison of UK residential house price indices. BIS Papers chapters, in: Bank for International Settlements (ed.), Real estate indicators and financial stability, volume 21, pages 212-227.

Appendix A

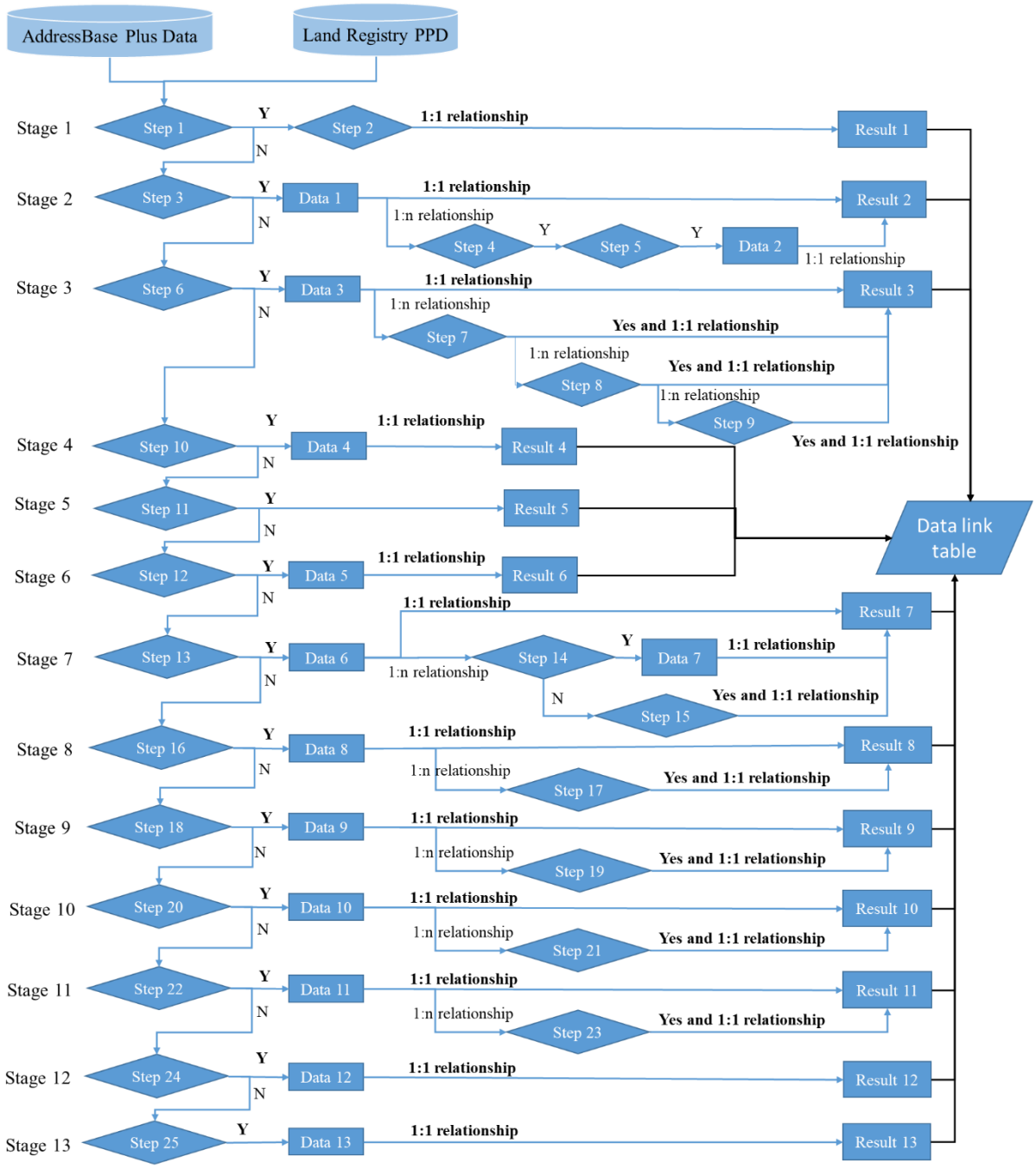


Figure A1 Master workflow of the 13 stages data linkage[10]

Figure A1 presents the master workflow of 13 stages of data linkage between Land Registry PPD and AddressBase Plus data. Each stage contains more than one step and each step contains more than one match rules. Details of the match rules for each step are listed in Table A1. For each step, we assess whether all corresponding matching rules listed in the table are satisfied. If yes, the matching process will go the branch marked "Y" in Figure A1; otherwise, the matching process follow the branch marked "N". Take the first match part in Stage 2 as an

[10] Land Registry Price Paid Data used here covers the whole transactions before 31/7/2017, the version of OS MasterMap is 21/09/2017. The version of AddressBase Plus is 06/12/2017

example, in the OS MasterMap Topography Layer, TOID[11] is a unique reference to identify the building feature. TOID contained in the AddressBase Plus data is named as ostopotoid. Meanwhile in the Land Registry Price Paid Data, each transaction has a unique identifier named as transactionid. In each step we loop the matching rules within the same postcode. When putting in the Data 1 in the matching process of step 3. Firstly, a function starts with creating a dataset which contains all the unique postcodes from Land Registry Price Paid Data (temp data1), then the function continue subset all the records from Land Registry Price Paid Data and AddressBase Plus from a given postcode unit in temp data1, then the match executes the match rules in step 3 (i.e. "test whether PAON of each transaction in Land Registry Price Paid Data is equal to buildingname in AddressBase Plus" or "test whether PAON of each transaction in Land Registry Price Paid Data is equal to buildingnumber in AddressBase Plus" or "test whether PAON of each transaction in Land Registry Price Paid Data is equal to bb in AddressBase Plus"), if the result is YES then transactionid and ostopotoid will directly link based on the match rules in step3 and restore in Data 1. After this, a new function will be used to identify if there is a one transactionid match one ostopotoid and if the result is YES and this tested link result will store in Result 1 dataset. Otherwise it will go to Stage 3 to conduct the match test in step 4. Following this all the successful 1:1 match link in Stage 2 will store in Result 2 dataset and final store in Data link table. All the matching process in Figure A1 works the same as described above and the final result is data link table. The data linkage job is conducted in RStudio.

Table A1 Details of 84 matching rules in 13 stages[12]

Stage No.	Step No.	Match rules
Stage 1	Step 1	PAON is NA[13]
	Step 2	SAON is equal to pp
		SAON is equal to saostartnumber
		SAON is equal to psao
		SAON is equal to paostartnumber1
		SAON is equal to paotext
		SAON is equal to buildingname
Stage 2	Step 3	PAON is equal to buildingname
		PAON is equal to buildingnumber
		PAON is equal to bb
	Step 4	SAON is not NA
	Step 5	PAON is equal to buildingname and SAON is equal to saotext
		PAON is equal to buildingname and SAON is equal to subbuildingname
		PAON is equal to buildingname and FLATSAON is equal to subbuildingname
		PAON is equal to buildingname and FLATSAON is equal to subbuildingnamenew
		PAON is equal to buildingname and SAON is equal to flatss
		PAON is equal to buildingnumber and SAON is equal to saotext
		PAON is equal to buildingnumber and SAON is equal to subbuildingname
		PAON is equal to paostarnumber and SAON is equal to saotext

[11] All the words coloured in grey shading are the fields name.

[12] In all the matching rule of this table, capital word coloured in grey stands for the address field in Land Registry, the capitalized word coloured in grey stands for the address field in AddressBase Plus data.

[13] The matching rule 'PAON is NA' means test the PAON in the Land Registry PPD is NA

Stage No.	Step No.	Match rules
		PAON is equal to saostarnumber and SAON is equal to saotext
		PAON is equal to pp1 and SAON is equal to saotext
		PAON is equal to pp1 and SAON is equal to saotext2
		PAON is equal to pp1 and SAON is equal to flatsao
		PAON is equal to pp1 and SAON is equal to subbuildingname
		PAON is equal to pp and SAON is equal to saotext
		PAON is equal to paotext and SAON is equal FLATSAON to saotext
		PAON is equal to paotext and SAON is equal SAON to flatsao
		PAON is equal to bb and SAON is equal to saotext
		PAON is equal to bb and SAON is equal to flatss
		PAON is equal to bb and SAON is equal to subbuildingname
		PAON is equal to bb and SAON is equal to ss
		SAONPAON is equal to buildingname
Stage 3	Step 6	PAON is equal to paostartnumber
	Step 7	PAON is equal to paostartnumber and SAON is equal to flatpao
	Step 8	PAON is equal to paostartnumber and SAON is equal to saotext
	Step 9	PAON is equal to paostartnumber and SAON is equal to saotext and STREET is equal to streetdescription
Stage 4	Step 10	PAON is equal to pp
Stage 5	Step 11	For flat transactions, do the direct match when there is only one ostopotoid in its postcode unit
Stage 6	Step 12	PAON is equal to paosao
Stage 7	Step 13	PAON is equal to pp1
		PAON is equal to pp2
Stage 7	Step 14	PAON is equal to pp2 and SAON is equal to saotext
		PAON is equal to pp2 and SAON is equal to ss
		PAON is equal to pp2 and SAON is equal to flatsao
		PAON is equal to pp2 and FLATSAON is equal to saotext
		PAON is equal to pp2 and SAON is equal to unitss
		PAON is equal to pp2 and SAON is equal to subbuildingname
		PAON is equal to pp3 and SAON is equal to subbuildingname
		PAON1 is equal to buildingname and SAON is equal to subbuildingname
	Step 15	For detached, semi-detached and terrace transactions: PAON is equal to pp1 and SAON is equal to ss
		For flat transactions: PAON is equal to pp2 and SAON is equal to ss1
Stage 8	Step 16	PAON is equal to paotext
		PAON is equal to sp
	Step 17	PAON is equal to paotext and SAON is equal to ss
		PAON is equal to paotext and FLATSAON is equal to saotext
		PAON is equal to paotext and SAON is equal to flatss
		PAON is equal to paotext and SAON is equal to saotext
		PAON is equal to paotext and SAON is equal to pp
		PAON is equal to paotext and SAON is equal to subss
		PAON is equal to paotext and SAONPAON is equal to saobui
Stage 9	Step 18	PAON1 is equal to buildingname

Stage No.	Step No.	Match rules
		PAON1 is equal to pp4
	Step 19	PAON1 is equal to buildingname and SAON is equal to subbuildingname
		PAON1 is equal to buildingname and SAON is equal to saotext
		PAON1 is equal to buildingname and SAON is equal to flatsub
		PAON1 is equal to buildingname and SAON is equal to ss
		PAON1 is equal to buildingname and FLATSAON is equal to subbuildingname
		PAON1 is equal to pp4 and SAON is NA
		PAON1 is equal to ppp and SAON is equal to ss
		PAON1 is equal to ppp and SAON is equal to flatss
		PAON1 is equal to ppp and SAON is equal to saotext
		PAON2 is equal to pp4 and SAON is equal to saotext
		PAON2 is equal to pp4 and FLATSAON is equal to saotext
		PAON2 is equal to pp4 and SAON is equal to subbuildingname
		PAON2 is equal to pp4 and SAON is equal to ssp
Stage 10	Step 20	STREET is equal to paotext
	Step 21	STREET is equal to paotext and PAON is equal to ss
		STREET is equal to paotext and FLATPAON is equal to saotext
		STREET is equal to paotext and UNITPAON is equal to saotext
		PAONSTREET is equal to paotext
Stage 11	Step 22	PAON is equal to saopp
	Step 23	PAON is equal to saopp and SAON is equal to flatss
Stage 12	Step 24	SAONPAON is equal to buildingname
Stage 13	Step 25	PAON is equal to ss and SAON is NA

Appendix B

Table B1 New address variables created from Land Registry PPD and EPC datasets

Type	Variable	Create method	Dataset
Capitalisation	ADD1	Capitalised the all the string in ADDRESS1, then remove leading and trailing whitespace	Domestic EPCs
	ADD2	Capitalised the all the string in ADDRESS2, then remove leading and trailing whitespace	Domestic EPCs
	ADD3	Capitalised the all the string in ADDRESS3, then remove leading and/or trailing whitespace	Domestic EPCs
	ADD	Capitalised the all the string in ADDRESS, then remove leading and/or trailing whitespace	Domestic EPCs
Subset string	ADD2NEW	Delete the '-' in the ADD2	Domestic EPCs
	ADDC	Delete all the '/' , ' .', ''' punctuation characters and blank space in ADD	Domestic EPCs
	ADDU	Delete the 'UNIT ' string in the ADD, then delete all the comma and blank space	Domestic EPCs
	ADDC3	Delete the comma in ADDC	Domestic EPCs
	ADDCC	Delete all the '-' , '/' , ' .', ''' punctuation characters and blank space in ADD	Domestic EPCs
	ADDC4	Delete all the '/' , ' .', '-' punctuation characters and blank space in ADD	Domestic EPCs
	ADDC6	Delete all the '' , ' ,' punctuation characters and blank space in ADD	Domestic EPCs
	ADDRE	Delete the blank space in ADD	Domestic EPCs
	ADDREC	Delete the comma in ADDRE	Domestic EPCs
	ADD1C	Delete all the '/' , ' .', ''' punctuation characters and blank space in ADD1	Domestic EPCs
	ADD1CC	Delete '-' punctuation characters in ADD1C	Domestic EPCs
	ADD1C2	Delete the comma in ADD1C	Domestic EPCs
	ADD1C3	Delete the comma in ADD1	Domestic EPCs
	ADD1C6	Delete the 'UNIT ' in ADD1, then delete all the comma and blank space	Domestic EPCs
	ADD1C4	Delete '' punctuation characters in ADD1C3	Domestic EPCs
	ADD1C5	Delete the '.' and blank space in ADD1	Domestic EPCs
	ADD1C7	Delete all the comma and blank space in ADD1	Domestic EPCs
	ADD1C8	Delete all the comma in ADD1C5	Domestic EPCs
	ADD1C9	Delete the all the blank space in ADD1	Domestic EPCs
	ADD1C10	Delete the '/' punctuation characters in ADD1	Domestic EPCs
	ADDRE2	Delete the blank space in ADD2	Domestic EPCs
	ADDRE3	Delete the blank space in ADD3	Domestic EPCs
	ADD12C2	Delete the comma in ADD12	Domestic EPCs

Type	Variable	Create method	Dataset
	ADD12C	Delete ‘ .’, ’’’ ,‘/’ punctuation characters in ADD12	Domestic EPCs
	ADD12C1	Delete ‘ .’, ’’’ ,‘/’ punctuation characters and comma in ADD12	Domestic EPCs
	ADD12C3	Delete all ‘ .’, ’’’ ,‘/’ , ‘-’ punctuation characters and comma in the ADD12	Domestic EPCs
	ADD12C4	Delete all the ‘ .’, ’-’ ,‘/’ and blank space in ADD12	Domestic EPCs
	ADD12C5	Delete all the ‘ .’, ’,’ and blank space in ADD12	Domestic EPCs
	ADD13C	Delete ‘ .’, ’’’ ,‘/’ punctuation characters in ADD13	Domestic EPCs
	ADD13C1	Delete the comma in ADD13C	Domestic EPCs
	ADD13C2	Delete the comma in ADD13	Domestic EPCs
	ADD23C	Delete ‘ .’, ’’’ ,‘/’ punctuation characters in ADD23	Domestic EPCs
	ADD23C1	Delete the comma in ADD23C	Domestic EPCs
	ADD161	For the ADD1 contain a comma, then select the string before the first comma	Domestic EPCs
	ADD162	For the ADD1 contain a comma and ‘-’ punctuation character, then select the string after the first comma, then delete the ‘-’ punctuation character	Domestic EPCs
	ADD165	For the ADD1 contain a comma and ‘.’ punctuation character, then select the string after the first comma, then delete the ‘.‘ punctuation character	Domestic EPCs
	add1sp	For the add2 is not start with number string and also does not contain a word with one character, select the string before the first blank space	Domestic EPCs
	add63	Delete ‘-’ and ‘.’ in add62	Domestic EPCs
	add1nnn	Delete ‘NO ’ string in ADD1, then delete all the comma	Domestic EPCs
	ADD1df1	Delete ‘FLAT ’ string in ADD1 , then select the string the first string before the first word boundary, then delete the comma	Domestic EPCs
	ADD1du	Delete the ‘UNIT ’ string in ADD1 , then delete all the comma and blank space	Domestic EPCs
	ADD163	Select the string before the first blank space in ADD1	Domestic EPCs
	add261	For the add2 contain a comma, then select the string before the first comma	Domestic EPCs
	add263	Select the string before the first blank space, then delete comma	Domestic EPCs
	add31	Delete ‘’’,’.’ and ‘/’ in ADD3	Domestic EPCs
	fladd1c	Delete all the blank space in fladd1	Domestic EPCs
	fladdc	Delete all the comma in the fladd	Domestic EPCs
	ADD1dff	For the ADD1 has ‘FLAT ‘, delete ‘FLAT ’ string in ADD1	Domestic EPCs
	add264	Select the string after the first blank space	Domestic EPCs
	apADD1	Delete ‘-’ , ‘/’ , ‘ .’, ’’’ ‘,’ punctuation characters and blank space in ADD	Domestic EPCs
	ADDr61	For the ADD contain a comma, then select the string before the first comma	Domestic EPCs

Type	Variable	Create method	Dataset
	ADDr62	For the ADD contain a comma and -punctuation character, then select the string after the first comma, then delete the '-' , '', '.'punctuation character	Domestic EPCs
	add361	For the ADD3 contain a comma, then select the string before the first comma	Domestic EPCs
	ADDC5	Delete all the '/' , ' .' punctuation characters and blank space in ADD	Domestic EPCs
	ADDC7	Delete all the '-' punctuation characters and blank space in ADD	Domestic EPCs
	ADDC8	Delete all the '.',''' punctuation characters and blank space in ADD	Domestic EPCs
	ADDC9	Delete all the '.',''' and '/' punctuation characters in ADD	Domestic EPCs
	ADDC10	Delete all the '-' , '/' , ' .', ''' ,'.' punctuation characters and blank space in ADD	Domestic EPCs
	ADD262	For the ADD2 contain a comma character, then select the string after the first comma	Domestic EPCs
	add1f61	For the ADD1 in EPC data has 'FLAT ' string, then delete the FLAT ' string, then subset the string before the first comma , then delete the all the comma	Domestic EPCs
	add1f61f2	combine 'FLAT ' and add1f61 , then combine ADD2 with a comma and a blank space, then delete all the blank space and comma.	Domestic EPCs
	adddap	Delete 'APARTMENT ' string in ADD , then delete all the blank space	Domestic EPCs
	saonn	Delete all the '/' punctuation characters in SAON	House price spatial data
	paonn	Delete all the '', '.' punctuation characters in PAON	House price spatial data
	paonn2	Delete comma and blank space in PAON	House price spatial data
	paonn3	Delete '-' and blank space in PAON	House price spatial data
	streetn	Delete all the '' punctuation characters in street	House price spatial data
	streetn1	Delete '-' , ' .', ''' punctuation characters and blank space in street	House price spatial data
	streetn2	Delete '-' , ''' punctuation characters and blank space in street	House price spatial data
	streetn5	Delete '/' , ' .', ''' punctuation characters and blank space in street	House price spatial data
	localityn	Delete all the '', '.' punctuation characters in locality	House price spatial data
	saonpaonstreet31	Delete the comma in saonpaonstreet3	House price spatial data
	saonpaonstreetn31	Delete the comma in saonpaonstreetn3	House price spatial data
	paon61	For the PAON contain comma, subset the string before the first comma	House price spatial data
	paon61c	Delete all the blank space in paon61	House price spatial data
	paon62	For the PAON contain comma, subset the string after the first comma	House price spatial data
	paon64	Subset the string before the first blank space in PAON	House price spatial data
	paon65	Subset the string after last blank space in PAON	House price spatial data
	paon65n	For the paonn contain comma, subset the string after last blank space in paonn	House price spatial data

Type	Variable	Create method	Dataset
	saon2	Delete ‘APARTMENT ’ string in SAON	House price spatial data
	flsaon	For the SAON start with number string , combine ‘FLAT’ string with SAON with a blank space	House price spatial data
	fldsaon	Delete ‘FLAT ’ string in SAON	House price spatial data
	saon7	Replace ‘FLAT ’ string to ‘APARTMENT ’ string in SAON	House price spatial data
	saon71	Replace ‘FLAT ’ string to ‘APARTMENT ’ string in SAONN	House price spatial data
	saonn4	Delete ‘FLAT ’ string in saonn	House price spatial data
	saon1	Replace ‘APARTMENT ’ string to ‘FLAT ’ string in saonn	House price spatial data
	saonn2	Delete ‘APARTMENT ’ string in saonn	House price spatial data
	saonn3	Delete ‘.’ And ‘ / ’ in SAON	House price spatial data
	ADD1num	Extract the number string in ADD1	House price spatial data
	saonn5	Replace ‘APARTMENT ’ string to ‘UNIT ’ in saonn	House price spatial data
	paonn61	For the paonn contain comma, subset the string before the first comma	House price spatial data
	sao1	Replace ‘APARTMENT ’ string to ‘FLAT ’ string in SAON	House price spatial data
	saon8	Replace ‘LOFT ’ to ‘FLAT ’ in SAON	House price spatial data
	saon4	Delete ‘FLAT ’ string in SAON	House price spatial data
	paon6164	Select the number string from paon61	House price spatial data
	paon6163	Select all the non-digitals from paon61	House price spatial data
	paon11	Delete all the comma in the PAON	House price spatial data
Combine	ADD12	Combine ADD1 and ADD2 with a comma and a blank space, then delete all the blank space	Domestic EPCs
	ADD12new	Combine ADD1 and add2new with a blank space, then delete the ‘-’ , ‘/’ , ‘ .’, ’’’ punctuation characters, blank space and comma	Domestic EPCs
	ADD13	Combine ADD1 and ADD3 with a comma and a blank space , then delete all the blank space	Domestic EPCs
	ADD23	Combine ADD2 and ADD3 with a blank space , then delete all the blank space	Domestic EPCs
	ADD66	Combine ADD161 and ADD162 with a comma and a blank space , then delete all the comma and blank space	Domestic EPCs
	ADD662	Combine ADD66 and ADD2 with a comma and a blank space ,then delete the comma and blank space	Domestic EPCs
	ADD67	Combine ADD161 and ADD165 with a comma and a blank space , then delete all the comma and blank space	Domestic EPCs
	ADDSP12	Combine add1sp and add2 with a comma and a blank space, then delete the comma and blank space	Domestic EPCs
	ADD68	Combine add61 and add63 with a comma and a blank space , then delete ‘’ and blank space	Domestic EPCs
	ADD69	Combine add1nn and ADD2 with a comma and a blank space, then delete all the blank space	Domestic EPCs
	ADD1632	Combine ADD163 and ADD2 with a blank space , then delete all the comma and blank space	Domestic EPCs

Type	Variable	Create method	Dataset
	flADD	Combine ‘FLAT ’ string and ADD with a comma and a blank space , then delete all the comma and blank space	Domestic EPCs
	ADD2611	Combine add261 and add1 with a comma and a blank pace , then delete all the comma and blank space	Domestic EPCs
	fladd1	Combine ‘FLAT ’ and ADD1 with a blank space	Domestic EPCs
	fladd	Combine ‘FLAT ’ and ADD with a blank space , then delete all the blank space	Domestic EPCs
	flADD13	Combine fladd1 and add31 with a blank space , then delete all the comma and blank space.	Domestic EPCs
	ADD5	Combine add263 and ADD1dff , then combine add264, , then delete all the blank space	Domestic EPCs
	apadd1	Combine ‘APARTMENT ’ and ADD1 with a blank space	Domestic EPCs
	ADDr66	Combine ADDr61 and ADDr62 with a comma and a blank space , then delete all the comma and blank space	Domestic EPCs
	ADD6	Combine ADD1 and ADD2 with a comma and a blank space , then combine add361 with a comma and a blank space ,then delete all the, ‘/’ , ‘ .’, ’’’ punctuation characters and blank space	Domestic EPCs
	add12643	Combine ADD1 and add264 with a comma and a blank space, then combine add3 with a comma and a blank space, then delete all the blank spaces	Domestic EPCs
	ADD1264	Combine ADD1 and add264 with a comma and a blank space , then delete all the blank space and comma	Domestic EPCs
	ADD8	Combine ADD1C10 and add2 with a comma and a blank space ,then delete all the blank space	Domestic EPCs
	ADD7	Combine ADD161 and ADD2 with a blank space , then delete all the blank space	Domestic EPCs
	ADD1num2	Combine ADD1num and ADD2 with a comma and a blank space , then delete, ‘/’ , ‘ .’, ’’’ punctuation characters	Domestic EPCs
	ADD1262	Combine ADD1 and ADD262 with a comma and a blank space , then delete all the blank space	Domestic EPCs
	ADD1262c	Combine ADD1 and ADD262 with a comma and a blank space , then delete all the blank space and comma	Domestic EPCs
	ADD1262cc	Combine ADD1 and ADD262 with a comma and a blank space , then delete all the blank space and ’’’	Domestic EPCs
	apadd1632	Combine ‘APARTMENT’ and add163 with a blank space , then combine with ADD2 with a comma and a blank space , then delete all the blank space and comma	Domestic EPCs
	saonpaonstreet	Combine SAON and PAON with a comma and a blank space , then combine street with a blank space, then delete all the blank space	House price spatial data
	saonpaonstreet5	Combine SAON and PAON with a comma and a blank space , then combine street with a blank space, then delete all the blank space and cooma	House price spatial data
	saonpaonstreet1	Combine SAON and PAON with a comma and a blank space , then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaonstreet2	Combine SAON and PAON with a blank space , then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaonstreetn	Combine saonn and paonn with a comma and a blank space, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	saonpaonstreetn1	Combine saonn and paonn with a comma and a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data

Type	Variable	Create method	Dataset
	saonpaonstreetn2	Combine saonn and paonn with a blank space , then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaonlo	Combine SAON and PAON with a blank space, then combine locality with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaonlon	Combine saonn and paonn with a blank space, then combine localityn with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaonstreet3	Combine SAON and PAON with a blank space, then delete combine street with a blank space, then delete all the blank space	House price spatial data
	saonpaonstreetn3	Combine saonn and paonn with a blank space, then delete combine streetn with a blank space, then delete all the blank space	House price spatial data
	saonpaonstreetlo	Combine SAON and PAON with a comma and a blank space , then combine street with a comma and a blank space, then combine locality with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaonstreetnlo	Combine saonn and paonn with a comma and a blank space, then combine streetn with a comma and a blank space, then combine localityn with a comma and a blank space,, then delete all the blank space	House price spatial data
	saonpaon1	Combine SAON and PAON with a blank space, then delete all the blank space	House price spatial data
	saonpaon2	Combine SAON and PAON with a comma and a blank space, then delete all the blank space and all the blank space	House price spatial data
	paonstreetlo	Combine PAON and street with a comma and a blank space, then combine locality with a comma and a blank space, then delete all the blank space	House price spatial data
	paonstreetnlo	Combine paonn and street nwith a comma and a blank space, then combine localityn with a comma and a blank space, then delete all the blank space	House price spatial data
	paonstreetlo1	Combine PAON and street with a blank space, then combine locality with a comma and a blank space, then delete all the blank space	House price spatial data
	paonstreetnlo1	Combine paonn and street nwith a blank space, then combine localityn with a comma and a blank space, then delete all the blank space	House price spatial data
	paonstreetlo2	Combine PAON and street with a blank space, then combine locality with a blank space, then delete all the blank space and comma	House price spatial data
	paonstreetn	Combine PAON and streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	paon66	Combine paon61 and paon62 with a comma and a blank space, then delete the blank space	House price spatial data
	paon65streetlo	Combine paon65 and street with a comma and a blank space, then combine locality with a comma and a blank space,, then delete all the blank space	House price spatial data
	paon65streetnlo	Combine paon65n and streetn with a comma and a blank space, then combine localityn with a comma and a blank space,, then delete all the blank space	House price spatial data
	paon65streetlo1	Combine paon65 and street with a blank space, then combine locality with a blank space,, then delete all the blank space and comma	House price spatial data
	paon61streetlo	Combine paon61 and street with a comma and a blank space, then combine locality with a comma and a blank space,, then delete all the blank space	House price spatial data

Type	Variable	Create method	Dataset
	paon61streetlo1	Combine paon61 and street with a blank space, then combine locality with a blank space,, then delete all the blank space and comma	House price spatial data
	paon61lo	Combine paon61 and locality with a comma and a blank space, then delete all the blank space	House price spatial data
	paon61street	Combine paon61 and street with a blank space, then delete all the blank space and comma	House price spatial data
	paon65street	Combine paon65 and street with a blank space, then delete all the blank space and comma	House price spatial data
	paon66streetlo	Combine paon62 and paon61 with a blank space, then combine street with a blank space, then combine locality with a blank space, then delete all the comma and blank space	House price spatial data
	paon65streetlo	Combine paon65 and street with a blank space, then combine locality with a blank space, then delete all the comma and blank space	House price spatial data
	paon61new	Combine 'THE' and paon61 with a blank space	House price spatial data
	paonstreetlo3	Combine PAON and street with a comma and a blank space, then combine locality with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	paonstreet	Combine PAON and street with a comma and a blank space, then delete all the comma and blank space	House price spatial data
	paonstreetn1	Combine PAON and streetn1 with a comma and a blank space, then delete all the comma and all the blank space	House price spatial data
	paonstreet1	Combine PAON and street with a comma and a blank space, then delete all blank space	House price spatial data
	paonstreet2	Combine PAON and street with a blank space, then delete all blank space	House price spatial data
	paon66streetlo1	Combine paon62 and paon61 with a blank space, then combine street with a comma and a blank space, then combine locality with a blank space, then delete all the blank space	House price spatial data
	paon62streetlo	Combine paon62 and street with a comma and a blank space, then combine locality with a comma and a blank space,, then delete all the blank space	House price spatial data
	paon62streetlo1	Combine paon62 and street with a blank space, then combine locality with a blank space,, then delete all the blank space and coma	House price spatial data
	paonflat	Combine 'FLAT' string and PAON with a blank space	House price spatial data
	paonfstreet	Combine paonflat with street with a comma and a blank space, then delete all the blank space	House price spatial data
	paonap	Combine 'APARTMENT' string and PAON with a blank space	House price spatial data
	paonapstreet	Combine paonap with street with a comma and a blank space, then delete all the blank space	House price spatial data
	paonfstreet1	Combine paonflat with street with a blank space, then delete all the blank space	House price spatial data
	paonfstreetn5	Combine paonflat with streetn5 with a blank space, then delete all the blank space	House price spatial data
	paonstreet3	Combine PAON and street with a blank space, then delete all blank space and comma	House price spatial data
	paonapstreet1	Combine paonap with street with a blank space, then delete all the blank space	House price spatial data
	paonapstreet2	Combine paonap with street with a blank space, then delete all the blank space and comma	House price spatial data
	paonapstreetn5	Combine paonap with streetn5 with a blank space, then delete all the blank space	House price spatial data

Type	Variable	Create method	Dataset
	paonstreet4	Combine the PAON and street with a blank space	House price spatial data
	paonfl1	Combine 'FLAT,' string and PAON with a blank space	House price spatial data
	paonf1streetn5	Combine paonfl1 with streetn5 with a comma and a blank space, then delete all the blank space	House price spatial data
	paonfstreetn6	Combine paonflat with streetn5 with a comma and a blank space, then delete all the blank space	House price spatial data
	flpaon3streetn5	Combine 'FLAT ' string and paonn3 with a blank space, then combine with streetn5 with a blank space then delete all the blank space and '-'	House price spatial data
	saonpaon65street	Combine SAON and paon65 with a comma and a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaon62streetn2	Combine SAON and paon62 with a comma and a blank space, then combine streetn with a blank space,, then delete all the blank space	House price spatial data
	saonpaon61street	Combine SAON and paon61 with a blank space, then combine street with a comma and a blank space,, then delete all the blank space and comma	House price spatial data
	saonpaon62streetn	Combine SAON and paon62 with a blank space, then combine streetn with a blank space,, then delete all the blank space	House price spatial data
	saonpaonn	Combine saonn and paonn with a comma and a blank space, then delete all the blank space	House price spatial data
	saon2street	Combine saon2 and street with a comma and a blank space, then delete all the blank space	House price spatial data
	saon2paon61street	Combine saon2 and paon61 with a blank space, then combine street with a comma and blank space, then delete all the blank space.	House price spatial data
	flsaonpaon	Combine flsaon and PAON with a comma and a blank space	House price spatial data
	flsaonpaon1	Combine flsaon and PAON with a blank space, then delete all the blank space	House price spatial data
	flsaonpaon2	Combine flsaon and PAON with a comma and a blank space, then delete all the blank space	House price spatial data
	flsaon1	Combine 'FLAT' string with saonn with a blank space	House price spatial data
	flsaon1paonstreetn2	Combine flsaon1 with paonn with a comma and a blank space, then combine the streetn2 with a comma and a blank space, then delete all the blank space	House price spatial data
	flsaonpaonstreet1	Combine flsaon with PAON with a blank space, then combine the street with a blank space, then delete all the blank space and comma	House price spatial data
	flsaonpaon62street1	Combine flsaon and paon62 with a blank space, then combine street with a blank space, then delete all the blank space and comma	House price spatial data
	fldsaonpaonstreet1	Combine fldsaon and PAON with a blank space, then combine street with a blank space, then delete all the blank space and comma	House price spatial data
	saon7paonstreet1	Combine saon7 and PAON with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saon7paonstreet2	Combine saon7 and PAON with a blank space, then combine street with a blank space, then delete all the blank space and comma	House price spatial data
	apsaon	Combine 'APARTMENT' string with SAON with a blank space	House price spatial data

Type	Variable	Create method	Dataset
	apsaonpaonstreet1	Combine apsaon and PAON with a blank space, then combine street with a blank space, then delete all the blank space and comma	House price spatial data
	saon7paonstreetn	Combine saon71 and paonn with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saon7paonn	Combine saon7 and paonn with a comma and a blank space, then delete all the blank space	House price spatial data
	saon7paon	Combine saon7 and PAON with a comma and a blank space, then delete all the blank space	House price spatial data
	saon4paonstreetn	Combine saonn4 and paonn with a comma and a blank space, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	saon4paonstreetn1	Combine saonn4 and paonn with a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	apsaonpaon6streetn	Combine apsaon and paon62 with a comma and a blank space, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	flsaonpaonstreetn	Combine flsaon and PAON with a comma and a blank space, then combine with streetn with a blank space, then delete all the blank space	House price spatial data
	saon4paonstreetn3	Combine saonn4 and paonn with a blank space, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	saon4paonstreetn4	Combine saonn4 and paonn with a comma and a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	saon1paonstreetn	combine saon1 and PAON with a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	saon1paonstreetn1	Combine saon1 and PAON with a comma and a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	saon1paonstreetn2	Combine saon1 and PAON with a blank space, then combine streetn with a blank space, then delete all the blank space and comma	House price spatial data
	saon2paonstreetn3	Combine saonn2 and paonn with a blank space, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	saon2paonstreetn2	Combine saonn2 and paonn with a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	saonn2paonn1	Combine saonn2 and paonn with a blank space, then delete all the blank space	House price spatial data
	saonpaon62street	Combine SAON and paon62 with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saon2paonstreetn	Combine saonn2 and paonn with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saonn3paonnstreet	Combine saonn3 and paonn with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saonn2paonn1streetn	Combine saonn2 and paonn with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saonpaon62streetn1	Combine SAON and paon62 with a comma and a blank space, then combine streetn with a comma and a blank space,, then delete all the blank space	House price spatial data

Type	Variable	Create method	Dataset
	saon1paonstreet6n	Combine saon1 and paon62 with a comma and a blank space, then combine streetn with a comma and a blank space,, then delete all the blank space	House price spatial data
	saon2paonstreetn4	Combine saonn2 and paonn with a comma and a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saon5paonstreetn1	Combine saonn5 and paonn with a blank space, then combine streetn with a comma and a blank space, then delete all the blank space	House price spatial data
	paonsaon2streetn	Combine paonn and saonn2 with, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	saonpaon61streetn	Combine saonn and paonn61 with a blank space, then combine streetn with a comma and a blank space,, then delete all the blank space	House price spatial data
	saonpaon66street	Combine saonn and paon62 with a comma and a blank space, then combine paon61 with a blank space, then combine street with a blank space,, then delete all the blank space	House price spatial data
	saon1paonstreetn3	Combine saon1 and PAON with a comma and a blank space, then combine streetn with a blank space, then delete all the blank space	House price spatial data
	saon1paonstreet	Combine sao1 and PAON with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saon2paonlo	Combine saon2 and PAON with a blank space, then combine locality with a comma and a blank space, then delete all the blank space	House price spatial data
	saon1paon	Combine sao1 and PAON with a comma and a blank space, then delete all the blank space	House price spatial data
	saon1paon61street	Combine sao1 and paon61 with a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saon1paon1	Combine sao1 and PAON with a blank space, then delete all the blank space	House price spatial data
	psaonpaonstreet	Combine paon64 and SAON, then combine paon65 with a blank space, then combine street with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	saon2paon62street	Combine saon2 and paon62 with a comma and a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saon2paonstreet	Combine saon2 and PAON with a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	flsaonpaonstreet	Combine flsaon with PAON with a comma and a blank space, then combine the street with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	psaon8street	Combine PAON and fldsaon , then combine street with a blank space then delete al the blank space and comma	House price spatial data
	saonstreet	Combine SAON and street with a comma and a blank space, then delete all the blank space	House price spatial data
	saonstreet1	Combine SAON and street with a blank space, then delete all the blank space and comma	House price spatial data
	saonstreet2	Combine SAON and street with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	saonstreet3	Combine SAON and street with a blank space, then delete all the blank space	House price spatial data
	saonstreetlo	Combine SAON and street with a comma and a blank space, then combine with locality with a comma and a blank space, then delete all the blank space	House price spatial data

Type	Variable	Create method	Dataset
	unsaonpaonstreet2	Combine 'UNIT' string with SAON with a blank space, then combing PAON with a blank space, then combine with street with a comma and a blank space and the delete all the blank space.	House price spatial data
	flsaonpaonstreet2	Combine flsaon with PAON with a blank space, then combine the street with a comma and a blank space, then delete all the blank space	House price spatial data
	saon7paon6street	Combine saon7 and paon62 with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saon8paonstreet2	Combine saon8 and PAON with a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	paonlo	Combine PAON and locality with a comma and a blank space, then delete all the blank space.	House price spatial data
	flsaonpaonstreet3	Combine flsaon with PAON with a blank space, then combine the street with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	flsaonpaonstreet4	Combine flsaon with PAON with a blank space, then combine the street with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	saonpaon62steet	Combine SAON and paon62 with a comma and a blank space, then combine street with a comma and street, then delete all the blank space	House price spatial data
	flsaonpaon61street	Combine flsaon with paon61 with a blank space, then combine the street with a comma and a blank space, then delete all the blank space and comma	House price spatial data
	saon4paonstreet	Combine saon4 with PAON with a blank space, then combine the street with a blank space, then delete all the blank space	House price spatial data
	saonpaon61street1	Combine SAON and paon61 with a blank space, then combine street with a comma and a blank space,, then delete all the blank space	House price spatial data
	flsaonpaonstreet5	Combine flsaon with PAON with a comma and a blank space, then combine the street with a comma and a blank space, then delete all the blank space	House price spatial data
	paonsaonstreet	Combine PAON and SAON, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	saonpaon61	Combine SAON and paon61with a comma and a blank space, then delete all the blank space	House price spatial data
	paonsaonstreet1	Combine PAON and SAON with a comma and a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	apsaonpaon	Combine apsaon and PAON with a blank space, then delete all the blank space	House price spatial data
	saon1paon62street	Combine sao1 and paon62 with a comma and a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	apsaonpaon62street1	Combine apsaon and paon62 with a comma and a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saon2paonstreet1	Combine saon2 and PAON with a blank space, then combine street with a comma and a blank space	House price spatial data
	apsaonpaonstreet2	Combine apsaon and PAON with a blank space, then combine street with a comma and a blank space, then delete all the blank space	House price spatial data
	psaonpstreet	Combine paon6164 and SAON, then combine paon6163 with a blank space, then combine paon62 with a comma	House price spatial data

Type	Variable	Create method	Dataset
	saonpaonstreet11	Combine SAON and paon11 with a blank space, then combine street with a blank space, then delete all the blank space	House price spatial data
	saonpaon65street1	Combine SAON and paon65 with a comma and a blank space, then combine street wihit a blank space, then delete all the blank space	House price spatial data

Appendix C

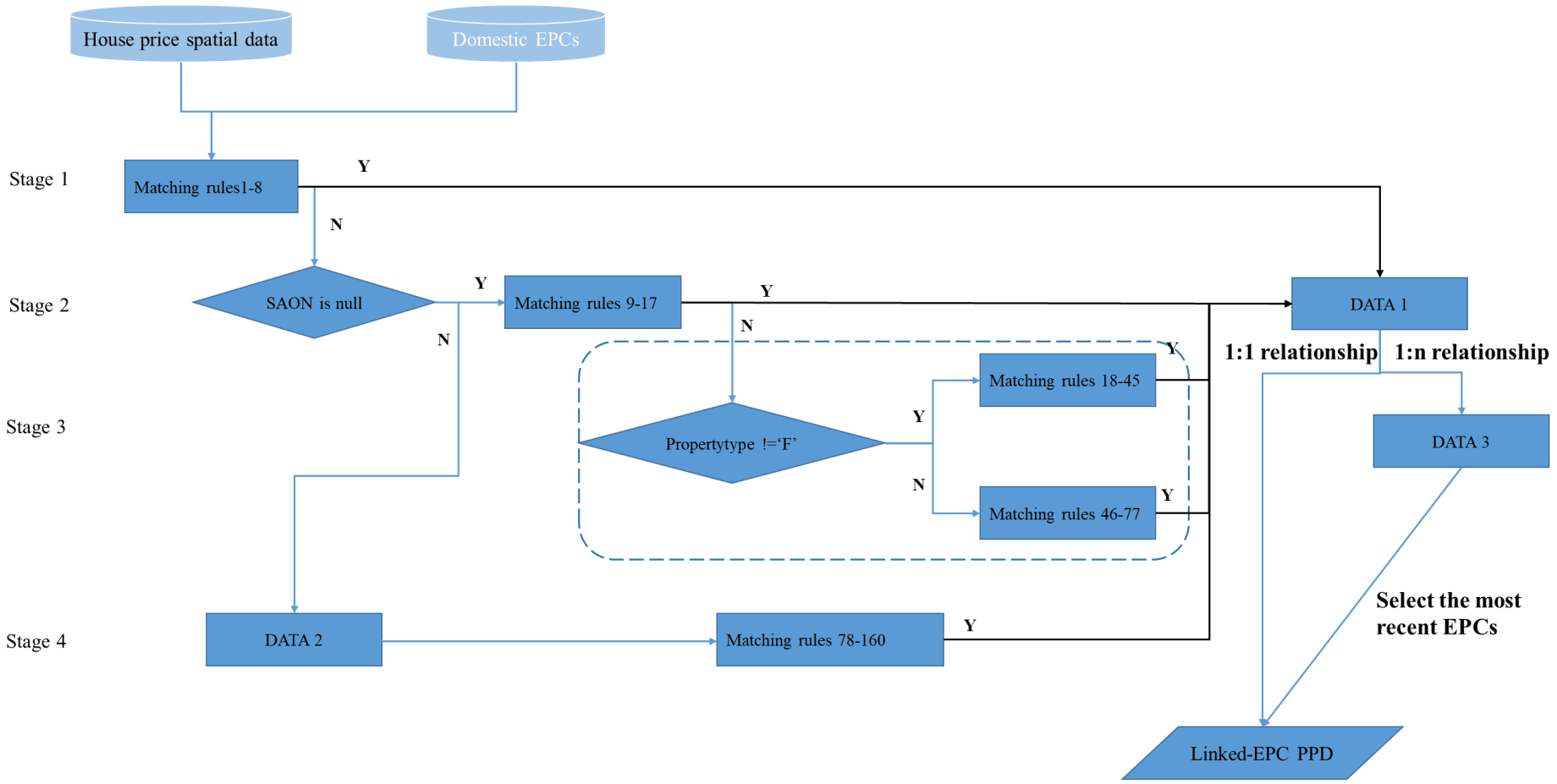


Figure C1 Master workflow of the 4 stages data linkage between house price spatial data and Domestic EPCs

Figure C1 demonstrates the data linkage workflow between Domestic EPCs and house price spatial data. Each stage contains more than one match rule. Details of the match rules for each Stage are listed in Table C1. In the Domestic EPCs, each record is created using a unique identifier with names of epcid. Each transaction in house price spatial data has a unique identifier named transactionid. The whole matching process between these two datasets is divided into four Stages. Take Stage one as an example of the matching process. All the matches are based on a "temple address strings" (i.e. postcode+saonpaonstreet) which is the combination of postcode and address variables. When Domestic EPCs and house price spatial data are put into the matching process, the process starts to link house price spatial data (transactionid) with epcid basing on the "temple address strings". For example, it tests whether postcode+saonpaonstreet in house price spatial data is equal to any postcode +ADDRE in Domestic EPCs. If the result shows yes and the epcid will direct link with transactionid and restore in Data 1, otherwise the transaction records will move to the other matching rules within the same Stage to conduct further matching tests, For the transactions that cannot be matching in Stage 1, they will move to Stage 2 to do the further matching tests. All the successfully linked transactions in Stage 1 are stored in DATA 1. However, in the real world, one property could have more than one Domestic EPCs in this matching process. The transaction property with only one successfully linked EPC will direct stored in linked-EPC PPD, transaction property with successful links to more than one EPC will be stored in DATA 3. A new function will be conducted to select all the Domestic EPCs for which total floor area is not null or 0 and will then keep the EPC's inspection date or lodgement date which is closest to the transaction date in the house price data. This result will then be stored in linked-EPC PPD. Stages 2 to 4 follow a similar process to Stage 1. Finally, linked-EPC PPD is the data linkage result. These data linkage results will firstly join back to Domestic EPCs according to the same epcid, then join with house price spatial data according to the transactionid. The data linkage process is conducted in RStudio.

Table C1 Details of matching rules in 4 stages[14]

Stage No.	Match rules No.	Match rules
Stage 1	1	(saonpaonstreet OR saonpaonstreet1 OR saonpaonstreet2 OR saonpaonlo OR saonpaonstreetlo OR saonpaonstreet3 OR saonpaon1) = ADDRE;
	2	(saopaonstreetn OR saonpaonstreetn1 OR saopaonstreetn2 OR saonpaonlon OR saonpaonstreetnlo or saonpaonstreetn3) = ADDC
	3	saonpaonstreet OR saonpaonstreet1 OR saonpaonstreet2 OR saonpaonstreet3 OR saonpaonstreetlo or saonpaonlo) = ADD12;
	4	(saopaonstreetn OR saonpaonstreetn1 OR saopaonstreetn2 OR saonpaonlon) = ADD12C;
	5	saonpaonlon = ADDCC;
	6	saonpaonstreetn3 = ADD12C1;
	7	saonpaonstreet31 = ADDREC
	8	saonpaonstreetn31 = ADDC3
Stage 2	9	(paonstreetlo OR paonstreetlo1) = ADDRE;
	10	(paonstreetnlo OR paonstreetnlo1) = ADDC;
	11	(paonstreetlo OR paonstreetlo1) = ADD12
	12	(paonstreetnlo OR paonstreetnlo1) = ADD12C

[14] In this table, all the address fields in house price spatial data is written in small letters and the address variable in the Domestic EPCs is written capital letters

	13	paonstreetlo2= ADD12C2
	14	paonstreetlo2= ADDREC
	15	paonstreetn=ADD12C1
	16	street is null and paonn3 =ADD1C
	17	For the PAON contain comma, then paon66=ADD1CC
Stage 3	18	paon65streetlo=ADDRE
	19	paon65streetlo=ADD12
	20	(paon65streetnlo OR paon66streetlo)=ADDCC
	21	(paon65streetlo1 or paon61streetlo1)=ADDREC
	22	paon61streetlo=ADDC
	23	(paon61streetlo1 OR paon65street) = ADDC3
	24	paon61streetlo1= ADD12C1
	25	paon61lo= ADD12C
	26	paon61street= ADD12C1
	27	paon61street= ADD13C1
	28	paon65street= ADD1C2
	29	paon66streetlo =ADD12C3
	30	For the propertytype in EPCs is not Flat or Maisonette, paon65streetlo =ADD23C1
	31	For the propertytype in EPCs is not Flat or Maisonette, paon61new=ADD1
	32	paonstreetlo3= ADD12new
	33	paonstreetlo3= ADD13C1
	34	paonstreetlo3 = ADD13C2
	35	paonstreet= ADD1C3
	36	PAON=ADD1
	37	paonstreetlo3 =ADD662
	38	paonstreet= ADD67
	39	For the street is null and the propertytype in EPCs is not Flat or Maisonette , paonstreet= ADDSP12;
	40	paonstreetn1=ADD1C4
	41	For the propertytype in EPCs is not Flat or Maisonette, paonstreet=ADDU
	42	paonstreet1=ADD6
	43	paonstreet1=ADD69
	44	For the address written different, (paonstreet1 OR paonstreet2) =ADD1C5
	45	For the address written different, paonn2=ADD1C6
	46	For the paon61 did not contain 'FLAT' string and 'FLOOR' string, then paon66streetlo1= ADDRE;
	47	For the paon61 does not contain FLAT' string and 'FLOOR' string and also not start with number, then paon62streetlo=ADD12
	48	paon65streetnlo=ADDCC;
	49	For property type in EPC is Flat/Maisonette, paon62streetlo1=ADDREC;
	50	For property type in EPC is Flat/Maisonette, (paon61streetlo OR paon61streetlo1)=ADDC;
	51	paon61streetlo1=ADDC3
	52	paon61streetlo1=ADD12C1
	53	For property type in EPC is Flat/Maisonette, paon61street= ADD13C1
	54	paon66streetlo= ADDCC
	55	paon66streetlo =ADD12C3
	56	paonfstreet= ADD12
	57	(paonfstreet OR paonapstreet OR paonfstreet1)= ADDRE
	58	paonstreet= ADD1C7
	59	paonstreetn1= ADD1C8
	60	For the address words written different, (paonstreet1 OR paonstreet2)=ADD1C5;

		PAON=ADD1df1; paonn2=ADD1du; paon61c=ADD1C9; paonfstreetn5=ADD1C3;
	61	(paonfstreetn5 OR paonstreet1)=ADD1C;
	62	For property type in EPC is Flat/Maisonette, paonstreet3=ADD1632;
	63	(paonapstreet1OR paonapstreetn5)=ADD12C1
	64	paonn2 OR paonstreet4)=ADDC3
	65	paonstreet3=flADD
	66	paonn2=ADD2611
	67	paonstreet3=flADD13
	68	paonstreet3=ADD13C2
	69	(paonfstreetn5 OR paonn2 OR paonstreet2)=ADD1C2
	70	paonf1streetn6= ADD12
	71	paonapstreet2=ADD12C2
	72	paonfstreetn6=ADD12C;
	73	For the add in EPC is not start with ‘number string, number stirng’ pattern, flpaon3streetn5=ADDC10
	74	paonstreet2=ADD5
	75	paonstreet2=apADD1
	76	paonapstreet3=ADD13C2
	77	paonstreet3=ADDr66
Stage 4	78	saonpaonstreet2=ADDRE
	79	saonpaonstreet2=ADD12
	80	saonpaonstreetn=ADDC
	81	saonpaon65street=add12C;
	82	saonpaon62streetn2=ADD13C
	83	saonpaonstreetn=ADD6
	84	saonpaonstreetn=ADDCC
	85	saonpaon61street=ADD12C2
	86	saonpaon61street=ADDREC
	87	saonpaon62streetn=ADD7
	88	saonpaonstreet1=ADD13C2
	89	saonpaon1=ADD1C9
	90	saonpaonn=ADDC4
	91	paonstreetn=ADDC4 and saon=‘FLAT’;
	92	paonstreetn=ADDC4 and saon did not contain ‘FLOOR’,’UPPER’,’BASEMENT’, ‘LOWER’ ,‘FLAT ’ or any number string
	93	For the property type is not Flats/Maisonettes;paonstreetn=ADDC4 , then delete keep the successful linkage whose properety type in EPC is not Flat or Maisonette
	94	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’ :saon2paon61street= ADDCC; fldsaonpaonstreet1=ADDREC
	95	saonpaonn=ADD12C;
	96	For propertytype in house price data is ‘F’, flsaonpaon=ADD; flsaon1paonstreetn2=ADDCC; flsaonpaonstreet1= ADDREC; flsaonpaon62street1 = ADDREC; saon7paonstreet1=ADDRE; saon7paonstreet2=ADDREC;saon7paonstreet2=ADD12C2 ; apsaonpaonstreet1=ADD12C2
	97	For propertytype in house price data is ‘F’ and SAON start with number string, apsaonpaonstreet1=ADDREC
	98	saon7paonstreetn=ADDC4
	99	saon7paonn=ADD12C4
	100	saon4paonstreetn=ADDC4
	101	For propertytype in house price data is ‘F’, PAON start with number string ,saon4paonstreetn1=ADDC4

	102	apsaonpaon6streetn=ADDC4
	103	For propertytype in house price data is ‘F’, flsaonpaonstreetn=ADDC4
	104	For the PAON start with number string , saon4paonstreetn3=ADDC5
	105	saon4paonstreetn4=ADD12C
	106	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’, saon4paonstreetn1=ADD12C
	107	For propertytype in house price data is ‘F’: saon1paonstreetn=ADDC; saon1paonstreetn=ADD12C; saon1paonstreetn2=ADDC3; saon1paonstreetn2=ADD12C1; saon2paonstreetn3=ADDC; saon2paonstreetn3=ADD12C
	108	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’, saon2paon61street=ADD12C
	109	For propertytype in house price data is ‘F’ and PAON start with number string: saon2paonstreetn2=ADDC; saon2paonstreetn2=ADD12C
	110	saonn2paonn1=ADDC
	111	saonpaon62street=ADD12C
	112	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’, saon2paonstreetn=ADD12C
	113	saonn3paonnstreet=ADD13C
	114	saonn2paonn1streetn=ADDC
	115	saonpaon62streetn1=ADDC
	116	saon1paonstreet6n=ADD12C
	117	For propertytype in house price data is ‘F’: paon62saonpstreet=ADDRE; saon2paonstreetn4=ADDC; saon2paonstreetn4=ADD12C
	118	For propertytype in house price data is ‘F’: saon2paonstreetn4=ADD1num2 and ADD1 in EPC does not contain a character pattern that consist of number strings with a character
	119	For propertytype in house price data is ‘F’: saon5paonstreetn1=ADDC; paonsaon2streetn=ADD1C; saonpaon61streetn=ADDC; saon2paonstreetn2=ADD13C
	120	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’, saonpaon62streetn2=ADD13C
	121	saonpaon66street=ADDC6
	122	For propertytype in house price data is ‘F’: saon1paonstreetn3=ADD12C
	123	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’: saon2street=ADDC; saon2paonlo=ADDRE; saon2paonstreet=ADD12
	124	saon1paonstreet=ADDRE
	125	saon1paon=ADD12
	126	For propertytype in house price data is ‘F’: saon1paon61street=ADD12; saon1paon1=ADD1; saon1paonstreetn2=ADD12C2; psaonpaonstreet=ADDRE
	127	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’: saon2paon62street=ADD12
	128	saon2paonstreet=ADD1262
	129	saonpaonstreetn2=ADD7
	130	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’: flsaonpaonstreet=add1f61f2; psaon8street=ADDREC; saonpaonstreet1=add12643
	131	For propertytype is not F,saonstreet=ADDRE
	132	saonstreetlo= ADDRE
	133	unsaonpaonstreet2=ADDRE
	134	For propertytype in house price data is ‘F’: flsaonpaonstreet2=ADD8; saon7paon6street=ADDRE; saon7paon6street=ADD12
	135	For property type in EPC is Flat/Maisonette and for propertytype in house price

		data is ‘F’: flsaonpaon1=ADD1C9; saonpaon1=fladd; saonpaon1=fladd1c; saonpaonstreet3=fladd
	136	saon8paonstreet2=ADDRE
	137	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’: saonpaonstreet2=fladd
	138	PAON start with number string, paonlo=add12;
	139	For propertytype in house price data is ‘F’: saonpaonstreet1=adddap; saonpaon2=fladdc; saonpaonstreet11=ADD12; saonpaon61street=ADD1262c and paon62 contain ‘-’ string; saonpaon61street=ADD1262c and add261 contain ‘-’ string; flsaonpaonstreet3=ADD12C5; flsaonpaonstreet4=ADD12C1; saonpaon62steet=ADDC7
	140	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’: saonpaon61street=fladdc; saonpaonstreet5 =apadd1632; saonstreet1=ADD1C7; saonpaonstreet1=add1f61f2; saonstreet2=ADD1264
	141	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’ and SAON start with number string,flsaonpaon61street=ADDREC
	142	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’: saon4paonstreet=ADD12
	143	For propertytype in house price data is ‘F’: saonpaon61street1=ADD1262; flsaonpaon1=ADDRE
	144	saonpaon1=ADD1
	145	For propertytype in house price data is ‘F’: saonstreet3=ADDC; flsaonpaon2=ADD12; flsaonpaonstreet5=ADD1262; saonstreet=ADD1264
	146	paonsaonstreet=ADDRE
	147	For propertytype in house price data is ‘F’: saonpaon61=ADD12C; saon7paon=ADD12C; paonsaonstreet1=ADD12C; flsaonpaon61street=ADD12
	148	For propertytype in house price data is ‘F’ and SAON start with number string, apsaonpaon=ADD12; apsaonpaon62street1=ADDC8
	149	saon1paon62street=ADD12C
	150	For propertytype in house price data is ‘F’ and PAON does not start with number string, For property type in EPC is Flat/Maisonette, saonstreet=ADDC5
	151	saonpaonstreet2=ADDRE 187
	152	For propertytype in house price data is ‘F’: saon2paonstreet1=ADDC9; apsaonpaonstreet2=ADD1262cc; psaonpstreet=ADDRE
	153	saonpaon65street1=ADD12C
	154	For propertytype in house price data is ‘F’: saon2paonstreetn3=ADDC
	155	saonpaonn=ADD12C
	156	saon1paonstreetn1=ADDC
	157	For property type in EPC is Flat/Maisonette and for propertytype in house price data is ‘F’, saon4paonstreetn1=ADDC
	158	For propertytype in house price data is ‘F’: saon1paonstreetn=ADDC
	159	saonpaonlon=ADDC
	160	For propertytype in house price data is ‘F’: saonpaon65street1=ADD12C