Enriching address-based data with UPRNs

Learning from Domestic EPC and Land Registry PPD datasets

Bin Chi, Nick Bailey, Mark Livingston, Adam Dennett

08 June 2022

Urban Big Data Centre

casa

About the Authors:

Bin Chi is a Research Associate at the Urban Big Data Centre at the University of Glasgow

Nick Bailey is a Professor of Urban Studies and Director of the Urban Big Data Centre at the University of Glasgow

Mark Livingston is a Senior Lecturer in Urban Studies and Associate Director of the Urban Big Data Centre at the University of Glasgow

Adam Dennett is a Professor of Urban Analytics at the Centre for Advanced Spatial Analysis at University College London

In this short study, we explore some of the challenges in working with address-based data.

Addresses are captured in many ways and with much scope for error or uncertainty: in the (mis-)spelling of place names; in the use of abbreviations or punctuation; in the separation of building and street name elements; or the writing of flat numbers, for example. This can make it difficult to link information from different sources together or to accurately identify the location of properties. Tagging address-based data with the standard reference number, the Unique Property Reference Number (UPRN), would support greater accuracy and efficiency in linking and analysing data. We explore how easy it is to attribute UPRNs to address information, looking at two property datasets in England and Wales, one of which has already had UPRNs added by the data owner.

Key points/summary

1. One dataset, the Domestic Energy Performance Certificates (Domestic EPCs), has had UPRNs attached to 93% of cases by the data owner. We show that this can be improved using a more extensive rules-based approach, up to a 96% match rate, albeit with increasing effort or diminishing returns.
2. With the second dataset, the Land Registry (LR) Price Paid Data (PPD), the same rate can be achieved (96%) but with less time and effort (and fewer rules) because the address information is more structured. However, this required the development of a new rules-based approach. The rules used to process one dataset could not simply be transferred to the other because of the differences in the ways that address information was captured.
3. In both datasets, there is a small proportion of addresses where no match will ever be possible because the address information is incomplete or incorrect.
4. Flats present a particular problem when it comes to identifying a UPRN. One factor here is that flat numbers can be written in several different ways with no common standard.
5. For some properties, UPRNs have a hierarchical structure with a 'parent' UPRN for the whole building or block, and 'child' UPRNs for individual units or flats within it. Errors can occur when the wrong level of UPRN is attributed to a property.

We make several recommendations for improving the capture of address-based information so that UPRNs can be more successfully added in future. We also provide two rule-based approaches which can be further developed in future.

Introduction

The current landscape for property-based data is fragmented by the lack of standardised address identifiers and the inconsistencies in how address information is captured. This creates a barrier to the linkage of data related to the same address, making research and analysis difficult. Even important national property datasets like the Department for Levelling Up, Housing and Communities (DLUHC) Domestic Energy Performance Certificates (Domestic EPCs) or the Land Registry (LR) Priced Paid Data (PPD) store address information using different conventions or structures. Combining data on house sales or rents with energy performance, and perhaps with energy usage, age of the property and other locational characteristics could help fill some of the most pressing gaps in our understanding of housing, but it is very challenging without standard identifiers. The problem is by no means unique to the UK.[1]

Much personal or individual data also contains address information which can be a valuable means to link related data. In Scotland, for example, address information within health records has been used to identify household relationships which are otherwise largely absent from these systems.[2] Similar work has been undertaken in Wales.[3] Here too, successful linkage depends upon the quality of address information and how it is structured.

We are aware of several different attempts to tackle the address-to-UPRN matching problem. As well as the deterministic approaches used by work mentioned in the previous paragraph, others have applied fuzzy matching techniques or a combination of the two. In this work, we focus on our own deterministic algorithm.

The current UK's geospatial strategic vision is to build up "a coherent national location data framework" by 2025. To this end, and in response to the Open Standards Board's call, OS have made Unique Property Reference Numbers (UPRNs) and Unique Street Reference Number (USRN) available under an Open Government Licence. Nowadays, UPRN and USRN are "the public sector standard for referencing and sharing property and street information"[4]. The change to both these geospatial identifier numbers has increased the research focus in this area, particularly in an academic context where UPRN and USRN linkage to historic data significantly extends the prospects for longitudinal build environment analyses related to, for example, energy performance improvements or the dynamics of housing affordability.

Ordnance Survey (OS) have a well-managed system which oversees the production of UPRNs for every addressable location in the UK, both residential and commercial. However, few systems which collect address-based data use these UPRNs from the outset. Instead, address information is stored in a variety of more or less structured ways. Unfortunately, there is no consistent format for the recording of address data in the UK, which makes accurate attribution of a UPRN difficult. Flats are

[1] One US example of a commercial service to provide standardised identifiers - https://www.placekey.io/blog/the-most-common-address-standardization-problems-and-what-you-can-do

[2] Clark D, Dibben C. A guide to CHI-UPRN Residential Linkage (CURL) file. Scottish Centre for Administrative Data Research and Public Health Scotland; November 2020. Available from: https://www.isdscotland.org/Products-and-Services/EDRIS/_docs/CURL-Report-November-2020.pdf

[3] Harper, G., Boomla, K., Robson, J., Stables, D., Ahmed, Z., Fry, R., and Dezateux, C. (2020) Allocating Unique Property Reference Numbers to Patient Addresses Using A Deterministic Address-Matching Algorithm: Evaluation of Accuracy, Match Rate and Bias, *International Journal of Population Data Science* 5 (5).

[4] Identifying property and street information: https://www.gov.uk/government/publications/open-standards-for-government/identifying-property-and-street-information

particularly problematic, with many ways of recording a flat's position in a building. Different organisations have developed or are developing a variety of methods to identify and add UPRNs to data, including fuzzy logic address matching, machine learning and rules-based approaches[5,6]. In Scotland, for example, the Improvement Service's Datahub offers a service for local authorities and others to attach UPRNs to any address-based information.[7]

This short project sets out to examine how address-based data can be integrated more effectively by identifying limitations of current address capture and making recommendations for the future. We will achieve this by applying rules-based approaches developed by one of the authors to two existing datasets, the Domestic EPCs and the LR PPD. We explore how far it is possible to reliably identify the correct UPRN. With the EPC data, we compare our results with those available from the data owner's efforts. The EPC data have recently had UPRNs attached retrospectively via an undisclosed algorithm containing a "combination of rules-based and machine-learning approaches"[8]. Since September 2021, EPC assessors have been asked to add UPRNs when creating a record.

Aims

- To apply rules-based approaches to attach UPRNs to Domestic EPC and LR PPD datasets in England and Wales.
- To compare matched UPRNs from DLUHC and our own methodologies, summarising and accounting for any differences.
- To understand what causes failed record linkage in Domestic EPC and LR PPD.
- To make suggestions for improvements in the recording of address-based data to improve future success rates in the processing of current address data.

[5] Office for National Statistics. ONS working paper series no 17 - Using data science for the address matching service. Available from: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsworkingpaperseriesno17usingdatasciencefortheaddressmatchingservice#authors

[6] Harper, G., Stables, D., Simon, P., Ahmed, Z., Smith, K., Robson, J., and Dezateux, C. (2021) Evaluation of the ASSIGN open-source deterministic address-matching algorithm for allocating Unique Property Reference Numbers to general practitioner-recorded patient addresses, *International Journal of Population Data Science* 6 (1).

[7] https://datahub.scot/home/

[8] https://news.opendatacommunities.org/energy-performance-certificates-now-include-uprn/

Attaching UPRNs to Domestic EPCs and LR PPD

We have taken two dynamic address-based datasets and sought to match each address to one (and only one) contained in OS's definitive list, the AddressBase Plus dataset. AddressBase Plus then provides the definitive UPRN.

The first is DLUHC's Domestic EPCs. We use the ninth version, published in November 2021. It records 21,857,699 Domestic EPCs between October 2008 and September 2021 in England and Wales[9]. Records are created by building surveyors when they conduct an EPC survey. Since September 2021, surveyors have been encouraged to include a UPRN when they register the EPC.[10] The second is the LR PPD which, at the time of our research, did not contain any UPRNs. The LR PPD was downloaded in March 2022. The dataset contains 26,883,169 transactions in England and Wales, between January 1995 and January 2022.

Both datasets record the property's full postcode and detailed address in a number of fields. However, the structuring of the address information is different. In simple terms, the Domestic EPC data for England and Wales contains three address fields with no consideration of the ordering or distribution of information between these. For example, flat position is often combined with the building number and street name in a single field.[11] With LR PPD, there are separate fields for the property name or number and for the street name, and a further field for a flat number where appropriate.[12]

To reduce the complexity that results from the UPRN's lifecycle and parent-child relationship[13], in this research we use the active UPRN in OS AddressBase Plus (Epoch Number 90) and then remove the parent UPRN where a property also has a child UPRN. As the address strings are differently structured, we create a separate rules-based linkage process for each dataset. More details are available in the full report but in summary:

- For the Domestic EPC addresses, we use a rules-based linkage method with 446 detailed matching rules; 336 of the rules are conducted at the postcode level, with the remaining 109 at postcode sector level or higher. After a series of customized cleaning processes to take care of one-to-many linkages, 96.31% of Domestic EPCs have been geo-tagged with one unique UPRN.
- For the LR PPD's address, we use a twelve-stage process with 142 matching rules all at the postcode level. After removing results which have a one-to-many relationship, 96.53% of PPD transactions have been geo-tagged with one unique UPRN.

Figure 1 shows the contribution of the top 20 matching rules for each linkage process to the cumulative match rate. The UPRN linkage for the LR PPD is easier than for the Domestic EPCs, with a higher level of matching achieved for any given number of rules. As noted above, one important

[9] The DLUHC's UPRN linkage has recently been updated https://twitter.com/owenboswarva/status/1535991296719986689?s=20&t=yy_wsHactUOGf905sF8MZg

[10] https://news.opendatacommunities.org/energy-performance-certificates-now-include-uprn/

[11] In Scotland, the EPC address data is more structured and therefore easier to link.

[12] Explanations of column headers in the PPD https://www.gov.uk/guidance/about-the-price-paid-data#explanations-of-column-headers-in-the-ppd

[13] UPRNs are created when buildings are constructed or undergo structural alterations affecting the number of units, and they are deleted when buildings are demolished or units combined. See: https://static.geoplace.co.uk/downloads/The-UPRN-lifecycle-V3-2015.pdf

reason for this is that the structure used for recording address information in the PPD is more closely aligned to the structure used in OS AddressBase Plus.

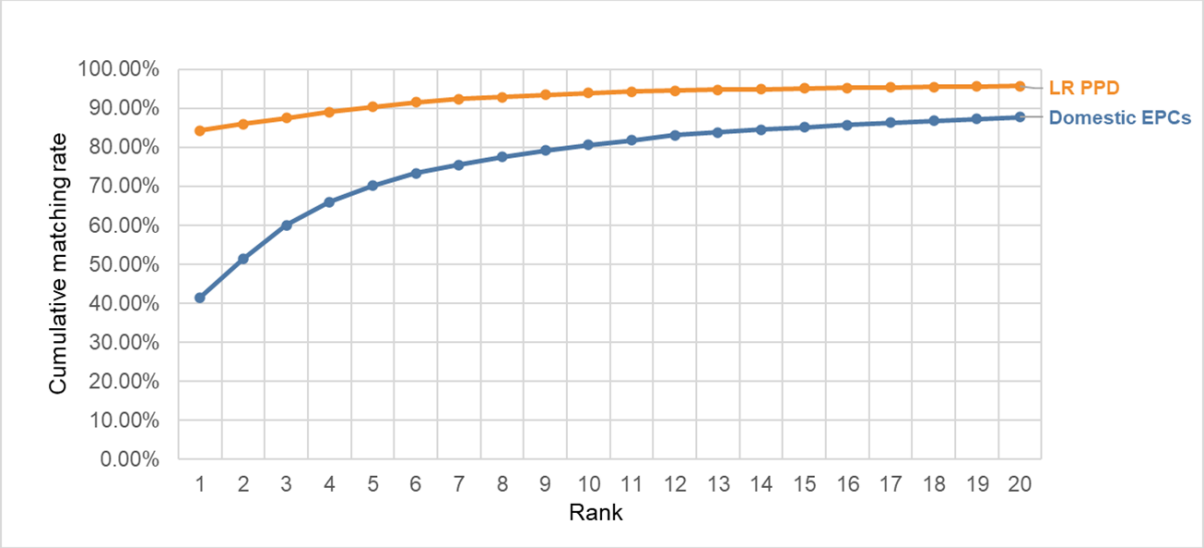

Fig 1. Cumulative match rates by number of matching rules for PPD and Domestic EPC datasets

Domestic EPCs

Figure 2 summarises the difference between the original DLUHC UPRN linkage and that using the UBDC method. DLUHC published the ninth version of the Domestic EPC dataset reporting a 92.52% successful linkage rate. Using the UBDC linkage method, the total match rate is 96.31%.

Overall, there is a great deal of consistency between the two. DLUHC attached a UPRN for 92.52% of EPCs. In the overwhelming majority of cases, UBDC identifies the same UPRN (91.27% of all EPCs). There are 0.70% of EPCs where DLUHC identifies a UPRN but UBDC does not, and 0.55% of cases where the two methods provide a different result. Of the remaining 7.48% of EPCs where DLUHC does not identify a UPRN, UBDC identifies one for 4.49% of cases. That leaves 2.99% of EPCs where neither method identifies a unique UPRN – equivalent to 654,085 records.

For the 0.70% of EPC records where DLUHC attach a UPRN but UBDC's method does not, the majority (54%) had multiple UPRNs. In these cases, DLUHC chose to allocate one of these to the address but UBDC chose not to allocate any since there did not appear to be one unambiguous match. For a further 12%, DLUHC was able to allocate a historical UPRN by using the OS AddressBase Premium database. UBDC was limited to using current UPRNs from AddressBase Plus. For the remainder (34%), DLUHC's method appears to be successful where UBDC's method is not. There are likely to be lessons which each approach can learn from the other.

For the 0.55% of cases where the two methods provide a different result, in the great majority of these cases (91%), the records had a parent-child relationship. DLUHC attached the parent UPRN while UBDC allocated the child UPRN.

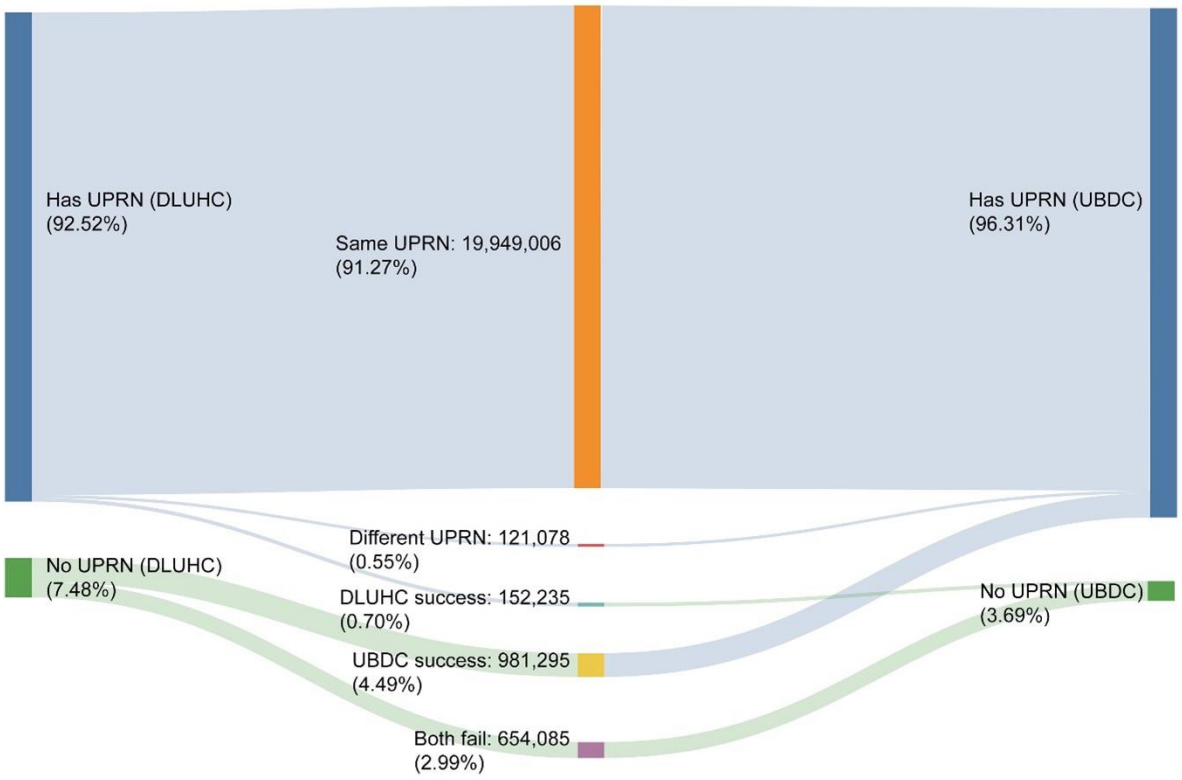

Fig 2. DLUHC and UBDC UPRN linkage comparison

LR PPD

The UBDC matching rate for the LR PPD is 96.53%. The PPD dataset includes property type which allows some disaggregation of results (Figure 3). With houses (Terraced, Semi-detached and Detached), UPRNs are successfully linked in at least 97% of cases but with Flat/Maisonettes, the matching rate falls to 93%. The main reason for the lower linkage rate for flats and maisonettes is the number of different ways in which flat number can be recorded ('Flat' or 'Apartment'; '9A' or '9/A' or '9, Flat A' and so on). With 'Other' properties, the linkage rate is lower still (65%), but these make up a very small proportion of the total. The reasons for this are complex but appear largely to be due to the variety of different types of properties, including caravans, lodges and barns, but also land parcels covered by the category.[14] Address information here may not easily fit the standard structures or there may be no addressable property.

[14] https://www.gov.uk/guidance/about-the-price-paid-data#explanations-of-column-headers-in-the-ppd

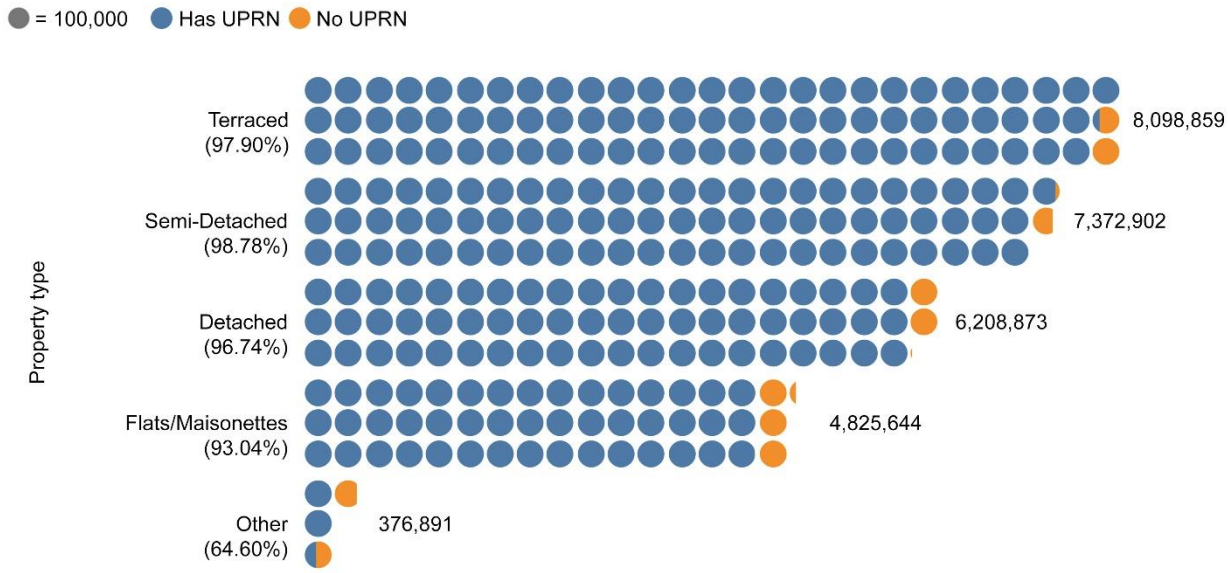

Fig 3. Matching rates by property type

Key problems and solutions

This exercise in linking UPRNs and making comparisons with the results of DLUHC's methodology provides insights into some of the key problems or issues for linking address data. Here we highlight some of these issues and the potential solutions. A more detailed discussion is provided in the main report.

Improving address capture

The low quality of the information in the address data is the main obstacle to achieving higher rates of accurate UPRN linkage. Controlling and standardising the quality of address curation in property-based databases offers the most effective way of improving UPRN linkage accuracy.

We offer four suggestions on how this might be achieved.

- First, it would be helpful to have guidance from GeoPlace on how the UPRN system works in terms of the lifecycle of UPRNs and the structures of parent and child relationships so that different data owners can be clear what their 'target' UPRN is in any situation.
- Second, data collection should be based on selection from fixed address lists which are derived from the OS AddressBase system so that UPRNs can be attached at the point of creating the records, not added through address matching. There will inevitably need to be arrangements for manual entry, not least to deal with lags between the creation of UPRNs and data capture systems being updated.
- Third, to improve manual entry, data owners should agree a standard methodology for collecting and storing address data. Both datasets we examined have examples of address components at the same level (like street name) which are recorded in different fields. To link data that is inconsistent in this way requires specific matching rules to be created to overcome the problem. Data validation and cleaning should be applied consistently to deal with e.g., punctuation and abbreviations.

- Fourth, OS should continue to work to improve the AddressBase product. While the quality of addresses stored in the AddressBase products is very high and is the industry standard, there are still some remaining errors or inconsistencies in the data. These are explained in detail in the full report, but the main issues are to do with inconsistencies in where some information, such as building or locality names, are stored. There are also questions about missing addresses or properties.

Improving address matching rules

Although perfect matching with address-based data is unlikely to ever be possible, there is more that can be learnt from comparing approaches which have been developed within different organisations. A confidence measure for the linkage would also be a useful addition, helping inform users of potential limitations in the data.

We discuss several detailed issues with address matching in the full report. One general observation here is that there does not appear to be an agreed way of dealing with parent-child relationships with UPRNs. This was one of the main sources of difference between DLUHC and UBDC/UCL methods in the assigned UPRNs. There should also be rules on how to deal with retired properties, i.e., the extent to which the target is the current or the contemporary UPRN.

Contact for further information

Bin Chi (bin.chi@glasgow.ac.uk)

Mark Livingston (mark.livingston@glasgow.ac.uk)

Funding and acknowledgements

This work is funded by the Ordnance Survey.

We gratefully acknowledge the support from the Ordnance Survey to do this work.

Image on front page: Benjamin Elliott on Unsplash.com