

**Grant Number**: CSC201708060184

**Sponsor:**

**Project title**: *A new attribute-linked residential property price dataset for England and Wales, 2011 to 2019*

The following files have been archived:

|  |  |
| --- | --- |
| File name | File description (Short description of content, sample size, format, any linking between different types of data, i.e. survey and interviews/focus groups) |
| [tranall2011\_19.csv](https://reshare.ukdataservice.ac.uk/854240/1/tranall2011_19.csv) | This is the new linked dataset from this project. It details 5,732,838 transactions in England and Wales along with each property's total floor area and number of habitable rooms. It also includes a new unique identifier (i.e. id) and all the other non-address fields (except LMK\_KEY field) in the Domestic EPC dataset. This dataset contains 105 fields written in upper and lower case. All the fields written in upper case come from the Domestic EPC dataset. Explanations of fields are provided in the Github (https://github.com/BINCHI1990/Link-LR-PPD-and-Domestic-EPCs) |
| epc\_id.zip | This is the non-address domestic EPC dataset, created by deleting the delete LMK\_KEY,ADDRESS, ADDRESS1, ADDRESS2, ADDRESS3 and POSTCODE fields from original Domestic EPCs dataset. A new unique identifier (id) created during the data linkage is included in the dataset, allowing us to link properties in the EPC dataset back to the linked data. The whole dataset contains 18575357 records with 79 fields. |
| pp-complete.csv | This is the original Land Registry PPD dataset with field names. It records 24,852,949 transactions in England and Wales between 1/1/1995 and 31/10/2019. |
| Read\_EPC\_csv.R | The is the R code to read all the EPC dataset into R and save the result data in PostGIS database . |
| Read\_LR\_PPD.sql | This the SQL code used in PostGIS to read all the Land Registry PPD into a datajournal database |
| Read\_NSPL.sql | This is SQL code used in PostGIS to read in the NSPL dataset |
| Data\_cleaning.sql | This the SQL code used in PostGIS to  clean the Land Registry PPD dataset before the linkage between Land Registry PPD and EPCs. |
| PPD\_EPC\_linkage.R | This is the R code to achieve the linkage between Land Registry PPD and EPCs. |
| Evaluation.R | This is the R code to evaluate the performance of data linkage and identify for which years information losses through matching are relatively small. |
| Data\_cleaning.R | This is R code used to clean up the untenable values in total floor area or number habitable rooms identified after conducting the linkage between Land Registry PPD and EPCs. |
| Output\_Area\_2011\_to\_Builtup\_Area\_Subdivision\_to\_Builtup\_Area\_to\_Local\_Authority\_District\_to\_Region\_December\_2011\_Lookup\_in\_England\_and\_Wales.csv | This is National Statistics Postcode Lookup (NSPL) dataset (November 2019). It is used to geo-reference the linked data during the linkage process. |
| rulechi.csv | This CSV file is used in PPD\_EPC\_linkage.R code. It records the mismatch address strings identified for the same property between Land Registry PPD and Domestic EPC datasets, such as 'WOODLANDS PARK' VS 'WOODLAND PARK' . |
| Readme.docx | This is the file to describe the datasets deposited in the UKDA. |

**Publications**:

Chi B, Dennett A, Oléron-Evans T, et al. (2020) *A new attribute-linked residential property price dataset for England and Wales, 2011 to 2019*. Scientific Data (accepted).