Data Documentation: NIC project

Mateusz Mysliwski, Lars Nesheim, Polly Simpson

Centre for Microdata Methods and Practice
IFS

March 5, 2018

Abstract

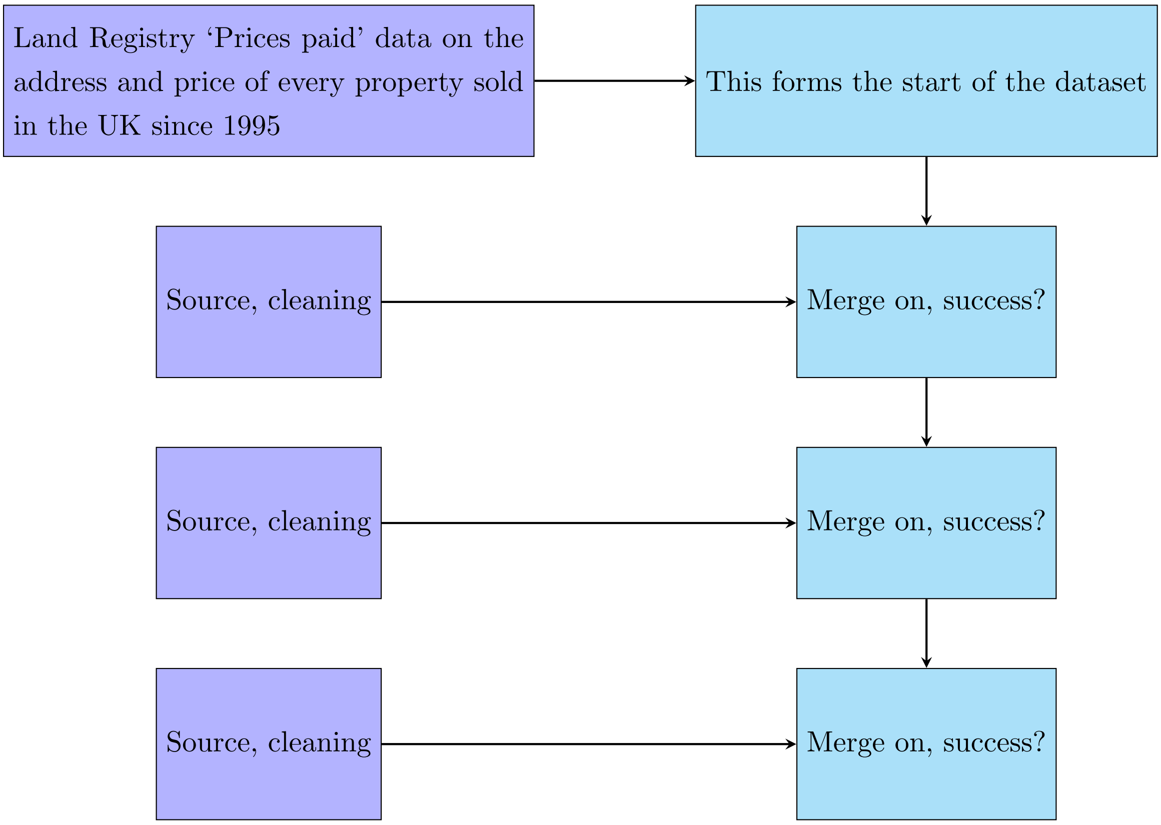
This documents provides details on the data used for the Land Value Tool.

Contents

1 Model overview including explanation of regions

1.1 Model regions

2 Overview of data construction


Land Registry 'Prices paid' data on the address and price of every property sold in the UK since 1995
This forms the start of the dataset
Source, cleaning
Merge on, success?
Source, cleaning
Merge on, success?
Source, cleaning
Merge on, success?

3 Data sources

3.1 Land Registry Price Paid

Brief description: Covers (almost all) residential property sales in England Wales from 1995 to the current day. Exceptions include properties sold or transferred for below market value (e.g. right to buy sales at a discount, gifts, compulsory purchase orders)

Variables: Information on transacted price as well as property characteristics, including:

- **pricepaid**: Sale price stated on the transfer deed
- **propertytype**: D = detached, S = semi-detached, T = terraced, F = flat, O = other
- **newbuild**: Y = new build, N = established residential building
- **tenure**: F = freehold, L = leasehold. (Leasehold is very strongly correlated with being a flat)
- For more information see: https://www.gov.uk/guidance/about-the-price-paid-data.

Source and version used: Available from gov.uk https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads. The model uses data from 2008 to November 2017.

Download details: Data is released monthly, but 'year-to-date' available in a single file. Although the data is available as a single file containing all years, it's advisable to download only a year at a time as processing data files of that size can be slow. Yearly files are collated on the date of the transaction/deed date rather than the date that the information was lodged with HM Land Registry.

License conditions: Contains HM Land Registry data©Crown copyright and database right 2017. This data is licensed under the Open Government Licence v3.0.

Limitations: Data doesn't contain sales of land or non-domestic property.

3.2 Addressbase Plus

Brief description: Addressbase Premium is the Ordnance Survey's flagship database of Addresses in the UK.

Variables: AB contains unique identifiers (TOIDs) which can be used to link addresses to OS Mastermap topography and transport layers. Addresses are structured on the basis of PAO/SAO (explained below). For example:

- pao_start_number - number on the street that the building is e.g. 15 Newbridge Road, comes after pao_text
- pao_start_suffix - e.g. the 'A' bit of 15A Newbridge Road
- pao_end_number - e.g. the 16 bit of 15-16 Newbridge Road
- pao_end_suffix

- pao_text - e.g. the 'Radley House' bit of Radley House, 15A-16 Newbridge Road

Source and version used: AddressBase Plus June 2017

Download details: Data for England comes in 36 csv files (15.4GB).

License Conditions: Normally, access to AddressBase Premium is subject to a fee. However it is available as part of the public sector mapping agreement.

Limitations:

Notes:

3.3 Energy Performance Certificate database

Brief description:DCLG energy performance certificate data is based on assessments of energy efficiency (EPC) which are mandated for almost all properties in England and Wales when constructed, sold or let. Since 2008 assessors have been required to submit the underlying data for these assessments to DCLG. DCLG publishes this data, except where:

- the holder of the energy certificate has 'opted-out' of disclosure
- energy certificates are excluded on grounds of national security
- energy certificates are marked as 'cancelled' or 'not for issue'
- DECs that can be identified as 'voluntary' (some organisations choose to have a DEC produced even though they are not required to do so by the regulations) are also excluded

Variables: As well as address and assessment data, the data also contains information on:

- Total floor area
- Window glazing
- Heating systems
- Energy efficiency
- Story of the building that the property is on

Source and version used: Available to download here:

https://epc.opendatacommunities.org/. Last downloaded for the project 01/12/2017.

Download details: Bulk download comes as one file per local authority.

License Conditions: The data is at the address level and is therefore considered personal data subject to the Data Protection Act. The guidance explicitly gives permission for the use of this data for research purposes, does not place specific restrictions on linking this data, and DCLG have confirmed that we don't need permission to link the data.

Limitations: The address field is structured in a different way to Land Registry/OS data.

Notes:

3.4 ONS Local authority to Region lookup

Brief Description: Geographic lookup from local authority code to statistical region name and code.

Source and version: Local Authority District to Region (December 2015) Lookup in England. https://ons.maps.arcgis.com/home/item.html?id=17a30c262cf44ad0b9780d97c5d4a856

Download Details: Available as csv file.

3.5 National Statistics Postcode Lookup (NSPL)

Brief Description: Geographic lookup from postcode to most other levels of administrative/statistical geography in the UK. Includes population-weighted centroids for each postcode.

Source and version: Available to download from ONS Open Geography Portal. Latest version (as used for project) is November 2017.

http://geoportal.statistics.gov.uk/datasets/national-statistics-postcode-lookup-latest-centroids

3.6 Local plans

Brief Description: Local authorities are required to produce local plans for future housing development. DCLG have collated details of the published plans by local authority. The data includes the total number of planned houses over the lifetime of each authority's plan, and the annual equivalent number of houses.

Source and version: Correspondance with DCLG (now MHCLG). Current as of June 2017.

3.7 Number of dwellings by tenure and district, England

Brief Description: Total number of dwellings by local authority district in England.

Source and version: Table 100 from the DCLG live tables on dwelling stock. Current as of July 2017.

https://www.gov.uk/government/statistical-data-sets/live-tables-on-dwelling-stock-including-vacants

3.8 DCLG LA-level land constraints data land

Brief Description: DCLG statistics on the area of land in a local authority that is a) Greenbelt b) Built-up c) AONB/SSSI/National Park etc. Includes total land area for calculation of percentages.

Source and version: Correspondence with DCLG in July 2017. Current as of 2015.

3.9 Generalised Land Use Database 2005

Brief Description: In 2001 and 2005, DCLG produced (as experimental statistics) the Generalised Land Use Database which classified all land in England into 10 land categories:

- Domestic buildings
- Domestic gardens
- Non-domestic buildings
- Roads
- Paths
- Rail
- Greenspace
- Water
- Other land uses (largely handstanding)
- Unclassified

The data is available as multiple geographical levels: Census Output Areas (OAs), Lower Layer Super Output Areas (LSOAs), Middle Layer Super Output Areas (MSOAs), Local Authorities (LAs), and Government Office Regions (GORs). The area of coverage is England.

Source and version: Census ward level data for 2005 downloaded from the national archives

http://webarchive.nationalarchives.gov.uk/20060206095010/http://www.odpm.gov.uk/index.asp?id=1146084

Download details:Available as spreadsheets that give the area of land (thousands of square metres) in each administrative area. I think it's also available in a way that means you can do more specific mappings but that you have to already have a public sector license to use OS MasterMap.

Notes: You can find useful guides to the data here:

http://www.andywightman.com/docs/GLUD_2005.pdf

http://www.andywightman.com/docs/GLUD_2006.pdf

3.10 Population Density

Brief Description: Population density at the output area level collected for the 2011 Census

Source and version: Downloaded from NOMIS (ONS official labour market statistics portal)

https://www.nomisweb.co.uk/

3.11 Index Multiple Deprivation

Brief Description: The index of multiple deprivation (IMD) is a measure of very local level deprivation across multiple dimensions including health, income and education. Various summary measures of the IMD are available. For this analysis we use IMD decile at the output area level.

Source and version: IMD 2015 data downloaded from NOMIS (ONS official labour market statistics portal)

https://www.nomisweb.co.uk/

3.12 Open Flood Risk

Brief Description: "Open Flood Risk by Postcode combines the Environment Agency's Risk of Flooding from Rivers and Sea with English postcodes from Open Postcode Geo. Each postcode is supplemented with the information from the flood risk area in which it is located. Postcodes are treated as a single point in space, allowing for the postcode to be looked up and the flood risk determined for that point. Note the value of SUITABILITY for a given area for how suitable the risk level is for a given purpose."

Categorizes floodrisk into the following categories, recorded in the prob_4band variable:

- High - Each year, there is a chance of flooding of greater than 1 in 30 (3.3%).
- Medium - Each year, there is a chance of flooding of between 1 in 30 (3.3%) and 1 in 100 (1%).
- Low - Each year, there is a chance of flooding of between 1 in 100 (1%) and 1 in 1000 (0.1%).
- Very Low - Each year, there is a chance of flooding of less than 1 in 1000 (0.1%).
- None - This value is added by GetTheData to indicate a postcode which is not in a flood risk area.

Source and version: OpenFloodRisk available https://www.getthedata.com/open-flood-risk-by-postcode, current as of Febuary 2016. Details of original floodrisk information are available https://www.getthedata.com/docs/RoFRS_Product_Description_vs1.5.pdf

Download details: Available as a single CSV file, one observation per postcode

License Information: "Derived from Risk of Flooding from Rivers and Sea. Attribution required: Contains Environment Agency data licensed under the Open Government Licence v3.0. Derived from Open Postcode Geo (https://www.getthedata.com/open-postcode-geo) which requires attribution."

3.13 Greenbelt

Brief Description DCLG produce a greenbelt dataset, in polygon form, of greenbelt boundaries in England. They do this by merging together information from different local authorities. They caveat that due to differences in local authorities digitisation methods the accuracy of the given boundaries may vary.

Source and version: Latest data available here: https://data.gov.uk/dataset/english-local-authority-green-belt-dataset4. We use latest data available as of June 2017.

Download Details: Download is a shapefile (.shp).

3.14 Road Noise

Brief Description: The EU Environmental Noise Directive requires the UK government to measure the noise exposure of large settlements and the area around major roads, railways and airports.[1] This was last undertaken in 2012. The Environment agency publishes the resulting assessment as open data. Multiple measures of road noise exposure are available, we use LAeq, 16h - an indicator of the annual average noise levels for the 16-hour period between 0700 2300 (in db).

Source and version: Details about the requirements and data, as well as download links are available here: https://www.gov.uk/government/publications/open-data-strategic-noise-mapping.

Download Details: Data is available in TAB, GML or shapefile format (we use shapefile).

3.15 Google Maps API

Brief Description: Google Maps Distance Matrix API allows users to request travel times between a pair of co-ordinates or addresses, with varying assumptions on the travel mode used and traffic conditions. Requests can be made using various programming languages including Python and R. We use the R package [insert name] to make geo-coding and travel time requests. More details on how we use the google maps data are in section [x] on travel time estimation. Google's online documentation can be found here: https://developers.google.com/maps/documentation/distance-matrix/

We'll need to add a whole section on how the travel times model works

3.16 VOA Rating list

Brief Description: The Valuation Office Agency is responsible for valuing all non-domestic property in England and Wales in order to determine the tax liability of occupants (specifically, Business Rates, a local tax on business property). They maintain a database containing the estimated market rent of almost all non-domestic property in England and Wales. These esti-

[1] http://ec.europa.eu/environment/noise/directive_en.htm

mated rents are called 'rateable values' because they are used as the basis of business rates. The data covers 2 million commercial properties in England and Wales as of April 2017. Property coverage is almost, but not quite, complete. For example it doesn't cover:

- “Some types including hotels, public houses, universities, schools and hospitals are covered in the Rating List but usually have no floor area data.
- “Very large industrial hereditaments are valued separately, and are omitted from both Rating List and SMV.
- “Three classes of activity are exempt from rates altogether: agricultural premises, places of worship and properties of Her Majesty the Queen. This last group besides Royal palaces includes the Ministry of Defence estate, which is extremely large. Floor area data in all these cases must be estimated or obtained from other non-VOA sources. Some area data are available from Display Energy Certificates (DECs) and Energy Performance Certificates (EPCs) (Department for Communities and Local Government, 2015).”[2]

Version and Source: Data available from the VOA. If you have agreement from VOA data can be downloaded here: https://voaratinglists.blob.core.windows.net/html/rlidata.htm. The latest available data is based on the 2017 revaluation (release April 2017). Comprehensive property valuation carried out in 2010 and 2017, but incremental updates are carried out continuously as appeals take place.

Licensing conditions: VOA data has restricted licensing conditions (terms available her: https://www.tax.service.gov.uk/business-rates-find/terms-and-conditions). It cannot, legally, be downloaded and used for research without making a data sharing agreement with the VOA. For this project a data sharing agreement was made with the VOA. [ADD DETAILS OF LICENSING CONDITIONS]

Details of data structure: There are two available datasets of rateable values from the VOA:

The Rating List: Covers 2 million properties, giving addresses and primary activities. Each observation is a different property. There is only one observation per property.

The Summary Valuation database: Covers a smaller number of properties (100,000 fewer), giving sub-activities and floor areas for each of these sub-activities.

The Summary Valuation database is structured slightly strangely. V1 is an indicator telling you the type of record of the observation. These range from 01 to 07. A 01 record is the main record type, and is essentially the same as in the rating list. It tells you what assessment and what property subsequent observations are going to relate to so will be what we use to link the detailed info to the ratings list.

Subsequent records (before the next 01 record) all relate to this property. These are numbered 02 to 07. The vast majority of records are type 2. These tell you about each individual room of a 'property' (e.g. the most common are "Office", "Internal Storage", "Retail Zone A/B/C", "Staff Toilets", "Kitchen", "Workshop", "Store", "Warehouse"), what floor it's on, their floor space and contribution to the rateable value.

[2] Detail from Evans et al 2017 guide to 3DStock, http://journals.sagepub.com/doi/pdf/10.1177/0265813516652898

Type 3 records are 'additional items'. Includes things like bowling green, clubhouse, car wash, helipad, swimming pool etc.

Type 4 records are plant and machinery - only value is given.

Type 5 records are car parking which gives the amount and value of car parking.

Type 6 records are adjustments. These are explained here:

http://app.voa.gov.uk/corporate/publications/manuals/ratingmanual/ratingmanualvolume4/sect7/d-rat-man-vol4-s7-app1.html.

Basically, adjustments to value are often made at the end of valuation that reflect factors which will change the overall valuation of the property. The top factors here are listed below and are not insignificant determinants of property prices (median adj is -5%)

- Access
- Adjustment for Quality
- Divided or split unit
- Merged units
- Non-standard frontage
- Planning restriction
- Poor access
- Services
- Shape
- Shared access
- Size or quanitty allowance
- Toilet facilities
- Vairations in floor level
- Width to depth ratio

Type 07 gives you the total value of the property before adjustments and the cash value of the adjustment.

Notes: Guide to valuation https://www.gov.uk/guidance/how-non-domestic-property-including-plant-and-machinery-is-valued

4 Data construction - domestic

4.1 Glossary

Primary Addressable Object (PAO): The best way to store address information is not immediately obvious. The Ordnance Survey chose to break down addresses into their most basic building blocks - this creates a large number of variables but can be very precise. Central to their system are primary and secondary addressable objects (PAO/SAO respectively). Take the address Flat 4, 17 Stapleton Road, Bristol. Here the primary addressable object is 17. Stapleton road is the street or thoroughfare. Flat 4 is the secondary addressable object (as it is a subset of a larger building). PAO and SAO can both either be a number, or text, or both.

4.2 Clean prices paid

Brief summary: This file reads in prices data and restructures the address variables so that it can be matched to AddressBase.

More detail:

- Import prices paid data for 2008 to 2017
- Renames, labels and destrings variables
- Drop properties in Wales and garages/plots of land.
- Other than basic cleaning, the objective is to restructure the address data so it is structured as similarly as possible to the addressbase data. It won't be perfect because it's difficult to know how to break up the given information into the different components of an address. We focus on identifying, from the given address information, the primary addressable object (see definition above).
- Split postcode into postcode_left (the first word) and postcode_right (the second word)
- Rename the 'street' variable 'thoroughfare'.
- Clean each year at a time to get primary addressable object (PAO), noting whether the PAO is purely numeric or contains non-numeric characters (e.g. 'Flat' or 'House')
- Splits each year's dataset into two files, one containing properties with 'numeric' type PAO, one containing properties with non-numeric PAO.
- Appends the annual datasets into two files covering 2008-2017 (one with for numeric type addresses, one for non-numeric addresses)
- Note contains some code for cleaning SAO but this is commented out - delete.

Relevant do file: Clean_PricesPaid2.do

Input data file(s): Raw prices paid data from land registry (split by year)

Output data file(s):

- PricesPaid0817_vclean_nums.dta - All properties for 2008 to 2017 that have a numeric PAO
- PricesPaid0817_vclean_string.dta - All properties for 2008 to 2017 that have a string (non-numeric) PAO

4.3 Clean addressbase

Brief summary: We match the prices paid data to addressbase in order to get co-ordinates (latitude/longtitude) for each property. When we perform the match the addressbase data must be unique at the level of the variables we match on. Differences in the structure of addresses between the Land Registry and AddressBase mean that this match won't be perfect so try it twice - once when we try to match at the street number/building name (PAO) level and once when we simply match by street name and postcode. To facilitate that this do file makes 3 clean versions of address base - one unique at the pao number, street, postcode level. A second unique at the pao text, pao number, street, postcode level. A third unique at the street, postcode level.

Relevant do file: Clean_AddressBase_England.do

Input data file(s): Raw addressbase download from the Ordnance Survey - split into 36 files.

Output data file(s): 36 x 3 data files of AddressBase unique at different levels of address (house number, house name, street)

4.4 Add co-ordinates to land registry

Brief summary: Merges together PricesPaid and AddressBase data. Does 4 attempts at merging, two each for the number and string type addresses. One which uses the PAO info, one that just uses the street and postcode.

Relevant do file: M1_PPAddressBase.do

Input data file(s): AddressBase (clean), Price Paid data for 2008-2017.

Output data file(s): Merged data file, m1_PPAB.dta

Limitations: 2.9% of properties don't match and therefore don't have a set of co-ordinates. Solution: Most of the properties that don't match do have postcodes - could give these properties co-ordinates based on their postcode?

4.5 Create model-region lookup

Brief summary: Creates a look-up from 2015 local authority codes and names to 'model region' (region as defined for our model) and 'london plus' (london broadly defined). It starts from a normal LAD-¿ region lookup, then defines CaMKOx, Cornwall and Devon, and 'london plus'.

Relevant do file: LA_ModelRegion_Lookup.do

Input data file(s): Standard local authority to region lookup

Output data file(s): Adapted local authority to model region lookup

4.6 Add model regions to master data

Brief summary: Cleans the local authority district variable in the master data (which comes from the Land Registry) so that it is consistent with changes to local authority boundaries that happened in 2009. Then merges in model region lookup based on this local authority district variable (ladistrict).

Relevant do file:M2_ModelRegions.do

Input data file(s): Master data - m1_PPAB.dta, local authority to model region lookup

Output data file(s): Merged data file, m2_modelregions.dta

4.7 Clean Energy Performance Certificate data

Brief summary: From Energy Performance Certificate data we can get an accurate measure of property floor area. In order to match this to our existing data we need to make sure the address information is structured in a similar way and that there is only one observation per address in the EPC data. The address field in the EPC data is structured so differently to our other data sources that we have taken the approach of combining the whole address into one variable and matching on that. We also have to remove repeat observations for each property (e.g. old assessments).

More detail: There's one EPC file for each local authority. This do file:

- Loops over each local authority and cleans/labels the data.
- Drops any observations with zero total_floor_area.
- Creates split postcode, and a 'clean address' which is the address field without gaps, commas etc.
- Deletes observations so there is one per property (postcode, clean address combo).
- First keeps the most recent observations.
- Then drops randomly from within these if total_floor_area is the same for remaining properties.
- Then randomly drops from the remaining observations if there are still doubles.

Relevant do file:Clean_EPCRough.do

Input data file(s): Raw EPC data at the local authority level

Output data file(s): Cleaned EPC data at the local authority level

4.8 Add floor area from EPC to master data

Relevant do file: M3_EPC.do

Brief summary: Matches master data to the EPC data using an address variable that combines all house number/street name information into one string called 'clean address'. To take some account of differences in address structure the file makes three attempts to match the master data to the EPC with three different versions of 'clean address'. As the merge is done, the data is split by year, so the output is one merged file for each year.

Input data file(s): Master data - m2_modelregions.dta, EPC datafiles for each local authority.

Output data file(s): m3_epc.dta x 10 (one for each year)

Limitations: The match rate is around 80% in most years. Lower in 2008 and 2017.

4.9 Rearrange data into regional files

Relevant do file: M4_Model1RegionalData.do

Brief summary:Takes the yearly M3_EPC files and merges/splits them to create one file for each region in our model. It then makes three versions of the data

1. m4_regionclean: Has everything we need for further matching to model 2 data
2. model1_regionclean: Has just the variables needed for model one
3. coordinates_regionclean: Contains 1 observation for every set of co-ordinates in the data.

Input data file(s): Master data (m3) split by year

Output data file(s): Data files for each model region, covering all years.

4.10 Clean national postcode lookup

Relevant do file: Clean_PostcodeLookup.do

Brief summary: Renames and labels the latest national postcode-lookup. Keeps just the English postcodes.

Input data file(s): Raw National Statistics Postcode Lookup

Output data file(s): Clean postcode lookup, csv file of postcode centroids

4.11 Add output areas to master data

Relevant do file: M5_OutputAreas.do

Brief summary: Merges in output areas and other administrative geographies based on postcodes from the postcode database.

Input data file(s): Master data (m4) at the model region level

Output data file(s): Merged data (m5) at the model region level

Limitations: **Potentially not everything matched ==> double check**. New missing_oa variable indicates if it didn't match.

4.12 Clean data on local housing plans

Relevant do file: Clean_LocalPlans.do

Brief summary: This file cleans data on local plan numbers and the dwelling stock in each local authority.Then the two data sources are combined to calculate the 'local plan rate', which is the number of houses planned annually as a share of the 2012 dwelling stock in a local authority.

More detail: In order to use local authorities future housing development plans to measure the eagerness of different authorities to build, it seems sensible to express this as a share of the authority's existing housing stock. However, the plans run over different periods, so it's not clear which year to use as the base. Following consultation with DCLG we chose to use the 2012 dwelling stock as this was the median year in which the plans were adopted. This gives the 'local plan rate', which is the number of houses planned annually as a share of the 2012 dwelling stock in a local authority.

Input data file(s): Raw data on local plan numbers and the dwelling stock at the local authority level.

Output data file(s): One file - local plan rate at the local authority level.

4.13 Clean data on local land restrictions (e.g. green belt)

Relevant do file: Clean_DCLGLandConstraints.do

Brief summary: Labels and lightly cleans LA-level DCLG data on the share of land in a local authority with particular restrictions on construction. busyland is land that's already built on. restrictedland is AONB, NPs, SSSIs, Greenbelt. fz3 means also including floodzone 3.

Input data file(s): Raw data from DCLG on land constraints

Output data file(s): Cleaned version of the same data

4.14 Add local authority level variables to master data

Relevant do file: M6_LAcharacteristics

Brief summary:Merges based on local authority code to data on LA level land constraints and local plan rate.

Input data file(s): Master data (m5) at the regional level. Clean data on land constraints. Clean data on local plan rate.

Output data file(s): Merged data (m6) at the regional level.

Limitations: Uses the local authority code from the postcode lookup, not the local authority in the original land registry data, so only properties that successfully matched to the postcode lookup will match to the LA characteristics data.

4.15 Clean data on local land use

Relevant do file: Clean_GLUD.do

Brief Description: Firstly cleans and labels data from the 2005 generalised land use database. Then it calculates percentage figures for the share of land in each ward that has each use. Finally, because the data is at the 2003 census ward level, matches this to 2011 output area so that it can be merged to the master data. Note that because of the change in geography, the area (non-percentage) measures at the OA level are not accurate and shouldn't be used.

Input data file(s): Raw GLUD 2005 data at the ward level, 2003 ward to 2011 OA lookup.

Output data file(s): Cleaned GLUD data at the 2011 OA level.

4.16 Add local land use to master data

Relevant do file: M7_GLUD.do

Brief Description: Merges master data to the GLUD by output area (oa11cd).

Input data file(s): Master data at the regional level (m6), GLUD data at the OA level.

Output data file(s): Merged data at the regional level (m7)

Limitations: OA11CD comes from the postcode database, so only properties that match to that will match to the GLUD.

4.17 Clean Population Density Data

Relevant do file: Clean_PopulationDensity.do

Brief Description: Reads in raw data on population density, labels and lightly cleans. Saves in stata .dta format.

Input data file(s): Output area and lower super output area level data on population density from the 2011 census.

Output data file(s): Two cleaned data files, one at the OA level, one at the LSOA level.

4.18 Add population density to master data

Relevant do file: M8_PopulationDensity.do

Brief Description: Merges master data to population density data by output area (OA11CD) and lower super output area (LSOA11CD)

Input data file(s): Master data (m7) at the regional level. OA level data on population density. LSOA level data on population density.

Output data file(s): Merged data (m8) at the regional level.

4.19 Clean data on the index of multiple deprivation (IMD)

Relevant do file: Clean_IMD.do

Brief Description: Gets raw data on IMD at the LSOA level. Labels. Saves as .dta.

Input data file(s): Raw data on IMD

Output data file(s): Lightly cleaned data on IMD.

4.20 Add IMD to master data

Relevant do file: M9_IMD.do

Brief Description: Merges master data to IMD at the LSOA level.

Input data file(s): Master data (m8) at the regional level, clean data on IMD at the LSOA level.

Output data file(s): Merged data (m9) at the regional level.

4.21 Clean open flood risk data

Relevant do file: Clean_OpenFloodRisk

Brief Description: Labels and lightly cleans data from Open Flood Risk on flood risk by postcode.

Input data file(s): Raw data from open flood risk

Output data file(s): Cleaned data on flood risk by postcode

4.22 Add flood risk to master data

Relevant do file: M10_floodrisk.do

Brief Description: Merges master data to floodrisk at the postcode level.

Input data file(s): Master data. Cleaned floodrisk data.

Output data file(s): Merged data (m10) at the regional level

4.23 Use GIS for greenbelt and roadnoise measures

Relevant do file: N/A

Brief Description: Uses GIS mapping software to compare property co-ordinates to known boundaries of local authority greenbelt and EA assessments of roadnoise to ascertain whether these apply at those locations.

More detail:

- Create local authority boundary shapefiles for each region by joining the shapefile to the lad_modelregion lookup, toggle editing, dropping parts of the shapefile that have the wrong region. Note that these files are using British National Grid (BNG) based co-ordinates.
- Reproject greenbelt and road noise shapefiles (if necessary) to BNG.
- Create separate greenbelt/roadnoise shapefiles for each region using the 'clip' function (Vector/GeoProcessing/Clip, with input layer = Greenbelt and Clip Layer = region boundary)
- Make sure it the new clipped files are recorded as being in BNG
- Load in the co-ordinates of properties as a csv file
- Use 'join attributes by location' with target vector layer = coordinates and join vector layer = greenbelt. Keep all records (regardless of whether they match) and save the joined layer as a csv file. This creates a file with every pair of co-ordinates and the new characteristics variables (which equal NA if missing)
- NOTE there is no greenbelt in Cornwall and Devon so instead of following the above process it was sufficient to create a variable greenbelt = 0 for all properties in that region

Input data file(s): Greenbelt and EA roadnoise shapefiles, local authority boundary shapefiles, model region lookup, property co-ordinates.

Output data file(s): CSV file for each region with co-ordinates of all properties in that region as well as new greenbelt and roadnoise variables.

4.24 Clean new GIS variables

Relevant do file: Clean_QGISoutput.do

Brief Summary: Reads in CSV file outputted from QGIS, renames/labels variables and saves as a .dta file.

Input data file(s): Regional level CSV files from QGIS

Output data file(s): Regional level .dta files of co-ordinates, greenbelt and roadnoise variables.

4.25 Add greenbelt and road noise to master data

Relevant do file: M11_QGIS.do

Brief summary: Matches greenbelt status and roadnoise exposure to master data by property co-ordinates.

Input data file(s): Master data, .dta form of QGIS output (co-ordinates matched to greenbelt status and roadnoise exposure)

Output data file(s): Merged data for each region (m11) - the final files!

5 Data construction - non-domestic

All dofiles in P:\NICProject\stata\DataCleaning\NonDomesticModel

5.1 Clean basic VOA data

Relevant do files:

Clean_VOAEntryList.do

Clean_VOASummaryValuation.do

Clean_VOACombined

Input data file(s): Raw VOA entry list

M:\NICProject\ValuationOfficeAgency\originals\nondomesticrates_file1_2017.csv

Output data file(s): Cleaned entry list

M:\NICProject\ValuationOfficeAgency\clean\VOAEntryList2017.dta

Brief summary:

Cleaning Notes:

- Data is asterik delimited, must specify delimiter("*") when importing into stata.
- 1500 properties missing postcodes
- 18000 properties missing rateable values
- 8500 properties have postcodes but don't match to postcode database, mostly isolated postcodes but some clusters which could be cleaned up
- **Data Accuracy**:"Analytical work carried out on VOA records and drawings in the 1990s showed that these measurements are very accurate, no doubt in part because they are open to challenge by ratepayers (Gakovic et al., 1993)."[3]
- **Linking to OS data**:"[...] VOA hereditaments can be matched to map polygons representing building footprints, by their respective addresses. Ordnance Survey Address Base holds a link in many (although not all) cases to unique address reference numbers (UARNs) in the VOA Rating List."[4]

5.2 Clean postcode database

See above section, already cleaned for domestic property model

[3] Detail from Evans et al 2017 guide to 3DStock, http://journals.sagepub.com/doi/pdf/10.1177/0265813516652898

[4] Detail from Evans et al 2017 guide to 3DStock, http://journals.sagepub.com/doi/pdf/10.1177/0265813516652898

5.3 Merge VOA to postcode database

ND1_VOAPostcodeDatabase.do

93% matched to postcode database (giving all tiers of administrative geography and postcode centroids)

When I checked properties that didn't match, Royal Mail didn't seem to believe a lot of their postcodes existed. Is it possible to match based on the first 5 letters of a postcode or something?

5.4 Make model region lookup

see above, already done for domestic model LA_ModelRegion_Lookup.do

saved in "M:\NICProject\Lookups\clean"

5.5 Merge in model region

Everything matches as long as it has a local authority. ND2_ModelRegion.do

5.6 Clean LA characteristics

see above, already made for domestic property

5.7 Merge in la characteristics

ND3_LACharacteristics.do

everything matches as long as it has a local authority code

5.8 Clean GLUD

see above, cleaned for domestic property in Clean_GLUD.do

5.9 Merge in GLUD

1,065 properties with output areas from the postcode database don't match the GLUD. I can only assume it's because they're new output areas? A real mystery.

5.10 Clean population density

see above, cleaned for domestic property in Clean_PopulationDensity.do

5.11 Merge in population density

ND5_PopulationDensity.do

Everything with an OA/LSOA code successfully merged

5.12 Clean IMD

see above, cleaned for domestic property in england rollout, Clean_IMD.do

5.13 Merge in IMD

ND6_IMD.do

Everything with an lsoa code matched

5.14 Clean floodrisk

see above, already cleaned for domestic property in Clean_OpenFloodRisk.do

5.15 Merge in floodrisk

ND7_Floodrisk.do

This loses us an extra 55,000 properties that have postcodes created since this version of floodrisk data was created. Can match in using nearest neighbour (did lars do this for the domestic property? I can probably do it after I've done my course)

5.16 TEMPLATE

Relevant do file:

Input data file(s):

Output data file(s):

Brief summary:

Limitations: