What’s in the laundromat? Mapping and characterising offshore owned domestic property in London

Jonathan Bourne *,1, Andrea Ingianni2, and Rex McKenzie2

1University College London, The Bartlett Faculty of the Built Environment
2Kingston University Department of Economics

July 25, 2022

Abstract

The UK, particularly London, is a global hub for money laundering, a significant portion of which uses domestic property. However, understanding the distribution and characteristics of offshore domestic property in the UK is challenging due to data availability. This paper attempts to remedy that situation by enhancing a publicly available dataset of UK property owned by offshore companies. We create a data processing pipeline which draws on several datasets and machine learning techniques to create a parsed set of addresses classified into six use classes. The enhanced dataset contains 138,000 properties 44,000 more than the original dataset. The majority are domestic (95k), with a disproportionate amount of those in London (42k). The average offshore domestic property in London is worth 1.33 million GBP collectively this amounts to approximately 56 Billion GBP. We perform an in-depth analysis of the offshore domestic property in London, comparing the price, distribution and entropy/concentration with Airbnb property, low-use/empty property and conventional domestic property. We estimate that the total amount of offshore, low-use and airbnb property in London is between 144,000 and 164,000 and that they are collectively worth between 145-174 billion GBP. Furthermore, offshore domestic property is more expensive and has higher entropy/concentration than all other property types. In addition, we identify two different types of offshore property, nested and individual, which have different price and distribution characteristics. Finally, we release the enhanced offshore property dataset, the complete low-use London dataset and the pipeline for creating the enhanced dataset to reduce the barriers to studying this topic.

*corresponding author: jonathan.bourne.15@ucl.ac.uk
1 Introduction

The UK, and in particular London, is hub for international money laundering using property [1]. The volumes of money laundered through the UK capital are believed to be so high the city is sometimes referred to as ‘The London Laundromat’. A substantial part of the reason for this is that in the UK it is possible for offshore shell companies to purchase property. This allows those with illicit incomes to buy property without being identified [2]. In the ‘National risk assessment of money laundering and terrorist financing 2020’ The risk of money laundering using property purchases was rated as high [3], with domestic property being a particular focus due to the large amounts of money that can be moved in a single transaction and the lifestyle benefits such properties bring. In addition transparency international have found that 75% of money laundering cases involving property use offshore shell companies [4]. However, attempts to control money laundering have struggled for decades [5], for an in depth discussion of money laundering and property see [6, 7]. In 2015 a list of offshore owned properties in England and Wales was leaked to, and made public by, the ‘Private Eye’ a British publication [8]. Since then the UK government has published a regularly updated dataset of offshore owned properties called the ‘Overseas companies that own property in England and Wales - Use land and property data’ (OCOD) [9].

The OCOD dataset has the potential to be a great resource for those wishing to understand a the distribution and characteristics of offshore owned property in the UK. However, although the dataset has been used by researchers [10, 11], the dataset is challenging to use with addresses recorded as free text. These addresses can be incomplete, contain nested properties, where multiple properties exists within a single row or title number, it also contains no information on whether the property is domestic, business or something else. Such poor quality data makes understanding the distribution and characteristics of offshore owned property in the UK challenging.

In addition to offshore owned domestic property, London has large amounts of short term lets and low-use/empty property. In this paper the Offshore domestic property, airbnb property and low-use property, will be grouped under the term ‘Unconventional Domestic Property’ (UDP). Short term lets of the sort facilitated by Airbnb, One Fine Stay and Vrbo have received a quantitative exploration In both London and other major cities [12, 13, 14]. Low-use property, i.e. property with no permanent resident, has also received coverage [15, 16, 17], but less than the short term lets market due to difficulty obtaining data which is seldom publicly available due to privacy concerns. In the UK data gathering is further complicated by the distinct sub-national jurisdictions. For example, the register of property sales is public in England and Wales [18] but not in Scotland. To counter some of the difficulties with property data the UK government developed the Unique Property Reference Number (UPRN) [19]. The UPRN is a referencing system designed increase the ease of linking property data across government datasets. However, whilst the use of the UPRN is mandated, almost no government department uses it meaning linking the data...
requires advanced data processing skills [20, 21].

A partial solution to the data issues within OCOD and other property datasets, is the use of address parsers. Recently open source address parsers have been developed to convert free text addresses to a tabular and normalised format, which is much easier to analyse [22, 23, 24]. Address parsing is a form of Named Entity Recognition (NER) which is a subfield of Natural Language Processing (NLP) which is itself a subfield of Machine Learning (ML). NER uses statistical techniques to identify certain words within a sentence or larger block of text which represent a predefined entity and labels them with a given entity class. For address data, an example of a specific class could be ‘street name’ or ‘postcode’. The named entity is defined by its ‘Span’ which is the position number of the start and end characters as well as the label of the entity class.

Although very valuable the existing address parser’s are not without problems. Many of the parsers are are trained using Open Street map data [25], this has the advantage that there are many millions of stored addresses to use as training data, and algorithms can be trained to parse addresses from countries across the world using multiple languages and scripts. However, it also means that even though training data has noise induced to increase robustness, competing algorithms are all trained on the same dataset. This training dataset may be subtly different from the actual data the resulting algorithms are applied to. This can lead to weaknesses, for example the most well established parser libpostal [23] has been shown to have issues parsing incomplete addresses [22], it also struggles with addresses in formats it is not used to [26]. A more fundamental issue may be that the algorithms are designed on the assumption that they are parsing a single address, the OCOD dataset contains large numbers of nested addresses where a single free text line may contain tens or even hundreds of properties. As such a custom parser must be created to deal with the issues presented by this specific dataset. Given that there is no real training data creating such a parser will need to take advantage of recent developments in ML and NLP.

A major difficulty in any machine learning project is the obtaining of labelled data. This is because the majority of machine learning tasks use supervised learning. In supervised learning, the algorithm is instructed to learn a specific task such as classification or regression, for which the training data supplies the correct answer, the algorithm is considered trained when the error between the predictions and ground truth is minimised. The data used in supervised classification tasks typically has to be labelled by a human. This is time consuming, and can be impractical for small teams or those without funding to pay others using ‘crowd sourcing’ [27, 28]. A developing solution to this issue is ‘Weak learning’ [29], in this approach data is labelled using simple rules, this can produce a large number of labels which are often correct, but can also overlap and contradict one-another. A denoising process is then applied to the data using some statistical method such as a hidden markov model [30] or matrix completion [31]. The denoising process removes overlapping labels to minimise the error with the ground truth. An obvious critique of weak learning is that the true labels are not known bringing prediction quality into question,
however one of the most powerful aspects of weak learning is that it can produce models that are almost equal in performance to hand labelled data [32], indeed the quality of the model scales according the the size of the unlabelled dataset [29]. In addition recent research has highlighted that most datasets have errors even those used as a benchmark machine learning datasets such as MNIST, ImageNet and Amazon reviews [33]; this suggests that datasets created either through weak learning or hand labelling are on an error scale but can both produce high quality models.

This paper proposes a new method of identifying and classifying offshore owned property in England and Wales and uses this method to answer the research question; How is offshore domestic property in London characterised in terms of financial value and spatial distribution? How does this differ from low-use property, airbnb property and conventional homes? Does offshore domestic property appear to be broadly homogeneous, or does it contain distinct subgroups?

2 Method

The method is broken into three parts. The first part introduces the data, the second part details how the OCOD dataset is cleaned and transformed into the enhanced OCOD dataset. The third part describes the steps used in the analysis.

It should be noted that although address parsing is a major part of this project, the goal is not to create an universal address parser. As such the focus of the method is the process of creating the enhanced OCOD dataset and the resulting analysis.

In this project geographical analysis is performed on discrete units. These units were created by the Office for National Statistics (ONS) for use in census geography [34]. We use 3 levels of census geography which are (from smallest to largest) the Output Area (OA), Lower Super Output Area (LSOA) and Middle Super Output Area (MSOA), In addition we use the Local Authority District (LAD), which is the smallest level of local government. On average a LAD contains 30 MSOA although there can be substantial variance. The use of the OA system provides a standard interface with government data and prevents data privacy issues.

2.1 Data and code

This work uses several publicly available datasets. They key data set is the ‘Overseas companies that own property in England and Wales - Use land and property data’ (OCOD) [9]. The data cleaning process uses the Land Registry Price Paid dataset [18], The VOA ratings listing dataset [35] and the Office of National Statistics Postcode Directory (ONSPD) [36]. The Airbnb data is from the website InsideAirbnb [37]. The airbnb properties in this dataset only include entire homes which are let to the extent it is not possible to have a permanent
The low-use dataset uses data gathered using Freedom of information (FOI) requests, the majority of which was collected for [15], but the outstanding London boroughs have since been added using the same FOI process (see github repo isoa foi template maker). The definition of low-use is the same as [15] which is broadly that the property has no permanent resident, but this is through choice. Price data used in the inference and analysis section comes from the Price Paid dataset. The London Low-use dataset and the Enhanced Offshore ownership dataset are both available along with the code used at the github repository for this project JonnoB empty homes london. The data description of the Enhanced OCOD can be found in supplementary material.

It should be noted that the nested properties highlight another nuance with the OCOD dataset. Nested properties are part of a title that is owned by an offshore company but that title may refer to the freehold as opposed to the leasehold. This means that the right to live in the property has been sold to an individual for some period of time (e.g. 100 years) although the property still belongs to the freeholder. Leaseholders usually pay ‘ground rent’ to the freeholder. This leaseholds are typically case with flats and apartments. We included both freehold and leasehold properties on the basis that in either case the benefit either in terms of sale or 'ground rent' asset value belongs to the offshore company.

2.2 Creating the Enhanced OCOD dataset

The enhanced OCOD dataset is created by putting the original OCOD dataset through a process of cleaning, labelling, parsing and classification. The end result is a dataset that can be easily analysed and follows tidy data principles [38] in that each property is an observation/row in the data. The OCOD dataset is a CSV file, it contains several columns however for the purpose of this paper two are of importance, the title number and the property address. The title number is the unique reference given to each ‘property’ although in reality a title can contain many properties or none. The property address contains the free text which will be parsed. The data cleaning and preparation process is broken into five sections, Label, Parse, Expand, Classify and Contract, in addition quality testing of the labelling and classification is performed. A schematic of the process is shown in figure 1.

The address data has a form similar to the three examples below. As the data is all addresses there is a degree of structure and predictability to it. The addresses are broken into comma separated parts; e.g. if present, street number or unit ID is always first, and postcode is always last, (see table 1).

| Example | Address                                      |
|---------|----------------------------------------------|
| 1       | flat 6, chartfield house, babel road, london |
| 2       | 5 to 15 (odds only) babel road, london (w1 8ap) |
| 3       | 5 babel road, london and parking 3.5 w1 8ap  |

Table 1: Example address formats
Figure 1: A schematic of the data cleaning process, from loading the OCOD dataset to outputting the final dataset ready for analysis.
After loading the OCOD dataset light pre-processing was performed such as ensuring a space after comma and removing double spaces, changing to lower case etc. The data was then imported to programmatic \[39\] where entities were labelled using regex (Regular expressions) functions. There were around 50 functions used to describe the 8 types of named entity (Regex functions can be found here https://github.com/JonnoB/empty_homes_london/tree/main/humanloop_1_2/export). These entities are unit ID, unit type, building name, street number, street name, city, postcode, filter type. The entity classes are described in table 2.

| Entity class      | Description                                                                                                                                 |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Unit ID           | Describes a sub-unit such as a flat number or parking space ID. Example one would have ‘6’ and example three would have ‘3.5’ as unit id. Unit Id is not always a number |
| Unit type         | Describes the type of unit if available. Example one would have ‘flat’ whilst example three would have "parking"                               |
| Building Name     | Example one would have ‘chartfield house’, the field would not be present for the other two examples                                      |
| Street Number     | The street number of the property if available, would be ‘5 to 15’ in example two and ‘5’ in example three. Street number is not always a number |
| Street Name       | Self explanatory, would be ‘babel road’ in all three examples                                                                                   |
| Filter Type       | When multiple properties are included in the address a filtering condition is often used, this is because in the UK odd and even numbers are often on opposite sides of the road. Or a company may not own all the flats in an apartment block. Example two would have ‘odd’ |
| City              | Self explanatory, would be London for all three examples                                                                                       |
| Postcode          | Self explanatory. In almost all cases the postcode is in parenthesis. In addition UK postcodes follow a pattern which can be extracted using regex, making them easy to label |

Table 2: Definitions of the entity classes used in the labelling process, example formats are found in table 1.

Once the labels have been added to the data the whole dataset is extracted as a json file. In some cases label spans overlap these are removed either using a simple rule that keeps only the largest span, or a denoising process that uses
a hidden markov model, following the principles of weak labelling. In addition, using the output of Programmatic, we create a predictive model using spaCy [40], the model uses RoBERTa [41] as the underlying algorithm. RoBERTa is a type of transformer [42], currently one of the most successful approaches to solving problems in NLP.

Once the labels have been denoised a ground truth data set is hand labelled using 1000 randomly sampled observations. An additional dev set of 2000 examples is also labelled using the same process. This ground truth dataset is used to asses the quality of the NER labelling, using programmatic and spaCy.

2.2.1 Address parsing

With a single address the parsing process is effectively over when the the entities have been labelled. However, in the case when there are multiple properties the parsing process is more involved the label dependency becomes important. As an example consider the nested address ‘flat 5, chartfield house and flat 16, zebra house, babel road, london (w1 8ap)’. In this example there are two flats in two separate buildings on a single street and with a common postcode, as such there are two distinct addresses with some shared labels. We parsed the addresses by assigning each character span its own row and each label class its own column. We then propagated the columns backwards to create complete address rows. Certain labels were classed as terminators and all non-terminator rows were removed from the data. Terminators were hierarchically structured such that if the most high priority label class was not available the next label class in the hierarchy was used. The terminator list in hierarchical order is unit type, unit ID’s, building name, street number. The presence of a span in a label column would block the backfilling of data from later in the sentence, this prevented situations where houses would be assigned the wrong road. When parsing was complete and address rows without terminators were removed and a tabular dataset of parsed addresses was returned.

As mentioned earlier and shown in table 1, some title numbers contained multiple addresses using an abbreviated format. To have a single address per line required that the dataset be expanded.

2.2.2 Address Expansion

Expanding the dataset so that each row represents a single property required machine reading of the unit ID label and the filter number label. A regex expression was used to identify unit ID’s or street numbers that covered multiple addresses the first and last numbers were then extracted and the numbers between them filled in making a contiguous list of numbers from i to j. If a filter number label was present it was used to identify whether the numbers were odd or even and the list of numbers were filtered to so only the appropriate number type remained. With the data now parsed and expanded such that a single property address was represented on a single line the data could now be localised to a specific OA or LSOA and classified into a specific property type.
2.2.3 Address Classification

The classification phase of the cleaning process includes classifying the property types but also includes localising the properties within the UK governments geographical area system. Initially all postcodes are cross referenced with the Office of national Statistics' Postcode Database (ONSPD). This links the postcode and the OA system used by the ONS for census and other demographic analysis. However, not all addresses have a postcode, in these cases the parsed addresses are cross referenced with the land registry price paid dataset and the VOA company ratings list dataset, if a match is found the OA and LSOA is added to the OCOD address. For nested addresses the same OA and LSOA is added to all addresses within the property title. The data is classified into one of six classes, in a sequential process (see figure 2). The 6 classification groups are airspace, business, domestic, land, parking and ‘unknown’. There are two classification processes Type 1 which has twelve classification steps, where each step identifies some aspect of a single property class, (for details on the process see classifying_the_ocod_dataset.ipynb, section ‘Classify Property type’). Any properties which cannot be identified are classed as ‘unknown’, these properties are almost entirely a mix of domestic and business. Once the Type 1 classification process is complete the Type two process is applied. As shown in Figure 2 the Type two process builds on the output of Type 1 and deduces whether a property is domestic or business by finding which addresses can be identified. Type 2 has several sequential classification steps which filter for the level of completeness of an address. When class deduction is possible the Type two process changes the ‘unknown’ class of the Type one process to ‘domestic’ or ‘business’ as appropriate. However addresses which do not contain enough information to accurately locate remain in the ‘unknown’ class. The outputs of the processes are provided separately in the dataset as the Type 1 process is more secure as it primarily relies on positively identifying classes whereas Type two uses deduction to work out what the property is not. The analysis will use Type two classification.

2.2.4 Address Contraction

This is the final and most straight forward of all the cleaning steps. As it is not unusual for commercial premises to occupy more than one ‘street address’ non-domestic properties that are within a nested property group are reduced back to single address. As such only domestic properties are ever within a nested property title.

2.2.5 Test set and Accuracy metrics

Whilst the previous sections have described the steps by which the data is prepared, they do not describe any way to measure the quality of the labelling or classification processes. This is essential as the for the final analysis to be considered reasonable the data needs to be of high quality. The data is tested by randomly sampling 1000 addresses from the original dataset these samples are
Figure 2: The classification process during the preparation of the enhanced OCOD dataset.
then hand labelled and held out from any further analysis until the end of the cleaning process. Figure 1 shows that the ground truth is created after the initial labelling, this is for practical reasons to speed up the hand labelling process as many of the labels will already be complete and correct. Finally the classification quality is assessed. The manual checking of the class is done by looking up the address using map software, looking at pictures of the property, and doing an internet search of the address to see if it is a domestic or a commercial property. This process is certainly not infallible but without the use of the UPRN system there is no other way within a reasonable time frame.

Once the ground truth has been created, the datasets produced by the processes in this paper and alternative parsing algorithms can be compared. Performance will be measured using the precision, recall and F1 metrics. Precision is the probability of a span exactly matching the ground truth test set. In the binary class case precision is defined as

\[
\text{Precision} = \frac{tp}{tp + fp},
\]

where \(tp\) is the total number of true positives where the span matches the test set and \(fp\) is total number of false positives where span does not. Recall is the fraction of matched spans which are also given the correct class, it is defined by

\[
\text{Recall} = \frac{tp}{tp + fn},
\]

where \(fn\) is the total number of false negatives. The F1 score is

\[
\text{F1} = \frac{2tp}{2tp + fp + fn},
\]

making it the harmonic mean of precision and recall. However in order to adjust the measure for the multiclass data used in this paper we use the micro-averaged values. Microaveraging takes the sum of \(tp, fp, fn\) across all classes, and uses those summed values to calculate the final precisions, recall, F1 values. As an example micro precision can be defined as \(\bar{tp}_\mu = \frac{\sum_i tp_i}{\sum_i tp_i + \sum_i pp_i}\), where \(i\) is the property class and \(m\) is the total number of properties. It should be noted that although the micro average can produced biased results in imbalanced data, the output put is more easily interpretable as it shows in simple terms the fraction labels that are correct.

Once the final enhanced dataset has been created the process of analysis can begin.

### 2.3 Analysing the offshore dataset

The analysis of this paper is performed on each domestic property class. Where the domestic property classes are conventional homes, low-use homes, airbnb properties, offshore homes. The purpose of the analysis is to understand the differences in the behaviour between these property classes and the implications
for the housing market. Three core analyses are performed in this paper, inferring property price, entropy of property and spatial autocorrelation. Having described the outline of the method the next section describes how the total number of UDPs will be calculated.

2.3.1 Calculating the total number of unconventional domestic property

The three property types airbnb, low-use that make up UDP are not mutually exclusive and so cannot simply be added together. Instead the joint probability will be found under two conditions; The first is that data are independent; The second that offshore property is always low-use. The first case is the null hypothesis and will act as the upper bound, whilst the second case is based on offshore properties being inherently luxury purchases and the owner having multiple homes. The second case reduces the total number of classes as the offshore properties become a subset of the low-use properties. The probability of a UDP in a given LSOA is defined as

\[ P(U|m) = 1 - \prod_{i=1}^{m}(1 - P(i|m)), \]  

That is 1 minus the joint probability of a property not being any UDP type. Although the number of properties will vary if sampled from the distribution the variance will be small compared to the difference between the two cases where UDP types are independent or not. Therefore for simplicity the total number of UDPs in an LSOA is calculated to be the product of the total number of homes and the probability of being an UDP in that LSOA. With the number of UDPs calculated the next section will describe the sampling process applied to the number of UDPs to infer their mean prices.

2.3.2 Inferring the prices of properties in London

In [15] The value of domestic property and low use property was inferred using a simple graphical model. The same technique will be used in this work to infer the value of the distinct UDP types as well as for the total number of UDPs. Due to low the number of samples in certain areas, the LSOA counts will be aggregated up to MSOA level.

The distribution of the mean price by property type is calculated as shown in

\[ \langle C \rangle_{m=i} = \frac{1}{z} \sum_{j} C_{j} P(C_{j} \mid m = i), \]  

Where \( \langle C \rangle_{m=i} \) is the mean price for property type \( m \), \( C_{j} \) is the vector of possible prices for \( j \)th property and \( P(C_{j} \mid m = i) \) is the probability price distribution for \( C_{j} \). The sum across all values of \( j \) is divided by the total number of properties \( z \). Clearly \( \langle C \rangle_{m=i} \) is simply the mean price of the MSOA of \( j \) as such equation 5 is the Bayesian equivalent mean of the mean MSOA price weighted
by distribution of homes. Both methods can be used and for large values will produce similar results, however equation 5 is able to provide an indication of the variance statistical significance and so will be used here. The equation is implemented as form of ancestral sampling from the distribution, in order to obtain a representative mean price, the sampling is repeated 501 times to produce a distribution of the mean, this is similar to the bootstrap technique developed by [43].

2.3.3 Analysing the entropy of domestic property

In information theory Shannon entropy [44] is used to describe the amount of disorder within a system of a discrete fixed number of states and is measured in the amount of bits of information required to store it. The higher the number of bits the more disorder there is in a system and conversely the lower the number of bits the more order there is in a system and the easier it is to describe. The equation for Shannon entropy is

\[ E = - \sum_{i=1}^{k} p_i \log p_i, \]  

(6)

where \( k \) is total number of states, and \( p_i \) is the probability of the \( i \)th state.

In the geospatial analysis being performed here the entropy describes how concentrated the offshore property types are. The discrete states in this case are the 4835 LSOA in London. The entropy would be minimised if all of a certain property class would be found in a single LSOA whilst entropy would be maximised if the properties were spread evenly across all LSOA. As such it can be said that entropy represents how concentrated the property classes are.

Although understanding the concentration of the property is very valuable it lacks any spatial information, for this we need to use a measure of spatial auto-correlation.

2.3.4 Analysing the spatial auto-correlation of domestic property

Spatial auto-correlation measures the strength of the correlation between objects in physical space and a given property. That is, if objects close together have very similar values but objects far away are very different, there will be a high auto-correlation value; whilst if objects that are close by have a very different value there will be a very low value-of auto-correlation. In this paper we will use Moran’s I [45] as the measure for spatial auto-correlation. Moran’s I is described as

\[ I = \frac{k}{W} \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij}(x_i - \bar{x})(x_j - \bar{x}) \sum_{i=1}^{k} (x_i - \bar{x})^2, \]  

(7)

where \( k \) is the total number of spatial indexes in this case LSOA, \( x \) is the target characteristic e.g. number of offshore properties, \( \bar{x} \) the mean characteristic value, \( W \) is the matrix of weights and \( w_{i,j} \) is element \( i,j \) of the weights matrix.
Similar to the more commonly used, non-spatial, Pearson correlation, Moran’s I varies between -1 and 1, where 1 shows perfect positive correlation.

Having described both the data processing and the analysis methods, we will now introduce the results of both parts of this method.

3 Results

The Results section is broken into two parts the first describing the data processing and the second the analysis of the enhanced OCOD dataset.

3.1 Labelling and parsing performance

Table 3 shows the labelling results of the NER task. These results are similar for both the basic process and using a trained neural network. Overall the model scored an f1 of 0.96 although this was somewhat unevenly distributed across the categories. In order to accurately locate and class the addresses building name, street number, street name and unit id are required (postcode is already supplied), using just these metrics the mean f1 is 0.95.

Once the labelled addresses have been parsed and expanded the process of locating the properties and classifying them is performed. The f1 of the classification process is 0.94 across all classes and 0.96 for domestic properties with a precision of 0.93 and a recall of 0.99, as shown in table 4. This means that whilst almost all the domestic properties are being correctly tagged the process is slightly over-estimating the number of domestic properties and under estimating the total number of businesses. Counting only unique title numbers instead of properties reduces the f1 to 0.92 as a large number of nested domestic properties are discounted.

| class          | precision | recall | fscore | support |
|----------------|-----------|--------|--------|---------|
| building name  | 0.93      | 0.89   | 0.91   | 383.0   |
| city           | 1.00      | 0.97   | 0.98   | 947.0   |
| number filter  | 1.00      | 1.00   | 1.00   | 23.0    |
| postcode       | 1.00      | 1.00   | 1.00   | 768.0   |
| street name    | 1.00      | 0.95   | 0.97   | 1029.0  |
| street number  | 0.98      | 0.98   | 0.98   | 678.0   |
| unit id        | 0.98      | 0.92   | 0.95   | 370.0   |
| unit type      | 1.00      | 0.97   | 0.98   | 488.0   |

Table 3: The NER labelling task is uneven across the entity types but overall provides an f1 score of 0.95-0.96, depending on whether only core labels or all labels are considered.

As the labelling and classification f1 scores show good quality results subsequent analysis of the data can be treated as having a reliable foundation. The next section will use the enhanced OCOD dataset to provide insight into...
| class     | precision | recall | fscore | support |
|-----------|-----------|--------|--------|---------|
| airspace  | 1.00      | 0.93   | 0.96   | 14.0    |
| business  | 0.97      | 0.78   | 0.87   | 311.0   |
| carpark   | 1.00      | 0.96   | 0.98   | 26.0    |
| domestic  | 0.93      | 0.99   | 0.96   | 918.0   |
| land      | 1.00      | 0.99   | 1.00   | 179.0   |
| unknown   | 0.00      | 0.00   | 0.00   | 9.0     |

Table 4: The performance of the classification task shows that the most challenging type are domestic and business.

UDPs in London. Overall the enhanced ocod dataset has 138 thousand entries 44% more than the 94 thousand in the original dataset. Table 5 provides the breakdown by property use of the enhanced OCOD dataset using the type 2 classification.

| class     | counts | percentage |
|-----------|--------|------------|
| 1         | airspace | 1667       | 1          |
| 2         | business | 21396      | 16         |
| 3         | carpark  | 2438       | 2          |
| 4         | domestic | 94576      | 69         |
| 5         | land     | 16384      | 12         |
| 6         | unknown  | 1124       | 1          |

Table 5: Breakdown of property type using type 2 classification

### 3.2 Data analysis

Analysing the subset of the enhanced OCOD dataset that is domestic property in London and comparing it to Airbnb, low-use and conventional property reveals certain characteristics of these property types. A summary of the quantitative analysis can be seen in table 6. Figure 4 shows the total value of the offshore properties is approximately £56 Billion, this is slightly than the total value of airbnb property of £42 Billion, however both are less than the value of the low-use property in London which is £85 Billion. Although the total amounts are interesting, it is more insightful to compare the mean value of the property types. There are significant differences in the mean values of the the property types as shown in Figure 3. Mean value of the Offshore property is £1.33 million, substantially higher than Airbnb properties which have a mean value of £1 million, this is followed by Low-use properties at £900 thousand. Overall the mean value of the UDPs is £1.02 million. However, all UDPs are much more expensive than the mean conventional property price of £600 thousand.

Figure 5 shows the distribution of the different property types across the
capital. The map shows that offshore property is the most concentrated with a large fraction focused in the centre of the city. Airbnb properties are not as concentrated but appear to have high levels of spatial auto-correlation. In contrast the distribution of empty homes appears relatively widely dispersed and with lower levels of spatial auto correlation. The visual interpretation of Figure 5 is supported by the quantitative analysis using entropy and Moran’s I. Figure 6 shows that the entropy of the offshore property is substantially higher than that of the other property types. The mean entropy of the different property types is 9.21 for offshore, 10.8 for airbnb, 11.4 for low-use and 11.2 for all UDPs. This entropy ranking matches the mean value ranking of the property types, however this is not the case with the spatial auto-correlation. Table 6 shows that offshore property has a low spatial auto-correlation (0.47) compared to the other UDP types (airbnb 0.72, low-use 0.61). It should be noted that figures 4 and 3 have very small variance, there is little doubt the differences between the categories is statistically significant.

| type     | total value (£B) | counts (K) | mean value (£M) | bits   | moran’s I |
|----------|------------------|------------|-----------------|--------|------------|
| 1 airbnb | 42.30            | 41         | 1.03            | 10.81  | 0.72       |
| 2 low_use| 85.10            | 93         | 0.91            | 11.39  | 0.61       |
| 3 offshore | 56.10        | 42         | 1.33            | 9.21   | 0.47       |
| 4 homes  | 2287.70          | 3553       | 0.64            | 12.19  | 0.33       |
| 5 UDP    | 173.80           | 170        | 1.02            | 11.20  | 0.70       |

Table 6: The complete metrics list for the analysis of the main property types

3.2.1 Nested vs single offshore properties

The notably low spatial auto-correlation value for offshore property compared to the other UDP types requires further investigation. One of the interesting aspects of the OCOD dataset is that it contains nested addresses. When the nested and single addresses are analysed separately a clear separate dynamic emerges. The single offshore properties show relatively high spatial auto-correlation with a value of 0.656 and a similarly low entropy as previously. However, the nested properties have a dramatically lower spatial auto-correlation of 0.112 and a much lower entropy of 7.35 bits. There is also a substantial difference in the estimated values of these properties with the individual properties having a higher value of £1.5 Million whilst the nested properties are closer to the value of conventional housing with a mean value of £890 thousand. Table 7, provides the full comparison of this analysis, whilst figure 7, maps the differences between the two types of offshore property. Looking at the price paid for individual properties provided in the original OCOD dataset, shows the mean to be £1.8 Million, removing all properties that are suspiciously low and the mean value increases to £3 Million. However, the prices in the OCOD dataset have a lot of quality issues and businesses that have been miss-classified will easily skew the total. Having said some properties within the dataset do have confirmed sales prices
### Comparison of Mean Property Price by Type Across 501 Simulations

| Property Type | Mean Price per Property |
|---------------|-------------------------|
| AirBnB        |                         |
| Low-Use       |                         |
| Offshore      |                         |
| Unconventional|                         |
| Total         |                         |

Figure 3: All three of the property types under analysis are worth substantially more than regular homes in the capital. AirBnB properties are worth more than low-use homes, highlighting the importance of location for this property type that exceeds that of general low-use property. However, offshore property is by far the most expensive of all the property types.
Figure 4: Due to the quantity of low use homes, their total value dwarfs that of the offshore domestic property, even though offshore domestic property is substantially more expensive
Figure 5: Looking at the maps clearly shows that there is a substantial difference in the distribution of the three property types. Offshore domestic property has some extremely high concentrations where an entire housing development is owned by an offshore company. Grey areas have no properties of that type present.
Comparison of mean entropy by property type across 501 simulations

Figure 6: The entropy of the property types appears inversely proportional to the value. Indicating that the offshore, airbnb and low-use properties are not only more concentrated geographically that normal homes but are more concentrated into high value areas.
Figure 7: Comparing the nested vs normal property shows very different spatial patterns. These patterns correspond to large average differences in value. Grey areas have no properties of that type present of over £100 million. Overall even though it may be slightly conservative the sampling method prevents unrealistically high estimates.

| type  | total value (£B) | counts (K) | mean value (£M) | bits | morans I |
|-------|------------------|------------|-----------------|------|----------|
| 1     | individual       | 77.60      | 51              | 1.51 | 9.55     | 0.66     |
| 2     | nested           | 17.30      | 16              | 1.06 | 7.35     | 0.11     |

Table 7: The complete metrics list comparing nested and single offshore homes

3.2.2 Country of Incorporation

There are clear patterns in the London domestic property market with regards country of incorporation of offshore companies. Whilst there are properties associated with 133 different countries or territories, 50% of the properties are related to companies incorporated in the three crown dependencies of Jersey, Guernsey and the Isle of Man. A further 30% of properties are associated with one of The British overseas territories. In fact only 4 territories, British Virgin Islands, Jersey, Guernsey and The Isle of Man, are associated with 78% of all properties. Looking at the relationship with nested properties, The crown
dependencies have an over representation of nested properties with The crown dependencies and the BVI representing 74% of all nested titles and 91% of all nested properties. Jersey alone represents 52% of all nested properties. More detail can be found in supplementary material.

4 Discussion

The overall labelling quality was high, an f1 score of 0.96 suggests that the data is being effectively labelled and provides confidence in the quality of the final addresses. This also shows that the deduction used in the type two classifier is a robust approach. On top of this the final classification values also had strong performance across the categories although separating businesses and domestic properties was problematic at times. Improving the performance of the NER process would require a more in depth analysis of why buildings were not being captured at the same level as the other entities possibly including large scale visual data and as such is beyond the scope of this paper. Recall was notably lower for the buildings labels. This was partially due to the complexity of building names, but the high precision suggests that there may be some patterns which have not been properly addressed. A failure to identify evenly patterns from across all building names highlights a drawback of the weak learning approach and a risk of inducing systemic bias. It was surprising that the RoBERTa based spaCy model did not improve performance, this may be because the structure of the addresses is so simple and regular that the attention abilities of the transformer could not add any value. The lower recall and precision scores for Business and Domestic properties was mostly due to these two classes being difficult to separate. This challenge highlights the importance of the UPRN system in the UK. It should be noted that a freedom of information request was sent to both the VOA and the Land Registry requesting they provide the UPRN data, however, both FOI’s were rejected.

Although undoubtedly the most appropriate choice for the OCOD dataset, the parser developed for this paper is a niche product designed for a single use case and is unlikely to offer the robustness or flexibility of already available general purpose parsers such as Deepparse [22] and libpostal [23]. We would not recommend it as a first choice for parsing data other than the OCOD dataset and possibly other UK government address data. The code for creating the enhanced OCOD dataset is publicly available under a GNU GPLv3 license. In addition the supplementary material provides a more in depth description of the training details. Having the trained pipeline available means those who wish to use the enhanced OCOD dataset can do so for any past or future release of the Land Registry’s OCOD dataset.

The results of analysing the enhanced OCOD dataset show that there is a substantial amount of UDP in London (Property that is either low-use, Airbnb or offshore owned). There is an estimated £145-174 billion UDP in London. This is approximately 7.5% of the total domestic value, of this there is approximately £56 Billion offshore owned property representing 42 thousand dwellings. We
found that there were two distinct types of offshore property single and nested. The single properties may be under valued as an analysis of the prices in the OCOD dataset suggest a mean value of £3 Million whilst the empirical sample only returns £1.5 Million. However, due to data quality concerns with the values reported in OCOD the mean reported value should only be considered as context. In the dataset nested properties, those properties sharing a title number, are mostly entire housing complexes and have very different spatial patterns to the individual offshore properties. Individual offshore properties are very expensive even by the standards of UDP, in addition they are concentrated on the centre of London with strong spatial auto correlation. In contrast nested offshore property is somewhat less concentrated on central London but more highly concentrated in general, there is also almost no spatial correlation. There was a brief analysis relating to country of incorporation. Four territories control 78% of all offshore owned domestic property London. These territories are the three UK Crown Dependencies (Jersey, Guernsey, The Isle of Man) and The British Virgin Islands, which is a British Overseas Territory. Although we do not suggest that any specific property is being used to launder money, throughout the project certain properties did show some unusual characteristics; However a deeper inspection of these properties is beyond the scope of this paper.

5 Conclusion

This paper proposes a new method of identifying and classifying offshore owned domestic property in England and Wales. Using the dataset created by this method we have shown that Offshore owned domestic property in London is not a homogeneous group, but contains at least two distinct types, characterised by measurable quantities and appearing to relate to purpose and country of incorporation. We have also shown the distribution and value of three different types of UDP and that all three property types are more concentrated and more expensive than conventional domestic property. Whilst this paper does not draw any causal relations or make any policy proposals it provides for the first time a clear map to understanding the behaviours and value of Offshore owned domestic property and Low-use property in London. This data can be used by sociologists, economist and policy makers to ensure that attempts to reduce money laundering and high property prices are based on detailed data that reflect the real situation. In addition to the findings of the analysis the paper has created a high quality open-source dataset that allows further research as well as a code pipeline to create new datasets as required. It has been said that fighting money laundering requires openness and transparency [5], the OCOD dataset was a step in that direction, we hope that the Enhanced OCOD dataset is the next.
5.1 Future work

The enhanced OCOD dataset can be used for a wide variety of projects. Future work can expand the offshore analysis to the entire of England and Wales. Building on the work of [11], and analysis of the nested offshore properties can be performed to explore whether they are vehicles for gentrification. Some of the properties appeared to have unusual characteristics, work of [7] could be used to identify if these properties may be being used for money laundering.

6 acknowledgements

We would thank humanloop.com for giving us access to the ‘Programmatic’ beta and providing technical support during the NER labelling phase of the project.

7 funding

The authors would like to thank Trust for London (grant number MAIN-S2-06.10.2020-8792(6930) ) for funding the research. We would also like to thank Kingston University for providing additional research funding.

References

[1] TIUK. Stats reveal extent of suspect wealth in UK property and Britain’s role as global money laundering hub, 2022.

[2] Ben Cowdock and Steve Goodrich. The Cost of Secrecy. Technical report, Transparency International, London, 2018.

[3] HM Treasury. National risk assessment of money laundering and terrorist financing 2020. Technical report, UK Government, London, 2020.

[4] Matteo de Simone, Lloyd Barthropp, and Nicl Maxwell. Corruption on your doorstep: How corrupt capital is used to buy property in the UK. Technical report, Transparency International, 2015.

[5] Austin Mitchell, Prem Sikka, and Hugh Willmott. The Accountants’ Laundromat. Association for Accountancy & Business Affairs, Basildon, 1998.

[6] Milind Tiwari, Adrian Gepp, and Kuldeep Kumar. A review of money laundering literature: the state of research in key areas. Pacific Accounting Review, 32(2):271–303, January 2020. Publisher: Emerald Publishing Limited.

[7] Brigitte Unger and Joras Ferwerda. Money Laundering in the Real Estate Sector. Edward Elgar Publishing, 2011. Publication Title: Books.

[8] Private Eye. Selling England (and Wales) by the pound, 2015.
[9] Land Registry. Overseas companies that own property in England and Wales - Use land and property data, 2022.

[10] Rex McKenzie and Rowland Atkinson. Anchoring capital in place: The grounded impact of international wealth chains on housing markets in London. Urban Studies, 57(1):21–38, January 2020. Publisher: SAGE Publications Ltd.

[11] Filipa Sá. The Effect of Foreign Investors on Local Housing Markets: Evidence from the UK. SSRN Scholarly Paper 2877263, Social Science Research Network, Rochester, NY, November 2016.

[12] Zahratu Shabrina, Elsa Arcaute, and Michael Batty. Airbnb and its potential impact on the London housing market. Urban Studies, 59(1):197–221, January 2022. Publisher: SAGE Publications Ltd.

[13] Helen X. H. Bao and Saul Shah. The Impact of Home Sharing on Residential Real Estate Markets. Journal of Risk and Financial Management, 13(8):161, August 2020. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[14] Miquel-Àngel Garcia-López, Jordi Jofre-Monseny, Rodrigo Martínez-Mazza, and Mariona Segú. Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. Journal of Urban Economics, 119:103278, September 2020.

[15] Jonathan Bourne. Empty homes: mapping the extent and value of low-use domestic property in England and Wales. Palgrave Communications, 5(1):1–14, February 2019. Number: 1 Publisher: Palgrave.

[16] Youngre Noh, Galen Newman, and Ryun Jung Lee. Urban decline and residential preference: The effect of vacant lots on housing premiums. Environment and Planning B: Urban Analytics and City Science, 48(6):1667–1683, July 2021. Publisher: SAGE Publications Ltd STM.

[17] Hiroki Baba and Kimihiro Hino. Factors and tendencies of housing abandonment: An analysis of a survey of vacant houses in Kawaguchi City, Saitama. JAPAN ARCHITECTURAL REVIEW, 2(3):367–375, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2475-8876.12088.

[18] Land Registry. How to access HM Land Registry Price Paid Data, 2022.

[19] OS. OS Open UPRN, November 2021.

[20] Bin Chi, Adam Dennett, Thomas Oléron-Evans, and Robin Morphet. A new attribute-linked residential property price dataset for England and Wales, 2011 to 2019. UCL Open: Environment Preprint, May 2021. Publisher: UCL Press.
[21] Sam Comber and Daniel Arribas-Bel. Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. *Transactions in GIS*, 23(2):334–348, 2019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12522.

[22] Marouane Yassine, David Beauchemin, François Laviolette, and Luc Lamontagne. Leveraging Subword Embeddings for Multinational Address Parsing. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pages 353–360, June 2020. ISSN: 2327-1884.

[23] Al Barratine. libpostal: international street address NLP, 2017. original-date: 2015-03-03T00:21:31Z.

[24] Selman Delil, Birol Kuyuncu, Cüneyt Aksakalı, and İsa Semih Akçıra. Parsing Address Texts with Deep Learning Method. In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, October 2020. ISSN: 2165-0608.

[25] OSM. OpenStreetMap, 2004.

[26] Helen Craig, Dragomir Yankov, Renzhong Wang, Pavel Berkhin, and Wei Wu. Scaling Address Parsing Sequence Models through Active Learning. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL ’19*, pages 424–427, New York, NY, USA, November 2019. Association for Computing Machinery.

[27] Victor S. Sheng and Jing Zhang. Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9837–9843, July 2019. Number: 01.

[28] Jing Zhang, Xindong Wu, and Victor S. Sheng. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4):543–576, December 2016.

[29] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data Programming: Creating Large Training Sets, Quickly. arXiv:1605.07723 [cs, stat], January 2017. arXiv: 1605.07723.

[30] Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. Named Entity Recognition without Labelled Data: A Weak Supervision Approach. arXiv:2004.14723 [cs, stat], April 2020. arXiv: 2004.14723.

[31] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training Complex Models with Multi-Task Weak Supervision. arXiv:1810.02840 [cs, stat], December 2018. arXiv: 1810.02840.
[32] Jason A. Fries, Ethan Steinberg, Saelig Khattar, Scott L. Fleming, Jose Posada, Alison Callahan, and Nigam H. Shah. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1):1–11, April 2021. Number: 1 Publisher: Nature Publishing Group.

[33] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv:2103.14749 [cs, stat]*, November 2021. arXiv: 2103.14749.

[34] ONS. Census geography, 2022.

[35] VOA. VOA rating list downloads, 2017.

[36] ONS. ONSPD, 2022.

[37] Murray Cox, John Morris, and Taylor Higgins. Inside Airbnb, 2022.

[38] Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59:1–23, September 2014.

[39] Humanloop. Humanloop: Programmatic, 2022.

[40] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. arXiv: 1907.11692.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. arXiv: 1706.03762.

[43] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*. Chapman & Hall/CRC, Philadelphia, PA, 1993.

[44] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. Conference Name: The Bell System Technical Journal.

[45] P. A. P. Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17–23, 1950. Publisher: [Oxford University Press, Biometrika Trust].