Investigation into the nature of Homelessness and its related Government solutions

1st Linda Lomenčíková
University of Bristol
Computer Science
fv19959@bristol.ac.uk

2nd Lokhei Wong
University of Bristol
Computer Science
uy19297@bristol.ac.uk

3rd Yuze Ma
University of Bristol
Computer Science
on19961@bristol.ac.uk

4th Haoran Lin
University of Bristol
Engineering Mathematics
ib19228@bristol.ac.uk

5th Yuyang (Eric) Liu
University of Bristol
Engineering Mathematics
gf19229@bristol.ac.uk

***Abstract*—The main objective of this project is to investigate the past, present and future of homelessness in the UK and to quantify the relationship between homelessness and the relevant housing market policies in the country. Using England's data tables provided by the Government Statistical Service (GSS), the investigation distinguished the provided data into two types: time series and categorical, with data selected, imputed and explored accordingly. Firstly, the project established a strong correlation between the time series of homelessness and multiple government policies' data, such as the size of the social housing waiting list. Then, the correlation studies into the categorical data of homelessness have provided sufficient evidence to show the size of local population would be a important factor correlated with local homeless population. Last but not least, the prediction of homelessness based on the housing policies' statistics have also been successful and forecasts some positive changes in the situation of the homeless. At the same time, however, the size of social housing waiting list will continue to grow. We also acknowledge that the prediction might only be valid in the short term due to the irregularities in policy decisions. Overall, the findings in this project support the idea that homelessness data recorded for England is highly dependent on changes in the housing market and relevant policies. For reproduction, all codes and results are stored in the project's Github page[1].**

I. Introduction

By 2019, around 60,000 households were recorded as homeless in England. 71% of local councils across the UK reported an increase in the homeless population. With constant travelling and temporary addresses, the homeless population in the UK forms a marginalised minority group seriously threatened in both health and social identity, as discussed by Jeremy S. Godfrey in his research on modern homelessness[7].

Among many other researchers exploring the topic of homelessness, data-driven analysis has already been attempted by economists in British academia, who linked the housing instability in homelessness to the population's disadvantage in medication and poor health conditions[10][16]. The British government is also seeking to improve the living conditions of homeless citizens with policies such as Help to Buy, Affordable Housing and Prevention duties. However, there is sufficient evidence to suggest that the provisions are under-provided[17][16].

Therefore, the investigation into the homeless population would not only be valuable sociologically but also in the interest of the policymakers seeking to improve their responses to the ongoing homelessness crisis. The new interactive housing statistics tool provided by the Government Statistical Service (GSS), together with modern data analysis via machine learning techniques, can quantify the strength of the connection between markets and government policies and the homeless population.

Thereby, two main objectives can be established. The first is to investigate the past, the present and the future of British homelessness and its related market and policy statistics in order to establish a more quantified understanding of the ongoing situation. The second objective is to use machine learning tools to quantify the linkage between homelessness and the relevant factors.

Regarding the ethics of the project, established government statistical teams collect the data from the households and personnel that have been anonymised through aggregation, and all data was collected with consent by local authority councils.

II. Data Selection and Preparation

GSS housing database provides a range of data produced by the Government including housing and planning, homelessness, and rough sleeping statistics. In the UK, housing policy is a devolved matter. As such, we do not aim to combine the statistics from the respective countries, and seek only to explore those of England.

A. Selection of Data based on the Objectives

First of all, among the 123 available datasets in the GSS database, we selected those relevant to our topic by exploring the briefs, which we then categorised into three types:

1) Data Directly Related to Homelessness: Among these datasets, the most informative one was selected: the "Statutory homelessness: Detailed local authority-level tables". The selected table contains information about the different types of homelessness, including the most conclusive feature - the total households owed a duty.

2) Government Housing Policies' Data: This classification includes policies directly related to social housing or those that influence housing supply and demand and market pricing. Examples of tables selected for this category would be "Social Housing Waiting List statistics" and "Housing Statistics Tables", directly related to homelessness. We also include other measures, such as the Council Tax scheme and House prices, in order to establish links with homelessness.

3) The others: The only table in this category is the ”2018-based household projections” table, which is an indicator of the size of consumers in the housing.

After selecting the tables, they are split into time series and categorical data. Due to the pandemic, many entries in the categorical datasets were missing. This is addressed in the data preparation stage. In contrast, our time series data was consistently formatted and had no missing data values. As such, no data preparation was required, aside from merging the datasets.

B. Data Preparation of the Categorical Data - Merging, Aggregation and Imputation

Our categorical data is organised by local authorities, which can be distinguished by the ONS code (geocodes). This is useful for merging and aggregating our data. For convenience, we first merged relevant features of our categorical data into a dataframe. This is done using DataFrame's *merge()* method from the Pandas library. Following this, we aggregated the council tax bands in order to categorise housing from cheap and affordable to expensive and luxury.

To deal with the missing values, we use data imputation. We used two methods to do this. The first of which was to drop every instance where the number of missing features exceeded 15% of the total number of features for the concerning dataset. The next strategy was to impute missing values using the k-Nearest Neighbours algorithm. Scikit-learn provided us with a function, ”KNNImputer()”, which allowed each missing value of the sample to be imputed using the mean value from the k-nearest neighbours.

III. Data Exploration

A. Time Series Data Exploration

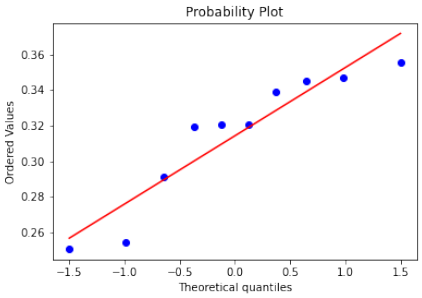


Fig. 1: *Q-Q plot for the homelessness dataset.*

As our data was clean, we then performed data exploration. Using standard methods, visualisation and summary statistics, we were able to uncover patterns and trends in our data. The outcome of doing data exploration is to gain a deep and full understanding of the data at hand and be able to select features from the datasets which are the most insightful for the problem.

The techniques used in the project were calculating summary statistics and plotting Q-Q plots and box plots in order to test for normal distribution, so that we know which methods are appropriate for our data during our analysis. As exemplified in Figure 1, we discovered a lot of our data was not normally distributed, such as that of the homelessness dataset.

We have a combination of data recorded every year or every quarter. We separated the data we were exploring into categories by the frequency of their entries.

1) Quarterly Data: We chose three datasets: homelessness statistics, the average dwelling price and people in temporary accommodation (TA). All three are related to our topic and are recorded quarterly.

The next step was to do some initial feature selection. Firstly, the “Number of households in temporary accommodation at the end of quarter by type of Temporary Accommodation (TA) provided in England, 1998 Q1 to 2021 Q3” dataset, details only one feature.

However, feature selection for the “Housing market: simple average house prices, by new/other dwellings, type of buyer and region, United Kingdom, from 1992 (quarterly)” dataset was more complicated. The three features of this dataset were: the simple average price of all dwellings, simple average house price for first-time buyers and simple average house price for former owner occupiers. For our purposes we were not interested in the difference between a first-time buyer and repeat buyer. Instead, we only considered the average dwelling price.The same approach was applied to the Homelessness Statistics dataset, where we selected the total main duty owed as a relevant feature. In both cases, the other features in the dataset include subsets of the total, which we were not interested in.

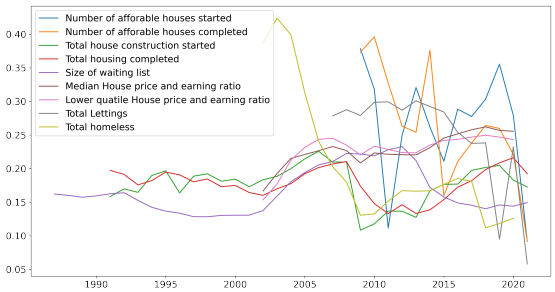


Fig. 2: *The datasets selected for the yearly time series data analysis.*

2) Yearly Data: Five datasets were chosen for time series analysis of data recorded yearly. Their differences are best highlighted by plotting them (Figure 2). This graph shows the datasets' variation in terms of the number of data points. Each of these datasets are independent and do not record the same value. Therefore we do not need to exclude any of them and we keep this collection of datasets as our working database of yearly time series.

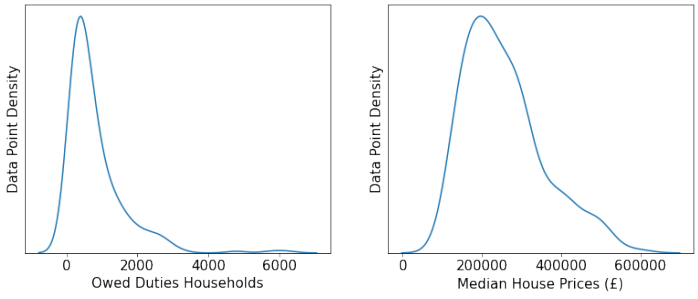
B. Categorical Data Exploration

Before performing data analysis on the categorical data, it is important to select the features which would be suitable for modelling. The objective of categorical analysis is to establish quantified relationships for data related to homelessness and housing for the present situation. Therefore, the "Total Owed a Prevention or Relief Duty" (Total Duty Owed) from the homeless dataset has been selected as the target feature, which is the index representing total homelessness in the country introduced by the data brief.

After deciding the target, all other homelessness related indexes were excluded in further analysis, while the remaining indexes would be the features for further modelling. Some further 6 features are removed because they are the aggregation of information contained by other indexes.

It shall be noted that, unlike the time series, the categorical data exploration will not focus on finding the overall trend. Instead, we explored the distribution and outliers of the data, from which it can be seen that the minimum and maximum can vary greatly.

The datasets selected have an average and median skewness of 2.71 and 2.20 respectively, which suggests that the majority of the datasets are highly skewed distributions. Consequently, this is an aspect we had to consider when selecting our model.



(a) **(b)**

Fig. 3: *Examples of the selected features and target's data distribution, 3a Total Duty Owed Homelessness data is more skewed than 3b Median House Price.*

As shown in Figure 3, the data of 3a has a skewness of 2.90 and 3b's data has only 0.91, thus the latter's data points are less concentrated with fewer outliers. However, normalising the data distribution by cleaning the data points is not possible.

Due to the nature of how data are matched in categorical analysis, removing the more spread out data points of 3b will also remove the entire entry for those local councils. This would lead to a significant loss of data and break the consistency of the analysis. Thereby, one available solution would be to apply appropriate analysis methods capable of modelling data with highly skewed distributions.

IV. Time Series Data Analysis with Trend Studies and Correlation Analysis

With the time series data, we aimed to do three things - to gain an insight into how certain events in the real world impact these figures, to forecast their future values, and lastly, to explore the correlation between different measures of homelessness and related features. We did this by a combination of univariate analysis [IV-A], forecasting[IV-C], and multivariate analysis [IV-E].

A. Univariate Analysis

Time series analysis refers to a variety of methods used to analyse a sequence of data points recorded at consistent intervals of time. The time variable provides additional information as it shows how data changes, evolves or adjusts over time.

There are many different models of time series. This includes but is not limited to the models used in this project - descriptive analysis, explanative analysis and forecasting. Descriptive analysis attempts to identify patterns in time series data, which could be the trend, cycles, or seasonal variation. Explanative analysis tries to understand the data and the relationships within, which include the cause and the effect relationship. Forecasting is explored further in [IV-C].

1) Time Series Decomposition: A time series is made out of several components, trend (the increasing or decreasing value in the time series), seasonality (the repeating cycle in the time series),level and noise.

Decomposing a time series involves thinking of it in terms of a combination of these components. All series have a level (the average value in the time series), and noise (the random variation in the time series), but the trend and seasonality components are optional. The way these components are combined can be thought of in two ways. In additive combination, the time series is composed by adding all the components together. A multiplicative time series on the other hand, is comprised of the components being multiplied together.

Time series decomposition is a useful tool. Each of the components of a time series is an important factor in time series analysis and forecasting, and needs to be addressed at all data analysis stages. Due to the nature of real-world time series data, it is not always possible to cleanly decide a series to be multiplicative or additive.

To figure this out, plotting the time series and or doing some basic summary statistics is a natural starting point. By looking at the plot, it is possible to estimate whether a series was produced by aggregating or summing its components. More precisely, if the seasonality and noise components are independent of the trend of the series, then the series is additive. However, if these two components fluctuate on trend, the series is multiplicative.

2) Time Series Stationarity: Removing the trend and seasonality components of a times series makes it stationary. These fixed components can be removed in order to explore other significant signals, on which one can make predictions. After removing these components, the mean, variance and autocorrelation are no longer time dependent and therefore do not change over time. Autocorrelation is the measure of similarity between a given time series and a lagged version of itself.

The Augmented Dickey-Fuller test is a type of statistical test called a unit root test. The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend. If $p - value <= 0.05$ the data does not have a unit root and is stationary.

B. Time Series Descriptive and Explanative Analysis

For our analysis, we decided to pick two time series datasets, one from each frequency category. As homelessness is our main topic, for quarterly recorded data, the total number of homeless households was chosen. On the other hand, for the analysis and forecasting of yearly data, we chose to study the social housing waiting list, considering that it is a measure of the availability of affordable housing. Both of these time series datasets are the biggest in terms of the amount of data recorded in their respective category.

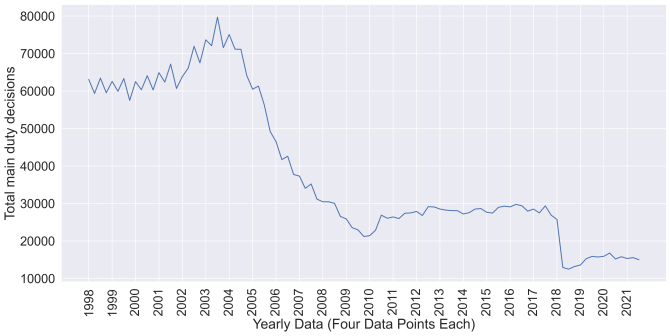


Fig. 4: *The homeless datasets recording the total main duty decisions for an applicant household from 1988 until 2021 (data recorder quarterly).*

1) Quarterly Data - The Homelessness Dataset: Figure 4 shows the chosen time series plotted. At a first glance, the series does not seem to be seasonal but has a decreasing trend overtime. The number of homeless households peaked in 2003 and since then it has not come close to reaching the same number. One possible explanation for this trend change could be attributed to The Homelessness Act introduced in 2002, focusing on homelessness prevention by providing advice and assistance [4].

The total number of homeless households was more or less constant from 2009 with a slightly increasing trend until 2018, where there was a sudden drop. This can be attributed to the introduction of the Homelessness Reduction Act in 2017, which focused on even more prevention and applicants who do not belong to any priority group [14].

When attributing certain policies with achieving the goal of decreasing homelessness, it is important to consider the possibility that the reductions, or any other changes, are a direct result of these acts and policies changing the requirements for being legally recognized as homeless.

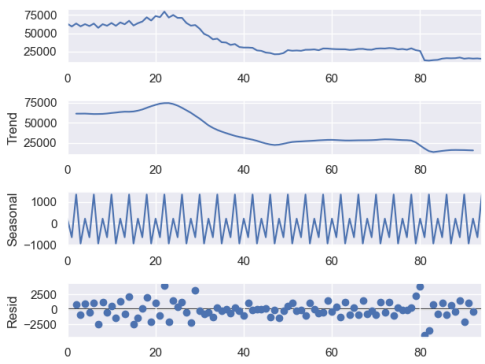
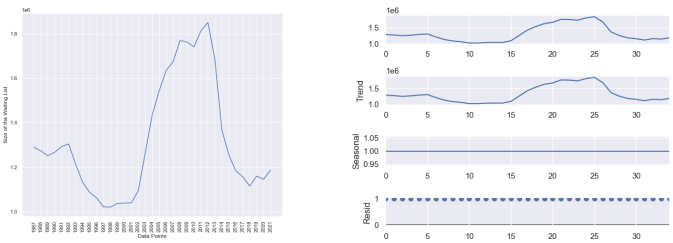


Fig. 5: *Time series decomposition on the homeless dataset, showing the trend, season and residual components of the series.*

For the time series decomposition (Figure 5), the multiplicative model was assumed. The decomposition show that the time series is seasonal and noisy and that our data has an obvious trend and level, suggesting that our data is not stationary. Running the ADF test, the resulting P-value is 0.655808, which confirms this assumption.

2) Yearly Data - The Waiting List:



(a) **(b)**

Fig. 6: *The size of the social housing waiting list, 6a The data recorded once a year from 1987 until 2021 and its corresponding 6b decomposition into trend, season and residual components.*

Social housing waiting list reached its peak in 2012 (Figure 6a). Since then it has been on a steady decline. Looking at Figure 2 temporary accommodation has been increasing until reaching its peak in 2020. These two observations can be explained by looking at the context behind the 2011 Localism Act [5]. Local authority councils were given more flexibility in how they organise their waiting lists. And, presumably, due to insufficient supply of social housing, purged the waiting list by limiting it to people who lived locally for a certain period of time. From 2012, this put more people into temporary accommodation, as there is not enough social housing to go around.

From Figure 6b, we can see that the series is not seasonal and has varying trend. Calculating the ADF test returns a P-value of 0.98552.

C. Time series forecasting

Developed in the 1950s, exponential smoothing has proven to be effective in forecasting time series throughout the years with solid statistical foundations[9] and sound record in performance among various statistical competitions[12]. A more recent example would be a group of researchers implementing the Smoothing method in Cellular Network Traffic Prediction[15].

1) Simple Exponential Smoothing: Exponential smoothing is a forecasting method, using the exponential window function. Unlike the simple smoothing average, this utilises a more sophisticated approach, where predictions use a weighted sum of past observations with exponentially decreasing weight for past observations. We use simple exponential smoothing (SES) as an initial step into forecasting. This model assumes that there is no clear trend or seasonality, as given in the following relationship:

$$s_t = \alpha x_t + (1-\alpha)s_{t-1}, \tag{1}$$

,where s_t is the smoothed value, and $0 < alpha \leq 1$. The smoothed statistic is therefore given by a weighted average of the current observation, x_t, and the previous smoothed statistic, s_{t-1}.

2) Double Exponential Smoothing: As SES does not take into account trends, we seek to expand on this method, by using Double Exponential Smoothing (DES). This involves adding a second smoothing model to capture the trend.

$$\begin{aligned} level: s_t &= ax_t + (1-\alpha)(s_{t-1} + bt - 1), \\ trend: b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}. \end{aligned} \tag{2}$$

To forecast beyond x_t is given by the approximation:

$$F_{t+m} = s_t + m * b_t, \tag{3}$$

The level equation follows a similar logic to that of SES, while the trend equation shows that b_t is a weighted average of the estimated trend at time t based on the previous estimate of the trend.

3) Triple Exponential Smoothing: We further build on exponential smoothing by considering Triple Exponential Smoothing (TES). TES adds support for seasonality, by similarly adding an additional gamma parameter that controls the influence of the seasonality component. We assume that both of our series are multiplicative [IV-B], and therefore show the multiplicative TES model:

$$\begin{aligned} s_0 &= x_0; \\ level: s_t &= \alpha\frac{x_t}{c_t - L} + (1-\alpha)(s_{t-1} + b_{t-1}); \\ trend: b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}; \\ seasonality: c_t &= \gamma\frac{x_t}{s_t} + (1-\gamma)c_{t-L}; \\ F_{t+m} &= (s_t + mb_t)c_{t-L+1+(m-1) \bmod L}, \end{aligned} \tag{4}$$

where $\alpha(0 \leq \alpha \leq 1)$ is the *data smoothing factor*, $\beta(0 \leq \beta \leq 1)$ is the *trend smoothing factor*, and the $\gamma(0 \leq \gamma \leq 1)$ is the *seasonal change smoothing factor*.

D. Results of Forecasting

1) Homelessness Predictions: The DES and TES models were applied to forecast the future number of homeless households. First, using DES methods, we modelled the series by treating the trend component as both additive and multiplicative. In addition, it was found that applying a damping factor decreases the sum squared error (SSE) of the predictions. Applying the damped trend method involves reducing the size of the trend over time until it converges to a straight line. For DES, the model with the best prediction was the multiplicative one with the trend damped. This confirms our assumption from the analysis [IV-B] about data having multiplicative trend. From our analysis we also claimed that our data is seasonal. We modelled the homeless dataset with TES by setting the two studied components, trend and seasonality, to be both additive and multiplicative and switching between damped and non-damped trend. The best combination of features was setting both components as multiplicative and damping the trend. The SSE for this model was 5.382343×10^8. The DES multiplicative damped model that performed the best for DES methods had an error of 8.392230×10^8. Comparing the two best prediction models for DES and TES, the results confirm our analysis that the data is multiplicative with a trend and seasonality.

Disregarding the high error difference, both methods of forecasting produced similar looking predictions. Figure 7 shows the data following a slightly decreasing trend. According to our model prediction, the number of homeless households should decrease in the following years.

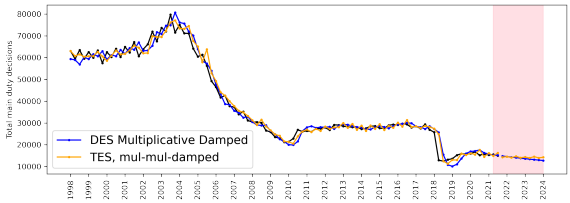


Fig. 7: *TES and DES damped forecasts for the year 2022 until 2026 predicting the number of main duty decision from the homeless dataset.*

2) Waiting List Results : Firstly, the social housing waiting list dataset was modelled with the SES model. Optimising the alpha coefficient at 0.995 produced the best predictions. The SES model resulted in an SSE of 2.787887×10^{11}. Then, we modelled the data with DES, setting the trend to be modelled as additive or multiplicative and adding damping. The model with the smallest SSE result of 1.756359×10^{11} was once again obtained by setting these features to multiplicative and including damping. For DES, the additive model results had a difference of more than 0.1×10^{11} with the multiplicative model, confirming the previous assumption IV-B that the waiting list series data is multiplicative from the decomposition. Figure 8 best showcases the stark difference of modelling with and without trend, as the SES method essentially produced a

flat prediction. Meanwhile, the DES forecasts a steady increase of people applying for social housing in the next few years.

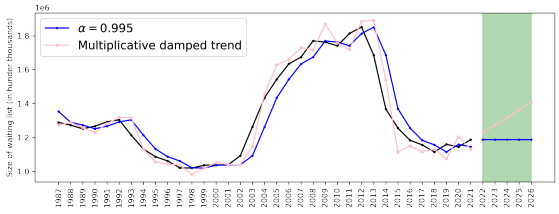


Fig. 8: *The DES and SES prediction for the size of the social housing waiting list dataset, from 2022 until 2026.*

Overall, it has to be pointed out that the irregular events proposed in Section IV-B are prevalent in the studied data. One could say it is the core of the data, as the definition of homeless and the requirements to be put on the social housing waiting list are all dependent on the current acts and policies in place. The predictions made by our analysis could be very relevant in the short term. However, irregular events like new acts and policies or political change, which could occur at any time, would render our predictions irrelevant. Considering that we are currently experiencing one of the most significant and damaging irregular events, an economic crisis caused by a pandemic, we should acknowledge that reality will probably fail to meet the proposed predictions.

E. Multivariate Analysis

1) Spearman's Rank Coefficient: Arguably, the most well-known correlation method is the Pearson Coefficient Correlation (PCC) method. However, our data fails to conform to its assumptions, and in particular, its restriction for normally distributed data. Instead, we use rank correlation methods to quantify the association between 2 variables. One such method is the Spearman's Rank Correlation (SRC), as suggested by multiple literature[18]. A further advantage of SRC is that it associates variables with a monotonic function, and therefore does not assume a linear relationship.

The SRC's coefficient ρ is given by[18]:

$$\boldsymbol{\rho} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{5}$$

$\boldsymbol{\rho}$ = spearman's rank correlation coefficient;
d_i = difference between the two ranks of each observation;
n = number of observations.

Intuitively, SRC calculates a Pearson's correlation based on the rank values. However, this method does not provide insights into signal dynamics and is therefore unable to differentiate between the independent and dependent variable. We therefore use cross-correlation to gain more insight into this aspect.

2) Cross-Correlation: Cross-correlation measures similarity between two series as a function of the displacement of one relative to the other. It is therefore useful for determining the time delay between two signals in our time series analysis. It is calculated by computing the cross-correlation coefficient, creating a lag by shifting the series, and repeating these 2 steps until necessary. For discrete data, such as ours, cross-correlation is given by:

$$\begin{aligned} Y(n) &= \sum_{i=-\infty}^{\infty} h(n-1)x(i) \\ &= \sum_{i=-\infty}^{\infty} h(n)x(n-i) = h(n) * x(n) \end{aligned} \tag{6}$$

where h and x are the two time series data with a time lag of i. A restriction on cross-correlation is that the data must be evenly sampled. However, this is not an issue as all our time series data meets this criteria.

F. Cross-Correlation Results

SRC (P-values)	LQ-House Prices	LQ-Household Income	LQ-Affordability
Waiting List	-0.3515 (0.1340)	-0.3193 (0.1827)	-0.2344 (0.3340)
Homelessness	-0.7225 (0.0004763)	-0.7544 (0.0001901)	-0.2915 (0.2260)

TABLE I: *SRC coefficient and their corresponding P-values for housing affordability and homelessness*

For our multivariate analysis, we aim to explore the correlation between housing affordability for the population in the lower quartile (LQ) and homelessness. We use lower quartile as a measure, because we want to focus on the population who struggle to afford housing. Table I shows the SRC values. Surprisingly, all the values show negative correlation. We take our null hypothesis to be that there is no correlation between the data, with a p-value of 0.05. The p-values given in Table I suggest there is little evidence to reject our null hypothesis and we are therefore unable to conclude that the data are correlated. However, by plotting the normalised trend during data exploration, we discovered that there seems to be strong correlation between waiting lists and the affordability data for a subset of the dataset, so we perform SRC again, but only on the correlating subset, namely the years prior to 2013:

SRC (P-values)	LQ-House Prices	LQ-Household Income	LQ-Affordability
Waiting List	0.7581 (0.004272)	0.8671 (0.0002598)	0.8242 (0.0005302)

TABLE II: *SRC values for only housing affordability*

The new results show strong correlation between the number of households on the social housing waiting list and the corresponding variables. One may expect lower quartile earnings and households on the waiting list to be negatively correlated, but this is not the case. This is due to a disproportionate

increase in the price of rent relative to earnings. As a result, income is not a good indicator of housing affordability. A more informative measure would therefore be the lower quartile ratio between rent price and income, as it indicates affordability of rent. In contrast to the first SRC analysis performed, the p-values here are all significantly smaller than 0.05. As a result, there is significant evidence to suggest that the affordability of housing is positively correlated with the waiting list for social lets. One should also note that although these variables are highly correlated, there is no evidence to suggest a causation as these factors may only be indirectly related, for instance. While this subset yielded results showing strong correlation, it brings the question as to why the data stopped correlating from 2013 onwards. This can be traced back to the Localism Act [5], which came into effect into 2012, subsequently removing thousands from the waiting list. While the size of the waiting list dropped, despite the worsening situation of housing affordability, this is likely due to the new policy change, rather than an improvement in the housing situation.

Moreover, while correlation for waiting lists improved, this was not the case for homelessness. This is due to external factors, namely policy and government changes affecting the way data is collected, as mentioned in section [IV-B].

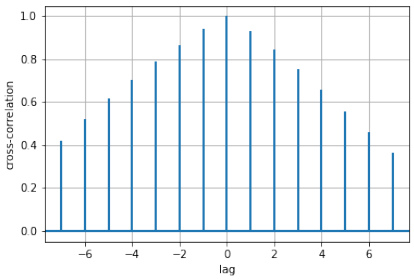


Fig. 9: *Cross correlation between lower quartile ratio and waiting list*

Performing cross-correlation on our datasets, we found that there was no time lag as the highest correlation occurred at a time lag of 0. One may have expected a lag as it takes time for affordability of rent to affect the livelihood before households start applying for social housing. However, this is not reflected in our results, perhaps due to the sparsity of our data. Each data point is a year apart and so if the time lag is a shorter time frame, our data will be unable to detect this.

V. Categorical Data analysis with Machine Learning Models

As discussed in Section III-B, we needed a method to bypass the high skewness in the distribution of the data points in the indexes. We therefore decided to apply two variations of the tree based models to analyse the categorical data in this project. They are the Random Forest Models and Extreme Gradient Boosting Models (XGBoost).

The two variations were chosen for their ability to quantify the relationship between the features into a set of values named ”importance scores”. This is effectively the quantification of the correlations between the features and the Total Duty Owed Homelessness, helping us achieve the second objective of this project, as mentioned in Section I.

Before moving onto the more intricate details of modelling, it should be noted that the randomness in the computation process can be significant to the tree based models. When the models are tested to optimise over 20 iterations, the randomness caused the overall mean R^2 score of the models to be decreased as low as below 50%, which showed no sufficient evidence of correlation between model's estimation to the corresponding data recorded.

Therefore, the purposes of implementing two models in this section were twofold. Firstly, it maximised the reliability of the results by cross comparing the accuracy of the models. Secondly, we were able to seek different angles in assessing the situation by comparing the differences between the set of importance scores generated by the models.

A. The Random Forest Model

The random forest algorithm, proposed by L. Breiman in 2001 [3], has been extremely successful as a general-purpose classification and regression method. Many studies have applied it to research the correlation between related indexes, such as an investigation of factors of rental prices[8].

Regarding the approach of random forest modelling, it combines several randomised decision trees and aggregates their predictions by averaging. This has shown excellent performance in settings where the number of variables is much larger than the number of observations. Thus, the structure of a random forest model can be divided into two parts: decision trees and bagging.

1) Decision tree: Random forests are built on the basis of decision trees. Tree learning comes closest to meeting the requirements as an off-the-shelf program for data mining. It is invariant to scaling and various other transformations of eigenvalues, robust to the inclusion of irrelevant features, and yields a checkable model. However, they are rarely accurate.

In particular, trees that grow very deep tend to learn highly irregular patterns: they overfit the training set, i.e. have low bias, but very high variance. Random forest is a method of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing variance. This comes at the cost of a small increase in bias and some loss of interpretability. The process implemented is the bagging of trees.

2) Bagging: The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

1. Sample, with replacement, n training examples

from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \left(\frac{1}{B}\right)\sum_{b=1}^{B} f_b(x'), \tag{7}$$

or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of multiple trees is not. Moreover, bootstrap sampling is a way of decorrelating the trees by showing them different training sets.

3) From Bagging to Random Forest and to the Model: The above procedure describes the original bagging algorithm for trees. Random forests also include another type of bagging scheme. They use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features[3]. This process is sometimes called "feature bagging".

Thus, the new sample weights after the i^{th} iteration of random feature bagging, the weighting of features would be as follows:

$$\hat{w_i} = \hat{w_{i-1}} * e^{\alpha}, \tag{8}$$

with α assessing the influence throughout different iterations, based on the measurement of uncertainties for the feature bagging in the i^{th} iteration.

As such, the feature bagging procedure can output a more accurate estimation of the features' importance weighting $\hat{w_i}$, i.e. the importance score. Then, via operating multiple iterations of random forest modelling over different random seeds, the effect of randomness can be minimised by choosing the most optimal model.

Thus, the algorithm of exhaustive search, the *GridSearchCV*, has been implemented to optimise the parameters, and we looped through 20 random seeds to optimise the performance. Some of those less significant indexes are removed, because their influence in the overall model are near zero to none.

B. XGBoost Model

Developed in 2014, XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable [19]. The major difference between random forests and XGBoost application, is that XGBoost uses boosting [6] and the Gradient-Boosted Decision Trees (GBDT) method instead of bagging. Related works include a investigation of the testing accuracy of Chronic Kidney Disease Diagnosis[13] and risk assessment of flash flood in southeastern Yunnan[11].

1) Boosting: In boosting, weak learners are trained sequentially, which means a series of models are built. In each training iteration, the samples with incorrect prediction in the previous model were weighted more than those correctly predicted. This strategy helps the algorithm to know which parameters need to be improved. Similar to bagging, for the final prediction, it takes the weighted average of all weak learners with its own re-weighted data x as:

$$\hat{f} = \left(\frac{1}{B}\right)\sum_{b=1}^{B} w_b \cdot x_b. \tag{9}$$

.

2) From GBDT to XGBoost: Similar random forests, GBDT uses decision trees as base learner. For one tree at time step t of the model's operation, the algorithm aims to build a decision tree to predict not the target variable's values, but the error calculated between the predictions of last time step $t-1$ and the true values. At the end, the sum of every single weak learner will become the final prediction.

XGBoost, the improved version of GBDT, introduces various regularisation techniques to reduce overfitting and model complexity and improve overall performance. Runtime is also improved as XGBoost can be implemented in parallel. Despite XGBoost having those benefits, there are more hyperparameters that need to be tuned as it introduces more concepts. This makes the tuning process become much more difficult in order to get the best set of hyperparameters and requires users to have a deep understanding on the fundamental theory of XGBoost. However, XGBoost's boosting method of feature selection and ranking has proven to have overall improvements on the models' accuracy. Consequently, when applied together with the Random Forest, it provided a more accurate quantification of the features' correlation to the Duty Owed Homelessness.

The tree model is generally powerful but still has its limitations. For example, r forests may suffer from overfitting on certain datasets that have large noise. For XGBoost, its complicated theory means that it consumes lots of memory, especially for high-dimensional features. Additionally, heavy cache miss may affect the performance of XGBoost.

C. Results and Discussions of the Two Models

The random forest model is used to predict the number of homeless as it can assess the importance of each feature. In our implementation, we used a 80-20 training and test split. Moreover the output score was be amplified to 100 times the original coefficients output by the final $\hat{w_i}$, to emphasise the differences. The random forest results in a r2-score of 0.7661 and the rank of importance of features are shown in Figure 10.

From Figure 10, the most important factor is the population in the local authority council. There are 5 features ordered by importance: Starting Affordable Home Ownership Houses,

Completed Affordable Home Ownership Houses, Starting Affordable Rent Housing, Completed Affordable Rent Housing, Completed Market Supplied Affordable Housing. Evidently, features relating to affordable homes have a significant impact on homeless.

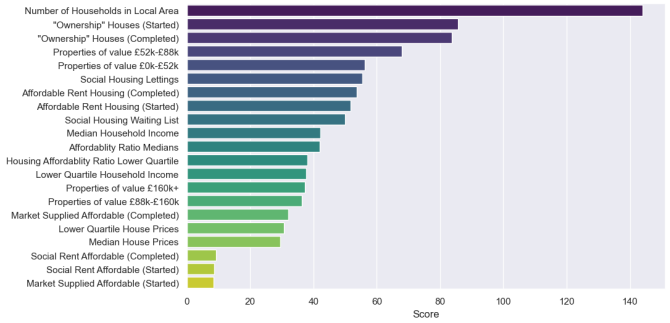


Fig. 10: *Feature Importance Weighting Based on the features' respective Importance Score in the Random Forest Model with optimal accuracy.*

XGBoost used the same source dataset as the random forests model, resulting in an r2-score, of 0.8312, a improvement of 8.49% compared to the previous model. The rank of importance of features created by XGBoost is shown in Figure 11. Features with a higher significance score means that it is more correlated with the goal of the prediction, namely, the number of homeless.

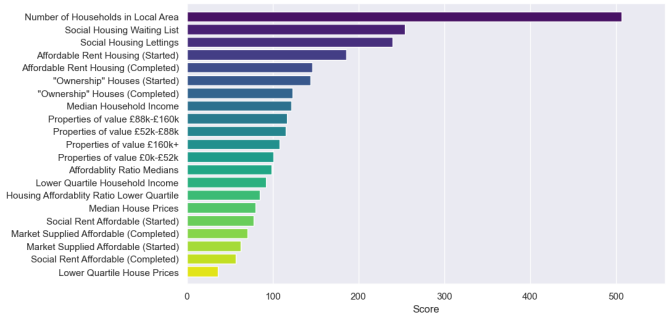


Fig. 11: *Features' Importance Score Based on the features' respective weights in the XGBoost model.*

The feature that ranked first is the number of households in the area which is two times greater than the second most important feature - the number of people on the social housing waiting list. The next most crucial feature is the number of lettings in social housing, followed by the number of rentable and affordable housing that has started being built.

By comparing the rankings of feature importance between random forests and XGBoost, it is clear that the order of crucial features is not the same. However, the difference is acceptable because these two models use different strategies to weigh the importance of features. Although the rank is not the same, we can still find similarities between the two. Features that were ranked with high importance by the random forest model can also be found at the top of the XGBoost rankings.

Combining the information given by both models, it is easy to find out the most important feature is the number of households in the local. This makes sense because the number of homeless inevitable increases as the size of the city gets larger. Government should formulate each policy to flexibly respond to population growth by considering the relationship between the number of homeless and the crime rate.

The features associated with the constructions of social and affordable housing by the government can be found at the top in both charts. This suggests that the problem with homelessness can be solved by increasing the supply of affordable housing. In fact, other academic research has suggested that in order to solve the homelessness problem, the government needs to pay attention to policies relating to affordable housing if they hope to achieve long-term success in helping homeless households[2].

The size of the social housing waiting list and the number of social housing lets are also ranked higher among the other features. This potentially indicates that to decrease the number of homeless, the government can introduce policies to attract housing associations to construct more social housing.

VI. Conclusion and Evaluation

In conclusion, multiple pieces of evidence have suggested clear correlations between the progress of the government housing policy and the positive changes related to the homeless population in England throughout the past and present.

The Spearman's Rank Correlation and cross-correlation have both shown to be excellent and successful measurements in quantification. For the time series analysis, one significant finding would be associated with the high positive correlation between the social housing waiting list and the affordability of housing. In combination with the forecasting data, which predicts a growing increase in the waiting list, this is likely a sign of missing government action in maintaining the supply chain of the social housing scheme.

Conversely, the exponential smoothing forecasting on the homeless population predicted a decreasing trend. It should be acknowledged that the datasets analysed are relatively small, which limits the reliability of our results. The limitations of the application of forecasting methods on our data also relate to their dependency on irregularities such as policy change, one example being the sharp trend change in the homeless population before and after 2003 [4].

For the categorical analysis, the two models have shown high accuracy based on the R-Squared test and thus exhibits sufficient evidence of correlation between the features and the duty owed homelessness data. It supports the time series by showing that the social housing waiting list data has a high importance score and, therefore, a high correlation with the target.

There were some other notable findings. Firstly, the categorical analysis shows that homelessness in England is more common among more populated regions by rating the number of households as the most relevant feature among both models. Secondly, the features relating to the construction of the policy

housing, such as the "Market Supply Affordable Housing", could signify those schemes are poorly targeted or designed. Last but not least, the value of properties, the region's household income, and the affordability of housing have average or low correlation measurements, suggesting the occurrence of homelessness in one area may not be directly related to their local economy.

As for the limitations, both models perform the same analysis in different mathematical procedures. As a result, the importance score may not lead to the same conclusion.

For both time series and categorical analysis, if more time and resources were available, they can be improved in the following ways. First of all, this project's research has been restricted to the content of the GSS housing statistics interactive tool. Thus, one may go beyond the tool, searching other statistics in fields other than housing, such as the national economy and the distribution of ethnic groups across the country. Secondly, one may extend the categorical and time series analysis models to Scotland, Wales and Northern Ireland, then discuss the nature of homelessness and related statistics across the UK. Lastly, one may test and implement different models for better performance, such as for the time series analysis.

Overall, both the time series and the categorical models built with data analysis regarding the housing market policies and homelessness statistics, have successfully investigated their past and explained their present. But the predictions as part of the analysis of the future are limited by the irregularities of government policy decisions. Therefore, concerning the future of the homelessness problem, the findings of the project urge the government to pay more attention to the durability of the policies such as social housing and affordable housing scheme. In such a way, the changes shown in the predictions will be more than just a short term estimation based on irregularities of policy decisions.

References

[1] *ADS year3 project 8: Codes, Results and Data Storage*. GitHub, Apr. 2022. URL: https://github.com/SIthodox2286/ADS_year3_project_8.

[2] Jocelyn Apicello. "A paradigm shift in housing and homeless services". In: *The Open Health Services and Policy Journal* 3.1 (2010).

[3] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[4] Crisis. *Crisis — Together we will end homelessness*. Crisis, 2009. URL: https://www.crisis.org.uk/ending-homelessness/the-plan-to-end-homelessness/.

[5] Dawn Foster. *Why council waiting lists are shrinking, despite more people in need of homes*. May 2016. URL: https://www.theguardian.com/housing-network/2016/may/12/council-waiting-lists-shrinking-more-need-homes.

[6] Yoav Freund and Robert E. Schapire. "Experiments with a New Boosting Algorithm". In: *International Conference on Machine Learning*. 1996, pp. 148–156.

[7] Jeremy S Godfrey. *Rewriting homeless identity : writing as coping in an urban homeless community*. Lexington Books, 2016.

[8] Lirong Hu et al. "Monitoring housing rental prices based on social media". In: *Land Use Policy* 82 (2019), pp. 657–673. ISSN: 0264-8377. DOI: https://doi.org/10.1016/j.landusepol.2018.12.030.

[9] Rob Hyndman. *Forecasting with exponential smoothing : the state space approach*. Springer, 2008.

[10] Dan Lewer and et al. Aldridge. "Health-related quality of life and prevalence of six chronic diseases in homeless and housed people". In: *BMJ Open* 9.4 (2019). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2018-025192.

[11] Meihong Ma and et al. Zhao. "XGBoost-based method for flash flood risk assessment". In: *Journal of Hydrology* 598 (July 2021), p. 126382. DOI: 10.1016/j.jhydrol.2021.126382.

[12] et al. Makridakis. "The accuracy of extrapolation (time series) methods: Results of a forecasting competition". In: *Journal of Forecasting* 1 (Apr. 1982), pp. 111–153. DOI: 10.1002/for.3980010202.

[13] Adeola Ogunleye and Qing-Guo Wang. "XGBoost Model for Chronic Kidney Disease Diagnosis". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.6 (2020), pp. 2131–2140. DOI: 10.1109/TCBB.2019.2911071.

[14] Full Fact team. *Homelessness in England*. Full Fact, Oct. 2016. URL: https://fullfact.org/economy/homelessness-england/.

[15] et al. Thanh Tran. "Cellular Network Traffic Prediction". In: *Journal of Information and Communication Technology* 18 (Jan. 2019), pp. 1–18. DOI: 10.32890/jict2019.18.1.1.

[16] M Tong et al. "Factors associated with food insecurity among older homeless adults: results from the HOPE HOME study". In: *Journal of Public Health* 41.2 (Apr. 2018), pp. 240–249. ISSN: 1741-3842. DOI: 10.1093/pubmed/fdy063.

[17] et al. Tsai Jack. "Needs of homeless veterans: 5 years of the CHALENG Survey 2012–16". In: *Journal of Public Health* 41.1 (May 2018), e16–e24. ISSN: 1741-3842. DOI: 10.1093/pubmed/fdy076.

[18] Stéphane Tufféry. *Data mining and statistics for decision making*. Wiley, 2011.

[19] *XGBoost Documentation — xgboost 1.6.0 documentation*. Readthedocs.io, 2022. URL: https://xgboost.readthedocs.io/en/stable/.

VII. Refelctive Discussion, Yuyang(Eric) Liu, Student No. 1982734

A. Introduction

The process of developing our group's project, as discussed extensively in the report, can be divided into six stages. The topic introduction, data collection, data selection, data exploration, analysis with modelling and results discussion with conclusion. Among all stages, the management of datasets have also been practiced, and the group evaluated the project's achievements and limitations at the end.

B. Related back to the Unit's Lecture

The taught techniques of our units are practised at each stage. For example, in the topic introduction, my groupmates and I identified the type of the data stored in the datasets to be overall numerical but could be distinguished further according to how they are formatted in their tables: Time-series and Categorical (each entry referring to one local authority). Another example could be managing and storing our data with GitHub Web Application. Moreover, multiple machine learning tools are implemented for analysis purposes, including Spearman's Rank Coefficient, Exponential Smoothing and Tree-based models.

However, some of the data exploration tools and taught techniques of the unit are not practiced, such as Query perturbation for privacy protection. Because the design of the GSS database and the new housing data exploration tool, which involves professional working codes in the storage and anonymizing of datasets.

C. Contributions along the way

Regarding the work distributed and completed during the project, my contribution includes:

1. I set-up our GitHub page and the Overleaf project for storing our work and datasets we interested (which are downloaded once selected), then I was in charge of maintaining them;

2. Raise my idea about what should be our project's theme and objectives by reading through the database, discuss them with the group (Although in the end, we decided to focus on Linda's idea about studying social housing and homelessness);

3. Select and explore relevant datasets, choosing effective and informative means of machine learning models for them. This was when I noted the categorical datasets could be analyzed on a "Local Authority Council" basis;

4. Model building and tests. This was when I noticed the skewed distribution of the categorical datasets, and then decided to try tree-based algorithms. Firstly, I built the Random Forest model and then Yuze Ma introduced and practiced the XGBoost;

5. Together with the team, I recorded the video (including the writing of my scripts, as others wrote theirs), whilst writing the report;

6. Edit the video and the report with the group.

For the group, the tasks were shared equally among the members in the earlier parts of the project, then split into two teams focused on time-series and categorical analysis, respectively. However, the distribution of the analysis tasks are not restrictions. To be more precise, Linda and Lokhei were in the time-series team, while me, Yuze and Haoran were in the categorical team. Moreover, people in different team would still help each other, such as in debug and seeking others' opinions.

D. Strength and Weakness

As a group, we managed to not keep doubts and disagreements among us, but always make our plan at each stage clear and sound among all members. We had our long and fruitful meeting twice every week, whilst daily text message and GITHUB update ensured everyone would be up to date. On the other hand, the project has been successful in obtaining detailed evidence of correlations among the homelessness dataset and housing policy statistics via modelling. We tried a range of analysis and modelling techniques. Each of the group member had their contribution in the data collection, then worked with each other even closer when we divided our team into two, to do the time-series analysis and categorical analysis, respectively.

However, we spent more time and effort in the technical in-depth of the report, covering all essential parts of our modelling process, whilst less effort has been spent on explaining the results in detail. Such as explain and explore what social housing does right to make its data highly correlated to the duty owed homelessness.

E. Personal Reflections

Personally, beyond the contributions described among the points listed earlier, I managed to organize and complete my tasks in a well organized schedule. During the teamwork part of the project, I have also managed to make good communication with the other members and staying to be active during our meeting with our TA. Especially for the earlier parts of the report, with less machine learning and more statistical analysis.

Reflectively, in this project, I learnt a lot about how to keep a team working under a tough schedule, with multiple deadlines approaching. Especially for those who were taking the 40 credit end of year report. As a member of the Engineering Mathematics department, working with people from Computer Science have also been a valuable experience to me. I'm especially interested in how they did to keep improve their models by always look for better methods and techniques. Furthermore, the project have also been a valuable experience for me to practise our studies in lectures in researches.

Overall, the group managed to establish sound understanding on our data with sufficient technique in-depth. On the other hand, our objectives raised in the introduction of the report have all been achieved to a sufficient degree of technical in-depth.