Fly-Tipping in the Lancaster City Region

Group 4 - Lancaster City Council

Arthur Deng [*], Avi Holden[†], Hengliang Tang[‡], Prathamesh Kulkarni[§], Rhys Peploe[¶] and Ryan Noonan [‖]
[*]35997444, [†]36031730, [‡]35811652, [§]36004026, [¶]36020881, [‖]36051936

Explanation of member's inputs:

- Arthur - Data visualisation, Conclusion in report.
- Avi - Exploratory analysis, Data visualisation, 'Top 4/ Bottom 4 Deep Dive', Pre-processing & Deep dive report sections.
- Hengliang - Data visualisation and Future work report sections.
- Prathamesh - Clustering, Interactive Maps, Video for presentation.
- Rhys - Cover page, Time series analysis.
- Ryan - Introduction section in report, initial clustering experiments (not included in report).

Fig. 1. Fly-tipping on Bold Street, Heysham North. Photo from Cllr Roger Cleet

School of Computing and Communications
Lancaster University

LANCASTER CITY COUNCIL
Promoting City, Coast & Countryside

Lancaster University

I. Introduction

Fly-Tipping poses great danger to the environment and human health, and is continuing to grow at an unprecedented rate. Statistics from (1) reveal that in England alone, between 2018 and 2019, a total of 1.072 million fly-tipping incidents were reported, highlighting the scale of the problem. Moreover, many of these fly-tipping incidents are for household waste, making up 65% of all fly-tips. To the average fly-tipper, dumping a little rubbish may seem harmful, but it can have far reaching consequences on the environment and public health. The environment can be affected in various ways, including choking hazards for other animals, rodent infestations due to the availability of food and waterway pollution. All of these factors can make our lives more difficult in the long term, reducing biodiversity and making it easier for disease to spread. In the short term, however, the presence of fly-tips can affect the health of those who interact with fly-tips. A study by (2) measured the incidence of Hepatitis B (HBV) among waste workers and the general population in the United States. Waste workers were far more likely to test positive for HBV than members in the general population and this was attributed to various aspects of waste collection, including being spiked by needles. According to worldwide estimates (3), there were 290 million people living with chronic HBV in 2019, and there were 800,000 deaths from complications (including liver cancer and cirrhosis). Although, the risk of infection needn't be limited to waste workers as it could also affect small children who are unaware of the risks. These risks motivate the work undertaken in this project.

Due to the unprecedented scale of the problem, the scope of our project is limited to Lancaster, England, and data from the Lancaster City Council (LCC) will inform our analysis. Exploratory Data Analysis was used to identify the areas, fly-tip sizes and fly-tip types which are most common, to recommend where LCC should prioritise it resources. Using the raw fly-tipping dataset, an interactive graph was created to allow the council to visualise hotspots and find common locations (a particular part of a river, for example). These techniques form the basis of our methodology, using visualisation, time series analysis and clustering techniques. There are many findings from our work. For example, household waste appears to be the most common form of fly-tip, which is consistent with the national statistics. However, back alleyway fly-tips appear to be more common in Lancaster than the national statistics suggest; highway fly-tips are more common in the national statistics. Furthermore, our data deep dive suggests that proximity to recycling centres and infrequent bin collection days may have significant effect on fly-tipping incidents and is therefore a recommended future area of exploration.

Research Questions/ Objectives:

1) What are the hotspots?
2) What do they have in common?
3) Can locations be predicted?
4) Is there a link between the type of tipping and a location?
5) Should there be areas to which enforcement officers should pay particular attention?

II. Methodology

This report is based on data contained in the *Easting-NorthingArcGISTEST.csv* file provided by the Lancaster City Council. The file contains 23 columns and 17734 rows, displaying information regarding fly-tipping incidents in the local area from September 2015 to October 2021. All of the necessary information was contained in this file, so data collection or integration was not needed for our analyses.

Before we could work with the data to achieve our objectives, the file had to be pre-processed. We decided to use R Studio for this due to the wide range of appropriate built-in functions. When reading the file in, we made sure that all missing values were recognised in the dataframe by using the *na.strings* parameter within the *read.csv* function.

The first step in our pre-processing involved deciding which, if any, data columns would not be useful in our analysis and could therefore be removed. As we were not interested in when each incident was responded to and closed, we first removed the columns *ResponseDate* and *ClosedDate*. Then, as a large proportion of the entries in the *Postcode* column were empty, we decided to remove the column as no useful analysis could be conducted with it. Finally, we chose to remove the columns *SRequestId*, *LPIKey*, *UPRN* and *Fiscal_Qtr*, as all of the information contained in these was already known from other columns and/or was not necessary (e.g., *SRequestId* was just the unique identifier of each incident).

Then, we used the *as.date* function to convert the format of the *ReceivedDate* to the class "Date". This also disregards the times that were included in the entries from this column, which were all "*00:00:00*" (hence why we decided that they did not need to be included).

Next, we observed that some entries in the *Year_Received* column were not matching up to what we knew about the data, so the corresponding rows needed to be removed. For example, there were 4 entries with *Year_Received*=1753, which was clearly not the case as the council had only provided us with information between the years 2015 and 2021.

Furthermore, we identified a very small amount of missing data entries in the *Waste_Type*, *Land*, *Size* and *WardCode* columns. As it was agreed that these would be important factors for our exploratory analysis, we concluded that also removing the corresponding rows of data in these cases was the most suitable course of action.

A. Clustering

Clustering is the process of categorising data. Clustering is an unsupervised process. Data points in the same category are closer to other data points in the category, while data points in other categories are different. Essentially, it's a group of objects arranged according to their similarity and difference. A wide variety of applications use clustering; here, it serves as a data summarisation technique. Summarising data can reduce the amount of information needed to be present in

an application and make it easier to understand (4) (5). The data were clustered using the K-Means algorithm. The K-means algorithm selects centroids k at random, calculates the distance between each data point and the centroid, and updates centroids' positions. The process is repeated until all data points are assigned to their nearest centroid. In the end, there are k clusters. Instead of selecting k randomly, it is possible to identify its value. The elbow method is one such method. The elbow method plots the average squared distance between each data point and its nearest centroid as a function of the number of clusters. This plot shows the point in the curve where the inertia is no longer dropping fast. This point refers to the optimal value k. A model was constructed based on the latitude and longitude of each fly-tipping incident. The objective of clustering was to identify high fly-tipping areas based on two variables. Thus, fly-tipping reports from locations with higher levels of fly-tipping are grouped.

B. Time Series

Time series analysis provides a useful tool in forecasting future events and in the content of fly-tipping, we have taken the series to be the counts of incidences per month over the years that data is available for. The Autoregressive Integrated Moving Average (ARIMA) method is utilised as autoregressives terms model on the variable's own lagged points; uses the past to predict the future, which is relevant since high prevalence is likely to influence the future occurrence of tipping in the same area. Additionally, the moving average element of the ARIMA function incorporates the dependency between a new data point and the residual error applied to lagged observations (6). The model also allows for differencing the values to achieve a stationary series, where the difference between one point and the next becomes the data points in the function. Removing non-stationarity is required in order to implement the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), which will assist in discovering the appropriate model parameters (7). A stationary model requires a constant mean, a finite variance and every lag has a constant variance (8), and can be formally determined by the Augmented Dickey Fuller (ADF) test which performs a hypothesis test with the null being the series is non-stationary and the converse is the alternative. These three parts are summarised in the model coefficients ARIMA(p, d, q), where there is p autogressive terms, q moving average terms and differenced d-times. The Akaike's Information Criterion (AIC) and log likelihood are used for model selection, as well as diagnostic checks on the residuals of the model to ensure the data is reasonably fitted.

Two wards were selected to conduct the time series prediction on, Heysham North and Harbour, since they are amongst the top locations for incidences across the region. The choice to model on wards and not the region or area was to be able to localise the forecast and attempt to provide usable, focused information.

III. Results

A. Interactive Map and Clustering

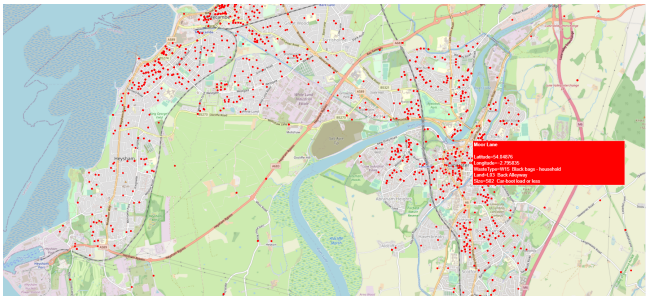

Fig. 2. Screenshot of the interactive map, where an example is shown to showcase how the information related to that data point is displayed by hovering over it.

This project utilizes interactive maps in three different ways. First, to showcase all instances of fly-tipping; second, to showcase any information that is related to a given location; and third, to display the different clusters. It is possible to use these maps to meet various requirements of a specific location, including the size of the bin that needs to be installed, the size of the truck required to collect garbage, and the type of waste generated. The code can easily be adapted to display any type of information. By using clustering, these maps make it easier to identify areas with more cases. This allows decisions to be made regarding areas with more fly-tipping.

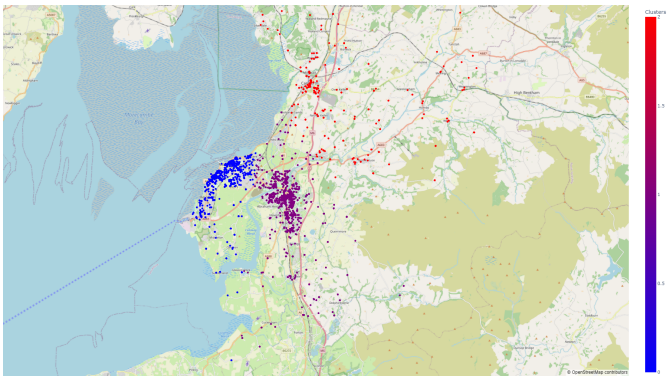

Fig. 3. The different clusters formed after clustering the latitude and longitude data. The three clusters represent Lancaster, Morecambe, Carnforth.

B. Time Series

Initially, by inspecting the graphs of grouped incidences by time, the two wards clearly display a non-stationary series, as the counts per month vary wildly in places. After applying the ADF test to the raw observations, they produced p-values of 0.0597 (Heysham North) and 0.1731 (Harbour), which are insignificant at a 5% level, thus implying that insufficient evidence exists to reject the null hypothesis and labelling them as non-stationary. As a result, differencing is required; by repeating the test on transformed data, the p-values are <0.01

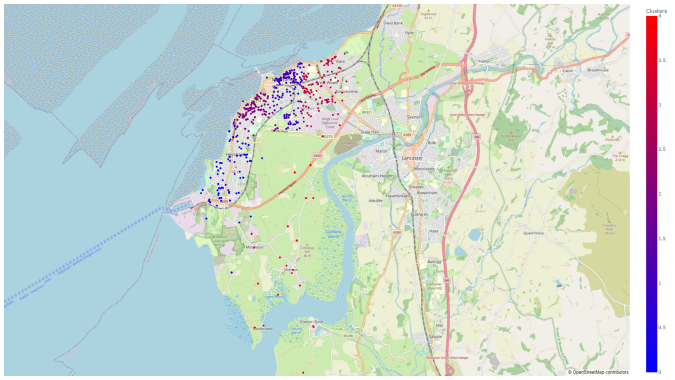

Fig. 4. The figure shows the clusters within the Morecambe cluster. It can be seen that the clusters are more dense along the seaside areas of Morecambe, indicating a higher number of cases of fly tipping in those areas.

for both wards and, hence, significant, so the ARIMA model should include $d = 1$.

Following that, the ACF and PACF plots are used to suggest the model parameters p & q, where a ACF plot experiences a sharp drop off after lag q and, conversely, the peaks in the PACF suggest possible values for p (9). For the Heysham North ward, the PACF indicates 1, 2 and 10 as good candidates for the number of autoregressive terms in the models, while the ACF showcases significant drops between lags 3 and 4, as well as 9 and 10. With Harbour, the only significant peaks are at 1 and 14 for the PACF, and the ACF has a very gradual decline, so looking for sizable drops, the first one occurs as late as between lag 13 and 14. Armed with this, various models were fitted to find a model with a powerful performance, while not excessively complicating it or overfitting the data; the final models given this are (2, 1, 10) for Heysham North and (6, 1, 14) for Harbour, with the model statistics given in Table I. Diagnostic tests were conducted to measure the fit of these models to the respective data, which included a plot of the standardised residuals, which showed no skew or abnormal behaviour; the ACF of the residuals showed a large peak at lag 0 and insignificant points thereafter, as expected; thirdly, the p-values for the Ljung Box statistic were all above 0.9 showing a good fit.

Ward	Model	Log Likelihood	AIC
Heysham North	(2,1,10)	-363.20	750.41
Harbour	(6,1,14)	-306.98	653.96

TABLE I
MODEL STATISTICS FOR THE OPTIMAL SOLUTIONS FOUND

The models can now be used to forecast the incidences of fly-tipping in each region for the near future. 10 months of actual data are withdrawn to use as test datapoints, and 15 months will be predicted, including the ten held back, meaning that this will forecast from January 2021 to March 2021, with November onwards having unknown actuals. From Figure 5, the model seems to reflect the test points well, generally over-estimating the number of incidences per month, and producing a fairly conservative prediction for upcoming months. The forecast generated implies that there will be around 35-40 records per month with a slightly upward trend, furthermore, the upper bound for this time period is also increasing, staying under the 110 mark though, so upon repeated sampling, it is expected that 95% of the time, the number of incidences will be under this value. More autoregressive terms would mean the series would take observations further in the past to dictate what will happen in the future, so the true model may be different than (2,1,10), but, this would be investigated with more data as the council continues to track fly-tipping.

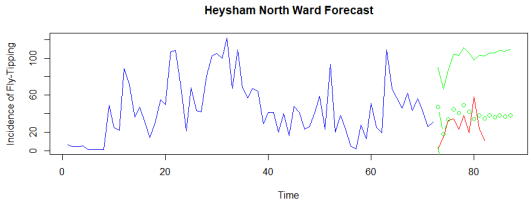

Fig. 5. Time series forecast for the Heysham North ward. The blue line indicates actual observations used in the forecasting, the red is actual data but used as test points. Green circles are the expected count of incidences, along with green 95% confidence intervals.

Turning to Figure 6, the pattern that the expected values takes is very similar to what actually happened, with peaks and troughs copied almost perfectly. Although, the model to begin with over predictions, and then underestimates the data, so there are improvements. However, it does inspire confidence in the forecasted points for November to March, which suggests that at least 10 incidences will be reported each month, with peaks in December and March where the count would be closer to 25 for these two months. The upper limit fluctuates with the data, but is showing that the total count per month should not exceed 60, though it could reach higher in March in line with the expected peak in the same month.

This method can be expanded to include other wards, areas and streets, as well as done much more regularly so that the council have a rolling estimate of the upcoming months, given only minor changes to the code in terms of pre-processing. Additionally, seasonal effects can be looked into and incorporated in the SARIMA model, which may produce better results since there are monthly trends as explored later in Figure 9. Once refined, time series analysis could provide

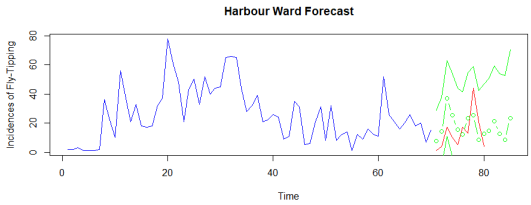

Fig. 6. Time series forecast for the Harbour ward. The blue line indicates actual observations used in the forecasting, the red is actual data but used as test points. Green circles are the expected count of incidences, along with green 95% confidence intervals.

another crucial weapon in the arsenal for the LCC to combat fly-tipping, by being able to predict when a certain location will experience peaks and troughs, thus allowing the council to divert resources in advance and begin to provide a proactive approach for tackling this issue.

C. Data Visualisation

From the fly-tipping data provided by Lancaster City Council from September 2015 to October 2021, this project builds three graphs to illustrate the incidents regarding waste types, ward codes and land types. Figure 7 shows the waste type of fly-tipping incidents per year. The most significant waste type in 7 years is household black bags (W15) shown in the red line, following other household wastes (W12) shown in the purple line. Other waste types present a relatively optimistic number of events, typically less than 250 incidents per year. Overall, the fly-tipping incidents have generally improved since 2017, dropping from 2400 in 2017 to 1000 in 2021. However, household black bin bag waste is still the main object that results in fly-tipping incidents.

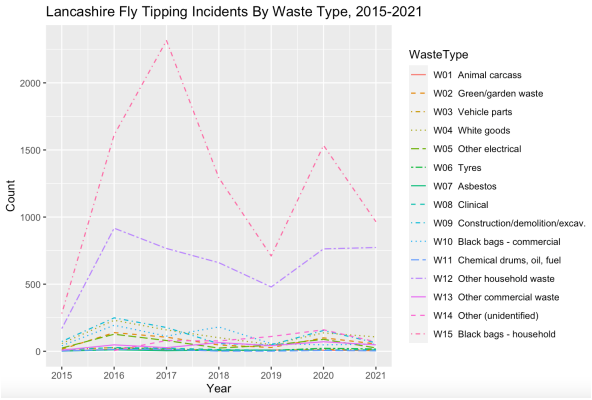

Fig. 7. The figure shows the waste type of fly-tipping per year from September 2015 to October 2021

To discover the fly-tipping hotspot locations and potential links between the type of tipping and a location, this project divides the incidents by waste types into their corresponding ward codes and builds a bar chart shown in Figure 8. The top 3 severe locations are respectively Heysham North (HEN), Poulton-Le-Fylde (POL) and Harbour (HAR). It is worth noting that Heysham North occurred approximately 3500 incidents from 2015 to 2021, which exceeds 900 incidents compared to the second-highest location Poulton-Le-Fylde.

Figure 9 plots the line graph of monthly fly-tipping incidents by land type. It is obvious to observe from the graph that the back alleyway (L03) is the most frequent land location that arises incidents, which happens 900 monthly incidents on average shown by the brown line. The number of incidents in the back alleyway rapidly increases from 800 in July to 1050 in October and dramatically decreases to 750 in February. Although other land locations show a gentle number of incidents, they obtain a similar monthly trend as the back alleyway.

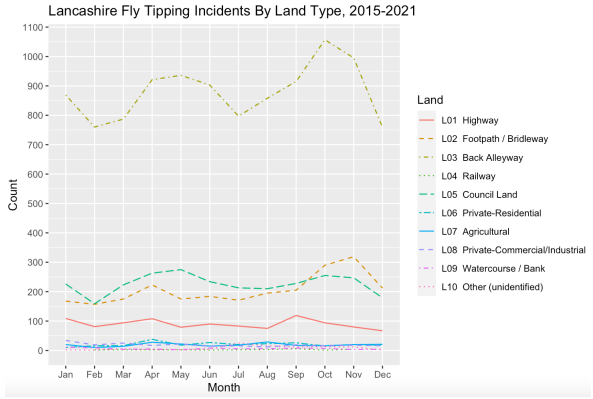

Fig. 9. The figure shows the monthly fly-tipping incidents from September 2015 to October 2021

By combining these three graphs, it can be observed that household wastes are the main waste types in most locations. Those household wastes usually occur in the back alleyway. Moreover, incidents in the back alleyway increase from July to October per year. The reason is probably because of the summer vacation. Hence, the City Council staff could put more emphasis on fly-tipping in October.

D. Top 4/ Bottom 4 Deep Dive

To help us further answer the first, second and fifth research questions, we decided to conduct a deep dive into the four 1st address lines with the most fly-tipping incidents recorded between the studied period against four of the 1st address lines with only 1 incident reported in this period. Before carrying out this analysis, we were careful in locating and dealing with any duplicate street names. Following on from this, the 'Top 4' were identified as *Clarendon Road East* (Morecambe), *Cavendish Road* (Heysham), *Sefton Road* (Heysham) and *Alexandra Road* (Morecambe). There were a number of street names with only 1 incident to their name, so we decided to chose one from each of Morecambe, Lancaster, Carnforth and Heysham. These were *Africa Drive* (Lancaster), *Anstable Road* (Morecambe), *Berwick Way* (Heysham) and *Back Market Street* (Carnforth).

The deep dive was executed using (10), which, after inputting a postcode or 1st address line, displays relevant corresponding information including bin round details, local recycling sites and the type of waste they accept, and street cleaning details. Interestingly, the nearest recycling sites for each of the 'Top 4' are identical - Morrisons Car Park and Sainsburys Morecambe. Neither of these sites currently accept plastic bags, which was found to be the most common type of fly-tip between 2015 and 2021. Hence, one inference that could be made from the deep dive is that the council needs to focus on either increasing the capacities of these critical recycling sites, or introducing more sites in close proximity to these target areas. In contrast, two of the 'Bottom 4' were located near actual recycling centres, further implying that the capacities of nearby recycling sites to the 'Top 4' is a factor in

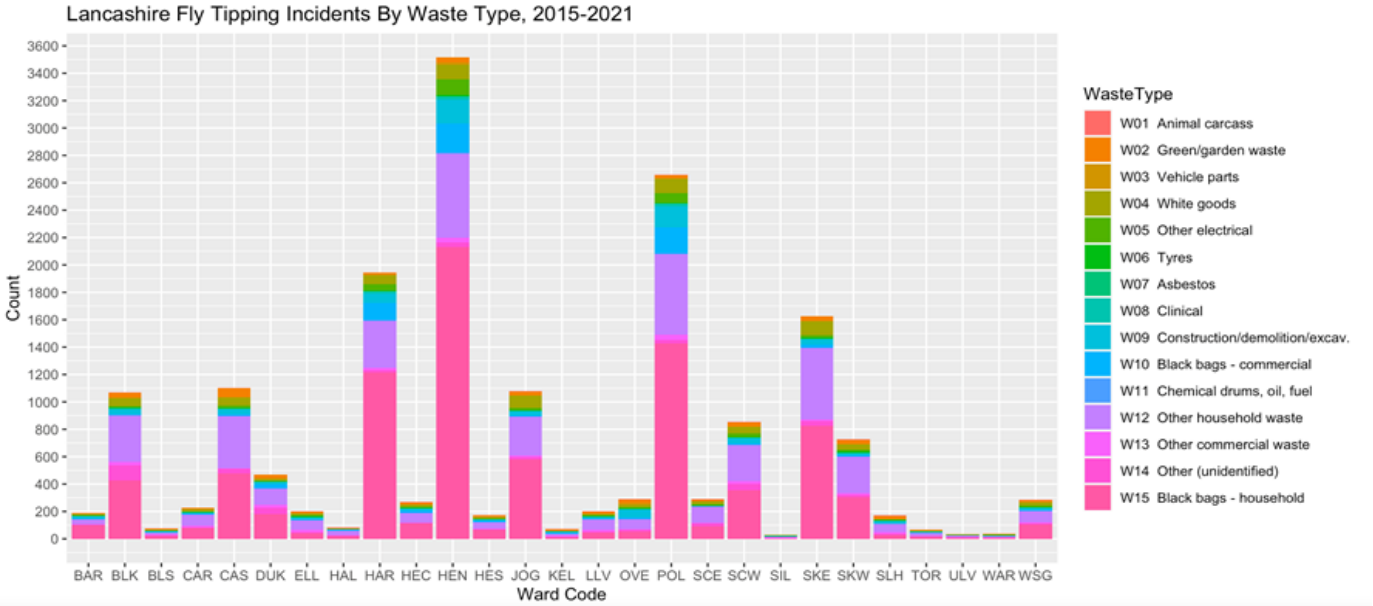

Fig. 8. The figure shows the waste type of fly-tipping by ward code

the high frequency of fly-tipping incidents reported amongst them.

Additionally, it can be observed that on average, the bin collection days for the 'Top 4' are earlier in the week than those for the 'Bottom 4'. This suggests that there may be some correlation between the amount of fly-tipping incidents and where the bin collection day falls within the week. It would be interesting to explore this relationship further by comparing a larger amount of locations in the local area.

Finally, it can be seen that for the 'Top 4', street cleaning is carried out daily and a sweeping vehicle is used either daily or weekly, depending on the location. In comparison, three of the 'Bottom 4' location have no street cleaning at all. This suggests that while the council are clearly aware of the locations of fly-tipping hotspots, street cleaning has not proved to be very effective in reducing the amount of incidents. Hence, our suggestion is that the council focus on other preventative measures instead.

E. Bias and Validity

As we were completing our analyses, there were a few areas in which we identified potential biases, which could have affected the validity of our work.

Firstly, the data in columns representing the attributes of fly-tipping incidents are subjective to whoever reported and detailed the incident. With the size of the tip, for example, it is hard to differentiate between *S02 Car-boot load or less* and *S03 Small van load*. Hence, two different council worker's opinions on the size of a tip may differ, affecting the observed results as a whole. This is an example of response bias, which may also be applicable in other columns of data.

Furthermore, the 'deep dive' into the 1st address lines with the most/least incidents to their name (see Section III-D) was conducted using only current information on these addresses. There is a high chance that some of this information may have changed over the 6 years that the data spans (e.g. recycling sites may have changed/been upgraded, bin rounds and street cleaning may have changed days etc.). Hence, there is a potential for information bias here.

Finally, the clustering was done using the K-Means algorithm, which is a distance-based clustering algorithm, but if the aim of clustering is taken into consideration (which is to find out areas where the cases of fly-tipping are high), then the more appropriate technique would have been to use density-based clustering, using algorithms like DBSCAN. However, the use of the DBSCAN algorithm gave very poor results as compared to the K-Means algorithm. The clusters formed using the K-Means algorithm are very well defined.

To assess the validity of the study, there are four pillars that will be discussed:

- Internal
- External
- Construct
- Conclusion

Beginning with internal validity, which represents the extent to which the observed results represent the truth (11), since the dataset we have received is directly from LCC, then we know that it is factually correct, although there is a degree of subjectiveness involved as mentioned in with the biases, however this is not an issue with the study and could not be further verified. Another possible threat to the internal validity of our results is the potential information bias in Section III-D, again mentioned in the discussion of biases. If this section were to be expanded on in future work, greater care would be needed to satisfy the demands of this validity. Aside from these two potential threats, we could not see any other reason

to question the internal validity of our results.

Externally, we are concerned with generalising the findings more globally, so for example, would it be the case that household waste is the most commonly dumped category in neighbouring regions? Our methods would be applicable in other scenarios and the process is replicable for other datasets. Conclusions made here would likely be similar in parts, and different in others, such as the question just posed, it is likely black bin bags (W15) are the most reoccurring waste in other city regions too, due to the high population density and residential buildings. Other findings such as the time series and land type plots would depend on the location, amongst other factors which we did not have information on such as income levels and number of Houses in Multiple Occupation (HMOs) in the area, so our results may show some trends seen nationally, however to confirm this would require data from other authorities too and was not in the scope of the project.

Construct validity refers to the level to which inferences can legitimately be made from the operationalisations in your study to the theoretical constructs (12). Our brief for the study was predominately based on no prior information, as such there was no agreed upon position that we could test. Instead, we took what existed and made sense of it through clustering and analysis because of the fact we are not judging conclusions against a theorised outcome.

Finishing off this section, conclusion validity is surrounded by the idea whether the claims made are reasonable. With the interactive map, we can see dense clusters centered in the heart of Morecambe and Lancaster, and a smaller group in Carnforth, which comes at little surprise given these places have large numbers of people and buildings closely packed together. Additionally, with the various plots and 'Top 4/ Bottom 4', we discuss potential reasons for peaks and differences between time, location and type, all of which are not unusual so we can confidently say the results are valid, with the conclusions made from them are reasonable.

IV. Conclusions

This research was based on the data provided by the Lancaster City Council and aimed to investigate the fly-tipping problems in order to locate and predict the hot spots in the Lancaster city area. Interactive maps, which include the fly-tipping cases' location and relevant information, and various clusters, are provided. These maps facilitates future prediction, sorting, tackling and prevention of fly-tipping problems for the council. In addition, time series analysis for Heysham North and Harbour cases gave reliable predictions for the near future. The Council can calculate expected case numbers using this model, which can be used to arrange local preventative measures. Furthermore, the correlations between waste type, waste size, land type, location, and time have been found and explored using data visualisation techniques. Finally, the learning from the 'Top 4' and 'Bottom 4' locations evaluates the performance of existing preventative measures and gives a greater insight into areas enforcement should pay particular attention. It is worthy to have a deeper analysis into different hotspots to suggest appropriate measures.

We feel that the approaches taken to the analyses were all appropriate, as we were able to answer either partially or in full each of the research questions given.

Future work can contribute an Application Programming Interface (API) to directly access to the interactive map. The API would allow users to access the data and communicate. Thereby, the Lancaster City Council staff could more conveniently supervise the fly-tipping incidents through computers, phones and tablets. In terms of prediction, we could apply more statistical approaches based on the characteristics of different areas. Bayesian inference is an efficient mathematical technique used to achieve the posterior probability based on prior knowledge. In the fly-tipping case, future incidents could be predicted by building the relevant distribution, using the given sample.

References

[1] GOV.UK, "Fly-tipping statistics for england, 2020 to 2021," Dec 2021. [Online]. Available: https://www.gov.uk/government/statistics/fly-tipping-in-england/fly-tipping-statistics-for-england-2020-to-2021

[2] G. Rachiotis, D. Papagiannis, D. Markas, E. Thanasias, G. Dounias, and C. Hadjichristodoulou, "Hepatitis b virus infection and waste collection: Prevalence, risk factors, and infection pathway," *American Journal of Industrial Medicine*, vol. 55, no. 7, p. 650–655, Apr 2012.

[3] "Hepatitis b," Jul 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b

[4] C. K. R. Charu C. Aggarwal, "Data clustering," *Chapman and Hall/CRC*, August 2013.

[5] A. Géron, "Hands-on machine learning with scikit-learn, keras, and tensorflow, 2nd edition," *O'Reilly Media, Inc.*, September 2019.

[6] A. Hayes, "Autoregressive integrated moving average (arima)." [Online]. Available: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

[7] C. S. Ding, "Multidimensional scaling," in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, ser. Oxford Library of Psychology. Oxford University Press, 2013.

[8] S. Palachy, "Stationarity in time series analysis." [Online]. Available: https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322

[9] E. Zvorniconin, "Choosing the best q and p from acf and pacf plots in arma-type modeling." [Online]. Available: https://www.baeldung.com/cs/acf-pacf-plots-arma-modeling

[10] LANCASTER.GOV.UK, "Lancaster my neighbourhood tool," May 2020. [Online].

Available: https://www.lancaster.gov.uk/information/my-neighbourhood
[11] C. M. Patino and J. C. Ferreira, “Internal and external validity: can you apply research study results to your patients?” [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6188693/
[12] P. W. M. Trochim. Research methods knowledge base.