Case study: Footfall in Leeds city centre

1 Introduction

Measuring the number of people that enter a specific place in a given time, along with other features such as people's psychology and the season of the year, allows us to improve and forecast the staff scheduling, sales volumes and also the potential for crowd trouble. Recently, the technology made the development predictive models possible which allows predicting customer behaviour. For example, if we ask ourselves, why are the expensive products at eye level while the economic ones are in less visible places or why do they offer so many loyalty schemes? The answers are found in the data that detect people's patterns, suggesting them to do so (Murali, Pugazhendhi, and Muralidharan 2016). Currently, a large number of companies are trying to collect and use as much data as possible to increase both, the user experience and their own profits.

Focusing on footfall, the ability to predict this has a large range of applications for both governments and businesses. However, obtaining a correct estimate is difficult as each situation requires a different mathematical model. In addition, the data used to make these models can have different origins such as CCTV cameras, cellular networks, maps services, etc. In this study, we will analyse the estimates of people obtained by Leeds City Council (LCC) through various CCTV cameras and create several models to predict the count.

The contributions of this study are:

- We analyse how different features affect the count of people.
- We propose additional elements which can affect and consequently, improve the accuracy of the estimate.
- We evaluate different models using the root mean squared error (RMSE).

The report is organised as follows: we will introduce how the footfall estimation has been used lastly and propose several events which will help to improve the models' accuracy in Section 2. In Section 3, we will explain how the data was collected and processed. We then analyse all available variables that we can use for modelling in Section 4. In Section 5, we create different models and all our findings are discussed in Section 6, and Section 7 concludes the report.

2 Background Research

The Leeds City Council placed the eight cameras at the locations shown in Figure 1. As we can see, the locations in which there are more intersections, there are usually two cameras deployed facing opposite directions to deal with the limitations on spatial coverage due to the limited viewing angles of ordinary camera devices (Ng, Pei, and Jin 2017). By contrast, the council only installed a single camera in straight streets.

The weather has been previously identified as a significant influence on the footfall and as one of the major factor affecting the sales of retailers (Martínez-de-Albéniz and Belkaid 2020), (Badorf and Hoberg 2020). Therefore, we will analyse if this relationship can be applied to our scenario and, in case where it is applicable we will utilise the weather to better improve our models.

On the other hand, we believe that other events can also affect the number of pedestrians such as the Leeds United games. Leeds United has had approximately an average attendance of 25,000 people throughout the period of study (Football Web Pages 2019). Thus, the total footfall across the city centre is likely to be greater on game days than for regular days.

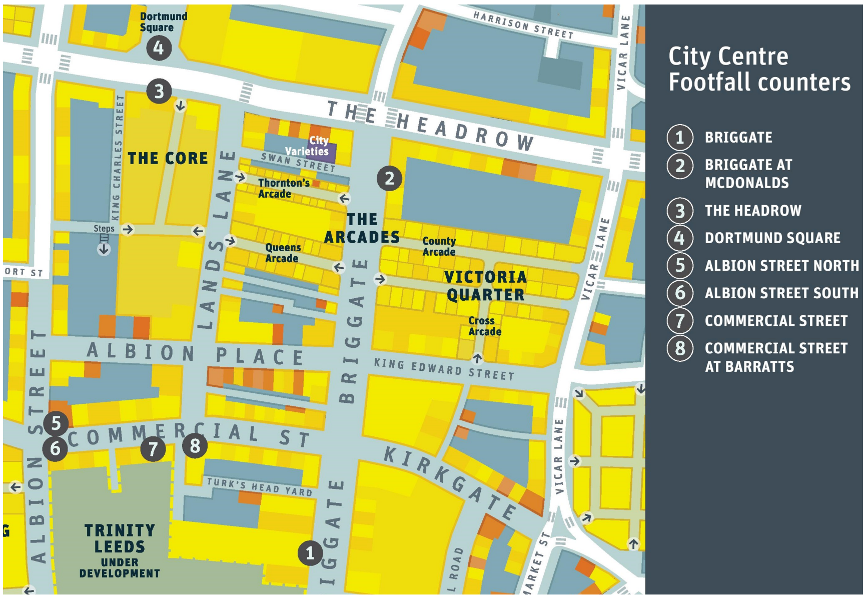


Figure 1: Camera locations.

3 Data Preparation

The cameras data set was acquired from Minerva (University of Leeds 2019) containing data from 2007 to 2014; however, the original data was from the LCE (Leeds City Council 2020) in which we could access to up to date data (April 2020). The data contained the same variables in both sources: the date and hour identifying the actual day of the record, the location of the camera, the estimated count of people at that location, the day of the week and three variables (week number, month and year) with the British Retail Consortium (BRC) calendar format which slightly differs from the regular calendar.

Before transforming the data to our desired format, we checked the authenticity of our data. To do so, we downloaded the necessary CSV files from the original source and ensured that we had the same data by comparing all records at each location and checking each column had the same value. Once the data was verified, we combined the date and hour into a single variable since this would simplify later linkage with other data sets and, as our study is not focused on retail, we also converted the variables week number, month and year into the regular calendar what will ease the analysis.

In addition, as commented in Section 2, we included in our study two other features: the weather and the Leeds United fixtures.

Firstly, the weather data was collected from World Weather Online (WWO) website (World Weather Online 2020). WWO is a popular website for weather forecasts that also stores historical hourly data, which can be accessed through their powerful REST API. Through this API and an authorization key, we were allowed to access weather conditions from 1st July 2008 up until the present time; therefore, we obtained the hourly weather in Leeds from this date to the latest date in our data, 30th January 2014. We only stored the most significant data from WWO: the date and time, the temperature in Celsius, the wind speed in kilometre per hour, the precipitation in millimetres, the humidity and the cloud cover in percentage.

Secondly, we scrapped the worldfootball website (HEIM:SPIEL 2020) in order to acquire the Leeds United games from 2007 to 2014. We created a data set containing the date, the time of the kick-off, the place (home, away or neutral), the opponent and the result.

Finally, we joined all three data sets using the date and time, this allowed us to perform a deeper analysis of what affects the original count of people in the Leeds city centre.

4 Analysis

In the section Background Research, we have seen that the Leeds Council placed eight cameras at different locations and estimated the amount of people per hour in each of them. Additionally, we proposed two additional sets which can be related to this estimation; therefore, in this section, we perform a deep analysis for all three data sets. Firstly, we evaluate the count estimation without any other additional variables and, secondly, we analyse whether the weather or the Leeds United game days have any correlation with the estimation.

As a first insight into the data, we create a summary of the most significant information for each location and also the correlation between each location. We use the Pearson's Correlation Coefficient which measures the strength of the linear relationship between two variables and is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{1}$$

When the value is zero, there is no linear relationship, while when it is close to one (in absolute value) there is a perfect correlation.

Location Name	First Date	Mean	Med.
Briggate	20/07/2007	1269.57	505
Briggate at McDonalds	27/06/2008	648.55	301
Headrow	27/08/2008	611.93	275
Dortmund Square	27/08/2008	1009.15	389
Albion St. North	16/08/2008	837.63	232
Albion St. South	16/08/2008	1207.42	342
Commercial St.	27/06/2008	1307.24	238
Commercial St. at Barratts	17/07/2009	891.58	259

Table 1: Summary of each camera location.

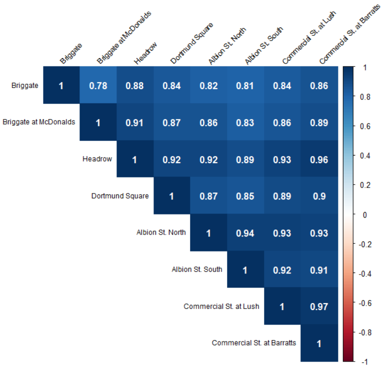


Figure 2: Correlation between each camera location.

As we can see from Table 1, the cameras were deployed (or the council started making estimations) on different dates. Most of the camera records started in 2008, with the exceptions of the Briggate and Commercial Street at Barratts (2007 and 2009, respectively). Furthermore, the Commercial Street has the greatest mean whilst the Briggate has the largest median; from this, we can infer that the amount of people at Commercial Street changes considerably throughout the day and that Briggate has a constant flow.

It can be seen from Figure 2 that all locations have a strong correlation because they are always over 0.8, except the cameras at Briggate which are separated by a wide distance. Moreover, the correlation between both cameras located at Commercial

Street is 0.97, which means that people usually walk between them without turning into another street.

A deeper understanding of the data can be achieved by focusing on the different dates; for this reason, the figures 3 and 4 show the average count over the months and the hours, respectively. Additionally, we filtered the records recorded before 2009 because not all locations data were not available.

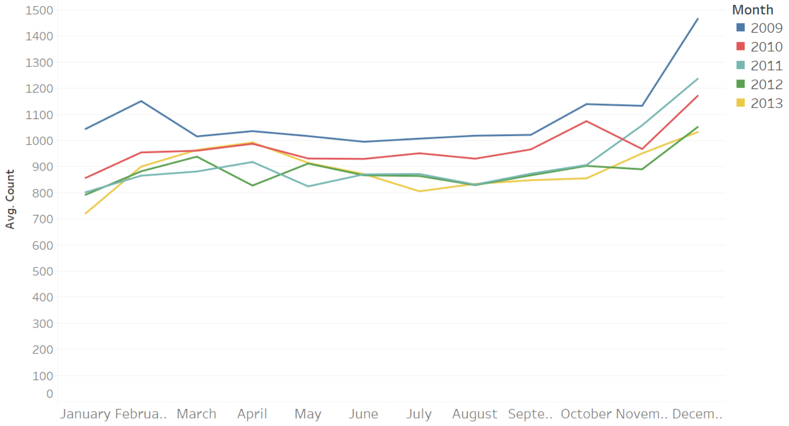


Figure 3: Average count throughout the study period.

Figure 3 shows that the same trend is followed throughout the years. During the first two-three months, there is a constant increase followed by a decrease for the next two months. From this point until September, the average fluctuates with the lowest value in July or August, which is probably caused by the summer holidays. Finally, the last three months of the year experience a gradual growth until reach the maximum in December. However, although they all follow a similar trend, 2009 was the year in which the centre of Leeds was busier.

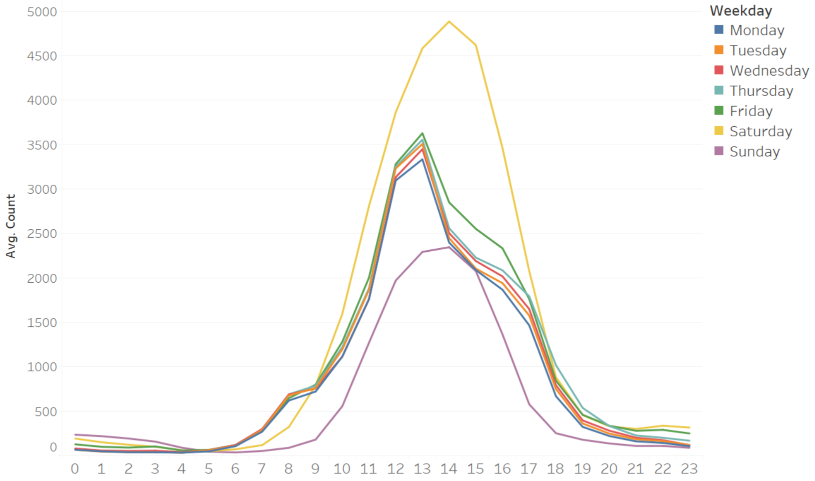
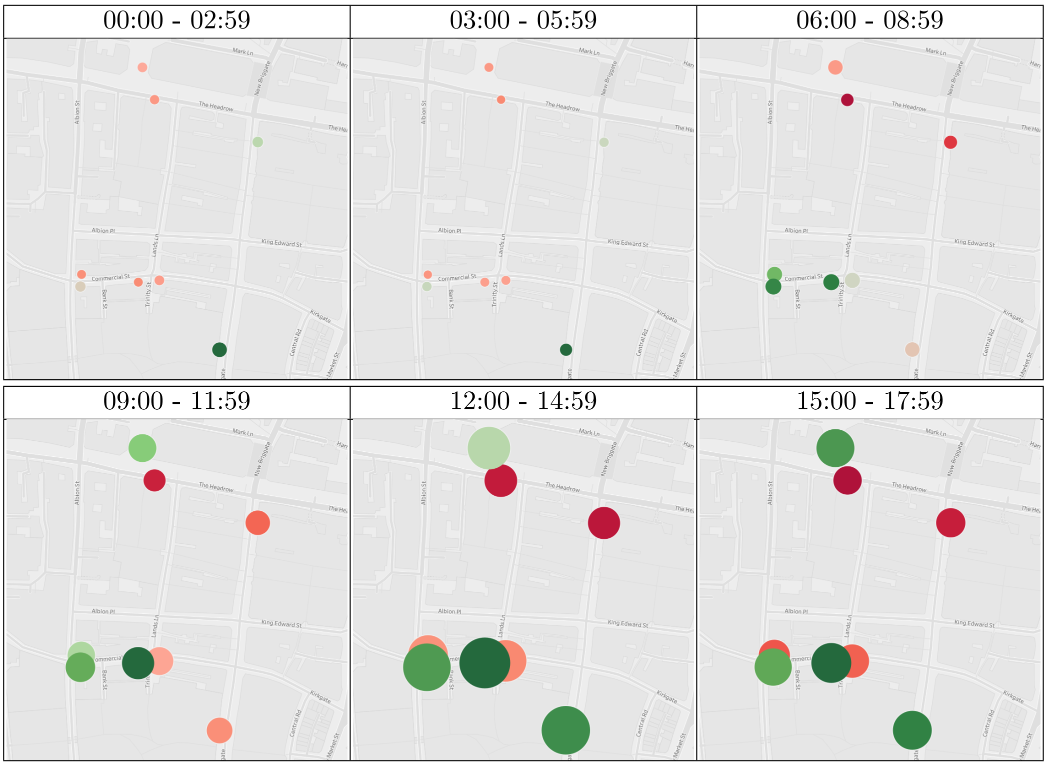


Figure 4: Hourly average count by location.

The hourly graph clearly shows a clear difference between weekdays and weekends. With regards to day time, all weekdays follow the same tendency with a maximum of approximately 3,500 people, but there is an exception on Friday evenings which are slightly greater than the other weekdays (around 500 people more); nonetheless, Saturdays are the most active days with a peak of almost 5,000 people and Sundays are dullest days.

On the other hand, focusing on night time, we can see that Friday and Saturday nights are the busiest because they have around 500 people during most of the night while the other days vary between 0 and 100 people. In addition, Saturday and Sunday mornings start later compared to the other days with less nightlife.

Finally, a more intuitive way to see how the situation is different at each location (Figure 1) is to overlay the average count with ranges of 3 hours a map. To do so, we altered the colour and size of the nodes, the size according to the average count and the colour according to Z-scores, which represent how far away the data point is from the mean by using standard deviation (Credera 2020). In other words, the colour will help us to interpret which locations are below or above the mean of that hour with red and green colours, respectively.



00:00 - 02:59

03:00 - 05:59

06:00 - 08:59

09:00 - 11:59

12:00 - 14:59

15:00 - 17:59

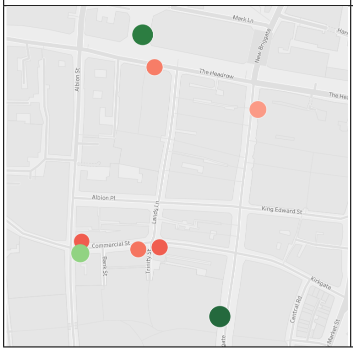
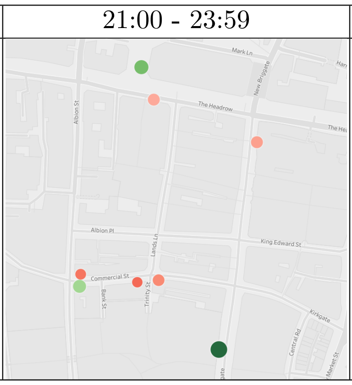
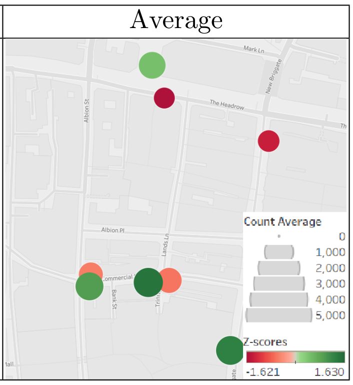
18:00 - 20:59	21:00 - 23:59	Average
 Mark Ln New Briggate The Headrow Albion St Albion Pl King Edward St Commercial St Lands Ln Bank St Trinity St Kirkgate Central Rd Market St	 Mark Ln New Briggate The Headrow Albion St Albion Pl King Edward St Commercial St Lands Ln Bank St Trinity St Kirkgate Central Rd Market St	 Mark Ln New Briggate The Headrow Albion St Albion Pl Commercial Lands Ln Bank St Count Average 0 1,000 2,000 3,000 4,000 5,000 Z-scores -1.621 1.630

Table 2: Hourly map by location.

As we can see from Table 2 that there is a clear transition from day into the night. We can see that both cameras located in Briggate are above average within the period from 00:00 to 5:59 due to the surrounding nightclubs. Nevertheless, over the early hours of daylight (from 06:00 to 11:59), both locations decrease their importance and from mid-day onwards only the camera at the bottom of Briggate recovers its significance, being above the mean.

By comparison, all other locations, except The Headrow, gain relevance between the period from 6:00 to 17:59 with two locations standing out over the others, Dortmund Square and Albion Street South, which will also keep a greater influence during the transition from day to night (from 18:00 to 23:59).

Furthermore, looking at the average map, we can see that the deductions from Table 1 were correct, the busiest location is the lower part of Briggate where there is a constant flow of people, followed by the stretch between Albion Street and Lands Lane of Commercial Street. Finally, Dortmund Square is slightly above the average.

Once we analysed how the date and the location affect the count, we evaluate how other events also influence the estimation. Firstly, the Figure 5 shows the average count over each weather feature and add a trend line to help the interpretation. Secondly, Figure 6 shows a bar plot to compare the daily average depending on whether there was a Leeds United game or not.

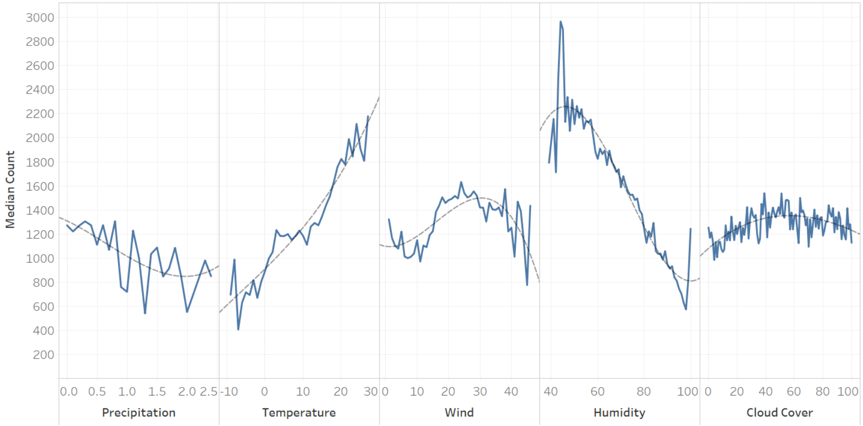


Figure 5: Average count over the weather features.

It can be seen from Figure 5 how the different features influence the average count. The two most significant relationships are for the variables temperature and humidity, both have an approximately linear relationship. The temperature has a positive relationship so as the temperature rises, the number of people also increases; conversely, the humidity has a negative linear relationship decreasing the count as the humidity declines. Moreover, the precipitation follows a less clear positive linear relationship with strong fluctuations from 1mm.

On the other hand, the wind and the cloud cover follow a higher-order relationship in which the average count has a peak for intermediate values. We have to emphasize that the cloud cover contains few variations compared to the wind, which has large fluctuations that can affect its predictive utility.

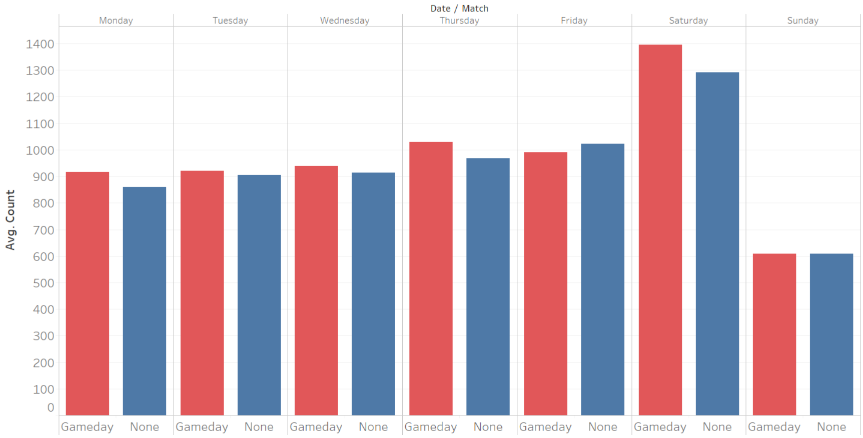


Figure 6: Daily average count splitting by game day.

As we can see from Figure 6, game days have a slightly greater average count than the regular days. This increase is almost imperceptible except on Saturdays when the difference is around 100 more people; however, on Sundays, there is no distinction. Therefore, we decided to rule out including this attribute in our models.

5 Modelling

5.1 Methodology

A. Regression

We decided to use two different types of regressions, polynomial and random forest, for obtaining a count estimation of people at each location in Leeds city centre. A polynomial regression establishes a relationship between a dependent variable (Y) and independent variables (X), which power is more than 1. This regression is represented by the equation:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_n x_m^p + \epsilon. \tag{2}$$

It is similar to multiple linear regression but it allows to achieve more accurate results. To model this type of regression, we use the function "*lm*" (Fitting Linear Models) from the library "stats".

On the other hand, Random Forest is an ensemble of decision trees exploiting collective wisdom to generate very accurate predictions. It uses bootstrapping to create different samples from the original data to create different decision tree models and finally compute the average over the predictions of all trees when making predictions. Figure 7 shows a diagram of how a random forest looks like. In this case, we use the library "ranger", which is faster and requires less computational resources than other libraries such as "RandomForest" or "Random Jungle"; however, it also reduces the accuracy of the model (Wright and Ziegler 2017).

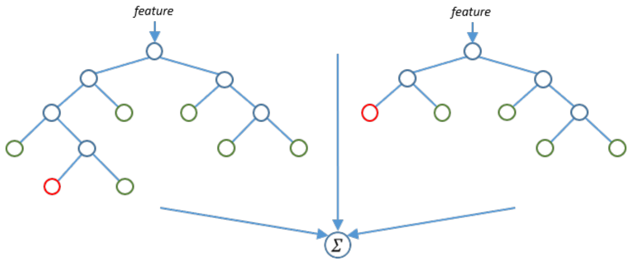


Figure 7: Random Forest diagram.

According to our findings in Section 4, we propose five different models to evaluate if the additional features improve the prediction or not. Two models are polynomial, the first one using only the original camera data, and the second one using the camera data as well as all five features from the weather data set. The other models are random forests, two of them have the same inputs as the polynomials but we also create another one including all-weather features except for the wind. We can find a summary of these models in Table 3.

Regression Type	**Name Abbreviation**	**Input Variables**
Polynomial	Without Weather	Location + Year + Week Num. + Week day + Hour
	All	Location + Year + Week Num. + Week day + Hour + Precip. + Temp. + Wind + Humidity + Cloud Cover
Random Forest	Without Weather	Location + Year + Week Num. + Week day + Hour
	All without Wind	Location + Year + Week Num. + Week day + Hour + Precip. + Temp. + Humidity + Cloud Cover
	All	Location + Year + Week Num. + Week day + Hour + Precip. + Temp. + Wind + Humidity + Cloud Cover

Table 3: Model types summary.

B. Error Metrics

To evaluate the accuracy of the proposed footfall estimation methods, we use two error metrics:

- The Root Mean Square Error (RMSE) is the standard deviation of the residuals and is defined by the equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Prediction_i - Actual_i)^2}{n}}. \tag{3}$$

The best value possible is 0; however, the RMSE from only one model does not provide significant information. We have to compare the RMSE obtained using the same data with different models and the one with the smallest value, it is the more ideal to fit that data (Barnston 1992).

- Correlation, as it was introduced in Equation 1, describes the strength of the linear relationship between predictions and their corresponding observations. Ideally, the values predicted by the model would obtain a perfect correlation (equal to 1) (ibid.).

5.2 Results

To evaluate our proposed methods, we split the data set randomly into train and test sets (80% and 20%, respectively). We ran this procedure 10 times and computed the average of each error metric.

Table 4 shows a summary of the predicted data and the average error metrics for each model. We can see that all five models obtained a close average to the original data; however, models using polynomial regression predicted negative values and their maximums were drastically lower than the original. By contrast, all three models using random forest obtained minimum values close to 0 and were able to predict some counts greater than the average.

The error metrics prove that models based on random forest outperform those based on polynomial because the RMSE is halved. In addition, it allows us to observe a significant variability in the estimations depending on the variables used. As expected, aggregating all the weather inputs improved the performance (from 382.26 to 350.30); however, we highlight the increase of the predictions performance when the wind variable is not used (from 350.30 to 332.61). This enhancement is mainly due to the wind fluctuation we observed in Section 4.

	Original	P Alone	P All	RF Alone	RF All	RF All without Wind
Average	949.57	949.57	949.58	949.56	951.19	950.90
Minimum	0	-824	-928	5	2	2
Maximum	19,820	4,553	4,631	12,149	14,658	14,639
Median	322	502	504	351	333	333
RMSE	-	748.54	736.31	382.26	350.30	332.61
Correlation	-	0.8389	0.8399	0.9609	0.9673	0.9703

Table 4: Summary of the models.

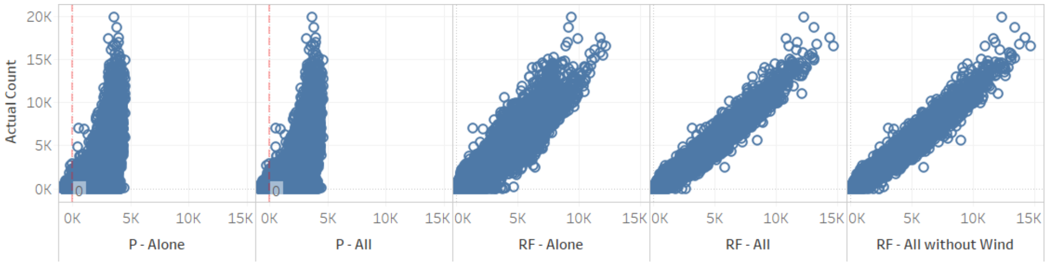


Figure 8: Correlation graph of each model against the actual count.

It can be seen from Figure 8 along with Table 4 how the accuracy was improved by being closer to the ideal correlation line (slope equal to 1). The most perceptible increases are the changes from polynomial to random forest and the aggregation of weather; nevertheless, the improvement previously observed whether using the wind or not is not very observable since the correlation slightly increased from 0.9673 to 0.9703 (0.0030).

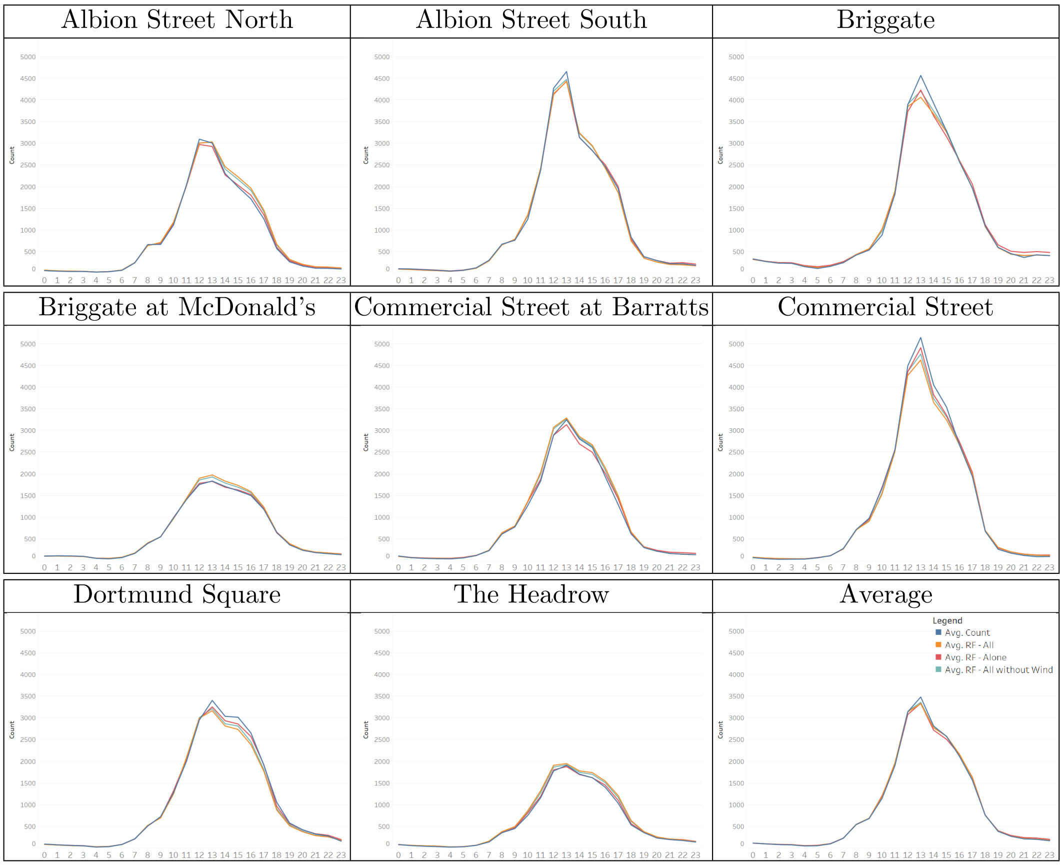


Table 5: Actual and predicted count by the hour for each model and location.

The figures in Table 5 show the hourly average predictions for the best three models and the original average count for each location. Overall, all three models average prediction are close to the original one throughout the day and particularly better during the night; however, they underestimate the number of people during the peak hours (from 13h to 16h) at all locations except at Briggate at McDonald's and The Headrow where their peaks do not reach 2,000 people.

On the other hand, the gap between the model with all-weather variables and the one without the wind is minimal, but in most charts, the model without the wind is slightly closer to the original.

6 Discussion

The study performed a deep analysis of the hourly number of people in Leeds city centre from 2007 to 2014, and researched about what other features could influence this count. Therefore, the findings can be used to estimate the number of people on a future date given the weather conditions; however, the model is not completely accurate and it is only useful for approximation. To improve the accuracy, we should investigate more events and if we find a relationship between an event and the count, add it to the model. Additionally, for a proper application of this analysis it would be best to use the most recent data sets available.

On the other hand, to make the study more alluring and profitable to companies, extra work could be done to improve the model. For example, I would propose to investigate the people's purpose while about the city centre via a short survey. This survey would ask non-personal information such as if they are going to work, walking around or shopping, and if so, where they were shopping. We may also ask other questions such as what kind of retailers they may be interested in. This could be used to create a model for shopping behaviour and estimate a customer conversion rate from the observed pedestrian count according to the area and shop type (Graham, Khan, and Ilyas 2019). Therefore, entrepreneurs could be better informed to decide their business strategy and whether their investment is profitable or not.

7 Conclusion

The study focuses on the analysis of the different features which can affect the number of people in Leeds city centre and the creation of several models to estimate this count given a date and some optional variables.

The analysis shows a clear transition from day into the night as during day time people are spread around three locations (Commercial Street, Briggate and Dortmund Square) and as it gets dark, people are gathered only at Briggate. Although this changeover, there is a strong correlation among the different locations, being particularly high between Commercial Street and The Headrow.

Overall, February, October and December are the months with higher average over the studied period. The first three Saturdays of December are always busier than in any other month making December the most crowded month of the year. Focusing on the time, we observed a similar trend on weekdays when during the peak hour with around 3,500 people; by contrast, Saturdays are the most active days with a peak of almost 5,000 people and Sundays are dullest days with less than 2,500 people.

According to the events, we found out a relationship between different weather conditions but we discarded the other hypothesis as we did not find enough a clear influence of the Leeds United games on the number of people in the city centre. With this findings, we created several models to predict the count of people and obtained that the model which fits better in this scenario would be a random forest using the date, weather (without the wind speed) and location. Overall, this model would provide a good approximate estimation for any time; however, these predictions are slightly lower than the actual counts from 13h to 16h (difference of 50-200 people, as average).

References

Badorf, Florian and Kai Hoberg (2020). “The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores”. In: *Journal of Retailing and Consumer Services* 52, p. 101921. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2019.101921. URL: http://www.sciencedirect.com/science/article/pii/S0969698919303236.

Barnston, Anthony G. (1992). “Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score”. In: *Climate Analysis Center, NMC/NWS/NOAA, Washington, D.C.* DOI: https://doi.org/10.1175/1520-0434(1992)007<0699:CATCRA>2.0.CO;2. URL: https://journals.ametsoc.org/doi/abs/10.1175/1520-0434%281992%29007%3C0699%3ACATCRA%3E2.0.CO%3B2.

Credera (2020). *Tableau Workaround Part 5: Improve Map Coloring in Tableau.* URL: https://www.credera.com/ (visited on 05/16/2020).

Football Web Pages (2019). *Leeds United – Home Attendances.* URL: https://www.footballwebpages.co.uk/ (visited on 05/17/2020).

Graham, Charles, Kamran Khan, and Muhammad Ilyas (2019). “Estimating the value of passing trade from pedestrian density”. In: *Journal of Retailing and Consumer Services* 46, pp. 103–111. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2017.10.005. URL: http://www.sciencedirect.com/science/article/pii/S0969698917304769.

HEIM:SPIEL (2020). *Leeds United Fixtures Results.* URL: https://www.worldfootball.net/ (visited on 05/17/2020).

Leeds City Council (2020). *Leeds City Centre Footfall Data.* URL: https://datamillnorth.org/ (visited on 05/17/2020).

Martínez-de-Albéniz, Victor and Abdel Belkaid (2020). “Here comes the sun: Fashion goods retailing under weather fluctuations”. In: *European Journal of Operational Research.* ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2020.01.064. URL: http://www.sciencedirect.com/science/article/pii/S0377221720301028.

Murali, S., S. Pugazhendhi, and C. Muralidharan (2016). “Modelling and Investigating the relationship of after sales service quality with customer satisfaction, retention and loyalty – A case study of home appliances business”. In: *Journal of Retailing and Consumer Services* 30, pp. 67–83. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2016.01.001. URL: http://www.sciencedirect.com/science/article/pii/S0969698916300042.

Ng, Y., Y. Pei, and Y. Jin (2017). “Footfall Count Estimation Techniques Using Mobile Data”. In: pp. 307–314.

University of Leeds (2019). *Minerva.* URL: https://minerva.leeds.ac.uk/ (visited on 05/17/2020).

World Weather Online (2020). *World Weather Online — World Weather — Weather Forecast*. URL: `https://www.worldweatheronline.com/` (visited on 05/17/2020).

Wright, M. N. and A Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++and R". In: *Journal of Statistical Software*. ISSN: 0094-1190. DOI: `https://doi.org/10.18637/jss.v077.i01`. URL: `https://arxiv.org/abs/1508.04409`.