# Forecasting tourist counts with historical counts and external features

CIE5050-09 Additional Graduation Work, Research Project

Wang Xinyi

5130974

Front page picture:

https://www.tripsavvy.com/thmb/IxenggMQ6RCJcmFhpoqPpYejxC4=/3865x2174/smart/filters:no_upscale()/netherlands--amsterdam--de-wallen--oudezijds-achterburgwal-at-twilight-707442219-5972ba8ab501e800115572ed1.jpg

# Contents

# 1 Introduction

## 1.1 Research motivation

Amsterdam has always been one of the most popular European cities among tourists from all over the world. With a large number of visitors traveling the city, especially on weekends and holidays, overcrowding has become inevitable at major tourist attractions, which threatens the safety, comfort, and travel efficiency of visitors. To prevent overcrowding and provide effective crowd management, crowd monitoring and forecasting crowdedness have become increasingly important. An accurate forecast of crowdedness tells when and where a high level of crowdedness is expected and indicates corresponding crowd management approaches.

However, the pedestrian demand is influenced by multiple factors and fluctuates in the short term and long term, which makes the forecast of crowdedness a complex task. For instance, the demand patterns at residential neighborhoods and tourist attractions, during the day and the evening, in the week and on weekends, can be completely different, as people's travel behavior changes at different locations and periods of time. Different from the field of traffic state estimation, where a huge amount of studies have been done, crowdedness forecasting remains to be developed. Contrary to vehicular traffic, which always travels uni-directionally, pedestrian flows are characterized by additional degrees of freedom, context, and functional description of space (Papadimitriou, Yannis, & Golias, 2009). For instance, pedestrians in a corridor and at traffic lights have different behavioral rules. As a result, the crowdedness pattern is location-dependent and it is impossible to develop forecast models by referring to related studies.

Therefore, to provide forecast to pedestrian counts at one of the major tourist attractions in Amsterdam, short-term prediction models will be built in this research.

## 1.2 Research questions

The objective of this research is to develop a short-term prediction model within the limited time period that provides accurate predictions of the tourist counts at 14 camera locations around the red light district in Amsterdam. To achieve the objective, firstly, the model structures applied in related studies will be reviewed and the model structures to be applied in this research will be determined. Secondly, the influence factors of pedestrian behavior will be reviewed and the relations between factors and the counts in the data will be analyzed. Thirdly, forecast models featuring tourist counts will be developed based on the historical counts and external feature data and compared. To provide a reference to applying crowd management in time, the prediction horizon is set as 15 minutes to 30 minutes.

The main research question is formulated as:

Do ARIMA models provide a more accurate prediction of tourist counts for a 30-minute prediction horizon compared to multiple linear regression models in the area around Amsterdam red light district?

The following subquestions are derived to answer the main research question:

1. What factors influence tourist counts and to what extent?

This question helps select related parameters and determine the importance of parameters.

2. How to train and validate ARIMA and multiple linear regression models?

This question specifics the methods for training and validation of ARIMA and multiple linear regression models.

3. How to evaluate and compare the performances of different models?

This question provides the evaluation approach to standardize the performances of models.

4. What are the performances of the developed models?

By comparing the performances of models, the main research question can be answered.

## 1.3 Outline

This report continues as follows. In chapter 2, an overview of pedestrian-related prediction approaches applied in existing studies will be presented, as well as the review of influence factors on pedestrian behavior. In chapter 3, the data used for model development will be described and fused. In chapter 4, the relations between influence factors and the counts will be analyzed, as well as the autocorrelation in the count number itself. In chapter 5, the methodology of model training, validation, and evaluation will be described. In chapter 6, the performance results of multiple linear regression and ARIMA models will be presented and compared. In chapter 7, the results will be discussed and analyzed. In chapter 8, the conclusion of this research is presented and some limitations and recommendations are provided.

# 2 Literature review

In this chapter, the pedestrian-related prediction approaches are reviewed in sub-chapter 2.1 and the influence factors of pedestrian behavior used in previous studies are reviewed in sub-chapter 2.2.

## 2.1 Prediction approaches

Prediction aims to estimate a future state based on the current and past data samples. A prediction function f(x) is defined based on the input historical data and its output is the predicted value for a certain prediction horizon (Sapankevych and Sankar, 2009).

Predictions can be classified as linear or non-linear, depending on the exact mathematical relationship between the parameters and the output. Both types of models aim to find the optimal parameter set and defining the criteria for finding the optimal set of weights. The methods to assign or tune the parameters can be classified as supervised, unsupervised, and reinforcement learning (calibration) methods (DEL, TSS, USF, & AIZ, 2016).

Supervised learning maps an input to an output based on given input-output pairs (Russell and Norvig, 2002). The input is the values of parameters and the output can be understood as a 'label' that is assigned to the input by humans. During training, the output is compared to the desired output value and accordingly used to update the weights or parameters to reduce the global error (DEL et al., 2016). In unsupervised learning, the human 'label' is left out, which means that the training only looks into the patterns of the input data and requires no human supervision (Hinton, Sejnowski, & Poggio, 1999). All the criteria for updating the weights being determined internally within the training method (DEL et al., 2016). It is mainly used for classification problems and is not considered for the forecast models in this research. Reinforcement learning does not need the 'label' as well. It tunes the parameters by assigning a positive score if the prediction from the iteration is accurate and a negative score if wrong. The predictor learns after each iteration to perform better (DEL et al., 2016).

A large variety of data-driven approaches have been applied in the field of traffic state prediction. However, not all of them are applicable to pedestrian-related predictions. In this subchapter, the prediction approaches that have been used in pedestrian-related studies will be reviewed, as well as the way that the parameters are tuned.

### 2.1.1 Multiple linear regression (MLR)

Linear regression is an approach that models the linear relation between the prediction variable x (independent variable) and the outcome variable y (dependent variable). When multiple independent variables are considered in the model, it is called multiple linear regression (Freedman, 2009). Following is the mathematical formulation of MLR. $\beta$ denote the parameter and $\varepsilon$ is the error term that adds 'noise' to the linear relation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Multiple linear regression models have been widely used due to their simple structures and high computational efficiency. The main restriction of this method is the forced linear relation, as it only generates good results when a linear relation is expected. When the relation between the prediction variable and the outcome variable cannot be represented linearly, the prediction performs poorly. MLR models have been used as a baseline model for various prediction studies, such as for bus ridership prediction (Roosmalen, 2019) and travel time prediction (Nikovski, Nishiuma, Goto, & Kumazawa, 2005), and showed better performance than even more complicated models such as decision trees.

### 2.1.2 ARIMA models

The AutoRegressive Integrated Moving Average (ARIMA) model is a popular type of linear regression model which have been applied extensively for traffic prediction. It is adapted from the ARMA model which contains two parts, the AutoRegressive (AR) and Moving Average (MA) model. The former relates the value of a variable in one period to its values in previous periods. The latter relates the variable to the residuals from previous periods. In the ARMA model, the outcome variable $y_t$ is predicted through a weighted linear function of the past observations yt-n of that variable (AR) and past error terms et (MA) (Shekhar and Williams, 2007). The two parts, AR(p) and MA(q), are combined as follows. $\epsilon$ is the error term. $\phi$ and $\theta$ denote AR and MA parameters.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}$$

By using the backshift operator L, the ARMA model can be rewritten as follows.

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \epsilon_t$$

$$(L y_t = y_{t-1}) \quad (L^S y_t = y_{t-S})$$

The successful application of the ARMA method requires a process to be stationary, which is not the case for the volatile traffic states. Therefore, for non-stationary traffic prediction, an extra term for structural trends in the data is incorporated in the model, which leads to the ARIMA (p, d, q) model with p autoregressive lags, q moving average lags, and the difference in the order of d (Shekhar and Williams, 2007).

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(1 - L)^d y_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \epsilon_t$$

ARIMA models have been developed for short-term pedestrian counts prediction in the city of Melbourne and outperformed MLR and SVR models on modeling weekday and weekend patterns (Wang, Liono, McIntosh, & Salim, 2017). Though few other studies implemented ARIMA models for pedestrian-related prediction, time series prediction studies based on historical data have much in common. ARIMA models also showed reliable predictions for short-term traffic predictions, such as cycle-length prediction in actuated traffic-signal control (Moghimi, Safikhani, Kamga, & Hao, 2018). More advanced models have been adapted from ARIMA to incorporate more data trends and achieve better prediction such as seasonal ARIMA models (Smith, Williams, & Oswald, 2002) and subset ARIMA models (Lee and Fambro, 1999).

### 2.1.3 Support vector regression (SVR)

Support Vector Machine (SVM) maps the non-linear training data from the input space into a higher dimensional feature space via a kernel function and then construct a separating hyperplane with maximum margin in the feature space. Support vector regression (SVR) is the application of SVM to time-series forecasting (DEL et al., 2016). The choice of the kernel function is the key to SVR application, as it provides the capability of mapping non-linear data into essentially linear feature spaces, where the optimization process can be duplicated as in the linear case. Linear, polynomial, and radial basis functions, including Gaussian, have been commonly used as kernel functions for SVR (Roosmalen, 2019).

SVR has been used for various studies, including predicting bus ridership (Roosmalen, 2019) and pedestrian counts in the city (Girzadas, 2020), where it showed better performance over multi-layer perceptron models. Adapted versions have been developed and showed improved prediction accuracy, for instance, the least squares support vector machine (LSSVM), which uses linear least-squares criterion for the loss function instead of inequality constraints (Wang and Hu, 2005).

The advantage of SVR lies in effective solutions to problems of nonlinearity, small samples, local minima, and over-fitting. However, the limitation is that the performance of the method depends greatly on the choice of the kernel function rather than the algorithm itself (DEL et al., 2016).

### 2.1.4 Neural networks

Neural networks (NN) represent the type of algorithm that mimics the nerve cells of the brain: the output is calculated by propagating the input signal(s) through a network of neurons. A neuron is linked to all neurons in the next layer, with links of different weights. The neurons of the next layer calculate their values by entering the summed weighted inputs in an activation function. The process is illustrated in Figure 2.1. In the case of regression analysis, the last layer only has one neuron: the dependent variable. The number of (hidden) layers, the number of neurons in a cell, the learning parameter, and the activation function are all subject to hyperparameter tuning. The weights can be derived using backpropagation and stochastic gradient descent (Roosmalen, 2019).

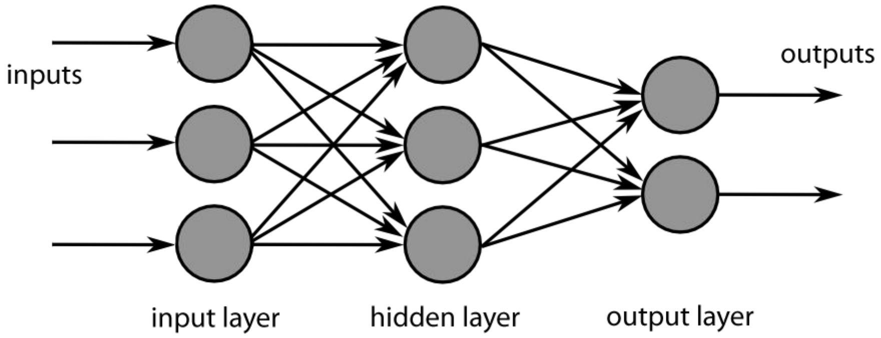


*Figure 2.1 Neural networks (DEL et al., 2016)*

Different types of neural networks have been developed and applied, which mostly differ in structure. The most used ones in transport are the basic artificial neural network (ANN) and recurrent neural networks (RNN) with supervised learning. From existing studies, ANN, for instance, the multi-layer perceptron model (MLP), has been applied for traffic volume prediction from hourly traffic data, bus ridership prediction and showed promising results (Siddiquee and Hoque, 2017; Roosmalen, 2019). However, for pedestrian counts prediction, its performance was not as good as expected with an incomplete input data set (Girzadas, 2020).

NN provides a method of deep learning which represents traffic features without prior knowledge, leading to good performance for traffic prediction. The downside of NN is its limited inherent explanatory power, in other words, the inherent inability to produce a unique solution to a problem (DEL et al., 2016).

### 2.1.5 Decision trees

Decision trees fit the model using recursive partitioning, whereby the data are successively split along coordinate axes of the predictor variables, which are seen as the branches of the tree. The split is done so that, at any node, the response variable is maximally distinguished in the left and right branches. The splitting continues until data are too sparse for each node. Then the tree is "pruned" using cross-validation. Terminal nodes are called "leaves," whereas the initial node is called the "root" (DEL et al., 2016).

Among various tree methods, mainly classification and regression trees (CART) and random forests (RF) have been used for traffic prediction (Xu, Kong, & Liu, 2013; Roosmalen, 2019). CART is used for analyzing classification issues for either categorical or continuous dependent variables. When the dependent variable is categorical, CART produces a classification tree and continuous variables produce regression trees (DEL et al., 2016). RF is an extension of a decision tree, where multiple trees are fitted. The final prediction is made by taking the weighted average of the prediction of each tree (Roosmalen, 2019). Nikovski et al. (2005) applied a simple regression tree for traffic time prediction and found that the model accuracy for the short-term horizon was poorer compared to MLR.

Tree methods are relatively simple but the limitations are the risk of overfitting and poor performance on incomplete data (DEL et al., 2016). The risk of overfitting can be limited by setting a minimum number of samples in each end node of the tree and restricting the number of levels (depth) of the tree (Roosmalen, 2019).

### 2.1.6 Comparison

The above methods have been widely used in traffic prediction models and in some studies, applied at the same time and compared through performance metrics of the root mean square error (RMSE: the Euclidian distance) and $R^2$ (the adjusted coefficient of determination). The RMSE matrix is used to determine to what extent the trends in the time series are similar. A smaller Euclidian distance identifies that two time series are more similar. $R^2$ measures the proportion of variance that is explained by the model can be used to compare the results with other researches.

Some studies showed RF outperformed other methods (Roosmalen, 2019), others found the best performance from RNN (Duives, Wang, & Kim, 2019). However, as mentioned above, these approaches only provide site-based predictions based on data with certain features and cannot be generalized to predict in other situations. Therefore, the results from previous studies should not play an important role in the model choice.

Each method has its limitations. For most approaches, the limitations can be overcome by adapting or cooperating with other methods. For instance, traffic volatility can be explained when ARIMA is cooperated with GARCH. However, it also requires extra efforts to be taken for model development. Considering the limited time-frame of this study, the multiple linear regression and ARIMA models are preferred due to their simple structures, promising results, and high computational efficiency, explicit modeling. The type and completeness of the input data also influence the choice of the methods, as some of the other models perform much worse when the data set is incomplete. There are existing toolboxes of the above-mentioned models in Matlab that shorten the time for model development, which is important to the limited time frame of this research.

## 2.2 Influence factors

Pedestrian trip decisions are influenced by multiple factors. They can be categorized into time and date factors (including holidays), weather factors, spatial features, and social factors.

It is proven by various studies that pedestrian traffic patterns change according to seasons, days of a week, time of day, festivals, and holidays (Wang et al., 2017; Ohler, Krempels, & Möbus, 2017). Wang et al. (2017) observed distinctively different patterns for weekdays and weekends in pedestrian counts in the city and explained that people become more active and likely to have more outdoor activities during the evening at the weekend. The above-mentioned elements have been included in prediction models in related studies (Roosmalen, 2019). In some studies, some of these elements are considered but not eventually included in the model because of the incomplete data set, in which case will lead to poor model performance (Girzadas, 2020). These time-related factors are usually included in prediction models as dummy variables (Roosmalen, 2019), which take only the value 0 or 1 to indicate the absence or presence of the factor.

Active mode trips are also influenced by weather as pedestrians have no shelter from extreme temperature, wind, or rain compared to other mode users such as car and PT users. Previous studies showed that people's mode choice for walking is negatively influenced by cold temperature below 5 degrees; wind speed and precipitation also negatively influence pedestrian trips, however, the extent of which is less than the influence of cyclists (Saneinejad, Roorda, & Kennedy, 2012). In previous pedestrian prediction studies, mostly only rain data has been considered and other elements are assumed to be not highly related to the trip choice or are omitted for computational efficiency (Roosmalen, 2019). Other studies that do not include weather data, convert the single value prediction into a prediction range to tackle the fluctuation of output caused by multiple exogenous factors (Girzadas, 2020).

Pedestrian behavior in an area is influenced by spatial features, for instance, physical boundaries, infrastructure completeness, functionality, and connection to other destinations st the data collection locations (Hess, Vernez Moudon, Catherine Snyder, & Stanilov, 1999). For instance, the physical boundaries limit the maximum number of pedestrians and influence the way people move in the area (Hoogendoorn and Daamen, 2005). The infrastructure completeness influences the comfort of tourists and therefore influences people's travel behavior. Most importantly, the demand for the area is determined by the functionality and connection to other destinations. Tourist attractions expect more visitors on weekends and holidays compared to a residential neighborhood. Roads connecting tourist attractions and public transport stations are more likely to be crowded with people compared to the roads without connection function. However, these important features are difficult to be quantified, which makes previous pedestrian-related studies constrained within the studied areas (Wang et al., 2017).

Social factors, such as demography and economy, also have an influence on pedestrian traffic. Age, gender, income, etc. Influence an individual's trip preference (Saneinejad, 2012). In previous pedestrian traffic studies, social factors have been taken into account by building location-based models (Roosmalen, 2019).

# 3 Data description and processing

In this chapter, the data used for this research will be described, as well as data processing and fusion.

## 3.1 Data description

Chapter 3.1 will present a description of the data featured in this research. This data comprises counting data (chapter 3.1.1.), weather-related data (section 3.1.2), and time-related data (section 3.1.3).

### 3.1.1 Counts data

The counts data is provided by the department Kennis & Kaders of the municipality of Amsterdam, collected through the Crowd Monitoring System Amsterdam (CMSA).

The data is collected around the red light district in Amsterdam, from 7 locations with the corresponding camera numbers, as illustrated in Figure 3.1. The counts in two directions are collected - the same as and the reverse to the arrows shown in Figure 3.1, resulting in 14 locations and directions in total. The data includes 58752*14 records of counts of every 5 minutes for nearly 7 months, from 01/07/2018 00:05:00 to 21/01/2019 00:00:00.

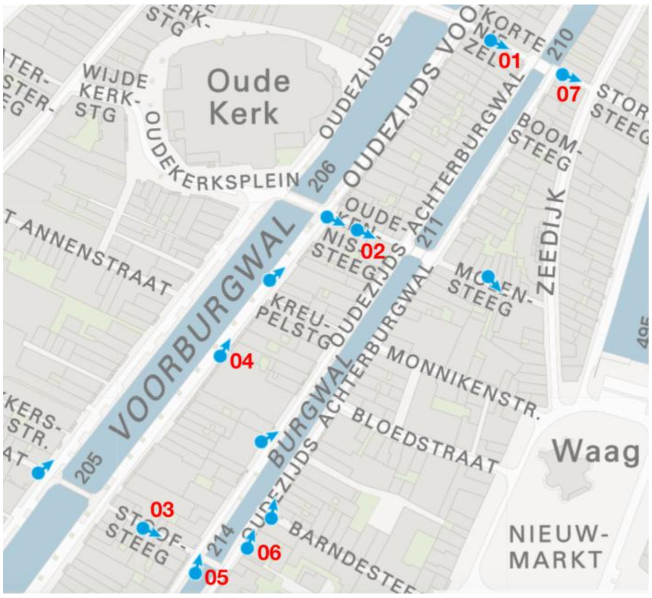


*Figure 3.1 Camera locations located at the Red Light District in Amsterdam*

Among 7 locations, camera 2, 5, 6 are located along the busiest touristy street Oudezijds Achterburgwal; camera 1, 7 are located on Korte Niezel and Korte Stormsteeg street which connect the red light district and China town to Amsterdam Central Station; camera 2, 3 are located on narrow allies of Oudekennissteeg and Stoofsteeg which connect the red light district to other tourist attractions, such as de Oude Kerk and de Bijenkorf; camera 4 is located on a less busy street Oudezijds Voorburgwal, where hotels, galleries, clinics, and local companies are situated.

From the counts data with corresponding time, the features of hour of day and day of week are also derived.

The data trends will be analyzed in chapter 4.

### 3.1.2 Weather-related data

The weather data is obtained from the publicly accessible website of the Royal Netherlands Meteorological Institute (KNMI) (http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi), which provides the hourly recorded weather data in different areas across the Netherlands. For this research, the data set from station Schiphol is used, as it is the nearest weather station to the studied area. It is important to note that the weather station is around 12km from the studied area, measured from Google Maps (https://www.google.com/maps). The weather conditions in the red light district might be different from station Schiphol, especially with showers during summer, which usually only cover a small area and move with the clouds. Therefore, it might influence the accuracy of the prediction.

The original data includes 22 weather elements, among which, the following 10 elements that are assumed to be relevant to pedestrian behavior are listed in Table 3.1.

*Table 3.1 Weather factors*

| Nr. | Elements | Category | Assumption |
|---|---|---|---|
| 1 | Dummy rain | Rain | Tourists are reluctant to travel in the rain. |
| 2 | Duration of rain per hour (0.1 hour) | | |
| 3 | Hourly sum of precipitation (0.1 mm) | | |
| 4 | Temperature (degrees Celsius) | Temperature | Tourists are more willing to travel when the temperature is in a comfortable range, not too high or too low. |
| 5 | Duration of sunshine in an hour (hour) | Sunshine | Tourists are more willing to travel in the sun. |
| 6 | Dummy snow | Snow | In severe weather conditions, tourists are reluctant to travel. |
| 7 | Dummy thunderstorm | Thunderstorm | |
| 8 | Dummy ice formation | Ice formation | |
| 9 | Average wind speed in an hour (m/s) | Wind speed | When the wind is very strong, tourists are less willing to travel. |
| 10 | Maximum wind speed in an hour (m/s) | | |

As the data collection period lasts from July to January, all the factors listed in Table 3.1 have occurred during the period. To provide a forecast as accurate as possible, these elements will firstly be fused with the counts data and their correlations and data trends will be analyzed before model development in the next chapter.

### 3.1.3 Time-related data

As people make trip plans according to the types of day and time periods of the day, the following time-related elements are considered to influence the tourist travel behavior.

1. Hour-of-day: more people travel in the afternoon and evening.
2. Day-of-week: people tend to travel on weekends rather than in the week.
3. Holidays: more tourists are expected on holiday.
4. Seasonality: people are assumed to travel more in summer than winter.

In this research, they will firstly be fused with the counts data and the correlations will be analyzed and used to decide which elements to be taken into account in the models.

Hour-of-day and Day-of-week features are derived from the timestamps of the counts data itself. For seasonality, which is also included in the counts data, the seasonal trend will be analyzed and used to decide which months to include in the next chapter. For holidays, the observance and national holidays in the Netherlands during the counts collection period are included in the data, as listed in Table 3.2.

*Table 3.2 Holidays*

| Date | | Name | Type |
|---|---|---|---|
| 5-Dec-18 | Wednesday | St Nicholas' Eve/Sinterklaas | Observance |
| 6-Dec-18 | Thursday | St Nicholas' Day | Observance |
| 24-Dec-18 | Monday | Christmas Eve | Observance |
| 25-Dec-18 | Tuesday | Christmas Day | National holiday |
| 26-Dec-18 | Wednesday | Second Day of Christmas | National holiday |
| 31-Dec-18 | Monday | New Year's Eve | Observance |
| 1-Jan-19 | Tuesday | New Year's Day | National holiday |

(https://www.timeanddate.com/holidays/netherlands/2019).

## 3.2 Data processing

A few steps need to be taken before an exhaustive data set is ready for use. Firstly, the counts, weather, and holiday data need to be processed and merged. The original counts data includes the camera location, camera direction, date, time, and counts number. They are separated into 14 data sets with corresponding camera location and direction. In the remaining part of the report, they will be referred to as data set 1, 1R, 2, 2R, 3, 3R, 4, 4R, 5, 5R, 6, 6R, 7, 7R, where data set 1 refers to the data set collected from camera 1 in the original direction as shown in Figure 3.1 and data set 1R refers to the data set collected from camera 1 in the reverse direction. The day-of-week and hour-of-day features are derived from the timestamps, adding 7 and 24 dummy variables to the counts data set. Dummy variables take only the value 0 or 1 to indicate the absence or presence of the factor. The original weather data is at an interval of 1 hour. It is expanded at the 5-minute interval by repeating each row of an hour data by 12 times and generating a weather data set which has the same number of rows as counts data. The weather data is the same for 14 counts data sets. 10 weather variables are then fused with the counts data sets. The holiday data is generated as a dummy variable where the dates of the holidays are made 1 and the rest remains 0. 1 holiday variable is then added to the counts data.

Secondly, the completeness of the data set is checked. Data set 1 to 5R are complete along the data collection period. For data set 6 and 6R, 440 records from 09/07/2018 22:10:00 to 12/07/2018 11:45:00 are missing. For data set 7 and 7R, 685 records from 18/08/2018 04:10:00 to 20/08/2018 13:10:00, 284 records from 23/08/2018 16:30:00 to 24/08/2018 16:05:00, 408 records from 10/10/2018 03:15:00 to 11/10/2018 13:10:00, 251 records from 12/12/2018 12:45:00 to 13/12/2018 09:35:00, and 268 records from 02/01/2019 18:30:00 to 03/01/2019 15:55:00 are missing. The missing data only takes up 0.75% and 3.2% of the records of data set 6, 6R and 7, 7R. The chosen models, MLR and ARIMA, are not sensitive to a small amount of missing data. However, ARIMA modes are sensitive to the chronological order. Therefore, to keep the time series in a continuous chronological order, the missing data is kept in the data set.

# 4 Feature selection

This chapter will present the process of feature selection. Chapter 4.1 elaborates on the trends of counts data within the data collection period. Chapter 4.2 looks into the correlations between all the external features and counts. In chapter 4.3, the autocorrelation of the counts is analyzed. Chapter 4.4 provides the conclusion of selected features.

## 4.1 Trends of the data

To select significant features to be included in the model, the trends of counts over pre-selected features are studied. Though the scale of counts is different at different cameras, the related features are assumed to be similar, as they are located in the same area. Similar trends along with features of date, time, holidays, and weather are found for 14 data sets. The following analysis will be presented with findings from data set 1 and is applicable for other data sets as well.

Firstly, the trend of time of day is illustrated in Figure 4.1. In the figure, the counts of every 5 minutes are averaged from the data collection period of 204 days. It is observed from the plot that the counts pattern is highly related to time of day. The counts increase from early morning (06:45:00) to late evening (22:15:00) and decrease from late evening to early morning. The trend peaks at 22:30 and drops to almost 0 from 5:00 to 8:00. The average counts per 5 minutes are above 100 persons from 19:40:00 to 23:05:00 and below 20 persons from 03:10:00 to 10:25:00. The daily trend is similar in other locations, with different volumes of counts.

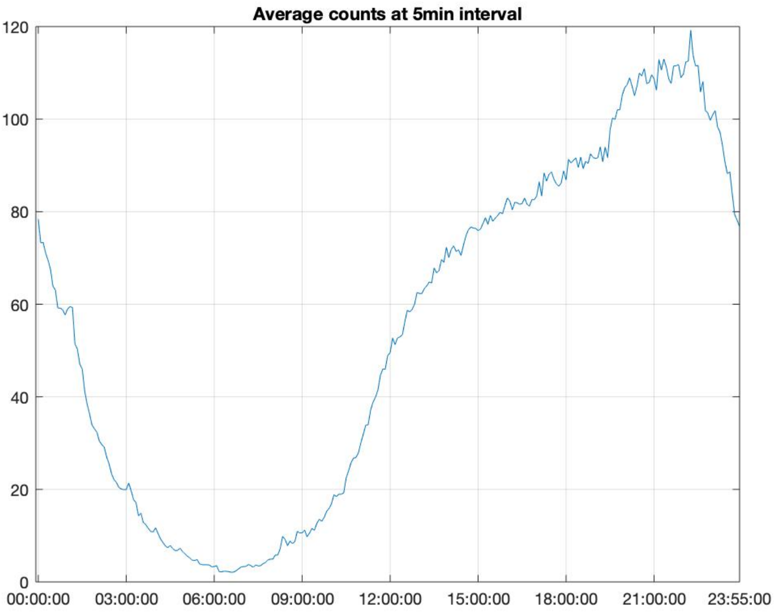


*Figure 4.1 Average counts every 5 minutes at camera 1*

Secondly, the weekly and monthly trend is plotted in Figure 4.2. In the figure, the counts of every 5 minutes are averaged from data of every day. The black lines in the stem plot show the average counts on weekdays (Monday to Thursday). The green lines indicate Friday and Sunday, red Saturday. It is observed that there is a weekly cycle in the collected data. The counts on weekends are higher than those on weekdays, with the biggest values on Saturday.

A drop is also observed in counts from November to January, except for the new year holiday. The reason may be that people travel less in the winter for the bad weather, cold temperature and short duration of daytime. Though the data set does not cover every month in a year, the seasonality can still be incorporated in the model in order to explain the drop from November to January. The seasonality cannot be replaced by including multiple weather factors, as it might reveal other preferences of tourists that are not accounted for by weather factors.

As is shown in Figure 4.2, the average counts on holidays in Table 3.2 do not display a significantly different pattern, except for the new year holiday. Therefore, only the new year holiday will be included in the holiday feature.

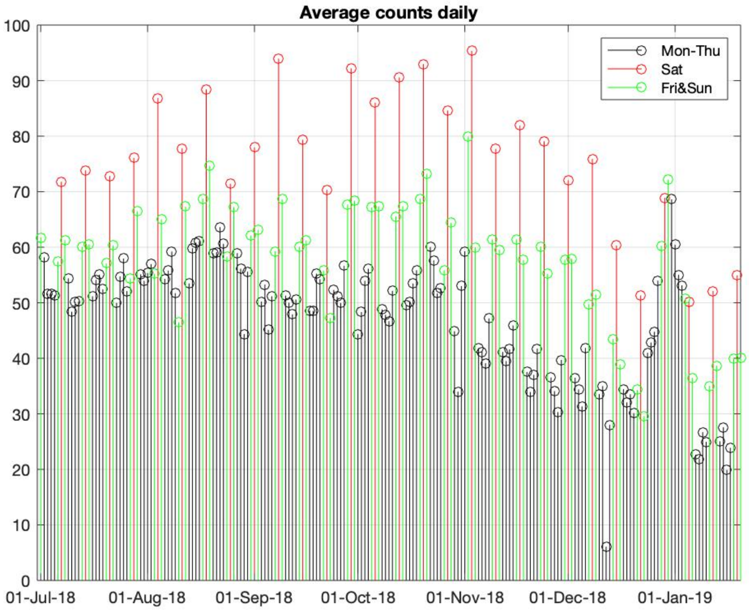


*Figure 4.2 Average counts every day*

## 4.2 Feature correlation

As the selected prediction approaches, MLR and ARIMA, are both linear methods, the linear correlations between features and counts will be tested in this section to determine to what extent are the counts linearly correlated to the features. The linear correlations are tested with correlation coefficients and p-values. A correlation coefficient ranges from -1 to +1, where ±1 indicates the strongest possible correlation and 0 no correlation. A p-value represents the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct (Wasserstein and Lazar, 2016). In this case, the null hypothesis refers to no correlation exists between the feature and counts. When the p-value is smaller than the significance level of 0.05, the null hypothesis is rejected, which means there is a linear correlation.

Data set 1, collected at camera 1 in the original direction, is used for the correlation analysis. The resulting correlation coefficients and p-values are shown in Table 4.1.

*Table 4.1 Feature correlation*

| Nr | Feature | Type | Correlation coefficient r | p-value |
|---|---|---|---|---|
| 1 | Hour 0 | Dummy | 0.0548 | 0 |
| 2 | Hour 1 | Dummy | -0.0359 | 0 |
| 3 | Hour 2 | Dummy | -0.1321 | 0 |
| 4 | Hour 3 | Dummy | -0.1767 | 0 |
| 5 | Hour 4 | Dummy | -0.2092 | 0 |
| 6 | Hour 5 | Dummy | -0.2262 | 0 |
| 7 | Hour 6 | Dummy | -0.2346 | 0 |
| 8 | Hour 7 | Dummy | -0.2290 | 0 |
| 9 | Hour 8 | Dummy | -0.2092 | 0 |
| 10 | Hour 9 | Dummy | -0.1898 | 0 |
| 11 | Hour 10 | Dummy | -0.1459 | 0 |
| 12 | Hour 11 | Dummy | -0.0673 | 0 |
| 13 | Hour 12 | Dummy | 0.0064 | 0 |
| 14 | Hour 13 | Dummy | 0.0545 | 0 |
| 15 | Hour 14 | Dummy | 0.0875 | 0 |
| 16 | Hour 15 | Dummy | 0.1101 | 0 |
| 17 | Hour 16 | Dummy | 0.1263 | 0 |
| 18 | Hour 17 | Dummy | 0.1469 | 0 |
| 19 | Hour 18 | Dummy | 0.1655 | 0 |
| 20 | Hour 19 | Dummy | 0.1932 | 0 |
| 21 | Hour 20 | Dummy | 0.2452 | 0 |
| 22 | Hour 21 | Dummy | 0.2546 | 0 |
| 23 | Hour 22 | Dummy | 0.2488 | 0 |
| 24 | Hour 23 | Dummy | 0.1620 | 0 |
| 25 | Monday | Dummy | -0.0635 | 0 |
| 26 | Tuesday | Dummy | -0.0718 | 0 |
| 27 | Wednesday | Dummy | -0.0768 | 0 |
| 28 | Thursday | Dummy | -0.0493 | 0 |
| 29 | Friday | Dummy | 0.0243 | 0 |
| 30 | Saturday | Dummy | 0.1937 | 0 |
| 31 | Sunday | Dummy | 0.0428 | 0 |
| 32 | Seasonality (Nov-Jan) | Dummy | -0.1444 | 0 |
| 33 | Holiday (New year eve - New year) | Dummy | 0.0225 | 0 |

| | | | | |
|---|---|---|---|---|
| 34 | Average wind speed in an hour | Number | -0.0647 | $1.5517e^{-55}$ |
| 35 | Maximum wind speed in an hour | Number | -0.0505 | $1.4823e^{-34}$ |
| 36 | Temperature | Number | 0.0966 | $1.0634e^{-121}$ |
| 37 | Duration of sunshine in an hour | Number | -0.1179 | $7.2052e^{-181}$ |
| 38 | Duration of rain in an hour | Number | -0.0745 | $4.34345e^{-73}$ |
| 39 | Precipitation of rain in an hour | Number | -0.0420 | $2.4016e^{-24}$ |
| 40 | Rain | Dummy | -0.06178 | $8.7203e^{-51}$ |
| 41 | Snow | Dummy | -0.0049 | 0.2376 |
| 42 | Thunderstorm | Dummy | 0.0044 | 0.2895 |
| 43 | Ice formation | Dummy | -0.0274 | $3.1640e^{-11}$ |

As is shown in Table 4.1, hourly, weekdays' and seasonality features are highly correlated with p-values of 0. Features in orange show small p-values, but their absolute correlation coefficients are relatively small, which indicates their linear influences are relatively small. Features in red have p-values bigger than 0.05, indicating they are not correlated to the counts. The duration of sunshine feature in blue is negatively correlated to the counts, which is on contrary to the assumption that people tend to travel more in the district during the sun. The reason is that more people travel in the district during the night when there is no sunshine.

## 4.3 Autocorrelation of counts

To look into the correlation between the counts and the historical data and determine the horizon of historical data to be considered in prediction models, the autocorrelation of counts is studied. Autocorrelation is the correlation of a time series with a delayed copy of itself as a function of delay. It represents the similarity between observations with lags and shows the repeating patterns in the time series. It can be measured by an autocorrelation function, which shows the correlation between $y_t$ and $y_{t+k}$, where k = 0,...,K and $y_t$ is a stochastic process. The autocorrelation for lag k is $r_k = \frac{c_k}{c_0}$ . where $c_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_t + k - \bar{y})$; $c_0$ is the sample variance of the time series (Box, Jenkins, & Reinsel, 2011).

The autocorrelation of counts from camera one in the original direction is plotted in Figure 3.9.1 to 3.9.3 with up to 60 lags and 600 lags (5 minutes per lag). As is shown in Figure 3.9.1, the autocorrelation of counts within 12 lags (1 hour) decreases slowly but still remains above 0.8. With more lags, the autocorrelation decreases faster and drops to around 0.25 in 60 lags (5 hours). As is shown in Figure 3.9.2, the counts are highly correlated to the counts from the previous days at the same time, which implies a repeating daily pattern of counts. The autocorrelation remains 0.8 from 24 hours ago and still above 0.7 from 48 hours ago. The periodicity of the weekly pattern is clearly shown in Figure 3.9.3, as the autocorrelation of the counts from 1 week and 2 weeks ago is higher than on the other days. The counts from 1 week ago have an even higher autocorrelation than that of 1 day ago, implying a repeating weekly pattern. A fluctuation in the ACF is observed in Figure 3.9.2 and 3.9.3, that the autocorrelation decreases to negative values and increases to a maximum positive value every 24 hours, which shows that the counts have a positive correlation with counts from several hours around the observation time but a negative correlation with counts from around 12 hours earlier. This pattern indicates that the counts value fluctuates in the form of a sine curve every 24 hours, which is in line with the daily counts pattern plotted in Figure 4.1.

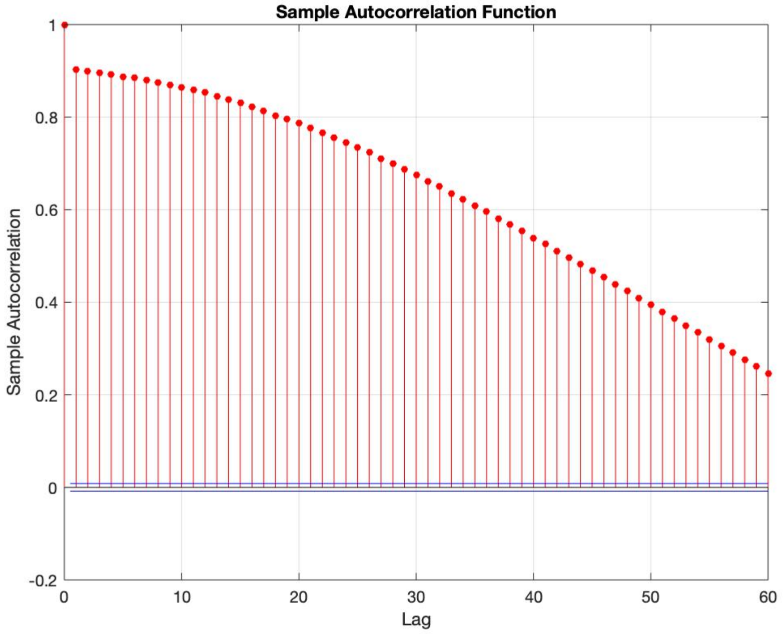


*Figure 4.3.1 Autocorrelation of counts within 60 lags*

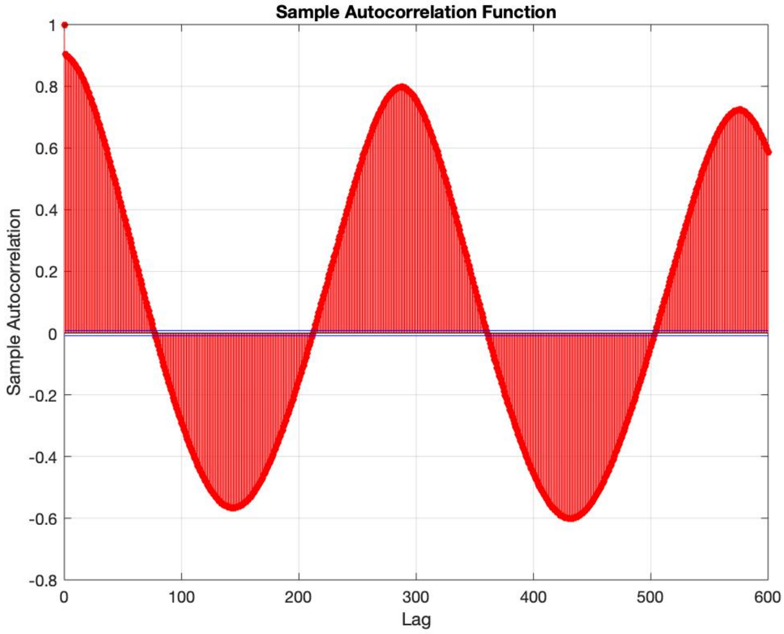


*Figure 4.3.2 Autocorrelation of counts within 600 lags*

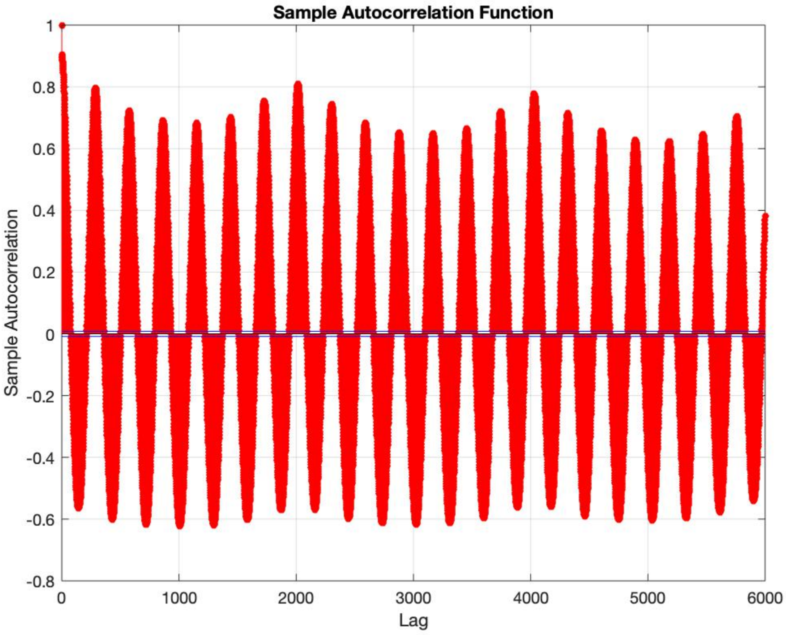


*Figure 4.3.2 Autocorrelation of counts within 6000 lags*

## 4.4 Conclusion

According to the linear correlation analysis in chapter 4.2, feature snow and thunderstorm have a p-value bigger than 0.05 and therefore do not show a significant influence on the counts. They will not be included in the model. For day-of-week and hour-of-day features, one out of the same category needs to be deleted from the variables to provide a reference to other features in the same category. Monday from 7 days of a week and 12:00:00 from 24 hours in a day are chosen to be the reference, as the counts on Monday and at 12:00:00 are non-zero and less fluctuating over the time period of data collection. After deleting 4 features from Table 4.1, there are 39 features left.

Features derived from the autocorrelation also needs to be included in the model. To correspond to the short-term prediction horizon of 30 minutes, the autocorrelation of counts from 7 lags earlier are looked into. The autocorrelation of counts from lags up to 36 remains above 0.6. Therefore, for MLR, the counts from 30 minutes earlier/the average counts from 30 minutes to 1h earlier/the average counts from 30 minutes to 3h earlier will be included in the model as an extra feature to explore which scale of historical data performs better for short-term prediction. Therefore, there will be 40 features used in the model development. For ARIMA, the autocorrelation of counts is included in the model itself. The number of lags taken into account will be determined during model training. The data set for ARIMA will include 39 features.

# 5 Methodology

This chapter presents the methodology of model development, validation and performance evaluation. In chapter 5.1, three types of linear regression models will be described. Chapter 5.2 introduces 2 types of ARIMA models to be developed in this research. Chapter 5.3 provides the validation methods and chapter 5.4 presents the methods of performance evaluation.

## 5.1 Linear regression models

Three types of linear regression models will be trained with the Regression Learner app in Matlab: multiple linear regression, interactions linear regression, robust linear regression.

The multiple linear regression model uses a constant term and linear terms in the predictors.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

The interactions linear regression considers a constant term, linear terms, and interaction terms between the predictors, which include the possible dependency of pairs of two features. The number of beta in the model is much larger than the basic linear regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \ldots + \beta_{(n-1)n} x_{n-1} x_n$$

The robust linear regression uses a robust objective function and makes the model less sensitive to outliers. Its fitting method automatically assigns lower weights to data points that are more likely to be outliers.

Among the three models, the basic linear regression and robust linear regression provide very low flexibility, while interactions linear regression provides medium flexibility by considering interactions between features. Therefore, interactions linear regression is expected to perform better than the other two models.

## 5.2 ARIMA models

ARIMA and ARIMAX models will be trained with the Econometric Modeler app in Matlab. The ARIMA models make predictions through the weighted linear function of the past observations $y_{t-n}$ of that variable (AR), an extra term for structural trends in the data (I), past error terms $\varepsilon_t$ (MA). The degree of integration is set as 1 to obtain linear relations.

$$(1 - \phi_1 L - \ldots - \phi_n L^n)(1 - L) y_t = c + (1 + \theta_1 L - \ldots - \theta_n L^n)\varepsilon_t$$

The ARIMAX models also take into account external features (X).

$$(1 - \phi_1 L - \ldots - \phi_n L^n)(1 - L) y_t = c + X_1 \beta_1 + \ldots + X_m \beta_m + (1 + \theta_1 L - \ldots - \theta_n L^n)\varepsilon_t$$

The lags to be included in the models will be determined after the training of the MLR models, which provides the reference of the significance of autocorrelation from 30 minutes/30 minutes to 1 hour/30 minutes to 3 hours. The external features to be included will be determined during the training of ARIMAX models, as many features can be represented by the trends in the time series themselves. The ARIMAX models with the most lags are expected to perform better, as more information is taken into account.

## 5.3 Model validation

Validating models with data that is not included during training gives better insight into the generalization error. The technique which separates data for training and testing is called cross-validation, Cross-validation has been used for prediction models for estimation of how accurately a predictive model will perform in practice. Different methods exist for cross-validation. To avoid overfitting, k-fold cross-validation will be used. By this method, the data are split randomly in k folds with equal sizes, with these folds k models will be trained where for each model (k-1) folds are used for training and the other fold is left for testing.

For MLR models, k-fold cross-validation with 5 folds is used, which is also achieved through the Regression Learner app in Matlab. For ARIMA models, the random split is not feasible as the time series has to be kept in the chronological order to obtain daily, weekly and monthly trends. Therefore, to guarantee the model quality, the data is separated in the chronological order every 4 weeks into 3 weeks for the training set and 1 week for testing. To make sure the ARIMA models capture as many trends in the time series as possible, the data is split from 07:00:00 in the morning, where the smallest counts are observed. For the records of 58752, 7*3 weeks of data will be used for training and 7*1 weeks for testing. Besides the complete training and testing data sets, there are 8 days left in the processed data set, giving flexibility of validation. Out of the 8 days, the data split is randomly done five times for training and testing. However, by keeping data in the chronological order, the 5 testing sets do not cover the whole time period of data collection. The possibility of overfitting is higher than k-fold validation.

## 5.4 Performance evaluation

The performance of models is compared by Root Mean Squared Error (RMSE) and $R^2$ score. RMSE is the standard deviation of the residuals (prediction errors), which measures the differences between values predicted by a model and the observed values. It is used to compare forecasting errors of different models developed from the same data set (Hyndman and Koehler, 2006). $R^2$ is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by variables in a regression model and can be used to compare the results with other researches (Barten, 1987). Smaller RMSE and bigger $R^2$ indicate a more accurate prediction of the model. The best model will be determined by comparing the values of RMSE and $R^2$.

# 6 Model development and results

In this chapter, MLR and ARIMA models will be developed and their performance will be evaluated. Chapter 6.1 provides the development and performance of MLR models. ARIMA models are presented in chapter 6.2. A comparison of model performance is made in chapter 6.3 and is analyzed in chapter 6.4.

## 6.1 MLR models

The historical data used for MLR models will be selected in section 6.1.1. The model results will be presented in section 6.1.2.

### 6.1.1 Historical data selection

To decide which scale of historical data performs better for short-term prediction, data collected from camera 1 is used for model training. The time periods to be tested are determined in chapter 4, according to the autoregression pattern. The results are shown in Figure 6.1.1 and 6.1.2.

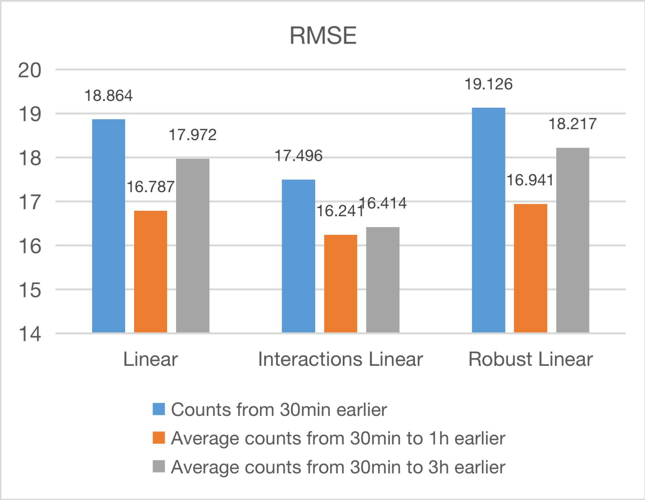


*Figure 6.1.1 RMSE of models with different historical data*

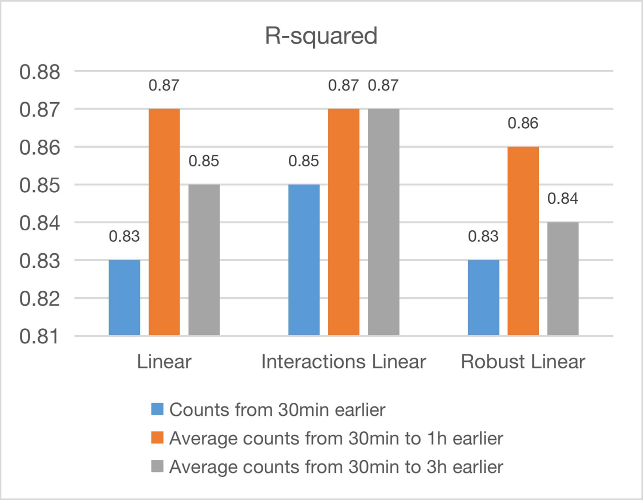
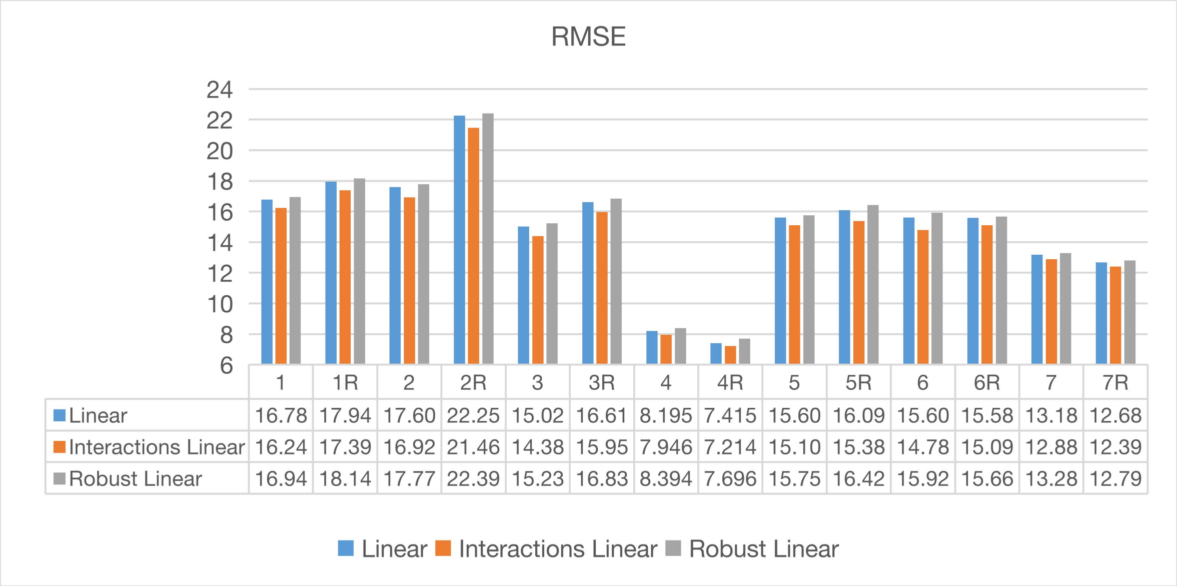


*Figure 6.1.2 $R^2$ of models with different historical data*

As is shown in the figures, models trained from the historical data of the average counts from 30 minutes to 1h earlier have smaller RMSE and bigger $R^2$, which indicates the average counts from 30 minutes to 1h earlier correlate better with the predicted counts. It is also observed that interactions linear regression has a better performance over the other two models which do not consider the dependency among features.

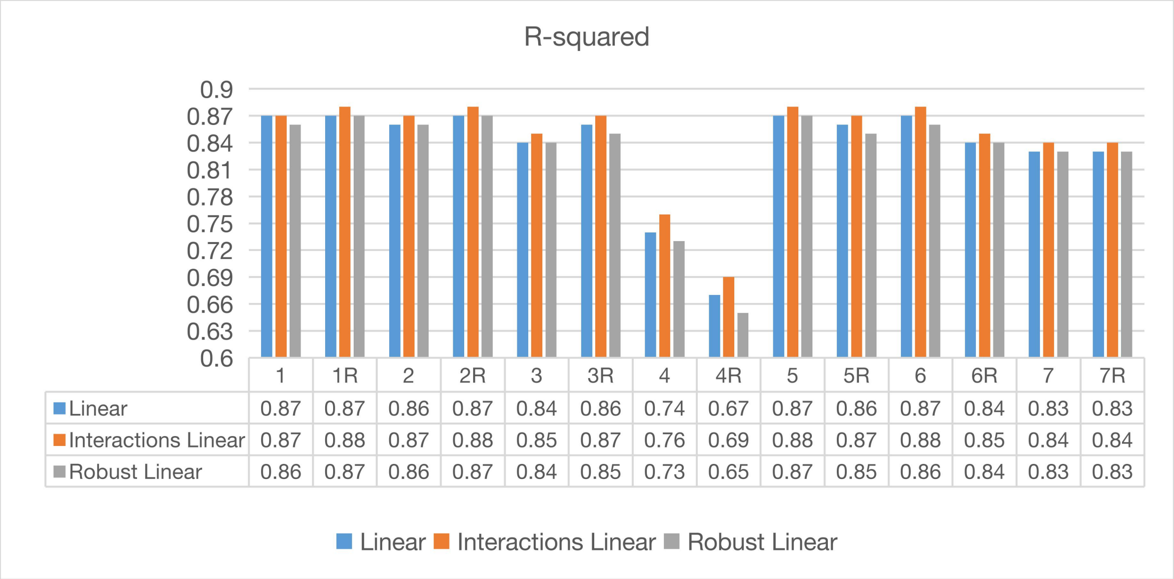
### 6.1.2 MLR model results

Three types of linear regression models are trained with 14 data sets, resulting in 42 models. All the models use the same 40 variables, as determined in chapter 4. The RMSE and $R^2$ are illustrated in Figure 6.2.1 and 6.2.2.



| | 1 | 1R | 2 | 2R | 3 | 3R | 4 | 4R | 5 | 5R | 6 | 6R | 7 | 7R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Linear | 16.78 | 17.94 | 17.60 | 22.25 | 15.02 | 16.61 | 8.195 | 7.415 | 15.60 | 16.09 | 15.60 | 15.58 | 13.18 | 12.68 |
| Interactions Linear | 16.24 | 17.39 | 16.92 | 21.46 | 14.38 | 15.95 | 7.946 | 7.214 | 15.10 | 15.38 | 14.78 | 15.09 | 12.88 | 12.39 |
| Robust Linear | 16.94 | 18.14 | 17.77 | 22.39 | 15.23 | 16.83 | 8.394 | 7.696 | 15.75 | 16.42 | 15.92 | 15.66 | 13.28 | 12.79 |

*Figure 6.2.1 RMSE of 3 models with 14 data sets*



| | 1 | 1R | 2 | 2R | 3 | 3R | 4 | 4R | 5 | 5R | 6 | 6R | 7 | 7R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Linear | 0.87 | 0.87 | 0.86 | 0.87 | 0.84 | 0.86 | 0.74 | 0.67 | 0.87 | 0.86 | 0.87 | 0.84 | 0.83 | 0.83 |
| Interactions Linear | 0.87 | 0.88 | 0.87 | 0.88 | 0.85 | 0.87 | 0.76 | 0.69 | 0.88 | 0.87 | 0.88 | 0.85 | 0.84 | 0.84 |
| Robust Linear | 0.86 | 0.87 | 0.86 | 0.87 | 0.84 | 0.85 | 0.73 | 0.65 | 0.87 | 0.85 | 0.86 | 0.84 | 0.83 | 0.83 |

*Figure 6.2.2 $R^2$ of 3 models with 14 data sets*

As is observed from Figure 6.2.1, the interactions linear model has a smaller RMSE among 3 linear models with all the data sets, which indicates a better prediction result with the interactions linear model. However, as the models are trained with different sets of data, the performance of RMSE cannot be used to compare the prediction accuracy among 14 data sets. $R^2$ provides the proportion of the variance in the model and can be used to compare the prediction accuracy among 14 data sets. What is illustrated in Figure 6.2.2 is in line with the finding from Figure 6.2.1, namely almost all the interactions linear models have the smallest $R^2$ among 3 types of models. The $R^2$ of all the interactions linear models, except the two at camera 4 and 4R, are equal or bigger than 0.84, with the maximum reaching 0.88. These high $R^2$ values imply that the models are able to capture the trends in the counts data relatively accurately. The predictions at camera 4 and 4R are not as good as other locations as the $R^2$ of interactions linear models is 0.76 and 0.69, indicating that the models predict for camera 4 and 4R with a larger variation to the observations.

## 6.2 ARIMA models

The number of lags considered in ARIMA models will be determined in section 6.2.1. Section 6.2.2 presents the process of external feature selection. The model results will be provided in section 6.2.3.

### 6.2.1 Defining number of lags

According to the historical data selection of MLR models, it is assumed that counts from up to 1 hour ago provide accurate predictions. Besides, two more features (AR and MA) are introduced into the model with each lag. Therefore, to limit the number of features, the lags of both AR and MA to be taken into account in the model will be defined within 1 hour: 12/6/3 lags. From the autocorrelation plot in Figure 4.3.1, it is observed that the nearest lags are the most related. However, from the practical point of view, predictions made in advance help better with crowd management. Therefore, lag 4 to 6 (15 minutes to 30 minutes)/lag 7 to 12 (30 minutes to 1 hour) will also be considered in the model for both AR and MA.

To compare which range of lags provides the most accurate prediction, data 1 is used for model training. The RMSE and $R^2$ are illustrated in Figure 6.3.1 and 6.3.2.

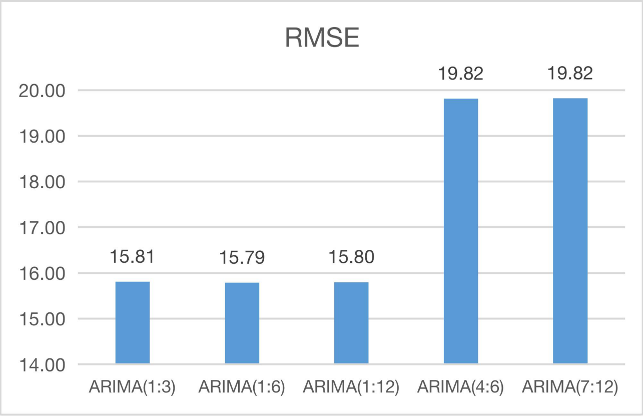


*Figure 6.3.1 RMSE of ARIMA models with different lags*

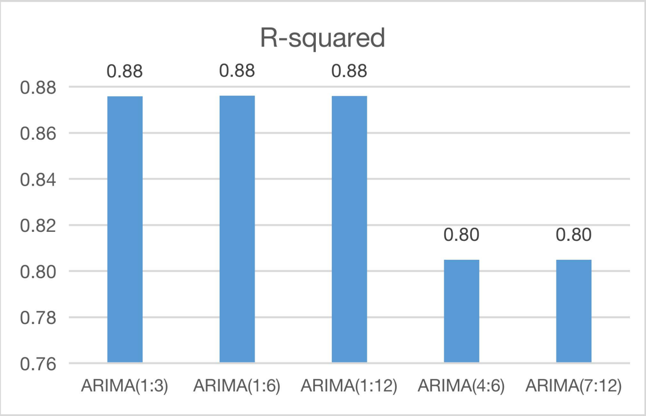


*Figure 6.3.2 $R^2$ of ARIMA models with different lags*

As is shown in Figure 6.3.1 and 6.3.2, the ARIMA models with 12/6/3 lags perform almost the same for counts prediction and have smaller RMSE values and bigger R-squared values, which suggests that the trend of the time series can be captured by taking into account the nearest lag, while adding further lags does not improve the prediction accuracy. The models that predict with a few lags behind, including 4 to 6 lags and 7 to 12 lags, also have the same model performance, which is not as good as the models that consider the nearest lags. RMSE and $R^2$ of ARIMA(4:6) and ARIMA(7:12) is by 4 higher and by 0.08 lower than that of ARIMA models with 12/6/3 lags, losing 9% of $R^2$, which indicates that the nearest lags are more important than counts from a few lags behind. Among the models with the highest $R^2$ value, ARIMA(1:3) is chosen for further model development, as the performance is as good as models with more lags, but it considers the smallest number of lags, which reduces the number of features in the model.

### 6.2.2 Defining external features

As external features are very likely to be captured by modeling the time series in ARIMA, some pre-selected features might be not significant in the ARIMAX model. Therefore, the correlations of features and counts need to be analyzed and significant features need to be selected. The selection of external features is by the stepwise method, which firstly adds 39 features to the ARIMAX, then deletes the features with a p-value bigger than 0.1 and runs the model with the left features to once again deletes the features with a p-value bigger than 0.1. It is observed from practice that when the p-value boundary is set as 0.05, the model variance grows after deleting the features. However, when it is set as 0.1, the model variance decreases, which shows a better fit to the data. Therefore, the p-value boundary is set as 0.1.

Data 1 is used for the external feature selections. The ARIMAX model is developed with 3 lags of autoregressive order and moving average order. During the iterations of deleting features, it is observed that with more insignificant features being deleted, the p-value of AR and MA decreases, which indicates that the deleted features are captured by AR and MA features. After several iterations, 27 external features are selected, as shown in Table 6.1. It is assumed that the significant features for 14 data sets remain the same.

*Table 6.1 Selected features*

| Nr | Feature | Type | Beta value | p-value |
|---|---|---|---|---|
| 1 | Hour 0 | Dummy | 0.0712 | 0.0026 |
| 2 | Hour 1 | Dummy | -1.7621 | 1.13E-41 |
| 3 | Hour 2 | Dummy | -2.4345 | 6.22E-60 |
| 4 | Hour 3 | Dummy | 0.4663 | 8.03E-06 |
| 5 | Hour 4 | Dummy | -0.2682 | 0.0018 |
| 6 | Hour 5 | Dummy | -0.4193 | 9.20E-09 |
| 7 | Hour 6 | Dummy | -0.5304 | 1.70E-11 |
| 8 | Hour 7 | Dummy | -0.6009 | 2.22E-14 |
| 9 | Hour 8 | Dummy | -0.3934 | 1.03E-07 |
| 10 | Hour 9 | Dummy | -0.7494 | 3.18E-20 |
| 11 | Hour 11 | Dummy | 0.2167 | 0.0013 |
| 12 | Hour 13 | Dummy | -1.42 | 1.01E-22 |
| 13 | Hour 14 | Dummy | -0.6053 | 1.02E-17 |
| 14 | Hour 15 | Dummy | -0.5023 | 1.29E-12 |
| 15 | Hour 16 | Dummy | -1.44 | 2.11E-55 |
| 16 | Hour 17 | Dummy | -2.4673 | 1.13E-67 |
| 17 | Hour 18 | Dummy | -1.3377 | 2.13E-16 |
| 18 | Hour 19 | Dummy | -1.0022 | 9.23E-07 |
| 19 | Hour 20 | Dummy | -0.9473 | 0.0022 |

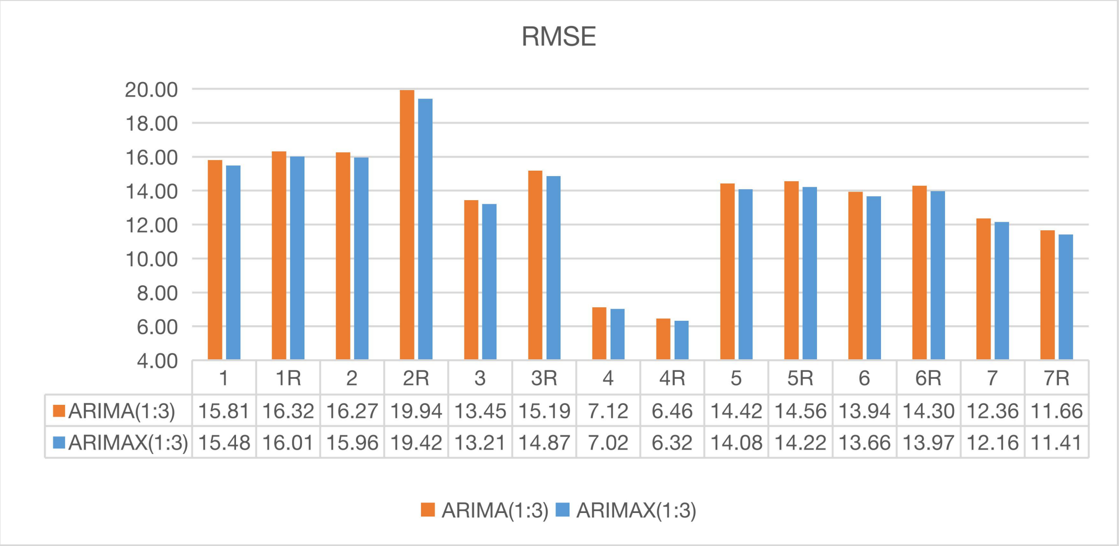
| | | | | |
|---|---|---|---|---|
| 20 | Hour 21 | Dummy | -0.7771 | 0.0367 |
| 21 | Hour 22 | Dummy | -0.5977 | 0.0596 |
| 22 | Hour 23 | Dummy | -0.4796 | 0.0291 |
| 23 | Friday | Dummy | -0.4291 | 0.0052 |
| 24 | Saturday | Dummy | 0.037 | 0.0685 |
| 25 | Sunday | Dummy | -0.1157 | 2.20E-06 |
| 26 | Precipitation of rain in an hour | Number | -0.0052 | 0.0589 |
| 27 | Duration of rain in an hour | Number | -0.0185 | 0.0029 |

As is shown in Table 6.1, hour 10 is deleted because of the resemblance of counts at hour 12. Tuesday to Thursday are removed as the detected pattern is similar to Monday. Holiday and seasonality are not significant to ARIMAX models. It might be because the holiday pattern is similar to the weekend pattern and therefore, omitted by the model. Almost all the weather-related features are covered by the trends in the time series, except precipitation and duration of rain in an hour.

Therefore, the ARIMA models will be trained with 6 features - 3 AR and 3 MA features. The ARIMAX models will be trained with 34 features - 3 AR, 3 MA features and 27 external features.

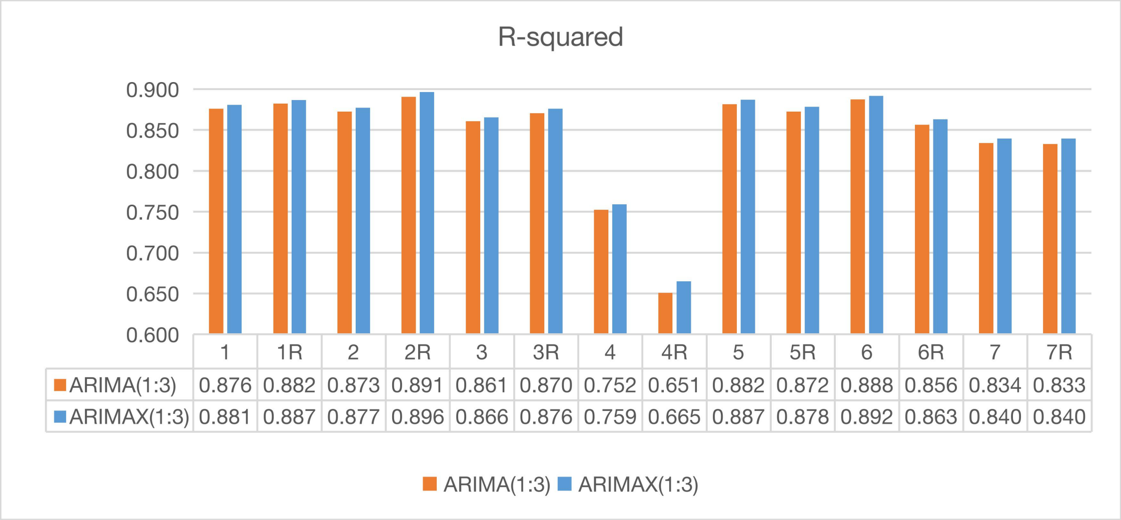
### 6.2.3 ARIMA model results

14 ARIMA models and 14 ARIMAX models are trained and tested with 14 data sets, all with 3 lags of autoregressive order and moving average order. The RMSE and R$^2$ of the models are illustrated in Figure 6.4.1 and 6.4.2.



| | 1 | 1R | 2 | 2R | 3 | 3R | 4 | 4R | 5 | 5R | 6 | 6R | 7 | 7R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARIMA(1:3) | 15.81 | 16.32 | 16.27 | 19.94 | 13.45 | 15.19 | 7.12 | 6.46 | 14.42 | 14.56 | 13.94 | 14.30 | 12.36 | 11.66 |
| ARIMAX(1:3) | 15.48 | 16.01 | 15.96 | 19.42 | 13.21 | 14.87 | 7.02 | 6.32 | 14.08 | 14.22 | 13.66 | 13.97 | 12.16 | 11.41 |

*Figure 6.4.1 RMSE of ARIMA and ARIMAX models*



| | 1 | 1R | 2 | 2R | 3 | 3R | 4 | 4R | 5 | 5R | 6 | 6R | 7 | 7R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 🟧 ARIMA(1:3) | 0.876 | 0.882 | 0.873 | 0.891 | 0.861 | 0.870 | 0.752 | 0.651 | 0.882 | 0.872 | 0.888 | 0.856 | 0.834 | 0.833 |
| 🟦 ARIMAX(1:3) | 0.881 | 0.887 | 0.877 | 0.896 | 0.866 | 0.876 | 0.759 | 0.665 | 0.887 | 0.878 | 0.892 | 0.863 | 0.840 | 0.840 |

*Figure 6.4.2 $R^2$ of ARIMA and ARIMAX models*

The ARIMA and ARIMAX models show a similar pattern of RMSE and R-squared to the MLR models. The RMSE peaks at camera 2R and reaches the minimum at camera 4 and 4R. The R-squared is also smaller at camera 4 and 4R, which shows the prediction accuracy of camera 4 and 4R is not as good as other locations. Out of expectation, ARIMAX models will 27 external features only reduce RMSE by 0.3 and increase $R^2$ by 0.05 on average, which indicates the external features are mostly captured by the autoregression of the time series itself and adding external data leads to overfitting.

## 6.3 Comparison of results

The performances of all the models are compared in Figure 6.5.1 and 6.5.2.

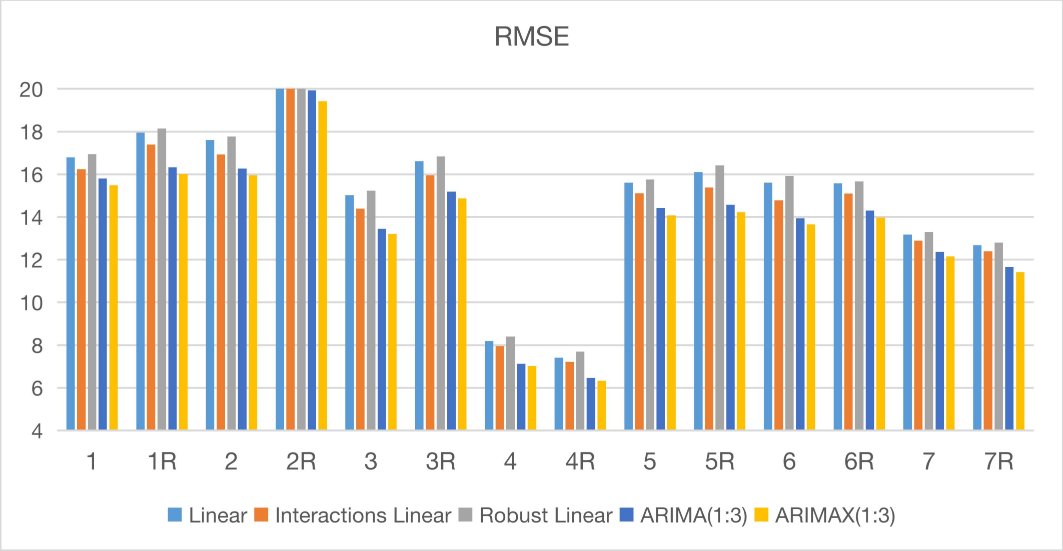


*Figure 6.5.1 RMSE of all the models*

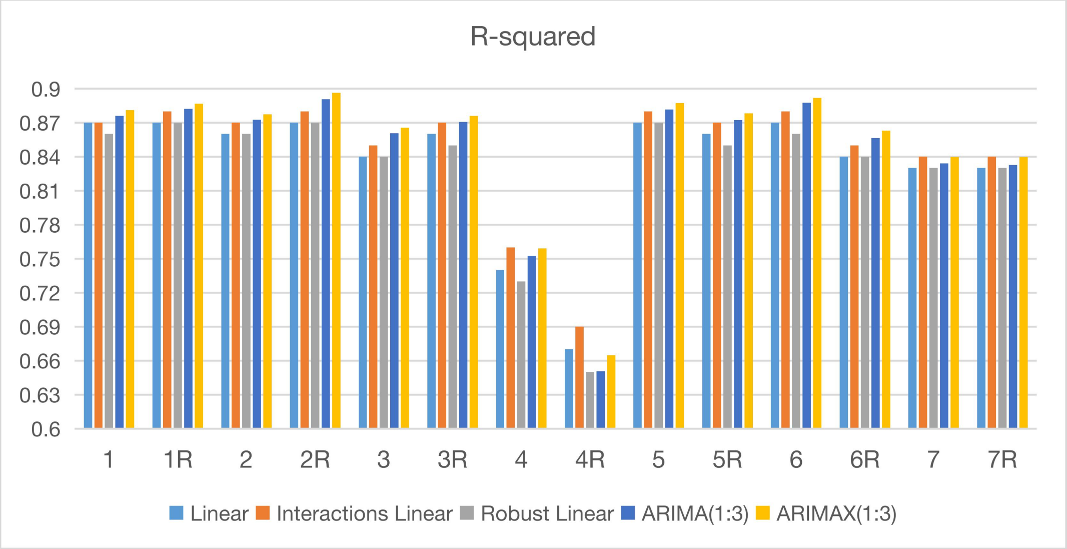


*Figure 6.5.2 $R^2$ of all the models*

As shown in the figures, the RMSE of ARIMA and ARIMAX models is smaller than that of MLR models at every location. However, $R^2$ of ARIMA and ARIMAX models are not always larger than that of MLR models. At camera 4R, the interactions linear model performs better, which indicates including the dependency between features or simply including more external features that may lead to better predictions at camera 4R. At camera 7 and 7R, the interactions linear models have the same $R^2$ as ARIMAX models. It may be because of the most missing data is included in data 7 and 7R, leading to poor predictions of ARIMA and ARIMAX models. For all the models, the lowest $R^2$ occurs at camera 4R, followed by camera 4. At most locations, ARIMAX models provide the most accurate predictions, with $R^2$ larger than that of interactions MLR by 0.1, which means that ARIMAX models are able to make a better prediction with fewer external features compared to MLR models.

## 6.4 Performance analysis

As it is observed from the results that all the models provide relatively poorer predictions at camera 4 and 4R, the residuals of model predictions and autocorrelation of data sets are analyzed.

Firstly, the residuals plots of camera 1, 4, and 4R of interactions linear regression and ARIMAX models are compared. Camera 1 acts as the reference for ‘good performance’. Figure 6.6.1 to 6.6.3 plot the residuals along with the record number, which is the number of data in the chronological order. For the MLR model, the residuals from cross-validation are plotted. For the ARIMAX model, the residuals of one the validation set are plotted. Figure 6.7.1 to 6.7.3 plot the residuals along with the true response, which is the value of the observed data in the validation sets. For the interactions linear regression model, the residuals from cross-validation are plotted. For the ARIMAX model, the residuals of one the validation set are plotted.

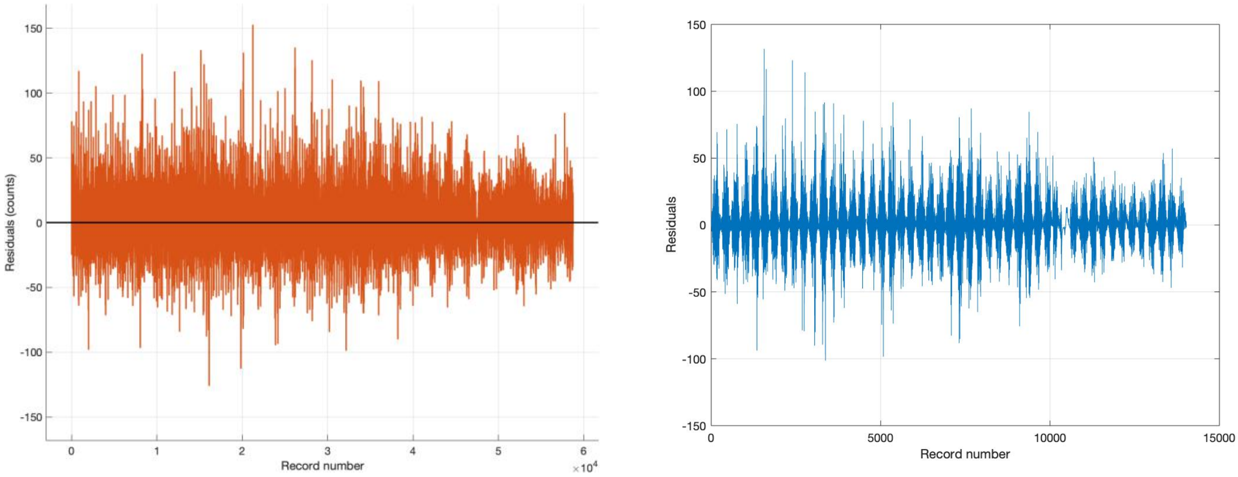


*Figure 6.6.1 Residuals at camera 1 along with record number (Interactions linear and ARIMAX)*

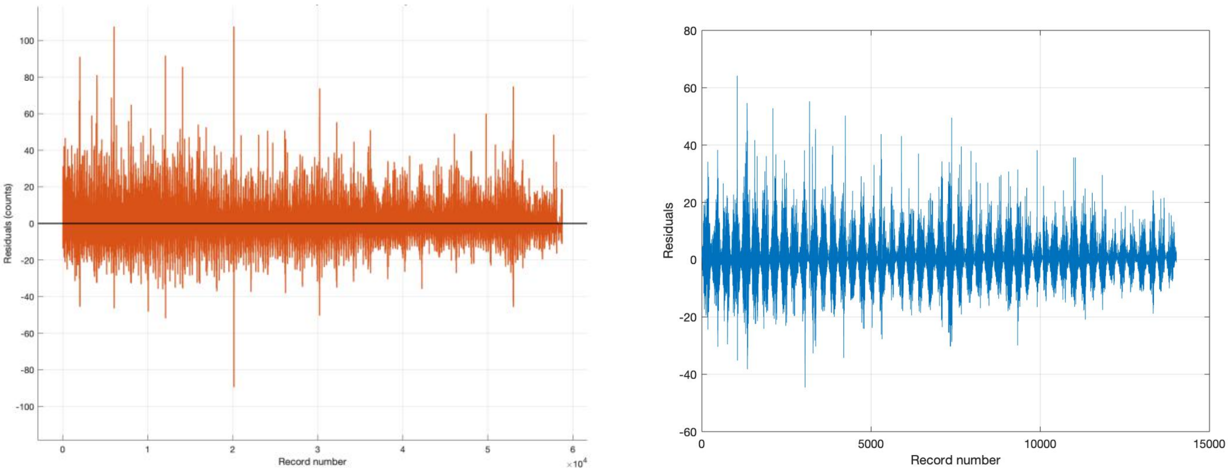


*Figure 6.6.2 Residuals at camera 4 along with record number (Interactions linear and ARIMAX)*

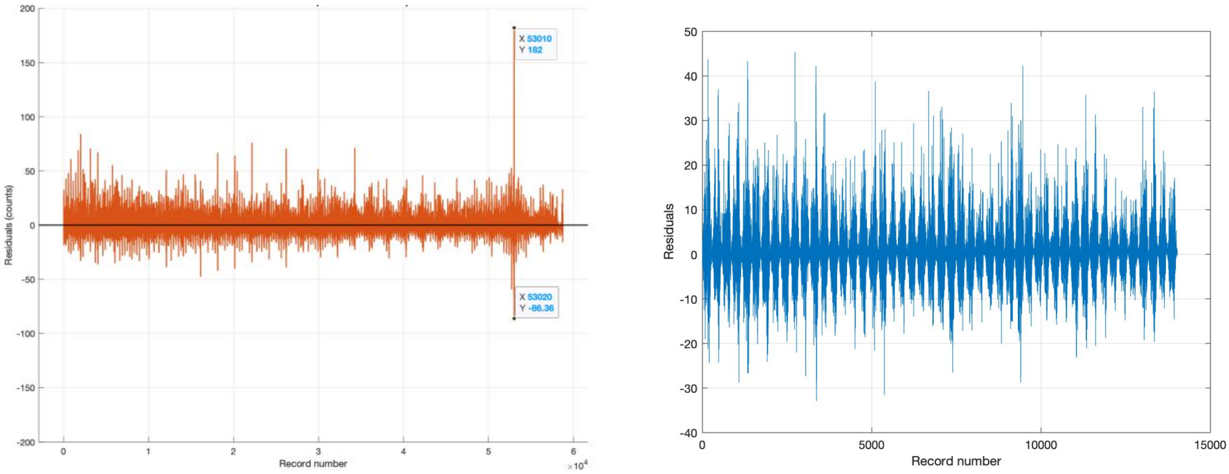


*Figure 6.6.3 Residuals at camera 4R along with record number (Interactions linear and ARIMAX)*

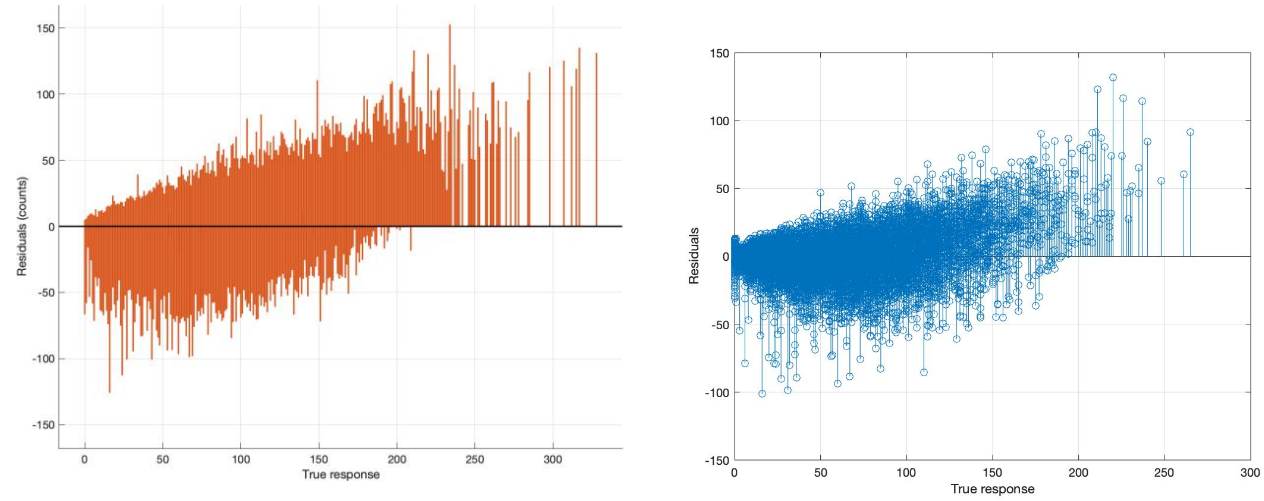


*Figure 6.7.1 Residuals at camera 1 along with observed data(Interactions linear and ARIMAX)*

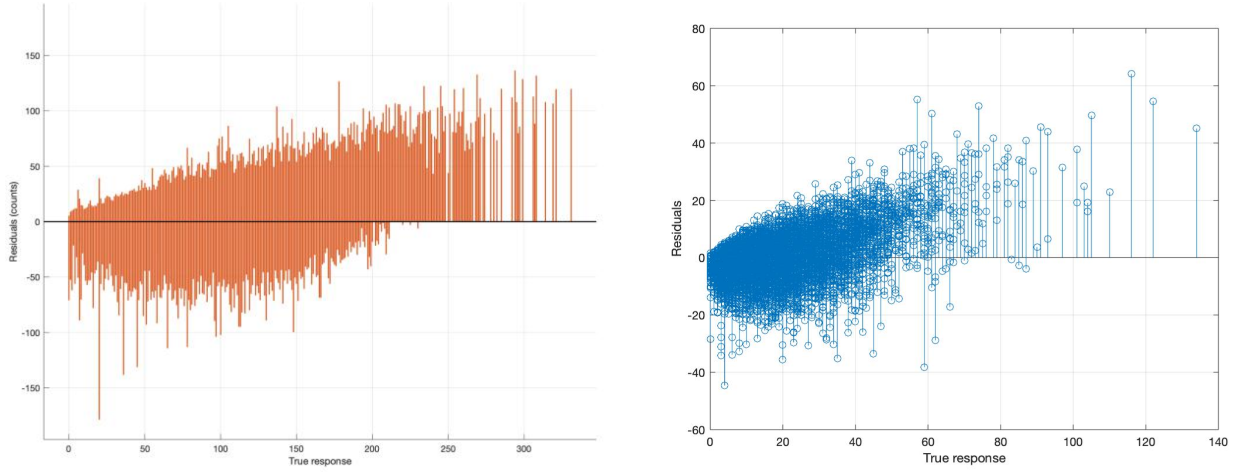


*Figure 6.7.2 Residuals at camera 4 along with observed data(Interactions linear and ARIMAX)*

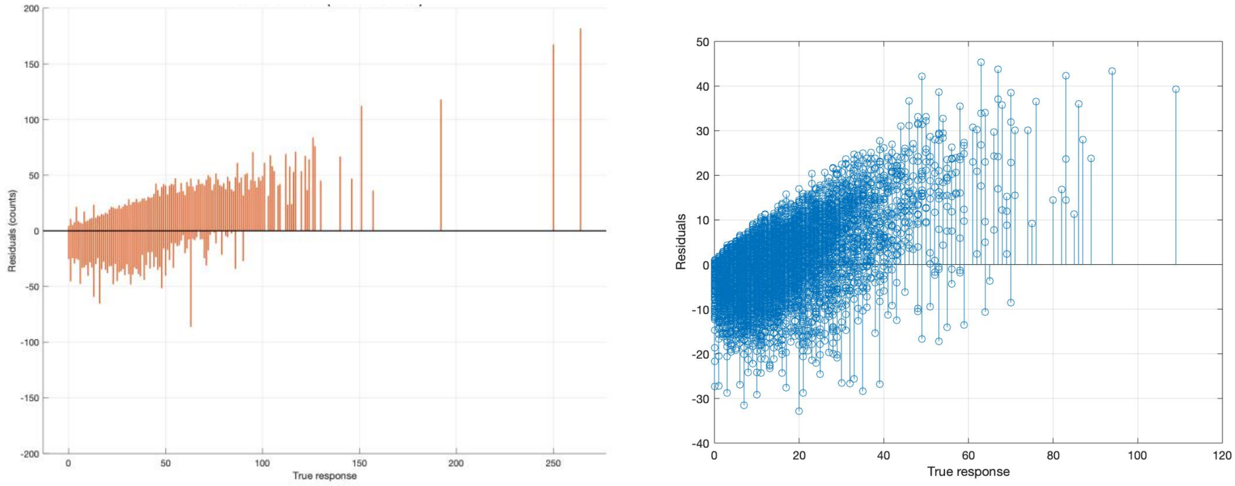


*Figure 6.7.3 Residuals at camera 4R along with observed data(Interactions linear and ARIMAX)*

As is shown in Figure 6.6.1 to 6.6.3, more outliers are observed in the models of camera 4 and 4R compared to camera 1R, of which the residuals are less scattered. The outliners of models of different cameras occur in different time periods. However, 2 residuals from MLR models at around number 53010 and 53020 are much higher compared to other numbers for all the models, especially for the model of camera 4R. Number 53010 and 53020 corresponds to the time of 01:30:00 and 02:20:00 on 01.Jan.2019, which fall into the holiday and hour-of-day feature, indicating these features might not be accurately captured in the model - their relations with counts might be non-linear. A daily pattern is observed from the residuals plots of both interactions linear regression and ARIMAX model results, that the residuals increase on both sides of the axis from early morning (03:00 - 9:00) to evening (18:00 - 00:00) and decreases to around 0 in the early morning every 24 hours. The daily pattern of residuals suggests that both types of models are able to capture the data trend that the counts remain a small number in the early morning (03:00 - 9:00). More variances, both positive and negative, are observed for the prediction of counts during the day, the absolute values of which grow to the maximum in the evening (18:00 - 00:00), indicating the models predict with a bigger variation to the counts when the counts are big.

In Figure 6.7.1 to 6.7.3, it is illustrated that for all models, the residuals along the value of observed data do not display a symmetrical distribution around 0. More negative residuals correspond with small values of observed data and decrease when the value increases. While more positive residuals occur with large values of observed data and increase when the value increases. The pattern indicates these models tend to estimate smaller counts when the counts are large and larger counts when the counts are small, which implies an underestimation of the variations in the counts. The reason can be that the relations between some of the considered features and the counts are not linear or important predictors are missing (Everitt and Skrondal, 2010). It can also be that the performance is constrained by the flexibility of the model, in which case models with higher flexibility would perform better, such as regression trees.

However, except for more outliers in the residual plots, the poor performance of the models of camera 4 and 4R is not explained. Therefore, the pattern of the data itself has to be analyzed separately from camera 1. By looking into the autocorrelation of the two data sets shown in Figure 6.8.1 and 6.8.2, it is observed that the autocorrelation of the data itself is not as good as that of camera 1, which is shown in Figure 4.3.1. The autocorrelation of the nearest lag is only around or even smaller than 0.8. With more than 6 lags (30min), the autocorrelation decreases to below 0.8. The counts from camera 4R are less correlated to historical counts compared to camera 4. Models developed with data set 4R also has smaller $R^2$ than those with data set 4. The poorer autocorrelation makes short-term historical counts less significant for the prediction, which partly accounts for the poor performance of models, especially ARIMA and ARIMAX models.

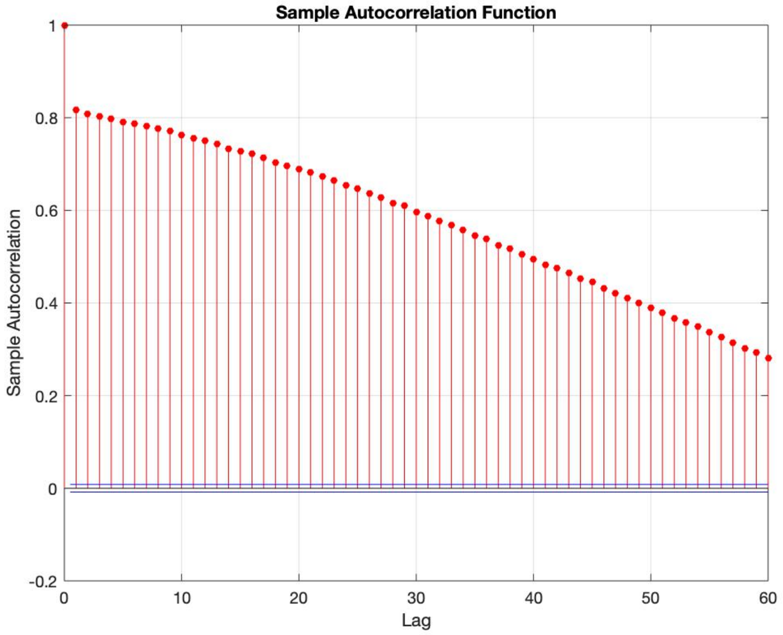


*Figure 6.8.1 ACF of data set camera 4*

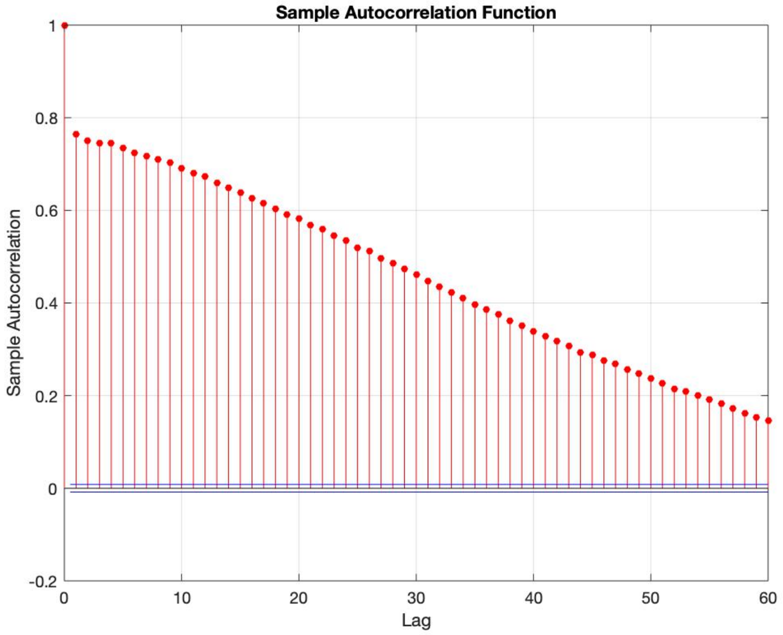


*Figure 6.8.2 ACF of data set camera 4R*

# 7 Discussion

In this chapter, the findings of this research will be discussed and recommendations regarding the limitations of findings will be made.

Firstly, the better predictions by ARIMA and ARIMAX models with fewer features compared to MLR models indicate that the external features can be captured within the changes of time series itself. With only 3 nearest lags, the prediction accuracy can be as good as or even better than MLR models with an exhaustive feature set. However, the limitation of ARIMA and ARIMAX models are obvious. When counts from a few lags behind are taken into account, the model performance decreases by 9%. To predict more accurate counts with ARIMA and ARIMAX models, the nearest lags have to be available, limiting the prediction horizon to 5 minutes, the time of which is too short to prepare effective crowd management approaches. To make the models more practicable, aggregation of counts in 30-minute intervals may provide good predictions in a broader prediction horizon. On the other hand, the external features added by ARIMAX models do not improve the model performance as much as expected but tend to overfit the data, which indicates the trends of the data can be mostly captured by the last lag.

Secondly, the poor performances of both MLR and ARIMA models at camera 4 and 4R indicate different counts pattern at this location. As analyzed in chapter 6.4, the residuals of models developed from data set 4 and 4R present more outliers and the autocorrelation of the time series is not as good as that at other locations. The reason may be that the functionality of this location is different than others. As described in chapter 3.1.1, no tourist attractions are located on the street of camera 4 and the street does not connect tourist attractions and public transport stations. The street provides more residential functions, such as local hotels, galleries, and clinics, rather than attractions to tourists. Therefore, the behavior pattern of visitors is different from other locations, which may be the reason that the models with the same variable set as other locations developed from data of camera 4 and 4R do not perform as well as in other locations. It is also observed that the number of people recorded by camera 4 and 4R is about 1/4 of the counts at other locations, which corresponds to the assumption of different functions at location 4. In addition, compared to other streets, the street of camera 4 is wider, indicating a higher possibility of people wandering around rather than walking in the same direction, which makes the original counts data less autocorrelated and less reliable. What's more, at camera 4R, the performance of ARIMA and ARIMAX models are poorer than interactions MLR models by 0.02 of $R^2$, indicating that interactions between features may capture the structural trends in data set 4R slightly better than ARIMA and ARIMAX models. To make better predictions at location 4, non-linear models may help capture the relations between features and counts.

Thirdly, both types of models tend to underestimate the variations in the data, which means the data structure is not adequately captured by the model. It might be because of missing parameters or fitting non-linear data to linear models. In the case of missing parameters, to make the model more accurate, the trends with more features should be looked into. For instance, only the Dutch holidays are selected from holidays and no big events are included in the features. However, Amsterdam attracts tourists from all over the world and the events, such as concerts and football matches in Amsterdam also influence the number of pedestrians around the red light district. By looking into trends with international holidays and local events, the model might fit the data better.

# 8 Conclusion

In this chapter, the conclusions are made by answering research questions and summarizing the limitations of this research.

## 8.1 Research questions

The answers to subquestions are as follows.

1. What features influence tourist counts and to what extent?

From the literature review in chapter 2, it is found that historical counts, time-related, weather-related, spatial, and social features have an influence on tourist counts. The spatial features are difficult to quantify and therefore 14 models are built to distinguish the spatial differences of time series. Among other features, as analyzed through feature selection in chapter 4, the short-term historical counts are the most important, observed from the parameter values and p-values, followed by the hour-of-day, day-of-week, holiday, and seasonality. The weather-related features have the least influence on the counts. The nearer the historical counts are, the more important are they for the prediction. During the development of ARIMA models in chapter 6, only the hour-of-day, weekends, and rain duration and precipitation are important to the prediction, while other features are captured by AR and MA in 3 lags.

2. How to train and validate ARIMA and multiple linear regression models?

The MLR models are validated by 5-fold cross-validation, by which the data set is randomly divided into 5 folds and the training and validation are done 5 times, each time with 4 folds for training and 1 fold for validation. The ARIMA models are trained and validated for 5 times as well. However, to keep the time series in chronological order, the data is separated also in chronological order, every 3 weeks for training, and 1 week for validation. The separation is done randomly for 5 times, which does not guarantee all the data in the set can be validated.

3. How to evaluate and compare the performances of different models?

The performances of different models are evaluated and compared by RMSE and $R^2$. RMSE measures the differences between values predicted by a model and the observed values. $R^2$ represents the proportion of the variance for a dependent variable that's explained by variables in a regression model and can be used to compare the results with other researches.

4. What are the performances of the developed models?

In general, the ARIMAX models with 3 lags of AR and MA and 27 external features, provide the most accurate prediction, followed by ARIMA models with 3 lags of AR and MA, interactions MLR models with 40 features, robust MLR models with 40 features, and MLR models with 40 features. However, at camera 4R, interactions MLR models outperform ARIMAX models, indicating interactions between features capture the structural trends in data set 4R better. At camera 7 and 7R, interactions MLR models perform as good as

ARIMAX models, indicating the ARIMAX models are more sensitive to missing data. However, the performances of ARIMAX models only improve by around 0.05 of $R^2$, when adding 27 external features, which leads to overfitting of the data. Therefore, ARIMA models with the nearest lag of AR and MA already capture the trends in the data. Considering limiting the number of features included, ARIMA models outperform other models. It is important to notice that though ARIMA models make an accurate prediction based on the nearest lags, the performance becomes worse when lags from 30 minutes ago are used for prediction. On the prediction horizon of 30 minutes, the interactions MLR models have the highest $R^2$ values.

The main research question is answered after answering the above subquestions.

Do ARIMA models provide a more accurate prediction of tourist counts for a 30-minute prediction horizon compared to multiple linear regression models in the area around Amsterdam red light district?

For the prediction horizon of 30 minutes, ARIMA models do not predict as good as multiple linear regression models. However, when the nearest historical counts (from 5 minutes ago) are considered, ARIMA models perform better than multiple linear regression models, which indicates that in the area around Amsterdam red light district, the pedestrian counts are highly correlated with the observations of 5 minutes ago, but less correlated with the earlier observations. Among MLR models, interactions linear regression models perform the best, with 40 parameters including average counts from 30 minutes to 1 hour, 23 out of 24 hours of day, 6 out of 7 days of a week, a seasonal influence from November to January, new year holiday, 8 weather-related factors including average wind speed in an hour, maximum wind speed in an hour, temperature, duration of sunshine in an hour, duration of rain in an hour, precipitation of rain in an hour, rain and ice formation. These parameters contribute to the variations of pedestrian numbers in the studied area.

## 8.2 Other findings and limitations

Due to the limited duration of this project, the exploration of the data set has not been done exhaustively. However, listing the work that could not be done can make a reference to further researches into related data sets.

Firstly, the SVR models could be trained with the Regression Learner app in Matlab. Whereas the training of SVR models requires more computational efforts, the models with 40 features could not be developed. It is observed from several tryouts that when the feature number is reduced to 20 and the principal components analysis (PCA) is on, the models could be trained within a reasonable time. However, by excluding too many features, the model performance is relatively poor. For further researches, to train SVR models with limited computational efforts, the set of features included in the model can be set smaller initially. By PCA, the insignificant features can be selected and removed from the set, and the rest features can be added and analyzed in the next iteration until only significant features are included in the model.

Secondly, the sensitivity analysis was not done during ARIMA and ARIMAX model development, which could provide a reference for the evaluation of how the parameters influence the model output. Instead, the lags are chosen based on the findings from MLR models and the ARIMA model tests with data set 1. It might be that in this research, the chosen number of lags do not capture the data structure accurately, as important lags are not completely taken into account. In addition, the ARIMA model tests were done with data set 1 only. The trends and important features might be different at each location, especially at location4, where the functionality is different than other locations. For future researches, the sensitivity analysis should be done to select a significant number of lags and external features.

Thirdly, the missing data was not removed during the training of models to maintain the same number of records used for model development and the chronological order of time series. However, including 0 counts in the training and testing set leads to the poor performance of the models. For MLR models, removing the missing data might help improve the models. For ARIMA and ARIMAX models, though the counts are missing, other features are still available. What could be possible is the imputation of the missing data, based on trends discovered from the earlier time series. However, the methods to tackle missing data do not guarantee good results. To develop reliable models, the data set should always be as complete as possible.

Fourthly, in the research, $R^2$ was used for model comparison instead of adjusted $R^2$, which may lead to biased evaluation. Due to the mathematical nature of $R^2$, its value increases with the increase of the number of independent variables, even when they are not correlated to the output. When comparing models with a different number of independent variables, adjusted $R^2$ modifies $R^2$ according to the number of predictors in the models to avoid the improvement brought by adding meaningless predictors (Shieh, 2008). However, in this research, the number of predictors is different for ARIMA, ARIMAX, and MLR models. The comparison by $R^2$ may therefore be biased. For further research, the changes in the number of predictors should be taken into account by evaluating models with adjusted $R^2$.

Finally, further researches can look into non-linear prediction models and analyzing more external features, as both MLR and ARIMA models tend to underfit the data in this research.

# 9 Reference

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). Time series analysis: forecasting and control (Vol. 734). John Wiley & Sons.

DEL, TSS, USF, & AIZ. (2016, July 29). Deliverable 4.1 Exploring Prediction Perspectives, Retrieved July 19, 2020, from http://staffwww.dcs.shef.ac.uk/people/L.Moffatt/Seta_Deliverables/D4.1_Final.pdf

Duives, D. C., Wang, G., & Kim, J. (2019). Forecasting pedestrian movements using recurrent neural networks: An application of crowd monitoring data. Sensors, 19(2), 382.

Everitt, B., & Skrondal, A. (2010). Standardized mortality rate (SMR). The Cambridge Dictionary of Statistics, 409.

Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.

Girzadas, A. (2020). Adapting and employing smart city sensor data for strategic planning (Bachelor's thesis, Radboud University).

Hess, P. M., Vernez Moudon, A., Catherine Snyder, M., & Stanilov, K. (1999). Site design and pedestrian travel. Transportation research record, 1674(1), 9-19.

Hinton, G. E., Sejnowski, T. J., & Poggio, T. A. (Eds.). (1999). Unsupervised learning: foundations of neural computation. MIT press.

Hoogendoorn, S. P., & Daamen, W. (2005). Pedestrian behavior at bottlenecks. Transportation science, 39(2), 147-159.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.

Moghimi, B., Safikhani, A., Kamga, C., & Hao, W. (2018). Cycle-length prediction in actuated traffic-signal control using ARIMA model. Journal of Computing in Civil Engineering, 32(2), 04017083.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

Nikovski, D., Nishiuma, N., Goto, Y., & Kumazawa, H. (2005, September). Univariate short-term prediction of road travel times. In Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005. (pp. 1074-1079). IEEE.

Ohler, F., Krempels, K. H., & Möbus, S. (2017, April). Forecasting Public Transportation Capacity Utilisation Considering External Factors. In VEHITS (pp. 300-311).

Papadimitriou, E., Yannis, G., & Golias, J. (2009). A critical assessment of pedestrian behaviour models. Transportation research part F: traffic psychology and behaviour, 12(3), 242-255.

Roosmalen, J. J. (2019). Forecasting bus ridership with trip planner usage data: a machine learning application (Master's thesis, University of Twente).

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Saneinejad, S., Roorda, M. J., & Kennedy, C. (2012). Modelling the impact of weather conditions on active transportation travel behaviour. Transportation research part D: transport and environment, 17(2), 129-137.

Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine, 4(2), 24-38.

Shekhar, S., & Williams, B. M. (2007). Adaptive seasonal time series models for forecasting short-term traffic flow. Transportation Research Record, 2024(1), 116-125.

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. Organizational Research Methods, 11(2), 387-407.

Siddiquee, M. S. A., & Hoque, S. (2017). Predicting the daily traffic volume from hourly traffic data using artificial neural network. Neural Network World, 27(3), 283.

Smith, B. L., Williams, B. M., & Oswald, R. K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C: Emerging Technologies, 10(4), 303-321.

Wang, H., & Hu, D. (2005, October). Comparison of SVM and LS-SVM for regression. In 2005 International Conference on Neural Networks and Brain (Vol. 1, pp. 279-283). IEEE.

Wang, X., Liono, J., McIntosh, W., & Salim, F. D. (2017, November). Predicting the city foot traffic with pedestrian sensor data. In Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (pp. 1-10).

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose.

Xu, Y., Kong, Q. J., & Liu, Y. (2013, June). Short-term traffic volume prediction using classification and regression trees. In 2013 IEEE Intelligent Vehicles Symposium (IV) (pp. 493-498). IEEE.