Analysis and prediction of bus dwell times and bus demand in case of special events

Luca Furlanetto

Abstract—Along with the increment of population in the urban area, the demand of public transport increased in the recent years causing more frequent overcrowding situations; special events have been demonstrated to be one of the possible cause of critical situation. Te demand occurring on six venues in Copenhagen on 2017 has been analyzed and the results showed a positively correlation between the occurrence of events and bus demand. Furthermore, it has found out that often situations of abnormal demand occurred in correspondence of events. The same analysis has been carried out considering bus dwell time, the results obtained showed a lower correlation with events occurrence respect. Moreover, short term prediction models have been developed to forecast both bus demand and bus dwell time. Then, the models have tested on real case and the results demonstrated that, in many cases, including event data in regression model can improve the quality of the forecast.

Index Terms—Public transport demand prediction, Dwell time prediction, Impact of Special event on bus service, Gaussian Process, Bayesian models, Linear regression.

1 Introduction

Along with the increment of population in the urban area, the demand of public transport increased in the recent years causing more frequent overcrowding situations. Also, one of the factor influencing busses travel time is the dwell time and unexpected variations may cause delays. Special events can be the cause of particularly high public transport demand and longer dwell times.

This paper analyzes real data occurred in six venues in Copenhagen on 2017 to find correlation with the occurrence of event. Furthermore, the events has been categorized by its typology, in order to analyze if different events have different impacts.

It is assumed that the Public Transport provider can be interested on understanding the main factors of very high level of demand and, for this reason, an analysis on abnormal demand and dwell time is carried out.

Since, it has been showed that an adequate public transport demand forecast can reduce operational cost and to increase the service quality. Three short term regression models for predicting the number of passengers alighting and boarding have been developed and tested on real case. As well, three regression models for predicting dwell time and two for predicting bus stopping are developer and tested on real case.

2 Literature Review

2.1 Public transport demand

Along with the increment of population in the urban area, the demand of public transport increased in the recent years causing more frequent overcrowding situations [1]. As demonstrated by A. Tirachini et al. [2] when the bus service is not able to satisfy all the demand, crowding busses have a negative impact on travel time and passengers satisfaction, with the consequence of increasing the marginal cost of travelling. An inadequate demand forecast can be a possible reason of overcrowded vehicles while a reliable one is expected to reduce operational cost and to increase the service quality [1].

Public transport demand is usually predicted using parametric or non parametric models. Parametric models are a common and effective approach to the problem and many works involves the application of auto-regressive integrated moving average (ARIMA) models also known as Box Jenkins models [3] [4]. In other case, the auto-regressive model is implemented with a Holt Winters procedure in order to capture seasonality patterns [5]. With an increasing trend in the recent years, non-parametric approaches have been widely used for demand prediction. The forecast can be based, for example, on Support vector machine models [6] or Kalman filters to implement two auto-regressive models [7] or neural networks [8] [9]. Generally, both parametric and non-parametric models tend to work well, however non-parametric models are much more flexible and, in some cases, outperform the parametric ones, capturing complex pattern in the data.

In the literature, one of the main problem that has been recognized is the abnormal variations of demand that an event can cause, it has been recognized as a challenge to forecast [10]. Special events are known to attract a large number of people in a small area, this often leads to a high concentration of traffic flows and to a greater demand for public transport, causing traffic congestion and transit overcrowding [11] [12] [13]. For what regards mega-events such as FIFA world cup or Olympics, many works have been done such as Managing Travel for Planned Special Events [12] that offers a series of recommendations and policies to

• *This paper has been written with the Department of Management Engineering, Technical University of Denmark, DK for acquiring a Master of Transport and Logistics.*
E-mail: furlanettoluca92@gmail.com
Student ID: s161890
Supervisor: Francisco Camara Pereira

management and monitoring mobility for planned special mega-events. However, not only mega-event can cause an abnormal demand but also concerts, sport events or exhibitions can attract a number of people that may exceed the public transport service capacity [14]. Despite the potential impact of event participants on public transport, special events are not always considered in transport demand analysis. Few works consider directly the occurrence of events. One example is C. Zhou et al. [10] that state that one of the major challenges for predicting passengers demand for bus service is seasonal bursty, in other words, the unexpected demand causes by planned events. The study deals with this problem only making short time prediction. On the other hand, F. C. Pereira et al. [15] presented a Bayesian additive model, with Gaussian process components, that considers a base routine component and an event component that are summed to obtain the total demand in case of event. The main point of this research is demonstrating that using data about the occurrence of events provides substantially better forecast of demand in the area where the events would take place.

2.2 Dwell time prediction

One of the factor influencing busses travel time is the dwell time and consequently unexpected variations may cause delays. For this reason, many researches have focused the attention on dwell times prediction. Most of the developed regression models consider the number of passengers boarding or alighting as a major factor of dwell times both using parametric models [16] [17] [18] and non-parametric models such as neural networks [19] [20]. Other researches have demonstrated that also payment method [21], bus doors size and doors number [18] have relevant impact on dwell times. Interesting, the research *Estimation of bus dwell time using univariate time series models* by S. Rashidi et al. [22] has compared four models and showed that ARIMA models was performing better than exponential smoothing, moving average or random walk models. Noteworthy, R. Rajbhandari et al. [23] propose two linear and two non linear models for dwell times prediction that consider the number of passengers alighting and boarding, the number of standees and the time of the day. The research shows that not only the dwell time is correlated to the number of passengers boarding and alighting, as other research did, but that the number of passengers boarding has an higher impact on the dwell time and that the relation between them is non linear.

2.3 Topic modelling

Today, it can be found online a massive amount of information in form of text from which it can be extrapolated various types of information. This process is known as text mining and it can be applied in various fields such as: software engineering, political science, medical and linguistic science [24]. Topic modeling algorithms are statistical methods that analyze the words in a text to discover the themes that run through it [25]. The most common model for topic modelling is called Latent Dirichlet allocation. Given a set of text documents, it represents each document as a linear combination of a small number of topics [26]. Topic modeling has been also applied to forecast transport arrivals in case of special event [15]. In that case, textual information of event has been processed to assign to each event a topic assuming that similar type of events would share similar characteristics. The results show that the quality of the models' output increases taking into consideration this information.

3 Data

In this chapter, an overview and a general description of the data utilized are given. All the data analyzed have been recorded in Copenhagen (Denmark) and can be divided in four groups: bus service demand, dwell times, events and weathers data.

In order to focus the attention on the impact of special events on bus service, only observations occurred in location close to the main venues of the city have been taken into consideration. Those venues have been selected based on: events frequency and venue capacity. As result, the following six hot spots are identified. Royal arena, located in Ørestad, is known for hosting concerts of popular artists. Koncerthuset is located in København S, it hosts mainly classic music concerts. Vega is located in Vestebro and it is focused on underground and alternative music concerts. Bella center and Forum Copenhagen mainly host conferences and exhibitions of various genre, the first one is located in Ørestad and the second in Frederiksberg. Lastly, Telia Parken is the stadium of the main football team of Copenhagen and it is seldom used for concerts or other kind of events, it is located in Østebro. The yellow pins in Figure 1 shows the location of each venue.

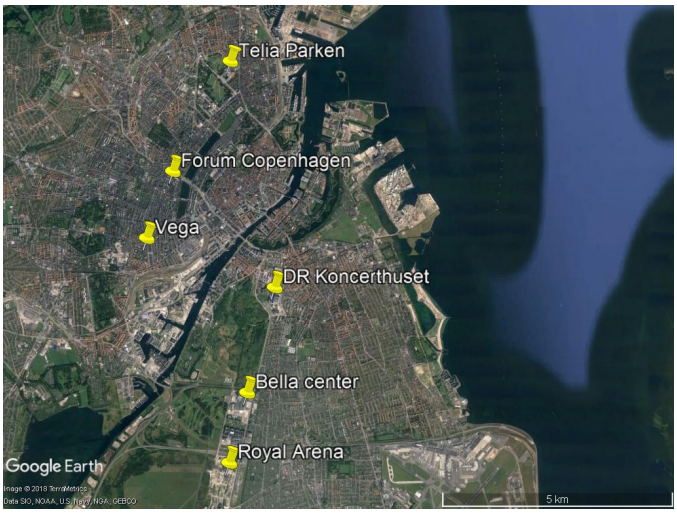

Figure 1: Location of venues

For each venues the closest stops are considered since it is assumed that the event participants want to alight as close as possible to the venue's entrance. It follows a list that indicates the selected stops for each venue.

When dwell time is analyzed, any stop closed to Royal arena is considered because the closest stop is a timing point, which mean that busses stop longer in case they are ahead of schedule; those kind of stops are not considered in the analysis. The other closest stops are quite distant from

Table 1: Stops and venues

Stops identification number	Venue
27999	Bella Center
860	DR Koncerthuset
678	Forum
1365	Parken
30941	Royal Arena
1586	Vega

the entrance and are served by the same bus line so it is reasonable thinking that only the closest is the one used to reach the venue by event participants.

3.1 Bus service demand

One of the main data set utilized regards bus service demand, the observations have been collected by the Public Transport Authority of the area (Movia) for one year starting from the first of January 2017 to 31st December 2017. It contains ticketing data, specifically Smart Card (Rejsekort) check-ins and check-outs. The check-in observations indicate the number of passengers boarding and the check-out the number of passengers alighting. Thus, passengers using other tickets are not counted (e.g. season pass, single-ride tickets, etc.) so this data represent only a sample of the reality. Given the unavailability of public data, specific information about the usage of Rejsekort has been required to the public transport operator. The information received shows that the trip made using this system in 2017 are 26% of the total journeys. It is assumed that people that travel to attend events, are not more or less likely to use different systems from Rejsekort than usual, which means that, this rate is assumed to be equal along all the observations. The total number of passengers alighting and boarding analyzed are respectively 336,589 and 342,148. The data report the time window in which the observations are aggregated, the number of passengers alighting or boarding and the stop where the observation has been recorded. An example of alighting passenger data is shown in Table 2.

Table 2: Example of alighting passengers data

Time Steps	Stop number	Alighting passenger
01-01-17 10:00	678	5
01-01-17 10:30	678	2
01-01-17 11:00	678	1
01-01-17 11:30	678	28

3.2 Dwell times

Similarly, observations of dwell times have been collected from the first of January to 31th of December 2017 by the public transport operator. As dwell time is intended the difference between arrival time and departure time at one stop. The number of observations recorded is 738,813 and each observation represents a single dwell time .The data indicates: arrival time, departure time at each stop, dwell time, the stop identification number and if the stop is the first or the last one of the journey or if it is a timing point. Timing points are used for longer stop in case the bus is running ahead of schedule. In this analysis are not considered stops that are the first or the last of a journey or timing point since the dwell time in these stops can be impacted by factors that are not part of this study. Notice that, if the bus passed a stop point without stopping, the dwell time has been calculated equal to 0 second. An example of the data is given in Table 3.

Table 3: Example of dwell time data

Arrival time	Departure time	Stop number	First Stop	Last Stop	Timing Point	Dwell seconds
26-01-17 16:28:35	26-01-17 16:28:47	678	0	0	0	12
26-01-17 16:28:35	26-01-17 16:28:35	30942	0	0	1	0
26-01-17 16:28:36	26-01-17 16:29:06	1365	0	0	0	30
26-01-17 16:28:37	26-01-17 16:28:37	1586	0	0	0	0

It has been noticed that 48,060 observations was affected by recording error which means that, the dwell time was missing and therefore those have been removed from the database. Similarly, it has been remove 83,266 because recorded in timing point stop.

3.3 Event

Event information have been utilized in order to predict possible traffic disruption caused by special events. In practise, this information is used to identify the observations of dwell time or number of passengers occurred in proximity of an event such as: before the starting time, after the end or during an event. The data have been collected from three websites: allevents.in, timeout.com and eventful.com. The data report the location, the starting time, ending time and a brief description of each event. The total number of events in the data set is 467 divided as: Koncerthuset 110, Forum Copenhagen 31, Vega 192, Royal Arena 51, Bella center 33, Telia Parken 50. Figure 2 shows the number of events per venue.

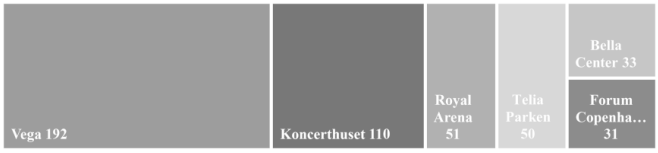

Figure 2: Number of event per Venue

3.3.1 *Event topic*

Each time an observation is identified as in proximity to an event, it is also specified the topic of that event. The topics have been defined based on the description of each event in two ways: first using Latent Dirichlet allocation (LDA) technique and, then, semi-manually.

LDA is a generative probabilistic model that, given a corpus, finds semantics and significant intra-document statistical structures. As a result, each document in the corpus is assigned a distribution over the topics which are in form of distributions over words. The topics are indented to represents the concept contained in each document, it can be used as summary, to categorize or in prediction models [26]. The result obtained from the application of LDA method on the events descriptions is shown in Table 4.

The reason why it has been chosen 8 topics each containing 8 words is that, after different combinations have been tried, this was the one that best explain the corpus without loosing information or containing useless information. Each

Table 4: LDA topic description

Topic 0	Topic 1	Topic 2	Topic 3
0.127* rock	0.110* tour	0.171* weekend	0.110* udsolgt
0.034* finger	0.045* forum	0.156* fc	0.101* venteliste
0.034* five	0.031* sting	0.134* store	0.035* kalkbrenner
0.034* death	0.031* europe	0.038* friend	0.035* ekstrakoncert
0.034* punch	0.026* genetics	0.038* klub	0.027* grellier
0.034* weekend	0.026* eshg	0.036* dj	0.027* koncertsalen
0.028* center	0.026* floyd	0.023* match	0.022* college
0.028* conference	0.026* marcus	0.023* typhoon	0.022* phlake
Topic 4	**Topic 5**	**Topic 6**	**Topic 7**
0.035* mew	0.053* metallica	0.061* live	0.041* europa
0.030* series	0.043* soundsystem	0.037* kid	0.041* league
0.030* pro	0.043* lcd	0.037* festival	0.031* gun
0.030* blast	0.035* concert	0.033* aerosmith	0.028* rose
0.025* live	0.029* sean	0.033* suspekt	0.023* place
0.025* herbie	0.025* paul	0.029* dillon	0.021* scorpion
0.025* hancock	0.025* show	0.029* dizzy	0.018* quickly
0.020* fair	0.022* jamie	0.025* music	0.018* mahler

topic indicates a set of words and a set of numbers. The numbers state how much a word is relevant to assign to that topic an event that contains that word in description. One description can contain words of different topics and so, belonging to different topics. For this reason, the final output is a matrix that state how much every event belongs to each topic. However, the words of each topic should describe an event typology and should express a general idea but the ones obtained do not have these characteristics. The words in each topic seem to not have any relationship to each others and they do not express any real topic. This can be due to the fact that the description text is not suitable for this process. Often the description is too specific or lacks of information. For example, an event described as "Tiny house - Abent hus, Forum København" took place in Forum Copenhagen on 1st of November 2017. This event was an exposition focused on tiny houses. On 2nd of November 2017, an event described as "Building green" took place in the same location. Both of them shared the same theme but they do not share any common words in description. Because of that, the model was un-capable to consider them in the same group. Another example is given, two events described as "Sådan får du penge til dit projekt som ung entreprenør" and "Konference om Faglig ledelse for offentlige ledere 2017" took place in Bella Center. The first event was focused on entrepreneurship and the second on management and public speaking. Even if they shared a similar topic and they can attract similar participants, the two descriptions do not share any common words. Additionally, the description are written both in English and in Danish. Those are the main reasons why this approach did not produce a useful result. After that, a semi-manual topic modeling has been performed. In this case, each event is defined with one or two words based on its description. Music related event are described based on music genre, this process has been done automatically using Gracenote Music API, an Application Programming Interface that gives the music genre given an artist. The list of topics created is: Education, Electronic Music, Alternative Punk Music, Business, Children, Classic Music, Pop Music, Culture, Design, Entertainment, Football, Outdoor festival, Independent Music, Lifestyle, Medicine, Metal Music, Music, Religious Music, Rock Music, Sport, Stage Musicals Music, Traditional Music, Urban Music. In this way, all the created topics well explain the general theme of each event. Figure 3 shows the number of event per topic.

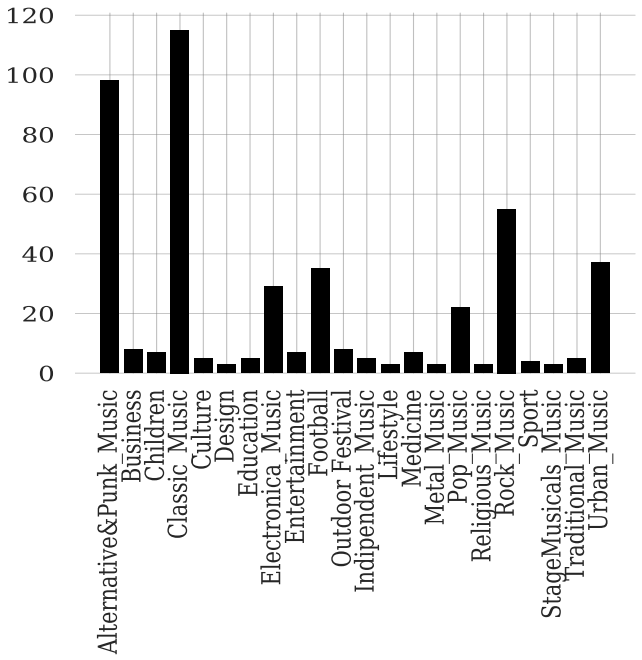

Figure 3: Number of event per Topic

3.4 Weather

The last database indicates the quantity of precipitation in millimeters and the temperature in Celsius degrees for every day in 2017. It reports one observation per day, so, the temperature is the mean of all the day and the precipitation is the sum of the day. The station where these observations are recorded is located in Viberup, south of Copenhagen. All the venues are located within an area of 15 km from the station; this distance is considered enough small to assume that the weather observed in the station is the same in all the venues.

4 Explanatory analysis Bus service demand

This section aims to provide a description of main characteristics of three main data sets.

4.1 General statistics

First at all, the observations of alighting and boarding passengers are analyzed from a general point of view. Mean and the most relevant quantiles are reported in Table 5.

Table 5: Basic information

	Alighting	**Boarding**
Mean	3.4	3.9
25th Quantile	1	1
50th Quantile	2	2
75th Quantile	4	4
90th Quantile	8	10
99th Quantile	21	23

The average of both the data sets is between 3 and 4 passenger per 30 minutes. This values might appear low

but it has to be remembered that these observations are a sample of the real demand which is almost 4 times higher.

Figure 4 and Figure 5 show the distributions of the observation of passengers alighting and boarding respectively. In both the graph it is highlighted the distribution of the observation occurred in proximity of event. In this section, the observation of passengers alighting occurred in proximity of event are studied and compared with the ones occurred in normal days. As proximity to an event is considered the time window starting from 3 hours before until 2 hours after the starting time. About, boarding passengers, proximity to an event it is considered 3 hours after the event ending time, however, if an event is particularly long (more than 6 hours) also the two hours before the ending time are included.

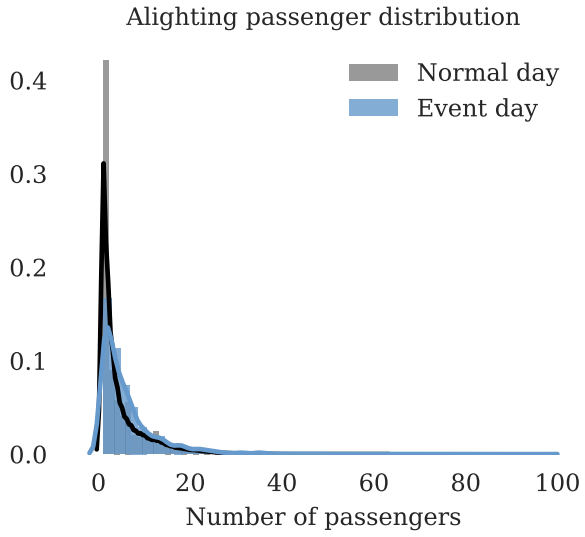

Figure 4: Distribution of the observation of alighting passengers

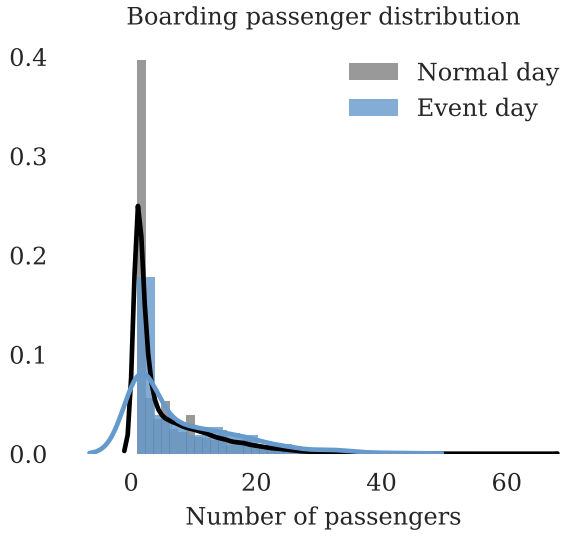

Figure 5: Distribution of the observation of boarding passengers

In both the cases, the distribution of the observation recorded in proximity of events takes more time higher value. Both distributions have a long tail which mean that there are few observations that report a very high demand. In the next sections, those very high observation will be analyzed and along with the main factor influencing the demand.

4.2 Daily pattern

In order to investigate on the presence of a seasonal pattern over the day, the average numbers of passengers during the day have been calculated and the results are shown in Figure 6.

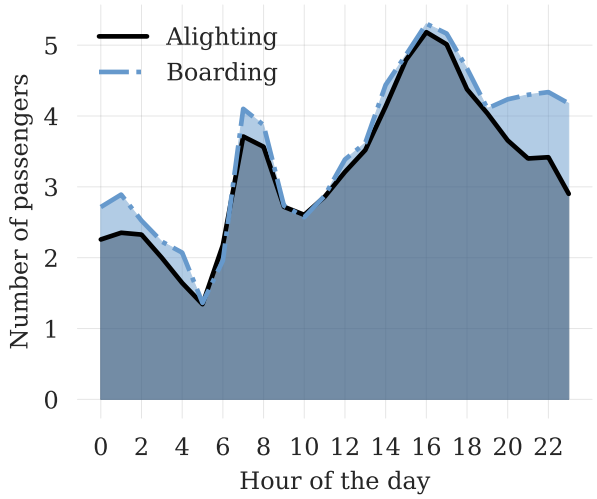

Figure 6: Average demand per hour

Both the passengers alighting and boarding curves have two peaks corresponding to the morning and the afternoon rush hour and have their minimum during the night.

4.3 Example on event day

It follows an example of the demand during an event day to show that events can cause an abnormal demand increment. The venue considered is Forum Copenhagen and the day is 7th of January. On this day, it took place an event called "Loppemarked" ("Garage sale") which is an exhibition of antiques and retro objects from all over the world. The event started at 11am and ended at 17pm. Figure 7 and 8 show the trend of, respectively, passenger alighting and boarding on stop 678, which is located closed to Forum Copenhagen. The blue dotted line represents the observations on that day and the black line represents the average number of passengers alighting calculated on all the weekends of January. The two dotted vertical lines represent the starting and ending time of the event.

In both the case,the curves share a similar trend in the morning but considerably diverge after the start stay above the average, until a couple of hours after the end of the event. The reason why of this substantial difference from the mean is considered to be the event.

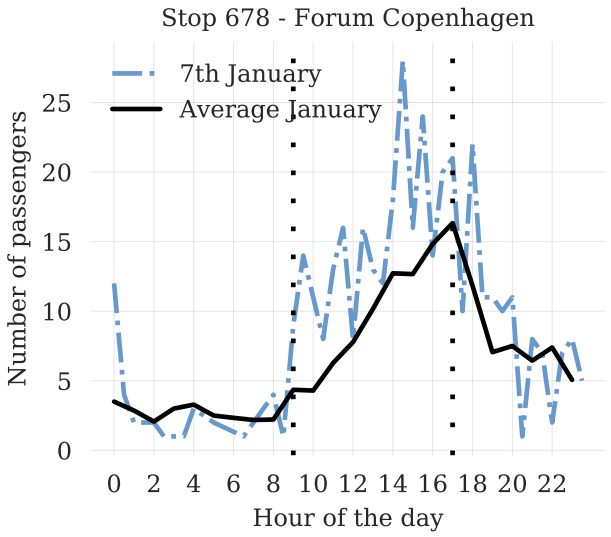

Figure 7: Comparison between the passengers alighting on an event day with the average demand of the same month

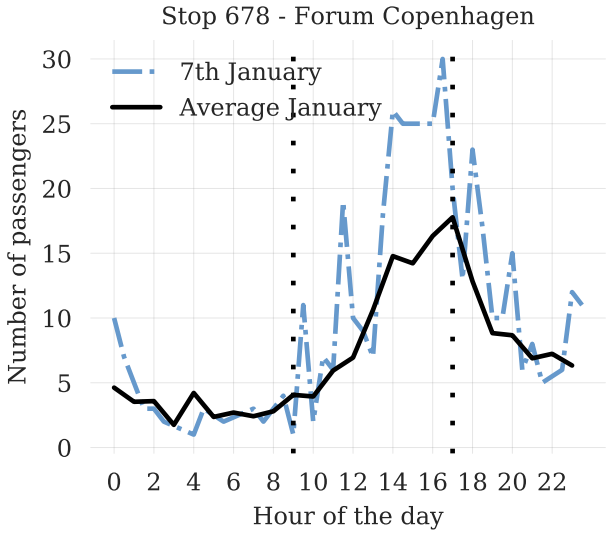

Figure 8: Comparison between the passengers boarding on an event day with the average demand of the same month

4.4 Correlation analysis - Passengers alighting

This section aims to provide an accurate description of the correlation between passengers alighting and all the factors that are considered potentially influencing. Those factors are also called input features and are listed below.

- Time of the day: the time when the observation occurred. It is divided in time frame of 4 hours.
- Time series: the previous observations in that stop. It indicates the observation occurred at the same stop in the previous time step. These features are named *lag* plus a number that indicates the number of intervals between the observations. For example, lag_1 indicates the previous observation (30 minutes before), lag_2 the one before the previous (one hour before) and so on.
- Event proximity: it indicates if an observation has been recorded in proximity of the starting time or the ending time of an event. In particular, the time interval taken into consideration is from 3 hours before the starting time to 3 hours after the ending time.
- Event topic: it indicates the topic of the event.
- Weather: quantity of precipitation in millimeters and the temperature in Celsius degrees for every day.

To calculate the correlation, the following formula has been used:

$$Correlation = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y})^2}}$$

Basically, it represents the covariance of the two variables normalized by the product of their standard deviations. It expresses how strong is the relationship between the two variables. For example, a correlation equal to 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.

Table 6 shows the correlation values between passengers alighting and all the input features.

Table 6: Correlation values of Alighting passengers

Input feature	Correlation value
Alighting Passenger_lag1	0.71
Alighting Passenger_lag2	0.65
Alighting Passenger_lag3	0.58
Alighting Passenger_lag4	0.53
Alighting Passenger_lag5	0.48
Hours 16-20	0.14
1h before event starting time	0.06
Is event	0.06
2h before event starting time	0.06
Hours 12-16	0.04
3h before event starting time	0.03
Precipitation	0.02
1h after event starting time	0
2h after event starting time	-0.01
Is weekend	-0.01
Temperature	-0.02
3h after event starting time	-0.02
Hours 20-24	-0.03
Hours 0-4	-0.07
Hours 4-8	-0.07
Hours 8-12	-0.08

It appears clear that the time series features are the ones that have the highest correlation values meaning that the previous observation can inform about the next one. Then, afternoon hours have a positive correlation while night and morning hours have a negative correlation. This partly confirm the daily trend shown before, it seems that the morning peaks, that should lead to a positive correlation, is not perceived, probably, because of the division in time frame. The hours before an event have a positive correlation and the ones after the start, have a negative correlation or null. In conclusion, weathers data are poorly correlated with alighting passengers.

4.4.1 *Event proximity correlation*

Since, event proximity represents one of the main feature; a further analysis is carried out to analyze the impact of the

event on the demand in each stop. The results are showed in Figure 9.

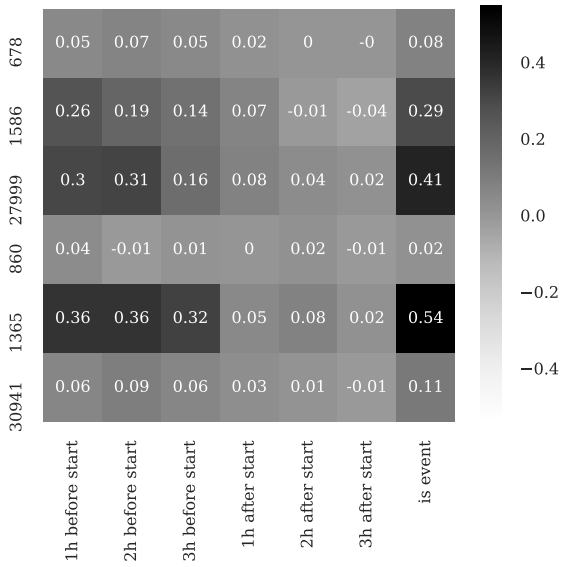

Figure 9: Correlation between alighting passengers and event proximity for each stop

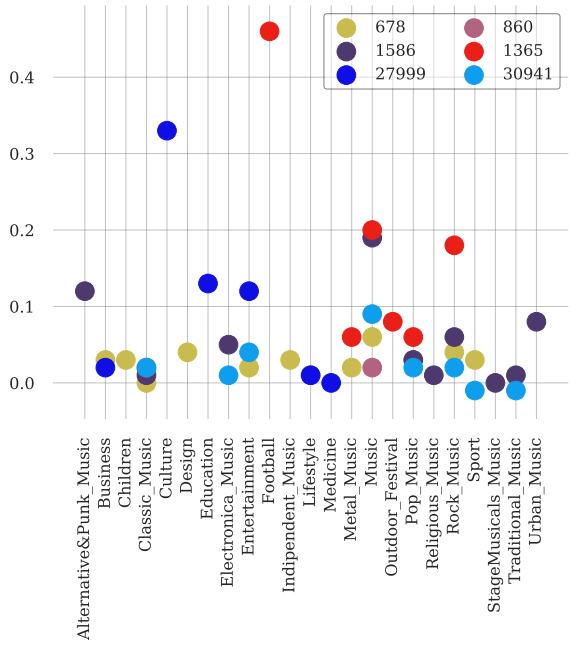

Figure 10: Correlation between alighting passengers and event topic

It appears clear that some stops are more correlated with the events than others, those are the number 1365, 27999 and 1586. The first is closed to Telia Parken, the second to Bella Center and the third to Vega. This means that the impact of event is not the same everywhere but that it depends on the venue. On the other side, the fact that an event is starting in 3 hour or less, is the most impactful feature which means that the people tend to arrive at the venue before the starting time and not after. The feature called *is event* whenever an observation is recorded between 3 hours before and 3 hours after the starting time, in some sense, it includes all the six other features. It is the highest correlated feature, which means that the participants not always arrive at the same time. For some stops such as 860, 30941 (Royal Arena), 678 (Forum Copenhagen) events seem to have any impact. This can be due to the fact that participants alight in different stops or that busses are not commonly taken to reach that venue.

4.4.2 *Event topic correlation*

As already mentioned, also the topic of the event have been considered as an input feature and since that, its correlation is analyzed. Figure 10 shows the correlation of each topic for each stop with the number of passenger alighting.

The correlation between event topics and passengers alighting is generally quite low, below 0.1. However, in same other cases the correlation is higher with some specific topics. For example, the observations recorded in stop 1365, which is located near Telia Parken, have a stronger correlation with football related event rather than concert related event. Similarly, the observation recorded at stops 27999 (Bella center) are more correlated with Culture related event rather than others. This is a clear example of the fact that different event types impact differently on the demand.

4.4.3 *Auto-correlation*

Since time series are the highest correlated features, a further analysis is made. In this case, instead of be limited to five time series, the lags considered are 336, one week. This is known as Auto-correlation and aims to express the degree of similarity of the observations and to identify repeating patterns. Figure 11 shows the Auto-correlation curve.

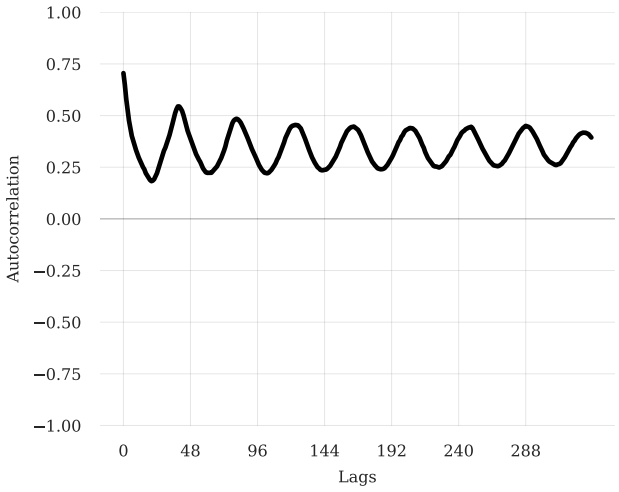

Figure 11: Alighting passengers auto-correlation

Before it has been showed that the correlation was lower further was the observation considered, e.g. correlation with lag_1>lag_2>lag_3>lag_4. However, what the auto-correlation curve shows is that an observation is more correlated with the one occurred the day before rather than the one occurred 12 hours before. Additionally, the results show

a seasonal pattern repeating every day. It can be noticed that the peaks are not repeating every 48lags as it would be reasonable thinking. This is due to the lack of observation when the demand is 0. This implies that, in particular at night, some time steps are missing causing, on average, one observation is distant 40 lags from the one of 24 hours before instead of 48.

Summarizing, the feature that is most correlated with passengers alighting is the time series. Not only the number of passenger alighting is highly correlated with the ones observed in the previous 30 minutes but also the observations recorded 24 hours before circa result to have a high correlation. Then, the observation recorded at stops number 1365, 1586,27999 results to be higher correlated with events compared with the average. In addition, correlation is higher when the 3 hours before the event starting time are considered. Moreover, it has been noticed particularly correlation between the demand and the occurrence of certain type of event. For example football related events are more correlated with the observation recorded in stop 1409 rather than music related event. As well, the observations recorded at stops 27999 are more correlated with Culture related event rather than others.

4.5 Correlation analysis - Passengers Boarding

Similarly, an analysis of the correlation between passengers boarding and the input features are carried out. Table 7 shows that correlation value between the input features.

Table 7: Correlation values of Boarding Passengers

Input feature	Correlation value
Boarding Passenger_lag1	0.73
Boarding Passenger_lag2	0.69
Boarding Passenger_lag3	0.63
Boarding Passenger_lag4	0.58
Boarding Passenger_lag5	0.53
Hours 16-20	0.11
Is event	0.04
Hours 20-24	0.03
5h after event starting time	0.03
6h after event starting time	0.03
Hours 12-16	0.02
1h after event ending time	0.02
4h after event starting time	0.01
Precipitation	0.01
3h after event ending time	0.01
2h after event ending time	0.01
Is weekend	-0.01
Temperature	-0.02
Hours 0-4	-0.05
Hours 4-8	-0.06
Hours 8-12	-0.09

Time series features are the highest correlated features meaning that the previous observations can inform about the next one. Regarding the time of the day, the correlation is positive after midday and negative before midday. This partly confirm the daily trend shown before, it seems that the morning peaks, that should lead to a positive correlation, is not perceived, probably, because of the division in time frame. Then, event information is positively correlated even if the correlation is not high. Finally, the correlations of weathers data are closed to zero.

4.5.1 *Event proximity correlation*

As before, the correlation between the boarding passengers and the event information has been computed for each stop. This information is shown in Figure 12.

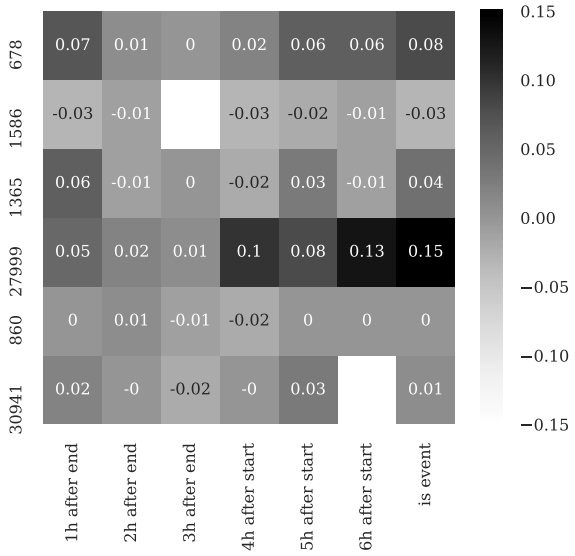

Figure 12: Correlation between boarding passengers and event proximity for each stop

What the heat map shows is that boarding passengers are low correlated with the occurrence of an event except for stop 27999 (Bella center). This means that, people usually do not take the bus after the end of an event. It can be noticed that some data are missing, this means that any observation was recorded in that stop in correspondence of that parameter.

4.5.2 *Event topic correlation*

The correlation has been computed also for each event type in order to check further correlation between event and passengers boarding. The results are shown is Figure 13.

While the correlation between demand in term of passenger boarding and event topic is always low, it can be noticed a quite high correlation on stop 27999 (Bella center) with culture and entertainment related event.

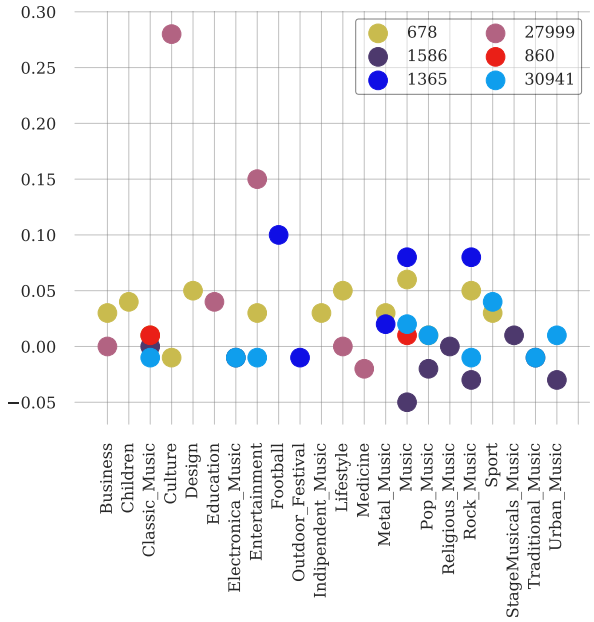

Figure 13: Correlation between boarding passengers and event topic

4.5.3 *Auto-correlation*

Also in this case, time series are the most correlated features. For this reason, it has been computed an auto-correlation analysis. Figure 14 shows the boarding passengers auto-correlation curve.

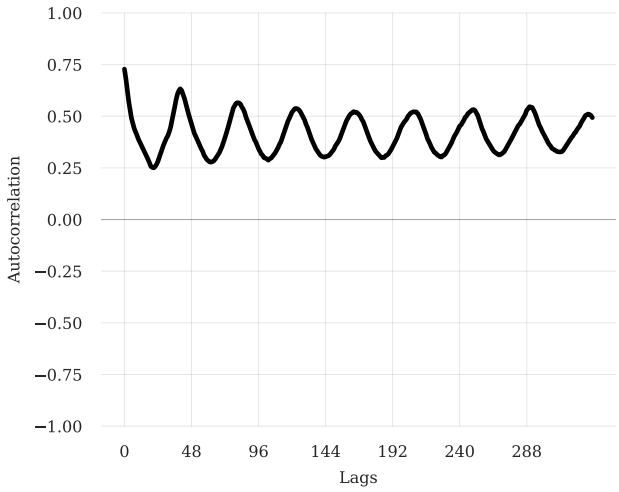

Figure 14: Boarding passengers auto-correlation

Auto-correlation of boarding passengers has the same characteristics of the alighting one. The observations are high correlated with the one occurred at the same time in the previous day.

Summarizing, the feature that is most correlated with passengers boarding is the time series. Not only the number of passenger alighting is highly correlated with the ones observed in the previous 30 minutes, but also the observation recorded 24 hours before, results to have a high correlation. Then, only the observation recorded at stop number 1409 one hour before the events, results being higher correlated with the occurrence of event compared with the average. Moreover, the correlation is higher at stop 1409 when the event are outdoor festival or football related while the demand at stop 27999 is higher correlated with culture and entertainment related events.

4.6 Abnormal observation and Event

It is assumed that the public transport authority of Copenhagen can be interested on understand the reason behind abnormal demand. The main goal of this section is discovering if events have caused abnormal demand in the venues taken into consideration on 2017. Since, it does not exist a formal definition of abnormal demand it has been considered from two points of view. First, it is considered as abnormal a value of demand over the 98th or 99th percentile, in other words, the highest values of the data set. Secondly, it is considered as abnormal an observation that is one and half times or twice the average calculated in the same day and in the same time of the observation (an observation occurred on Monday at midday is compared with the average computed on all the observations occurred on Monday at midday). In other words, is abnormal a situation when the demand is a way higher than the average.

4.6.1 *Passengers alighting*

The goal of this section is analyzing the highest observation of demand and calculating how many of them occurred in proximity of events. As proximity to an event is considered the time window starting from 3 hours before until 2 hours after the starting time.

First at all, the observations occurred in proximity of events are 8.8% of the total. But, if only the 2% highest observations are considered(observation over 17 passengers which are 189) the rate of event observations is 19,7%. If only the 1% highest observations are considered(observation over 21 passengers which are 108) the rate of event observation is 25,5%.

What this means is that the highest values of demand registered are occurred in proximity of events. Since the demand can depend on the venue, the same analysis has been done for each venue and the result are shown in Table 8.

Table 8: Analysis on how many observation of the highest observation are occurred in proximity of events

Venue	Event obs.	98th q.	Over 98th q.	99th q.	Over 99th q.
Forum	2%	26	4%	39	4%
Vega	15%	10	56%	13	59%
Bella Center	3%	10	35%	14	48%
Koncerthuset	9%	4	12%	5	11%
Parken	8%	15	74%	20	86%
Royal Arena	5%	4	17%	4	17%
All the data set	9%	17	19%	21	25%

What Table 8 says is that often when the demand is high, an event is taking place in the venue nearby. In some cases, such as: Vega, Bella center and Telia Parken, most of the highest observations occurred in proximity of events. Considering the venue Forum Copenhagen, the rate of observations recorded in proximity of events, in the entire data

set, is almost the same of the one calculated on the highest values. This means that an abnormal situation is not more likely to happen in proximity of events. For Royal Arena and Koncerthuset, both the 98th and 99th quantiles are pretty low meaning that, the demand in these two locations is quite stable and it is not often subject to abnormal situations.

After that, it has been considered as abnormal observation when the demand was twice or 50% higher than the average. The average has been computed for weekdays, Saturday and Sunday and it has been divided on time windows in order to compare each observation with the average calculated on similar observations. An example of what is intended as abnormal situation is given in Figure 15. It shows the demand occurring at Telia Parken on 10th of September. On this day the football match between FC Copenhagen and Midtjylland started at 16pm. The graph shows four line: the black one represents the demand on that day, the blue one the average demand, the green one the average increased of 50% and the red one the average increased of 100%.

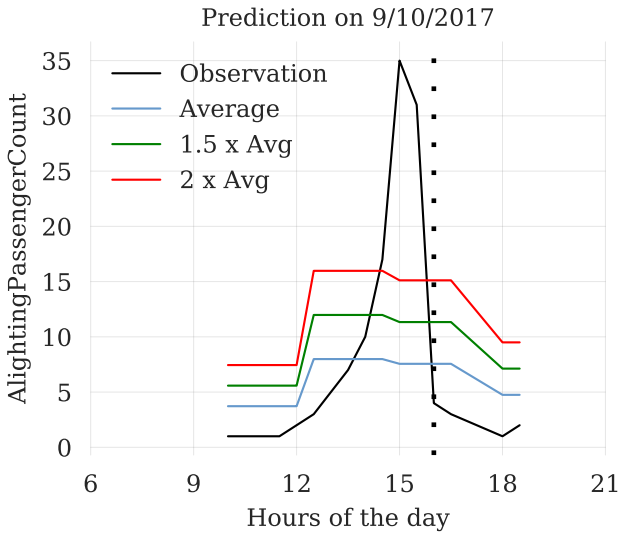

Figure 15: Example of abnormal demand

It can be seen that, right after the start of the event, the demand rose over all the three lines, reaching a value that is almost five time higher than the average.

Another example in given in Figure 16, it regards the venue Vega and reports the observation occurred on 2nd of October. On this date, a concert of the band Megadeth started at 20pm.

It can be seen that the demand was closed to the average for, almost, all the day but in correspondence with the event starting time, it was more than three times higher than the average. So, it has been calculated for every venues how many situations like those occurred in proximity of events. Table 16 shows, first, the rate of observation occurred in proximity of events on 2017, then, how many observations over 50% of the average occurred on events proximity and, finally, how many observations twice the average occurred on events proximity.

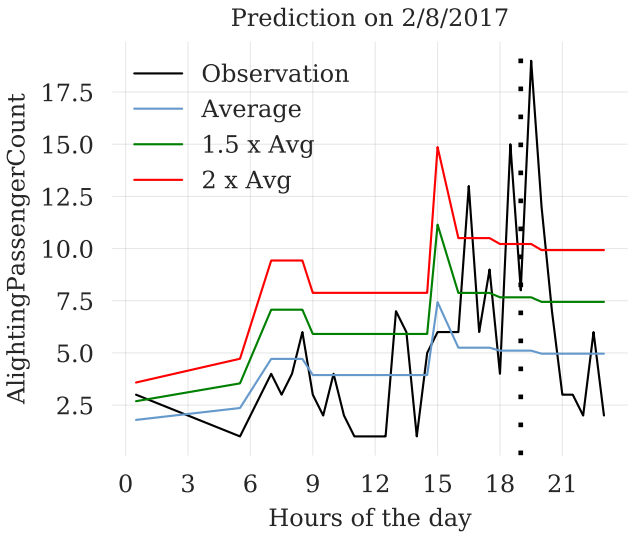

Figure 16: Example of abnormal demand

Table 9: Rate of alighting passengers observation over the average per stop in case of event

Venue	Event observation	Over 50%	Over 100%
Forum	2%	2%	2%
Vega	15%	40%	47%
Bella Center	3%	12%	18%
Koncerthuset	9%	0%	0%
Parken	8%	64%	67%
Royal Arena	5%	0%	0%

Most of the observation that substantially exceed the average in Vega, Bella Center and Telia Parken have occurred in correspondence of events. About Forum Copenhagen, the rate of observations remains stable in all the three cases meaning that an abnormal situation is not more likely to occur during the event day than in a normal day. In case of Koncerthuset and Royal Arena any observation substantially over the average have been registered.

4.6.2 Boarding passengers

The goal of this section is to analyze the highest observation of demand and to calculate how many of them occurred in proximity of events. As proximity to an event, it is considered the 3 hours after the event ending time, however, if an event is particularly long (more than 6 hours) also the two hours before the ending time, are included.

First of all, the observations occurred in proximity of events are 1.8% of the total. But, if only the 2% highest observations are considered (observations over 25 passengers which are 31), this rate increases to 4.6%. If only the 1% highest observations are considered (observations over 28 passengers which are 21), this rate increases to 5.8%.

What this means is that, the highest values of demand registered are slightly more likely to occurred in proximity of events. Since the demand can depend on the venue, the same analysis has been done for each venue and the results are shown in Table 10.

Table 10: Analysis on how many observation of the highest observation are occurred in proximity of events

Venue	Event obs.	98th q.	Over 98th q.	99th q.	Over 99th q.
Forum	2%	28	6%	32	8%
Vega	1%	6	0%	6	0%
Royal Arena	2%	3	4%	3	4%
Bella Center	3%	4	18%	5	28%
Koncerthuset	1%	3	1%	4	4%
Parken	2%	4	0%	5	0%
All the observation	2%	25	5%	28	6%

What Table 10 says is that seldom abnormal situation occurred in correspondence of events. In all the cases except for Vega, the 98th and the 99th quantiles are pretty low and, in these cases, observation exceeding those quantile cannot be considered abnormal. About Vega, the rate of event observations do not change that much in the three cases meaning that the impact of event is not that relevant. It can be noticed that the 98th and the 99th quantiles of the entire data set are way more similar to the ones of Vega than the others, this happen because half of the observations of the data set are recorded in Vega.

After that, it has been considered as abnormal observation when the demand was twice or 50% higher than the average. The average has been computed for weekdays, Saturday and Sunday and it has been divided on time windows in order to compare each observation with the average calculated on similar observations. An example of what is intend as abnormal situation is given in Figure 17. It shows the demand occurring at Vega on the first February. On this day the concert of the band Mew ended at 23pm.

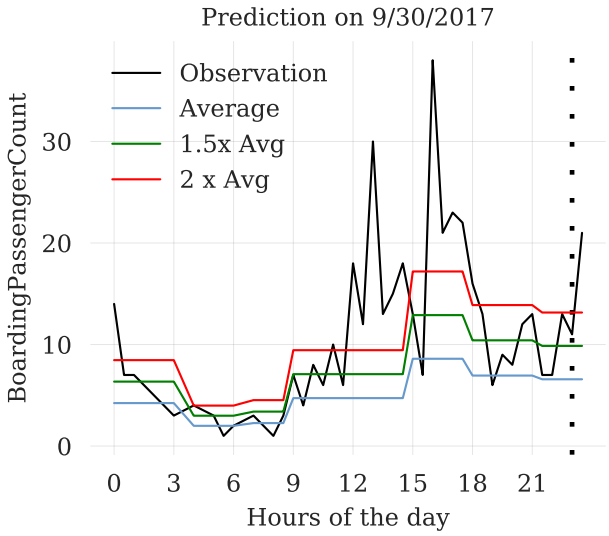

Figure 17: Example of abnormal demand

What appear is that in correspondence with the end of the event the demand increased until reach a level three times higher the average. So, it has been calculated for every venues how many situation like this occurred in proximity of events. Table 9 shows, first, the rate of observations occurred in proximity of events, then, how many observation over 50% of the average occurred on events proximity and, finally, how many observation twice the average occurred on events proximity.

Table 11: Rate of boarding passengers observation over the average per stop in case of event

Venue	Stop	Event observation	50% more	100% more
Forum Copenhagen	678	0.01	0.04	0.04
Vega	1586	0.03	0.00	0.00
Parken	1365	0.00	0.00	0.00
Bella Center	27999	0.01	0.50	0.33
Koncerthuset	860	0.00	0.00	0.00
Royal Arena	30941	0.00	0.00	0.00

From this analysis it can be seen that many of the abnormal demand in terms of boarding passengers observed on Bella center occurred in proximity of events. On the other hand, it has been found out that any of the abnormal observation occurred in the other venues are related with events.

5 Explanatory Analysis Dwell times

5.1 General statistics

In this section the main characteristics of dwell times observation are analyzed. Dwell time is characterized by the fact that the bus can not stopping. For this reason the dwell time observation has a particular distribution which is shown in Figure 18.

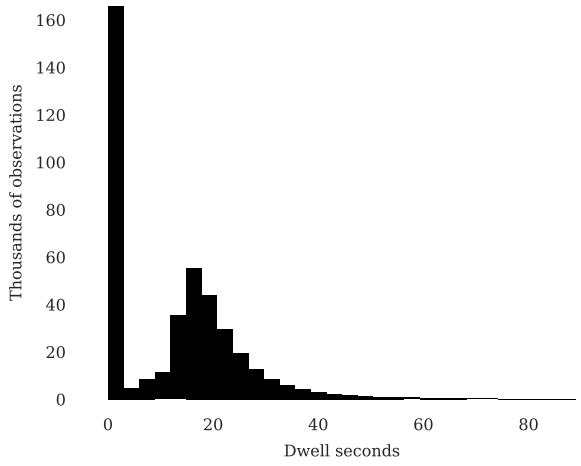

Figure 18: Dwell time distribution

A large quantity of observation (40%) are equal to 0 meaning that very often the bus do not stop at a planned stop. In addition, it can be seen that if the bus stop the dwell time has a certain distribution with mean 21 seconds.

5.2 Daily pattern

Furthermore, the trend that dwell time has over the day is analyzed from two points of view. The first regards the rate of non-stop to answers the question: how often the bus stop during the day. Secondly, the average dwell time in case of stop to answer the question: if the bus stop how long is the dwell time. Figure 19 compares this two trends. The black line represents the Non-stop rate and the dotted blue line the average dwell seconds.

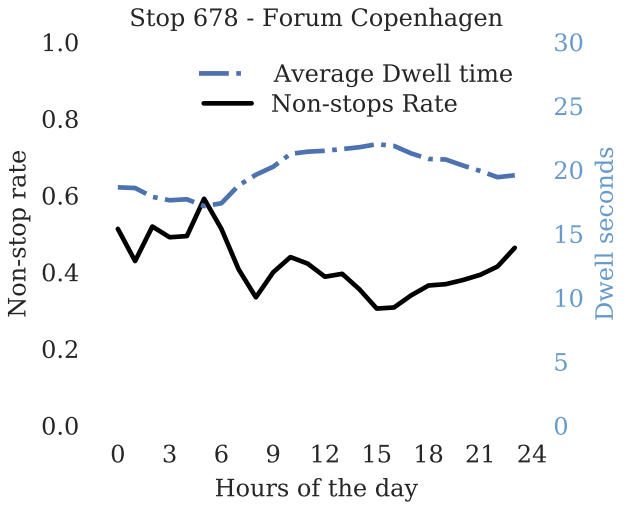

Figure 19: Comparison between average dwell time and non-stop rate over the day

It is clear that the two curves have an opposite trend meaning that when the busses stop more frequently they are also more likely to stop for longer time.

5.3 Example on event day

As it has been done before, an example of the distribution of dwell time over an event day is given. The location and the date is the same as before, 7th of January on stop 678. Figure 20 shows the average dwell time per hour on the event day and the average dwell time per hour occurred on the weekend on January. The two dotted vertical line represent the starting and ending time of the event.

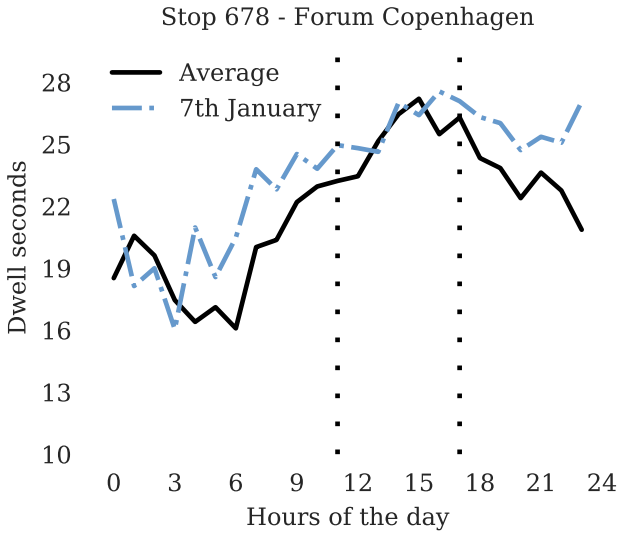

Figure 20: Dwell time during event compared with the average

The graph shows that in proximity of the starting and ending time of the event the dwell time is above the average of about 2 seconds.

A similar analysis can be made for what regards the rate of stopping, assuming that in case of an event the busses stop more frequently. Figure 21 shows the average non-stopping rate per hours during the event date and the average non-stopping rate occurred in January.

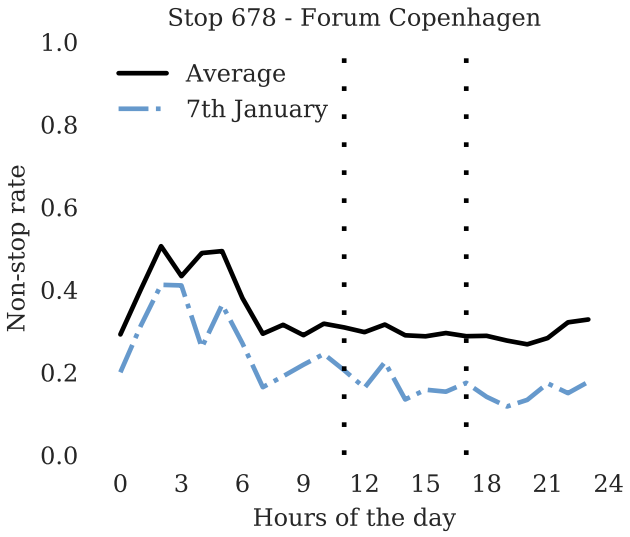

Figure 21: Non-stop rate during event compared with the average

As expected, the busses stop more frequently during the event compared with the month average.

5.4 Correlation analysis

This section aims to provide an accurate description of the correlation between dwell time and all the factors that are considered potentially influencing. Those factors are also called input features and are listed below.

- Time of the day: the time when the observation occurred. It is divided in time frame of 4 hours.
- Time series: the previous observations in that stop. It indicates the observation occurred at the same stop in the previous time step. These features are named *lag* plus a number that indicates the number of intervals between the observations. The time distance between an observation and the previous one is not fixed and range between 8 and 60 minutes depending on the stop and the time of the day.
- Event proximity: it indicates if an observation has been recorded in proximity of the starting time or the ending time of an event. In particular, the time interval taken into consideration is from 3 hours before the starting time to 3 hours after the ending time.
- Event topics: it indicates the topic of the event.
- Weather: quantity of precipitation in millimeters and the temperature in Celsius degrees for every day.

The correlation values have been computed for the dwell time observation. The results are shown in Table 12.

Table 12: Correlation values of Dwell time

Input feature	Correlation value
Dwell Seconds_lag1	0.22
Dwell Seconds_lag2	0.22
Dwell Seconds_lag3	0.21
Dwell Seconds_lag4	0.21
Dwell Seconds_lag5	0.21
Hours 12-16	0.07
Hours 16-20	0.07
Is event	0.05
1h before event starting time	0.04
2h before event starting time	0.03
3h before event starting time	0.03
1h after event starting time	0.02
2h after event starting time	0.01
Temperature	0.01
1h after event ending time	0.01
6h after event starting time	0.01
7h after event starting time	0.01
2h after event ending time	0.01
5h after event starting time	0
4h after event starting time	0
2h after event ending time	0
Precipitation	-0.01
Is weekend	-0.01
Hours 8-12	-0.01
Hours 20-24	-0.02
Hours 0-4	-0.05
Hours 4-8	-0.1

Time series features are the one most correlated with the dwell time meaning that the previous observation can inform about the next one. However, differently from what it has been seen for passengers demand there is not difference between the five lags. Also, afternoon hours are positive correlated while night and morning hours are negatively correlated. Event proximity have a positive impact on the dwell time even if the correlation is pretty low. The impact of the weather on dwell time is almost null.

5.5 Event proximity correlation

A further analysis on the correlation between the event features and the dwell time is done for each stop in order to discover if any stop is higher correlated with events. The results are shown in Figure 22.

The only two stops that result in being affected by the occurrence of an event are number 1365 (Telia Parken). In the first case, the hours before the event are the most relevant ones while in the second case the hours during the event are the most relevant. Moreover, the feature *is event* that indicates that the observation happened within 3 hours before to 3 hours after the event, is in both cases the highest meaning that the participants not always arrive at the same time. For all the other stops the correlation is pretty much similar to the one shown in Table 12 which means quite low.

5.6 Event topic correlation

In order to investigate if different type of event have different correlation, it has been computed the correlations between the dwell time and the event topics. The results are shown in Figure 23

Even considering the typology of the event, most of the values obtained are quite low, below 0.1. But the results shows that dwell time on stops 1365 is more influenced by football event rather than others.

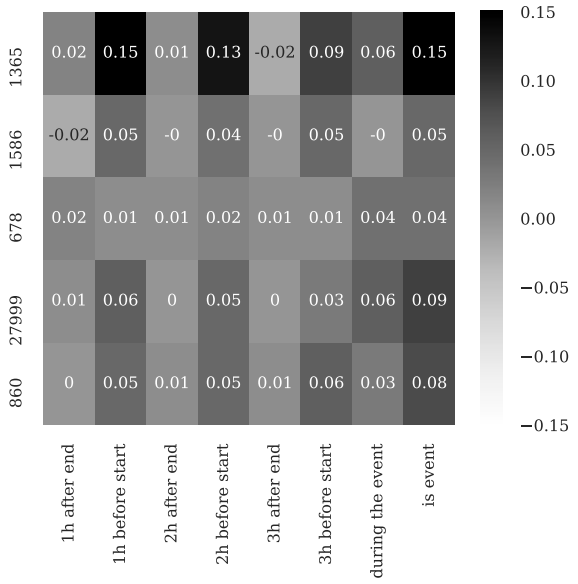

Figure 22: Correlation between dwell time and event proximity for each stop

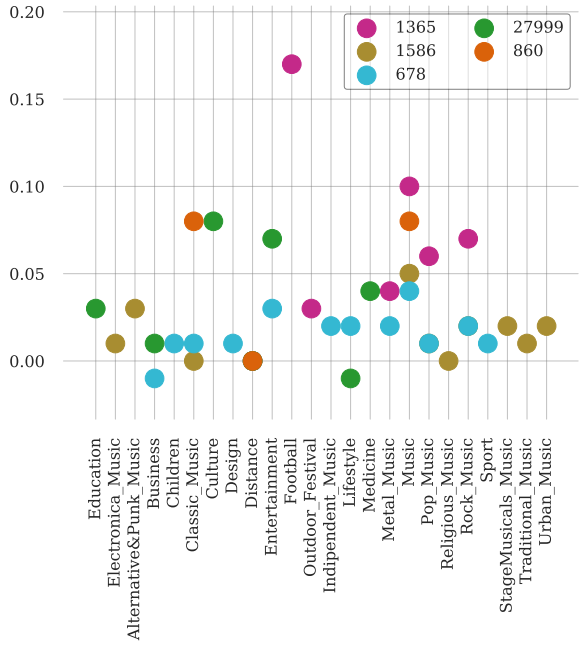

Figure 23: Correlation between dwell time and topic event

5.7 Auto-correlation

As before, an auto-correlation analysis is made since the time series is the highest correlated feature. Figure 24 show the auto-correlation values over 336 lags.

In this case, dwell times do not present any kind of pattern or seasonality and the auto-correlation value is stable around 0.2 for all the lags. This might be because of the different time interval between each observation or because of the randomness of the dwell time.

Summarizing, the features that are most correlated with dwell time is the time series. In this case, the closest lags

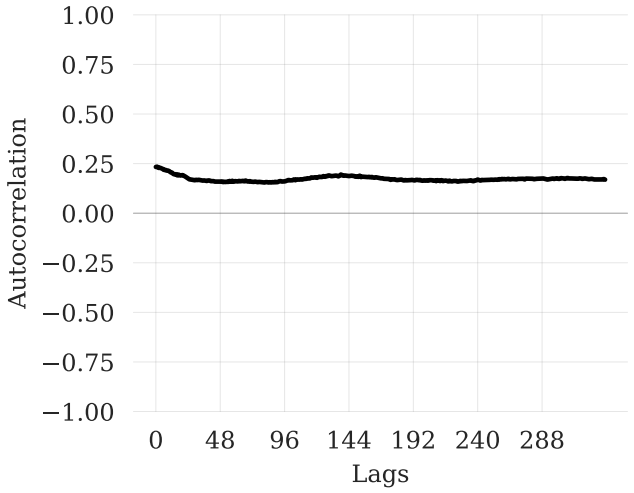

Figure 24: Dwell time auto-correlation

have a slightly higher correlation compared with the others and, differently from before, it is not possible find any kind of seasonality based on the time series. Only the observations recorded at stops number 1365 one or two hours before the events results to be higher correlated with the occurrence of events compared with the average. Regarding the correlation with events, only stops number 1365 shows a correlation over the average.

5.8 Abnormal observation and Event

The goal of this section is analyze the highest observation of dwell time and calculated how many of them occurred in proximity of events. As before, an abnormal situation is intended in two ways. First it is considered as abnormal a dwell time over the 98th or 99th percentile, in other words, the highest values of the data set. Secondly, it is considered as abnormal an observation that is one and half or twice the average calculated in the same day and in the same time of the observation (an observation occurred on Monday at midday is compared with the average computed on the all the observations occurred on Monday at midday). In other words, is abnormal a situation when the demand way higher than the average.

First at all, the observations occurred in proximity of events are 9% of the total. In this case, the number of event observations that report a dwell time over the 98th and the 99th percentile are the same proportion. Since each venue can be different a further analysis is made. Table 13 presents first the ratio of observations occurred in proximity of event, then the 98th quantile and the how many observation occurred in proximity of event are over that threshold and, then, the same for the 99th quantile.

In two cases, Parken and Bella Center, the ratio of event observations substantially increase if only the observation over the 98th and the 99th quantile are considered. This shows that abnormal situation are more likely to occur in case of events. While, in the other cases, the ratio remain stable or even decrease meaning that an abnormal situation is not more likely to occur in case of events.

Table 13: Analysis on how many observation of the highest observation are occurred in proximity of events

Venue	Event obs.	98th q.	Over 98th q.	99th q.	Over 99th q.
Parken	6%	30	35%	37	48%
Vega	17%	35	15%	40	14%
Forum	5%	62	7%	71	8%
Bella Center	8%	34	17%	45	16%
Koncerthuset	11%	32	15%	38	12%
All the observation	9%	47	10%	38	10%

After that, it has been considered as abnormal observation when the demand was twice or 50% higher than the average. The average has been computed for weekdays, Saturday and Sunday and it has been divided on time windows in order to compare each observation with the average calculated on similar observations. An example of what is intend as abnormal situation is given in Figure 25. It shows the dwell time occurring at Telia Parken on 19th of February. On this day the football match between FC Copenhagen and Brøndby took place between 13.30 and 15.30pm. The graph shows four line: the black one represents the dwell time occurred over the day, the blue one the average dwell time, the green one the average increased of 50% and the red one the average increased of 100%.

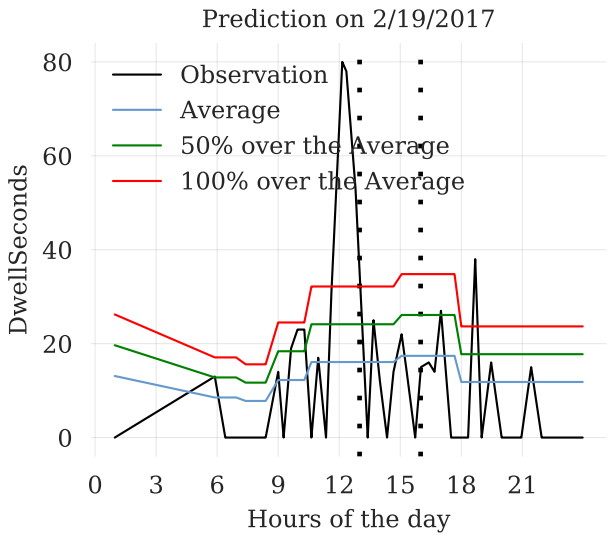

Figure 25: Example of abnormal dwell time

It can be seen that in right before starting time of the match, the bus stopped for longer time, 4 times longer than the average. So, it has been calculated for every venues how many situation like this occurred in proximity of event. Table 14 shows, first, the rate of observation occurred in proximity of events, then, how many observation over 50% of the average occurred on events proximity and, finally, how many observation twice the average occurred on events proximity.

Table 14: Rate of dwell times observation over the average per stop in case of event

Venue	Event observation	50% more	100% more
Parken	6%	12%	22%
Vega	17%	10%	8%
Forum	5%	5%	5%
Bella Center	8%	10%	11%
Koncerthuset	11%	7%	7%

From this analysis it can be seen that only in Parken an abnormal situation is more likely to occurs in correspondence of event rather than on regular days. On all the other cases,

6 PREDICTIVE MODELLING

In this chapter, the prediction models developed are presented and described. Then, the criteria to evaluate the models' performances are presented.

6.1 Models description

Three short term prediction models have been developed: the first one is based on linear regression, the second on Bayesian linear regression, the last is on Gaussian process. Regarding dwell time, the problem has been divided in two part, the first aims to predict whenever the bus is going to stop or not and the second the length of the stop. For the first step, a Support vector machine and a classification Bayesian model are used while for the second part the three models mentioned before are utilized.

To evaluate the goodness of the predictions it would be ideal comparing the results with the current forecast models. However, the bus service provider does not produce any kind of real-time forecast but only yearly based estimation [27]. So, it has been developed an elementary model based to the historical average in order to evaluated the models' performance.

The data sets are divided in two subsets: a training data set and a test data set. The first one is used to fit the model and it consist on all the observation from January to August which means two third of the total observations. The second is used to provide an evaluation of model fit, it consists on all the observations from September to December, one third of the main data set.

During the model developing process, it has been noticed that weather information did not have any impact on the quality of the prediction. For this reason and given the low correlation with the target variables shown in the paragraph Explanatory analysis, these information have not been included in the models.

In conclusion, the input features considered in the models are: Time series, time of the day, typology of event and event proximity.

6.1.1 Historical average

The purpose of this model is creating a baseline to evaluate the goodness of the prediction models. It simply computes the mean of all the observations in the training data set.

6.1.2 Linear regression

Linear regression is a method used to find relationship between an independent variable (input features) and a dependent variable (target variable) and then to make prediction. The form of the model developed is the following:

$$TV = \beta_0 + \beta_1 Lags + \beta_2 H + \beta_3 EP + \beta_4 TE + \beta_5 W + \epsilon$$

Where TV represents the target variable that can be both the passenger alighting, boarding and dwell time. $Lags$ indicates the time series features. H indicates the hour of the day. Then, EP and TE represents respectively the event proximity and the event topic. The weather features is indicated as W. β_0 represent a real number that is commonly defined as intercept and the others β represent the coefficients of the input features. Finally, ϵ represents the error which is assumed to be a random variable normally distribute with mean 0. The method used to calculate the coefficients is the ordinary Least Squares method, it computes the coefficients of the observed variables minimizing the sum of the squared of the residuals. As residual is intended the difference between actual value of the dependent variable and the values predicted by the model. It follows the formula of the residual sum of squares.

$$\sum_{i=1}^{n}(y_i - \beta^T x_i)^2$$

Where y_i indicates the real values, βs the coefficient and x the input variables.

6.1.3 Bayesian linear regression

Before the coefficients of the linear regression have been obtained as a point estimation, in this case the coefficients are retrieved from a probability distributions. In this case, the target variable is assume to normally distribute with mean equal to the product between the input variables and their weights and a certain deviation standard.

$$y_i \sim N(\beta^T X, \sigma I)$$

As results, this approach computes the posterior distribution of the model parameters instead of giving the a single value. The posterior distributions are computed based on Bayes' theorem.

$$P(\beta|y, X) = \frac{P(y|\beta, X) \cdot P(\beta|X)}{P(y|X)}$$

Where, $P(y|\beta, X)$ is the likelihood of the function, it represents the probability of getting the target variable as a function of $beta$ and X. The Maximum likelihood estimation is used to determine the parameters and getting the maximum probability. Then, the prior $P(\beta|X)$ includes the domain knowledge and represents the belief of what the model parameters should be. $P(y|X)$ is the normalization term. Then, the posterior probability distribution, $P(\beta|y, X)$, is the distribution of the coefficients based on the data and the prior. The distribution of the parameters gets updated whenever new data points are observed.

6.1.4 Gaussian Process

Gaussian process is a non-parametric approach that produces a probability distribution over functions. This process is based on Bayesian theorem that can be wrote as

$$p(f|y,X) = \frac{p(f) \cdot p(y|f)}{p(y|X)}$$

Such that, the prior $p(f)$ is represented by the Gaussian process and can be seen as a collection of random function f. Gaussian process is specified by a mean function and a covariance matrix which has been defined using the squared exponential function, it imposes a smoothness to the functions ensuring that the points that are close together in input space have higher covariance. Then, the likelihood constraints the functions to go trough the observed data points. As more observation became available the number of possible functions became smaller. The joint distribution of the posterior is:

$$p(f|y,X) \propto \mathcal{N}(y|f,\sigma^2 I)\mathcal{N}(f|0,K)$$

The Gaussian process posterior represents the probability distribution over all the possible functions that are consistent with the observations. One of the main characteristics that make Gaussian process different from the linear regression is that this approach allows non linear relationship between the dependent and the independent variables.

6.1.5 Support vector machine classifier

Support Vector Machine is an algorithm that has been used as classifier to predict whenever the bus is stopping or not. Support Vector Machine works in a way that plots in a n-dimensional space each observation where n is the number of input features. The classification is performed finding the hyper-plane that best differentiate the two classes.

6.2 Evaluation criteria

To evaluate the regression models the root-mean-square error (RMSE) and the R-squared measures are considered. The first one is calculated as the square of the difference between the predicted values and the real ones and indicates how much the prediction differs from the real values.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y_i})^2}{N}}$$

Where N is the number of observation, $\hat{y_i}$ the prediction and y_i the real values. Then, R-squared indicates how much the variance of the target variable is explained by the prediction.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \mu_y)}$$

Where N is the number of observation, $\hat{y_i}$ the prediction, y_i the real values and μ_y is the mean of the values.

In case the classification models, accuracy and F1 score are used as evaluation criteria. The performance of classification models is commonly evaluated dividing in four group the output: True positive, True negative, False positive and False negative. Where true/false indicates the case in which the prediction is right or wrong and positive/negative indicates the class predicted. Accuracy indicate the rate of True positive and True negative over all the observations. Accuracy states the quantity of observation that have been predicted correctly. To explain F1 score it has to be stated what the Precision and the Recall measurements are. The precision is defined as the number of True Positives divided by the number of True Positives and False Positives. While the Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. Finally, F1 score is the harmonic average between the Precision and the Recall.

$2 * ((precision * recall)/(precision + recall))$

F1 score ranges between 0 and 1 where 1 indicates an optimal prediction.

7 Results

The performance of the models have been tested on event days in order to evaluate their capacity to predict the demand in correspondence of events. The section is divided in four parts: first passengers alighting, second passenger boarding, third bus stopping and lastly dwell time. For each part, all the predictions have been computed twice: first not considering event features and, then, including them. In this way, it is possible understand and evaluate the impact of these feature on the models' output. Notice that, at the end of each part a short summary is provided.

7.1 Passengers Alighting

Firstly, the number of passengers alighting has been forecast considering as inputs the time series features and the time of the day. Table 15 show the errors in terms of passengers of the different models.

Table 15: Errors of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	7.9	5.3	6.3	5.3
Telia Parken	1365	12.0	7.2	8.3	8.2
Vega	1586	5.0	3.7	4.8	3.8
Bella Center	27999	23.1	15.1	19	20.5
Royal Arena	30941	1.3	1.0	1.0	0.8
Koncerthust	860	0.9	0.9	1.0	0.7

Linear regression and Gaussian process models are the ones that perform better. In terms of RMSE, these model outperform the baseline of more than 30% in some cases.

Table 16 reports the R squared values.

Table 16: R squared values of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	0	0.47	0.28	0.49
Telia Parken	1365	0	0.26	0.25	0.26
Vega	1586	0	0.03	0	0
Bella Center	27999	0	0.23	0.23	0.10
Royal Arena	30941	0	0	0	0
Koncerthust	860	0	0	0	0

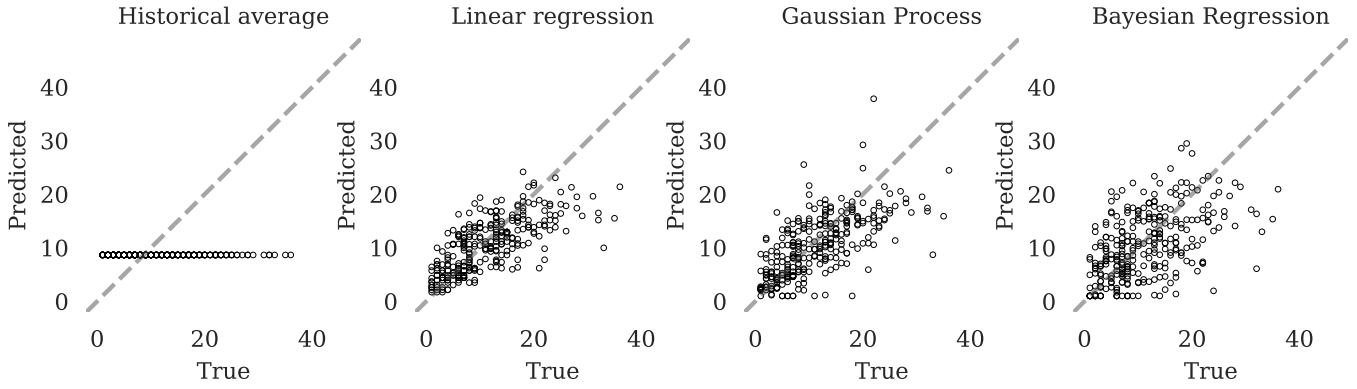

Figure 26: 45 degree line graph to compare predicted value and real observation of alighting passengers at venue Forum

It can been immediately noticed that historical average forecast explains none of the variability of the real data. On the other hand, Linear regression and Gaussian process models are more capable to express the variability of the real observations. Notice that, even if the errors calculated for the venues Royal Arena, Koncerthuset and Vega are pretty low, the R squared values are null for all four models. This means that the demand on average is very low and that the models are not able to forecast the few times that the demand change.

7.1.1 Considering event data

The results shown in this subsection have been obtained from models that include the also event information. The performance in term of RMSE are shown in Table 17 and in term of R squared in Table 18.

Table 17: Errors of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	8.0	5.1	6.5	5.5
Telia Parken	1365	12.0	6.0	7.5	7.1
Vega	1586	5.1	3.2	4.2	3.2
Bella Center	27999	23.1	9.6	19.1	10.4
Royal Arena	30941	1.3	1.0	1.2	1.5
Koncerthust	860	0.9	1.0	1.0	1.0

Table 18: R squared value of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	0	0.50	0.25	0.43
Telia Parken	1365	0	0.43	0.39	0.20
Vega	1586	0	0.16	0	0.15
Bella Center	27999	0	0.33	0.22	0.22
Royal Arena	30941	0	0	0	0
Koncerthust	860	0	0	0	0

Regarding the errors in term of number of passengers, the models' performance have improved considerably. In case of Bella Center, the error of the prediction decreased of 30% circa for the linear regression model and of 50% for Gaussian process model. In term of R squared, considering the event features allows the models to express better the variability of the real observations capturing the variation of demand occurring in proximity of the events. Particularly interesting is the difference of the performance obtained for the venue Vega: in the first case the R squared value was zero for all the model but, in the second case, the it increased substantially. This mean that the demand diverge from the mean in a particular way in proximity of an event and that the models can predict these variations if one of the input indicates that. Also in this case the models perform poorly in case of Royal Arena and Koncerthuset meaning that the demand in these two locations is independent on the occurrence of events.

7.1.2 Performance example

In this subsection, an example of the model performance is given. Firstly, Figure 27 shows the prediction of the three regression models compared with the real values and the historical average. The example chosen refer to venue Forum Copenhagen on the day 2nd of November, on this day, the event named *Building Green* started at 9.00am. This day has been chosen as example because is one of the most representative. Notice that the vertical dotted line indicates the starting time of the event.

It can be noticed that right before the event starting time the number of passengers alighting at the stop increased substantially. The graph shows that the prediction based on Gaussian process and on Linear regression predict partially the peak on time while the Bayesian regression responds too late. Not surprisingly, a prediction based on the historical average do not provide a reliable forecast since the demand varies all over the day.

Another example is given regarding the venue Forum Copenhagen. The prediction is compared with the real observation in a 45 degree line graph, Figure 26. More the points are close to the diagonal line and more precise is the prediction. As seen before, the two models that approach most the diagonal line are Linear regression and Gaussian Process while Bayesian regression tent to diverge more.

7.1.3 Summary

In the previous part it has been shown that the models that are more precise are the ones based on Linear regression and Gaussian Process both in term of RMSE and R squared. Additionally, the performance of the models considerably

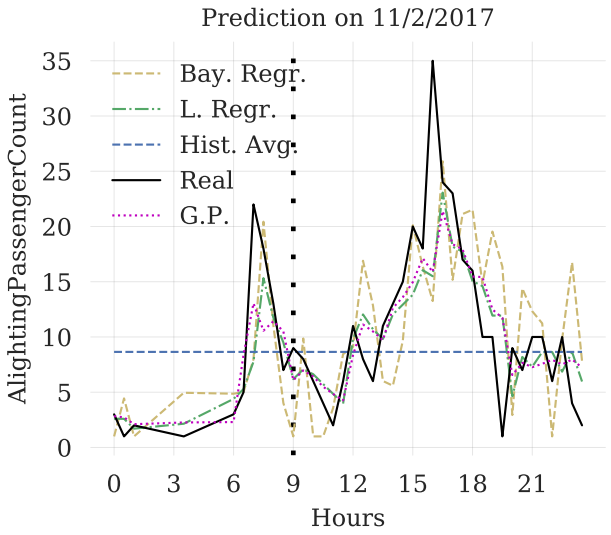

Figure 27: Example of prediction in Forum Copenhagen on 2nd of November

increased if event data are considered. However, not in all the cases considered the forecast obtained can be considered reliable (the case in which R squared is very low). Even if the error in term of number of passenger is quite low, the prediction made for Royal Arena and Koncerthuset obtained a R squared value equal to zero, even considering event information, meaning that the variance of the real observation is not expressed by the forecast.

7.2 Passengers Boarding

The number of passengers boarding has been forecast considering as inputs the time series features and the time of the day. Table 19 shows the errors in terms of passengers of the different models and Table 20 the R squared values.

Table 19: Errors of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	9.0	7.4	8.8	7.4
Telia Parken	1365	0.9	0.9	1.1	0.9
Vega	1586	1.1	1.1	1.7	1.1
Bella Center	27999	2.5	6.3	6.4	5.6
Royal Arena	30941	0.7	0.6	0.9	0.7
Koncerthust	860	0.7	0.7	0.8	0.7

Table 20: R squared values of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	0	0.27	0	0.26
Telia Parken	1365	0	0	0	0.02
Vega	1586	0	0.02	0	0
Bella Center	27999	0	0.49	0.47	0.61
Royal Arena	30941	0	0	0	0
Koncerthust	860	0	0.18	0	0

Noteworthy, the error associated with the historical average model is lower than the ones associated with the others models, except for venue Forum Copenhagen. In case of R squared, the Linear regression model have obtained quite high value in case of the venue: Forum Copenhagen, Bella Center and Koncerthuset. Gaussian Process model obtained quite high R squared values in case of the venues Bella center and Forum Copenhagen while the Bayesian regression model only for Bella Center. In conclusion, the demand in terms of boarding passengers is less dependent on the input factors rather than on alighting passenger. Also, the models developed are less capable to predict the variation of demand causing large errors and, sometimes, they predict a high value when is not happening in reality causing a high errors.

7.2.1 *Considering event data*

Table 21 and Table 22 shows the results obtained taking into consideration event information.

Table 21: Errors of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	9.0	7.4	8.0	8.0
Telia Parken	1365	0.9	0.9	1.1	0.9
Vega	1586	1.1	1.1	1.6	1.1
Bella Center	27999	2.5	6.3	5.4	2.6
Royal Arena	30941	0.7	0.7	1.4	2.4
Koncerthust	860	0.7	0.6	1.0	2.7

Table 22: R squared value of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	0	0.27	0.14	0.13
Telia Parken	1365	0	0	0	0
Vega	1586	0	0.01	0	0
Bella Center	27999	0	0.49	0.45	0.61
Royal Arena	30941	0	0	0	0
Koncerthust	860	0	0.25	0	0

Comparing with the previous case, the RMSE values associated with the first four venues presented in the tables slightly decrease or remain stable while for the last two increased. This means that event information are useful only in few locations. Regarding R squared values, it can be noticed an improvement of the Bayesian regression model at Forum Copenhagen and for Linear regression model at Koncerthuset. This means that part of the variation of the demand occurs in proximity of events and it can be forecast taking into consideration events occurrence. Surprisingly, the R squared value of the Gaussian process model decrease in venue Forum Copenhagen. This can be due to a low correlation between demand and event and by the fact that those new features are "confusing" the model.

7.2.2 *Performance example*

In this subsection, an example of the performances of the models is given in Figure 29. The example regards the same station and the same day as before, 2nd of November in Forum Copenhagen. The vertical dotted line represent the end of the event.

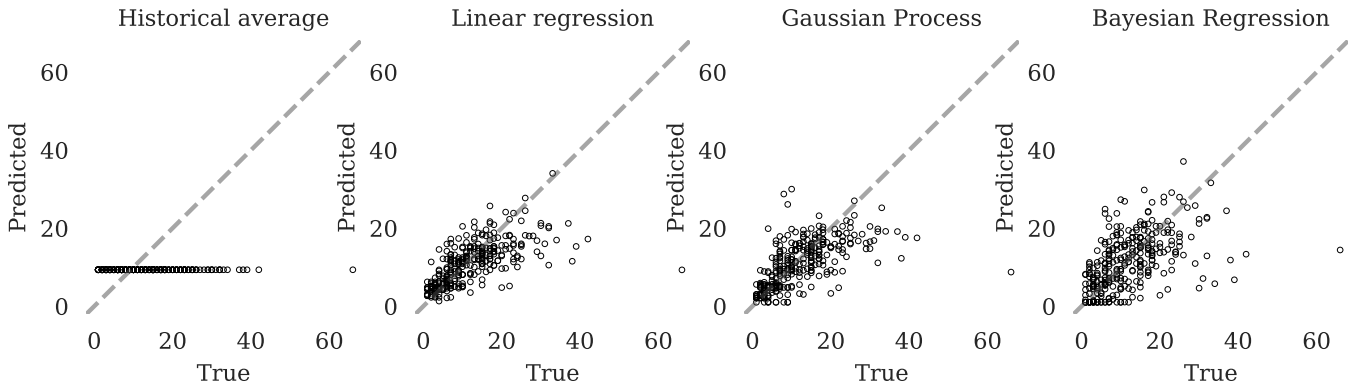

Figure 28: 45 degree line graph to compare predicted value and real observation of boarding passengers at Venue Forum

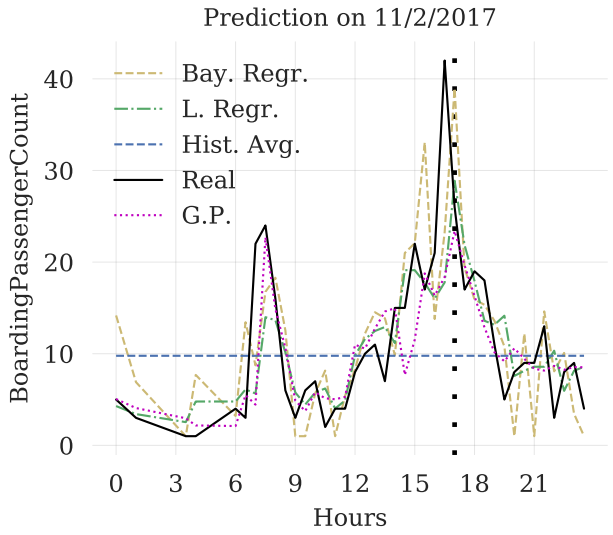

Figure 29: Example of prediction in Forum Copenhagen on 2nd of November

This example has been chosen because one of the most representative. As often happen, the peak in demand occurring at the end of the event is only partially predicted and all the models tent to respond too late at the variations in demand. This is one of the main problem on working with time series, a sudden increase is hardly predicted on time.

Another example is given on Figure 28, it compares the prediction models' output with real observation in a 45 degree line graph and it regards the venue Forum Copenhagen. As seen before, the two models that approach most the diagonal line are Linear regression and Gaussian Process while Bayesian regression tent to diverge more.

7.2.3 Summary

Linear regression and Gaussian Process are the models that are more precise both in term of RMSE and R squared. Additionally, not always the performance of the models increased if event data are considered meaning that the correlation between the demand in term of passenger boarding is low. Even if the error in term of number of passenger is quite low, the prediction made for Royal Arena, Vega and Telia Parken obtained a R squared value equal to zero, even considering event information, meaning that the variance of the real observation is not expressed by the forecast.

7.3 Stopping

This section presents the results of prediction the bus stopping. Table 23, Table 24 report respectively the accuracy and the F1 values of the prediction.

Table 23: Accuracy of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	SVM	Bay. Class.
Forum	678	0.63	0.71	0.72
Telia Parken	1365	0.50	0.60	0.60
Vega	1586	0.64	0.77	0.71
Bella Center	27999	0.48	0.55	0.52
Koncerthust	860	0.55	0.72	0.71

Table 24: F1 values of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	SVM	Bay. Class.
Forum	678	0.78	0.83	0.81
Telia Parken	1365	0.46	0.62	0.64
Vega	1586	0.75	0.87	0.80
Bella Center	27999	0.51	0.51	0.55
Koncerthust	860	0.68	0.83	0.77

In general, both the accuracy and the F1 scores are higher respect to the historical average. The results obtained for the venues Forum Copenhagen, Vega and Koncerthuset are the ones with the highest accuracy and F1 values representing an high dependency with the inputs.

7.3.1 Considering event data

Here, the results obtained considering event information are presented. Table 25 and Table 26 report respectively the accuracy and the F1 values of the predictions.

Table 25: Accuracy of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	SVM	Bay. Class.
Forum	678	0.63	0.71	0.70
Telia Parken	1365	0.50	0.59	0.57
Vega	1586	0.64	0.77	0.77
Bella Center	27999	0.48	0.55	0.45
Koncerthust	860	0.55	0.72	0.72

Table 26: F1 values of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	SVM	Bay. Class.
Forum	678	0.78	0.83	0.82
Telia Parken	1365	0.46	0.61	0.48
Vega	1586	0.75	0.87	0.87
Bella Center	27999	0.51	0.51	0.29
Koncerthust	860	0.68	0.83	0.84

Surprisingly, the scores of the prediction based on support vector machine model do not change. Likely because of the low correlation with events. The scores of the Bayesian model decreased in two cases and increased in one. This means that event information are not always relevant for this kind of prediction.

7.3.2 Example of performance

To investigate on how the models works, all the confusion matrices have been investigated. The confusion matrix give an overview of the performance of the models providing the number of True positive, True negative, False positive and False negative prediction. Where true/false indicates the case in which the prediction is right or wrong and positive/negative indicates the class predicted. It follows the confusion matrix of models for Forum Copenhagen, Figure 30 shows the output from the historical average model, Figure 31 from the Support vector machine models and Figure 32 from the Bayesian model.

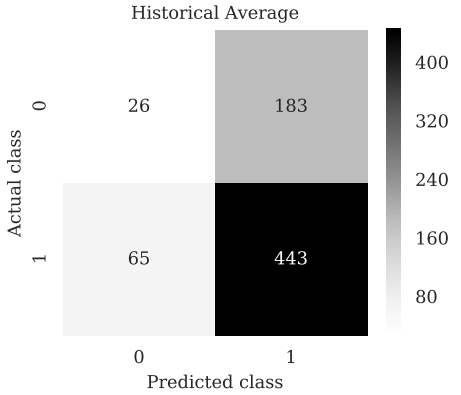

Figure 30: Confusion matrices of Historical average model on Forum Copenhagen

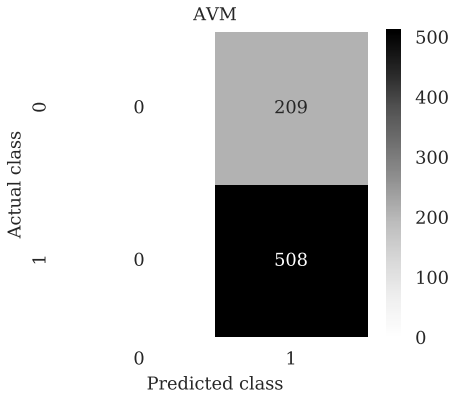

Figure 31: Confusion matrices of Support vector machine model on Forum Copenhagen

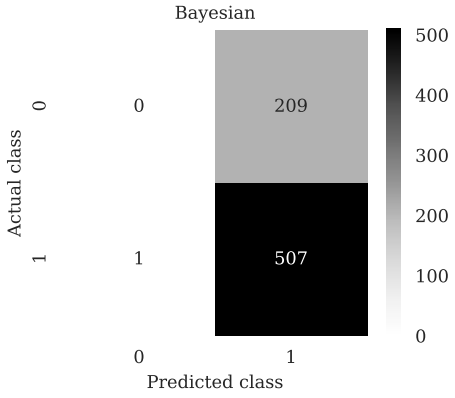

Figure 32: Confusion matrices of Bayesian model on Forum Copenhagen

The reason why these examples have been chosen is because, even if the accuracy and the F1 score are pretty high, these forecast are not able to express any of the variance of the real observations. This happens because busses stop very often in case of event and for this reason the model tent to always predict stopping. An example of a prediction balanced is given, Figure 33 shows the output from the historical average model, Figure 34 from the Support vector machine models and Figure 35 from the Bayesian model.

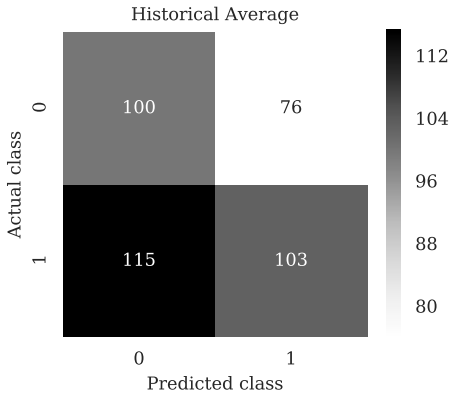

Figure 33: Confusion matrices of Historical average model on Telia Parken

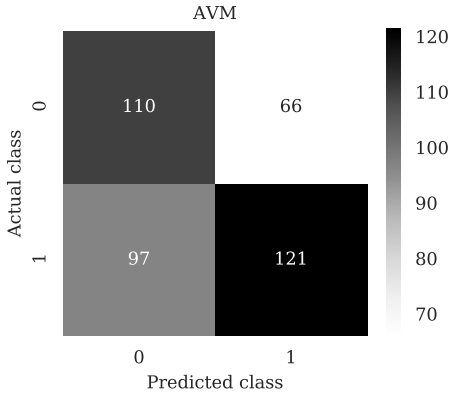

Figure 34: Confusion matrices of Support vector machine model on Telia Parken

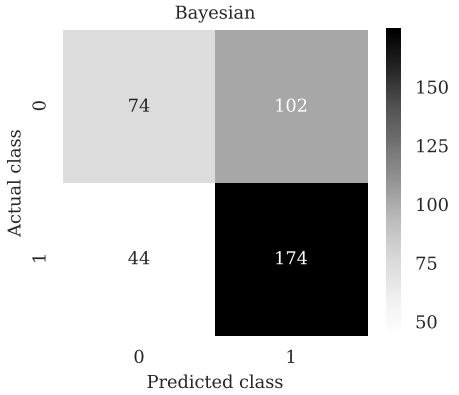

Figure 35: Confusion matrices of Bayesian model on Telia Parken

In this case the prediction is much more balanced. The confusion matrix of the models developed presents a higher concentration of prediction of True positive and True negative. Also, those model outperform the baseline showing that the model developed are effective.

7.3.3 Summary

It has been noticed that it is possible to provide a reliable prediction on bus stopping considering the time series and the time of the day as features. Also, event information do not have any significant impact on the quality of the prediction. The forecast on Forum and Vega is very unbalanced and is not able to predict any non stopping observation and,so, these two forecast have any relevance.

7.4 Dwell times

Here, the results of the prediction of dwell times are presented. In this case, the predictions take into consideration only observation in which the bus stooped and, so, removing all the zero seconds ones. The results report only the RMSE values which are shown in Table 27. The reason why the R squared values are not shown is because all of them are almost equal to zero. This means that none of the model are able to express the variance of the real observation of the dwell time due to the high variance of the observation and the low correlation with the input features.

Table 27: Errors of the prediction models computed during event days without event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	11.3	13.6	19.1	14.8
Telia Parken	1365	12.5	12.2	13.0	15.0
Vega	1586	6.8	7.0	11.1	7.2
Bella Center	27999	9.4	10.9	15.6	14.1
Koncerthust	860	5.6	5.9	8.9	7.1

The only model that perform better than the historical average is the linear regression but, also this model, does not present relevant improvement.

7.4.1 Considering event data

Here, the results of the prediction that consider event data are shown in Table 28. Also in this case R squared values are not presented for the same reason as before.

Table 28: Errors of the prediction models computed during event days and considering event data

Venue	Stop	Hist. Avg.	L. Regr.	Bay. Regr.	G. P.
Forum	678	11.3	9.5	18.7	14.6
Telia Parken	1365	12.5	11.8	11.8	12.1
Vega	1586	6.8	10.0	10.0	7.0
Bella Center	27999	9.4	15.0	15.0	12.0
Koncerthust	860	5.6	8.6	8.6	6.3

In this case, event information do not have almost any impact on the performance of the model meaning that the dependency between dwell time and events is very low.

7.4.2 Performance example

In this subsection, an example of the performance of the models is given. The example regards Telia Parken on 22nd of October. On this date, the football match between FC Copenhagen and Aarhus started at 18pm. Figure 36 compare the prediction with the real observations, the vertical dotted lines represent the starting and the ending time of the event.

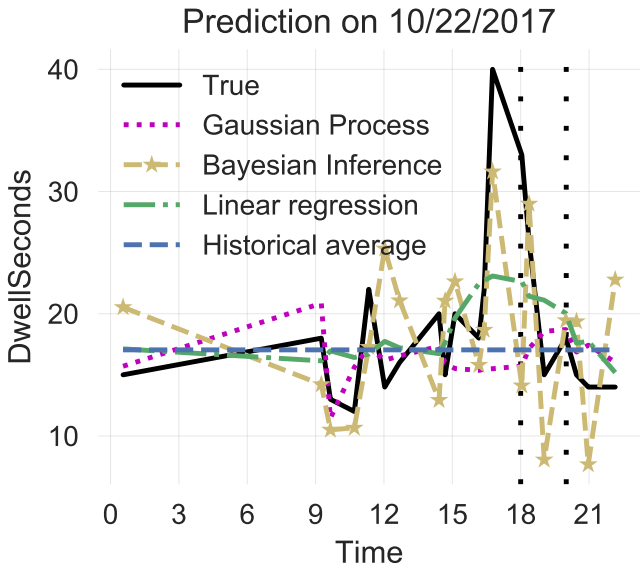

Figure 36: Example of prediction in Forum Copenhagen on 2nd of November

None of the prediction models' output is able to forecast precisely the real dwell time. In addition, the high observation occurring right before the match starting time is forecast only partially by the Bayesian model while the linear regression underestimate it.

7.4.3 Summary

None of the models obtained decent R squared values and none of them outperform the baseline. The results obtained show a poor performance of the models, mainly because of the that fact that the variance of the observations is hardly predictable due to the fact that the dwell time is low dependent on the input features. Additionally, event features do not improve the models' quality.

8 Discussion of results

In the previous chapter the result have been divided in four groups: passengers alighting, passengers boarding, bus stopping and dwell time. Each of them have shown particularity and interesting results.

First, the models developed to forecast passenger alighting shown to work pretty well, outperforming the base line and often obtaining good R squared values except for the venues Koncerthuset and Royal Arena. This means that the demand in this two locations is low correlated with the input features while in the other cases the correlation is higher. Additionally, considering event features considerably improve the models' performance.

The performance of the models developed to forecast passengers boarding are poor both in term of RMSE and R squared. Comparing the models developed with the historical average, the RMSE values are similar but that R squared value are, generally, higher. This means that the variation of the observations is partly expressed by the models but, also, that the predictions often diverge substantially from the real observations. The best results have been obtained for the venue Forum Copenhagen and Bella Center including event information in the models.

Based on the accuracy and F1 scores, the performance of the models developed outperform the baseline. However, in some case such as Forum Copenhagen and Vega, the prediction is not able to express the variability of the real observations. In addition, event information did not improve the models' performance in this case.

The performance of the models developed to forecast dwell time are poor both in term of RMSE and R squared for mainly two reason: Dwell time is characterized by a quite low auto-correlation meaning that two consecutive observations can be quite different. Secondly, the forecast tent to be an average of all the previous observations and therefore not able to express any of the variation.

9 Conclusion

This paper provides an analysis and short term prediction models of bus demand and bus dwell time in case of special events. This paper shows that, in some location, the demand is positively correlated with the occurrence of events and that event information can be used in prediction model to improve the forecast.

In the fist part, the demand of public transport in terms of passengers alighting and boarding is analyzed in order to find relevant pattern and correlation. The observation of passengers alighting presents a strong auto-correlation and shows a pattern repeating 24 hours. The analysis of the correlation between event information and alighting passengers shows that the demand in Telia Parken, Bella Center and Vega is high correlated with the occurrence of event and, in particular, the three hours before the event starting time are the most relevant. Additionally, the analysis demonstrated that similar events share similar characteristics and that not all the events have the same impact: Football related event are more impactful on the demand in Telia Parken and culture related event are more impactful in Bella Center.

Then, the demand has been considered from the point of view of passengers boarding. The auto-correlation analysis shows not only a strong correlation with the observation occurred short time before but also it shows a patter repeating every 24 hours. What results from the correlation analysis with events occurrence is that only the demand in Bella Center is highly correlated with events and that culture related events are more impacful that others.

Since the Public Transport provider can be interested on understanding the main factor of very high level of demand, an analysis on the abnormal observations has been carried out. It has been shown that most of the abnormal observation of passengers alighting, occurred on Vega, Bella Center and Telia Parken, happened in proximity of events.

Regarding the demand in term of passengers boarding, it has been observed the same situation on Bella Center.

After that, the observation of dwell time has been analyzed from two points of view: the rate of stopping and the duration of the dwell time. An analysis of the autocorrelation has shown that an observation is positively correlated with the previous ones. Also, it has been demonstrated that dwell times occurred in Telia Parken are more correlated with football events than with other kind of event. Then, further analysis has shown that abnormal dwell time on Telia Parken and Bella Center is likely to happen in correspondence of events.

In the second part of the paper, three short term prediction models for forecasting the bus demand have been developed and tested on real cases and their quality has been evaluated. Generally, the models developed provide a lower error compared to the baseline and those are also capable to express good part of the variability of the demand. However, it has been noticed that the models are performing poorly in case of Royal Arena and Koncerthuset meaning that the demand in these two location is slightly correlated with the input features. Also, it has proved that including event information on regression model can increase the quality of the prediction. Those information has been included both as proximity with the starting or ending time and as event typology. The case in which the event information have led to the most relevant improvement are: Telia Parken, Vega and Bella Center for passengers alighting. Then, it has been presented two classification models for predicting the bus stopping and three regression models for predicting the duration of the dwell times. Overall the classification models obtained good accuracy and F1 values. However, the models tested on Vega and Forum Copenhagen was not capable to express the variability of the real observation. The models developed to forecast the duration of the dwell time have performed poorly meaning a low correlation between the input features and the dwell time. Both in case of stopping and dwell time, including the event information did not improved the models' performance.

Acknowledgements

I would like to thank Francisco Camara Pereira for the supervision of this thesis, his guidance and for all the support received. Also, I want to thanks all the department of Management engineering, Transport modeling and, in particular, Filipe Rodrigues, Inon Peled, Niklas Christoffer Petersen and Ioulia Markou for the help and the advises that they gave me.

Inoltre, ringrazio Elena per tutto il suo impegno e i suoi sacrifici. Ringrazio anche la mia famiglia e mia zia Paola per tutto l'affetto e il supporto che mi hanno dato.

Appendix

The code written in the prepossessing phase, for the explanatory analysis and to develop and test the models have been written in in Python using Jupyter Notebooks 2. All the file are attached separately.

References

[1] F. Qin. Investigating the In-Vehicle Crowding Cost Functions for Public Transit Modes. Mathematical Problems in Engineering. 2014.
[2] A. Tirachini, D. A. Hensher, J. M. Rose. Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand. Transportation Research Part A. 2013.
[3] S. Anvari, S. Tuna, M. Canci, M. Turkay. Automated BoxâĂŞJenkins forecasting tool with an application for passenger demand in urban rail systems. Journal of Advanced Transportation. 2015.
[4] M. Gan, Y. Cheng, K. Liu, G. Zhang. Seasonal and trend time series forecasting based on a quasi-linear autoregressive model. Applied Soft Computing. 2014.
[5] M. Milenkovic, L. Svadlenka, V. Melichar, N. Bojovic, Z. Avramovic, SARIMA modelling appoach for rail way passenger flow forecasting. Transport. 2015.
[6] Y. Sun, B. Leng, W. Guan. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. Neurocomputing. 2015.
[7] J. Guo, W. Huang, B.M. Williams. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification, Transportation Research Part C: Emerging Technologies. 2014.
[8] , Y. Bai, Z. Sun, B. Zeng, J. Deng, C. Li. A multi-pattern deep fusion model for short-term bus passenger flow forecasting. Applied Soft Computing. 2017.
[9] F. C. Pereira, F. Rodrigues, M. Ben-Akiva. Using data from the web to predict public transport arrivals under special events scenarios. Journal of Intelligent Transportation Systems. 2015.
[10] C. Zhou, P. Dai, R. Li. The Passenger Demand Prediction Model on Bus Networks. 2013 IEEE 13th International Conference on Data Mining Workshops. 2013.
[11] F. Potier, P. Bovy, C. Liaudat, Big events: Planning, mobility management. European transport conference. 2003.
[12] S. P. Latoski, W. M. Dunn, Jr., B. Wagenblast, J. Randall, M. D. Walker. Managing Travel for Planned Special Events. 2006.
[13] J. Skolnik , D. Stern , R. Chami , A. Lane, C.S. Kulesza, M. Walker. Planned Special Events: Cost Management and Cost Recovery Primer. U.S. Department of Transportation Federal Highway Administration. 2009.
[14] F. C. Pereira, F. Rodrigues, E. Polisciuc, M. Ben-Akiva. Why so many people? Explaining Nonhabitual Transport Overcrowding With Internet Data. 2015.
[15] . C. Pereira, F. Rodrigues, S. S. Borysov, B. Ribeiro.A Bayesian Additive Model for Understanding Public Transport Usage in Special Events. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017
[16] S. Jaiswal, J. Bunker, L. Ferreira. The influence of platform walking on BRT station bus dwell time estimation. Journal of Transportation Engineering. 2010.
[17] Transportation research board of the national academies. Transit Capacity and Quality of Service. 2003.
[18] V. R. Vuchic. Urban Transit: Operations, Planning, and Economics. 2005.
[19] J. Amita, S.S. Jain. P. K. Garg. Prediction of Bus Travel Time using ANN: A Case Study in Delhi. ScienceDirect. 2014.
[20] R. Jeong, L.R. Rilett. Bus Arrival Time Prediction Using Artificial Neural Network Model. Intelligent Transportation Systems Conference. 2004.
[21] M. Milkovits. Modelling the factors affecting bus stop dwell time. Journal of transpor tresearch board. 2008.
[22] S. Rashidi, P. Ranjitkar. Estimation of bus dwell time using univariate time series models. Journal of advanced transportation. 2015.
[23] R. Rajbhandari, S. I. Chien, J. R. Daniel. Estimation of Bus Dwell Times with Automatic Passenger Counter Information. Transportation Research Record. 2003.
[24] H. Jelodar, Y. Wang, C. Yuan, X. Feng. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.
[25] D. M. Blei. Probabilistic Topic Models. Communications of the acm. 2012.
[26] D. Blei, A. Ng, M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research 3. 2003.
[27] Interview with a Movia employee. 10/05/2018.