Fare revenue forecast in public transport: a comparative case study

Jonas Krembsler[a], Sandra Spiegelberg[b], Nicki Lena Kämpf[b], Thomas Winter[b], Nicola Winter[a], Robert Knappe[a]

[a] *Hochschule für Wirtschaft und Recht, Alt-Friedrichsfelde 60, Berlin, 10315, Berlin, Germany*
[b] *Berliner Hochschule für Technik, Luxemburger Str. 10, Berlin, 13353, Berlin, Germany*

Abstract

This paper presents results from a case study of fare revenue forecast in public transportation in Berlin base d on monthly fare revenues aggregated for different product segments. The forecast is generated with the intention of automating revenue controlling and implementing data-driven decision-making within the existing controlling processes.

The following prediction methods are applied in order to obtain suitable and reliable predictions: autoregressive methods and exponential smoothing - and methods able to include exogenous variables - such as SARIMAX, MLR, LASSO, and Ridge. The data concerning exogenous variables are freely available and cover tourism data, labor market development, and weather data.

We compare the prediction results of the forecast methods. The goal is to evaluate a wide range of methods in order to decide in which situations they perform best. We apply automatic feature selection, discuss the interpretability of the results and the performance of the different approaches.

Keywords: public transport, forecast, revenue, time series, regression, revenue controlling

Acknowledgements

This research was funded by the Berlin Institute for Applied Research (IFAF) as part of the research project ReComMeND (Revenue Controlling mit mathematischer Modellierung auf Basis empirischer Daten für Nahverkehrs-Dienstleistungen). ReComMeND started in 2019 involving the project partners Berliner Hochschule für Technik, Berlin School of Economics and Law, Berliner Verkehrsbetriebe AöR, International Association of Controllers (ICV) work group Berlin-Brandenburg, Lufthansa Systems GmbH & Co. KG, and Lufthansa Industry Solutions GmbH & Co. KG. We would like to thank all project partners for their contributions to the research work.

Email addresses: Jonas.Krembsler@hwr-berlin.de (Jonas Krembsler), Sandra.Spiegelberg@bht-berlin.de (Sandra Spiegelberg), NickiLena.Kaempf@bht-berlin.de (Nicki Lena Kämpf), Thomas.Winter@bht-berlin.de (Thomas Winter), Nicola.Winter@hwr-berlin.de (Nicola Winter), Robert.Knappe@hwr-berlin.de (Robert Knappe)

Fare revenue forecast in public transport: a comparative case study

Abstract

This paper presents results from a case study of fare revenue forecast in public transportation in Berlin base d on monthly fare revenues aggregated for different product segments. The forecast is generated with the intention of automating revenue controlling and implementing data-driven decision-making within the existing controlling processes.

The following prediction methods are applied in order to obtain suitable and reliable predictions: autoregressive methods and exponential smoothing - and methods able to include exogenous variables - such as SARIMAX, MLR, LASSO, and Ridge. The data concerning exogenous variables are freely available and cover tourism data, labor market development, and weather data.

We compare the prediction results of the forecast methods. The goal is to evaluate a wide range of methods in order to decide in which situations they perform best. We apply automatic feature selection, discuss the interpretability of the results and the performance of the different approaches.

Keywords: public transport, forecast, revenue, time series, regression, revenue controlling

1. Introduction

Accurate forecasts of expected passenger traffic and the resulting revenue from ticket sales are important for planning and operation of public transport systems. We focus on a prediction problem arising in management accounting and revenue controlling of ticket sales in Berlin at the example of BVG, the largest public transport company in Berlin.

Every year a prediction of the expected revenue for the next fiscal year is to be made. This forecast will be checked and adapted on a monthly basis during the year. The goal is to receive a forecast that is as accurate as possible.

The main difficulty in the prediction problem is that Berlin is an evolving city with an increasing number of inhabitants and hence an increasing number of users of the public transport systems. In addition, the mobility patterns of the users are changing which is also supported by political measures like the Berlin mobility act. Besides this, there are further factors impacting the traffic demand and ticket sales in a positive or negative way, for instance job tickets, the employment rate, fuel prices, weather conditions, as well as unexpected events like strikes or changes in the product structure.

Hence, the resulting prediction problem is a prediction problem where the impact of exogenous factors needs to be taken into account. Moreover, the expectation from the controllers and higher management is that the impact of these exogenous factors can be quantified or at least to some extend be explained.

In this comparative case study, we apply different prediction models and methods from time series theory and regression theory to sales data and exogenous data available from 2005/2012 to 2018 or 2019, respectively. We compare the results of the methods including exogenous factors to methods that are only based on revenue data for different product segments. It becomes apparent that in most cases the consideration of exogenous factors does not result in an higher accuracy of the prediction results. This is mostly true when there are no unexpected external effects so that the regular external effects like trends and seasonal impacts are already included in the revenue data. However, methods with exogenous factors perform best when it comes to predicting the total revenue, which is the most important forecast for the liquidity planning and management.

The paper is structured as follows. We start with a brief review of revenue prediction in public transport, in particular concerning the application of time series models and the use of exogenous data. Then, we describe the prediction problem for the different product groups and the underlying data in further detail. In section 4, we introduce the different forecast models and methods considered. We distinguish between methods using exogenous variables and methods which are not able to deal with such variables. In section 5, we show the results of these methods applied to our data for different evaluation periods. We conclude with a brief summary of the results.

2. Brief literature review

Autoregressive and moving average models have been introduced already in the 1950s. They were considered for modeling and prediction purposes in practice beginning in the 1970s and more intensified since the 1990s [8, 9]. ARIMA and seasonal ARIMA models have been applied to various applications in tourism, epidemics, ecology and environmental studies etc. As additional forecasting methods, exponential smoothing methods are introduced and extended in the 1950s [10, 19, 20] and the beginning 1960s [11, 33]. They have been applied quite successfully to various applications, for instance in the tourist sector [23] or in comparison to (S)ARIMA-based approaches in environmental modeling [2] and concerning air traffic demand in [13].

Methods and models including exogenous variables like SARIMAX models have been considered as well in various applications like [32], [31], and [6].

Prediction methods are widely used for traffic and revenue prediction in public transport. However, most traffic prediction is carried out for origin-destination pairs, for instance in [30]. Most articles on modeling and prediction the traffic demand in public transport ignore which kind of ticket is used and which revenue is yield. Revenue prediction is mostly done on a strategic level predicting the overall evolution of traffic, for instance in [5]. For a comprehensive overview on forecast methods and principles we refer to [22]. When it comes to a more short-term forecast of the revenue gained by public transport services including the impact of potential external impacts only a few studies are available.

A comparison of univariate and multivariate time series models has been carried out in Tsai, Mulley, and Clifton for the metropolitan area of Sydney, Australia [29]. Tsai, Mulley, and Clifton consider the application of ARIMA models and the partial adjustment model (PAM). They fit both models to monthly data derived from train and bus services. Using the PAM model, they estimate demand elasticities considering a number of public transport determinants as exogenous variables such as tourist demand and fuel prices.

In a recent paper Su and Su consider the application of optimized ARIMA models to public transport data in the city of Istanbul for rail and bus services [28], however without focusing on external factors.

3. Problem Description and Data Preprocessing

The goal of this study is to predict the fare revenues of the BVG. However, this is not a straightforward task. Besides the common challenges of forecasting problems, such as finding an accurate prediction model, defining the forecast metric etc., there are further obstacles inherent in the revenue management of public transport systems. One obstacle is that multiple public transport companies operate in the same areas in Germany, meaning that the revenue in one operating area needs to be redistributed between the companies based on earnings. In the subsection "‘Segmented Modeling"’, we describe how we account for the redistribution.

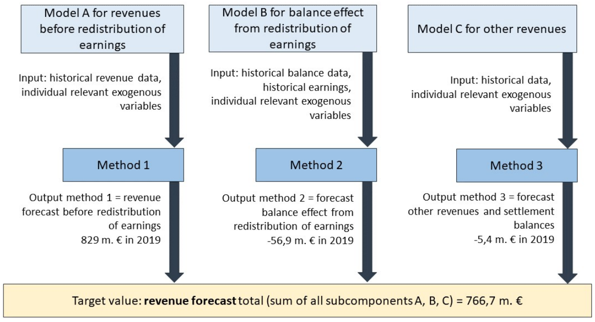
Another obstacle is that the revenue in the public transport sector is highly driven by exogenous factors. Therefore, these factors need to be identified as well as preprocessed and aligned with the granularity of the internal data from the BVG.

One important step in every predictive analytics project is the data preprocessing. This process is described in detail in the subsection "‘ Preprocessing"’. Before the prediction methods can be applied, the data need to be preprocessed. In the case of fare revenue prediction in public transport, some problems concerning the data preprocessing arise from different data granularity with respect to quantity and time. In particular, the data granularity of the impact factors varies from hourly to monthly to quarterly to yearly. Other problems result form changes of the data structure due to system changes as well as from changes in the product (ticket) structure itself, for example due to the political measures as the introduction of the free Berlin school student ticket, or due to postponed accounting entries. The data are provided by the BVG or are publicly available.

3.1. Segmented Modeling

In Berlin, there exists a redistribution of earnings with between the transport companies due to a special redistribution key according to the transport association contract. This is the result of the fact that in Berlin the ticket of one transport company (e.g. BVG) is also valid for the other transport companies operating in Berlin (like S-Bahn or DB Regio). This redistribution of earnings has effects on the revenue of each transport company.

While the relevant indicator for the revenue controlling is the revenue after the redistribution, it cannot be forecasted easily. Multiple forecasts are necessary to account for the redistribution of direct earnings and other revenue resulting from redistribution. To deal with these aspect, a segmented forecasting approach is developed as shown in fig. 1. In this paper, we focus on model A.



Model A for revenues before redistribution of earnings
Model B for balance effect from redistribution of earnings
Model C for other revenues
Input: historical revenue data, individual relevant exogenous variables
Input: historical balance data, historical earnings, individual relevant exogenous variables
Input: historical data, individual relevant exogenous variables
Method 1
Method 2
Method 3
Output method 1 = revenue forecast before redistribution of earnings 829 m. € in 2019
Output method 2 = forecast balance effect from redistribution of earnings -56,9 m. € in 2019
Output method 3 = forecast other revenues and settlement balances -5,4 m. € in 2019
Target value: **revenue forecast** total (sum of all subcomponents A, B, C) = 766,7 m. €

Figure 1: segmented forecasting approach for the public transportation in Berlin

Figure 1 shows that the forecast of the total revenue is splitting into three parts: model A as forecast of the revenue before the redistribution, model B as forecast of the balance effect resulting from the redistribution, and model C as forecast of revenue of other products which are part of other small redistributions. In model A, we use historical revenue data and relevant exogenous variables as an input for method 1. The output of this method is the forecast of the revenue before redistribution. In model B, the approach is quite similar. Only the input data differ, as historical balance data and individual relevant exogenous variables are used. The output is the forecast taking into account the balance effect from the redistribution. Model C considers some simple methods like the naïve method or a fixed value prediction.

Because of focusing on model A in this study, the revenue is considered and not the earnings. In the fig. 2, the difference between the revenue and the earnings is negligible: the peaks of the earnings have been smoothed out in the revenues, for example, by distributing annual payments over the relevant months. Furthermore, fig. 2 shows that from 2005 until 2008 the revenue is higher than the earnings because of the monthly tickets and the student ticket. This is due to a change in the data handling for monthly tickets and because the earnings include the student tickets only beginning from 2008.

3.2. Data Preprocessing

The provided data include monthly ticket sales per product from 2005 to 2019. Over the years, there were changes in the product structure or the terms of use, so that the data needed to be cleaned an preprocessed. For some tickets we have only incomplete time series data.

The products were grouped in two different ways: the product data were divided according to product groups (single tickets, daily tickets, monthly tickets, yearly tickets, etc.) and according to user groups (pupils, employees, tourists, seniors, etc.), respectively. In this paper, we consider only the product groups because not all tickets could be clearly assigned to the user groups. In addition, the monthly revenue data were only available at product group level.

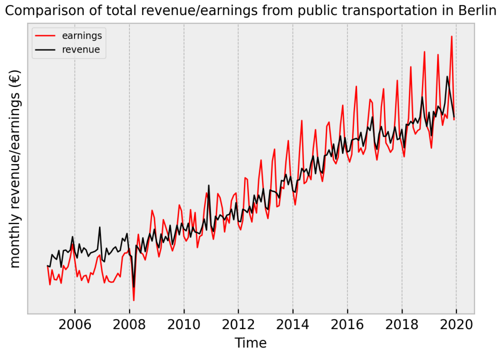


Figure 2: Comparison of the time series of the total revenue and earnings for public transportation in Berlin

Introducing product groups help us to include the impact of exogenous factors. Furthermore, correlations such as user migrations, for example, transitions from single one-way tickets to 4-trip tickets resulting from tariff changes are difficult to model. Since both tickets are in the same product group, we can neglect transitions within a product group.

Similar to almost every time series data set, our data has gaps due to missing values, different spellings or blanks in the tables. We adapted the data by interpolation, when appropriate, and replaced unknown missing values by zero so that the time series of the product groups are complete and the models do not break off at these points.

In 2019, a free Berlin school student ticket was introduced. This had the effect that many school students changed their tickets from monthly tickets to a subscription, what the free school student ticket is classified. Compensation payments from the Berlin government were payed, which generated more fare revenue than before, since almost every school student demanded this ticket. In order to absorb this shock, the whole data was cleaned by removing all tickets that could be clearly attributed to school students.

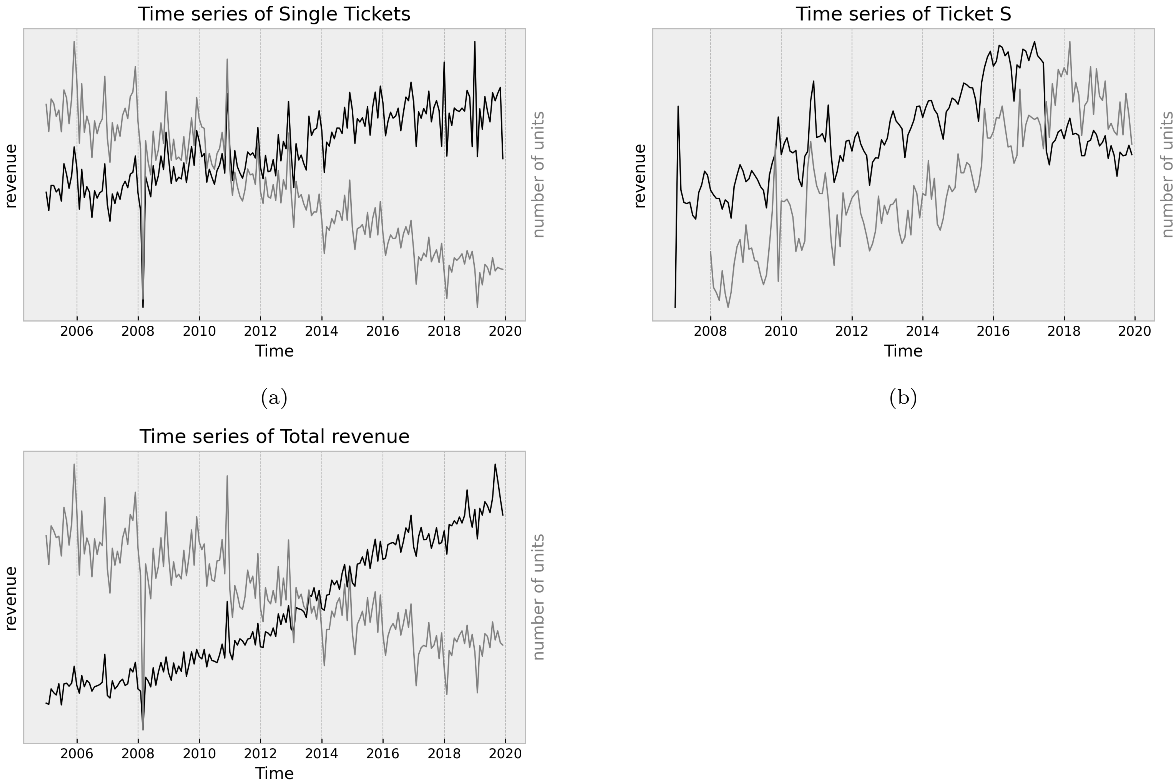
The price development of the individual products caused some user migration between the products without significant effect on total revenue.

Due to confidentiality the y-axes in figure fig. 3 do not show the actual values. As a first observation, the large drop at the beginning of 2008 visible for the single tickets as well as the total market was due to a strike in public transport at this time. As a consequence, an exogenous variable for strikes is introduced. More information can be founded in section 3.2.1, *BVG Strikes*.

The number of tickets sold in the product group "‘single tickets"’ (cf. fig. 3 (a)) decreases whereas the revenue increases. This is caused by a tariff change: Four-way tickets became relatively cheaper compared to one-way tickets. So customers switch from one-way tickets to four-way tickets. Additionally, a four-way ticket is considered as one unit so that the number of units sold decreases. The graph of ticket S in fig. 3 (b) shows that in 2013 the number of units decreased due to a price increase and in June 2017 there was a price decrease so the revenue decreased as well. These two examples for price changes show that there is almost no effect on the total revenue (seefig. 3 (c)).

As tickets are aggregated in product groups, it is difficult to determine an average price per unit sold. One reason is that the price development was not available for all tickets for the entire time series. Further details of the price development is shown in section 3.2.1, *Prices*. Therefore, the price is not considered directly as exogenous variable in this case study.

A list of all product groups, their percentage share of total revenue and the number of products within each group is shown at the example of 2019 in table 1 and fig. 4. In this case study, the forecast of the total market and the four largest product groups (subscriptions, single tickets, monthly tickets and daily tickets) are considered with respect to their percentage share.



(a)

(b)

(c)

Figure 3: time series of revenue (black) and number of units (gray) for the product groups (a) single tickets (b) ticket S and (c) total market

product groups	**shares in %**	**number of products**
subscriptions	38.63	106
single tickets	22.17	25
monthly tickets	11.26	20
daily tickets	6.77	17
ticket S	5.42	1
student tickets	4.89	1
company tickets	4.37	18
tourist tickets	1.85	194
weekly tickets	1.51	3
yearly tickets	0.94	21
other tickets	2.20	17
Sum	100	423

Table 1: list of product groups for 2019

3.2.1. Exogenous variables

The exogenous data need to be available with the same granularity and quality as the revenue data. i.e., on a monthly basis preferably from 2005 to 2019. Potential exogenous variables used in this case study are shown in the following table 2. Furthermore, the table shows some statistical details about these exogenous variables: the granularity, minimum, maximum, mean, the sources, and since when they are available.

The following four exogenous variables are explained in more detail: overnight stays (cf. fig. 5), population (cf. fig. 6), apprentices (cf. fig. 7) and strikes (cf. fig. 8).

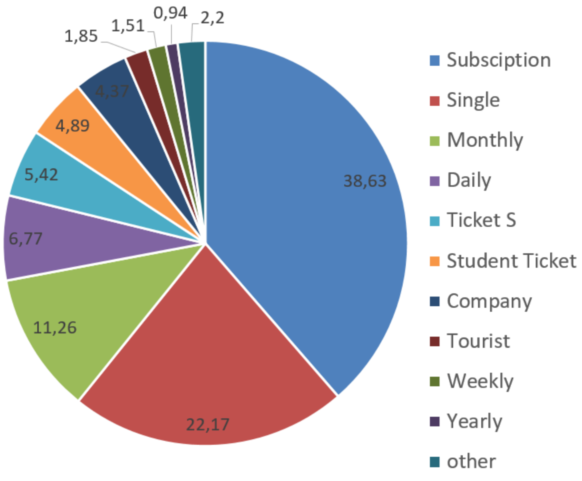


Figure 4: product group shares in % of total revenue

Overnight stays. The number of overnight stays is used to represent the number of tourists in Berlin. The data is available on a monthly basis without gaps. There is a clear visible seasonality of 12 months (cf. fig. 5). The data does not include all touristic stays in Berlin. Only the stays in hostels and hotels with 10 or more beds are included but no Airbnb or similar company data. The data is provided by the Federal Statistical Office of Germany.

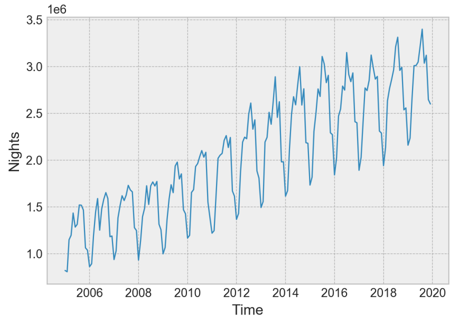


Figure 5: time series for the exogenous variable "‘overnight stays"’

Population. Reliable population data are only available from 2012 on. It was adapted by the census in 2011. Before 2012, the population data was biased and was only available on a quarterly basis. The commuter data is also only available from 2012. The population data for Berlin is provided by the statistics office. In fig. 6 it can be seen that the population of Berlin is continuously increasing.

Apprentices. The apprentices data are also provided by the statistics office, but on a annual basis. In order to receive a monthly granularity, a step function was used as shown in fig. 7. We assume that the number of apprentices in a one-year apprenticeship remains relatively constant. The figure shows also that the number of apprentices in Berlin is decreasing year by year.

BVG Strikes. We use Google Trends for generating a time series for modeling strikes. Google Trends is a free tool by Google providing the relative frequency of search terms in Google searches as a time series. This can be restricted to any time horizon and geographical area. The motivation for using the results of Google

Exogenous variable	**Granularity**	**min** **max** **mean**	**Available since**	**Source**
overnight stays (in Hostels and Hotels with 10 or more beds)	monthly	807,980 3,399,163 2,042,249	2005	Federal Statistical Office of Germany [27]
apprentices	annual (step function to monthly)	38,432 56,787 46,770	2005	Amt für Statistik Berlin-Brandenburg [3]
students (universities)	annual (step function to monthly)	132,822 195,799 157,899	2005	Federal Statistical Office of Germany [26]
number of unemployed	monthly	146,670 333,439 218,995	2005	Federal Statistical Office of Germany [12]
population	monthly	3,326,002 3,666,488 3,510,899	2012	Statistische Bibliothek [4]
monthly mean gasoline price in Cent	monthly	109.85 173.67 141.39	2005	Wirtschaftsverband Fuels und Energie e.V. (en2x)
12 monthly dummies	monthly	0 1 0.0833	2005	calendar
school holidays	daily (agg. to monthly)	0 31 7.6	2005	www.schulferien.org [24]
sundays and public holidays	daily (agg. to monthly)	4 8 5.04	2005	calendar
number of commuters from Berlin to Brandenburg	annual (interpolated to monthly)	69,770 88,743 81,804	2014 (extrapolated to 2012)	Federal Employment Agency [12]
number of commuters to Berlin	annual (interpolated to monthly)	266,810 337,949 292,418	2014 (extrapolated to 2012)	Federal Employment Agency [12]
days with snow[1]	hourly (agg. to monthly)	0 30 3.16	2005	Deutscher Wetterdienst [15]
days with rain[2]	hourly (agg. to monthly)	1 24 10.85	2005	Deutscher Wetterdienst [15]
(BVG) Strike	monthly	0 100 2.83	2005	Google Trends [16]

Table 2: List of exogenous variables used

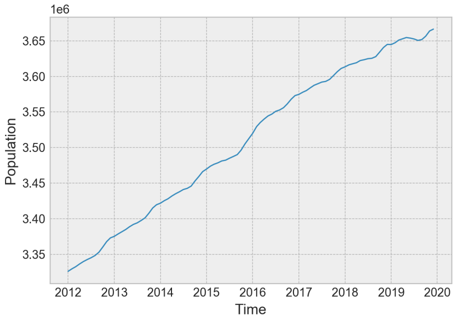


Figure 6: time series for the exogenous variable "‘population"’

Trends is that more people search for „BVG strike“ when a BVG strike takes place. We observe that Google Trends time series reflects the exogenous variable strike quite well because in the graph fig. 8 the peaks in

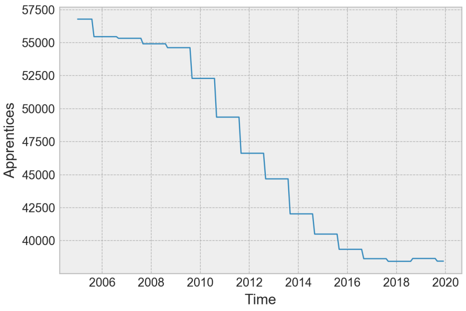


Figure 7: time series for the exogenous variable "‘Apprentices"’

2008, 2012 and 2019 match to the corresponding strikes but with a different intensity.

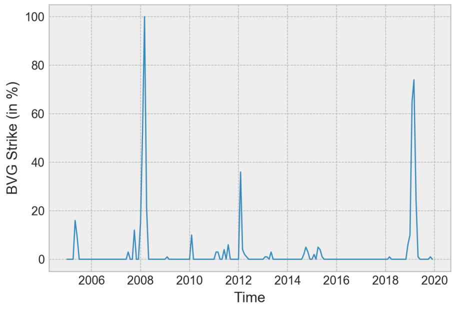


Figure 8: time series for the exogenous variable "‘BVG strike"’ in %

Prices. Since the product groups consists of different tickets with different tariffs, it is difficult to calculate a valid average price per unit sold for a product group over time. Another difficulty is that in addition to pure price changes the terms of use change as well, in most cases not for all products in a product group at the same time.

Over the years, tariffs have been adjusted, with prices mostly increasing. Price changes are not initiated by the public transport companies alone, because they need to be approved by the Berlin government. The main focus for a public transport company in Germany is on cost-covering financing of the service and not on profit maximization.

Price increases usually do not affect all products at the same or at the same amount. Migration between product groups due to price adjustments cannot be ruled out and are sometimes intended. From a global perspective, the fare increases do not lead to a decline in demand, as the price increases are typically below a 2 % price inflation rate.

In our models, the price is not included directly because of the previously mentioned reasons. However, price changes are indirectly included in all the revenue values.

4. Models and methods

4.1. Models without exogenous variables

4.1.1. ARIMA

The ARMA model combines autoregression and moving average models. The autoregressive part is characterized by an autoregressive model AR(p). p indicates the order of the autoregressive model AR(p), i.e., p past values are considered ([17]):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{1}$$

with ε_t as a white noise process. The moving average model MA(q) with the order q is characterized as follows ([17]):

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{2}$$

where the error term ε_t is a white noise process, $(\theta_1, \theta_2, \ldots, \theta_q)$ could be any real numbers and the mean $E(y_t) = \mu$. If, in addition, stationarity is ensured by d-times differencing $w_t = \Delta^d y_t$, this becomes the ARIMA model. To do this, the three orders p, d and q described above must first be determined according to the parsimony principle (i.e. as low as possible). This often done with an Akaike information criterion (AIC) optimization. The AIC basic formula is defined as ([1]):

$$AIC = (-2)\ log\ (maximum\ likelihood) + 2k \tag{3}$$

with k as the number of model parameters and the log-likelihood of the model fit. The lower the AIC, the better the model. Then, the coefficients ϕ and θ are estimated. The ARIMA model can thus be expressed as follows ([8]):

$$\begin{aligned} w_t =& c + \phi_1 w_{t-1} + \cdots + \phi_p w_{t-p} + \varepsilon_t \\ & -\theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \text{ with } w_t = \Delta^d y_t \end{aligned} \tag{4}$$

Using the backward shift operator, the previous equation can be written as follows

$$\phi_p(B)(1-B)^d y_t = \delta + \theta_q(B)\varepsilon_t \tag{5}$$

with $\mathit{\Phi}_p(B)$ as the autoregressive operator of order p, $\theta_q(B)$ the moving average operator of order q, and $(1-B)^d$ the differentiation operator of order d.

4.1.2. SARIMA

4.1.3. SARIMA(p,d,q)(P,D,Q)$_S$

The SARIMA model, as an extension of the simple ARIMA model, includes the effects of seasonality. As in the $ARIMA(p, d, q)$ models, p indicates the order of the autoregressive model AR(p) and q indicates the order of the moving average part $MA(q)$. If the time series y_t is considered as an integrated process, it can be differentiated as often as necessary until a new, stationary time series w_t is created. The parameter d indicates the order, i.e., how often the time series must be differentiated until the resulting time series w_t is stationary. S is the seasonality factor, which describes the number of periods before seasonality is repeated. For example, for yearly quarters $S = 4$, for monthly data $S = 12$. The capitalised parameters P, D and Q are equivalent in meaning to p, d and q, except that they refer to the seasonal component of the model (the coefficients Φ and Θ). The seasonal ARIMA model can be expressed as follows ([22]):

$$\begin{aligned} w_t =& \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} \\ & + \Phi_1 w_{t-S} + \Phi_2 w_{t-2S} + \ldots + \Phi_P w_{t-PS} \\ & + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \\ & - \Theta_1 \varepsilon_{t-S} - \Theta_2 \varepsilon_{t-2S} - \ldots - \Theta_Q \varepsilon_{t-QS} \end{aligned} \tag{6}$$

or, using the backward shift operator:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D y_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \tag{7}$$

with $\mathit{\Phi}_p(B)$ as the autoregressive operator of order p, $\theta_q(B)$ the moving average operator of order q, $(1-B)^d$ the differentiation operator of order d, $\Phi_P(B^S)$ the seasonal autoregressive operator of order P, $\Theta_Q(B^S)$ the seasonal moving average operator of order Q, and $(1-B^S)^D$ the seasonal differentiation operator of order D. Analogously to the ARIMA model, the orders p, d, q, P, Q and D are identified with the AIC criterion.

Since the monthly revenue and earning data are highly seasonal, this paper excludes the basic ARIMA models and focuses on SARIMA models.

4.1.4. Exponential smoothing: Holt-Winters method

The following description of the Holt-Winters method can be found in [7]. The Holt-Winters method as an extension of the Holt method for a linear trend and seasonality, requires three parameters $0 < (\alpha, \beta, \gamma) < 1$ and the following initial values: $l_t, b_t, s_1, ..., s_S$.

Holt [19] and Winters [33] expanded Holt's method to include seasonality. The Holt-Winters method includes the forecast equation and three smoothing equations for the level l_t, the trend b_t, and the seasonal component s_t with the corresponding smoothing parameters α, β, and γ as seen in eqs. (8) to (10) for the additive model. By S, we denote the period of seasonality, i.e., the number of seasonal parts in a year, for example, $S = 12$ for monthly dates and $S = 52$ for weekly dates.

There are two types of this method: additive and multiplicative. They vary with respect to the seasonal component. The additive method is favored if the seasonal variation is approximately constant across the whole time series, while the multiplicative method is favored when the seasonal variation changes proportionally to the level of the time series. In the additive model, the seasonal component is given in absolute terms on the scale of the observed time series. In order to adjust the time series for seasonality, the seasonal component is subtracted in the level equation. For each year, the seasonal component sums up to approximately zero. In the multiplicative model, the seasonal component is given in relative values (percentages) and in order to adjust the series for seasonality, the series is divided by the seasonal component in the level equation. For each year, the seasonal component will add up to approximately S.

Additive model.

$$\text{Level:} \quad l_t = \alpha(y_t - s_{t-S}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{8}$$

$$\text{Trend:} \quad b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{9}$$

$$\text{Season:} \quad s_t = \gamma(y_t - l_t - b_{t-1}) + (1-\gamma)s_{t-S} \tag{10}$$

$$\text{Forecast:} \quad f_{t+h|t} = l_t + h \cdot b_t + s_{t+h-S(k+1)} \tag{11}$$

where $f_{t+h|t}$ is the forecast for h periods ahead and k is the integer part of $\frac{h-1}{S}$, ensuring that the estimates of seasonal indices used in the forecast are from the most recent year of the sample. The level equation shows a weighted average between the seasonally adjusted observation $(y_t - s_{t-S})$ and the non-seasonal forecast $(l_{t-1} + b_{t-1})$ for time t. The trend equation is equal to Holt's linear method. The seasonal equation shows a weighted average between the current seasonal index $(y_t - l_{t-1} - b_{t-1})$ and the seasonal index of the same season last year (i.e. S time steps before).

According to [7], for the seasonal component the equation is often written as

$$s_t = \gamma^*(y_t - l_t) + (1 - \gamma^*)s_{t-S}$$

If l_t from the smoothing equation is used for the level of the component form above we get the following

$$s_t = \gamma^*(1-\alpha)(y_t - l_{t-1} - b_{t-1}) + [1 - \gamma^*(1-\alpha)]s_{t-S}$$

which is identical to the smoothing equation for the seasonal component where $\gamma = \gamma^*(1-\alpha)$. The usual parameter restriction $0 \leq \gamma^* \leq 1$ leads to $0 \leq \gamma \leq 1 - \alpha$.

Initialization. We set up the Holt Winters model in Python using the `ExponentialSmoothing` function of the `statsmodels` package and as minimizer of this function `basinhopping` . In order to find the best Holt Winters parameter set, we apply a grid search for the trend, seasonal, damped trend and to use `boxcox` resulting in 36 parameter options. The grid search is based on AIC optimization. As an additional restriction, we apply an upper and lower bound to the weight parameter α: $0.2 \leq \alpha \leq 0.8$.

As initialization of s_t for the first season period we use

$$s_t = \frac{y_t}{m_s} \text{ where } m_S = \frac{1}{S}\sum_{t=1}^{S} y_t \tag{12}$$

this means the first s_t-values indicate how the current value relates to the mean over the first period. b_S is set to 0, l_S is set to m_S, i.e., the mean. At $t = S + 1$ the recursive computation of the Holt Winters model starts.

4.2. Models integrating exogenous variables

4.3. Regression models

Multiple linear regression. In multiple linear regression, a linear equation is set up from two to k independent variables and one dependent variable. The coefficients represent the estimated change in y_i when the associated x_i value increases or decreases by one, holding the other variables constant:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{13}$$

The matching coefficients are found using the least squares method, for which we minimize the following equation (cf. [18]):

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij})^2 \tag{14}$$

A potential problem with this approach is over-fitting where too many independent variables are included in the equation. This leads to an unnecessarily high variance, which does not improve the prediction accuracy. Another problem is possible multicollinearity between the "‘independent"’ variables.

Ridge regression. Ridge regression is a method to avoid over-fitting and multicollinearity. For this purpose, a tuning parameter λ and a regularization term are introduced before determining the regression coefficients, which proportionally keeps the coefficients determined by minimizing the equation small and thus desensitizes the model to the test data. Note that the coefficients β_j cannot be reduced to 0 due to the RSS term. This regularization term consists of the sum of the squared coefficients. Therefore Ridge Regression and LASSO are also called shrinkage methods (cf. [18]):

$$RSS + \lambda\sum_{j=1}^{p}\beta_j^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{k}\beta_j^2 \tag{15}$$

The choice of the tuning parameter is decisive: the larger λ is chosen, the smaller the coefficients get. If they are too small, however, there is a risk of under-fitting. If $\lambda = 0$, the coefficients are the same as in a multiple linear regression. To find a suitable λ, cross-validation is usually applied. The aim is to find the best trade-off between bias and variance (cf. [18]).

LASSO regression. LASSO stands for "‘least absolute shrinkage and selection operator"’. LASSO regression is very similar to ridge regression, except that the regularization term is not the square but the absolute value of the coefficients. This leads to the fact that the coefficients of unnecessary parameters, which by the use of ridge regression at a high value of λ only result in almost zero, can really result in zero by use of LASSO regression. Thus, LASSO regression excludes unnecessary parameters from the model (cf. [18]).

$$RSS + \lambda\sum_{j=1}^{p}|\beta_j| = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{k}|\beta_j| \tag{16}$$

The optimal value for β is found in the same way as for ridge.

4.3.1. Standardization

For both Ridge and LASSO regression, the independent variables must be standardized in advance, since the squared or absolute coefficients are added up in the penalty term. If standardization is not performed, the coefficients that are large in magnitude would be shrunk first, regardless of their explanatory power. For standardization, the difference between a variable and its mean value in the data set is divided by the standard deviation. This puts variables of different magnitudes on the same scale.

$$x_j'^{(k)} = \frac{x_j^{(k)} - \mu_j}{s_j} \tag{17}$$

4.4. Regression with SARIMA errors (SARIMAX (p,d,q)(P,D,Q)$_S$(X))

As an alternative to modeling a time series y_t with a combination of only past values or as a regression model, y_t can be explained by both SARIMA and exogenous variables. In this study, the SARIMAX model is used to forecast the monthly time series using the Box-Jenkins SARIMA approach and multiple linear regression (MLR). The SARIMAX model is a SARIMA model with external variables, called SARIMAX (p,d,q) (P,D,Q)$_S$ (X), where X is the vector of external variables. The external variables can be modeled by a multiple linear regression equation expressed as follows:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \ldots + \beta_k X_{k,t} + \omega_t \tag{18}$$

where $x_{1,t}, x_{2,t}, ..., x_{k,t}$ are observations of k external variables, $\beta_0, \beta_1, ..., \beta_k$ are regression coefficients of the external variables, and ω_t is a so-called stochastic residual, i.e., the residual series is independent of the input series. The residual series ω_t can now be represented in the form of a SARIMA model after transforming equation 7 as follows:

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\mathit{\Phi}_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \tag{19}$$

The general SARIMAX equation can be obtained by substituting equation 19 into 18. It is then expressed as follows ([14]):

$$\begin{aligned} y_t =& \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} \\ &+ \left(\tfrac{\theta_q(B)\Theta_Q(B^S)}{\mathit{\Phi}_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t\right) \end{aligned} \tag{20}$$

In this case, the regression coefficients can be interpreted in the common way ([21]).

4.5. Genetic algorithm for SARIMAX

The SARIMAX described before includes the whole set of exogenous variables. To find a subset of the whole set, which might be a better forecasting model, we apply a genetic algorithm. For this we define a tuple (called chromosome) with the length of the number of the exogenous variables. The values of the genes in the chromosome can be 1 or 0. 1 indicates that the corresponding exogenous variable is included, 0 means that it is not used. The algorithm works briefly summarized as follows:

1. For the first round (called population) we start with 23 random chromosomes[3].
2. We find the optimal parameters (p,d,q)(P,D,Q) and the coefficients of equation (20).
3. We always keep the best chromosome based on the AIC and bring it directly to the next population.
4. For the missing 22 chromosomes we take the 10 best chromosomes of the last round (called parents) and cross them with each other. The first half of one chromosome comes from one parent, the second half from the other parent.
5. We then randomly mutate one gene of each of the 22 chromosomes: Change 1 to 0 or 0 to 1
6. We repeat this for 15 populations.
7. As we always take the best chromosome to the next population, we then chose the chromosome with the best AIC in the last population.

5. Comparison of revenue forecast models

5.1. Model training and optimization

The forecast focuses on the total revenue and the four most relevant product groups in terms of revenue, namely the subscriptions as well as the single, monthly and daily tickets. In the first step, the data are divided into a training and a test data set. Depending on which product group is predicted and depending on the availability of data for the relevant exogenous variables, the training data set starts from 2005 or

[3] We had 24 cores available, one for the system, so we could parallelize 23 processes.

2012. The population data is available on a monthly basis only beginning with 2012. It is a main impact factor for subscriptions and monthly tickets. Hence, the models for these two product groups are trained from 2012 until 2017 or 2018, respectively. In order to forecast 2018, the training data set ends with 2017. In order to forecast 2019, the training data is extended by 2018. The test data set is completely omitted from the training data and is used only to compare the methods (out-of-sample prediction). For the methods including exogenous variables, the real data of the exogenous variables were used for the forecasts, as these are already available for 2018 and 2019. In real forecasting situations, the actual data of the exogenous variables are not available at the time the forecast is calculated (s. 9) and are also to be forecasted.

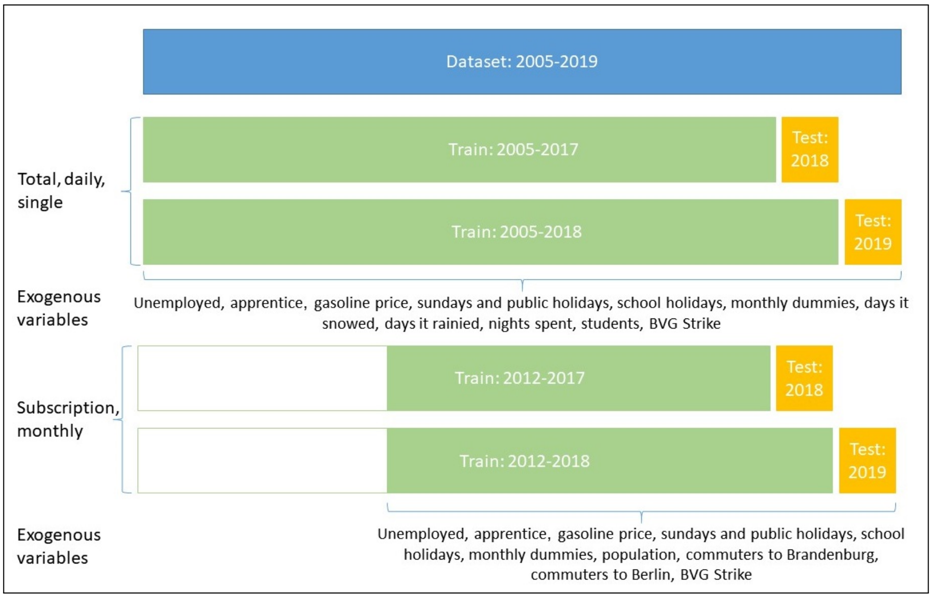

Dataset: 2005-2019
Total, daily, single
Train: 2005-2017
Test: 2018
Train: 2005-2018
Test: 2019
Exogenous variables
Unemployed, apprentice, gasoline price, sundays and public holidays, school holidays, monthly dummies, days it snowed, days it rainied, nights spent, students, BVG Strike
Subscription, monthly
Train: 2012-2017
Test: 2018
Train: 2012-2018
Test: 2019
Exogenous variables
Unemployed, apprentice, gasoline price, sundays and public holidays, school holidays, monthly dummies, population, commuters to Brandenburg, commuters to Berlin, BVG Strike

Figure 9: Set up of the different training and test horizons

In order to find the optimal parameters for the Holt-Winters, SARIMA SARIMAX models, the AIC is optimized on the training data set. To find the best set of exogenous variables for the SARIMAX, based on the AIC, a genetic algorithm was used in order to avoid iteratively testing all possibilities. To find the coefficients of the MLR, the least squares estimator was used in order to optimize the MSE of the training data set. To find the optimal alpha parameter for Ridge and LASSO in 15 and 16 k-fold cross-validation is applied to the training data. For k-fold cross-validation, the model data are divided into k parts of the same size from which $k-1$ parts are used as training set. The remaining part serves as validation set. When k estimates are done, each with a different part as validation set, the average of the errors in each validation set yields the cross-validation error. The alpha parameter is iteratively tested and the alpha corresponding to the lowest cross-validation error is selected. Using this alpha value, the coefficients are estimated for the complete training data set. Since the training data set is relatively small, a 3-fold cross-validation is applied.

5.2. Dimensions of model performance

The results of the different forecasting methods will be analyzed with regard to five criteria: scalability, accuracy, reliability, interpretability and justifiability. These five criteria are the requirements for a revenue forecasting models.

The underlying business case is to predict the monthly revenues for different products for one year in public transportation. There will be data updates during the year. In our case, new data are available every month or every quarter, depending on the data source. This is the reason why the models need to be scalable to predict the revenue of product groups with more than 400 products on a monthly basis.

Another inherent requirement is that the forecast model should be as accurate as possible. The mean absolute percentage error (MAPE) is used to evaluate the forecast accuracy on the test data set. The MAPE was chosen as the error metric, because it is easily interpretable and allows a comparison between different products of different scale. In the following, we will distinguish between the monthly MAPE (eq. 21), which is the average of the twelve monthly MAPEs of the year, and the annual MAPE (eq. 22), which is the MAPE of the actual yearly revenue and the sum of revenue predictions for a year.

$$MAPE_{month} = \frac{100\%}{12} \sum_{i=1}^{12} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{21}$$

$$MAPE_{annual} = 100\% \cdot \left| \frac{\sum\limits_{i=1}^{12} y_i - \sum\limits_{i=1}^{12} \hat{y}_i}{\sum\limits_{i=1}^{12} y_i} \right| \tag{22}$$

This distinction has two purposes: When there is no systematic under- or overestimation, the errors cancel each other out and a lower annual MAPE than a monthly MAPE serves as a first indicator hereof. The second purpose is that the annual MAPE is an indicator for an accurate yearly forecast which is needed for planning investment and cash-flows.

The three remaining criteria reliability, interpretability and justifiability are mutually dependent and important for the trust in the forecast model and, therefore, its use in practice. If one can explain how the predictions are created given the exogenous variables and if this is in line with the business understanding, then the model will be regarded as reliable.

5.3. Empirical results

				monthly					annual		
Exogenous factors	Method	Total	Single	Subs.	Monthly	Daily	Total	Single	Subs.	Monthly	Daily
without	Holt-Winters	5.64	6.41	🟩 0.49	🟩 1.82	5.83	5.72	3.69	🟩 0.35	🟩 0.09	5.50
	SARIMA	4.98	🟩 3.86	🟢 0.70	🟢 1.85	🟩 1.38	5.01	0.82	🟢 0.7	🟢 0.30	🟢 0.18
with	SARIMAX	🟩 2.72	6.31	1.12	2.43	2.51	🟩 1.66	3.35	1.10	1.40	0.21
	SARIMAX^{opt}	5.31	6.13	0.88	8.80	🟢 2.41	5.14	2.28	0.89	8.09	🟩 0.05
	MLR	3.17	4.94	2.77	2.48	5.71	2.56	🟩 0.12	2.8	1.08	5.33
	LASSO	3.17	4.91	2.77	2.57	5.59	2.56	🟢 0.19	2.79	0.13	5.20
	Ridge	🟢 2.99	🟢 4.84	2.77	2.55	5.10	🟢 2.27	1.59	2.80	0.60	4.62

🟩 most accurate method 🟢 2^{nd} most accurate method

Table 3: Monthly and yearly MAPE of revenue prediction for: all products (total), single tickets, subscriptions, monthly tickets and daily tickets for 2018

				monthly					annual		
Exogenous factors	Method	Total	Single	Subs.	Monthly	Daily	Total	Single	Subs.	Monthly	Daily
without	Holt-Winters	🟢 2.85	6.72	🟢 0.49	🟢 3.37	5.20	🟩 1.26	3.19	🟩 0.13	🟩 0.47	0.75
	SARIMA	🟩 2.41	🟩 4.36	🟩 0.31	🟩 2.80	5.49	🟢 1.29	🟩 1.70	🟢 0.19	0.70	3.04
with	SARIMAX	2.94	6.44	0.84	3.47	🟢 4.15	1.99	🟢 2.00	0.85	🟢 0.59	🟢 0.30
	SARIMAX^{opt}	3.49	7.63	1.49	4.00	🟩 4.06	2.73	6.28	1.50	1.44	🟩 0.25
	MLR	3.77	7.04	5.38	4.70	4.45	3.40	3.64	5.43	2.20	0.51
	LASSO	3.77	6.81	4.56	4.42	4.45	3.41	2.56	4.58	1.58	0.51
	Ridge	4.11	🟢 5.86	5.59	4.30	4.46	3.83	2.08	5.63	1.29	0.94

🟩 most accurate method 🟢 2^{nd} most accurate method

Table 4: Monthly and yearly MAPE of revenue prediction for: all products (total), single tickets, subscriptions, monthly tickets and daily tickets for 2019

Each of the five model performance criteria will be discussed in detail, taking into account the run-time of the models, the monthly and yearly MAPE as well as the model hyperparameters.

Scalability

The run time for all models except for the $SARIMAX^{opt}$ is less than five minutes on a Intel UHD Graphics 620 GPU. Hence, the models are scalable for more than 400 products for a monthly forecast. Training the $SARIMAX^{opt}$ model takes 12 hours, since the underlying genetic algorithm searches for an optimal feature subset. Once this feature subset is determined, it should not change significantly over time. Consequently, the $SARIMAX^{opt}$ model needs to be re-run only if there are unexpected significant changes in the underlying data (e.g. due to strikes in public transportation) which imply changes in the variables influencing the revenue. Otherwise, the optimized features and hyperparameters of the $SARIMAX^{opt}$ model can be applied in order to train a SARIMAX model for new data.

Since there are only 12 observations added per year, the scalability to large data sets is irrelevant. Hence, all models are appropriate to forecast the monthly revenue in public transportation.

Accuracy

In order to evaluate the accuracy, the monthly and annual MAPE of the forecasts for the years 2018 and 2019 will be examined. Surprisingly, the results for the years 2018 and 2019 show that most of the best-performing methods do not take into account exogenous factors. For the monthly MAPE across all product groups in 2018 and 2019, the methods without exogenous factors are the best-performing methods in 8 out of 10 times. For the monthly and including the annual MAPE across all product groups in 2018 and 2019, the methods without exogenous factors are the best-performing method in 14 out of 20 times. This is in contrast to the general opinion of experts that the revenue in public transportation is strongly driven by exogenous factors.

The better performance of methods without exogenous factors is due two aspects: the small data set and and the continuous development of the exogenous factors in this short period. Due to the small sample size, the models are not able to learn the relationship between the exogenous variables and the revenue. This is in particular apparent for the products with a smaller training set, namely the subscriptions and the monthly tickets where the difference between the MAPE for the models with exogenous variables and other models is worse than for other products. Moreover, there are little deviations from the trend and the seasonality in the historic revenues and the exogenous factors as can be seen in Table 10 and 11. This is why the models do not put emphasis on the exogenous factors but rely more on the historic revenues to learn the seasonality and the pattern.

However, for the most important category for the revenue management, the total revenue, the methods with exogenous factors, namely SARIMA and Ridge Regression, are outperforming the methods without exogenous factors in 2018. This seems to be surprising at a first sight, since one would expect the same method to perform well for the total market as well as for the individual product groups. One explanation for this effect is that the individual product groups are strongly influenced by a certain number of selected factors, whereas the total revenue is depending on various factors. In order to forecast the total market revenue accurately, the methods cannot solely rely on historic data but need to take into account exogenous factors.

In contrast to that, the two methods with the best MAPE for the total market in 2019, SARIMA and Holt-Winters, do not consider exogenous factors. However, the difference in the MAPE between the SARIMAX model and the best-performing SARIMA model for the total market revenue in 2019 is rather small. In addition to that, there have been changes in the revenue accounting that affected the revenue data in January and December in 2019. Since these accounting changes did not occur in the historical data, the models with exogenous factors are not able to learn this accounting shift. In these two months, the predictions of the SARIMAX model deviate more from the real values than those of the SARIMA model. The reason for the slightly worse performance of the SARIMAX model can thus be explained by unprecedented changes in the accounting and should not distort the evaluation of the prediction accuracy for the SARIMAX model.

For the other product groups, the methods without exogenous factors, especially SARIMA, are performing better in both years. Nevertheless, the value of the monthly MAPE in 2019 for the best and second-best method has increased substantially for the single, daily and monthly tickets. This increase in the MAPE is due to the aforementioned accounting shifts that affected the single, daily and monthly tickets. Another reason for the higher monthly MAPE for the monthly tickets is that neither the methods with or without exogenous factors were able to predict the higher revenue in August 2019, because the school holidays already ended in the beginning of August. With the school starting early in August, there are less people on vacation and thus more people buy a monthly ticket. The methods without exogenous factors are not capable of learning this relationship because in the previous years from 2014 to 2018 there have been at least has been 3 weeks of school holidays in August and thus a decline in revenue, too. The methods with exogenous factors could potentially learn the increase in monthly ticket revenue in August 2019 when there are only a few of school holidays. Due to the small training set, the model could not learn the true relationship although monthly dummies and school holidays were used as exogenous variables.

Taking everything into account, the SARIMAX model seems to be a promising model that predicts the total revenue with a satisfying MAPE of 2-3 %. For the individual product groups, the SARIMA model overall achieves the best results across all product groups.

Reliability, interpretability and justifiability

The interpretability and thus the ways of justifying and creating reliability differs between methods with and without exogenous factors. The methods used with exogenous factors allow for a direct quantification of the effects on revenue resulting from the exogenous factors. For the methods without exogenous factors, this is not possible. For these methods, the models can only be justified by interpreting how much emphasis they put on recent past values and whether they assume a trend in the data. Therefore, the choice of the model also depends on the desired degree of interpretability. In the case of the revenue forecasting at the BVG, the model for the total market revenue should be interpretable and identify the impact factors. For the individual product groups the interpretability of the exogenous factors is not required. Therefore, the focus is on justifiable results for the total market. The detailed results for the individual product groups are listed in the appendix.

As a general result, the better-performing models match with the business understanding. For the total market, we observe similar coefficients in value and sign for the exogenous factors. These coefficients are also reasonable. One example here is the coefficient for the exogenous factor "‘students"’ (see Table 5 and 7). To predict the total market revenue in 2018, the most accurate model SARIMAX states that each additional student results in 244.82 additional revenue. This is almost in line with the price for a semester ticket for students which costs 387.60 per year in 2018 ([25]). Further expected relations such as a negative effect of Sundays, public holidays, and school holidays as well as a positive effect of the overnight stays can also be observed for all methods with exogenous factors.

For the methods without exogenous factors, namely SARIMA and Holt-Winters, the interpretation is different. The SARIMA models for 2018 and 2019 only differ in the 24 months moving average term which is considered in the model for 2019. However, the coefficient is not statistically significant different from zero. Thus, taking the model with the best AIC leads to similar models across years. This is justified since there are no significant changes between the years which could be explained by seasonal variations in the training set. At first glance, the SARIMA models seem reliable since they account for seasonal fluctuations. This is not the case for the trend. Due to the upward trend of the total market revenues, one would expect that a drift with a positive coefficient is included in the SARIMA models. However, this is neither the case for 2018 nor 2019. Instead, the models learn the upward trend by adding up the revenues from the previous months and the revenue from 12 months ago. Then, this sum is adjusted downwards by the moving average terms. An interesting observation is that a SARIMA model with drift is one of the three best models based on the AIC optimization for both years. When applying these SARIMA models with drift, the MAPE improves compared to the SARIMA without drift for the years 2018 and 2019. Detailed results are included in the appendix.

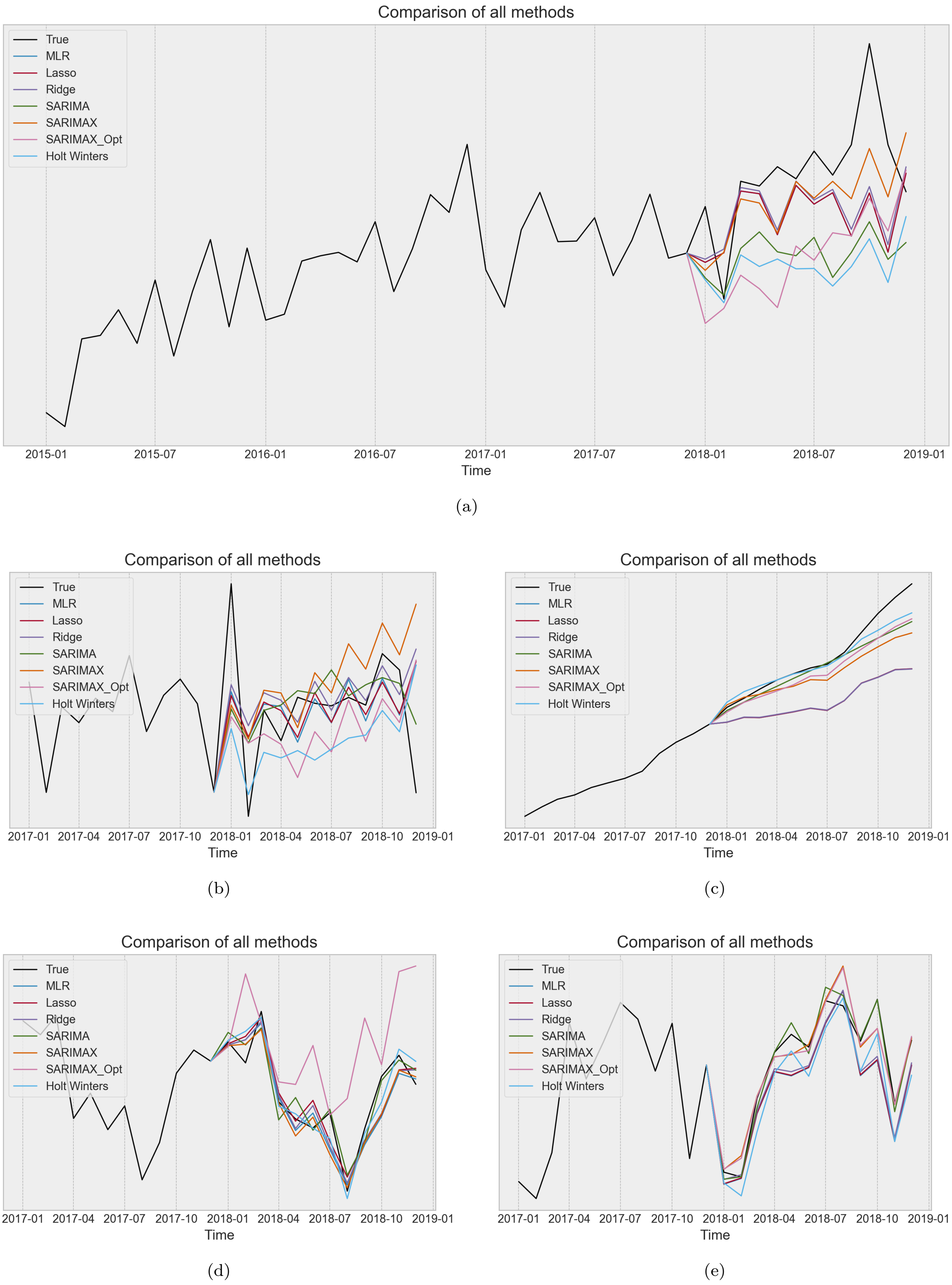
In contrast to the SARIMA models, there are clear differences between the best models resulting from the AIC optimization. In 2018, the Holt-Winters model is an additive model compared to a multiplicative model in 2019. It seems not clear if trend and seasonality in the revenues will develop in a linear or exponential way. Having two substantially different models in two subsequent years does not help to gain acceptance of an automated optimized model selection. Similar to the SARIMA models, the Holt-Winters model with the best AIC does not have the best MAPE. Detailed results are included in the appendix.

Taking everything into account, the AIC optimization does not result automatically in reliable and justifiable models. One reason is that the time series of the total market revenue is rather complex and subject to seasonal variations. In order to model this complexity properly, more parameters are necessary. Moreover, the AIC penalizes bigger models. In this case, this penalty might not be useful and models with more parameters will not necessarily lead to overfitting. We observe that models with more parameters are able to learn the relationship in our case study. This is supported by the fact that models matching with the business understanding, such as SARIMA models with a drift, generate more accurate predictions. We see a correlation between justifiability and accuracy in our case study. This also holds true for models which are less accurate. For example, the models with exogenous factors are less accurate for the monthly tickets. This lack of accuracy is accompanied by a lack of justifiability. For example, the models learn a negative relationship between the number of inhabitants of and commuters to Berlin and the revenues from the monthly tickets. This negative relation is not expected and not in accordance with the business understanding, since in general more inhabitants and commuters should generate more revenue.

5.4. Recommendations

Our work shows that data preprocessing is a crucial task when it comes to improving the prediction accuracy. An extensive data check with the help of business insights is necessary to identify data inconsistencies and deal with missing values. Another crucial step is identifying and finding data for relevant exogenous factors. In predictive analytics projects, we recommend to plan at least half of the project duration for these tasks. Before prediction methods are applied, the overall purpose of the prediction needs to be determined, i.e., whether the prediction should focus on accuracy only or should be interpretable or scalable to large data sets and daily forecasts. This limits the choice of suitable methods.

The recommendation for other public transport companies is to test first SARIMAX models which have shown promising results in our study. Especially when it comes to cost-to-benefit considerations, the SARIMAX models have favorable features. They are inherently interpretable, the code is easy to implement and to maintain, and SARIMAX models allow for a scenario analysis by adapting the values of the future exogenous factors to the underlying scenarios.



(a)

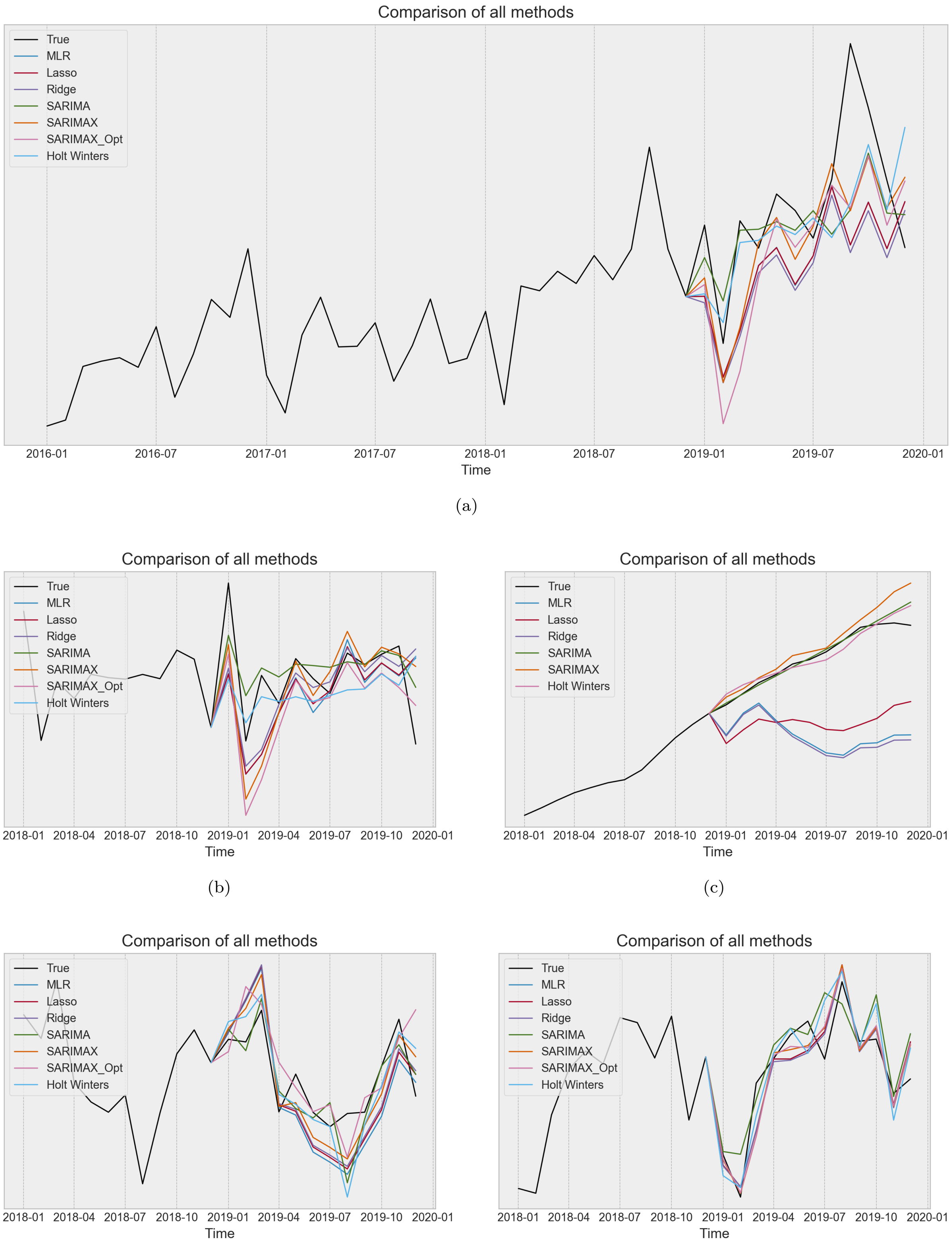
(b)

(c)

(d)

(e)

Figure 10: Forecast for 2018: (a) total revenue (b) single tickets (c) subscriptions (d) monthly tickets (f) daily tickets



(a)

(b)

(c)

(d)

(e)

Figure 11: Forecast for 2019: (a) total revenue (b) single tickets (c) subscriptions (d) monthly tickets (f) daily tickets

Appendix A: Detailed computational results for the best selected models

Factors	SARIMA	SARIMAX	SARIMAX^{opt}	MLR	Ridge	LASSO
	(1,0,1)x(1,0,1,12)	(1,0,2)	(0,1,1)x(0,0,[1,2,3,4],12)			
ar.L1	0.9935***	0.9832***				
ar.S.L12	0.7716***					
ar.S.L24						
ma.L1	-0.7193***	-0.6358***	-0.6609***			
ma.L2		-0.0667				
ma.S.L12	-0.3312***		0.0875			
ma.S.L24		61.7840	-0.0091			
ma.S.L36			-0.5212***			
ma.S.L48			-0.1583			
sigma2	2.149e+12***	9.157e+11***	7.507e+11***			
drift			-549.4846			
Intercept	6.581e+04***	1.932e+05***	1.614e+05**	61037443.78	61037445.81	56046015.70
numbers of unemployed		36.1531***	41.1568**	-27.32	-27.32	-29.34
apprentices		20.3925		-488.04	-488.04	-432.53
monthly mean gasoline price		-6252.3588		-85648.12	-85648.12	-82146.12
sundays and public holidays		-3.21e+05***	-5.536e+04	-347950.30	-347950.27	-337671.41
school holidays		-9.016e+04***	-8.999e+04***	-96366,11	-96366.10	-91585.41
January		1.438e+06***		200537.24	200536.99	248486.76
February		5.917e+05***		-645363.72	-645363.88	-576423.82
March		1.356e+06***	6.156e+05***	565482.72	565482.56	587512.76
April		1.388e+06***	4.961e+05**	1157097.29	1157097.12	1097322.64
May		6.501e+05***		624199.31	624199.17	570923.08
July		1.181e+06***	1.037e+06***	1398359.20	1398359.04	1255777.38
August		2.703e+05***		404033.28	404033.15	285647.13
September		3.289e+05***		-980347.06	-980347.11	-928467.39
October		1.183e+06***	1.854e+06***	821753.84	821753.69	743370.68
November		-1.255e+04***	7.387e+05*	-822447.43	-822447.54	-811179.88
December		4.017e+06	3.98e+06***	3133972.68	3133972.44	3078936.69
days with snow		6.937e+04***	7.678e+04***	67505.73	67505.73	65214.53
days with rain		2618.0083		-2780.42	-2780.41	-2651.85
students		244.8205***	120.9969**	141.79	141.79	156.19
overnight stays		4.1538***	3.4009***	3.60	3.60	3.62
BVG strike		-5.171e+04***	-6.61e+04***	-40279.26	-40279.26	-41673.37

*** -> 1%, ** -> 5%, * -> 10%,

Table 5: Coefficients for the total market revenue predictions 2018

Factors	SARIMA	SARIMAX	SARIMAX^{opt}	MLR	Ridge	LASSO
	(1, 0, 1)x(1, 0, [1, 2], 12)	(1, 0, 1)x(1, 0, [], 12)	(0,1,1)x(0,0,[1,2,3],12)			
ar.L1	0.9983***	0.7389***				
ar.S.L12	0.8931***	0.1282				
ar.S.L24						
ma.L1	-0.7328***	-0.5175**	-0.7651***			
ma.L2						
ma.S.L12	-0.4805***		0.2259***			
ma.S.L24	-0.1149		0.1280			
ma.S.L36			-0.1540			
sigma2	2.134e+12***	1.168e+12***	1.056e+12***			
drift						
Intercept	8218.6752***		1.277e+05***	33862408.1	33849911.99	40055319.21
numbers of unemployed		-31.7348***		-34.25	-34.22	-34.51
apprentices		126.2642*	38.8637**	-207.81	-208.45	-269.85
monthly mean gasoline price		-5.369e+04***		-71217.70	-71039.77	-72894.84
sundays and public holidays		-4.15e+05***	-1.991e+05**	-388747.68	-382881.59	-366055.56
school holidays		-8.12e+04***	-9.169e+04***	-91812.22	-90630.01	-89521.19
January		2.802e+06***		1082117.65	1042096.77	736156.70
February		1.332e+06***	-1.248e+06***	-201895.65	-224236.63	-486679.96
March		1.685e+06***		948890.31	920747.12	779941.02
April		1.526e+06***		1314497.03	1276859.09	1186004.84
May		8.68e+05		901405.86	867494.59	822272.99
June			-3.724e+05			
July		7.383e+05*	8.847e+05	1396639.92	1354337.88	1372353.61
August		-4.38e+05		311635.74	275367.34	325403.71
September		-2.012e+05		-466013.30	-474119.49	-540191.00
October		1.097e+06***	1.825e+06***	1210507.81	1175180.93	1142704.19
November		6.793e+05*		-314937.93	-332248.00	-493377.45
December		4.488e+06***	3.365e+06***	3498314.92	3452434.07	3218992.96
days with snow		5.742e+04***	6.617e+04***	53612.15	53443.55	54097.87
days with rain		-1.068e+04		-21256.12	-20774.14	-22431.62
students		291.8190***	150.9156***	224.78	224.99	210.09
overnight stays		6.0122***	2.9379***	3.97	3.95	3.64
BVG strike		-4.754e+04***	-5.198e+04***	-41526.82	-41580.09	-41235.70

*** -> 1%, ** -> 5%, * -> 10%,

Table 6: Coefficients for the total market revenue predictions 2019

Factors	SARIMA		SARIMAX		SARIMAX^{opt}		MLR		Ridge		LASSO	
	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019	2018	2019
	(1,0,1)x(1,0,1,12)		(1,0,2)									
ar.L1	0.9935***		0.9832***									
ar.S.L12	0.7716***											
ar.S.L24												
ma.L1	-0.7193***		-0.6358***									
ma.L2			-0.0667									
ma.S.L12	-0.3312***											
ma.S.L24			61.7840									
sigma2	2.149e+12***		9.157e+11***									
drift												
Intercept	6.581e+04***		1.932e+05***				61037443.78***		61037445.81		56046015.70	
numbers of unemployed			36.1531***				-27.32***		-27.32		-29.34	
apprentices			20.3925				-488.04***		-488.04		-432.53	
monthly mean gasoline price			-6252.3588				-85648.12***		-85648.12		-82146.12	
sundays and public holidays			-3.21e+05***				-347950.30***		-347950.27		-337671.41	
school holidays			-9.016e+04***				-96366,11		-96366.10***		-91585.41	
January			1.438e+06***				200537.24		200536.99		248486.76	
February			5.917e+05***				-645363.72		-645363.88		-576423.82	
March			1.356e+06***				565482.72		565482.56		587512.76	
April			1.388e+06***				1157097.29**		1157097.12		1097322.64	
May			6.501e+05***				624199.31		624199.17		570923.08	
July			1.181e+06***				1398359.20		1398359.04		1255777.38	
August			2.703e+05***				404033.28		404033.15		285647.13	
September			3.289e+05***				-980347.06		-980347.11		-928467.39	
October			1.183e+06***				821753.84		821753.69		743370.68	
November			-1.255e+04***				-822447.43		-822447.54		-811179.88	
December			4.017e+06				3133972.68		3133972.44		3078936.69	
days with snow			6.937e+04***				67505.73		67505.73		65214.53	
days with rain			2618.0083				-2780.42		-2780.41		-2651.85	
students			244.8205***				141.79		141.79		156.19	
overnight stays			4.1538***				3.60		3.60		3.62	
BVG strike			-5.171e+04***				-40279.26		-40279.26		-41673.37	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Coefficients for the total market revenue predictions 2018 and 2019

References

[1] Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723. doi:10.1109/TAC.1974.1100705.

[2] Alonso, G., Villaverde, A., Quan, A., Ruiz, M., 2021. Comparison between sarima and holtwinters models for forecasting monthly streamflow in the western region of cuba. doi:10.1007/s42452-021-04667-5.

[3] Amt für Statistik Berlin-Brandenburg, a. Auszubildende in Berlin 1991 bis 2020 nach Ausbildungsjahr. URL: `https://www.statistik-berlin-brandenburg.de/`. last visited 2022-05-04.

[4] Amt für Statistik Berlin-Brandenburg, b. Statistischer Bericht / A / I / 7 / A / II / 3 / A / III / 3 : Bevölkerungsentwicklung und Bevölkerungsstand in Berlin. URL: `https://www.statistischebibliothek.de`. last visited 2022-05-04.

[5] Arnold, M., Hajos, B., Busch, T., 2013. Long term transport demand and financial forecast for a large scale regional public transport network in germany. Economics .

[6] Arunraj, N., Ahrens, D., Fernandes, M., 2016. Application of sarimax model to forecast daily sales in food retail industry. doi:10.4018/IJORIS.2016040101.

[7] Athanasopoulos, G., Hyndman, R.J., 2021. Forecasting: Principles and practice. URL: `https://otexts.com/fpp3/`.

[8] Box, G., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. Prentice-Hall.

[9] Brockwell, P.J., Davis, R.A., 2009. Time series: Theory and methods.

[10] Brown, R.G., 1956. Exponential smoothing for predicting demand.

[11] Brown, R.G., 1963. Smoothing forecasting and prediction of discrete time series.

[12] Bundesagentur für Arbeit, . Pendlerverflechtungen der sozialversicherungspflichtig Beschäftigten nach Kreisen - Deutschland (Jahreszahlen). URL: `https://statistik.arbeitsagentur.de`. last visited 2022-05-04.

[13] Chudy-Laskowska, K., Pisula, T., 2017. Seasonal forecasting for air passenger trafic. doi:10.5593/sgemsocial2017/14/S04.089.

[14] Cools, M., Moons, E., Wets, G., 2009. Investigating the variability in daily traffic counts through use of arimax and sarimax models: Assessing the effect of holidays on two site locations. Transportation Research Record 2136, 57–66. doi:10.3141/2136-07.

[15] Deutscher Wetterdienst (DWD), . Climate data center (cdc). URL: `https://cdc.dwd.de/portal/`. dWD Climate Data Center (CDC): Historische stündliche Stationsmessungen der Niederschlagshöhe für Deutschland, Version v21.3, 2021.

[16] Google Trends, . Bvg strike / corona. URL: `https://trends.google.com/trends/`. last visited 2022-05-04.

[17] Hamilton, J.D., 1994. Time Series Analysis. 1 ed., Princeton University Press.

[18] Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

[19] Holt, C.C., 1957. Forecasting trends and seasonal by exponentially weighted averages.

[20] Holt, C.C., 2004. Forecasting trends and seasonal by exponentially weighted averages. doi:10.1016/j.ijforecast.2003.09.015.

[21] Hyndman, R.J., 2010. The arimax model muddle. URL: `https://robjhyndman.com/hyndsight/arimax/`.

[22] Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice. Otexts, Melbourne, Australia. URL: `https://otexts.com/fpp3/`.

[23] Jere, S., Banda, A., Kasense, B., Siluyele, I., Moyo, E., 2019. Forecasting annual international tourist arrivals in zambia using holt-winters exponential smoothing. doi:10.4236/ojs.2019.92019.

[24] schulferien.org, . Schulferien Berlin, Deutschland. URL: `https://www.schulferien.org/deutschland/ferien/berlin`. last visited 2022-05-04.

[25] asta.tu-berlin.de, . Ticket price. URL: `https://asta.tu-berlin.de/semtix/ticket-preise/`. last visited 2022-05-19.

[26] Statistisches Bundesamt (Destatis), a. 21311-0005: Studierende: Bundesländer, Semester, Nationalität, Geschlecht. URL: `https://www-genesis.destatis.de`. last visited 2022-05-04.

[27] Statistisches Bundesamt (Destatis), b. 45412-0025: Ankünfte und Übernachtungen in Beherbergungsbetrieben: Bundesländer, Monate. URL: `https://www-genesis.destatis.de`. last visited 2022-05-04.

[28] Su, E., Su, O.A., 2021. Public transport demand forecast using arima: The case of istanbul. URL: `https://ssrn.com/abstract=4020644`, doi:10.2139/ssrn.4020644.

[29] Tsai, C.H., Mulley, C., Clifton, G., 2014. Forecasting public transport demand for the sydney greater metropolitan area: A comparison of univariate and multivariate methods.

[30] Tsekeris, T., Tsekeris, C., . Demand forecasting in transport: Overview and modeling advances. Economic Research 24, 82–94. doi:10.1080/1331677X.2011.11517446.

[31] Vagropoulos, S., Chouliaras, G., Kardakos, E., Simoglou, C., Bakirtzis, A., 2016. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. doi:10.1109/ENERGYCON.2016.7514029.

[32] Vimal, M.R., Naseem, S.M.B., 2020. Time series analysis: Forecasting with sarimax model and stationarity concept. URL: `http://www.jetir.org/papers/JETIREJ06034.pdf`.

[33] Winters, P.R., 1960. Forecasting sales by exponentially weighted moving averages. doi:10.1287/mnsc.6.3.324.