




# Using Machine Learning to Predict POI Occupancy to Reduce Overcrowding

Jessica Bollenbach <sup>1</sup>, Stefan Neubig <sup>2</sup>, Andreas Hein<sup>3</sup>, Robert Keller <sup>4</sup> and Helmut Krcmar<sup>5</sup>


**Abstract:** Due to the rapid growth of the tourism industry, associated effects like overcrowding, overtourism, and increasing greenhouse gas emissions lead to unsustainable development. A prerequisite for avoiding those adverse effects is the prediction of occupancy. The present study elaborates on the applicability and performance of various prediction models by taking a case study of beach occupancy data in Scharbeutz, Germany. The case study compares different machine learning models once as supervised machine learning models and once as time series models with a persistence model. XGBoost and Random Forest as time series demonstrate the most accurate prediction, followed by the supervised XGBoost model. However, the short prediction span of time series models is a disadvantage for longer-term visitor management to avoid the explained unsustainable effects through steering measures, so depending on the use case, the XGBoost model is to be favoured.


**Keywords:** Beach Occupancy, Time series Forecast, XGBoost, Random Forest, Support Vector Regression, SARIMA, Tourism Demand.

## 1 Introduction


In recent decades, the international tourism sector has experienced significant growth from 25 million cross-border arrivals in 1950 to more than 1.3 billion arrivals in 2017, which is expected to continue up to 1.8 billion tourists in 2030 [Wo18]. Despite all the benefits, this upward trend is also driving unsustainable tourism, with effects such as overtourism, local overcrowding, and increasing greenhouse gas emissions due to the close link of tourism and mobility [Ca19], [Hø00]. Most leisure trips are made by greenhouse gas emitting cars, as touristic points of interests (POIs) in rural areas are often poorly accessible with public transport and people seek self-determination in their leisure time [Ei22], [GG18]. Overcrowding effects at the POI increase these emissions due to the

---

<sup>1</sup> University of Applied Sciences Kempten, WTZ, Bahnhofstraße 61, 87435 Kempten (Allgäu), & Fraunhofer Institute for Applied Information Technology FIT, Augsburg, jessica.bollenbach@hs-kempten.de,   
<https://orcid.org/0000-0001-9554-1640>

<sup>2</sup> Technical University of Munich, Boltzmannstr. 3, 85478 Garching, & Outdooractive AG, Missener Str. 18, 87509 Immenstadt, stefan.neubig@tum.com, <https://orcid.org/0000-0002-3794-7260>, 

<sup>3</sup> Technical University of Munich, Boltzmannstr. 3, 85478 Garching, andreas.hein@tum.de

<sup>4</sup> University of Applied Sciences Kempten, WTZ, Bahnhofstraße 61, 87435 Kempten (Allgäu), & Fraunhofer Institute for Applied Information Technology FIT, Augsburg, robert.keller@hs-kempten.de,   
<https://orcid.org/0000-0001-7097-1724>

<sup>5</sup> Technical University of Munich, Boltzmannstr. 3, 85478 Garching, helmut.krcmar@tum.de

increased chance of congestion and long search time for a parking lot [Pa22]. In addition, a lack of parking lots at overcrowded POIs lead to environmentally harmful illegal parking. Further, the mass of people at overcrowded POIs are responsible for pollution, noise, and wildlife disturbance. Such overcrowding effects occur especially in free public spaces like a beach, lake, city center, or mountain trail, as access cannot be controlled since no closure during peak occupancy occurs. Further, no final and predefined maximum occupancy exists, unlike, for example, in a parking garage. This leads to uncontrolled peak occupancy which is unsustainable and unpleasant for any visitor. To establish a way to more sustainable tourism, Schmücker et al. [Sc22] propose the demand for an active visitor management (AVM) system which guides visitors to more sustainable behavior. For example, steering measures enabled by an AVM system target the even distribution of visitors between open accessible POIs to reduce overcrowding or target the controlled filling of parking lots to reduce congestion and search time. A key element of such an AVM system is the prediction of POI occupancy because only with this knowledge can effective and target-oriented steering measures be initiated. POI occupancy prediction, or more generally, tourism demand prediction, is a well-established research field [WSS17]. However, existing and utilized data often lacks information on day tourism and tends to focus on overnight stays, thus creating an incomplete picture [Ne22]. Furthermore, the considered geographical areas are mainly larger regions [YZ19] or smaller, mostly closed areas like parking lots [APB21] and lack the examination of (semi-) open terrain of a specific POI. Hence, occupancy prediction models need to be developed for several days in advance to enable steering measures in an AVM system. These models need to cover a fine-grained temporal observation of a (semi-) open accessible terrain, including day visitors and overnight stays, which is, to the best of our knowledge, currently a research gap.

In this work, we evaluate different prediction model types, including various machine learning (ML) models, to predict POI occupancy of a local, semi-open terrain. The development of the prediction model is based on a use case study with occupancy data collected via sensors of the Bay of Lübeck in Scharbeutz at the Baltic Sea in northern Germany. Developing such a prediction model is crucial for an AVM system with steering measures to enable more sustainable tourism and mobility. We contribute to both theory and practice by evaluating different prediction models concerning their accuracy and elaborating about how this prediction can be used as an effective countermeasure to overcrowding and unsustainable tourism.

## 2 Related Work

Research in tourism occupancy prediction has grown rapidly since 2006 [Li19]. In the following, we provide an overview considering the perspectives of (i) data sources, (ii) prediction models, and (iii) spatiotemporal granularity.

**Data Sources:** Traditionally, occupancy prediction was conducted with one-dimensional

historical data capturing a certain timespan. Such data can be collected manually (e.g., based on surveys) or automatically (e.g., based on cameras, smartphones, or advanced sensors) [APB21]. Suitable data sources include tourist arrivals (e.g., entries to national parks [Ab21] or countries [Ki21]), historical parking data [Ch19], web traffic [PY17], booking data (e.g., historical room allocation in hotels [PS21], [Zh18]), and payment data [APB21]. In order to improve predictions, several studies added supplementary information, such as hotel room prices [TB20], as well as weather and holiday information [Bi21]. Considering the importance of online behavior of potential visitors [GV20], recent research addresses the use of behavioral online data [WSS17], such as search engines and Google trends [Vo19], [Di19], [BL19], [Fe19], online reviews [Hu22], Facebook [ÖGG20], and sentiments [ÖGS19].

**Prediction Models:** Previous research [SQP19], [JC19], [WSS17] classifies occupancy prediction algorithms in the three categories (i) time-series models, (ii) econometric and statistical models, and (iii) ML-based models. Time series models remain the most widely used techniques [MR16], although ML-based technologies have proliferated significantly in recent years. However, ML models such as gradient boosting with XGBoost, which have shown promising results in other applications like taxi demand [Va18], bus demand [SSD21], or parking [APB21], are still rarely used in POI occupancy prediction.

**Spatiotemporal Granularity:** Regarding time and space, occupancy predictions address different levels of granularity. Considering their temporal resolution, existing approaches address seasonal [Ch15], to monthly [BL19], weekly [Zh18], [PY17], daily [TB20], [Bi21], [Ki21], [PS21], or even hourly [ZHL21] prediction. Prilistya et al. [PEF20] observed that monthly granularity is the most widely used data frequency in tourism occupancy prediction, but due to the more widespread use of ML, Jiao and Chen [JC19] observed a trend towards smaller time granularities. In terms of spatial resolution, approaches range from predicting tourist arrivals in countries [Ki21], regions [YZ19], or cities [Cl20], [TB20] to predicting occupancy of smaller areas. However, small areas usually do not refer to open or semi-open POIs like a beach, but rather address closed environments with well-defined entrances, such as parking lots [APB21], [Ch19], [ZZ20] or hotel rooms [PS21], [Zh18].

Overall, previous research rarely focuses on semi-open POIs to predict overcrowding situations in an AVM system [Kh20]. Further, the literature review indicates advanced ML models like XGBoost are not yet applied and compared with other approaches to predict touristic POI occupancy. Our work relies on real-world sensor data to predict POI occupancy of an exemplary German destination with a fine-grained temporal prediction up to ten days in advance. The comparison of a classical time series model with various ML models of different types illustrates the use case specific usability of each model. We discuss the results of the models regarding their application as a countermeasure to overcrowding and unsustainable tourism.

### 3 Methodology and Case Study

The methodology to develop an occupancy prediction model is based on the Data Science Trajectories (DST) map by Martinze-Plumed et al. [Ma19], which is a more sophisticated version of the classic CRISP-DM approach [Ch00]. The DST-map is a more flexible approach to data science projects that integrates further steps regarding exploratory and data management activities around the CRISP-DM model.

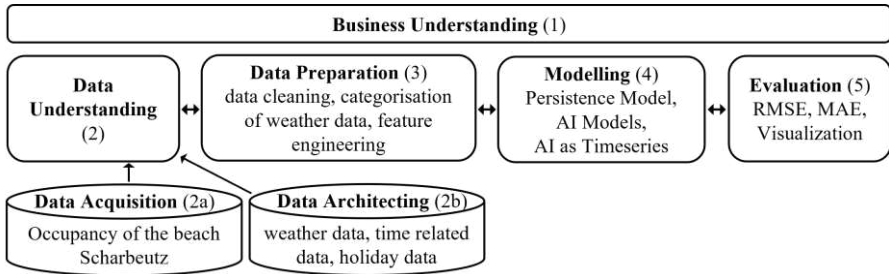


Fig. 1: The DST Map illustrates all steps within this study from business understanding to evaluation

Fig. 1 presents the DST map with the individual adapted steps for this study. In the Business Understanding part, we first provide an overview of the case study objective, followed by the examination of the utilized data. We divide the data understanding part in the acquisition of occupancy data and the architecting of external influential data which are both prepared cyclically in steps two and three. Subsequently the fourth step presents different model types with their inherent ML models, followed by the evaluation approach.

#### 3.1 Business understanding: Predicting the occupancy of the beach in Scharbeutz

To develop a prediction model for the occupancy of a POI, we conduct a case study of the Bay of Lübeck in Scharbeutz at the Baltic Sea in northern Germany. The Baltic Sea is the most popular tourist destination of the German population with a continuously increasing trend [Ar22]. This trend was reinforced by the increase in domestic travel due to the Covid-19 pandemic [In21]. The high proportion of day visitors, accounting for around  $\frac{2}{3}$  of the total visitors, has the effect that relatively scarce information is available on occupancy leading to unpredictable peaks in beach occupancy [De19], [De16]. To close the existing information gap, sensors have been installed at the entrances to the beach to count people entering and leaving the beach. The current state of data utilization is the representation of the current occupancy of the beach, which does not yet include a future prediction [To22]. However, for an active visitor management, a prediction, especially about occupancy peaks, is necessary to initiate possible steering measures in advance. Steering measures can be roughly divided into short-term measures like for example the selection of another parking lot and medium-term measures such as choosing another POI. Therefore, the different objectives of steering measures require different prediction spans

and accuracy. In the following case study, we use the collected occupancy data of the beach in Scharbeutz to develop, evaluate, and discuss the applicability and performance of different prediction models.

### 3.2 Data understanding and data preparation

The data acquisition phase covers the collecting of occupancy data of the beach in Scharbeutz which is also available as open data [Mi22]. The raw data contains quarter-hourly summed numbers of people entering and leaving the beach over the entire Scharbeutz beach section from 19-08-2020 to 23-03-2022. This sum of people entering and leaving is intended to represent the occupancy of the beach. A first analysis reveals annual seasonality with higher occupancy in the summer months and a daily seasonality between day and night. Further, some days demonstrate a high and irregular peak occupancy. Since the entrances of the adjoining beach sections are not equipped with sensors, the absolute number of people at the beach is probably not exact, but the general trends and peak occupancies can be identified. Data preparation of the occupancy data as the target value includes elimination of missing values and aggregation to larger time periods. To eliminate gaps in the time series due to missing data, we created a continuous time series dataframe and filled the small amount of missing data with zeros. Since the fine-grained partitioning into 15-minute time periods cause increased noise, we aggregated the data to larger 6 hour time periods per day, starting at midnight (see Tab. A.1).

The data architecting phase covers researching and compiling of possible factors influencing the target variable of occupancy. We identified three areas of possible influencing factors: time-related features, holiday-related features, and weather-related features. The features' data are required historically and in the near future, as historic data is necessary to train and test the prediction models, whereas future data is required for a real-world application of the developed prediction model. The time-related features are created to include and identify the possible timing of seasonality in the prediction model as an input variable. The holiday-related features are included as a possible reason for the peak demand since the population's leisure time should have an influence on the beach occupancy. Due to the high proportion of day visitors who presumably live near the considered POI, school holiday of Schleswig-Holstein is a specific feature. Since the Baltic Sea is a popular destination for domestic travel, the general school holiday density in Germany is covered as a feature to may represent increased occupancy by overnight visitors [Ar22]. Since visiting a beach is an outdoor-only activity, we assume that the occupancy is highly dependent on the weather. The collected weather data is from the open data hub of the German Weather Service which provides historical data as well as weather forecasts up to 10 days [Ge22b]. Similar to the occupancy data, we aggregated the raw weather data to 6-hour time periods. Following Studer et al. [St21], we converted the continuous raw data values into categorical features to ensure a good performance of the prediction models and to ensure that we use the same input features for each model. The categorization of the weather data follows the standards of the German Weather Service such as the classification of temperature, precipitation intensity, and wind class

[Ge22c], [Ge22a], [Ge22d] (see Tab. A.2). Precipitation form is not directly available for the historical and future weather data which is why we created the feature via feature engineering by combining input variables (see Tab. A.3). After categorizing the four weather-related features of wind, rain, temperature, and precipitation form, we transformed the features from categorical to binary features using one-hot encoding [St21]. Tab. 1 illustrates all input features, for the structure of the final input dataset see Tab. A.4.

<b>Time-related</b>	<b>Holiday</b>	<b>Weather</b>
year	weekend	wind category
month	bridging day	rain category
day of year	public holiday (bank holiday)	temperature category
quarter of year	regional school holiday	precipitation form
day of month	German school holiday density	
calendar week		
day of week		
hour		

Tab. 1: Overview of the features after categorization and feature engineering

### 3.3 Modelling

In this case study, we aim to identify a model for predicting the occupancy of the beach in Scharbeutz and therefore consider four different model types. The first model type (1) is a simple persistence model whose results are used as a baseline. It simply takes a historical value with a predefined lag as the prediction. To obtain a reasonably good baseline, we tested lags from one time step up to 1460 time steps, which equals exactly one year. The second type are supervised ML models (2) that aim to predict the target variable  $y$ , here beach occupancy per time period, based on several input features  $X$  [JM15]. Tab. 2 illustrates a brief overview of the ML models. We implemented the ML models in Python using the scikit-learn packages, and we avoided excessive tuning of the models to allow for an even comparison. The third model type are ML time series models (3), which merely consider the historical occupancy data as an input variable for prediction [Ki10]. For comparison, we use the same supervised ML models as in the second model type. To apply ML models as time series prediction models, we first transformed the time series into a supervised learning model. We created the required input features by shifting the occupancy data backwards by a predefined lag. To improve the prediction quality, 10 input features were created by backward shifting the data with a lag from 1 to 10. The target variable  $y$  remains the original occupancy of the beach data. Prediction takes place with a walk forward validation, where exactly one time step ahead is predicted, and the observed values are added to the training data set over time. The fourth model type is a classical approach (4) with the statistical, Seasonal Autoregressive Integrated Moving Average model (SARIMA( $p, d, q$ )( $P, D, Q, m$ )) [HX98], [LI10]. The parameters of the SARIMA model were determined using the automatic optimization of pmdarima with the seasonal

parameter set to one day corresponding to four timesteps, resulting in final parameters of (0,1,1)(0,1,1,4) [Sm22]. Analogous to the ML models as time series, we performed the prediction with SARIMA with a walk forward validation.

ML model	Short description	Source
Extreme Gradient Boosting (XGBoost)	The XGBoost model is a system optimized implementation of gradient boosting, whereby we included early stopping to avoid overfitting. Weak learners of decision trees are gradually improved and merged to obtain a strong one as the final prediction model.	[Fr01], [xg22]
Random Forest	The Random Forest model generates various independent trees by randomly selecting a choice of features and averaging over the results for prediction. Here, a maximum depth of four levels was used.	[Al12], [sc22a]
Support Vector Regression (SVR)	The SVR model creates an inherent function for predictions based on a kernel function by penalizing extreme values that are not within a specified radius. By testing the different kernels, the linear kernel indicated the best overall results here.	[AK15], [Va00], [sc22b]

Tab. 2: Overview of the considered ML models

### 3.4 Evaluation

To evaluate the performance of the different models, we split the occupancy data into a training and a testing data set, whereby the training data set includes the target variable to fit the model. With the test data set, the fitted model generates predictions which are then compared to the expected target values. In this case study, we split the time series data on the 1<sup>st</sup> of November 2021 which results in a training data set of 74% of the overall data set and a testing data set of 26%. With the single split, we aim to enable an even comparison of the different models and model types. To compare the predicting performance of the various models, we used the performance metrics root mean squared error ( $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$ ) and mean absolute error ( $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$ ). However, since in this case study especially the prediction of days with peak occupancy rates are relevant, a graphical interpretation is included as well.

## 4 Results

Tab. 3 illustrates the performance metrics RMSE and MAE for the testing dataset of the case study for each developed model and model type. For each model type (1)-(4) the best performance of RMSE and MAE is highlighted green. If a model performed worse than the best performance of the persistence model, the value of the performance metric is

highlighted red. The persistence model type (1) indicated the best results with a lag of one or four timesteps, whereby the lag of 1460 time steps, which corresponds to exactly one year, being the worst. This effect may be explainable by the limited amount of data which was mainly collected during the Covid-19 pandemic with many unusual effects. Thereby a smaller lag of 1 or 4 timesteps led to better results. For the ML models (2), the XGBoost model indicated the best results in both metrics. The Random Forest model performed the worst and was even below the baseline of the persistence model in the metric MAE. For the ML models as time series (3) XGBoost and Random Forest were particularly close and performed best across all model types. In comparison, the SVR models performed nearly the same in the model types (2) and (3). Due to the walk-forward approach by predicting one time step after another, all of the ML time series models (3) took considerably longer than the ML models (2). Since the Random Forest model takes a relatively long time to fit the model, it was the slowest model. The SARIMA model (4) performed only slightly better than the simple persistence model in the RMSE, yet worse in the MAE measurement. Since SARIMA is a statistical model, the prediction time was relatively fast and comparable to ML models (2) despite the walk-forward approach.

Model Type	Models		
<b>(1) Persistence Model</b>	lag = 1 (6 hours)	lag = 4 (1 day)	lag = 1460 (1 year)
RMSE	161	162	285
MAE	93	79	127
<b>(2) ML Models</b>	XGBoost	Random Forest	SVR
RMSE	117	140	137
MAE	60	94	80
<b>(3) ML as Time series</b>	XGBoost	Random Forest	SVR
RMSE	96	98	128
MAE	51	46	64
<b>(4) SARIMA Time series Model</b>			
RMSE	130		
MAE	82		

Tab. 3: RMSE and MAE of the prediction models of all model types

For an interpretation of the performance metrics, Fig. 2 illustrates the predicted and expected values for November 2021 for the model types (2), (3), and (4). The XGBoost model and the Random Forest model demonstrate similar effects in both types (2) and (3). As ML models (2), XGBoost and Random Forest demonstrate a continuous, recurring daily trend, but successfully identify days with peak occupancy. The difference in the performance metrics is explainable as the XGBoost model differentiates between low and medium occupancy, whereas the Random Forest model predicts the same relatively high daily trend resulting in poorer performance metrics. The ML models as time series (3), demonstrate exceptionally good predictions in November 2021, with both models correctly identifying peak occupancy days as well as lower occupancy days, and only slightly exceeding the expected values. The still relatively high error rates are related to days such as New Year's Day, which display a strongly deviating pattern from 2020/21 to



2021/22 due to the Covid-19 pandemic restrictions that are not yet covered in the trained model. The poor performance of the SVR model in both model types (2) and (3) is immediately apparent, as only the slight daily trend is anticipated, but it detected no occupancy peaks. The penalization of extreme values by creating the internal function of the model probably causes this effect. For predicting days with occupancy peaks, this model seems unsuitable. The SARIMA model correctly predicts the seasonal trend between day and night but does not identify days with peak occupancy in advance. The walk-forward validation only anticipates the upward or downward trend over time.

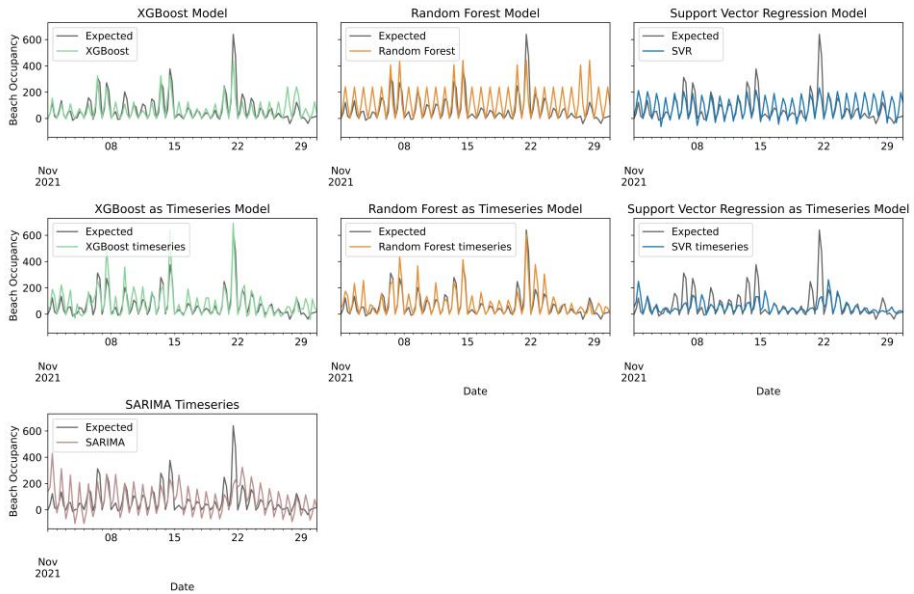


Fig. 2: Predicted versus expected occupancy data for November 2021 of all prediction models

## 5 Discussion

The performance metrics and graphical interpretation indicate a preference for the time series models with XGBoost or Random Forest. However, the time series models predict only one time step ahead, which is equal to 6 hours. In contrast, the supervised ML models predict 40 time steps which is equal to 10 days and limited only by the weather forecast as an input parameter. Since this case study aims to develop an occupancy prediction model to enable active management of beach visitors in Scharbeutz, the prediction of only 6 hours in advance is probably too short. Steering measures to avoid peak occupancies at the beach are more likely to be medium-term to convince some visitors to choose another POI. In this scenario, the still quite good performance of XGBoost as supervised ML model is in favor of the slightly more accurate ML time series models.

Further, the performance of the XGBoost model may improve with an increasing amount of available data. For example, the special situation of the Covid-19 pandemic and its resulting restrictions should become relative as more data with a regular history become available. In this context, a feature that indicates the presence of restrictions like contact restrictions and curfew gives the model more information about the special situation at New Year 2020/21. Since no data was available in this case study, the mentioned Covid-19-specific feature would not have influenced the predictive performance as no comparison to non-pandemic years was possible. Regardless of the Covid-19 pandemic, integrating additional features like events at the beach, felt weather temperature, or water temperature might improve model performance.

Overall, the selection of prediction models itself highly depends on the use case. On the one hand, if the prediction is utilized in steering measures to avoid overcrowding, like in this use case in Scharbeutz, a medium- to long-term prediction is required, since visitors need to be addressed already in the planning phase of their trip. As a countermeasure to overcrowding, the predicted occupancy information can be utilized in an AVM system by recommending alternative POIs to potential visitors. Here, minor deviations in the prediction are not decisive since there is no specific and limiting occupancy number for an open POI like a beach. On the other hand, steering measures at the POI, such as a controlled filling of parking lots requires a short-term and accurate prediction.

## 6 Conclusion

In the present paper, we elaborate on the applicability and performance of various model types, including different ML models, to predict the occupancy of tourist destinations. In a case study of beach occupancy in Scharbeutz, northern Germany, the time series models with XGBoost and Random Forest demonstrated the lowest error, followed by the supervised XGBoost model. However, due to the limited prediction of only one time step in advance, the time series prediction is quite limited and only suitable for steering visitors to a limited extent. In such cases, the supervised XGBoost model may be in favour since peak occupancy days were still predicted correctly. Overall, the case study highlights the importance of selecting the right model depending on the use case and the associated objectives. In future research, we plan to integrate more features on a larger dataset to improve the performance of the ML model, and plan to evaluate it with relative measures and time series cross-validation to obtain a more reliable estimate. Subsequently, the most important features might be identified by using SHAP values, which would allow explaining the underlying effects of peak occupancy days. In addition, the model can be further validated by integrating parking data near the beach to verify whether a full parking lot indicates a high beach occupancy. Finally, the transferability of the developed models to other regions and POIs, such as the Allgäu region in the south of Germany, might be investigated.

## 7 Acknowledgements

The authors gratefully acknowledge the financial support of the Project “AIR” (67KI21005G) by the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection of Germany (BMUV) and the financial support of the Project “FEB-NAFV” (19F2198A) by the Federal Ministry for Digital and Transport (BMDV) as part of the mFUND innovation initiative.

## Bibliography

- [Ab21] Abu, N. et al.: SARIMA and Exponential Smoothing model for forecasting ecotourism demand: A case study in National Park Kuala Tahan, Pahang. *Journal of Physics: Conference Series* 1/1988, pp. 12118, 2021.
- [AK15] Awad, M.; Khanna, R.: Support Vector Regression. In (Awad, M.; Khanna, R. Eds.): *Efficient Learning Machines*. Apress, Berkeley, CA, pp. 67–80, 2015.
- [Al12] Ali, J. et al.: Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)* 9, pp. 272, 2012.
- [APB21] Assemi, B.; Paz, A.; Baker, D.: On-Street Parking Occupancy Inference Based on Payment Transactions. *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2021.
- [Ar22] Arbeitsgemeinschaft Verbrauchs- und Medienanalyse (VuMA): *Verbrauchs- und Medienanalyse - VuMA 2022. Freizeit, Urlaub, Reisen, 2022.*
- [Bi21] Bi, J.-W. et al.: Forecasting Daily Tourism Demand for Tourist Attractions with Big Data: An Ensemble Deep Learning Method. *Journal of Travel Research*, 2021.
- [BL19] Bokelmann, B.; Lessmann, S.: Spurious patterns in Google Trends data - An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management* 75, pp. 1–12, 2019.
- [Ca19] Capocchi, A. et al.: Overtourism: A Literature Review to Assess Implications and Future Perspectives. *Sustainability* 12/11, p. 3303, 2019.
- [Ch00] Chapman, P. et al.: *CRISP-DM 1.0 step-by-step data mining guide*, 2000.
- [Ch15] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K.: Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4), pp. 1–4, 2015.
- [Ch19] Chawathe, S. S.: Using Historical Data to Predict Parking Occupancy: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, New York City, NY, USA, pp. 534–540, 2019.
- [Cl20] Claude, U.: Predicting Tourism Demands by Google Trends: A Hidden Markov Models Based Study. *Journal of System and Management Sciences*, 2020.
- [De16] Deutschen Wirtschaftswissenschaftlichen Institut für Fremdenverkehr e. V.: *Wirtschaftsfaktor Tourismus für das Reisegebiet Ostsee (Schleswig-Holstein)*, 2016.

- [De19] Deutschen Wirtschaftswissenschaftlichen Institut für Fremdenverkehr e. V. (dwif): Wirtschaftsfaktor Tourismus für das Reisegebiet Ostsee 2019, 2019.
- [Di19] Dinis, G. et al.: Google Trends in tourism and hospitality research: a systematic literature review. *Journal of Hospitality and Tourism Technology* 4/10, pp. 747–763, 2019.
- [Ei22] Eisele, J. et al.: Besucherlenkung und Reduktion des motorisierten Freizeitverkehrs - das Potential datengetriebener und flexibler Busangebote. In (Teich, T. et al. eds.): *Innovation und Kooperation auf dem Weg zur All Electric Society. Emergenzen für neue Geschäftsprozesse*. Springer, Heidelberg, 2022.
- [Fe19] Feng, Y. et al.: Forecasting the number of inbound tourists with Google Trends. *Procedia Computer Science* 162, pp. 628–633, 2019.
- [Fr01] Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 5/29, pp. 1189–1232, 2001.
- [Ge22a] German Weather Service: Weather and climate - Deutscher Wetterdienst - Glossary - N - Precipitation intensity. <https://www.dwd.de/DE/service/lexikon/Functions/glossar.html?lv2=101812&lv3=101906>, accessed 17 Mar 2022.
- [Ge22b] German Weather Service: Climate Data Center (CDC). [https://opendata.dwd.de/climate\\_environment/CDC/](https://opendata.dwd.de/climate_environment/CDC/), accessed 17 Mar 2022.
- [Ge22c] German Weather Service: Formulations of the Weather Elements. [https://www.dwd.de/DE/service/lexikon/begriffe/W/Wetterelementeformulierungen\\_pdf.pdf?\\_\\_blob=publicationFile&v=3](https://www.dwd.de/DE/service/lexikon/begriffe/W/Wetterelementeformulierungen_pdf.pdf?__blob=publicationFile&v=3), accessed 17 Mar 2022.
- [Ge22d] German Weather Service: Weather and climate - German Meteorological Service - Glossary - B - Beaufort scale. <https://www.dwd.de/DE/service/lexikon/Functions/glossar.html?lv2=100310&lv3=100390>, accessed 17 Mar 2022.
- [GG18] Gross, S.; Grimm, B.: Sustainable mode of transport choices at the destination – public transport at German destinations. *Tourism Review* 3/73, pp. 401–420, 2018.
- [GV20] Gowreesunkar, V. G.; Vo Thanh, T.: Between Overtourism and Under-Tourism: Impacts, Implications, and Probable Solutions. In (Séraphin, H.; Gladkikh, T.; Vo Thanh, T. eds.): *Overtourism*. Springer International Publishing, Cham, pp. 45–68, 2020.
- [Hø00] Høyer, K. G.: Sustainable Tourism or Sustainable Mobility? The Norwegian Case. *Journal of Sustainable Tourism* 2/8, pp. 147–160, 2000.
- [Hu22] Hu, M. et al.: Tourism demand forecasting using tourist-generated online review data. *Tourism Management* 90, p. 104490, 2022.
- [HX98] Ho, S. L.; Xie, M.: The use of ARIMA models for reliability forecasting and analysis. *Computers & Industrial Engineering* 1-2/35, pp. 213–216, 1998.
- [In21] Institut für Demoskopie Allensbach: Allensbacher Markt- und Werbeträger-Analyse - AWA 2021. *Urlaub und Reisen*, 2021.
- [JC19] Jiao, E. X.; Chen, J. L.: Tourism forecasting: A review of methodological developments over the last decade. *Tourism Economics* 3/25, pp. 469–492, 2019.
- [JM15] Jordan, M. I.; Mitchell, T. M.: Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)* 6245/349, pp. 255–260, 2015.

- [Kh20] Khatibi, A. et al.: Fine-grained tourism prediction: Impact of social and environmental features. *Information Processing & Management* 2/57, p. 102057, 2020.
- [Ki10] Kitagawa, G.: *Introduction to time series modeling*. Chapman & Hall/CRC, Boca Raton, 2010.
- [Ki21] Kim, D.-K. et al.: A Daily Tourism Demand Prediction Framework Based on Multi-head Attention CNN: The Case of The Foreign Entrant in South Korea: 2021 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, Orlando, FL, USA, pp. 1–10, 2021.
- [LI10] Loganathan, N.; Ibrahim, Y.: Forecasting international tourism demand in Malaysia using Box Jenkins Sarima application. *South Asian Journal of Tourism and Heritage* 2/3, pp. 50–60, 2010.
- [Li19] Liu, H. et al.: Hot topics and emerging trends in tourism forecasting research: A scientometric review. *Tourism Economics* 3/25, pp. 448–468, 2019.
- [Ma19] Martinez-Plumed, F. et al.: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering* 8/33, pp. 3048–3061, 2019.
- [Mi22] Ministerium für Energiewende, Landwirtschaft, Umwelt, Natur und Digitalisierung: Datensätze - Open-Data Schleswig-Holstein. [https://opendata.schleswig-holstein.de/dataset?q=&ext\\_startdate=&ext\\_enddate=&groups=tran&sort=score+desc%2C+metadata\\_modified+desc](https://opendata.schleswig-holstein.de/dataset?q=&ext_startdate=&ext_enddate=&groups=tran&sort=score+desc%2C+metadata_modified+desc), accessed 12 Apr 2022.
- [Ne22] Neubig, S. et al.: Data-driven Initiatives of Destinations Supporting Sustainable Tourism: Americas Conference on Information Systems (AMCIS) 2022, 2022.
- [ÖGG20] Önder, I.; Gunter, U.; Gindl, S.: Utilizing Facebook Statistics in Tourism Demand Modeling and Destination Marketing. *Journal of Travel Research* 2/59, pp. 195–208, 2020.
- [ÖGS19] Önder, I.; Gunter, U.; Scharl, A.: Forecasting Tourist Arrivals with the Help of Web Sentiment: A Mixed-frequency Modeling Approach for Big Data. *Tourism Analysis* 4/24, pp. 437–452, 2019.
- [Pa22] Paidi, V. et al.: CO2 Emissions Induced by Vehicles Cruising for Empty Parking Spaces in an Open Parking Lot. *Sustainability* 7/14, p. 3742, 2022.
- [PEF20] Prilistya, S. K.; Erna Permanasari, A.; Fauziati, S.: Tourism Demand Time Series Forecasting: A Systematic Literature Review: 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE). IEEE, Yogyakarta, Indonesia, pp. 156–161, 2020.
- [PS21] Phumchusri, N.; Suwatanapongched, P.: Forecasting hotel daily room demand with transformed data using time series methods. *Journal of Revenue and Pricing Management*, 2021.
- [PY17] Pan, B.; Yang, Y.: Forecasting Destination Weekly Hotel Occupancy with Big Data. *Journal of Travel Research* 7/56, pp. 957–970, 2017.

- [sc22a] scikit-learn developers: sklearn RandomForestRegressor. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>, accessed 24 Apr 2022.
- [sc22b] scikit-learn developers: sklearn svm.SVR. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>, accessed 24 Apr 2022.
- [Sc22] Schmücker, D. et al.: Digitales Besuchermanagement im Tourismus. Konzeptioneller Rahmen und Gestaltungsmöglichkeiten. In (Gardini, M. A.; Sommer, G. Eds.): *Digital Leadership im Tourismus. Digitalisierung und Künstliche Intelligenz als Wettbewerbsfaktoren der Zukunft*. Springer, Wiesbaden, 2022.
- [Sm22] Smith, T. G.: pmdarima 1.8.5 documentation. [pmdarima.arima.auto\\_arima. \[https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto\\\_arima.html\]\(https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto\_arima.html\)](https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html), accessed 24 Apr 2022.
- [SQP19] Song, H.; Qiu, R. T.; Park, J.: A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Annals of Tourism Research* 75, pp. 338–362, 2019.
- [SSD21] Stadler, T.; Sarkar, A.; Dünnweber, J.: Bus Demand Forecasting for Rural Areas Using XGBoost and Random Forest Algorithm. In (Saeed, K.; Dvorský, J. Eds.): *Computer Information Systems and Industrial Management*. Springer International Publishing, Cham, pp. 442–453, 2021.
- [St21] Studer, S. et al.: Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction* 2/3, pp. 392–413, 2021.
- [TB20] Tsang, W. K.; Benoit, D. F.: Gaussian processes for daily demand prediction in tourism planning. *Journal of Forecasting* 3/39, pp. 551–568, 2020.
- [To22] Tourismus-Agentur Lübecker Bucht: Strandticker Lübecker Bucht. <https://www.luebecker-bucht-ostsee.de/strandticker>, accessed 12 Apr 2022.
- [Va00] Vapnik, V. N.: *The Nature of Statistical Learning Theory*. Springer, New York, NY, 2000.
- [Va18] Vanichrujee, U. et al.: Taxi Demand Prediction using Ensemble Model Based on RNNs and XGBOOST: 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES). IEEE, pp. 1–6, 2018.
- [Vo19] Volchek, K. et al.: Forecasting tourist arrivals at attractions: Search engine empowered methodologies. *Tourism Economics* 3/25, pp. 425–447, 2019.
- [Wo18] World Tourism Organization et al.: 'Overtourism'? - Understanding and Managing Urban Tourism Growth beyond Perceptions, Executive Summary. UNWTO, 2018.
- [WSS17] Wu, D. C.; Song, H.; Shen, S.: New developments in tourism and hotel demand modeling and forecasting, p. 23, 2017.
- [xg22] xgboost developers: XGBoost Documentation — xgboost 1.5.2 documentation. <https://xgboost.readthedocs.io/en/stable/index.html>, accessed 13 Apr 2022.

- [YZ19] Yang, Y.; Zhang, H.: Spatial-temporal forecasting of tourism demand. *Annals of Tourism Research* 75, pp. 106–119, 2019.
- [Zh18] Zhang, M. et al.: Weekly Hotel Occupancy Forecasting of a Tourism Destination. *Sustainability* 12/10, p. 4351, 2018.
- [ZHL21] Zheng, W.; Huang, L.; Lin, Z.: Multi-attraction, hourly tourism demand forecasting. *Annals of Tourism Research* 90, p. 103271, 2021.
- [ZZ20] Zhao, Z.; Zhang, Y.: A Comparative Study of Parking Occupancy Prediction Methods considering Parking Type and Parking Scale. *Journal of Advanced Transportation* 2020, pp. 1–12, 2020.

## Appendix

Part of the day	Lower Limit	Upper Limit
0: night	$\geq 12:00 AM$	$< 6:00 AM$
6: morning	$\geq 6:00 AM$	$< 12:00 PM$
12: afternoon	$\geq 12:00 PM$	$< 6:00 PM$
18: evening	$\geq 6:00 PM$	$< 12:00 AM$

Tab. A.1: 6-hour time periods

Temp. category	Lower Limit [°C]	Upper Limit [°C]	Rain category	Lower Limit [kg/m <sup>2</sup> ]	Upper Limit [kg/m <sup>2</sup> ]	Wind category	Lower Limit [m/s]	Upper Limit [m/s]
-10		$< -10$	0.5		$< 0.5$	0		$< 0.3$
0	$\geq -10$	$< 0$	2.5	$\geq 0.5$	$< 2.5$	1	$\geq 0.3$	$< 1.6$
10	$\geq 0$	$< 10$	5.0	$\geq 2.5$	$< 5.0$	2	$\geq 1.6$	$< 3.4$
18	$\geq 10$	$< 18$	10.0	$\geq 5.0$	$< 10.0$	3	$\geq 3.4$	$< 5.5$
24	$\geq 18$	$< 24$	50.0	$\geq 10.0$	$< 50.0$	4	$\geq 5.5$	$< 8.0$
30	$\geq 24$	$< 30$	51.0	$\geq 50.0$		5	$\geq 8.0$	$< 10.8$
35	$\geq 30$	$< 35$				6	$\geq 10.8$	$< 13.9$
36	$\geq 35$					7	$\geq 13.9$	$< 17.2$
						8	$\geq 17.2$	$< 20.8$
						9	$\geq 20.8$	$< 24.4$
						10	$\geq 24.4$	$< 28.4$
						11	$\geq 28.4$	$< 32.6$
						12	$\geq 32.6$	

Tab. A.2: Temperature, rain, and wind categories

Precipitation form	Historical observation precipitation form	Mosmix forecast
0: no precipitation	NaN, 0, 4, 9	Snow water equivalent = 0 & Precipitation height < 0.5
1: rain	1, 6	Precipitation height > 0.5
2: snow	7, 8	Snow water equivalent > 0

Tab. A.3: Feature engineering of the precipitation form

Features	Type of Data	Value Ranges
year	Integer	[2020, 2021, 2022]
month	Integer	[1, ..., 12]
day of year	Integer	[1, ..., 365]
quarter of year	Integer	[1, ..., 4]
day of month	Integer	[1, ..., 31]
calendar week	Integer	[1, ..., 52]
day of week	Integer	[1, ..., 7]
hour	Integer	[0, ..., 23]
weekend	Binary	[0, 1]
bridging day	Binary	[0, 1]
public holiday (bank holiday)	Binary	[0, 1]
regional school holiday	Binary	[0, 1]
German school holiday density	Float	[0, ..., 1]
wind category	categorical	[0, 1, 2, 3, 4, 5, 6, 7,8, 9, 10, 11, 12]
rain category	categorical	[0.5, 2.5, 5.0, 10.0, 50.0, 51.0]
temperature category	categorical	[-10, 0, 10, 18, 24, 30, 35, 36]
precipitation form	categorical	[0, 1, 2]

Tab. A.4: Structure of the final data set of input features