CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Short-term Electric Load Forecasting Using Czech Data

Master's thesis

Author: Bc. Martin Řanda
Study program: Economics and Finance
Supervisor: prof. PhDr. Ladislav Krištoufek Ph.D.
Year of defense: 2023

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants permission to Charles University to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, May 2, 2023

Bc. Martin Řanda

Abstract

Forecasting electric load accurately is a critical prerequisite to dependable power grid operation. It is thus in the best interests of the responsible institutions to develop and maintain performant models for predicting load. In this thesis, we analyze Czech electric load data and execute three pseudo-out-of-sample forecasting exercises. We employ standard econometric as well as machine learning methods and compare the results to benchmarks, including the predictions published by the Czech transmission system operator. The results of the first task examining the predictability of minute loads using 11 years of data indicate that the high-frequency load series is predictable. In the second and third exercises, we utilize hourly loads with additional explanatory variables. We generate one-step-ahead and 48-hours-ahead forecasts on the 2021 out-of-sample set and evaluate the performance of several methods. In both exercises, the most accurate results are produced by averaging forecasts of our specified recurrent neural network and the seasonal autoregressive integrated moving average model, achieving a mean absolute percentage error of less than 0.5% on the out-of-sample set in the one-step-ahead analysis and 2.3% in the 48-hours-ahead exercise, outperforming the operator's predictions.

JEL Classification	C53, Q40, C58
Keywords	load forecasting, electricity, short-term forecasting, Czech data, time series analysis
Title	Short-term Electric Load Forecasting Using Czech Data

Abstrakt

Přesná předpověď elektrického zatížení je zásadním předpokladem spolehlivého provozu elektrické rozvodné soustavy. Je proto v nejlepším zájmu odpovědných institucí vyvíjet a udržovat výkonné modely pro předpovědi zatížení. V této práci analyzujeme data o zatížení elektrizační soustavy České republiky a provádíme tři pseudo-out-of-sample forecasting cvičení. Používáme standardní ekonometrické modely i metody strojového učení a výsledky porovnáváme s referenčními hodnotami, včetně předpovědí zveřejňovaných provozovatelem české přenosové soustavy. Výsledky první úlohy zkoumající předvídatelnost minutového zatížení na základě 11 let dat ukazují, že vysokofrekvenční časové řady zatížení jsou předvídatelné. Ve druhé a třetí úloze využíváme hodinovou zátěž s dalšími vysvětlujícími proměnnými. Vytváříme předpovědi na jeden krok a na 48 hodin dopředu na out-of-sample vzorku roku 2021 a vyhodnocujeme výkonnost několika metod. V obou cvičeních byly nejpřesnější výsledky získány zprůměrováním předpovědí námi specifikované rekurentní neuronové sítě a sezónního autoregresního integrovaného klouzavého průměru, které dosáhly průměrné absolutní procentní chyby menší než 0.5% na out-of-sample vzorku v analýze na jeden krok dopředu a 2.3% v úloze na 48 hodin dopředu, čímž překonávají předpovědi operátora.

Klasifikace JEL	C53, Q40, C58
Klíčová slova	předpověď zatížení, elektřina, krátkodobé předpovědi, česká data, analýza časových řad
Název práce	Krátkodobé předpovědi zatížení elektrizační soustavy s využitím českých dat

Acknowledgments

Firstly, I would like to gratefully acknowledge the valuable advice and guidance provided by prof. PhDr. Ladislav Krištoufek Ph.D. I would also like to extend my sincere gratitude to Bc. Matyáš Mattanelli, Bc. Pavel Pekárek, Bc. Jan Kubal, and one other colleague, who has chosen not to be named, for their assistance in proofreading as well as providing helpful suggestions. My appreciation further goes to V. W. for their (spiritual) support. Last but not least, I am thankful for the encouragement and support of my family, significant other, and friends.

Typeset in LaTeX using the IES Thesis Template.

Bibliographic Record

Řanda, Martin: *Short-term Electric Load Forecasting Using Czech Data*. Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2023, pages 110. Advisor: prof. PhDr. Ladislav Krištoufek Ph.D.

Contents

List of Tables

List of Figures

Acronyms

ADF	Augmented Dickey-Fuller (test)
AIC	Akaike information criterion
ANN	Artificial neural network
ARCH	Autoregressive conditional heteroskedasticity
ARIMA	Autoregressive integrated moving average; AR and MA sometimes referred to separately
°C	Degrees Celsius
ČEPS	Czech Transmission System Operator; ČEPS, a.s.
DM	Diebold-Mariano test
EIA	Energy Information Administration
EPA	Environmental Protection Agency
EU	European Union
EUR	Euro
hPa	Hectopascal
IPCC	Intergovernmental Panel on Climate Change
JB	Jarque-Bera (test or statistic)
km	Kilometer
kV	Kilovolt
LSTM	Long short-term memory
LTLF	Long-term load forecasting
m/s	Meters per second
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MLP	Multilayer perceptron
MW	Megawatt
MWh	Megawatt-hours
NOAA	National Centers for Environmental Information
OTE	Czech Electricity and Gas Market Operator; OTE, a.s.
Q1	First quartile
Q3	Third quartile
ReLU	Rectified linear unit
RMSE	Root mean square error
RNN	Recurrent neural network
SARIMA(X)	Seasonal autoregressive integrated moving average model (with exogenous factors)
SD	Standard deviation
SNAIVE	Seasonal naïve (method)
STLF	Short-term load forecasting
TWh	Terawatt-hours

Master's Thesis Proposal

Author	Bc. Martin Řanda
Supervisor	prof. PhDr. Ladislav Krištoufek Ph.D.
Proposed topic	Short-term Electric Load Forecasting Using Czech Data

Motivation In the 1920s, yearly electricity production on the territory of the Czech Republic reached almost 1300 gigawatt-hours according to the CZSO (2013). Ninety years later, such a number would cover only about one-sixtieth of the total demand for electric energy, which is undoubtedly a telling indicator of the drastic increase in technological progress in the region.

However, there are risks associated with our high reliance on electricity: perhaps the greatest challenge in the electric power industry is that the demand has to match the supply of electricity at all times. The issue is that large-scale electricity storage technologies have not yet reached a level of sophistication to be viable for this task, as Hong (2014) writes. Consequently, electricity generation needs to be balanced and coordinated in real-time, 24 hours a day. Because of these factors, coupled with the natural deterioration of infrastructure and the gradual shift towards "smart grids," forecasting of electricity load has been growing in importance (Hong and Fan 2016) and has relatively recently seen increased coverage in academic literature.

As reported by Hong and Fan (2016), the most widely used techniques for electricity load forecasting include models from the autoregressive-moving-average family, as well as generalized additive models, multiple linear regression, or exponential smoothing. Researchers also utilize machine-learning algorithms such as artificial neural networks, support vector machines, or fuzzy regression approaches. Kuster et al. (2017) provide a review of these electricity load forecasting models and their application to data at various scales (e.g., load data of buildings, districts, or countries) and frequencies (from sub-hourly to annual). Their findings suggest that most works were concerned with long-term forecasts and that there is no clear indication of which methods perform best in different circumstances. In a recently published paper by Lee and Cho (2022), the authors devise a model for nationwide peak electricity load forecasting in South Korea. They further claim that researchers have mostly concentrated on smaller-scale studies, meaning that works utilizing country-level data are, at the very least, in the minority.

Thus, one of the goals of this thesis is to contribute to this growing body of literature. In particular, we aim to provide an overview of the Czech energy sector, highlighting the responsibilities of the licensed transmission system operator ČEPS,

a.s. (Vlček and Černoch 2013). We will then analyze a highly granular and extensive dataset of country-wide electricity load obtained from the licensed operator of the Czech electricity transmission system. Finally, various models will be proposed, and their performance will be evaluated in a pseudo-out-of-sample forecasting exercise (as per, for instance, Fan and Hyndman (2012)).

Hypotheses

Hypothesis 1: *Very short-term forecasts* of electricity load outperform a naïve random walk model.
Hypothesis 2: Predictability *increases* with *lower* frequency data.
Hypothesis 3: Models with external variables generate more accurate forecasts than models based solely on historical load data.

Methodology We have obtained sub-hourly historical electricity load data from the transmission grid operator ČEPS, a.s. starting from 2010 until the end of 2021. Specifically, the dataset contains high-frequency electricity load statistics for the whole Czech Republic, as well as aggregated lower frequency data on an hourly and daily basis. Furthermore, at our disposal, we also have the official day- and week-ahead predictions of the variable in question.

As far as other external factors are concerned, we will also be working with electricity prices and weather data, both of which are commonly utilized in the literature (Kuster et al. 2017). Moreover, electricity load data is subject to a wide range of calendar variations, for example, day of the week or holiday effects (Fan and Hyndman 2012), which also need to be accounted for and treated properly.

The frameworks we plan on utilizing range from simple univariate models to more complex multivariate methods. In particular, we aim to utilize models from the autoregressive-moving-average family, as well as exponential smoothing and vector autoregression. In addition, we will also evaluate the forecasting accuracy of an artificial neural network, which has also been used in the short-term load forecasting literature (Hong and Fan 2016; Kuster et al. 2017).

For the high-frequency data, we will consider a naïve random walk model as the baseline—failing to outperform this specification in an out-of-sample forecasting exercise would imply unpredictability of the series in question. This is, in fact, the principal motivation behind the first hypothesis outlined above. Furthermore, for the lower frequency data, we will be comparing the results to the official forecasts published by the transmission system operator. Finally, forecasting accuracy will be assessed through standard measures such as root mean square error, mean absolute error, or the Diebold-Mariano test (Diebold and Mariano 2002), and a *ceteris paribus* analysis will be conducted.

Expected Contribution According to Malik et al. (2021), load forecasts play a vital role in the power industry as well as in the operation of the electricity grid (Hong 2014). Depending on the considered time frame, load forecasting is utilized for activities such as optimal supply planning, maintenance scheduling, or control over automatic power generation (Malik et al. 2021). Moreover, both an overestimation and an underestimation may lead to undesirable consequences—for example, in the

latter case, the entire system is put into a "vulnerable region to the disturbance" (Fan and Hyndman 2012), meaning that a "blackout" could occur (Lee and Cho 2022).

We believe that the added value of this thesis is in the use of high-frequency data on a national scale, especially in the case of the Czech Republic. While some works have used Czech electricity load data (e.g., Darbellay and Slama (2000) or Uher et al. (2015)), the frequency of their data and the overall approach differ. In terms of the lower frequency data and the planned direct comparison of our results to the official forecast, we think that such analysis could be helpful in assessing the quality of these predictions and the results could potentially improve the underlying model.

Outline

1. Introduction
2. Literature review
3. Background
    - Overview of the Czech energy sector
    - Electricity load
4. Data description and methodology
5. Results and discussion
6. Conclusion

Core bibliography

CZSO, 2013. "Historická ročenka statistiky energetiky – 2012." *Czech Statistical Office*. URL: https://www.czso.cz/csu/czso/8113-12-n_2012-01. [Accessed 2022-04-18].

Černoch, F. and Vlček, T., 2013. *The energy sector and energy policy of the Czech Republic*. Masaryk University Press.

Darbellay, G.A. and Slama, M., 2000. Forecasting the short-term demand for electricity: Do neural networks stand a better chance?. *International Journal of Forecasting, 16*(1), pp.71-83.

Diebold, F.X. and Mariano, R.S., 2002. Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), pp.134-144.

Fan, S. and Hyndman, R.J., 2011. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems, 27*(1), pp.134-141.

Hong, T., 2014. Energy forecasting: Past, present, and future. *Foresight: The International Journal of Applied Forecasting*, (32), pp.43-48.

Hong, T. and Fan, S., 2016. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting, 32*(3), pp.914-938.

Kuster, C., Rezgui, Y. and Mourshed, M., 2017. Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society, 35*, pp.257-270.

Lee, J. and Cho, Y., 2022. National-scale electricity peak load forecasting: Traditional, machine learning, or hybrid model?. *Energy, 239*, p.122366.

Malik, H., Fatema, N. and Iqbal, A., 2021. *Intelligent data-analytics for condition monitoring: smart grid applications*. Academic Press.

Uher, V., Burget, R., Dutta, M.K. and Mlynek, P., 2015, July. Forecasting electricity consumption in Czech Republic. In *2015 38th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 262-265). IEEE.

Chapter 1

Introduction

> *With every addition to the plant of an electric utility system there is presupposed, either consciously or unconsciously, some estimate of future loads, even if it be no more than a forecast that the enterprise will continue in business.*

– R. G. Hooke (1955)

In a report on historical energy usage within the present-day boundaries of the Czech Republic published by the Czech Statistical Office (2013), the earliest record enumerating annual electricity production dates back to 1919, according to which 1093 gigawatt-hours were generated. More than a century later, the country produces almost 80 times as much electricity each year, contains thousands of kilometers of power lines, and is part of the continent's largest electrical grid (Hofmann *et al.* 2020; ČEPS 2021; Ritchie *et al.* 2022).

For the system operator's seemingly never-ending task of maintaining electricity supply and demand in balance in order to ensure uninterrupted operation of the power grid, forecasting load accurately has been crucial (Liao & Tsao 2004). In fact, Krugman & Wells (2015, p. 698) argue that economic growth in many nations can be severely hindered by an unreliable electrical infrastructure, further underpinning its importance by stating that governments have to be politically disciplined to ensure its proper functioning. Moreover, the increase in adoption of renewable electricity sources has further heightened the significance of this field in recent years due to the inherent variability of output of some of these modern solutions (Poullikkas 2013; Hong & Fan 2016).

However, despite the current-day abundance of computational capacity and a considerable body of literature, generating load predictions remains far from

trivial (Wang *et al.* 2016; Kuster *et al.* 2017). Perhaps the main difficulty in modeling such data is rendered by the presence of multiple seasonal patterns, which include the day/night cycle, working/non-working days, season of the year, and public holidays (Weron 2006, p. 68, Fan & Hyndman 2012). Furthermore, because there seems to be a lack of consensus about the most performant techniques for load forecasting (Kuster *et al.* 2017), researchers employ a variety of methods, such as *artificial neural networks* (ANNs), *autoregressive integrated moving average* (ARIMA) models, or *hybrid approaches* (Hong & Fan 2016; Mamun *et al.* 2020).

While some academic papers have utilized data from the Czech Republic (Darbellay & Slama 2000; Khan *et al.* 2002; Uher *et al.* 2015; Lai *et al.* 2020), none of the works that we could find appeared to have investigated minute load series or contrasted their forecasts with those published by the system operator. Thus, this thesis aims to contribute to the existing load forecasting literature by analyzing recent Czech electric load series released by the country's transmission grid operator. Specifically, we execute three *pseudo-out-of-sample forecasting exercises*, i.e., a model's forecasting accuracy is tested on an *unseen* set of data reserved beforehand (Stock & Watson 2020, p. 575).

In the first minor task, we analyze the predictability of minute-ahead load using 11 years of recent high-frequency data, hypothesizing that the Czech minute load series is predictable, based partly on Taylor (2008). The other two exercises, in which we model hourly data, are concerned with producing one-step-ahead and up to 48-hours-ahead forecasts on an out-of-sample set of one year using several methods. In particular, the main techniques that we employ include *seasonal ARIMA with exogenous factors* (SARIMAX), *recurrent neural networks* (RNN), and *bagged regression trees*. For the multivariate methods, we incorporate additional predictors such as transformations of historical loads, weather, prices, or indicator variables. Moreover, in the latter multi-step analysis, we further compare our predictions to those published by the Czech transmission system operator to evaluate their accuracy. Generally, these short-term projections[1] are vital in multiple processes executed during the operation of a power grid (Khuntia *et al.* 2016), as we describe later.

The thesis is organized as follows. A review of academic literature on load forecasting, including a brief historical excursion, is presented in Chapter 2. The objective of Chapter 3 is to provide a general introduction to power grid

[1]Let us note that we use the terms *forecast*, *prediction*, and *projection* interchangeably as a stylistic choice.

management, the risks associated with inaccurate load forecasts, and the Czech electric power sector. In Chapter 4, we describe the acquired data and perform an exploratory analysis. Chapter 5 thoroughly details the methodological part of this thesis. The results of the three pseudo-out-of-sample forecasting exercises are reported and discussed in Chapter 6. Chapter 7 then summarizes our findings and suggests possible future steps. Appendices A & B display supplementary materials and additional results, respectively.

Chapter 2

Literature Review

This chapter is composed of five sections. Firstly, we provide a brief historical overview and early contributions to the field of load forecasting. We then outline the key design considerations researchers take before specifying a model. In the third section, we focus on the current methods utilized in academic literature. After describing a few works concerned with forecasting Czech electric load, we conclude this chapter by introducing three closely related fields that are also dependent on predictions.

2.1 A Historical Perspective

While electricity load forecasting is vital for the proper functioning of the power grid, it has only started garnering increased attention in the 1970s, according to Sachdev *et al.* (1977). They attributed the uptick in literature and an overall rise in the importance of load forecasting to the increase in fuel prices. Regardless, the authors list several prior works, the earliest of which dates back to 1918. Nearly three decades later, Dryar (1944) analyzes the impact of weather attributes such as temperature, wind speed, or the "degree of cloudiness" on the estimate of the power system load. Whether the paper was the first to articulate this relationship in academic literature or not (as alluded to by Wang *et al.* (2016)), it is a simple yet meaningful observation—today, weather variables are widely utilized in load forecasting (e.g., Feinberg & Genethliou (2005) or Kuster *et al.* (2017)) as well as electricity price forecasts (Weron 2014).

In this context, Heinemann *et al.* (1966) find it helpful to decompose the overall system load into two parts based on whether the weather has an impact on it or not. The component that is susceptible to weather fluctuations is in

part affected by the level of comfort that people are unwilling to sacrifice, as remarked by the authors. The paper itself details a method for making such a decomposition—this was especially important because of the increasing amount of electrical appliances found in homes, which significantly raised the mean system load as well as its volatility. For instance, in the United States, more than 50% of houses built in the 1960s had central air conditioning units installed, as reported by the *Energy Information Administration* (EIA 2009), meaning that the contribution of these appliances to total energy consumption on a particularly hot day was gradually becoming more impactful (Heinemann *et al.* 1966). Regardless of the demand side of this *equation*, as time passes, power systems grow in complexity due to the diversification of energy sources, decentralization, and grid interconnectedness (Pfenninger *et al.* 2014), arguably increasing the importance of load forecasting.

Hong & Fan (2016) state that due to the loosening of regulation in the utility sector in the 1980s, the development of *short-term load forecasting* (STLF) methods received a considerable amount of interest. One example of a paper from this period is that of Hagan & Behr (1987), which forecasts hourly loads using data from 1983 to 1984. The authors utilize three methods: an autoregressive integrated moving average model, a transfer function model, and its nonlinear extension. Each of these models was trained on 28 days of data and produced forecasts three weeks ahead. By comparing the predictions with a conventionally used method for load forecasting using *mean absolute percentage error* (MAPE), they found that all three models outperformed the standard approach, with the nonlinear procedure achieving the lowest average error.

Advancements in computing have undoubtedly played a significant role in shaping the modern-day state of the field. In the last decade of the 20th century, the growth of computer processing power has made the implementation of artificial intelligence and other computationally demanding methods viable and intensively researched in load forecasting (Alfares & Nazeeruddin 2002; Weron 2006, p. 68; Wang *et al.* 2016). Despite that, Hong & Fan (2016) write that only a limited amount of academic literature in the past thirty years has succeeded in generating industry-relevant findings, which is something they consider to be the main objective of the field. Therefore, they stress that contributions to load forecasting research should aim to be applicable in practice and bring innovation by, for example, utilizing new data or methods.

2.2 General Considerations in Model Specification

While the topic of electricity load forecasting may seem narrow enough to some, there are several distinct avenues to explore, each with its unique set of obstacles and considerations. For instance, longer forecast horizons are required for different purposes than short-term predictions in the overall administration of the power grid (Malik *et al.* 2021), which we further explore in Section 3.2. On that account, let us concentrate on three key elements in specifying a load forecasting model based on Kuster *et al.* (2017): resolution, scale, and variables.

2.2.1 Resolution

Observations that are captured at periodic intervals and ordered chronologically is what we typically imagine when we think of time series data (Lütkepohl & Krätzig 2004, p. 1). In this context, Kuster *et al.* (2017) refer to the frequency of data as resolution, which is then closely related to the forecast horizon.

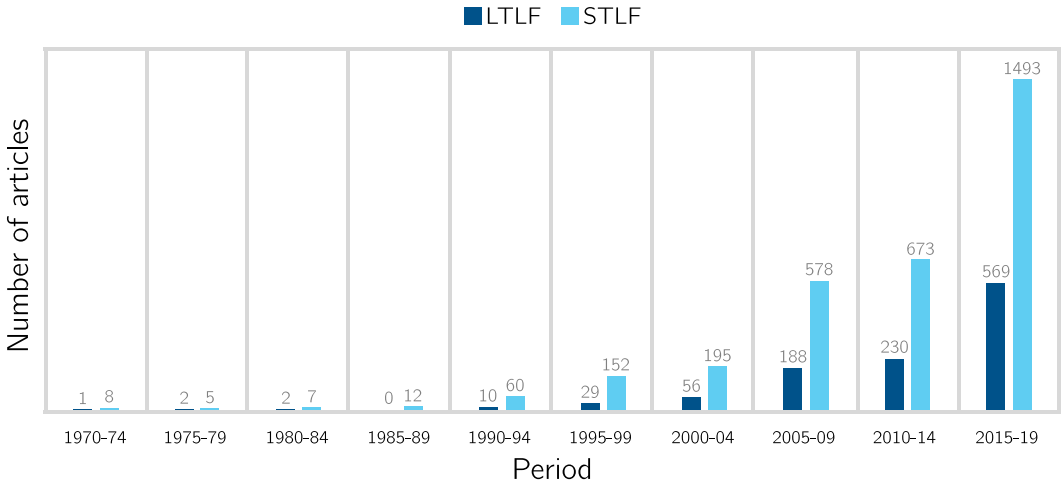
Researchers have commonly used load data with frequencies such as one minute, five minutes, half-hourly, hourly, daily, monthly, or yearly. For instance, Taylor (2008) utilizes over a half-year's worth of load data with a frequency of one minute obtained from the system operator in Great Britain, concluding that the most suitable specification for generating predictions of up to 30-minutes-ahead appears to be a particular exponential smoothing method. Guan *et al.* (2013), on the other hand, use five-minute loads in their model and produce forecasts for the next hour. Furthermore, in both Taylor (2012) and Fan & Hyndman (2012), the respective authors analyze half-hourly data on a national level, but their horizons vary (up to 24 hours and 7 days, respectively).

In several papers, hourly data have also been utilized—for example, Kwon *et al.* (2020) forecast loads one day ahead with a training set that consists of about a year of observations. Lee & Cho (2022), on the other hand, train their model on a 5-year dataset, but they work with daily data. Power system loads on a monthly frequency were analyzed by Ghiassi *et al.* (2006), forecasting up to 12-months-ahead. Finally, Bianco *et al.* (2009) utilize yearly data for their predictions, concluding that an increase in power consumption in the Italian market ought to be anticipated.

According to Hong & Fan (2016), shorter-term forecasts have garnered more attention throughout the years. To be more specific, while discussing forecast horizon lengths, the researchers support their argument with a possibly misla-

beled figure, which displays the number of articles on short-term and *long-term load forecasting* (LTLF) listed on the Web of Science in 5-year intervals since the 1970s. Thus, in Figure 2.1, we provide a revised and updated chart[1] similar to the original plot, which indeed illustrates the disparity in coverage between the two fields.

Figure 2.1: Articles on short- and long-term load forecasting listed on the Web of Science



Source: Data collected by the author, chart based on Hong & Fan (2016).

2.2.2 Scale

The overall area or the scope at which loads were gathered is referred to as the data's scale. Specifically, these levels range from buildings to city districts and up to the regional or state level (Kuster *et al.* 2017). For example, the paper of Marino *et al.* (2016) is one of the small-scale works that produces load forecasts for a single household. Yildiz *et al.* (2017), on the other hand, analyze loads generated by a university campus in Australia, stating that the electricity demand of large building complexes with complicated energy systems tends to be difficult to forecast, especially considering both internal and external factors such as weather, scheduling, or high energy requirements of specialized rooms (e.g., a research laboratory). Another factor Ruiz-Abellón *et al.* (2018) account for in their campus-level model are lower-activity periods related to the academic year. Moreover, Kim *et al.* (2019) forecast peak load for a 23-

[1]We collected the data by searching for keywords associated with LTLF and STLF in each 5-year period on the Web of Science, respectively, and saved the number of aggregated articles. Thus, these values should only be considered indicative of the overall trend.

building campus block in South Korea, though the methods that the authors use slightly differ.

One of the works that is concerned with load forecasts at the district and city level is that of Jung *et al.* (2020). Their dataset is composed of more than a decade of monthly loads in 25 districts of the South Korean capital, but they also have several weather variables at their disposal. Furthermore, regional-level data is utilized by Guan *et al.* (2013)—in particular, the researchers use high-frequency data obtained from the system operator in New England, and a significant part of their study focuses on data smoothing, which they found helpful for further modeling.

Finally, nationwide data are, for instance, analyzed in Liao & Tsao (2004), Taylor (2008), Papaioannou *et al.* (2016), Kwon *et al.* (2020), or Lee & Cho (2022). However, Lee & Cho (2022) note that most works have been concerned with data on a lesser scale, implying that studies utilizing country-level loads are in the minority, which is somewhat supported by Kuster *et al.* (2017).

2.2.3 Variables

As we have established earlier, the concept of employing additional predictors in load forecasting perhaps first emerged nearly 80 years ago (Dryar 1944). Apart from meteorological data, the current body of literature tends to also utilize other variables, such as calendar effects or demographic and socioeconomic factors, depending on the resolution of the study (Feinberg & Genethliou 2005; Weron 2006, p. 68; Mamun *et al.* 2020).

According to Feinberg & Genethliou (2005), one of the most popular weather-related factors used by researchers is temperature, followed by humidity. Wind speed appears to be employed in some approaches as well, either directly or indirectly—for instance, Khotanzad *et al.* (1998) transformed the three aforementioned weather variables into one as it enhanced the performance of their model. In fact, transformations of meteorological factors appear to be somewhat common in the load forecasting literature across data scale and resolution, as evidenced by, for example, Fan & Hyndman (2012), Yildiz *et al.* (2017), Elamin & Fukushige (2018), or Lee & Cho (2022). Last but not least, cloudiness is also occasionally incorporated by some (e.g., Kandil *et al.* (2006)), though with varying degrees of success.

Another frequently practiced procedure, especially in STLF, consists of controlling for specific days of the week, holidays, particular hours of the day, and

other seasonal trends. For instance, in their analysis, Yildiz *et al.* (2017) use a dummy variable for separating workdays from non-working days as well as indicators for the day of the week and for the hourly time of day. Additionally, similar control variables are also used in higher-scale studies, such as that of Fan & Hyndman (2012) or Jung *et al.* (2020).

Furthermore, some lower-resolution and higher-scale studies utilize demographic data—for example, Jung *et al.* (2020) account for, among other factors, migration and population density, which they consider reasonable due to the district-level scale and monthly resolution of the analysis. Finally, two of the variables that Bianco *et al.* (2009) also control for in their long-horizon model are gross domestic product and electricity price. In addition, prices are sometimes utilized in STLF specifications as well (Chen *et al.* 2001), though less commonly than other factors mentioned above.

2.3 State-of-the-Art Load Forecasting Methods

In spite of the relatively sizable amount of coverage of STLF in academic literature, no particular modeling approach is preferred in terms of its performance (Hong & Fan 2016; Kuster *et al.* 2017). As per Hong & Fan (2016), both machine learning and standard statistical methods are utilized in electric load forecasting. Some of the most popular techniques listed by the authors are ANNs, support vector machines, fuzzy regression, gradient boosting, ARIMA-family models, multiple linear regression, exponential smoothing, and generalized additive models. Mamun *et al.* (2020) further add that hybrid models, i.e., a combination of two or more methodologies, tend to be utilized and often produce better results than the underlying methods on their own. Finally, several papers report forecasts of custom techniques used by the grid operator relevant to the study (e.g., Taylor (2008)). These approaches are generally not thoroughly described in the respective papers themselves, but instead link to other works or documents that elaborate on the methodology.

In Table 2.1, we provide a categorized list of STLF literature we have cited thus far based on the utilized methods. As we may observe, the overwhelming majority of papers have used some form of an ANN, ranging from fuzzy logic to a wavelet neural network (Liao & Tsao (2004) and Guan *et al.* (2013), respectively). A few publications that we have reviewed also produce forecasts using hybrid models—we explore these articles in detail below due to their relevance to our research design: large-scale and *higher*-frequency data.

Table 2.1: Methods used within the load forecasting literature referenced in Section 2.2

Method	**Literature**
ANNs	Khotanzad *et al.* (1998), Chen *et al.* (2001), Liao & Tsao (2004), Kandil *et al.* (2006), Fan & Hyndman (2012), Taylor (2012), Guan *et al.* (2013), Papaioannou *et al.* (2016), Marino *et al.* (2016), Yildiz *et al.* (2017), Kim *et al.* (2019), Kwon *et al.* (2020), Lee & Cho (2022)
ARIMA-family	Taylor (2008), Taylor (2012), Papaioannou *et al.* (2016), Elamin & Fukushige (2018), Kim *et al.* (2019), Lee & Cho (2022)
Exponential smoothing	Taylor (2008), Taylor (2012), Papaioannou *et al.* (2016), Kim *et al.* (2019)
Generalized additive models	Fan & Hyndman (2012)
Hybrid models	Fan & Hyndman (2012), Taylor (2012), Lee & Cho (2022)
Multiple linear regression	Yildiz *et al.* (2017)
Tree-based models	Yildiz *et al.* (2017), Ruiz-Abellón *et al.* (2018)
Support vector machines	Papaioannou *et al.* (2016), Yildiz *et al.* (2017), Lee & Cho (2022)
Other (e.g., operator-specific)	Taylor (2008), Taylor (2012), Guan *et al.* (2013), Kwon *et al.* (2020)

Fan & Hyndman (2012) present an STLF model on a regional scale in Australia, which has consequently been applied in practice by the market operator. While the authors had the possibility of utilizing a larger dataset with more historical records, the final training set consisted of semi-hourly data from 2004 to 2008 as it produced more precise results. Fan & Hyndman (2012) then conduct an out-of-sample forecasting exercise by generating half-hourly predictions of three specifications with a horizon of up to one week. The approaches used by the researchers were composed of a semi-parametric additive regression, ANN, and a hybrid method, all of which incorporated a set of explanatory variables, including log-transformed lagged electricity demand, temperature, and several seasonal categorical or dummy variables. Between October 2008 and March 2009, the average monthly out-of-sample *mean absolute error* (MAE) of the additive model was, with the exception of January 2009, less than 100 *megawatts* (MW), while the other two specifications consistently recorded values above 110

MW. Similarly, the MAPE of the two approaches remained consistently 0.3–0.9 percentage points higher than that of the semi-parametric additive method.

A short-term load forecasting exercise utilizing half-hourly data from Great Britain and France is performed by Taylor (2012). The in-sample portion of the 3-year dataset spans from 2007 to 2008, while the out-of-sample set consists of observations in 2009. Multiple univariate exponential smoothing techniques, such as an extension of the Holt-Winters method or a singular value decomposition-based approach, are employed. The forecasts produced by these methods were compared with ARIMA and ANN benchmarks, as well as the predictions of the official model used by the operator and also employed in the author's previous work (Taylor 2008). For all horizons ranging from 30 minutes up to 24 hours ahead, the method that generated forecasts of British loads with the minimal MAPE was an unweighted average of the two aforementioned exponential smoothing approaches and the operator-utilized model.[2]

Lee & Cho (2022) evaluate the performance of several machine learning and hybrid methods in a peak load forecasting exercise using national-level data from South Korea. In particular, the analysis is based on roughly 6 years of daily loads (5 years of training and nearly 11 months of test data), spanning from 2014 to late 2019. In all their models, the authors control for the effect of holidays, average temperatures, humidity, and also *degree day* measures—*heating degree days* and *cooling degree days*, which indicate the temperatures required to heat or cool indoor areas to a "comfortable" level, according to the United States *Environmental Protection Agency* (EPA 2022). Out of the non-hybrid specifications, the most accurate method with respect to the error measures used was the *long short-term memory* (LSTM) neural network. Similarly, the hybrid approach consisting of a SARIMAX model and the LSTM RNN provided comparable levels of accuracy. Importantly, however, both procedures surpassed the official predictions produced by the power grid operator in South Korea in terms of all the applied error metrics.

2.4 Load Forecasting Using Czech Data

As we have shown earlier, the topic of load forecasting has received a great deal of attention in academic literature, especially in the last two decades. Despite that, it seems that only a limited number of studies have been con-

[2]The French dataset was not used in this part of the analysis.

cerned with forecasting electricity load within the Czech power grid. Among the papers that make use of such data, Darbellay & Slama (2000) have perhaps attracted the most interest. In their research article, the authors work with hourly 1994 and 1995 national-scale loads and forecast up to 36 hours ahead using ARIMA-family models as well as an ANN. However, their main objective was to contrast the sufficiency of linear and non-linear methods for predicting electricity loads in the Czech Republic. Importantly, they conclude that introducing non-linearity has no significant impact, though it is worth noting that their results could perhaps be improved by using hourly rather than daily temperature data in combination with hourly loads.

The data utilized by Khan *et al.* (2002) is also hourly and even overlaps with the dataset of Darbellay & Slama (2000), though it contains more observations as it covers the period from 1994 to the end of 2000. In contrast, however, Khan *et al.* (2002) motivate the usage of non-linear methods by stating that linear techniques are inefficient due to the nonstationarity of the underlying process as well as the lack of linearity between loads and meteorological variables. Consequently, the methods that they use consist of six different ANNs—the choice of utilizing these techniques is further supported by the usage of several weather-related factors such as humidity, wind velocity, or temperature.

In a conference paper by Uher *et al.* (2015), the authors work with hourly regional electricity consumption data in the Czech Republic from 2011 to 2014. Their analysis, however, is not concerned with employing meteorological variables; instead, five approaches are specified and trained on power consumption data with calendar effects as the only additional inputs. The performance of these techniques is then compared on a test set using *root mean square error* (RMSE)—out of the five considered methods, local polynomial regression attained the lowest error.

One of the more recent additions to the load forecasting literature using Czech data is a research article by Lai *et al.* (2020). In their analysis, Lai *et al.* (2020) use ten years of daily peak loads from 2006 in three European countries to present their approach based on deep learning and data augmentation, which the authors describe as a method that generates additional data points. Their technique is shown to outperform several standardly utilized approaches, such as an ARIMA model or an LSTM RNN, on all three datasets using the MAPE metric.

All in all, while it is true that several articles have been published using Czech data, we believe that there is still potential for improvement, particularly

with more recent and higher-frequency data. Moreover, as we have outlined above, some studies evaluate the performance of their methods directly on a test set rather than also employing a validation set (see, for example, Ripley (1996, pp. 7–8)). In this sense, we believe that the latter of the two approaches may provide a better understanding about the validity of the technique, should it be implemented by a grid operator. Additionally, none of the works that we have surveyed in this subsection model high-frequency data (e.g., minute frequency) or compare their forecasts to those produced by the official transmission system operator. Therefore, our goal is to contribute to this growing body of academic literature by addressing these issues.

2.5 Wind, Solar, and Price

Although predicting electricity load is a crucial task in the energy sector, the industry naturally engages in other forms of planning, too. For this reason, Hong *et al.* (2020) review wind and solar power generation as well as electricity price forecasting with the goal of encouraging cooperation across these fields, as they all frequently encounter similar obstacles and ultimately share a common objective. For example, the authors argue that all of these disciplines benefit from the adoption of machine learning approaches, and they also employ the same variables, such as those associated with weather. But they further add that there are also recurring problems across analyses, like the absence or insufficiency of model comparisons, the originality of the dataset, or the improper use of error measures.

As a result of unsustainable increases in man-made emissions, wind and solar power generation capacities have grown significantly each year in the past decade (Ahmed *et al.* 2020; Hanifi *et al.* 2020). The inherent volatility stemming from changes in weather that these two forms of renewable energy generation bring to the system presents the grid operator with a number of additional challenges as these technologies become more widely adopted, which further necessitates the use of forecasting (van Ackooij *et al.* 2018; Ahmed *et al.* 2020). For example, Boldiš (2013) describes that on particularly windy days, excess energy produced by German wind turbine farms gets distributed throughout the power system, and because of the interconnected European electricity grid, it could potentially lead to an overload of the transmission systems of bordering countries.

Furthermore, from the reviews of Ahmed *et al.* (2020) and Alkhayat &

Mehmood (2021), the connection that Hong *et al.* (2020) highlight becomes evident, as wind and solar power generation researchers appear to utilize much of the same techniques and tools as load forecasters do. For instance, Alkhayat & Mehmood (2021) report that hybrid models combining or incorporating ANNs and employing meteorological variables such as wind speed, temperature, humidity, or air pressure, show extensive usage in both domains.

Perhaps one of the most cited researchers in the electricity price forecasting literature is Rafał Weron, with two of the professor's most notable works being a book on price and load forecasting (Weron 2006) and a paper on the advancements in price forecasting (Weron 2014). In the article, the author predominantly reviews the usage of several methods in the literature, many of which are the same as those used in load forecasting, and provides suggestions for the future trajectory of the field, with one of the main ideas being a need for consistent model performance comparison and data reuse. More recently, Lago *et al.* (2018) compared the accuracy of forecasts produced by more than 20 different methods using Belgian data from 2010 until late 2016. In their forecasting exercise, they discover that the best-performing approaches are deep learning models; they further find that hybrid models fail to surpass the base techniques applied individually. We kindly refer the reader to the work of our colleague Křížová (2021) for further details on the topic of electricity price forecasting.

Chapter 3

Background

In this chapter, we start by briefly describing what an electrical grid is. We continue by explaining the principal responsibilities in power system operation. Grid-scale energy storage, renewable sources of electricity, and future issues are then discussed. This is followed by an examination of the role of varying time horizons in the context of load forecasting and power grid management. Next, we briefly investigate the economic costs of producing imprecise forecasts. Lastly, we review some aspects of the Czech electricity industry and its future.

3.1 Power Grid

The key components of a power system are production plants and stations connected by supply lines that constitute an electrical grid, the primary function of which is to *produce*, *transmit*, *transform*, and *distribute* electrical energy (Vlček *et al.* 2019, p. 147; ČEPS 2020). Put briefly, once electricity is generated, it is transformed to a higher voltage with the intention of being more effectively transmitted to supply points. Distribution lines then deliver electricity to consumers at a safer voltage (Crozier *et al.* 2020; EIA 2022).

Understandably, the scale of such systems is inherently enormous. Consider, for example, the size of the United States power grid, which is connected by millions of kilometers of power lines—due to its size, it could be regarded as the *largest machine* on the planet (Richardson 2022).

3.1.1 Maintaining Balance

While electricity shares much of the same characteristics as other commodities, it needs to be consumed almost at the exact moment it is produced, meaning

that the supply of electrical energy must always equal the demand at any given time (Hong 2014; D'Andrade 2017, p. 1). We refer to this condition as *energy balance* (Biggar & Hesamzadeh 2014, p. 60). Any other state in a power grid is highly undesirable, as both an over- and an under-generation may quickly lead to a partial or even a total power outage (Boldiš 2013; Vlček *et al.* 2019, pp. 153–154).

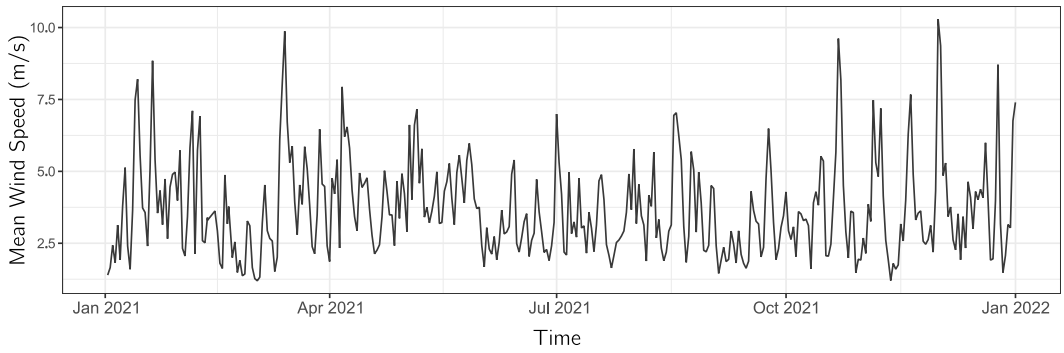
This act of constantly maintaining an equilibrium is further connected to another balancing process which involves keeping the frequency of the system as close to a specific fixed level (50 or 60 Hertz depending on the continent) as possible (Biggar & Hesamzadeh 2014, p. 60; Vlček *et al.* 2019, p. 154), much like maintaining an exchange rate in a fixed regime. Instead of the central bank, however, the entity we have referred to several times in Chapter 2 that is responsible for the operation of this system is called the *system operator* (Harris 2006, p. 13).

The deployment of high-capacity storage, which would be refilled during low-demand hours and then utilized in peak demand, would perhaps be the most straightforward solution to the aforementioned issue of energy balance, and this concept is, to a limited degree, applied in practice mainly through pumped storage hydropower (Soloveichik 2011; Kittner *et al.* 2020). However, coupled with other less employed energy storage methods such as flywheels or batteries, these technologies are becoming more relevant due to the increase in the adoption of renewable sources of energy (Poullikkas 2013), some of which are variable by their nature, as alluded to in Section 2.5. For example, consider wind power, which has a cubic relationship with wind velocity (Center for Sustainable Systems 2021): in Figure 3.1, the mean daily wind speed reported by the Prague-Ruzyně weather station in 2021 is plotted,[1] and it is clear that it varies greatly, creating a stability challenge should a (purely hypothetical) wind farm be built nearby.

However, while it may appear that integrating such a fluctuating source of electricity into the system would be ineffective or counterproductive, several papers, as well as practical experience, show that the introduction of renewables to robust power systems is safe even with no additional requirements for storage (Bowen *et al.* 2019). Additionally, as Bowen *et al.* (2019) state, grid-scale batteries are a versatile resource in a power system for several reasons, perhaps the

[1]The daily averages were calculated using half-hourly data. It is also worth noting that the Czech Republic generated less than 1% of total electricity using wind turbines in 2021 (Ritchie *et al.* 2022).

Figure 3.1: Daily average wind speed in Prague-Ruzyně (2021)



main one being their ability to respond almost immediately to unanticipated supply and demand movements. For instance, the Hornsdale Power Reserve in Australia is arguably one of the most well-known in this regard in part due to its almost immediate response to an instability challenge caused by an issue at a sizable coal plant which averted a "likely cascading blackout" (Bowen *et al.* 2019). This increase in flexibility is also a benefit of other electricity storage technologies (Poullikkas 2013), some of which may currently be more financially viable than batteries (Bowen *et al.* 2019).

These solutions further relate to the concept of a *smart grid*, which has received a great deal of attention relatively recently, and it generally refers to a vision of how the current power infrastructure could be enhanced utilizing contemporary technologies (Tuballa & Abundo 2016). It is also considered as a possible strategy in alleviating the effect of a broader adoption of electric vehicles on the grid (Liu *et al.* 2015), which is expected to represent a significant challenge for the whole system (Crozier *et al.* 2020).

3.2 Significance of Different Time Horizons

In Section 2.2.1, we examined data resolution as one of the parameters that varies across research articles, and to expand on this discussion, let us briefly review the nomenclature in load forecasting regarding time horizons. While there are no universally accepted definitions in the literature (Ahmed *et al.* 2020), Hong & Fan (2016) provide an informal categorization of load forecasting horizons with four classes, to which we attempt to adhere in this thesis: *very short-term* (up to 1 day ahead), *short-term* (1 to 14 days ahead), *medium-term* (14 days to 3 years ahead), and *long-term* (more than 3 years ahead). However, these four categories are frequently simplified to STLF and LTLF, with

two weeks serving as the midpoint. Having introduced the terms for varying horizons, let us now explore the significance of these time periods in practice.

3.2.1 Long and Medium Term

It seems reasonable to believe that obtaining the estimated level of electric load in five years might have different implications for the power system operator and other interested parties than knowing the approximate value in the next thirty minutes. Nonetheless, the overall impression that we got from relevant literature, such as Khuntia *et al.* (2016) or Malik *et al.* (2021), is that the ultimate goal appears to be shared—to ensure that electricity is reliably supplied to those who demand it.

Starting with long-term projections, activities such as component upkeep and transmission network development planning are undertaken (Khuntia *et al.* 2016), the latter of which may, for example, comprise the construction of new power lines (Mahdavi *et al.* 2019). However, it is worth noting that these grid investments occur in the medium term, too (Khuntia *et al.* 2016). Nevertheless, this task is an inherently complex problem in optimization, and it has received a substantial amount of attention in academic literature in the past 50 years, according to Mahdavi *et al.* (2019).

In the medium term, maintenance procedures are scheduled (Khuntia *et al.* 2016; Malik *et al.* 2021). The primary goal of these processes is to maximize equipment lifespan and minimize the likelihood of unanticipated interruptions (Khuntia *et al.* 2016). In this context, Harris (2006, p. 55) provides a model example that around 16% of uptime tends to be lost to outages (both scheduled and unplanned) in a standard two-decade-old coal plant.

3.2.2 Short and Very Short Term

One of the key procedures for which short-term load forecasting is crucial is *unit commitment* (Hong & Fan 2016; Malik *et al.* 2021). It can be described as a cost-minimization problem that, given a particular set of constraints, attempts to identify the *best* electricity generation schedule using the existing production resources (van Ackooij *et al.* 2018). If the choice of whether (and when) to turn power-generating units on or off is not considered, the problem is termed *economic dispatch* or *optimal power flow* in the context of a network (van Ackooij *et al.* 2018), both of which are mentioned by Malik *et al.* (2021), further adding that these activities are undertaken in the very short term as well. In

their review, van Ackooij *et al.* (2018) also state that there is a trade-off between accuracy and speed in unit commitment modeling, referring to the time period necessary to produce a solution as "unreasonably short."

In the very short term, *contingency analysis* is one of the critical tools in ensuring uninterrupted power grid operation by simulating the consequences of network component issues (e.g., power line outages) due to unanticipated failures, weather, or load variability (Khuntia *et al.* 2016; Coelho *et al.* 2019). Real-time system management then benefits from previous operational preparations, which aim to minimize as much uncertainty as possible during this process, as stated in Khuntia *et al.* (2016).

3.3 Cost of Inaccurate Load Forecasts

Overall, many of the stages mentioned in the preceding section involve load forecasting as a key component (Feinberg & Genethliou 2005; Liao & Tsao 2004). For example, borrowing from Fan & Hyndman (2012), predicting future loads determines how much *spinning reserve*—readily accessible electricity production capacities (Kirschen & Strbac 2004, p. 110)—should be prepared.

Moreover, Fan & Hyndman (2012), as well as Liao & Tsao (2004), emphasize the role of accuracy, stating that upward-biased forecasts may result in financial losses, particularly when *ex-ante* expectations would have dictated purchasing extra energy or an additional power-generating unit needing to be turned on (Ranaweera *et al.* 1997). In the opposite case, that is, when loads are *underpredicted*, Hobbs *et al.* (1999) state that costs can accumulate for reasons such as the increase in risk associated with lower reserves or the need for "uneconomic" electricity production as well as purchasing.

Unlike academic literature on load forecasting, economic assessments of prediction errors appear to be somewhat limited. Understandably, Hong *et al.* (2020) maintain that these estimates are "quite difficult, if ever possible" to conceive. Regardless, some researchers have tried to quantify the cost of producing imprecise forecasts; for instance, Hobbs *et al.* (1999) concentrate on day-ahead hourly predictions for the purposes of unit commitment. In their case study, they found that a MAPE equal to about 1% had essentially no effect on the cost of production, while a MAPE of 5% significantly increased these expenses by at least 0.35%, with the upper bound reaching around 0.5 percentage points higher in one of their settings. However, in more recent works, such as Wang & Wu (2017), the focus seems to be pivoted toward utilizing these potential costs

for generating better forecasts—the authors are also concerned with unit commitment, and they weigh predictions of various models based on their economic effect on the process.

3.4 Electric Power Sector in the Czech Republic

The energy industry in the Czech Republic has experienced a large degree of development following the country's transition to democracy, which was set in motion in 1989 (Vlček *et al.* 2019, pp. 28–29). Since 2003, the Czech Republic has consistently generated between 81 and 88 *terawatt-hours* (TWh) of electricity annually—roughly 30% more than in the 1990s (Ritchie *et al.* 2022). As a member of the *European Union* (EU), the country produced around 8 *megawatt-hours* (MWh) *per capita* in 2021, which was nearly 2 MWh above the Union's average, based on the data from Ritchie *et al.* (2022).

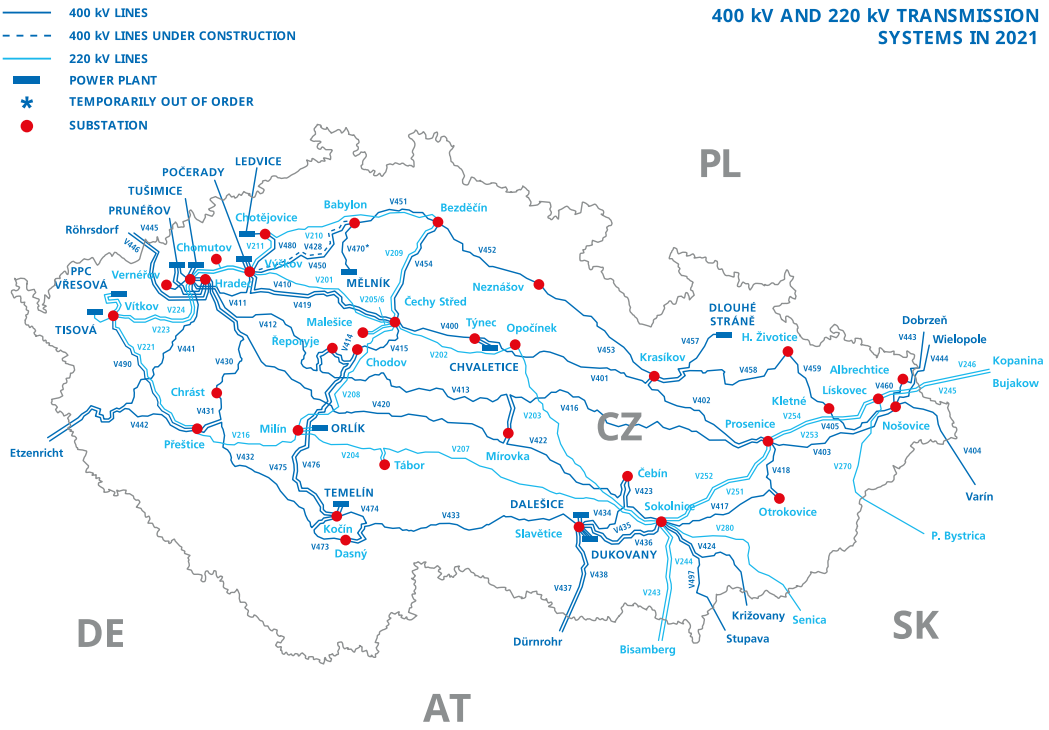
3.4.1 Transmission System

In the Czech Republic, the transmission part of the electricity grid is managed by ČEPS, a. s.,[2] which was established in 1998 and currently has a single shareholder—the Ministry of Industry and Trade (Vlček *et al.* 2019, p. 149). As the sole license holder, ČEPS has a legal obligation to maintain the transmission system, ensuring that it operates uninterrupted in a "safe and reliable" manner (ČEPS 2020).

In 2021, the transmission grid of the Czech Republic comprised 5703 kilometers of major power lines, around two-thirds of which were 400 *kilovolt* (kV) lines (ČEPS 2021). A recent illustration of the 400 and 220 kV networks can be seen in Figure 3.2, together with substations and the largest power plants in the country, most of which produce electricity by burning coal. The schematic in Figure 3.2 further includes 17 power lines that connect the Czech transmission system directly with all of its neighbors and indirectly with the rest of the nations that are a member of the continent's main subnetwork, known as the *synchronous grid of Continental Europe* (Hofmann *et al.* 2020). As Vlček *et al.* (2019, p. 149) describe, such a high degree of connectedness allows for considerable cross-border flows and electricity trading that ensure higher reliability while also introducing risks from unanticipated overflows or other difficulties—one of which we have described in Section 2.5.

[2]Hereinafter solely referred to as *ČEPS*.

Figure 3.2: Czech 400 kV and 220 kV electricity transmission networks in 2021



Source: Republished with permission from ČEPS (2021).

3.4.2 Electricity Mix and Decarbonization

Regarding the Czech electricity mix, which can be extracted and analyzed in context through the extensive report of Ritchie *et al.* (2022), around 40% of electrical energy is currently generated by coal-fired power plants, but this proportion has been steadily decreasing since 1991. On the other hand, the usage of nuclear power, which accounted for approximately 36% of total generation in 2021, has been more or less increasing in the past decade, which could also be said for renewable forms of electricity. Furthermore, according to the data compiled by the authors, only 11 countries in the world reported producing more than 30% of their total electricity using nuclear power plants in 2021, with the Czech Republic being one of the few where an increase in the adoption of this source of energy has been observed in the last couple of years. It is worth noting that the stance of many countries on nuclear power seemed to have begun to shift in 2022, as many started reconsidering it as a "part of the answer" to the issues related to energy security, which has been challenged after the Russian invasion of Ukraine, as well as the ever so pressing issue of climate change, as per The Economist (2022a).

According to a recent assessment by the *Intergovernmental Panel on Climate Change* (IPCC), a rise of 1.5 °C in comparison to the pre-industrial era is anticipated to materialize between 2030 and 2052 (IPCC 2022, p. 4). Notwithstanding its impact on ecosystems around the world, this increase is going to have a number of detrimental effects on human health and security, the authors warn. Perhaps the most direct example of the impact of global warming could be illustrated on small island states, like the Maldives, which will nearly cease to exist as a result of an increase in seawater levels (The Economist 2022b). Therefore, one of the goals that governments must continue to pursue is to reduce carbon emissions (IPCC 2022, p. 276).

In this context, the Czech Republic ranked second among its neighbors in terms of coal-generated power *per capita* in 2021 (3.236 kilowatt-hours), while Poland produced the most (3.437 kilowatt-hours) per person. In comparison to the rest of the world, the nation's proportion of electricity generated using low-carbon sources was about 10 percentage points higher than the global arithmetic mean—in this regard, the Czech Republic has been above average for almost two decades. Since at least 1985, however, the situation has more or less been the exact opposite in the European context, i.e., always roughly 10 percentage points below the mean of all the member states of the European Union (Ritchie *et al.* 2022). Furthermore, according to the European Parliamentary Research Service, the Czech Republic's pace in lowering emissions was slower in the past few years when compared to the EU average (Jensen 2021). In a recent McKinsey report, the company states that reaching carbon neutrality in 2050 would require a faster rate of decarbonization than currently proposed—a process that would entail introducing significant technological alterations to each sector of the economy (Hanzlík *et al.* 2020).

Chapter 4

Data Description

This chapter is divided into two parts: a description of the acquired datasets and an exploratory analysis. The first section gives an overview of the load, weather, and price data we gathered, as well as information on how missing variables were treated. The other part of this chapter presents summary statistics and data visualizations.

4.1 Datasets

In Section 2.2.3, we outlined the factors that researchers need to consider before modeling load. The general idea is that widely different variables can be employed depending on the scale and the resolution of the data in the analysis. However, we have also seen that, in some cases, no additional independent variables are utilized by researchers, meaning that they exclusively use past values and seasonal indicators. While this is also a possibility, Hong & Fan (2016) maintain that the utilization of, for example, meteorological data tends to be favorable as it often results in generating better forecasts, but it may be challenging for researchers to obtain these variables freely, especially high-frequency series, as we have learned.

In this analysis, we compiled data from three sources.[1] Firstly, we acquired Czech load data from ČEPS, the Czech transmission system operator (ČEPS 2022). Meteorological variables were collected from two weather stations in the country through the *Integrated Surface Dataset* published by the *National Centers for Environmental Information* (NOAA 2001). Finally, we further as-

[1]The links to the data sources can be found in the respective bibliographical entries.

sembled price data from the *Czech Electricity and Gas Market Operator* (OTE 2022). Let us discuss each of these datasets in more detail below.

4.1.1 Load

The electric load data used in our analysis was gathered over an 11-year period, starting on January 1st, 2011, and ending on December 31st, 2021. Our dataset initially included 2010 loads, but we decided to shorten the timespan by a year due to missing values of the other series. The data is at a national scale (i.e., the entire Czech grid), and it was collected at minute intervals, later aggregated by averaging to an hourly frequency. In the description of the data, ČEPS (2022) describe load as the "instantaneous amount of active power" in MW, further decomposing *gross load*, i.e., including losses and power consumption of plants, into:

$$\text{Load} = \text{Generation} + \text{Import} - \text{Export} - \text{Pumping}.$$

This data had a very high level of quality. The only issue we had to resolve was that two observations (four in the minute data) contained unreasonably large values. These were replaced using linear interpolation, which is a method we also utilized for weather variables, and we discuss it in more detail in the next subsection (Section 4.1.2).

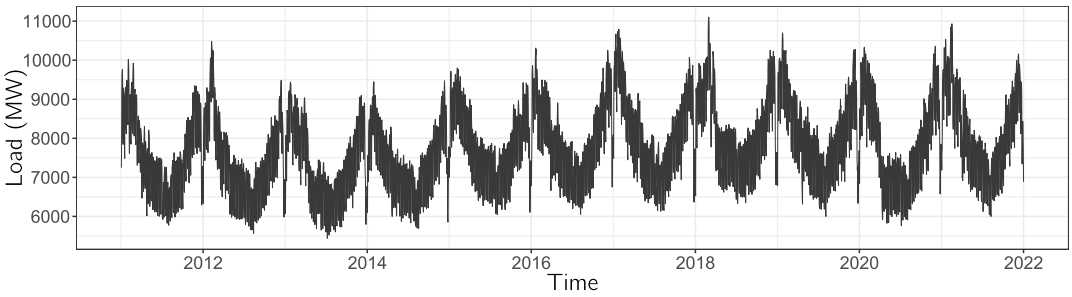
We display the entire load time series in Figure 4.1 as a line plot with daily load (aggregated for visual clarity) in MW on the vertical axis and time on the horizontal axis. Comparing Figure 4.1 to the daily peak loads in South Korea from 2014 to 2019, pictured in the article of Lee & Cho (2022), both series seem to have a slight upward trend from 2014 up to 2018, meaning that both countries gradually produced more electricity during this period, but this increase was later followed by a minor decrease or rather a stabilization in the case of our data. However, perhaps the most noticeable feature of the load data in Figure 4.1 is that the series exhibits multiple levels of seasonality.

To explore further, let us, for example, focus on the data from 2021 and expand the line plot by another dimension representing the hour of the day, creating a surface plot[2] (Figure 4.2). This visualization helps us identify several cycles in the load data, such as an increase in load during the winter and a decrease in the summer.

Returning to a two-dimensional graph, Figure 4.3 focuses on an arbitrarily chosen month (November 2021) and a particular day within that month (11th

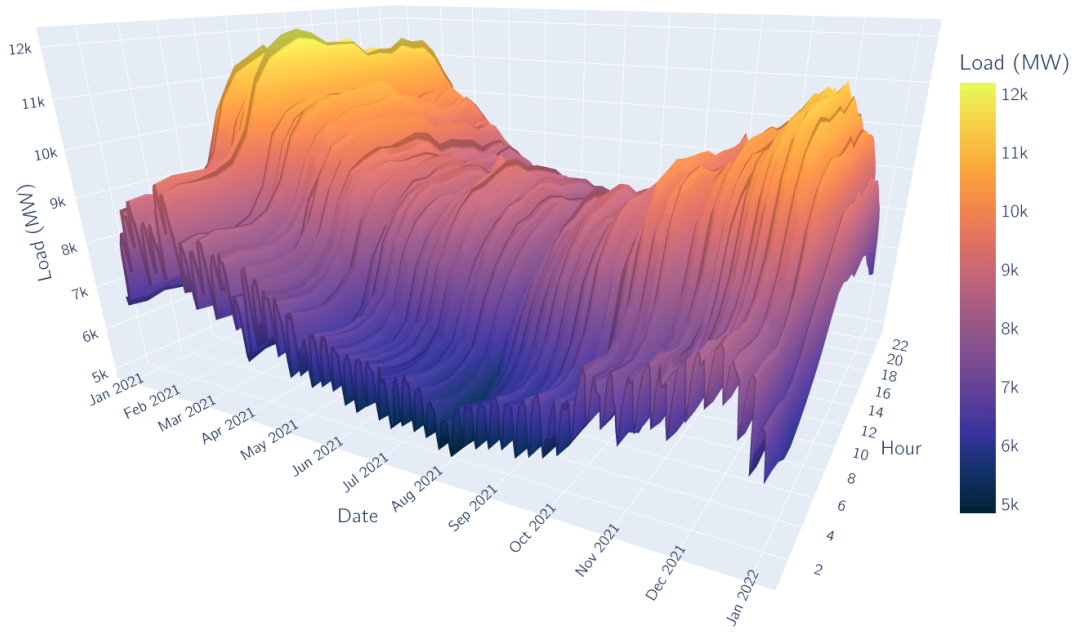
[2]Figure 4.2 was inspired by the surface plot of power consumption in Uher *et al.* (2015).

Figure 4.1: Daily load in the Czech Republic from 2011 to 2021



Note: Aggregated to daily data by averaging for visual clarity.

Figure 4.2: Daily load in the Czech Republic in 2021 expanded by hour of the day



of November 2021) for illustrative purposes. These charts display that, apart from a day/night pattern, there is a decrease in electricity consumption during the weekends. To better demonstrate the latter point, we plot a load heatmap in Figure 4.4 with hours of the day on the horizontal axis and days of the week on the vertical axis (similar to Wang *et al.* (2021)). Clearly, on average, electricity consumption was lower on Saturday and Sunday than during workdays based on the 2011 to 2021 national-level data.

Moreover, another calendar effect is perhaps best visible in the first load chart (Figure 4.1): there is a sharp decrease right before the end of each year, which is due to winter holidays—for instance, in their analyses, Fan & Hyndman (2012) and Yildiz *et al.* (2017) add dummy variables for other national holidays

as well, because loads exhibit a similar pattern during these days.

Figure 4.3: Seasonal patterns in hourly Czech electric load in 2021

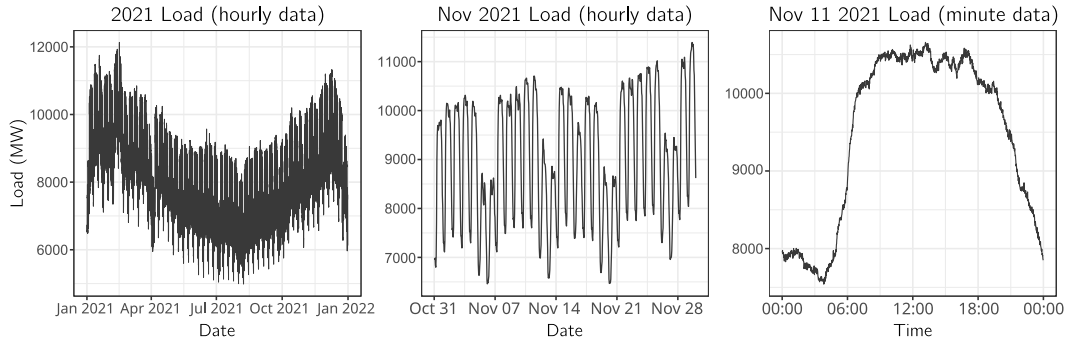
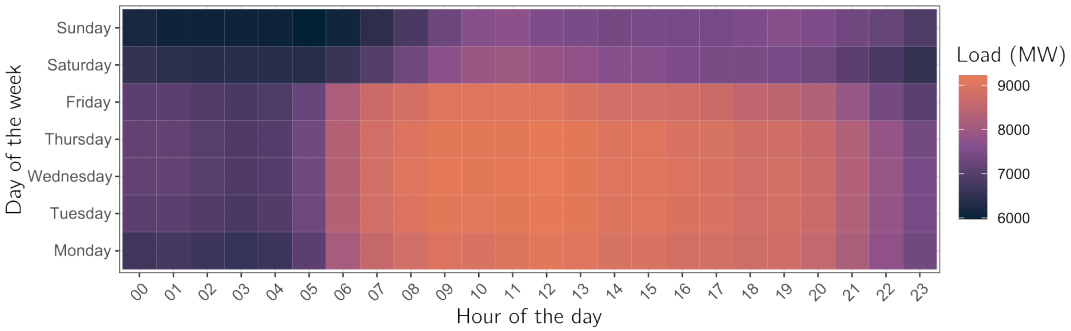


Figure 4.4: Mean hourly load by day of the week (2011 to 2021)



4.1.2 Weather

With respect to meteorological data, we collected weather variables from two stations in the Czech Republic, namely Ruzyně and Brno-Tuřany. There were three reasons for the choice of these two locations:

1. Unlike most stations in the Integrated Surface Dataset, both Ruzyně and Tuřany offer higher frequency (30-minute) weather data.
2. We believe that the two stations provide a more accurate representation of the weather across the entire nation because they are roughly on opposing sides of the country.
3. Finally, the approach of combining weather data from two stations has also been applied in other academic literature—namely Fan & Hyndman (2012).

For the data to be consistent, several pre-processing steps were required. Firstly, we needed to *synchronize* the data from the two weather stations by

finding missing values. Once these observations were identified for each of the two stations, we imputed the missing data (less than 0.5% of all values) in two ways. Either we *borrowed* observations from the other station if the first location was missing a particular record and the other one was not, or we used the best of three interpolation methods (spline, linear, and Stineman following Moritz & Bartz-Beielstein (2017)) to fill in the values. That is, for each variable, we selected the method that produced values closest to the 10-year mean of the observation in question in absolute terms. However, using this approach, the most accurate technique was linear interpolation in all cases. As per Lepot *et al.* (2017), this approach fits a line through two points to find the missing values. In other words, we may find x_b using x_a and x_c in the following way:

$$x_b = \frac{x_c - x_a}{c - a} \cdot (b - a) + x_a,$$

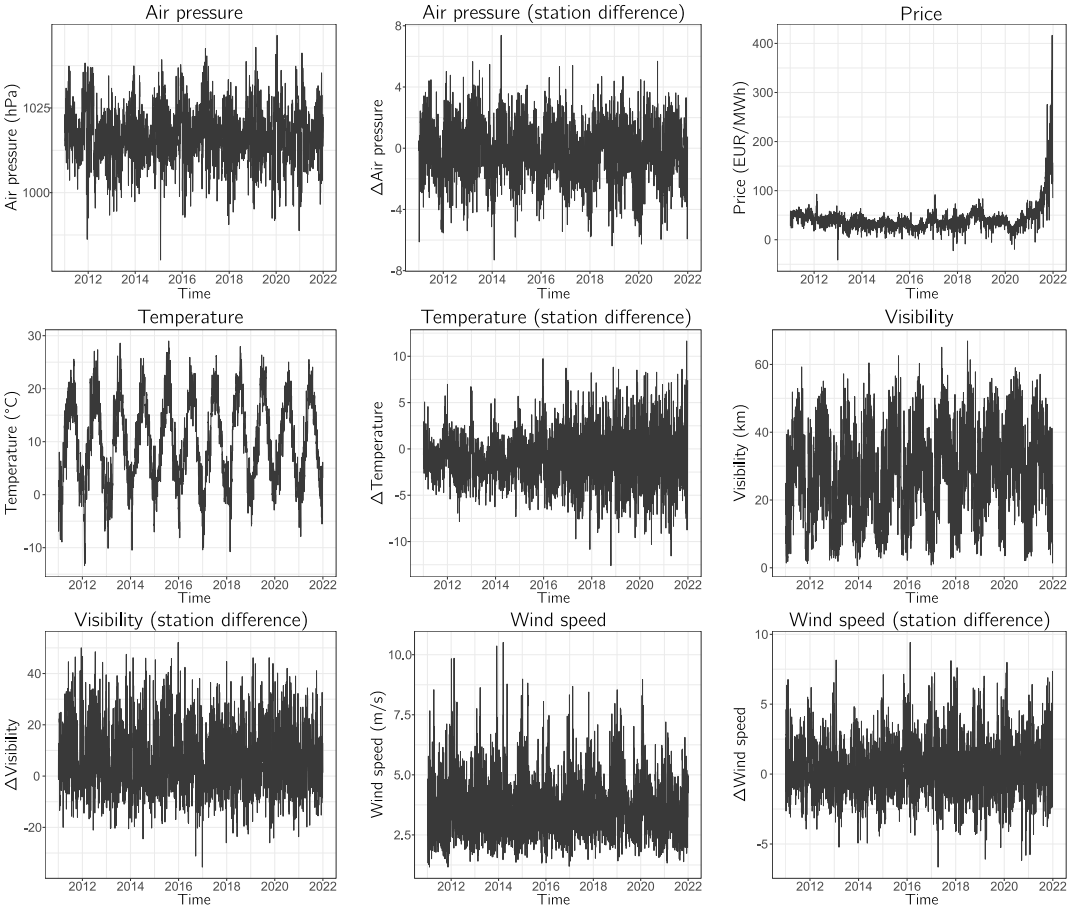
where, in the time series context, x_a and x_c are values that are observed before and after x_b, respectively. Due to the minimal number of missing values and the convenience of working with a complete series, we preferred this method over deleting the missing data.

Then, we aggregated the half-hourly data to hourly by averaging, mainly due to the fact that the predictions that we wanted to compare against were produced using hourly data. The other reason was that two variables (air pressure and visibility) were not reported on a half-hourly basis. After all of these processing steps were completed for each station, we combined the data from the two locations by calculating their means and differences, based on Fan & Hyndman (2012). Their reasoning is that highly similar data from two weather stations in relatively close proximity provide almost no new information. Additionally, the authors further motivate the usage of taking differences by stating that the newly produced data ought to be "almost uncorrelated" with the means. Thus, the final set of hourly meteorological variables included averages and differences between the two stations of the following series (plotted in Figure 4.5): air pressure in *hectopascal* (hPa), temperature in *degrees Celsius* (°C), visibility distance in *kilometers* (km), and wind velocity in *meters per second* (m/s).

4.1.3 Prices

As we described in Subsection 2.2.3, the use of price data in generating load forecasts does not seem to be prevalent in academic literature. Regardless,

Figure 4.5: Line plots of hourly price and weather data (2011 to 2021)



Note: Aggregated to daily data by averaging for visual clarity. "Δ" refers to the difference between 2 weather stations.

we decided to obtain price data to see whether their inclusion would help in producing more accurate predictions.

Therefore, we compiled day-ahead market price data (as per Weron (2014)) from the Czech electricity and gas market operator. These were available at an hourly frequency and were denominated in *Euros* (EUR) per megawatt-hour. We plotted the price data together with the weather variables in Figure 4.5—let us note that negative values can occur in electricity prices due to supply and demand imbalances (Sewalt & De Jong 2003).

4.2 Exploratory Analysis

Since we already discussed various characteristics of the load series, let us start this section by commenting on some of the features of the weather and price data pictured in Figure 4.5. Except for prices and the station-differenced se-

ries, all the raw meteorological variables seemed to expectedly exhibit seasonal patterns in the surveyed timeframe, with temperature being perhaps the most apparent exemplar of highly cyclical behavior. Moreover, this series appeared to develop quite similarly to load—in fact, the Pearson and the Spearman correlation coefficients of these two series were -0.47 and -0.46, respectively (Table 4.1), suggesting a negative relationship. Being mindful of the statistical mantra that correlation does not imply causation, in Section 2.1, we discovered that a link between these two variables was articulated decades ago, if not earlier.

The two-station average temperature readings ranged from -15.4 °C to 35.5 °C from 2011 to 2021, with the standard deviation being equal to 8.3 °C (Table 4.2). In addition, the middle chart in Figure 4.5, representing the difference in temperatures between the two locations, appears to have increasing variance over time, which we consider an intriguing revelation for which we fail to supply an explanation. The difference in wind speed, air pressure, or visibility does not seem to follow the same trend, however.

Table 4.1: Spearman and Pearson correlation coefficients of all variables (hourly data from 2011 to 2021)

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Load	—	0.10	-0.10	0.49	-0.46	0.08	-0.29	-0.13	0.07	0.05
2. Air pressure	0.09	—	0.17	0.06	-0.14	0.08	-0.11	-0.04	-0.08	-0.00
3. ΔAir pressure	-0.11	0.18	—	0.01	0.13	-0.08	0.07	-0.11	-0.02	-0.08
4. Price	0.35	0.04	-0.01	—	-0.07	0.02	-0.15	-0.08	-0.02	0.01
5. Temperature	-0.47	-0.14	0.12	-0.10	—	-0.11	0.45	0.16	-0.04	-0.07
6. ΔTemperature	0.09	0.07	-0.07	0.03	-0.10	—	-0.07	-0.02	0.01	0.34
7. Visibility	-0.29	-0.11	0.06	-0.10	0.43	-0.08	—	0.46	0.02	0.01
8. ΔVisibility	-0.13	-0.04	-0.11	-0.06	0.15	-0.03	0.47	—	0.01	0.05
9. Wind speed	0.08	-0.06	-0.03	-0.03	-0.06	0.02	0.01	0.01	—	0.02
10. ΔWind speed	0.06	0.00	-0.09	0.03	-0.08	0.34	0.00	0.04	0.07	—

Note: Spearman correlation coefficient is in the upper triangle, Pearson correlation coefficient in the lower triangular part. Further, "Δ" refers to the difference between 2 weather stations.

From Table 4.1, a relatively strong positive association can be observed between load and price—this relationship can likely be attributed to the difference in pricing between the base and the peak load (Morris 2013). Except for visibility, however, other weather variables displayed a lackluster level of correlation with load.

Regarding prices, from Figure 4.5 and Table 4.2, we may see that the series mostly fluctuated around 30 to 50 EUR/MWh for about ten years. In 2021, however, the day-ahead price seemed to have soared, reaching 620 EUR/MWh

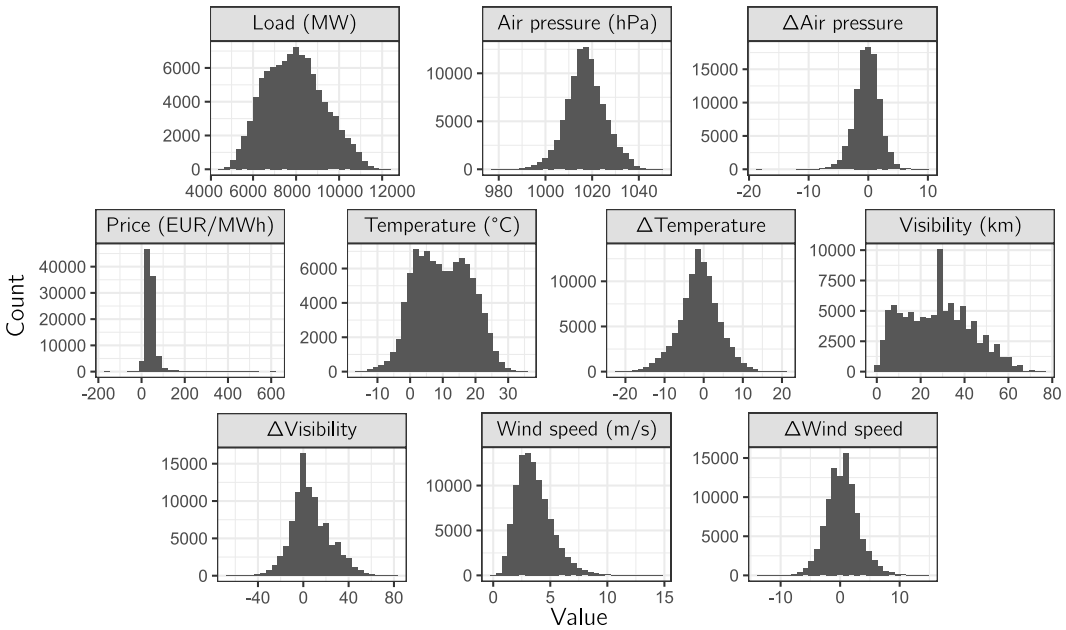
at one point (Table 4.2). Partly due to this increase, the distribution of price in Figure 4.6 is fat-tailed, a frequent property of financial time series.

Table 4.2: Summary statistics of all variables (hourly data from 2011 to 2021)

Variable	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**	**SD**
Load (MW)	4401.1	6888.9	7888.1	7928.5	8857.3	12132.8	1355.1
Air pressure (hPa)	976.7	1012.1	1017.0	1017.2	1022.4	1048.0	8.3
ΔAir pressure	-18.1	-1.4	-0.1	-0.2	1.1	9.8	2.0
Price (EUR/MWh)	-150.0	28.9	38.7	43.9	51.4	620.0	32.1
Temperature (°C)	-15.4	3.4	9.8	10.0	16.6	35.5	8.3
ΔTemperature	-22.3	-4.0	-1.0	-1.1	2.0	20.2	5.1
Visibility (km)	0.1	15.0	27.5	27.9	40.0	76.0	15.5
ΔVisibility	-65.0	-5.0	5.0	6.8	15.0	81.8	16.9
Wind speed (m/s)	0.2	2.4	3.3	3.6	4.4	14.8	1.6
ΔWind speed	-13.2	-1.2	0.3	0.4	2.0	14.8	2.7

Note: "Δ" refers to the difference between 2 weather stations.

Figure 4.6: Histograms of all variables (hourly data from 2011 to 2021)



Note: "Δ" refers to the difference between 2 weather stations.

Moreover, from 2011 to 2021, hourly loads in the Czech Republic ranged between 4401.1 MW to 12132.8 MW (Table 4.2). Although the mean and the median of the series seemed to be somewhat close, the distribution of loads visually appears to be slightly right-skewed and light-tailed (Figure 4.6). In Appendix B, we also provide the summary statistics for the 1-minute load

time series in Table B.1—we may observe that in the higher-frequency data, the minimum was almost 60 MW lower than in the hourly data, while the maximum recorded load was more than 400 MW higher.

Focusing further on Figure 4.6, an interesting feature that can be found in the weather data is that the temperature distribution seems to contain two local maxima—one slightly above zero and the other close to 15 °C. Moreover, from the distribution of visibility, values near 30 km tended to be reported the most by the two stations, on average. Because we have little to no way of verifying whether such a high number of occurrences of observations close to the number in question is sensible, we can only speculate that perhaps the value is some default setting reported by the sensors that record visibility measurements in weather stations.

Chapter 5

Methodology

In the first section of this chapter, we describe the statistical methods applied in our analysis. Afterward, the use of dummy variables and additional predictors is motivated and explored. The next section then outlines unit root testing and other pre-estimation procedures. In the fourth section, the three forecasting exercises are described. The two subsequent parts then outline parameter and variable selection, while the final part is concerned with forecast error measures.

5.1 Applied Methods

In this study, we use several univariate and multivariate time series frameworks based on the surveyed literature. Let us begin by introducing the simplest—yet occasionally *unbeatable* (e.g., in exchange rate forecasting (Rossi 2013))—model: the random walk.

First of all, we need to introduce the term *stationarity*. If the probability distribution of a time series y_t remains unchanged over time, then y_t is said to be a stationary time series (Stock & Watson 2020, pp. 561–562). In the words of Hyndman & Athanasopoulos (2021, sec. 9.1), the statistical properties of the series are not dependent on the timeframe at which it is examined. The random walk is a prime example of a nonstationary process, as its realizations typically exhibit abrupt increases and decreases, as well as prolonged timespans of upward or downward trends (Hyndman & Athanasopoulos 2021, sec. 9.1). Formally, y_t follows a random walk when

$$y_t = y_{t-1} + u_t, \ t = 1, 2, ..., T, \tag{5.1}$$

where u_t is a white noise term at time t, i.e., a sequence of independent and identically distributed random variables with zero mean and finite variance σ_u^2

(Tsay 2005, p. 64; Box *et al.* 2015, p. 28). It is straightforward to observe that the one-step-ahead prediction of such a model is simply equal to y_t. As a matter of fact, this holds for all forecast horizons (Tsay 2005, p. 64). Thus, this method is often referred to as *naïve* (Hyndman & Athanasopoulos 2021, sec. 5.8).

Applying first differences to the random walk renders the series stationary as Δy_t equals the white noise term, which is stationary by definition (Box *et al.* 2015, p. 28)—if a series needs to be differenced once to achieve stationarity, then it is *integrated of order one* and is said to contain a *unit root* (Brooks 2014, p. 360).

5.1.1 SARIMAX

The random walk model is a special case of an *autoregressive* model of order one, which is written as

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t,$$

with u_t being a white noise series (Tsay 2005, p. 32). Combining the autoregression with a *moving average*, which models y_t as a linear combination of white noise terms $u_{t-i},\ i = 0, 1, 2, ..., q$, and further incorporating orders of integration, d, we arrive at an ARIMA(p, d, q) model. If we utilize the backshift notation, i.e., $B^j y_t = y_{t-j}$ with $j = 0, 1, 2, ...$, we can write the ARIMA(p, d, q) model as

$$(1 - \phi_1 B - ... - \phi_p B^p)(1 - B)^d y_t = \phi_0 + (1 + \theta_1 B + ... + \theta_q B^q)u_t, \qquad (5.2)$$

where u_t is a white noise term, and ϕ_0 is the intercept (Brooks 2014, p. 256; Hyndman & Athanasopoulos 2021, sec. 9.2, 9.5).

Expanding the model further, we may include external regressors by adding them onto the right-hand side of Equation 5.2 (Hyndman 2010). Finally, seasonality can also be approached with an ARIMA-type framework by extending Equation 5.2 to an ARIMA$(p, d, q)(P, D, Q)_s$ model (Hyndman & Athanasopoulos 2021, sec. 9.9), often abbreviated as SARIMA. Combining all of the aforementioned specifications, we arrive at a SARIMAX$(p, d, q)(P, D, Q)_s$ model with k exogenous factors that can be formulated as[1]

$$\phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D y_t = \phi_0 + \theta_q(B)\Theta_Q(B^s)u_t + \sum_{i=1}^{k} \beta_i x_{it}, \qquad (5.3)$$

[1]Note that in the `forecast` R package (Hyndman & Khandakar 2008), the SARIMAX model is implemented in a way that allows the standard coefficient estimate interpretation (Hyndman 2010).

where u_t is a white noise series, ϕ_0 is a constant term, and $x_{1t}, ..., x_{kt}$ are external variables at time t (Hyndman & Khandakar 2008; Papaioannou *et al.* 2016; Lee & Cho 2022). Moreover, to elaborate:

- $\phi_p(B) = 1 - \phi_1 B - ... - \phi_p B^p$ represents the AR(p) part,
- $\theta_q(B) = 1 + \theta_1 B + ... + \theta_q B^q$ is the MA(q) polynomial,
- $\Phi_P(B^s) = 1 - \Phi_1 B^s - ... - \Phi_P B^{Ps}$ determines the order P seasonal AR,
- and $\Theta_Q(B^s) = 1 + \Theta_1 B^s + ... + \Theta_Q B^{Qs}$ being the seasonal MA part of order Q.

In Table 2.1, we mention two papers in which the SARIMAX specification was utilized for forecasting load: Papaioannou *et al.* (2016) and Lee & Cho (2022). For instance, in Papaioannou *et al.* (2016), the authors fit a SARIMAX$(4, 1, 1)(1, 1, 2)_7$ model while employing external factors such as a holiday dummy variable or lagged temperature data.

5.1.2 Regression Trees

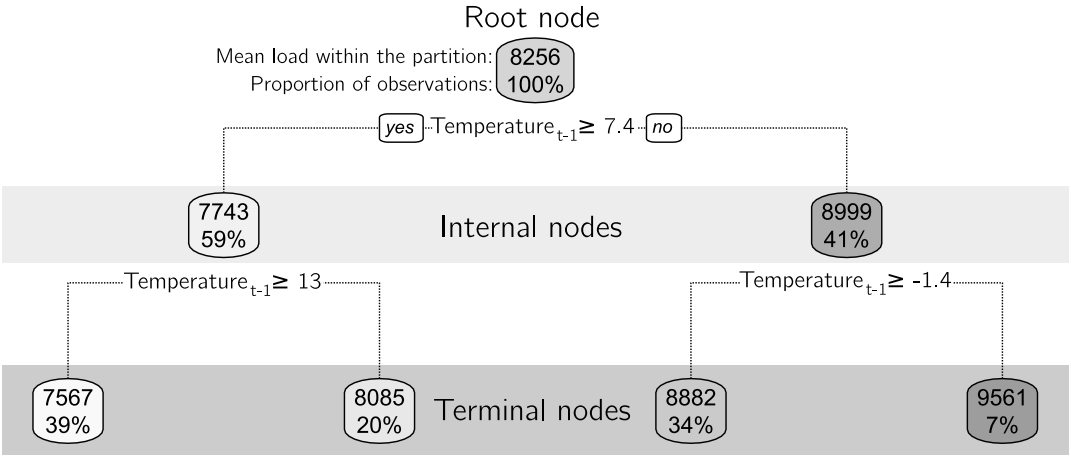
Classification and regression trees are methods that separate data into smaller regions and generate fitted values for the dependent variable, generally as an average for the specific partition (James *et al.* 2021, p. 327). In particular, James *et al.* (2021, p. 330) decompose the construction of regression trees into two stages: the first step aims to identify regions $R_1, ..., R_J$ minimizing the residual sum of squares $\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \bar{y}_{R_j})^2$, where $\bar{y}_{R_j}$ represents the mean of the response variable in region j. The initial stage is performed using an algorithm called *recursive binary splitting*, which finds a feature X_j and a cutoff value $c \in \mathbb{R}$ that reduce the residual sum of squares the most. In turn, this produces partitions $X_j \geq c$ and $X_j < c$ within the set of all possible values of $X_1, ..., X_p$, where p is the number of predictors (James *et al.* 2021, pp. 330–331).

Following James *et al.* (2021, pp. 330–331), the procedure outlined above is repeated up to a certain point (e.g., a preset minimum number of records resulting from a split is achieved), after which the second step is conducted. Since the first stage outputs regions $R_1, ..., R_J$, the fitted values of the dependent variable can simply be obtained as its mean within the particular R_j.

Simpler regression trees can be intuitively visualized—in Figure 5.1, we display an illustrative schematic produced by the `rpart.plot` package (Milborrow 2022) in R (R Core Team 2022) using our load & one-hour lagged temperature data, and we further include the nomenclature used in tree-based modeling. As

per Berk (2016, p. 132), the top part of the regression tree, the *root node*, encapsulates the whole dataset (as indicated by the proportion 100%), and splits into two *internal nodes* based on whether Temperature_{t-1} is larger than or equal to 7.4 °C. Finally, the four *terminal nodes* at the bottom of the schematic are the particular partitions $R_1, ..., R_4$ produced by the algorithm mentioned above (James *et al.* 2021, p. 329). Thus, for instance, using $\text{Temperature}_{t-1} = 15$ °C as the input leads to the predicted load at time t equal to 7567 MW.

Figure 5.1: Example of a regression tree predicting hourly load using lagged temperature data



According to Boehmke (2018) three parameters are typically optimized in regression trees. Firstly, *minimum split* refers to the smallest number of observations in order for a split to be executed. Secondly, *maximum depth* controls the allowed amount of internal nodes that can occur on a single path from a root node to any terminal node. Finally, the *cost complexity* parameter α, borrowing from Berk (2016, pp. 157–158), adds a penalty to the residual sum of squares for each terminal node, resulting in the following objective function being minimized

$$\sum_{m=1}^{|N|} \sum_{i:x_i \in R_m} (y_i - \bar{y}_{R_m})^2 + \alpha|N|,$$

where $|N|$ is the count of terminal nodes of tree N, and R_m is the m-th region (Boehmke 2018; James *et al.* 2021, pp. 332–333).

Moreover, James *et al.* (2021, pp. 340–341) add that standard regression trees tend to have issues with high variance. The authors maintain that it can be lowered using, for example, *bootstrap aggregation*, otherwise known as *bagging*. In this procedure, k different samples are taken from the training

set—on each of these, a regression tree is constructed. A prediction is then produced as an average of all the predictions generated by the k trees.

A useful *byproduct* generated by bagged regression tree models is *variable importance*. This metric indicates, typically in relative terms, the contribution of each explanatory variable to the reduction of the residual sum of squares (Berk 2016, p. 224, James *et al.* 2021, p. 343). However, as Berk (2016, p. 224) maintains, in-sample performance may not necessarily be reciprocated in an out-of-sample exercise.

Regression trees have been utilized for load forecasting by Yildiz *et al.* (2017), as we mentioned in Table 2.1. Ruiz-Abellón *et al.* (2018), on the other hand, employ more sophisticated tree-based methods such as random forests or boosting. In both cases, however, the authors use lagged inputs to be able to generate forecasts, and they further add several types of dummy variables to increase performance.

5.1.3 Artificial Neural Network

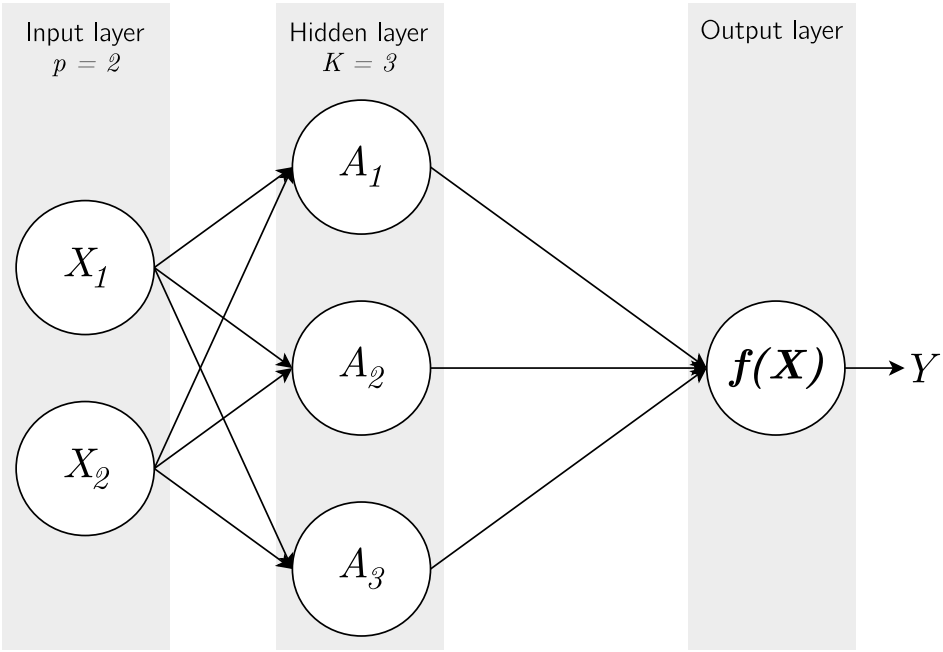
Berk (2016, p. 312) succinctly describes artificial neural networks as "a complicated $f(X)$ [...] approximated by a composition of many, far more simple functions." In this regard, James *et al.* (2021, p. 404) state that the specific *structure* of neural networks, displayed in Figure 5.2, is the key difference from other nonlinear methods. The ANN pictured in Figure 5.2 is called a *multilayer perceptron* (MLP) or a *feed-forward* neural network (Goodfellow *et al.* 2016, p. 5). Here, p predictors in the *input layer* are propagated to the *hidden layer* with K units, and finally to the output layer, producing Y. Thus, following James *et al.* (2021, p. 404), the model $f(X)$ can be written as

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k \underbrace{g(w_{k0} + \sum_{i=1}^{p} w_{ki} X_i)}_{A_k}, \tag{5.4}$$

where $g(z)$ is an *activation function* and $\beta_0, ..., \beta_K$ & $w_{10}, ..., w_{Kp}$ are parameters. An example of an extensively used function for $g(z)$ is the *rectified linear unit* (ReLU), which equals 0 if z is negative; otherwise, the function outputs z (Aggarwal 2018, p. 13).

Researchers tend to specify multiple hidden layers instead of just one, which often facilitates finding more performant models as opposed to increasing K within a single hidden layer (James *et al.* 2021, p. 407). Furthermore,

Figure 5.2: Feed-forward neural network structure example



Source: Based on James *et al.* (2021, p. 405).

when considering time series in the context of ANN modeling, a specific approach is appropriate for the temporal structure of the data (James *et al.* 2021, pp. 421–422). This type of information can be processed using a recurrent neural network—an extension of the feed-forward ANN architecture that adds a feedback loop (Goodfellow *et al.* 2016, p. 164).

In Figure 5.3, an illustration of an *unrolled*[2] RNN is shown. In contrast to the feed-forward neural network pictured in Figure 5.2, we may observe that the A_ℓ units not only process inputs X_ℓ, but also the previous information $A_{\ell-1}$ (James *et al.* 2021, pp. 422–423) or *hidden state* (Aggarwal 2018, p. 39). Additionally, every A_ℓ generates a prediction for Y, denoted as O_ℓ. Similarly to Equation 5.4, this output is calculated as

$$O_\ell = \beta_0 + \sum_{k=1}^{K} \beta_k A_{\ell k},$$

with $A_{\ell k}$ being equal to $g(z)$, where

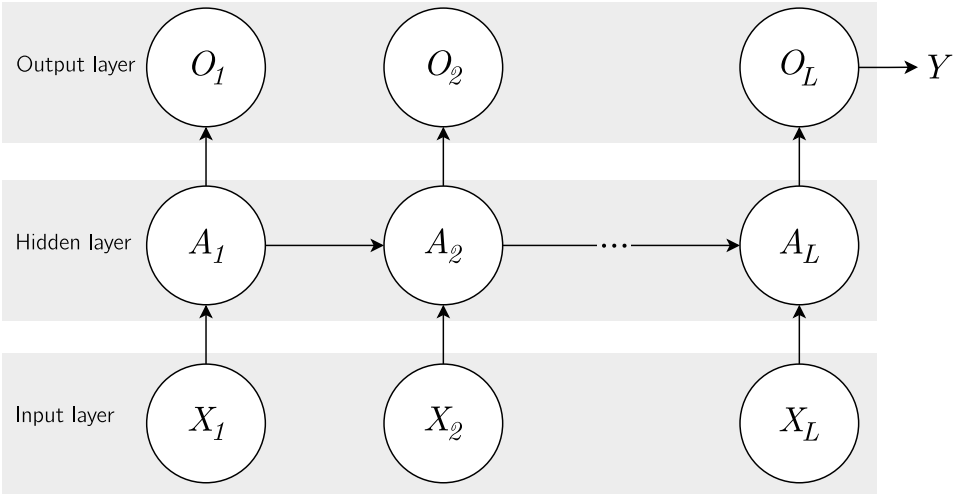
$$z = w_{k0} + \sum_{j=1}^{p} w_{kj} X_{\ell j} + \sum_{s=1}^{K} u_{ks} A_{(\ell-1)s},$$

supposing that $X_\ell^{'} = (X_{\ell 1}, ..., X_{\ell p})$, $A_\ell^{'} = (A_{\ell 1}, ..., A_{\ell K})$, with p being the num-

[2] A verbose, but arguably more descriptive, approach to illustrate the network.

ber of predictors, and K representing the number of units in the hidden layer (James *et al.* 2021, p. 423).

Figure 5.3: Simplified unrolled recurrent neural network structure



Source: Based on James *et al.* (2021, p. 422).

Due to the inherent parameter sharing in RNNs, a significant issue called the *vanishing* or *exploding* gradient problem may arise during training (Goodfellow *et al.* 2016, pp. 286–287, 396). Moreover, even under the assumption of stability, another obstacle in utilizing RNNs is that long-term relationships are given far less priority than short-term dependencies (Goodfellow *et al.* 2016, p. 396).

The long short-term memory RNN architecture is able to process long-term dependencies by augmenting the A_ℓ units from the standard RNN structure (Goodfellow *et al.* 2016, p. 405; James *et al.* 2021, p. 426). In Figure 5.4, the structure of a single LSTM cell is illustrated based on Olah (2015). One of the key additions is the *cell state* C_ℓ, which aims to preserve some of the memory from the previous cells (Aggarwal 2018, p. 293). Furthermore, an LSTM cell has *gates* that regulate the information flow using the sigmoid ($\sigma(z)$) and hyperbolic tangent ($tanh(z)$) activation functions (Aggarwal 2018, p. 293), the formulation of which is given in Appendix A.1.

As per Goodfellow *et al.* (2016, pp. 406–407) & Aggarwal (2018, pp. 293–294), the equations for forward propagation can be expressed as

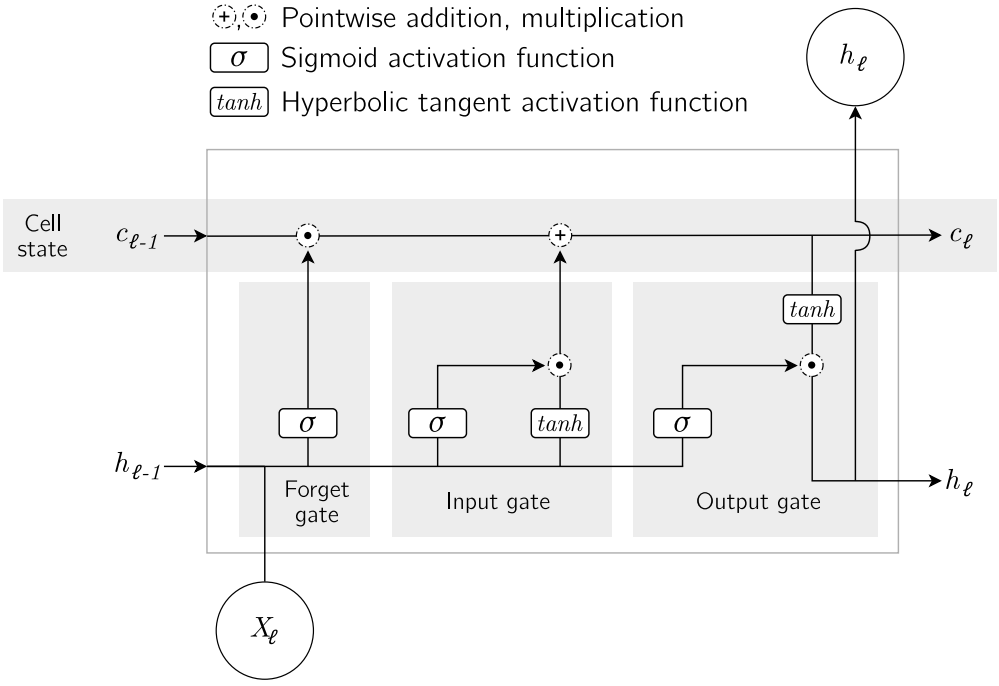
$$f_{\ell k} = \sigma\left(w^f_{k0} + \sum_{j=1}^{p} w^f_{kj} X_{\ell j} + \sum_{s=1}^{K} u^f_{ks} h_{(\ell-1)s}\right),$$

$$g_{\ell k} = \sigma\left(w^g_{k0} + \sum_{j=1}^{p} w^g_{kj} X_{\ell j} + \sum_{s=1}^{K} u^g_{ks} h_{(\ell-1)s}\right),$$

$$
\begin{aligned}
c_{\ell k} &= f_{\ell k}\ c_{(\ell-1)k} + g_{\ell k}\ \sigma\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_{\ell j} + \sum_{s=1}^{K} u_{ks} h_{(\ell-1)s}\right), \\
q_{\ell i} &= \sigma\left(w^{q}_{k0} + \sum_{j=1}^{p} w^{q}_{kj} X_{\ell j} + \sum_{s=1}^{K} u^{q}_{ks} h_{(\ell-1)s}\right), \\
h_{\ell k} &= tanh(c_{\ell k})\ q_{\ell k},
\end{aligned}
$$

where $f_{\ell k}$ represents the *forget gate*, which outputs values from 0 to 1 and acts as a weight, adjusting the amount of information to exclude from the cell state. Similarly, the input gate, $g_{\ell k}$, also produces values in the $[0, 1]$ interval, but controls additions to the cell state. Therefore, the current ℓ cell state, $c_{\ell k}$, is appropriately updated, with $f_{\ell k}$ and $g_{\ell k}$ containing a different set of biases and weights (hence the f & g superscripts, respectively). Additionally, $q_{\ell k}$ represents the output gate, which further dampens the amount of information to pass to the LSTM cell's output $h_{\ell k}$—in this final stage, the *tanh* activation produces a value from -1 to 1, in turn updating $h_{\ell k}$ (Goodfellow *et al.* 2016, p. 406; Aggarwal 2018, p. 293–295; Le *et al.* 2019).

Figure 5.4: Simplified schematic of a long short-term memory cell



Source: Based on Olah (2015).

As remarked by Aggarwal (2018, pp. 187–188), due to the vast array of parameters in neural network methods, the number of feasible specifications

tends to be enormous. Disregarding the input variables or in-sample length, in an LSTM RNN, it is possible to input different values of the following terms

- number of hidden layers,
- units in each layer,
- loss function,
- activation functions,
- number of time steps producing a prediction (*sequence length*),
- proportion of units to randomly omit (*dropout*),
- optimization algorithm or optimizer,
- number of observations to pick in the optimization algorithm (*batch size*),
- step size in the optimizer (*learning rate*),
- processed passes of the entire in-sample set (*epoch*),

among others (Goodfellow *et al.* 2016, pp. 83–84; Bouktif *et al.* 2020; James *et al.* 2021, pp. 436–439, 445). In terms of optimizers, once again, there are numerous options. To give an example of an established algorithm, *Adam* (Kingma & Ba 2014) tends to be viewed as "fairly robust" for different sets of parameters (Goodfellow *et al.* 2016, pp. 305–306).

LSTM RNNs are utilized in several fields, from translation to image recognition (Goodfellow *et al.* 2016, p. 404). From the load forecasting literature that we reviewed in Section 2.3, three works use this method: Marino *et al.* (2016), Kwon *et al.* (2020), and Lee & Cho (2022). Specifically, Marino *et al.* (2016) forecast several steps ahead by employing *sequence to sequence* learning (see, e.g., Aggarwal (2018, pp. 299–303)). In our analysis, following Keydana (2021a) and Keydana (2021b), we implement an LSTM RNN with a two-layered feed-forward ANN and (possible) dropout in between the MLP layers to generate forecasts multiple-steps-ahead. The same approach is applied in the one-step-ahead task, too.

5.2 Additional Predictors

5.2.1 Dummy Variables

Hyndman & Athanasopoulos (2021, sec. 7.4) provide several examples of context-agnostic independent variables that are regularly applied in time series analyses. Particularly, the authors list dummy variables as a powerful tool in accounting for holidays, outliers, but also seasonal patterns. Throughout this thesis, we have referenced many load forecasting papers (e.g., Fan & Hyndman (2012),

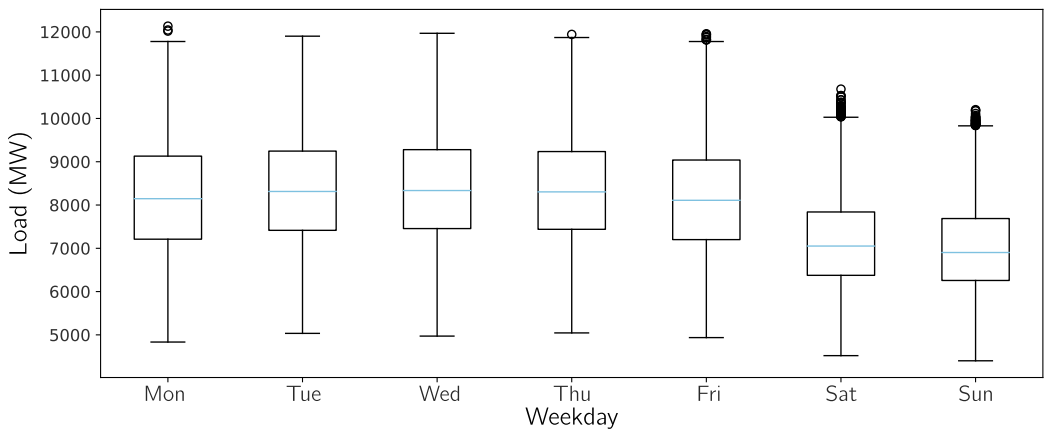
Yildiz *et al.* (2017), or Ruiz-Abellón *et al.* (2018)) that utilize dummy variables mainly to account for seasonality or specific events.

However, while controlling for working days or holidays is common across load forecasting works, each study seems to have a slightly different *mix* of these exogenous predictors. This is also the case in our analysis—the time period of the data that we have obtained intersects with the COVID-19 pandemic. According to Google's Mobility Reports in the Czech Republic, there was a significant decline in the movement of people in public spaces or places of work, especially during states of emergency (Seznam Zprávy 2022). To capture any potential effects of these events on load, we have created a dummy variable that is equal to one during states of emergency declared by the Government of the Czech Republic due to the pandemic and equal to zero otherwise.

The COVID-19 binary predictor was thus implemented into all the dummy variable specifications that we tested. Firstly, we started with a matrix of indicators similar to Ruiz-Abellón *et al.* (2018), which included predictors accounting for the day of the week, month of the year, holidays, and COVID-19. However, specifying a dummy variable for each day of the week did not appear to yield better results than treating Saturday and Sunday as non-working days without the split. This idea was based on Yildiz *et al.* (2017) and supported by the available data—earlier, we provided a heatmap of hourly loads during the week (Figure 4.4), which indicated that there is an apparent difference in the load profiles of weekdays and weekends.

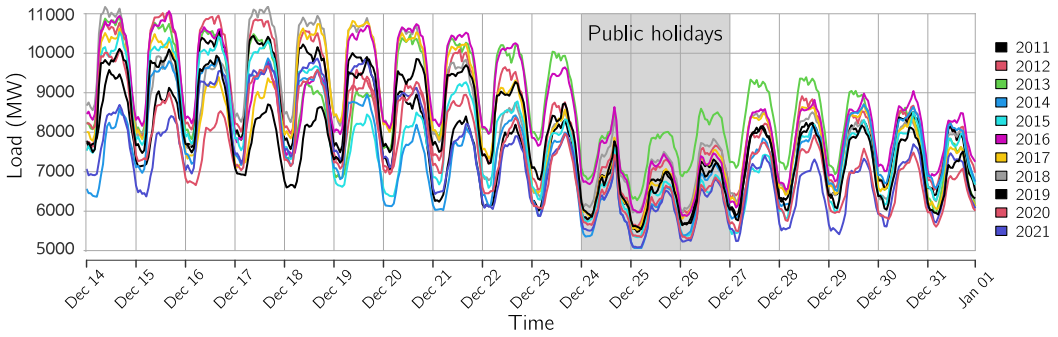
To illustrate this point further, we display boxplots of hourly loads for each day of the week in Figure 5.5. In this chart, we may observe that the difference in the distribution of Saturday and Sunday values is minimal, meaning that it might not be worthwhile including a dummy variable for both days.

Figure 5.5: Czech hourly load boxplots by weekday (2011 to 2021)



While a similar conclusion could be made for the working days, we discovered that separating the week into two exclusive groups seemed to generate worse results than including dummy variables for each day of the week (minding the *dummy variable trap* (Stock & Watson 2020, pp. 229–230)) and a combined predictor for the weekend. Furthermore, we treated both holidays and weekends as non-working days due to their relatively similar load profiles. In addition, as evidenced by Figure 5.6, while the Christmas holidays officially start on December 24 and end on the 26th, it appears that loads tend to continue to record lower levels in the subsequent days. Because of this, even though these final few days of December may have otherwise been classified as working days, we manually labeled them as non-working days.

Figure 5.6: Hourly Czech load each December from 2011 to 2021



Last but not least, we believed that by using a single indicator variable, our models were not able to adequately capture the substantial decrease in load that occurs on weekends and holidays. Strictly speaking, non-working days comprise more than 2/7 of all the observations, which we considered as a compelling argument for intervention. Therefore, comparably to Elamin & Fukushige (2018), we decided to capture the additional information provided by these observations by constructing an interaction term of hours of the day and non-working days. Thus, the final matrix of 63 dummy variables aimed to capture the following effects:

- month of the year,
- day of the week (for each working day),
- non-working days (a single variable for weekends and holidays),
- COVID-19 states of emergency,
- hour of the day,
- interaction of non-working days and hours of the day.

5.2.2 Feature Engineering

Apart from the set of dummy variables outlined in the previous section, we prepared several additional predictors by applying transformations to the variables described in Chapter 4. Following some of the ideas applied in Fan & Hyndman (2012), we transformed each of the base predictors (see Table 4.2) to create X_{t-1}, X_{t-24}, X_{t-48}, X_{t-72}, mean X_t in the past 7 days, and maximum X_t in the last 24 hours.

As a result, a substantial number of features were created.[3] Employing all the available variables would likely lead to overfitting on the in-sample set, which would negatively impact the out-of-sample performance, notwithstanding the computational burden. Thus, as described in Section 5.1.2, we fit a bagged regression tree model and extract the relative importance measure of each variable in order to provide a starting point in variable selection (in the neural network models). We supply more information in Section 5.6.

In future research using Czech data, we believe that it would be worthwhile investigating the performance of more complex weather variables such as heating and cooling degree days employed in Lee & Cho (2022). Alternatively, one could apply dimensionality reduction using principal components (as in Papaioannou *et al.* (2016), for example).

5.3 Pre-estimation Procedures

In standard econometric time series analyses, one of the steps frequently performed before estimating a model is a test for a unit root, the presence of which may be problematic for inference as well as forecasting (Stock & Watson 2020, pp. 584–586). One such approach, the *Augmented Dickey-Fuller test* (ADF), first examines the equation

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + u_t, \tag{5.5}$$

where y_t is the investigated series, and u_t is assumed to be a white noise term. The test then evaluates the hypothesis that $\gamma = 0$, meaning that a rejection of the null hypothesis would suggest no unit root (Brooks 2014, pp. 361–363; Enders 2015, p. 206–208). Furthermore, a constant term and a time trend may be added to Equation 5.5 to account for additional information (Enders 2015,

[3]Refer to Table B.4 in Appendix B for summary statistics of the 7-day averages and 24-hour maxima.

p. 206). In terms of lag length selection, Brooks (2014, p. 363) states that it is crucial to consider several lags and the sensitivity of the test results. Thus, we perform the three variations of the ADF test with up to 10 lags for all variables.

We report the results for each of the raw series in our dataset in an unconventional way in Table B.2 in the Appendix; that is, we display the number of times the null hypothesis was rejected for up to 10 lags to aggregate the large number of outputs. It is clear that, according to the test, a unit root was present in most of the series—for prices, for example, nonstationarity is expected based on Figure 4.5. Series that may not be affected by the presence of a unit root could include the hourly load series, difference in station temperatures, air pressure, and (possibly) raw temperatures (Table B.2).

However, we consider these results as a suggestion rather than a strict rule. One of the reasons is that, following Hyndman & Khandakar (2008), the design of the null hypothesis of the Augmented Dickey-Fuller test tends to *encourage* higher orders of differencing than necessary. This also seemed to be the case with our high-frequency data, as even first-order differencing did not seem to alleviate the supposed issue, but only for lag lengths up to 10—by setting the lag order to the default value suggested by the `tseries` R package (Trapletti & Hornik 2023), none of the previously problematic first-differenced series appeared be integrated of order two (see Table B.3).

Moreover, in neural networks, predictors are typically processed to increase model performance (Aggarwal 2018, p. 127). As per Aggarwal (2018, p. 127), one possible transformation, *standardization*, involves subtracting the mean of the data and dividing the resulting term by the standard deviation. Thus, following Keydana (2021a), we standardize our data before fitting our neural networks.

Seasonal differencing is another possible transformation that could be applied to our data (Hyndman & Athanasopoulos 2021, sec. 9.1). Nevertheless, according to the surveyed load forecasting literature, seasonality tends to be controlled using indicator variables (e.g., Taylor (2008), Fan & Hyndman (2012), or Yildiz *et al.* (2017)).

Finally, in some papers (e.g., Fan & Hyndman (2012) or Taylor (2012)), the authors log-transform their load series as a pre-estimation procedure. According to Taylor (2012), this is generally done to "stabilize the variance of each series." In other works, however, researchers do not seem to apply the logarithmic transformation (e.g., Darbellay & Slama (2000); Kandil *et al.* (2006)).

5.4 Research Questions and Forecasting Schemes

In this study, we develop three forecasting schemes, each addressing a different research question. Firstly, using the one-minute load data, we attempt to outperform a random walk model's forecasts in a pseudo-out-of-sample forecasting exercise, utilizing historical load data only. The remaining two analyses employ hourly data, with the shared goal of producing the most accurate out-of-sample forecasts—one-step-ahead and up to 48-hours-ahead. In the latter pseudo-out-of-sample exercise, our predictions are compared with the forecasts published by ČEPS.

5.4.1 Minute Data

The highest-frequency publicly accessible national load data provided by ČEPS have a frequency of one minute, as described in Section 4.1.1. At this resolution, obtaining any additional predictors, such as weather or prices, is difficult; thus, the outlined pseudo-out-of-sample exercise only utilizes historical loads. In Taylor (2008), the author also analyzes minute-frequency British load series, and implements several univariate forecasting methods, evaluating their performance on a ten-week-long out-of-sample set by generating predictions of up to 30 minutes ahead. One of the tested models was a driftless random walk. While it clearly produced the least accurate forecasts in terms of MAPE from approximately 5 to 30 minutes ahead, its one-step-ahead predictions have, on average, outperformed the official weather-based model used by the transmission system operator in Great Britain. Other methods, however, achieved lower MAPE than the random walk.

Our objective in the first scheme is to demonstrate that the high-frequency Czech load series is also predictable. In other words, following the results of Taylor (2008), we hypothesize that using the Czech load data, it is possible to consistently generate more accurate one-minute-ahead predictions than those generated by a naïve method in a pseudo-out-of-sample forecasting exercise. Naturally, from Figure 4.3, there are several visible patterns that may be exploited in modeling. However, our aim is to determine whether there is enough information to train a model (on a relatively short in-sample set and using detrended series) that outperforms a random walk in the very short term.

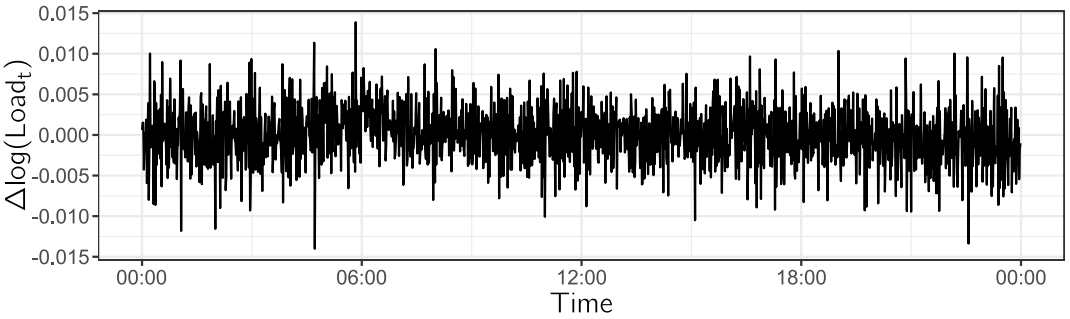
In particular, based on the results from Section 5.3, we are working with the first differences of logs of the entire minute load series spanning from 2011 to

2021—a small subset of the differenced data is pictured in Figure 5.7 for illustration. Thus, rearranging Equation 5.1 by subtracting y_{t-1}, the naïve forecast then predicts no change from the previous period. For $k \in \{2, 3, 4\}$, the overall scheme consists of the following steps:

1. Set k days as the in-sample set, starting at 00:00 on the first day and ending at 23:59 on the k-th day.
2. Set the $k + 1$st day of the minute load series as the out-of-sample set.
3. Using the Hyndman & Khandakar (2008) model selection algorithm[4] with maximum orders of p and q set to 10, fit an ARIMA(p, d, q) specification that minimizes the *Akaike information criterion* (AIC) on the in-sample set.[5]
4. Produce one-step-ahead forecasts of the chosen ARIMA model trained on k days of minute data and the random walk for the entire $k + 1$st day (the out-of-sample set).
5. Compare the predictions with actual values using MAE and RMSE. Additionally, save the results of relevant tests.
6. Move forward in time by one day and repeat the procedure until December 31st, 2021.

Finally, since the length of the in-sample set is also considered as a parameter (2, 3, and 4 days), we select the length that results in the most precise forecasts to simplify the evaluation and reporting of the results.

Figure 5.7: First differences of log-transformed Czech minute load series (Nov 11 2021)



Intuitively, following this scheme, a model is estimated every 24 hours on 2 to 4 days of historical data. The estimated coefficients are then used to

[4]We provide a short summary of the algorithm in the Appendix (Section A.2).
[5]If the lowest-AIC model is ARIMA$(0, 0, 0)$ with zero mean, fit ARIMA$(1, 0, 1)$, instead.

calculate one-step-ahead forecasts in real time—i.e., updating every minute to incorporate the newest load data point as it becomes available and generating a forecast for the next minute. After 24 hours, a new model is reestimated utilizing the latest 2 to 4 days of historical load data, and the process continues. Because the length of the in-sample set is fixed, we may refer to this approach as a *rolling* scheme.

5.4.2 Hourly Data

The main results of this thesis are intended to be produced by the second set of exercises, which focus on multivariate hourly load forecasting using the variables described in Chapter 4 & Section 5.2 and methods outlined in Section 5.1. From a researcher's standpoint, modeling hourly or daily data may be desirable due to the wide availability of other related time series.

Additionally, it appears that multi-step hourly-frequency forecasts might further be significant in the dispatch control of the transmission grid, at least in the Czech Republic—this observation is evidenced by some of the documents ČEPS shares on their website. In particular, in their monthly reports on the preparation of grid operation, one-month-ahead hourly predictions can be found in one of the document's sections (see, for example, ČEPS (2023b)). The use of these forecasts is further corroborated on the "Assessment of power system operation" page (ČEPS 2023a), where ČEPS states that its control center primarily evaluates, among other variables, electricity production, frequency, or system load, and they further add that the key information related to these quantities is made public—in fact, we utilize a part of this dataset in this thesis. Moreover, aside from historical load data, the company also releases hourly load forecasts for the next 48 hours, as well as longer-term predictions of up to a year ahead. When we requested additional information about this data from ČEPS, we were suggested to examine the documents available on their website. Thus, our investigation revealed that despite all the projections being hourly with a varying number of steps ahead, the initial periods where the forecasts intersected (e.g., the initial 48 hours in a 48-hour-ahead and a week-ahead forecast) do not share the same values, which implies that each series of the multiple-steps-ahead predictions is calculated differently. Secondly, as we were mostly interested in the two-day-ahead hourly forecasts, we learned that these are published at around 00:00 every day and calculated until 23:00 the next day. Lastly, in one of their press releases, ČEPS (2013) write about

the introduction of a new geospatial system in dispatch control, which shows "historical information and predictions [. . .] displayed for a time horizon of 48 hours," further solidifying the significance of this horizon.

Therefore, we aimed to produce pseudo-out-of-sample forecasts for the entirety of 2021 by rolling without re-estimation. In particular, these predictions were generated in two ways:

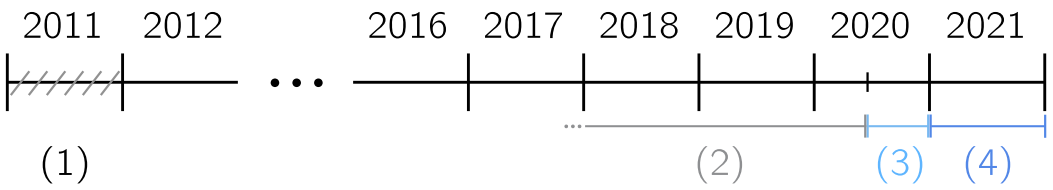
- One-step-ahead, i.e., predict next hour's load, move forward by one hour, repeat.
- Up to 48-hours-ahead, i.e., generate a sequence of 48 load forecasts, advance by two days, repeat.

In the 48-hours-ahead scheme, we then compared these projections of several models not only with the actual values of load, but also with the published predictions. Regarding the one-step-ahead exercise, this approach was performed by, for example, Elamin & Fukushige (2018). In this case, we were interested in generating one-step-ahead forecasts for one year and comparing the accuracy of several methods—as discussed in Section 3.2.2, very short term predictions have a variety of uses, such as real-time system management. Finally, in both schemes, we intended to combine forecasts from the utilized models to determine whether more accurate predictions would be generated, much like in related literature (e.g., Lee & Cho (2022)).

To describe each of the schemes in more detail, we present an illustration in Figure 5.8 outlining the length of the in-sample, validation, and out-of-sample sets in both scenarios.[6] The general objective was to fit each model on the in-sample set, find optimal parameters on the validation set, and produce forecasts out-of-sample. Note that although the underlying structure of the three sets is more or less shared across the two forecasting exercises, the best-performing specifications of the utilized methods should be expected to be different. Furthermore, apart from the obvious fact that the out-of-sample set is the same for all methods, the validation set should also be shared for total comparability. Finally, we see no issue with comparing the pseudo-out-of-sample performance of models with various in-sample lengths, in accordance with related literature (e.g., Dowell & Pinson (2016))—some methods might be expected to provide a better fit with more data while others may fail to utilize longer history. Thus, a minor part of our analysis involved the comparison of several in-sample set sizes.

[6]Note that the in-sample set is often referred to as a *training set*, while the out-of-sample set is frequently called a *test set* (Hyndman & Athanasopoulos 2021, sec. 5.8).

Figure 5.8: Subsets of hourly data used in the pseudo-out-of-sample forecasting exercises



(1) Discarded, *1st Jan 2011 – 31st Dec 2011*
(2) In-sample set, *? – 31st May 2020*
(3) Validation set, *1st Jun 2020 – 31st Dec 2020*
(4) Out-of-sample set, *1st Jan 2021 – 30th Dec 2021*

Regardless, in similar academic literature, the choice of the length of each of these sets varies. For instance, as mentioned in Section 2.3, the dataset of Fan & Hyndman (2012) originally spanned from 1997 up to March 2009, but the authors discovered that an in-sample set starting in 2004 did not compromise the forecasting performance of the models evaluated on an out-of-sample set of 6 months. In particular, they write that "increasing the size of the data set [...] is not always helpful" since the association between load and other factors could be gradually shifting in time.

As displayed in Figure 5.8, the 31st of May 2020 is assigned as the end date of the in-sample set. Directly following is the validation set, which we use for variable selection and also for finding optimal parameters of the utilized methods (refer to Section 5.5). Note that, in this sense, in-sample set length can be thought of as a parameter, too. Moreover, while it might seem sensible to consider the entirety of 2020 as the validation set, doing so would prevent us from capturing the effects of COVID-19 in the in-sample set; therefore, we consider the chosen number of months as a reasonable length for model calibration. Lastly, variable and optimal parameter selection is primarily performed by minimizing forecast error measures on the validation set while also considering in-sample fit (see, for example, Goodfellow *et al.* (2016, pp. 118–119)). These error metrics are described in Section 5.7.

In related academic literature, researchers typically produce forecasts on up to 1 year of *unseen* data (e.g., Taylor (2012) or Lee & Cho (2022)). Thus, as pictured in Figure 5.8, our out-of-sample set encompasses almost the entirety of 2021—the final day of the year was removed because it is the 365th day (i.e., not divisible by two). Finally, let us note that we also excluded observations

from 2011 due to the fact that the calculated features described in Section 5.2.2 naturally introduced missing values in the first month of the year.

5.5 Model Parameter Selection

Except for the random walk model, each of the methods described in Section 5.1 contains several parameters. While some of them can be selected manually—for example, it might be sensible to set the s parameter in the SARIMAX model (Equation 5.3) to 24 with hourly data to exploit the daily seasonal pattern (Elamin & Fukushige 2018; Hyndman & Athanasopoulos 2021, sec. 9.9)—however, more often than not, parameters are chosen in an automated manner (Goodfellow *et al.* 2016, pp. 422–423).

Thus, for the three methods that we utilize in this thesis to predict hourly data, i.e., SARIMAX, bagged regression trees, and LSTM RNN, we conduct a *grid search* (Goodfellow *et al.* 2016, p. 427) to determine the best-performing specifications on the validation set in both the one-step-ahead and the 48-hours-ahead schemes. Table 5.1 contains the final parameter sets we converged to through several rounds of manual searches. For instance, in the SARIMAX model, setting $d = 2$ generally did not appear to generate better results than lower values of the order of differencing ($D = 2$ as well). In addition, we tested different values of s, but achieved the most satisfactory results with $s = 24$, as alluded to in the paragraph above. Furthermore, it required numerous attempts to find reasonable parameter values in the bagged regression trees because some produced high-error forecasts on the validation set, while setting other values resulted in overfitting on the in-sample set.

Therefore, while the rightmost column in Table 5.1 provides the final number of estimated models, the actual amount could be two to three times larger for several additional reasons other than those described earlier, including tests of various in-sample sets or variables, as well as trial and error. In particular, with respect to the in-sample size, the grid search was conducted using data starting in 2017 (adhering to the scheme pictured in Figure 5.8). Later, the best-performing specifications were tested *horizontally* with longer and shorter lengths by year. Let us note that we are aware of the limitations of this heuristic approach—ideally, varying in-sample set lengths should be tested with all parameter combinations; however, this would drastically increase the computational time required to conduct this search.

Similarly, for feasibility, the number of bootstrap replications in regression

Table 5.1: Final sets of parameters tested in a grid search

Model	Parameter	Tested values	Count
SARIMAX	(p, d, q)	$(1, 0, 1) \rightarrow (4, 1, 4)$	256
	$(P, D, Q)_s$	$(0, 0, 0)_{24} \rightarrow (1, 1, 1)_{24}$	
Bagged trees	min split	18, 19, ..., 30	455
	max depth	16, 17, ..., 22	
	cost complexity	$10^{-8}, 2.575 \cdot 10^{-7}, ..., 10^{-6}$	
	bootstrap rep.	30	
RNN	hidden size	64, 128, 192	108
	LSTM layers	1, 2	
	linear units	128, 256, 512	
	learning rate	0.001, 0.0005	
	linear dropout	0, 0.2	
	LSTM dropout	0, 0.2	
	seq. length	120	
	loss function	mean square	
	MLP activation	ReLU	
	preprocessing	standardization	
	optimizer	Adam	
	batch size	128	
	epochs	up to 30	

Note: In-sample set starts in 2017. Combinations of input variables not included in the model count. Any other parameters remained at default values.

trees was set to 30 in the grid search. We tested several specifications with 100 replications; however, the decrease in the utilized error metrics (see Section 5.7) on the validation set was around 2.5% (e.g., from 100 RMSE to 97.5 RMSE). Thus, we believe that setting the number of bootstrap replications to 30 is an acceptable tradeoff. Moreover, the same number of repetitions would need to be applied to all 455 configurations to ensure fair competition, should one decide to increase the parameter value, which is the reason why we ultimately did not employ 100 bootstrap replications. Similarly, in the neural network grid search, some parameters, such as batch size or the loss function, were fixed due to the large number of possible combinations.

Finally, let us note that with respect to regression trees, sources such as Laurinec (2017) or Boehmke (2018) were helpful in determining the starting parameters. Similarly, as mentioned earlier, in the specification and implementation of the recurrent neural network, we adapted the approach outlined in Keydana (2021a) and Keydana (2021b).

5.6 Variable Selection

The total number of hourly variables used in this thesis is 132 (see Table B.9)—these included the raw series explored in Section 4.2, features created using these variables from Section 5.2.2, and a set of indicators detailed in Section 5.2.1. To avoid overfitting, we concluded that some form of variable selection was necessary. For example, in Fan & Hyndman (2012), the authors initially fit a model with all available predictors and continued removing variables until the out-of-sample performance no longer improved. However, this approach would be computationally unfeasible in our case due to the grid search.

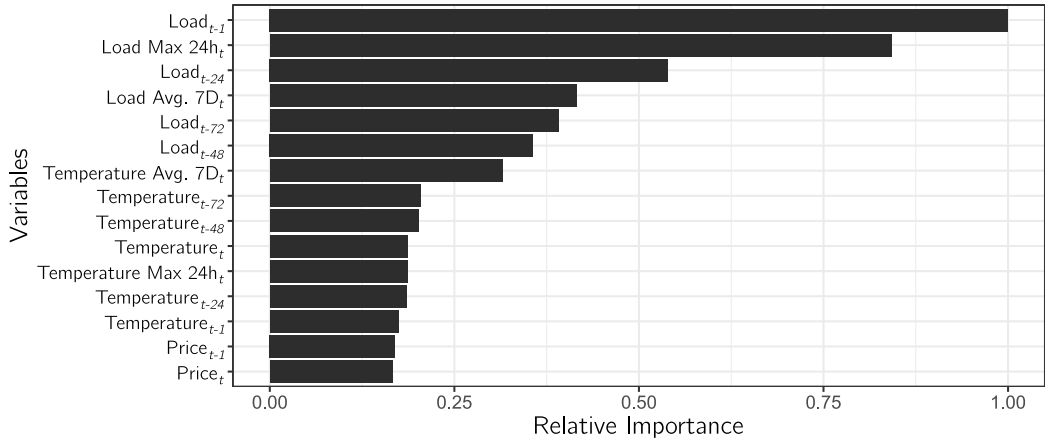
Therefore, to resolve this issue, we have decided to exploit the strengths of each utilized method. To elaborate, firstly, because ARIMA-family models are primarily associated with univariate time-series modeling (Brooks 2014, p. 251), we only utilize historical load data and dummy variables in the SARIMAX model. This further alleviates the problem in the multiple-steps-ahead forecasting task, where the future values of potential external regressors would need to be supplied.

Secondly, as outlined in Section 5.1.2, regression trees contain a variable selection *mechanism*, which solves the issue in this particular case. However, the procedure also requires future values of external inputs for longer forecasts. Thus, we use bagged regression trees only in the one-step-ahead forecasting exercise, ensuring that lagged variables are used as predictors.

Finally, in the neural network model, while the problem of multiple-steps-ahead multivariate forecasting is resolved using an MLP, the number of inputs needs to be regulated. Thus, we decided to fit a bagged regression tree model on the 2012 to 2020 dataset (i.e., excluding the 2021 out-of-sample set) and save the relative variable importance measures. We then tried to employ the top 15 predictors in the neural network models and added or discarded predictors depending on the improvement of the RNN's validation set performance. The relative variable importance measures (i.e., decrease in the residual sum of squares relative to the *best* variable) can be seen in Figure 5.9—after testing several specifications, the bagged regression tree model was fitted with 30 bootstrap replications, maximum depth of 18, minimum split of 21, and the cost complexity parameter equal to $5.05 \cdot 10^{-7}$. It is apparent that transformations of load, followed by temperature, seemed to be considered as the most important predictors.

Additionally, we used dummy variables as inputs to control for calendar

Figure 5.9: Bagged regression tree variable importance (top 15 predictors)



Note: Relative importance shows the decrease in the residual sum of squares relative to the *best* variable. "Avg. 7D" means average in the past 7 days. "Max 24h" refers to the maximum in the last 24 hours.

effects and seasonality because it is very frequently applied in academic literature, as we have previously observed. Furthermore, we later discarded the two price variables because their inclusion did not appear to considerably improve fit in our testing. Therefore, the final set of predictors included in the neural networks contained Load_{t-1}, Max Load in 24 hours$_t$, Load_{t-24}, Average Load in 7 days$_t$, Load_{t-72}, Load_{t-48}, Average Temperature in 7 days$_t$, $\text{Temperature}_{t-72}$, $\text{Temperature}_{t-48}$, Temperature_t, Max Temperature in 24 hours$_t$, $\text{Temperature}_{t-24}$, Temperature_{t-1}, and the indicator variables from Section 5.2.1.

5.7 Forecast Error Measures

Several error measures can be utilized to assess the accuracy of a model on an out-of-sample set in a pseudo-out-of-sample forecasting exercise. Denoting $f_{t,h}$ as the h-step-ahead forecast at time t, the following metrics, all of which we referred to in earlier parts of the thesis, are typically used to quantify error:

$$\text{RMSE}_h = \sqrt{\frac{1}{m}\sum_{h=1}^{m}(y_{t+h} - f_{t,h})^2},$$

$$\text{MAE}_h = \frac{1}{m}\sum_{h=1}^{m}|y_{t+h} - f_{t,h}|, \ \text{MAPE}_h = \frac{1}{m}\sum_{h=1}^{m}\left|\frac{y_{t+h} - f_{t,h}}{y_{t+h}}\right|,$$

where m is the length of the out-of-sample set (Brooks 2014, pp. 293, 298, Tsay 2005, p. 194).

Additionally, the *Diebold-Mariano test* (DM), modified by Harvey *et al.* (1997), approaches the comparison of forecasts in the form of a hypothesis test (Diebold & Mariano 1995). Following Harvey *et al.* (1997), if we consider $e_t^{(1)}$ and $e_t^{(2)}$ to be two different h-steps-ahead forecast errors (i.e., $e_t = y_{t+h} - f_{t,h}$), the null hypothesis of equal accuracy

$$E(g(e_t^{(1)}) - g(e_t^{(2)})) = 0,$$

where the loss function $g(\cdot)$ transforms the error in a pre-specified manner (e.g., by squaring), is tested against a one- or two-sided alternative (Harvey *et al.* 1997; Enders 2015, pp. 86–88).

Because the Diebold-Mariano test only examines errors from h-steps-ahead forecasts and not a sequence of predictions (Harvey *et al.* 1997; Hardy 2016), we apply the *Friedman and Nemenyi tests* (Demšar 2006; Svetunkov 2022) to compare predictions in the 48-hours-ahead exercise. Following Demšar (2006), the Friedman test assigns a rank to each forecast through analysis of variance measures and calculates the mean ranking for each method. It then tests the null hypothesis of equal average rank—if rejected, the Nemenyi test is consequently conducted. In this procedure, the following statistic is calculated

$$\text{Critical difference} = q_{\alpha,k}\sqrt{\frac{k(k+1)}{6N}}, \tag{5.6}$$

where $q_{\alpha,k}$ is a critical value for the Nemenyi test based on the significance level α along with the number of compared methods k, and N refers to the number of data points. That is, any two methods' performance is thought to significantly vary if their difference in mean ranking is larger than the value of the statistic from Equation 5.6 (Demšar 2006). These results are then reported following the visualization outlined in Koning *et al.* (2005).

The computational part of this thesis relied on the following R (R Core Team 2022) and Python (Van Rossum & Drake 2009) packages: `aTSA` (Qiu 2015), `caret` (Kuhn 2022), `DescTools` (Signorell 2023), `dplyr` (Wickham *et al.* 2023a), `forecast` (Hyndman & Khandakar 2008), `ggplot2` (Wickham 2016), `ggsci` (Xiao 2023), `imputeTS` (Moritz & Bartz-Beielstein 2017), `ipred` (Peters & Hothorn 2022), `isdparser` (Chamberlain 2020), `librarian` (Quintans 2021), `lubridate` (Grolemund & Wickham 2011), `matplotlib` (Hunter 2007), `mlr` (Bischl *et al.* 2016), `pandas` (McKinney 2010), `plotly` (Plotly Technologies Inc. 2015), `purrr` (Wickham & Henry 2023), `readxl` (Wickham & Bryan 2023), `rpart` (Therneau & Atkinson 2022), `rpart.plot` (Milborrow 2022), `rstudioapi` (Ushey *et al.* 2022), `tibble` (Müller & Wickham 2022), `tidyr` (Wickham *et al.* 2023b), `torch` (Falbel & Luraschi 2023), `tseries` (Trapletti & Hornik 2023), `tsibble` (Wang *et al.* 2020), `tsutils` (Kourentzes 2022), `xtable` (Dahl *et al.* 2019), `xts` (Ryan & Ulrich 2023), and `zoo` (Zeileis & Grothendieck 2005).

Chapter 6

Results and Discussion

This chapter comprises four sections. In the first part, we describe the results of the forecasting scheme employing high-frequency data. We then continue by reporting the two exercises utilizing hourly data, the first of which is one-step-ahead, while the other contains hourly load predictions of up to two days in the future. Each of these three sections contains a general overview of the results, followed by a detailed description, and ends with a summary of our findings. We conclude this chapter by discussing the contribution of this thesis, its limitations, and future research opportunities.

6.1 Minute Data

The objective of the first minor analysis was to investigate the predictability of high-frequency Czech load series in first differences of logs using in-sample sets of two to four days. This question was tested by iteratively fitting a large set of ARIMA models using the Hyndman & Khandakar (2008) algorithm, producing one-step-ahead forecasts for the next day, and comparing their accuracy to the predictions generated by a random walk model. Following the results of Taylor (2008), we hypothesized that the Czech minute load series is predictable, and we find some evidence in support of this claim.

6.1.1 Detailed Results

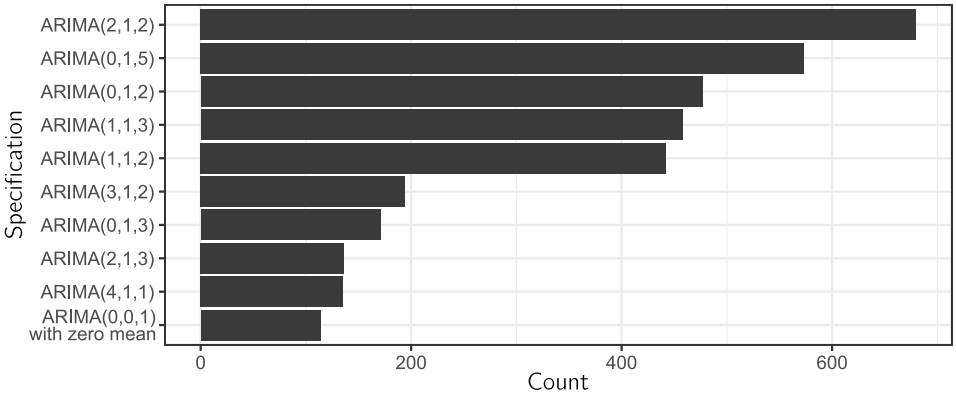
As reported earlier in Section 5.4.1, we conducted the one-minute-ahead pseudo-out-of-sample forecasting exercise with in-sample lengths of two to four days. Afterward, RMSE and MAE metrics were calculated for each out-of-sample day of minute data and compared across the three sizes. Averaging by day, both

of these measures recorded the lowest values using the two-day in-sample set (mean RMSE of 0.00348 & mean MAE of 0.00270). Therefore, we report the results for in-sample sets of two days ($T_i = 2880$) and one-day out-of-sample sets ($T_i = 1440$) of the minute data from 2011 to 2021 ($i = 1, 2, ..., 4016$).

Estimation

In Figure 6.1, we may observe the top ten most frequent specifications that were selected using the Hyndman & Khandakar (2008) algorithm. It is evident that, despite the fact that the input load series was in first differences of logarithms, the lowest-AIC specifications tended to be differenced one additional time. The reason for this might be that the algorithm performed a unit root test only on the two-day subset of the data instead of the entire 2011 to 2021 series. Regardless, it is apparent that with a total of 4016 models estimated, ARIMA$(2, 1, 2)$ was selected in nearly 17% of all cases, closely followed by an ARIMA$(0, 1, 5)$ model.

Figure 6.1: Ten most frequent lowest-AIC specifications fitting Czech minute load data (2011 to 2021)



Note: With two days of minute data used as the in-sample set, the final number of employed models was 4016.

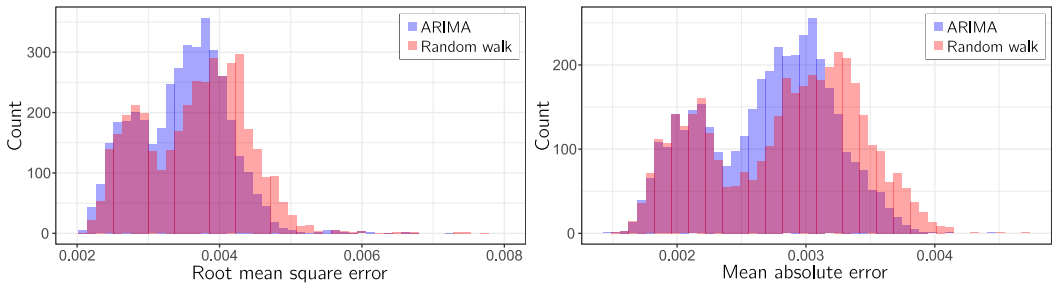
Evaluation

There were 4016 out-of-sample days of minute data, the overwhelming majority[1] of which contained 1440 one-step-ahead predictions generated by the ARIMA and random walk models. For each of these sets of forecasts, we calculated RMSE and MAE (we use MAE instead of MAPE because the load series is in first differences) and plotted the results in two ways.

[1]The outliers were daylight savings time days.

Firstly, in the left panel of Figure 6.2, we display the RMSE values of the one-step-ahead ARIMA & naïve forecasts, and while there is a large degree of overlap between the errors produced by the two methods, especially for lower values of root mean square error, the distribution of RMSE values generated by the ARIMA models seems to be considerably shifted to the left for RMSE values between 0.003 to 0.004. Similarly, observing the right panel of Figure 6.2, which displays the mean absolute error values, results in more or less the same conclusion.

Figure 6.2: Histograms of RMSE (left) and MAE (right) values of one-step-ahead out-of-sample forecasts of Czech minute load data (2011 to 2021)



Note: Two days were used as the in-sample set; the out-of-sample set's length was one day.

To provide a thorough understanding of whether these differences were statistically significant, we conducted the modified (one-tailed) DM test briefly described in Section 5.7, with the alternative hypothesis being that the random walk forecasts were less accurate than ARIMA forecasts. This was done for the entire series of predictions, as well as for each of the out-of-sample days of one-minute-ahead forecasts, respectively. Moreover, we performed the modified Diebold-Mariano procedure with absolute and squared loss.

The results of the one-tailed DM test utilizing the entire 2011 to 2021 set of one-step-ahead out-of-sample predictions can be seen in Table 6.1. At the significance level of 5%, we reject the null hypothesis, suggesting that the ARIMA forecasts were more accurate than the naïve predictions.

Table 6.1: Results of the modified Diebold-Mariano test comparing the random walk and ARIMA forecasts

Loss function	**Statistic**	**p-value**
Absolute	-254.98	< 0.01
Square	-163.14	< 0.01

Note: The alternative hypothesis is that the random walk forecasts were less accurate than the ARIMA forecasts.

As outlined above, we were further interested in conducting the procedure for each of the 4016 out-of-sample days. Thus, we saved the resulting DM test p-values of one-tailed tests, and we report the number of rejections of the null hypothesis at the significance level of 5% in Table 6.2. The modified DM test results reveal that in 2703 out of 4016 cases (absolute loss), the forecasts produced by the ARIMA models were significantly more accurate than those generated by a random walk. When using squared loss, the number of H_0 rejections was much higher—in 3569 instances. However, let us note that it would be problematic from a statistical standpoint to arrive at a general conclusion using these particular results. In this regard, the findings presented in Table 6.1 would be more suitable.

Table 6.2: Results of the modified Diebold-Mariano tests for each of the out-of-sample days

Loss function	H_0 **rejected***	**Failed to reject** $H_0{}^*$
Absolute	2703x	1313x
Square	3569x	447x

Note: *number of times the null hypothesis was/was not rejected at the significance level of 5%. The alternative hypothesis is that the random walk forecasts were less accurate than the ARIMA forecasts.

While the modified DM test is more insightful in determining which of the two forecasts were significantly more accurate, for completeness, let us also report the number of times the root mean square error and the mean absolute error were lower for the ARIMA forecasts. In terms of RMSE, the naïve method's predictions were more precise in 70/4016 instances. Regarding MAE, on the other hand, the comparison contained 3186 cases where the ARIMA forecasts yielded lower values of mean absolute error.

Finally, in Table 6.3, mean yearly values of RMSE and MAE metrics from the pseudo-out-of-sample forecasting exercise are shown for both the ARIMA and the random walk. Intriguingly, in the initial four years of our data, both methods produced more accurate predictions (on average) than in later periods—a notable increase in all error metrics can be observed from 2014 to 2015. However, since then, the difference in the measures across the two methods appeared to have increased in favor of the ARIMA models. In other words, despite producing higher error, ARIMA forecasts were comparatively more accurate from 2015 to 2021 than in the initial four years in contrast with the naïve predictions. Lastly, let us note that median RMSE and MAE values exhibited an analogous pattern.

Table 6.3: Average yearly RMSE and MAE of ARIMA and random walk one-minute-ahead forecasts (2011 to 2021)

	ARIMA Forecasts		**Naïve Forecasts**	
Year	**RMSE**	**MAE**	**RMSE**	**MAE**
2011	0.00313	0.00242	0.00320	0.00244
2012	0.00292	0.00224	0.00299	0.00225
2013	0.00275	0.00211	0.00280	0.00210
2014	0.00282	0.00214	0.00287	0.00213
2015	0.00394	0.00304	0.00416	0.00320
2016	0.00389	0.00302	0.00411	0.00319
2017	0.00375	0.00294	0.00400	0.00313
2018	0.00378	0.00296	0.00404	0.00317
2019	0.00377	0.00295	0.00403	0.00316
2020	0.00376	0.00295	0.00402	0.00315
2021	0.00379	0.00297	0.00410	0.00322

Diagnostics

Although the primary objective of this exercise was to determine whether it is possible to outperform a random walk model in one-step-ahead pseudo-out-of-sample forecasting, for completeness, we conducted three diagnostic tests on the residuals produced by each of the 4016 specified ARIMA models. Once again, let us reiterate that these results should be interpreted in the context of each fitted model rather than being used to form general conclusions.

Firstly, we performed the Ljung-Box test for serial correlation, described in Section A.3 of Appendix A, for lag lengths of 4 to 7. In the majority of cases (5% significance level), we were unable to reject the null hypothesis of no residual autocorrelation for every tested lag—the maximum number of H_0 rejections was 65 for 7 lags (Table B.6), suggesting a satisfactory fit in most instances.

Then, the presence of *autoregressive conditional heteroskedasticity* (ARCH) effects in the residuals was tested using the Portmanteau test on squared residuals (see Section A.3). At the significance level of 5%, the results (available in Table B.7) somewhat indicate that, for some of the models, there may be additional information contained in the residuals that could be further captured.

Finally, we also performed the *Jarque-Bera* (JB) test (explained in Section A.4) to assess whether the ARIMA residuals were normally distributed in each of the 4016 cases. Following the results in Table B.8, it seems that, in most instances, the normality assumption appeared not to hold (5% significance level). However, as mentioned by Brooks (2014, p. 210), the lack of residual

normality does not necessarily have to be problematic when the in-sample size is large.

Summary and Discussion

Overall, considerable improvements could be introduced to the modeling approach used in this exercise, such as accounting for ARCH effects in some cases. However, even with an extremely short in-sample set, we see that it is possible to produce more accurate very short term forecasts than those generated by a random walk model.

Therefore, we believe that the presented evidence provides support in favor of the hypothesis that high-frequency Czech load is predictable, especially following the overall results of the Diebold-Mariano test presented in Table 6.1. However, it is important to note that when observing the histograms in Figure 6.2, we noted that for lower values of the error metrics, there is a considerable overlap. Because the reported RMSE and MAE values quantify error for an entire day of one-minute-ahead forecasts, we later discovered that winter days were typically those where the random walk model produced predictions with accuracy closer to the forecasts generated by the ARIMA models. As outlined in Section 4.1.1, people typically demand more electricity during this season and load has higher variance, which might be a part of the reason why it appears to be more difficult to predict Czech minute load in winter.

On the contrary, even if we were to completely abstract from the findings of Taylor (2008), we believe that some degree of predictability of *any* national-level minute load series should be expected as even short-term in-sample periods contain a day/night cycle. Though, it should be noted that our load series was in first differences of logs, which means that such an exploitable pattern seemingly vanishes (see Figure 5.7).

6.2 Hourly Data: One-Step-Ahead

The first of the two major analyses utilizing hourly data was concerned with evaluating one-step-ahead forecasts of national-level loads on the 2021 out-of-sample set. Predictions produced by five methods were compared: SARIMAX, bagged regression trees, RNN, random walk, and RNN-SARIMAX. Overall, the combined forecasts generated by RNN-SARIMAX appeared to be the most accurate in comparison with other methods.

6.2.1 Detailed Results

In the one-hour-ahead pseudo-out-of-sample forecasting exercise, which we detailed in Section 5.4.2, the primary objective was to find the most accurate specification generating the lowest-error predictions of load on the 2021 out-of-sample set ($T = 8734$). Initially, three methods with optimized parameters were employed: SARIMAX, bagged regression trees, and LSTM RNN. Because official one-step-ahead load forecasts are not published by ČEPS, we added naïve forecasts as a benchmark. Moreover, we further averaged the predictions of the two most accurate models—the SARIMAX and the LSTM RNN—to produce even more precise forecasts. All these values were then compared using conventional error metrics (RMSE and MAPE) as well as the Diebold-Mariano test.

Estimation

Based on the results of the grid search and a subsequent test of several in-sample set lengths, optimal parameters of the three methods were selected for modeling load and can be seen in Table 6.4. For instance, one of the chosen specifications was SARIMAX$(1, 0, 1)(1, 1, 1)_{24}$, which contains Load_t as the dependent variable[2] that is shared across all the models. However, because the three approaches utilized different lengths of the in-sample set, we refrain from comparing their performance in this regard. Nevertheless, in Appendix B, we provide coefficient estimates of the SARIMAX model (left side of Table B.10) as well as a plot of in-sample and validation loss per epoch in the neural network (left panel in Figure B.1).

Moreover, since the SARIMAX coefficient estimates are *somewhat* interpretable (Hyndman 2010), let us briefly review some of the relationships displayed in Table B.10. Overall, most parameters are highly statistically significant. Some of those that do not appear to be different from zero include the coronavirus dummy variable, several months throughout the year, or indicators for two to four o'clock. Finally, it seems that the inclusion of hour-of-day and non-working day interactions, as motivated in Section 5.2.1, was a reasonable decision due to the (mostly) highly statistically significant negative parameter estimates. This is further corroborated by the observation that these per-hour coefficients do not *cancel* each other's effect—for instance, the parameter esti-

[2]Regular load, i.e., no transformations were applied. Applying natural logarithms to the dependent variable yielded mixed results across models.

mate of the interaction of non-working day and hour 20 is equal to −467.69, while the coefficient for hour 20 alone equals 1610.32 (both highly significant).

Table 6.4: Specifications used in the one-step-ahead forecasting exercise with hourly data

Model	Parameter	Selected values	Variables*
SARIMAX	(p, d, q)	$(1, 0, 1)$	Load and indicators
	$(P, D, Q)_s$	$(1, 1, 1)_{24}$	
	in-sample start	Jan 2018 ($T = 21162$)	
Bagged trees	min split	18	Load, weather, price, all transformations, and indicators
	max depth	21	
	cost complexity	$2.575 \cdot 10^{-7}$	
	bootstrap rep.	30	
	in-sample start	Jan 2013 ($T = 64976$)	
RNN	hidden size	192	Load, temperature, load transformations, temperature transformations, and indicators
	LSTM layers	1	
	linear units	256	
	learning rate	0.001	
	linear dropout	0	
	LSTM dropout	0	
	seq. length	120	
	loss function	mean square	
	MLP activation	ReLU	
	preprocessing	standardization	
	optimizer	Adam	
	batch size	128	
	epochs	22 (see Figure B.1)	
	in-sample start	Jan 2015 ($T = 47460$)	

Note: *Full list of variables available in Table B.9 of Appendix B. Bagged regression trees and neural networks contain random elements; thus, the *seed* was fixed for reproducibility. Any other parameters remained preset at default values.

Evaluation

In Table 6.5, the validation and out-of-sample RMSE & MAPE values for each utilized method's predictions are shown, including the benchmark naïve forecasts and the combined model. All in all, the averaged RNN-SARIMAX one-step-ahead predictions appeared to be the most accurate in terms of RMSE as well as MAPE in both the validation ($\text{RMSE}_1 = 51.43$, $\text{MAPE}_1 = 0.00492$) and the out-of-sample sets ($\text{RMSE}_1 = 53.77$, $\text{MAPE}_1 = 0.00484$).

Interestingly, with regard to the utilized out-of-sample error metrics, the most precise method from the three standard models was the SARIMAX, closely followed by the RNN. Finally, the root mean square forecast error of the

bagged regression tree model was equal to 101.03, while the random walk produced roughly three times as much error, both in terms of RMSE and MAPE. Thus, due to its poor performance, we ignore these naïve predictions in some further comparisons purely for practical reasons.

Table 6.5: One-step-ahead validation & out-of-sample forecasting accuracy results

	Validation		**Out-of-sample**	
Method	**RMSE**	**MAPE**	**RMSE**	**MAPE**
RNN	61.56	0.00604	65.19	0.00596
SARIMAX	58.12	0.00556	60.24	0.00540
Bagged trees	91.36	0.00883	101.03	0.00884
RNN-SARIMAX	51.43	0.00492	53.77	0.00484
Random walk	308.39	0.02977	305.79	0.02756

Note: The out-of-sample set spanned from Jan 2021 to Dec 2021.

Because demand for electricity varies throughout the year, forecasting accuracy may be expected to change as well. Thus, we provide monthly values of RMSE and MAPE of the 2021 out-of-sample set forecasts in a graphical as well as a tabular format in Figure 6.3 & Table 6.6, respectively. It is apparent that in nearly all months, both the root mean square and the mean absolute percentage error metrics were the lowest for the combined RNN-SARIMAX predictions, recording the minimum RMSE in August ($\text{RMSE}_1 = 40.99$) and the lowest MAPE in June ($\text{MAPE}_1 = 0.00428$).

The SARIMAX model's forecasts can perhaps be considered as the second-most accurate as measured by RMSE and MAPE in the pseudo-out-of-sample exercise—in fact, in February 2021, the method recorded slightly lower MAPE than the RNN-SARIMAX. Both the RMSE and MAPE of the bagged regression tree model's forecasts suggested that the method, using the parameters specified in Table 6.4, was the least fit for producing one-step-ahead forecasts on the 2021 data, further corroborating the results from Table 6.5.

To determine which of the methods produced the most accurate one-hour-ahead forecasts on the out-of-sample set, let us perform several modified Diebold-Mariano tests with absolute and square loss. The results are displayed in Table 6.7 and it is clear that none of the methods produced more precise forecasts than the RNN-SARIMAX at the 0.1% significance level, even after lowering the significance threshold using Bonferroni correction (James *et al.* 2021, pp. 564–565). The SARIMAX also appeared to yield significantly more accurate predictions

Figure 6.3: One-step-ahead out-of-sample forecast errors by month

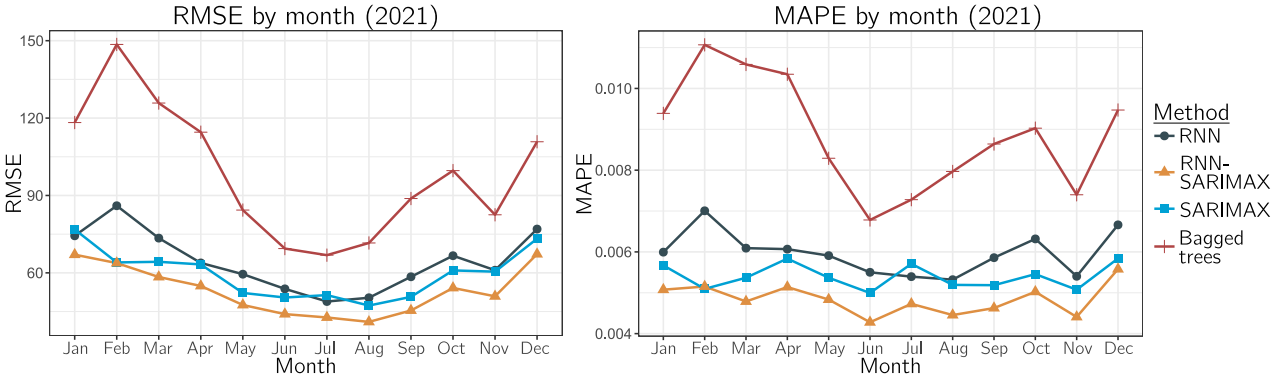


Table 6.6: One-step-ahead out-of-sample forecast errors by month

Month (2021)	**RMSE**				**MAPE**			
	RNN	SARIMAX	Bagged trees	RNN-SARIMAX	RNN	SARIMAX	Bagged trees	RNN-SARIMAX
Jan	74.36	76.83	118.27	67.05	0.00599	0.00567	0.00939	0.00508
Feb	86.01	64.04	148.55	63.80	0.00701	0.00509	0.01107	0.00515
Mar	73.47	64.28	125.83	58.39	0.00609	0.00537	0.01059	0.00479
Apr	63.85	63.27	114.53	54.92	0.00607	0.00584	0.01035	0.00514
May	59.48	52.21	84.34	47.54	0.00591	0.00537	0.00829	0.00483
Jun	53.80	50.44	69.41	44.00	0.00550	0.00499	0.00678	0.00428
Jul	48.89	51.37	66.86	42.70	0.00539	0.00571	0.00728	0.00473
Aug	50.35	47.32	71.59	40.99	0.00532	0.00519	0.00797	0.00445
Sep	58.48	50.69	88.82	45.40	0.00586	0.00519	0.00864	0.00463
Oct	66.65	60.94	99.64	54.15	0.00632	0.00546	0.00903	0.00502
Nov	61.02	60.46	82.51	50.91	0.00540	0.00507	0.00740	0.00441
Dec	76.98	73.24	110.82	67.26	0.00666	0.00584	0.00947	0.00558

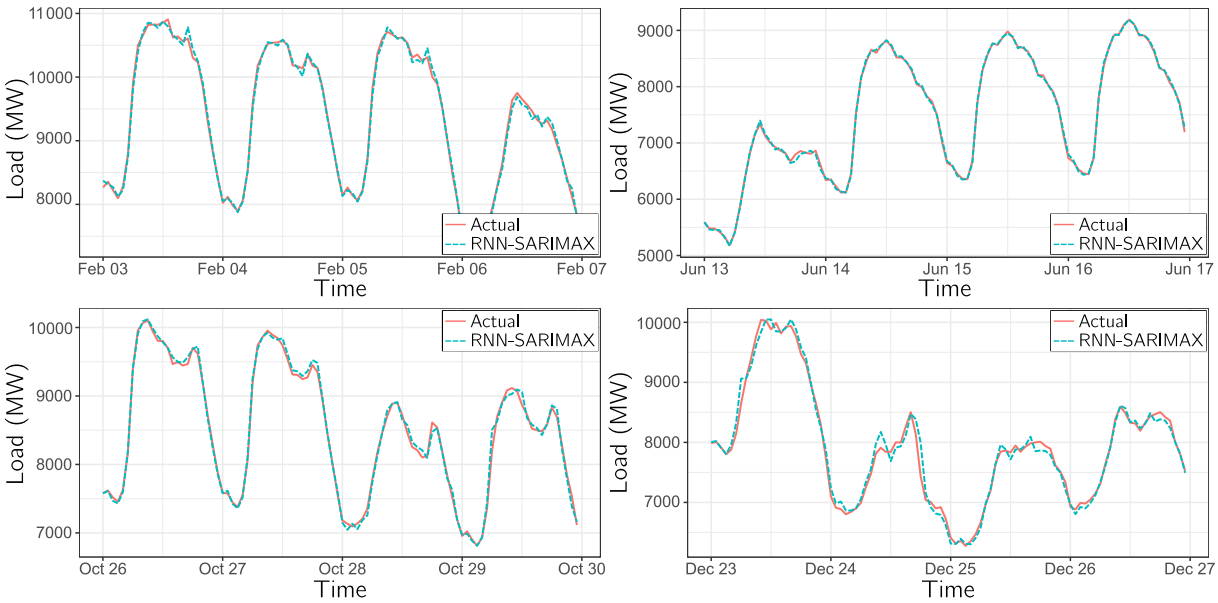
than the RNN on the 2021 out-of-sample load data, both with regard to absolute and square loss in the DM test (0.1% significance level).

Thus, in Figure 6.4, we provide examples of the most accurate one-step-ahead forecasts (RNN-SARIMAX) together with actual load in four periods of the 2021 out-of-sample set. We considered it important to include weekends (Feb 6, Jun 13, Dec 25–26) and holidays (Oct 28, Dec 24–26) in these panels, which are more challenging to forecast than regular days. Overall, except for the Christmas holidays, the one-hour-ahead RNN-SARIMAX predictions appeared to very closely approximate the actual load in the February, June, and October sample plots.

Table 6.7: Modified Diebold-Mariano test results in the one-step-ahead forecasting scheme

Modified DM Test p-values – Absolute Loss					
H_A : Column i method's forecasts are less accurate than row j method's forecasts, $i, j \in \{1, ..., 5\}$					
	RNN	SARIMAX	Bagged trees	Random walk	RNN-SARIMAX
RNN	—	>0.1	<0.001	<0.001	>0.1
SARIMAX	<0.001	—	<0.001	<0.001	>0.1
Bagged trees	>0.1	>0.1	—	<0.001	>0.1
Random walk	>0.1	>0.1	>0.1	—	>0.1
RNN-SARIMAX	<0.001	<0.001	<0.001	<0.001	—
Modified DM Test p-values – Square Loss					
H_A : Column i method's forecasts are less accurate than row j method's forecasts, $i, j \in \{1, ..., 5\}$					
	RNN	SARIMAX	Bagged trees	Random walk	RNN-SARIMAX
RNN	—	>0.1	<0.001	<0.001	>0.1
SARIMAX	<0.001	—	<0.001	<0.001	>0.1
Bagged trees	>0.1	>0.1	—	<0.001	>0.1
Random walk	>0.1	>0.1	>0.1	—	>0.1
RNN-SARIMAX	<0.001	<0.001	<0.001	<0.001	—

Figure 6.4: Sample plots of the best one-hour-ahead forecasts and actual load (2021)



Diagnostics

For completeness, we conducted several residual diagnostic procedures and tests for the SARIMAX, RNN, and bagged regression trees (in both the one-hour-ahead and 48-hours-ahead exercises), detailing our process in Appendix A, Section A.5, though we are unsure of the contribution of diagnostics in the lat-

ter two nonlinear methods. Thus, concentrating primarily on the results of the SARIMAX models (see Section A.5), serial correlation and ARCH effects appear to be present, suggesting that some information may not be utilized by the employed specifications (as per Hyndman & Athanasopoulos (2021, sec. 5.4)).

While this issue might be alleviated by including additional parameters, residual diagnostics are "not a good way to select a forecasting method" (Hyndman & Athanasopoulos 2021, sec. 5.4). Crucially, though, in the grid search outlined in Section 5.5, higher orders of p and q (see Equation 5.3) were tested and generated worse results on the validation set in terms of RMSE and MAPE. Nonetheless, we believe that these results should be somewhat expected because of the multiple seasonal patterns present in the load data, which are challenging to fully account for.

Summary and Discussion

In sum, we may observe that most employed models produced relatively highly accurate predictions of the 2021 Czech transmission grid load when compared to the benchmark random walk in the one-hour-ahead pseudo-out-of-sample forecasting exercise. With the exception of trees and the naïve method, every other model produced forecasts with MAPE far below 1% on a monthly basis. The most accurate technique in terms of RMSE, MAPE, and Diebold-Mariano tests involved averaging the predictions of the SARIMAX and the RNN models, yielding RMSE of around 53.77 and MAPE of 0.484% out-of-sample.

Given its performance, we believe that the RNN-SARIMAX model could be helpful in real-time management of the transmission grid. However, it is worth noting that we have no information about the actual models used by the system operator for one-hour-ahead projections. Nonetheless, we believe that the predictions of each of the models can be enhanced by, for example, including composite weather variables, as mentioned in Section 5.2.2. Another opportunity for improvement in the performance of these methods would involve conducting the grid search with higher-order parameters (for instance, more bootstrap replications in the bagged regression trees, larger seasonal AR & MA parameters of the SARIMAX model, etc.), but perhaps more importantly, executing the complete search for each expansion of the initial in-sample set. Lastly, for variable selection in the neural network model, the implementation of a more extensive step-wise selection procedure similar to Fan & Hyndman (2012) may further be utilized to increase out-of-sample accuracy of the RNN.

6.3 Hourly Data: 48-Hours-Ahead

In the second analysis utilizing hourly load, we generated forecasts of up to 48-steps-ahead through a rolling scheme and evaluated the performance of several models on the 2021 out-of-sample set. The SARIMAX, RNN, and *seasonal naïve* (SNAIVE) predictions were compared with the Czech transmission grid operator's forecasts (refered to as *official*). Additionally, RNN-SARIMAX, RNN-Official, and SARIMAX-Official averaged predictions were calculated and included in the comparison. All in all, while these combined forecasts appeared to be more precise than the standard methods as well as the official projections, the RNN-SARIMAX seemed to be the most accurate.

6.3.1 Detailed Results

The primary goal of forecasting load 48-hours-ahead out-of-sample and comparing the predictions to actual values was to assess the accuracy of the official projections published by ČEPS (see Section 5.4.2 for a detailed description of the scheme). As described in Section 5.6, regression trees were not utilized, unlike in the one-step-ahead analysis. Instead, SNAIVE was employed as a benchmark. Much like the naïve forecast, the SNAIVE method simply predicts the same values as in the previous *season* (Hyndman & Athanasopoulos 2021, sec. 5.2), which we set to the year before the start of the out-of-sample set in our case.[3] Furthermore, both the SARIMAX and LSTM RNN specifications were optimized for the 48-hours-ahead forecasting exercise, resulting in different parameters than in the first task.

Afterward, all three of the aforementioned methods were used to generate forecasts up to 48-hours-ahead on the 2021 out-of-sample set ($T = 8734$). Then, using simple averaging, RNN-SARIMAX, RNN-Official, and SARIMAX-Official predictions were created. All the multi-step forecasts were compared using RMSE and MAPE—overall and by month. The Friedman and Nemenyi tests (Section 5.7) were then employed for method comparison.

Estimation

The specifications selected based on the results of the conducted grid search are displayed in Table 6.8. Once again, Load_t is the dependent variable in all

[3]Practically, we simply superimposed 2020 data onto 2021, further aligning the load values by wekends instead of the correct dates.

tested models, and the final in-sample lengths of the SARIMAX and the RNN vary, complicating comparisons of in-sample fits. In Table B.10 of Appendix B, coefficient estimates of the SARIMAX$(2,0,2)(1,0,1)_{24}$ specification are shown, and the in-sample & validation loss per epoch of the RNN is displayed in the right panel of Figure B.1, the latter of which suggested choosing 17 epochs in the neural network model.

Focusing on SARIMAX parameter estimates in Table B.10, as in the first specification, most relationships appear to be statistically significant. Moreover, the coefficient estimates of the utilized dummy variables seem to be more or less equal across the two specifications, taking standard errors into account. For instance, the coefficient of non-working days (`wknd_or_h`) was negative in both cases and equal to approximately -170 in the two specifications. On the other hand, the parameter estimate of the coronavirus indicator now appeared to be highly statistically significant and equal to around -121. In standard linear regression, this would suggest that load during the states of emergency due to the pandemic was slightly lower than usual—in a SARIMAX model, however, the interpretation of this result might be less straightforward (Hyndman 2010).

Table 6.8: Specifications used in the 48-hours-ahead forecasting exercise with hourly data

Model	**Parameter**	**Selected values**	**Variables***
SARIMAX	(p,d,q)	$(2,0,2)$	Load and indicators
	$(P,D,Q)_s$	$(1,0,1)_{24}$	
	in-sample start	Jan 2018 ($T = 21162$)	
RNN	hidden size	128	Load, temperature, load transformations, temperature transformations, and indicators
	LSTM layers	1	
	linear units	512	
	learning rate	0.001	
	linear dropout	0	
	LSTM dropout	0	
	seq. length	120	
	loss function	mean square	
	MLP activation	ReLU	
	preprocessing	standardization	
	optimizer	Adam	
	batch size	128	
	epochs	17 (see Figure B.1)	
	in-sample start	Jan 2016 ($T = 38702$)	

Note: *Full list of variables available in Table B.9 of Appendix B. Neural networks contain random elements; thus, the *seed* was fixed for reproducibility. Any other parameters remained preset at default values.

Evaluation

Table 6.9 displays the forecast RMSE and MAPE metrics on the validation (where applicable) and out-of-sample sets of seven methods: RNN, SARIMAX, SNAIVE, RNN-SARIMAX, the official 48-hours-ahead predictions, and the RNN-Official & SARIMAX-Official forecast combinations. Concentrating on the out-of-sample error values, the lowest-RMSE predictions were generated by the simple mean of the RNN and official predictions ($RMSE_{48} = 269.76$). With regard to MAPE, the highest accuracy forecasts, considering the entire 2021 out-of-sample set, were generated by the RNN-SARIMAX models ($MAPE_{48} = 0.2300$). Furthermore, for 1 to 48-hours-ahead forecasts, the RNN on its own appeared to yield more accurate out-of-sample load forecasts than the SARIMAX in terms of both error measures.

Table 6.9: 48-hours-ahead validation & out-of-sample forecasting accuracy results

	Validation		**Out-of-sample**	
Method	**RMSE**	**MAPE**	**RMSE**	**MAPE**
Official	—	—	372.44	0.03953
RNN	300.25	0.02667	308.71	0.02499
SARIMAX	321.76	0.02946	326.00	0.02844
SNAIVE	557.49	0.05562	699.69	0.06153
RNN-SARIMAX	261.84	0.02395	271.60	0.02300
RNN-Official	—	—	269.76	0.02517
SARIMAX-Official	—	—	272.82	0.02525

Note: The out-of-sample set spanned from Jan 2021 to Dec 2021.

Similarly to the previous hourly forecasting exercise, we report monthly RMSE and MAPE measures to observe the development of forecast accuracy in time, and these results are shown in two sets of figures and tables for clarity. Firstly, in Figure 6.5 & Table 6.10, the performance of *standard* methods (i.e., no forecast combinations) is displayed together with the official predictions. On average, the forecasts of the RNN and the SARIMAX models were more accurate in terms of both error metrics in the majority of instances. With respect to RMSE, however, the official predictions were comparatively closer in some months, yet still seemed worse overall. Altogether, we discovered that December was the most challenging month to generate forecasts for, in part due to the Christmas holidays—specifically, as partially illustrated in the bottom-right panels of Figures 6.8 or B.2, the two methods had a tendency to overestimate load on some of the non-working days; on the other hand, the operator's predic-

tions were remarkably precise on December 24th, though less so in the following days.

However, perhaps the most surprising finding was the accuracy of the seasonal naïve method in August 2021 ($RMSE_{48} = 188.88$ & $MAPE_{48} = 0.02167$) as well as in later months. While the method produced highly inaccurate predictions in the first half of the year (see Figure 6.5 and Table 6.10), its forecasts in late summer and early fall were superior to the official projections in terms of MAPE.

Figure 6.5: 48-hours-ahead out-of-sample forecast errors of standard methods by month

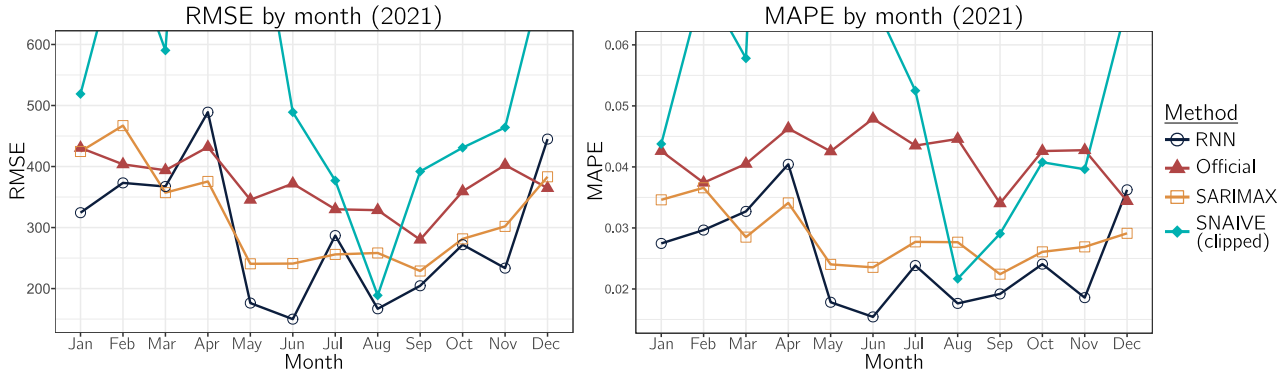


Table 6.10: 48-hours-ahead out-of-sample forecast errors of standard methods by month

Month (2021)	RMSE				MAPE			
	Official	RNN	SARIMAX	SNAIVE	Official	RNN	SARIMAX	SNAIVE
Jan	430.55	324.34	424.65	518.97	0.04260	0.02746	0.03461	0.04377
Feb	403.58	373.19	467.06	782.66	0.03741	0.02965	0.03658	0.07021
Mar	393.89	367.21	357.19	590.34	0.04050	0.03271	0.02849	0.05780
Apr	431.95	489.06	375.52	1538.43	0.04630	0.04043	0.03410	0.20609
May	345.48	176.27	240.62	895.50	0.04255	0.01782	0.02403	0.12106
Jun	372.02	149.93	240.98	489.00	0.04790	0.01542	0.02354	0.06726
Jul	329.92	286.86	255.96	376.97	0.04350	0.02384	0.02772	0.05249
Aug	328.49	166.90	258.35	188.88	0.04458	0.01763	0.02766	0.02167
Sep	280.13	204.62	228.78	391.94	0.03403	0.01919	0.02242	0.02906
Oct	359.30	271.93	281.47	430.86	0.04260	0.02407	0.02608	0.04076
Nov	402.51	233.52	301.92	464.08	0.04274	0.01858	0.02690	0.03961
Dec	364.70	444.96	383.19	741.95	0.03440	0.03625	0.02912	0.06666

Diverting our attention to Figure 6.6 & Table 6.11, which show the monthly results of the averaged (up to) 48-steps-ahead predictions on the 2021 out-of-sample set, it is evident that the official forecasts were now the least accurate every month both in terms of RMSE and MAPE. Moreover, in the summer,

the RMSE of the official two-day projections was more or less two-fold in comparison with the most accurate method (RNN-SARIMAX), while MAPE was approximately three times as large in some cases—intriguingly, the MAPE for the RNN-SARIMAX in June was at its minimum, while the MAPE for the official forecasts was the highest during this month. With regard to the other methods, the 48-hours-ahead RNN-SARIMAX predictions seemed to achieve the lowest error in terms of MAPE in 7/12 months, ranging from around 1.5% to nearly 3.3%. Combining the RNN or SARIMAX forecasts with the official projections appeared to yield comparatively better results in the context of RMSE rather than MAPE, with the lowest RMSE being recorded for the RNN-Official averaged predictions in September ($RMSE_{48} = 143.55$).

Figure 6.6: 48-hours-ahead out-of-sample forecast errors of averaged forecasts by month

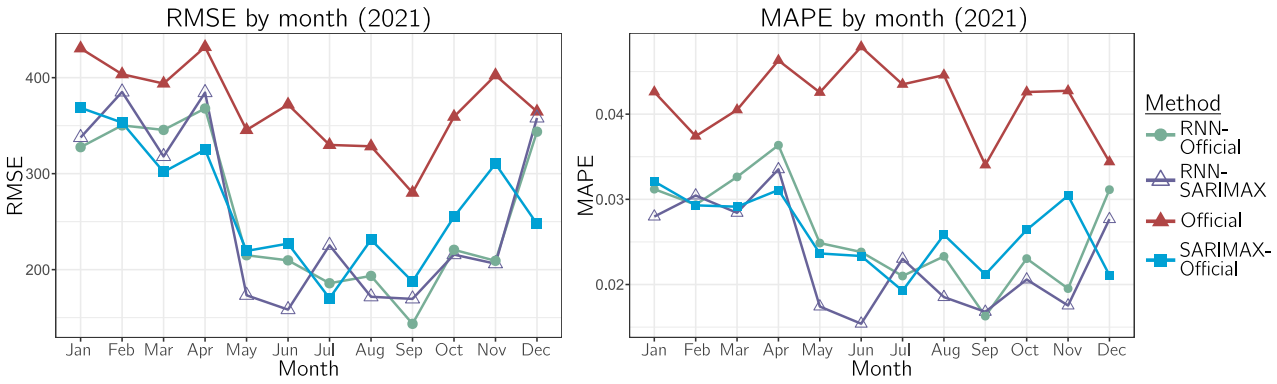
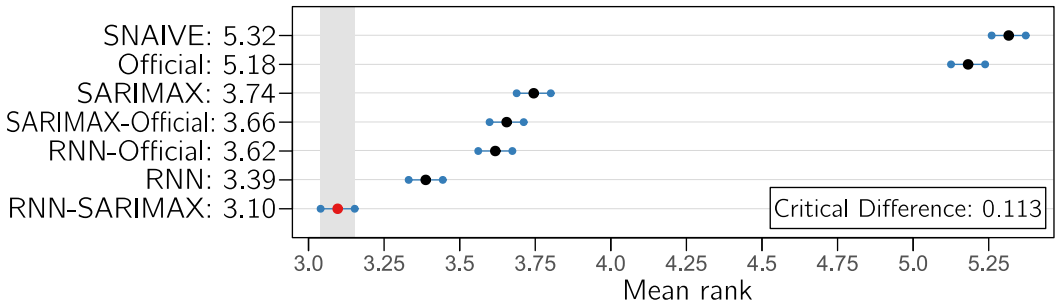


Table 6.11: 48-hours-ahead out-of-sample forecast errors of averaged forecasts by month

Month (2021)	RMSE				MAPE			
	Official	RNN-SARIMAX	RNN-Official	SARIMAX-Official	Official	RNN-SARIMAX	RNN-Official	SARIMAX-Official
Jan	430.55	337.68	327.60	368.63	0.04260	0.02798	0.03119	0.03210
Feb	403.58	385.12	350.29	353.19	0.03741	0.03045	0.02937	0.02929
Mar	393.89	317.90	345.55	301.78	0.04050	0.02842	0.03263	0.02914
Apr	431.95	384.47	367.90	325.07	0.04630	0.03353	0.03635	0.03109
May	345.48	173.30	215.00	219.49	0.04255	0.01742	0.02487	0.02365
Jun	372.02	158.36	209.69	227.12	0.04790	0.01540	0.02382	0.02335
Jul	329.92	225.40	185.83	170.18	0.04350	0.02302	0.02100	0.01927
Aug	328.49	171.74	193.47	231.12	0.04458	0.01852	0.02330	0.02585
Sep	280.13	169.52	143.55	187.56	0.03403	0.01679	0.01629	0.02118
Oct	359.30	215.66	220.66	255.08	0.04260	0.02059	0.02304	0.02642
Nov	402.51	206.19	209.21	310.30	0.04274	0.01754	0.01953	0.03043
Dec	364.70	358.03	343.71	247.64	0.03440	0.02768	0.03113	0.02109

Because we are working with a sequence of up to 48-hours-ahead forecasts rather than a vector of h-steps-ahead predictions, the Diebold-Mariano test should not be used for comparisons (as mentioned in Section 5.7). Instead, in Figure 6.7, we provide the results of the Friedman and Nemenyi tests visually—intuitively, the lower the method's mean rank, the better the forecasts. In our case, the most performant method, according to the results of these two tests, was the RNN-SARIMAX model, with the solo RNN in second (both ranks were statistically significant at the 5% level). The SNAIVE method was determined as the least accurate approach, narrowly outperformed by the official predictions.

Figure 6.7: Results of the Friedman and Nemenyi tests comparing the performance of methods in the 48-hours-ahead scheme



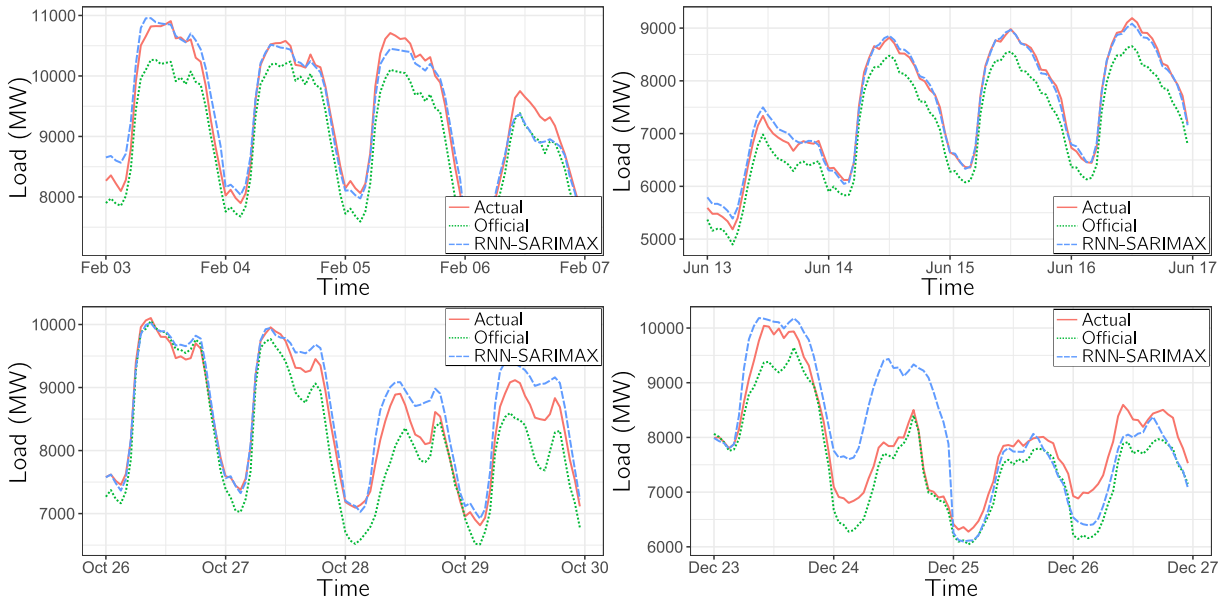
Finally, in Figure 6.8, we display four sample four-day periods with the RNN-SARIMAX forecasts, actual load, and official predictions.[4] As mentioned in Section 6.2.1, we emphasize weekends and holidays in the selected plots due to the higher difficulty in forecasting load accurately during these periods. If we were to only consider the working days in Figure 6.8, the accuracy of the RNN-SARIMAX predictions would seem unparalleled when compared to the official forecasts, especially in the four June days (top-right panel). However, for some public holidays like December 24th, the operator's projections were substantially more precise, as mentioned earlier.

Diagnostics

In terms of residual diagnostics, which are presented in Section A.5, identical conclusions are reached as in the one-step-ahead forecasting exercise—for instance, there were issues with autocorrelation and heteroskedasticity in most

[4] In Figure B.2 of Appendix B, we also provide sample forecasts of the remaining methods in this exercise.

Figure 6.8: Sample plots of the best (up to) 48-hours-ahead forecasts, official predictions, and actual load (2021)



models. To avoid repeating ourselves, we refer the reader to the explanation provided in the one-step-ahead analysis (Sections 6.2.1 and A.5).

Summary and Discussion

With the objective of forecasting load up to 48-hours-ahead on the 2021 out-of-sample set, we were able to specify models capable of outperforming the predictions published by ČEPS. Overall, averaging the outputs of the RNN and SARIMAX methods yielded the most accurate forecasts for a horizon of up to two days, producing MAPE of around 1.5% to 3.4%, whereas the official prediction's mean absolute percentage error ranged from 3.4% to 4.8% on a monthly basis.

However, one important observation from the sample plots in Figure 6.8 should be highlighted. All four panels show that the system operator's projections frequently underestimate the actual load. Nevertheless, when we compare this approach to the strategy of similar institutions around the world, such as the Australian Energy Market Operator for which Fan & Hyndman (2012) created a short-term load forecasting model, we do not observe the same pattern of underestimation. Perhaps it might be less complicated to increase electricity production in the very short term rather than managing excess load in the

Czech context. In this regard, we would strongly advise utilizing a different load forecasting method for holidays (e.g., expert judgement) due to the occasional lack of accuracy of the proposed RNN-SARIMAX model during these periods.

6.4 Contribution, Limitations, & Future Research

To reiterate, the results of the first task revealed that high-frequency Czech transmission system load is predictable, i.e., more accurate than a naïve guess in a one-step-ahead scheme, which is in accordance with related literature (specifically Taylor (2008)). We suppose that this question may have been tackled by operators themselves or monitoring systems manufacturers at some point. Regardless, based on our findings, it is possible that some structural shift may occasionally occur, as evidenced by the change in average error from 2014 to 2015 (Table 6.3), raising the importance of these types of exercises. Moreover, with the increasing adoption of renewable sources of electricity, we believe that the relevance of these analyses may further grow due to the stability challenge discussed in Section 3.1.1.

In the second task, we utilized hourly load data and produced forecasts for the 2021 out-of-sample set one-step-ahead. As reported, the best specification consistently generated predictions with MAPE fluctuating around 0.5% on a monthly basis, establishing itself as a potentially useful tool in real-time system management, we believe. Due to factors such as different sampling periods, it may be challenging or even incorrect to compare the results with those from related literature directly. Taking this into account, to provide at least some level of comparison, in Darbellay & Slama (2000), the authors list MAPE metrics of their one-step-ahead Czech load forecasts produced by an ANN and an ARIMAX model. For the two utilized techniques, the error measure was equal to around 1% both in 1994 & 1995. In our analysis, we were able to generate one-hour-ahead predictions with the lowest MAPE of less than 0.5% in 2021.

The third exercise expanded on the second task by forecasting up to 48-steps-ahead, and comparing our predictions with forecasts published by the Czech transmission system operator. Following the results, we believe that our thesis provides several ideas and opportunities to improve the official predictions. Specifically, we would highlight the performance of the developed indicator variables that could be utilized in future analyses, as they drastically improved fit, particularly in the case of the SARIMAX models. Though, let us emphasize that caution must be exercised with respect to load predictions

for certain holidays. Nevertheless, we have seen that even a simple seasonal naïve model was able to generate more accurate forecasts in some instances, indicating that the official predictions could perhaps be refined. Overall, we believe that the system operator should have a vested interest in maintaining accurate load forecasting models in order to facilitate compliance with its legal obligations (ČEPS 2020).

These improvements, as discussed in Section 3.3, which is concerned with the effects of inaccurate load forecasting, may lead to cost savings due to, for instance, decreased risk associated with the allocation of reserves or more efficient electricity production (Ranaweera *et al.* 1997; Hobbs *et al.* 1999). Following the findings of the third analysis, this might be especially relevant for all summer and early fall months.

With regard to the limitations of our research, we believe that more specifications could be surveyed in the grid search, especially with respect to varying in-sample sets (the grid search we conducted was performed on an in-sample set starting in 2017). Another potential enhancement to the outcomes of this thesis would likely be reached with the help of closer collaboration with the Czech system operator. Additionally, more complex weather predictors, as well as more elaborate variable selection schemes, would perhaps further improve model fit. Finally, the RNN-SARIMAX model's forecasts for 48-hours-ahead on particular public holidays, such as December 24th, significantly overpredicted actual load on the 2021 data—for this reason, different approaches, e.g., expert knowledge, should be used in forecasting load during these specific, albeit infrequent, days of the year.

In terms of directions of future research, one possibility would be to consider longer-horizon hourly predictions using the Czech data, such as one week or a month. Another option could involve evaluating different load forecasting frameworks on the dataset that we prepared, or even on newer load, weather, and price data published by ČEPS, NOAA, and OTE, respectively. Furthermore, Hong & Fan (2016) generally advocate for more extensive development of probabilistic forecasts with the objective of more thoroughly processing uncertainty, especially in the context of the industry's gradual shift towards sustainability. Finally, as maintained by Hong *et al.* (2020), cross-disciplinary cooperation as well as reproducibility are currently the key steps in generating valuable research in the field of load forecasting.

Chapter 7

Conclusion

The utility sector's deregulation in the 1980s jumpstarted the research on short-term load forecasting (Hong & Fan 2016). However, despite the field's substantial body of literature, generating accurate predictions remains a challenging endeavor (Kuster *et al.* 2017), yet an essential part of the modern-day power grid operation (Malik *et al.* 2021).

In this thesis, we analyzed minute- and hourly-frequency national-level electric load series from the Czech transmission system operator and conducted three pseudo-out-of-sample forecasting exercises to assess the performance of several univariate & multivariate methods. The objective of the first task was to determine whether high-frequency load is predictable by comparing rolling ARIMA one-step-ahead forecasts with predictions generated by a random walk. This exercise was performed using more than a decade (2011–2021) of first-differenced minute-frequency logarithmic loads, and we found evidence in support of predictability, which is consistent with related literature (Taylor 2008).

The second and third analyses utilized hourly loads as well as explanatory variables derived from weather data, electricity prices, and seasonal patterns of the modeled series. In both exercises, we conducted a grid search to determine optimal parameter combinations based on the performance of each model on the validation set (June 2020 to December 2020). Furthermore, a significant part of these two tasks was the development of a large set of indicator variables accounting for seasonal and calendar effects, as well as the creation of additional predictors by transforming the historical load data and the *raw* explanatory variables, following the paper of Fan & Hyndman (2012).

In one of the hourly analyses, we compared the one-step-ahead out-of-sample forecasting accuracy of several methods on one year of unseen data

(2021) using RMSE, MAPE, and the Diebold-Mariano test. Particularly, one-step-ahead forecasts generated by the following techniques were compared: SARIMAX, LSTM RNN, bagged regression trees, random walk, and averaged predictions of the SARIMAX and the LSTM RNN models. Overall, the most accurate next-hour forecasts were produced by combining the SARIMAX and the RNN, which consistently generated predictions with approximately 0.5% MAPE on a monthly basis.

The final exercise utilizing hourly data was concerned with evaluating the performance of various techniques in forecasting load up to 48-hours-ahead on the 2021 out-of-sample set, and comparing these predictions to those published by the Czech transmission system operator. In particular, we employed different SARIMAX & recurrent neural network models than in the one-hour-ahead task—both with optimized parameters for forecasting hourly load up to 48-steps-ahead. We further included a seasonal naïve method as well as averages of the SARIMAX, RNN, and the operator's projections. The combinations of these three approaches generally yielded the most accurate results in terms of MAPE and RMSE. Importantly, however, the RNN-SARIMAX forecasts produced lower RMSE as well as MAPE than the official predictions in every month of 2021, and generated around 50% less error in terms of both metrics in several months, especially with respect to MAPE. Finally, based on the results of the Friedman and Nemenyi tests, RNN-SARIMAX was determined as the most performant method.

While the proposed models in the hourly-load analyses can be improved by, for example, conducting a broader grid search or generating more accurate forecasts for some holidays, we believe that this thesis might offer some suggestions and ideas for enhancing the official predictions. Lastly, because we anticipate that the significance of generating precise load projections will rise as a result of the industry's transition towards sustainability, we find it important to engage in interdisciplinary cooperation (as argued by Hong *et al.* (2020)), which could be beneficial for all parties involved.

Bibliography

Aggarwal, C. C. (2018): *Neural Networks and Deep Learning: A Textbook*. New York, NY: Springer.

Ahmed, R., V. Sreeram, Y. Mishra, & M. D. Arif (2020): "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization." *Renewable and Sustainable Energy Reviews* **124, 109792**.

Alfares, H. K. & M. Nazeeruddin (2002): "Electric load forecasting: Literature survey and classification of methods." *International Journal of Systems Science* **33(1)**: pp. 23–34.

Alkhayat, G. & R. Mehmood (2021): "A review and taxonomy of wind and solar energy forecasting methods based on deep learning." *Energy and AI* **4, 100060**.

Berk, R. A. (2016): *Statistical Learning from a Regression Perspective*. New York, NY: Springer, second edition.

Bianco, V., O. Manca, & S. Nardini (2009): "Electricity consumption forecasting in Italy using linear regression models." *Energy* **34(9)**: pp. 1413–1421.

Biggar, D. R. & M. R. Hesamzadeh (2014): *The Economics of Electricity Markets*. Chichester: John Wiley & Sons.

Bischl, B., M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, & Z. M. Jones (2016): "mlr: Machine Learning in R." *Journal of Machine Learning Research* **17(170)**: pp. 1–5.

Boehmke, B. (2018): "Regression Trees." *UC Business Analytics R Programming Guide*. URL: `https://uc-r.github.io/regression_trees`. [Accessed 2023-02-25].

Boldiš, Z. (2013): "Czech electricity grid challenged by German wind." *Europhysics News* **44(4)**: pp. 16–18.

Bouktif, S., A. Fiaz, A. Ouni, & M. A. Serhani (2020): "Multi-Sequence LSTM-RNN Deep Learning and Metaheuristics for Electric Load Forecasting." *Energies* **13, 391(2)**.

Bowen, T., I. Chernyakhovskiy, & P. L. Denholm (2019): "Grid-Scale Battery Storage: Frequently Asked Questions." *Technical Report NREL/TP-6A20-74426*, National Renewable Energy Lab.

Box, G. E. P., G. M. Jenkins, G. C. Reinsel, & G. M. Ljung (2015): *Time Series Analysis: Forecasting and Control*. Hoboken, NJ: John Wiley & Sons, fifth edition.

Brooks, C. (2014): *Introductory Econometrics for Finance*. Cambridge: Cambridge University Press, third edition.

Center for Sustainable Systems (2021): "Wind energy factsheet." *University of Michigan, Pub. No. CSS07-09*. URL: `https://css.umich.edu/publications/factsheets/energy/wind-energy-factsheet`. [Accessed 2022-12-22].

ČEPS (2013): "Dispečeři Čeps mají nový informační systém pro situační zpravodajství." *ČEPS, a.s.* URL: https://www.ceps.cz/cs/novinka/dispeceri-ceps-maji-novy-informacni-system-pro-situacni-zpravodajstvi. [Accessed 2023-04-04].

ČEPS (2020): "Kodex přenosové soustavy (Revision 20, April 2020)." *ČEPS, a.s.* URL: https://www.ceps.cz/cs/kodex-ps. [Accessed 2022-12-02].

ČEPS (2021): "POCKET-BOOK 2021 Ensuring balanced energy flow." *ČEPS, a.s.* URL: https://web.archive.org/web/20221204170040/https://www.ceps.cz/download-data/?format=pagefile&path=9901/modules/files/66376_pocket-book-2021.pdf. [Accessed 2022-12-04].

ČEPS (2022): "All data." *ČEPS, a.s.* URL: https://www.ceps.cz/en/all-data. [Accessed 2022-06-02].

ČEPS (2023a): "Hodnocení provozu." *ČEPS, a.s.* URL: https://www.ceps.cz/cs/hodnoceni-provozu. [Accessed 2023-03-10].

ČEPS (2023b): "Měsíční příprava provozu." *ČEPS, a.s.* URL: https://web.archive.org/web/20230310135555*/https://www.ceps.cz/download-data/?format=pagefile&path=10934/modules/files/87845_mpp-01-2023.pdf. [Accessed 2023-03-10].

Chamberlain, S. (2020): *isdparser: Parse 'NOAA' Integrated Surface Data Files.* R package version 0.4.0.

Chen, H., C. Canizares, & A. Singh (2001): "ANN-based short-term load forecasting in electricity markets." In "2001 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No.01CH37194)," volume 2, pp. 411–415.

Coelho, E. P. R., M. H. M. Paiva, M. E. V. Segatto, & G. Caporossi (2019): "A New Approach for Contingency Analysis Based on Centrality Measures." *IEEE Systems Journal* **13(2)**: pp. 1915–1923.

Crozier, C., T. Morstyn, & M. McCulloch (2020): "The opportunity for smart charging to mitigate the impact of electric vehicles on transmission and distribution systems." *Applied Energy* **268, 114973**.

Czech Statistical Office (2013): "Historická ročenka statistiky energetiky." *Czech Statistical Office.* URL: https://www.czso.cz/csu/czso/8113-12-n_2012-01. [Accessed 2022-04-18].

Dahl, D. B., D. Scott, C. Roosen, A. Magnusson, & J. Swinton (2019): *xtable: Export Tables to LaTeX or HTML.* R package version 1.8-4.

D'Andrade, B. (2017): *The Power Grid: Smart, Secure, Green and Reliable.* London: Academic Press.

Darbellay, G. A. & M. Slama (2000): "Forecasting the short-term demand for electricity: Do neural networks stand a better chance?" *International Journal of Forecasting* **16(1)**: pp. 71–83.

Demšar, J. (2006): "Statistical Comparisons of Classifiers over Multiple Data Sets." *The Journal of Machine Learning Research* **7**: pp. 1–30.

Diebold, F. X. & R. S. Mariano (1995): "Comparing predictive accuracy." *, Journal of Business & Economic Statistics* **13(3)**: pp. 253–263.

Dowell, J. & P. Pinson (2016): "Very-Short-Term Probabilistic Wind Power Forecasts by Sparse Vector Autoregression." *IEEE Transactions on Smart Grid* **7(2)**: pp. 763–770.

Dryar, H. A. (1944): "The effect of weather on the system load." *Electrical Engineering* **63(12)**: pp. 1006–1013.

EIA (2009): "Residential Energy Consumption Survey." *United States Energy Information Administration*. URL: https://www.eia.gov/consumption/residential/reports/2009/air-conditioning.php. [Accessed 2022-09-18].

EIA (2022): "Electricity explained: How electricity is delivered to consumers." *United States Energy Information Administration*. URL: https://www.eia.gov/energyexplained/electricity/delivery-to-consumers.php. [Accessed 2023-01-14].

Elamin, N. & M. Fukushige (2018): "Modeling and forecasting hourly electricity demand by SARIMAX with interactions." *Energy* **165**: pp. 257–268.

Enders, W. (2015): *Applied Econometric Time Series*. Hoboken, NJ: John Wiley & Sons, fourth edition.

EPA (2022): "Climate Change Indicators: Heating and Cooling Degree Days." *United States Environmental Protection Agency*. URL: https://www.epa.gov/climate-indicators/climate-change-indicators-heating-and-cooling-degree-days. [Accessed 2022-09-22].

Falbel, D. & J. Luraschi (2023): *torch: Tensors and Neural Networks with 'GPU' Acceleration*. R package version 0.9.1.

Fan, S. & R. J. Hyndman (2012): "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model." *IEEE Transactions on Power Systems* **27(1)**: pp. 134–141.

Feinberg, E. A. & D. Genethliou (2005): "Load Forecasting." In J. H. Chow, F. F. Wu, & J. Momoh (editors), "Applied Mathematics for Restructured Electric Power Systems: Optimization, Control, and Computational Intelligence," Power Electronics and Power Systems, pp. 269–285. Boston, MA: Springer US.

Ghiassi, M., D. K. Zimbra, & H. Saidane (2006): "Medium term system load forecasting with a dynamic artificial neural network model." *Electric Power Systems Research* **76(5)**: pp. 302–316.

Goodfellow, I., Y. Bengio, & A. Courville (2016): *Deep Learning*. Cambridge, MA: MIT Press.

Grolemund, G. & H. Wickham (2011): "Dates and times made easy with lubridate." *Journal of Statistical Software* **40(3)**: pp. 1–25.

Guan, C., P. B. Luh, L. D. Michel, Y. Wang, & P. B. Friedland (2013): "Very Short-Term Load Forecasting: Wavelet Neural Networks With Data Pre-Filtering." *IEEE Transactions on Power Systems* **28(1)**: pp. 30–41.

Hagan, M. T. & S. M. Behr (1987): "The Time Series Approach to Short Term Load Forecasting." *IEEE Transactions on Power Systems* **2(3)**: pp. 785–791.

Hanifi, S., X. Liu, Z. Lin, & S. Lotfian (2020): "A Critical Review of Wind Power Forecasting Methods—Past, Present and Future." *Energies* **13, 3764(15)**.

Hanzlík, V., V. Javůrek, B. Smeets, & D. Svoboda (2020): "Pathways to decarbonize the Czech Republic: Carbon-neutral Czech Republic 2050." *McKinsey & Company*. URL: https://www.mckinsey.com/cz/our-work/pathways-to-decarbonize-the-czech-republic. [Accessed 2022-12-18].

Hardy, R. (2016): "Diebold-Mariano test for multiple prediction horizons." *Cross Validated*. URL: https://stats.stackexchange.com/q/231862. [Accessed 2023-04-04].

Harris, C. (2006): *Electricity Markets: Pricing, Structures and Economics*. Chichester: John Wiley & Sons.

Harvey, D., S. Leybourne, & P. Newbold (1997): "Testing the equality of prediction mean squared errors." *International Journal of Forecasting* **13(2)**: pp. 281–291.

Heinemann, G. T., D. A. Nordmian, & E. C. Plant (1966): "The Relationship Between Summer Weather and Summer Loads - A Regression Analysis." *IEEE Transactions on Power Apparatus and Systems* **PAS-85(11)**: pp. 1144–1154.

Hobbs, B., S. Jitprapaikulsarn, S. Konda, V. Chankong, K. Loparo, & D. Maratukulam (1999): "Analysis of the value for unit commitment of improved load forecasts." *IEEE Transactions on Power Systems* **14(4)**: pp. 1342–1348.

Hofmann, F., M. Schlott, A. Kies, & H. Stöcker (2020): "Flow Allocation in Meshed AC-DC Electricity Grids." *Energies* **13, 1233(5)**.

Hong, T. (2014): "Energy Forecasting: Past, Present, and Future." *Foresight: The International Journal of Applied Forecasting* **32**: pp. 43–48.

Hong, T. & S. Fan (2016): "Probabilistic electric load forecasting: A tutorial review." *International Journal of Forecasting* **32(3)**: pp. 914–938.

Hong, T., P. Pinson, Y. Wang, R. Weron, D. Yang, & H. Zareipour (2020): "Energy Forecasting: A Review and Outlook." *IEEE Open Access Journal of Power and Energy* **7**: pp. 376–388.

Hooke, R. G. (1955): "Forecasting the Demand for Electricity." *Transactions of the American Institute of Electrical Engineers. Part III: Power Apparatus and Systems* **74(3)**: pp. 993–1008.

Hunter, J. D. (2007): "Matplotlib: A 2d graphics environment." *Computing in Science & Engineering* **9(3)**: pp. 90–95.

Hyndman, R. (2010): "The ARIMAX model muddle." *Hyndsight blog.* URL: `https://robjhyndman.com/hyndsight/arimax/`. [Accessed 2023-02-07].

Hyndman, R. & G. Athanasopoulos (2021): *Forecasting: Principles and Practice.* OTexts, third edition.

Hyndman, R. J. & Y. Khandakar (2008): "Automatic time series forecasting: The forecast package for R." *Journal of Statistical Software* **27(3)**: pp. 1–22.

IPCC (2022): *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty.* Cambridge: Cambridge University Press.

James, G., D. Witten, T. Hastie, & R. Tibshirani (2021): *An Introduction to Statistical Learning*, volume 112. New York, NY: Springer, second edition.

Jensen, L. (2021): "Climate action in czechia: Latest state of play." *European Parliament: European Parliamentary Research Service.* URL: `https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)689329`. [Accessed 2022-12-18].

Jung, S.-M., S. Park, S.-W. Jung, & E. Hwang (2020): "Monthly Electric Load Forecasting Using Transfer Learning for Smart Cities." *Sustainability* **12, 6364**.

Kandil, N., R. Wamkeue, M. Saad, & S. Georges (2006): "An efficient approach for short term load forecasting using artificial neural networks." *International Journal of Electrical Power & Energy Systems* **28(8)**: pp. 525–530.

Keydana, S. (2021a): "Introductory time-series forecasting with torch." *Posit AI Blog.* URL: `https://blogs.rstudio.com/tensorflow/posts/2021-03-10-forecasting-time-series-with-torch_1`. [Accessed 2023-03-14].

Keydana, S. (2021b): "Torch time series continued: A first go at multi-step prediction." *Posit AI Blog*. URL: https://blogs.rstudio.com/tensorflow/posts/2021-03-11-forecasting-time-series-with-torch_2. [Accessed 2023-03-14].

Khan, M. R., A. Abraham, & C. Ondrusek (2002): "Soft computing for developing short term load forecasting models in Czech Republic." In A. Abraham & M. Koppen (editors), "Hybrid Information Systems," pp. 207–221. Heidelberg: Physica-Verlag Gmbh & Co.

Khotanzad, A., R. Afkhami-Rohani, & D. Maratukulam (1998): "ANNSTLF-Artificial Neural Network Short-Term Load Forecaster-generation three." *IEEE Transactions on Power Systems* **13(4)**: pp. 1413–1422.

Khuntia, S. R., B. W. Tuinema, J. L. Rueda, & M. A. van der Meijden (2016): "Time-horizons in the planning and operation of transmission networks: An overview." *IET Generation, Transmission & Distribution* **10(4)**: pp. 841–848.

Kim, Y., H.-g. Son, & S. Kim (2019): "Short term electricity load forecasting for institutional buildings." *Energy Reports* **5**: pp. 1270–1280.

Kingma, D. P. & J. Ba (2014): "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*.

Kirschen, D. S. & G. Strbac (2004): *Fundamentals of Power System Economics*. Chichester: John Wiley & Sons, first edition.

Kittner, N., O. Schmidt, I. Staffell, & D. M. Kammen (2020): "Grid-scale energy storage." In M. Junginger & A. Louwen (editors), "Technological Learning in the Transition to a Low-Carbon Energy System," pp. 119–143. London: Academic Press.

Koning, A. J., P. H. Franses, M. Hibon, & H. O. Stekler (2005): "The M3 competition: Statistical tests of the results." *International Journal of Forecasting* **21(3)**: pp. 397–409.

Kourentzes, N. (2022): *tsutils: Time Series Exploration, Modelling and Forecasting*. R package version 0.9.3.

Křížová, K. (2021): *Forecasting Electricity Pricing in Central and Eastern Europe*. Master's thesis, Charles University.

Krugman, P. & R. Wells (2015): *Economics*. New York, NY: Worth Publishers, fourth edition.

Kuhn, M. (2022): *caret: Classification and Regression Training*. R package version 6.0-93.

Kuster, C., Y. Rezgui, & M. Mourshed (2017): "Electrical load forecasting models: A critical systematic review." *Sustainable Cities and Society* **35**: pp. 257–270.

Kwon, B.-S., R.-J. Park, & K.-B. Song (2020): "Short-Term Load Forecasting Based on Deep Neural Networks Using LSTM Layer." *Journal of Electrical Engineering & Technology* **15(4)**: pp. 1501–1509.

Lago, J., F. De Ridder, & B. De Schutter (2018): "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms." *Applied Energy* **221**: pp. 386–405.

Lai, C. S., Z. Mo, T. Wang, H. Yuan, W. W. Y. Ng, & L. L. Lai (2020): "Load forecasting based on deep neural network and historical data augmentation." *Iet Generation Transmission & Distribution* **14(24)**: pp. 5927–5934.

Laurinec, P. (2017): "Using regression trees for forecasting double-seasonal time series with trend in R." *Peter Laurinec's Blog*. URL: https://petolau.github.io/Regression-trees-for-forecasting-time-series-in-R/. [Accessed 2023-03-10].

Le, X.-H., H. V. Ho, G. Lee, & S. Jung (2019): "Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting." *Water* **11, 1387(7)**.

Lee, J. & Y. Cho (2022): "National-scale electricity peak load forecasting: Traditional, machine learning, or hybrid model?" *Energy* **239, 122366**.

Lepot, M., J.-B. Aubin, & F. Clemens (2017): "Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment." *Water* **9, 796**.

Liao, G.-C. & T.-P. Tsao (2004): "Application of fuzzy neural networks and artificial intelligence for load forecasting." *Electric Power Systems Research* **70(3)**: pp. 237–244.

Liu, L., F. Kong, X. Liu, Y. Peng, & Q. Wang (2015): "A review on electric vehicles interacting with renewable energy in smart grid." *Renewable and Sustainable Energy Reviews* **51**: pp. 648–661.

Lütkepohl, H. & M. Krätzig (2004): *Applied Time Series Econometrics*. Cambridge: Cambridge University Press.

Mahdavi, M., C. Sabillon Antunez, M. Ajalli, & R. Romero (2019): "Transmission Expansion Planning: Literature Review and Classification." *IEEE Systems Journal* **13(3)**: pp. 3129–3140.

Malik, H., N. Fatema, & A. Iqbal (2021): "Intelligent Data Analytics for Time-Series Load Forecasting Using Fuzzy Reinforcement Learning (FRL)." In H. Malik, N. Fatema, & A. Iqbal (editors), "Intelligent Data-Analytics for Condition Monitoring," pp. 193–213. London: Academic Press.

Mamun, A. A., M. Sohel, N. Mohammad, M. S. Haque Sunny, D. R. Dipta, & E. Hossain (2020): "A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models." *IEEE Access* **8**: pp. 134911–134939.

Marino, D. L., K. Amarasinghe, & M. Manic (2016): "Building energy load forecasting using deep neural networks." In "IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society," pp. 7046–7051. IEEE.

McKinney, W. (2010): "Data Structures for Statistical Computing in Python." In Stéfan van der Walt & Jarrod Millman (editors), "Proceedings of the 9th Python in Science Conference," pp. 56–61.

Milborrow, S. (2022): *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. R package version 3.1.1.

Moritz, S. & T. Bartz-Beielstein (2017): "imputeTS: Time Series Missing Value Imputation in R." *The R Journal* **9(1)**: pp. 207–218.

Morris, C. (2013): "The flattening of peak and base prices." *Energy Transition*. URL: `https://energytransition.org/2013/05/the-flattening-of-peak-and-base-prices/`. [Accessed 2023-02-07].

Müller, K. & H. Wickham (2022): *tibble: Simple Data Frames*. R package version 3.1.8.

NOAA (2001): "Integrated Surface Dataset (Global)." *National Centers for Environmental Information*. URL: `https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00532`. [Accessed 2022-06-02].

Olah, C. (2015): "Understanding lstm networks." *Christopher Olah's Blog*. URL: `http://colah.github.io/posts/2015-08-Understanding-LSTMs`. [Accessed 2023-03-04].

OTE (2022): "Day-Ahead Market." *OTE, a.s.* URL: `https://www.ote-cr.cz/en/short-term-markets/electricity/day-ahead-market`. [Accessed 2022-06-02].

Papaioannou, G. P., C. Dikaiakos, A. Dramountanis, & P. G. Papaioannou (2016): "Analysis and Modeling for Short- to Medium-Term Load Forecasting Using a Hybrid Manifold Learning Principal Component Model and Comparison with Classical Statistical Models (SARIMAX, Exponential Smoothing) and Artificial Intelligence Models (ANN, SVM): The Case of Greek Electricity Market." *Energies* **9, 635(8)**.

Peters, A. & T. Hothorn (2022): *ipred: Improved Predictors*. R package version 0.9-13.

Pfenninger, S., A. Hawkes, & J. Keirstead (2014): "Energy systems modeling for twenty-first century energy challenges." *Renewable and Sustainable Energy Reviews* **33**: pp. 74–86.

Plotly Technologies Inc. (2015): *plotly: Collaborative data science*. Python package version 5.11.0.

Poullikkas, A. (2013): "A comparative overview of large-scale battery systems for electricity storage." *Renewable and Sustainable Energy Reviews* **27**: pp. 778–788.

Qiu, D. (2015): *aTSA: Alternative Time Series Analysis*. R package version 3.1.2.

Quintans, D. (2021): *librarian: Install, Update, Load Packages from CRAN, 'GitHub', and 'Bioconductor' in One Step*. R package version 1.8.1.

R Core Team (2022): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranaweera, D., G. Karady, & R. Farmer (1997): "Economic impact analysis of load forecasting." *IEEE Transactions on Power Systems* **12(3)**: pp. 1388–1392.

Richardson, H. (2022): "What does the world's largest machine do?" *TED-Ed*. URL: `https://www.ted.com/talks/henry_richardson_what_does_the_world_s_largest_machine_do`. [Accessed 2022-12-02].

Ripley, B. D. (1996): *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, first edition.

Ritchie, H., M. Roser, & P. Rosado (2022): "Energy." *Our World in Data*. URL: `https://ourworldindata.org/energy`. [Accessed 2022-12-04].

Rossi, B. (2013): "Exchange Rate Predictability." *Journal of Economic Literature* **51(4)**: pp. 1063–1119.

Ruiz-Abellón, M. D. C., A. Gabaldón, & A. Guillamón (2018): "Load Forecasting for a Campus University Using Ensemble Methods Based on Regression Trees." *Energies* **11, 2038(8)**.

Ryan, J. A. & J. M. Ulrich (2023): *xts: eXtensible Time Series*. R package version 0.13.0.

Sachdev, M., R. Billinton, & C. Peterson (1977): "Representative bibliography on load forecasting." *IEEE Transactions on Power Apparatus and Systems* **96(2)**: pp. 697–700.

Sewalt, M. & C. De Jong (2003): "Negative prices in electricity markets." *Commodities Now* **7(74)**: pp. 74–77.

Seznam Zprávy (2022): "Velký přehled: Dva roky s koronavirem v Česku." *Seznam Zprávy*. URL: `https://www.seznamzpravy.cz/clanek/fakta-velky-prehled-dva-roky-s-koronavirem-v-cesku-190958`. [Accessed 2023-03-04].

Signorell, A. (2023): *DescTools: Tools for Descriptive Statistics*. R package version 0.99.48.

Soloveichik, G. L. (2011): "Battery Technologies for Large-Scale Stationary Energy Storage." *Annual Review of Chemical and Biomolecular Engineering* **2(1)**: pp. 503–527.

Stock, J. H. & M. W. Watson (2020): *Introduction to Econometrics*. Harlow: Pearson Education Limited, fourth edition.

Svetunkov, I. (2022): "Forecasting and Analytics with ADAM." *OpenForecast*. URL: `https://openforecast.org/adam/`. [Accessed 2023-04-04].

Taylor, J. W. (2008): "An evaluation of methods for very short-term load forecasting using minute-by-minute British data." *International Journal of Forecasting* **24(4)**: pp. 645–658.

Taylor, J. W. (2012): "Short-Term Load Forecasting With Exponentially Weighted Methods." *IEEE Transactions on Power Systems* **27(1)**: pp. 458–464.

The Economist (2022a): "The French exception." *The Economist*. URL: `https://www.economist.com/leaders/2022/12/15/the-french-exception`. [Accessed 2022-12-18].

The Economist (2022b): "The world is going to miss the totemic 1.5°C climate target." *The Economist*. URL: `https://www.economist.com/interactive/briefing/2022/11/05/the-world-is-going-to-miss-the-totemic-1-5c-climate-target`. [Accessed 2022-12-22].

Therneau, T. & B. Atkinson (2022): *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.19.

Trapletti, A. & K. Hornik (2023): *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-53.

Tsay, R. S. (2005): *Analysis of Financial Time Series*. Hoboken, NJ: John Wiley & Sons, second edition.

Tuballa, M. L. & M. L. Abundo (2016): "A review of the development of Smart Grid technologies." *Renewable and Sustainable Energy Reviews* **59**: pp. 710–725.

Uher, V., R. Burget, M. K. Dutta, & P. Mlynek (2015): "Forecasting electricity consumption in Czech Republic." In "2015 38th International Conference on Telecommunications and Signal Processing (TSP)," pp. 262–265.

Ushey, K., J. Allaire, H. Wickham, & G. Ritchie (2022): *rstudioapi: Safely Access the RStudio API*. R package version 0.14.

van Ackooij, W., I. Danti Lopez, A. Frangioni, F. Lacalandra, & M. Tahanan (2018): "Large-scale unit commitment under uncertainty: An updated literature survey." *Annals of Operations Research* **271(1)**: pp. 11–85.

Van Rossum, G. & F. L. Drake (2009): *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vlček, T., F. Černoch, V. Zapletalová, E. Trmalová, J. Červinková, P. Brhlíková, T. Stašáková, L. Bodišová, G. Prokopová, & P. Bendlová (2019): *The Energy Sector and Energy Policy of the Czech Republic*. Brno: Masaryk University Press, second edition.

Wang, E., D. Cook, & R. J. Hyndman (2020): "A new tidy data structure to support exploration and modeling of temporal data." *Journal of Computational and Graphical Statistics* **29(3)**: pp. 466–478.

Wang, P., B. Liu, & T. Hong (2016): "Electric load forecasting with recency effect: A big data approach." *International Journal of Forecasting* **32(3)**: pp. 585–597.

Wang, Y. & L. Wu (2017): "Improving economic values of day-ahead load forecasts to real-time power system operations." *IET Generation, Transmission & Distribution* **11(17)**: pp. 4238–4247.

Wang, Z., T. Hong, H. Li, & M. Ann Piette (2021): "Predicting city-scale daily electricity consumption using data-driven models." *Advances in Applied Energy* **2, 100025**.

Weron, R. (2006): *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach.* Chichester: John Wiley & Sons.

Weron, R. (2014): "Electricity price forecasting: A review of the state-of-the-art with a look into the future." *International Journal of Forecasting* **30(4)**: pp. 1030–1081.

Wickham, H. (2016): *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer.

Wickham, H. & J. Bryan (2023): *readxl: Read Excel Files.* R package version 1.4.2.

Wickham, H., R. François, L. Henry, K. Müller, & D. Vaughan (2023a): *dplyr: A Grammar of Data Manipulation.* R package version 1.1.0.

Wickham, H. & L. Henry (2023): *purrr: Functional Programming Tools.* R package version 1.0.1.

Wickham, H., D. Vaughan, & M. Girlich (2023b): *tidyr: Tidy Messy Data.* R package version 1.3.0.

Xiao, N. (2023): *ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'.* R package version 3.0.0.

Yildiz, B., J. I. Bilbao, & A. B. Sproul (2017): "A review and analysis of regression and machine learning models on commercial building electricity load forecasting." *Renewable and Sustainable Energy Reviews* **73**: pp. 1104–1122.

Zeileis, A. & G. Grothendieck (2005): "zoo: S3 infrastructure for regular and irregular time series." *Journal of Statistical Software* **14(6)**: pp. 1–27.

Appendix A

Additional Definitions

In this section, we provide additional definitions and explanations referenced in the main body of the thesis.

A.1 Sigmoid and Hyperbolic Tangent

The sigmoid and hyperbolic tangent functions are widely used in artificial neural networks as activations, and can be written as

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$
$$tanh(z) = 2\sigma(2z) - 1,$$

following Goodfellow *et al.* (2016, p. 191) and James *et al.* (2021, p. 405).

A.2 ARIMA Model Selection

The Hyndman & Khandakar (2008) non-seasonal ARIMA(p, d, q) step-wise model selection algorithm first fits the following specifications: $(0, d, 1)$, $(1, d, 0)$, $(0, d, 0)$, and $(2, d, 2)$, where the order of differencing d is selected based on the results of the Kwiatkowski–Phillips–Schmidt–Shin test. If the d parameter is set to be equal to zero or one, a constant term is added, and a fifth specification with no intercept (ARIMA$(0, d, 0)$) is tested (Hyndman & Athanasopoulos 2021, sec. 9.7). Afterward, if the Akaike information criterion is specified as the goodness of fit measure, it is calculated as

$$AIC = -2log(L) + 2(p + q + k),$$

where L is the model's likelihood, p & q are the AR & MA orders, respectively, and $k = 1$ if an intercept term is included, $k = 0$ otherwise. The lowest-AIC model is then chosen as a temporary baseline.

In the second step of the algorithm, models with $p \pm 1$, $q \pm 1$, and an excluded or included intercept term are fitted based on the specification from the previous step. If a model with lower AIC is found, then it is treated as the new baseline. This stage is repeated until there is no *adjacent* specification with lower AIC. Finally, by default, the maximum order of p and q is set to 5.

A.3 Ljung-Box and ARCH Tests

The term *autocorrelation* refers to the linear relationship between u_t (in our case, u_t refers to the model's residuals) and its lags (Tsay 2005, p. 26). For lag ℓ, autocorrelation ρ_ℓ can be written as

$$\rho_\ell = \frac{Cov(u_t, u_{t-\ell})}{Var(u_t)},$$

where u_t is assumed to be weakly stationary, i.e., its mean, variance, and autocovariance (top term in the above equation) are constant (Tsay 2005, p. 26, Brooks 2014, pp. 252–253).

The Ljung-Box test augments the Portmanteau test, which is used to determine whether several autocorrelations of u_t are statistically different from zero, and defines the following test statistic

$$Q(m) = T(T+2)\sum_{\ell=1}^{m}\frac{\hat{\rho}_\ell^2}{T-\ell},$$

where T refers to the sample size, $\hat{\rho}$ is the sample autocorrelation, and the statistic asymptotically follows a chi-square distribution with m degrees of freedom (Tsay 2005, p. 27).

The test for ARCH effects, on the other hand, is concerned with squared residuals u_t^2, and tests the null hypothesis of $H_0 : \rho_1 = ... = \rho_m = 0$, i.e., no autocorrelation between u_t^2 and its lags. The test employs the original Portmanteau statistic, which is formulated as

$$Q^*(m) = T\sum_{\ell=1}^{m}\hat{\rho}_\ell^2,$$

where, with m degrees of freedom, $Q^*(m)$ is asymptotically chi-square dis-

tributed, T is the size of the sample, and $\hat{\rho}$ refers to the sample autocorrelation (Tsay 2005, pp. 26–27, 101).

A.4 Jarque-Bera Test

Following Tsay (2005, pp. 9–10), the *skewness* and the *excess kurtosis* of a normally distributed random variable are both equal to zero. The Jarque-Bera test utilizes these two facts and defines a test statistic that can be written as

$$JB = \frac{\hat{S}^2(x)}{6/T} + \frac{(\hat{K}(x) - 3)^2}{24/T},$$

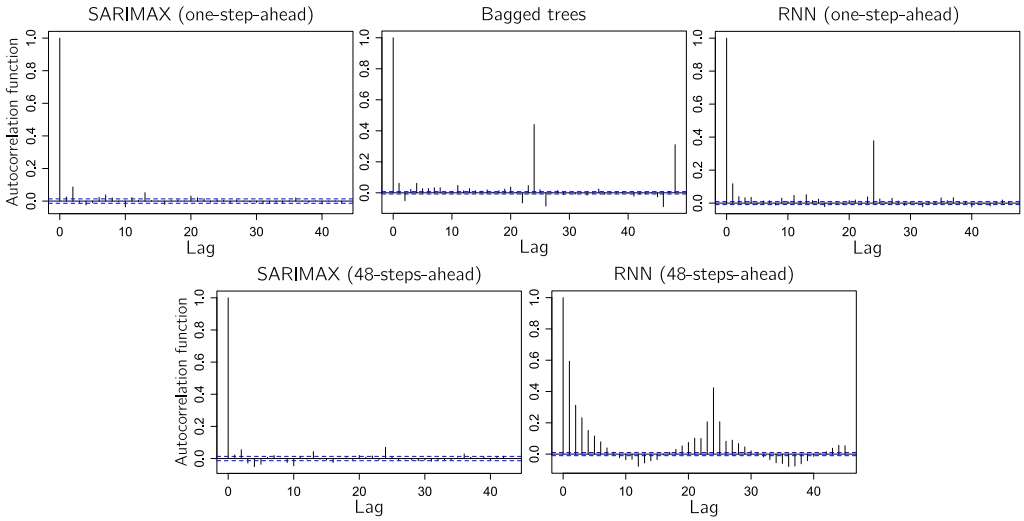
where $\hat{S}(x)$ & $\hat{K}(x)$ refer to sample skewness and kurtosis, respectively, while T is the sample size. Finally, the JB statistic is asymptotically chi-squared distributed with two degrees of freedom. The null hypothesis of the JB test is that the variable in question is normally distributed (Tsay 2005, pp. 9–10).

A.5 Hourly Data: Residual Diagnostics

Firstly, let us note that varying lengths of the in-sample set were used in most methods. Thus, pairwise comparisons of residual diagnostics should not be attempted. In Figure A.1, we may observe the autocorrelation functions of residuals for all standard methods across the two different hourly schemes. It is apparent that in both SARIMAX models, which are perhaps the only models where standard residual diagnostics are reasonable, a considerable degree of information appears to be accounted for. In all other models' residuals, there seems to be, at the very least, a spike at the 24th lag, which can be attributed to daily seasonality—the reason why this does not occur in the SARIMAX models is likely due to the s parameter being set to 24 (see Equation 5.3).

In terms of diagnostic tests, we conducted the Ljung-Box test for serial correlation, outlined in Section A.3, for lags of 1 to 7. For all models and for each of the seven lag lengths, the null hypothesis was rejected at the significance level of 5%, implying the presence of serial correlation. Similarly, in the ARCH tests (see Section A.3), for all tested lag lengths (1 to 7), the null hypothesis was rejected in all models (p-value below 5%). Let us note that in Papaioannou *et al.* (2016), which is one of the works where residual diagnostics are conducted (though only for SARIMAX), the authors fail to reject the null hypothesis in both the Ljung-Box and the ARCH tests. However, they do not appear to

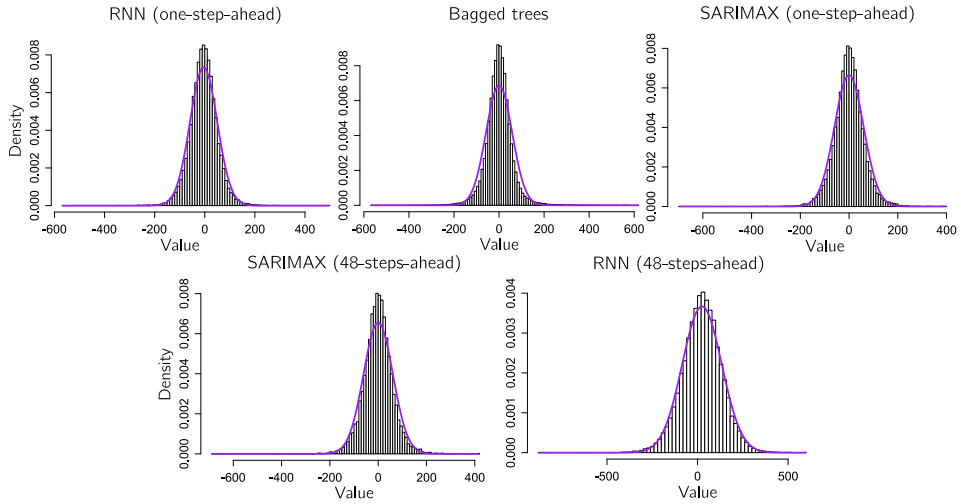
Figure A.1: Autocorrelation functions of residuals from all models



disclose the tested lag lengths, and furthermore, their analysis considered daily load data (from a different country). Moreover, in Kim *et al.* (2019), which is concerned with hourly building-level loads, the authors report mixed results in terms of residual autocorrelation with their ARIMA specifications.

Finally, in Figure A.2, we plot histograms of residuals with plots of the normal density function in each. In comparison with the normal distribution, it is immediately apparent that the residuals in all five panels exhibit excess kurtosis, suggesting a lack of normality. This is further corroborated by the results of the Jarque-Bera test (see Section A.4), as the null hypothesis of a normal distribution is rejected in each instance.

Figure A.2: Histograms of residuals from all models



Note: Normal density plot added for illustrative purposes.

Appendix B

Additional Results

This appendix includes additional tables and figures we refer to in the thesis.

Table B.1: Summary statistics of 1-minute load (2011 to 2021)

Variable	Min	Q1	Median	Mean	Q3	Max	SD
Load (MW)	4343	6881	7882	7923.5	8851	12569	1357.1

Table B.2: Augmented Dickey-Fuller test results on raw data (2011 to 2021)

	ADF test H_0 rejections*		
Variable	**None**	**Drift**	**Drift & Trend**
Load (minute)	0/10	0/10	0/10
Load (hourly)	4/10	5/10	5/10
Air pressure	0/10	9/10	9/10
Air pressure (station diff.)	0/10	0/10	0/10
Price	0/10	0/10	0/10
Temperature	3/10	3/10	3/10
Temperature (station diff.)	6/10	6/10	6/10
Visibility	0/10	0/10	0/10
Visibility (station diff.)	0/10	0/10	0/10
Wind speed	0/10	0/10	0/10
Wind speed (station diff.)	0/10	0/10	0/10

Note: *p-value below 5%; Hourly data taken since 2012. Rejection of the null hypothesis suggests that a unit root is not present.

Table B.3: Augmented Dickey-Fuller test results on a subset of differenced data (2011 to 2021)

	Standard ADF test	
Differenced variable	**Lag order**	**p-value**
Load (minute)	179	<0.01
Air pressure (station diff.)	45	<0.01
Price	45	<0.01
Temperature	45	<0.01
Visibility	45	<0.01
Visibility (station diff.)	45	<0.01
Wind speed	45	<0.01
Wind speed (station diff.)	45	<0.01

Note: Hourly data taken since 2012. Rejection of the null hypothesis suggests that a unit root is not present.

Table B.4: Summary statistics of additional predictors (hourly data from 2012 to 2021)

Variable	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Load Avg. 7D	6287	7256	7778	7948	8615	10 539
Load Max 24h	6176	8209	8934	8989	9743	12 133
Air Pressure Avg. 7D	995	1013	1017	1017	1021	1036
Air Pressure Max 24h	980	1015	1019	1020	1025	1048
ΔAir Pressure Avg. 7D	−4.623	−0.914	−0.070	−0.174	0.626	2.850
ΔAir Pressure Max 24h	−6.200	0.000	1.200	1.272	2.400	9.800
Price Avg. 7D	1.151	31.468	37.458	43.147	44.554	312.895
Price Max 24h	2.22	43.60	53.43	62.51	66.12	620.00
Temperature Avg. 7D	−12.281	3.530	9.993	10.082	16.944	27.739
Temperature Max 24h	−11.925	5.775	12.650	12.788	19.900	35.475
ΔTemperature Avg. 7D	−5.283	−2.055	−1.243	−1.117	−0.282	5.923
ΔTemperature Max 24h	−13.000	2.050	4.850	5.093	7.800	20.200
Visibility Avg. 7D	2.444	19.849	29.343	28.250	37.166	56.488
Visibility Max 24h	0.80	27.50	42.50	40.18	55.00	75.95
ΔVisibility Avg. 7D	−11.354	1.220	6.142	6.564	11.515	35.442
ΔVisibility Max 24h	−19.50	10.00	25.00	25.59	40.00	81.80
Wind Speed Avg. 7D	1.956	3.089	3.485	3.604	4.003	6.896
Wind Speed Max 24h	1.650	4.200	5.200	5.366	6.325	14.825
ΔWind Speed Avg. 7D	−3.212	−0.311	0.248	0.398	1.026	4.653
ΔWind Speed Max 24h	−4.800	2.050	3.300	3.739	5.100	14.750

Note: "Δ" refers to the difference between 2 weather stations. "Avg. 7D" means average in the past 7 days. "Max 24h" refers to the maximum in the last 24 hours.

Table B.6: Results of the Ljung-Box tests of ARIMA residuals

Lags	H_0 **rejected***	**Failed to reject** $H_0{}^*$
4	6x	4010x
5	25x	3991x
6	49x	3967x
7	66x	3950x

Note: *number of times the null hypothesis was/was not rejected at the significance level of 5%. The alternative hypothesis is that residuals exhibit autocorrelation.

Table B.7: Results of the ARCH tests of squared ARIMA residuals

Lags	H_0 **rejected***	**Failed to reject** $H_0{}^*$
4	2624x	1392x
5	2656x	1360x
6	2595x	1421x
7	2553x	1463x

Note: *number of times the null hypothesis was/was not rejected at the significance level of 5%. Rejection of the null hypothesis suggests the presence of ARCH effects.

Table B.8: Results of the Jarque-Bera tests of ARIMA residuals

H_0 **rejected***	**Failed to reject** $H_0{}^*$
3842x	174x

Note: *number of times the null hypothesis was/was not rejected at the significance level of 5%. The null hypothesis is that the residuals are normally distributed.

Figure B.1: Average in-sample and validation loss per epoch

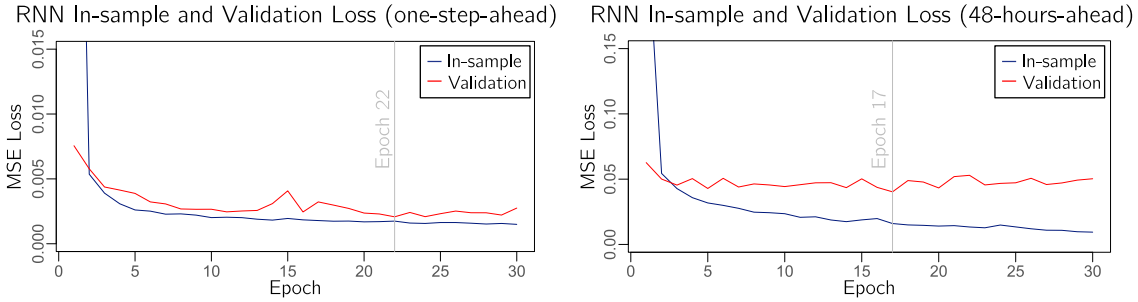


Figure B.2: Sample plots of all 48-hours-ahead forecasts and actual load (2021)

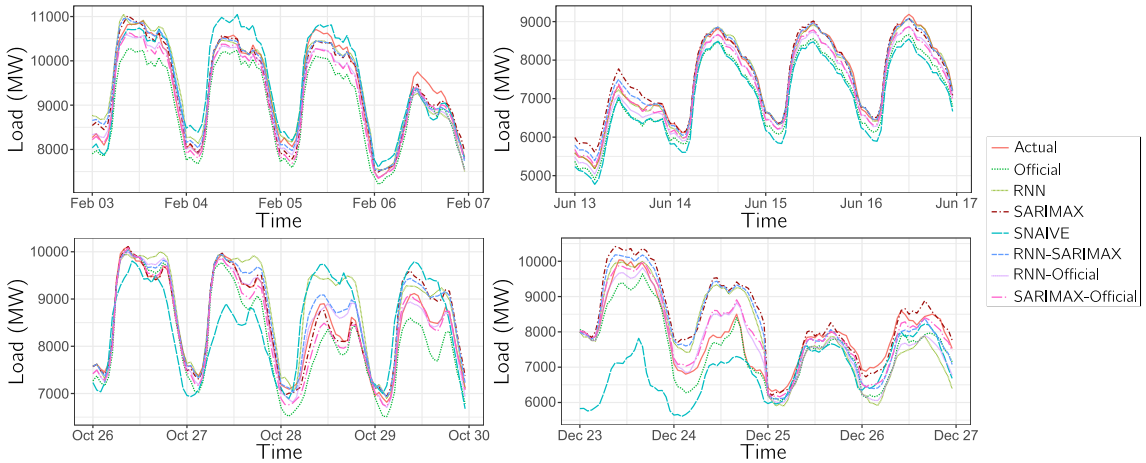


Table B.9: Explanatory variables used in this thesis

Variable name	Description	Directly used in	Variable name (cont.)	Description (cont.)	Directly used in (cont.)
air_pressure	Air pressure (hPa)	1.	wind_speed_diff_t72	Wind speed station diff. at t-72	1.
air_pressure_diff	Air pressure station difference	1.	wind_speed_diff_avg_7_days	Avg. wind speed stat. diff. in 7 days	1.
price_eur_mwh	Price (EUR/MWh)	1.	wind_speed_diff_max_24_hrs	Max wind speed stat. diff. in 24 hours	1.
temperature	Temperature (°C)	1., 2., 3.	mon_Feb	Month of February	1., 2., 3., 4., 5.
temperature_diff	Temperature station difference	1.	mon_Mar	Month of March	1., 2., 3., 4., 5.
visibility_distance	Visibility (km)	1.	mon_Apr	Month of April	1., 2., 3., 4., 5.
visibility_distance_diff	Visibility station difference	1.	mon_May	Month of May	1., 2., 3., 4., 5.
wind_speed	Wind speed (m/s)	1.	mon_Jun	Month of June	1., 2., 3., 4., 5.
wind_speed_diff	Wind speed station difference	1.	mon_Jul	Month of July	1., 2., 3., 4., 5.
load_mw_t1	Load (MW) at t-1	1., 2., 3.	mon_Aug	Month of August	1., 2., 3., 4., 5.
load_mw_t24	Load at t-24	1., 2., 3.	mon_Sep	Month of September	1., 2., 3., 4., 5.
load_mw_t48	Load at t-48	1., 2., 3.	mon_Oct	Month of October	1., 2., 3., 4., 5.
load_mw_t72	Load at t-72	1., 2., 3.	mon_Nov	Month of November	1., 2., 3., 4., 5.
load_mw_avg_7_days	Average load in 7 days	1., 2., 3.	mon_Dec	Month of December	1., 2., 3., 4., 5.
load_mw_max_24_hrs	Max load in 24 hours	1., 2., 3.	day_Mon	Monday indicator	1., 2., 3., 4., 5.
air_pressure_t1	Air pressure at t-1	1.	day_Thu	Tuesday indicator	1., 2., 3., 4., 5.
air_pressure_t24	Air pressure at t-24	1.	day_Tue	Wednesday indicator	1., 2., 3., 4., 5.
air_pressure_t48	Air pressure at t-48	1.	day_Wed	Thursday indicator	1., 2., 3., 4., 5.
air_pressure_t72	Air pressure at t-72	1.	covid_dummy	COVID-19 states of emergency	1., 2., 3., 4., 5.
air_pressure_avg_7_days	Average air pressure in 7 days	1.	hour_01	Indicator for 01:00	1., 2., 3., 4., 5.
air_pressure_max_24_hrs	Max air pressure in 24 hours	1.	hour_02	Indicator for 02:00	1., 2., 3., 4., 5.
air_pressure_diff_t1	Air pressure station diff. at t-1	1.	hour_03	Indicator for 03:00	1., 2., 3., 4., 5.
air_pressure_diff_t24	Air pressure station diff. at t-24	1.	hour_04	Indicator for 04:00	1., 2., 3., 4., 5.
air_pressure_diff_t48	Air pressure station diff. at t-48	1.	hour_05	Indicator for 05:00	1., 2., 3., 4., 5.
air_pressure_diff_t72	Air pressure station diff. at t-72	1.	hour_06	Indicator for 06:00	1., 2., 3., 4., 5.
air_pressure_diff_avg_7_days	Avg. air pressure stat. diff. in 7 days	1.	hour_07	Indicator for 07:00	1., 2., 3., 4., 5.
air_pressure_diff_max_24_hrs	Max air pressure stat. diff. in 24 hours	1.	hour_08	Indicator for 08:00	1., 2., 3., 4., 5.
price_eur_mwh_t1	Price at t-1	1.	hour_09	Indicator for 09:00	1., 2., 3., 4., 5.
price_eur_mwh_t24	Price at t-24	1.	hour_10	Indicator for 10:00	1., 2., 3., 4., 5.
price_eur_mwh_t48	Price at t-48	1.	hour_11	Indicator for 11:00	1., 2., 3., 4., 5.
price_eur_mwh_t72	Price at t-72	1.	hour_12	Indicator for 12:00	1., 2., 3., 4., 5.
price_eur_mwh_avg_7_days	Average price in 7 days	1.	hour_13	Indicator for 13:00	1., 2., 3., 4., 5.
price_eur_mwh_max_24_hrs	Max price in 24 hours	1.	hour_14	Indicator for 14:00	1., 2., 3., 4., 5.
temperature_t1	Temperature at t-1	1., 2., 3.	hour_15	Indicator for 15:00	1., 2., 3., 4., 5.
temperature_t24	Temperature at t-24	1., 2., 3.	hour_16	Indicator for 16:00	1., 2., 3., 4., 5.
temperature_t48	Temperature at t-48	1., 2., 3.	hour_17	Indicator for 17:00	1., 2., 3., 4., 5.
temperature_t72	Temperature at t-72	1., 2., 3.	hour_18	Indicator for 18:00	1., 2., 3., 4., 5.
temperature_avg_7_days	Average temperature in 7 days	1., 2., 3.	hour_19	Indicator for 19:00	1., 2., 3., 4., 5.
temperature_max_24_hrs	Max temperature in 7 days	1., 2., 3.	hour_20	Indicator for 20:00	1., 2., 3., 4., 5.
temperature_diff_t1	Temperature station diff. at t-1	1.	hour_21	Indicator for 21:00	1., 2., 3., 4., 5.
temperature_diff_t24	Temperature station diff. at t-24	1.	hour_22	Indicator for 22:00	1., 2., 3., 4., 5.
temperature_diff_t48	Temperature station diff. at t-48	1.	hour_23	Indicator for 23:00	1., 2., 3., 4., 5.
temperature_diff_t72	Temperature station diff. at t-72	1.	wknd_or_h	Non-working day indicator	1., 2., 3., 4., 5.
temperature_diff_avg_7_days	Average temp. stat. diff. in 7 days	1.	hour_01.wknd_or_h	01:00 on a non-working day	1., 2., 3., 4., 5.
temperature_diff_max_24_hrs	Max temp. stat. diff. in 24 hours	1.	hour_02.wknd_or_h	02:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_t1	Visibility at t-1	1.	hour_03.wknd_or_h	03:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_t24	Visibility at t-24	1.	hour_04.wknd_or_h	04:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_t48	Visibility at t-48	1.	hour_05.wknd_or_h	05:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_t72	Visibility at t-72	1.	hour_06.wknd_or_h	06:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_avg_7_days	Average visibility in 7 days	1.	hour_07.wknd_or_h	07:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_max_24_hrs	Max visibility in 24 hours	1.	hour_08.wknd_or_h	08:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_diff_t1	Visibility station diff. at t-1	1.	hour_09.wknd_or_h	09:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_diff_t24	Visibility station diff. at t-24	1.	hour_10.wknd_or_h	10:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_diff_t48	Visibility station diff. at t-48	1.	hour_11.wknd_or_h	11:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_diff_t72	Visibility station diff. at t-72	1.	hour_12.wknd_or_h	12:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_diff_avg_7_days	Average vis. stat. diff. in 7 days	1.	hour_13.wknd_or_h	13:00 on a non-working day	1., 2., 3., 4., 5.
visibility_distance_diff_max_24_hrs	Max vis. stat. diff. in 24 hours	1.	hour_14.wknd_or_h	14:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_t1	Wind speed at t-1	1.	hour_15.wknd_or_h	15:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_t24	Wind speed at t-24	1.	hour_16.wknd_or_h	16:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_t48	Wind speed at t-48	1.	hour_17.wknd_or_h	17:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_t72	Wind speed at t-72	1.	hour_18.wknd_or_h	18:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_avg_7_days	Average wind speed in 7 days	1.	hour_19.wknd_or_h	19:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_max_24_hrs	Max wind speed in 24 hours	1.	hour_20.wknd_or_h	20:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_diff_t1	Wind speed station diff. at t-1	1.	hour_21.wknd_or_h	21:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_diff_t24	Wind speed station diff. at t-24	1.	hour_22.wknd_or_h	22:00 on a non-working day	1., 2., 3., 4., 5.
wind_speed_diff_t48	Wind speed station diff. at t-48	1.	hour_23.wknd_or_h	23:00 on a non-working day	1., 2., 3., 4., 5.

Note: 1. Regression trees (1h), 2. Neural network (1h), 3. Neural network (48h), 4. SARIMAX (1h), 5. SARIMAX (48h)

Table B.10: Coefficient estimates of utilized SARIMAX models

SARIMAX$(1,0,1)(1,1,1)_{24}$			SARIMAX$(2,0,2)(1,0,1)_{24}$		
Coefficient	Estimate	Std. Error	Coefficient	Estimate	Std. Error
ar1	0.9816	0.00136	ar1	1.97464	0.00226
ma1	0.29207	0.00621	ar2	-0.97576	0.00222
sar1	0.13722	0.00877	ma1	-0.70651	0.00666
sma1	-0.79172	0.00482	ma2	-0.2573	0.00649
mon_Feb	-14.53115	29.91639	sar1	0.98588	0.00125
mon_Mar	-46.91535	40.90776	sma1	-0.72442	0.00529
mon_Apr	-86.12984	48.76624	intercept	7869.23319	239.53275
mon_May	-43.9137	54.07169	mon_Feb	-210.1423	30.65739
mon_Jun	-80.05926	59.18844	mon_Mar	-594.62601	42.2364
mon_Jul	-90.75683	61.5647	mon_Apr	-907.88537	50.56949
mon_Aug	-112.87219	61.86111	mon_May	-1195.24756	56.56638
mon_Sep	-156.15876	59.94916	mon_Jun	-1476.91027	62.29227
mon_Oct	-159.34261	55.579	mon_Jul	-1429.32722	64.55802
mon_Nov	-99.96327	48.30778	mon_Aug	-1349.05973	64.69478
mon_Dec	-91.37257	36.01687	mon_Sep	-1135.49513	62.21605
day_Mon	-3.15634	6.15528	mon_Oct	-822.44967	57.22503
day_Thu	10.85688	4.25464	mon_Nov	-533.59467	49.57709
day_Tue	5.30801	6.28638	mon_Dec	-156.31173	36.86929
day_Wed	15.86034	5.76355	day_Mon	-5.69198	6.64457
covid_dummy	9.89966	37.93814	day_Thu	-3.82966	4.42493
hour_01	72.2208	22.59185	day_Tue	-2.06219	6.49076
hour_02	13.20509	35.93579	day_Wed	4.65097	5.84415
hour_03	-73.29973	48.44715	covid_dummy	-120.94652	38.79617
hour_04	59.24605	64.52324	hour_01	56.65308	18.26381
hour_05	525.81975	78.47194	hour_02	0.96685	28.68193
hour_06	1482.08728	92.18742	hour_03	-77.22945	37.55772
hour_07	1949.01311	104.33286	hour_04	67.11282	48.17774
hour_08	2105.92383	114.99423	hour_05	537.22603	56.61774
hour_09	2207.1684	123.75633	hour_06	1482.41327	64.18285
hour_10	2157.49149	130.51679	hour_07	1951.9727	70.57672
hour_11	2148.08557	135.0981	hour_08	2134.67084	75.89824
hour_12	2266.58891	137.40265	hour_09	2279.46805	80.04869
hour_13	2148.91018	137.37309	hour_10	2278.05856	83.09656
hour_14	1893.88227	134.97634	hour_11	2321.21649	85.04645
hour_15	1836.11702	130.2044	hour_12	2464.97367	85.93511
hour_16	1694.95761	123.10496	hour_13	2367.20449	85.66136
hour_17	1651.72666	113.74489	hour_14	2131.43585	84.29881
hour_18	1591.44757	102.25132	hour_15	2102.91036	81.86197
hour_19	1799.97268	88.75227	hour_16	1987.30415	78.31976
hour_20	1610.31506	73.52462	hour_17	1930.22171	73.553
hour_21	1114.43474	56.62065	hour_18	1839.09151	67.50338
hour_22	712.18019	38.888	hour_19	2002.88041	60.1998
hour_23	322.48813	20.23573	hour_20	1757.23434	51.48703
wknd_or_h	-177.69826	4.68995	hour_21	1219.53327	41.20851
hour_01.wknd_or_h	-155.97375	4.25097	hour_22	781.09833	29.58879
hour_02.wknd_or_h	8.61972	6.63892	hour_23	336.86742	16.64512
hour_03.wknd_or_h	79.93399	8.2075	wknd_or_h	-166.87946	4.65089
hour_04.wknd_or_h	-28.84806	9.40411	hour_01.wknd_or_h	-146.05144	4.11637
hour_05.wknd_or_h	-501.81407	10.35838	hour_02.wknd_or_h	26.93947	6.42753
hour_06.wknd_or_h	-1330.08827	11.12514	hour_03.wknd_or_h	104.22893	8.00204
hour_07.wknd_or_h	-1497.12055	11.75103	hour_04.wknd_or_h	-0.85548	9.23552
hour_08.wknd_or_h	-1332.13097	12.25778	hour_05.wknd_or_h	-470.63372	10.23429
hour_09.wknd_or_h	-1064.63637	12.63922	hour_06.wknd_or_h	-1295.62085	11.0492
hour_10.wknd_or_h	-734.71323	12.91162	hour_07.wknd_or_h	-1458.53444	11.7114
hour_11.wknd_or_h	-652.47792	13.08355	hour_08.wknd_or_h	-1288.63659	12.23727
hour_12.wknd_or_h	-885.38552	13.14761	hour_09.wknd_or_h	-1017.62047	12.63921
hour_13.wknd_or_h	-916.13695	13.12275	hour_10.wknd_or_h	-684.0472	12.92466
hour_14.wknd_or_h	-845.49564	13.00069	hour_11.wknd_or_h	-598.65956	13.09897
hour_15.wknd_or_h	-857.92584	12.7889	hour_12.wknd_or_h	-830.06437	13.16621
hour_16.wknd_or_h	-798.70977	12.47532	hour_13.wknd_or_h	-859.37479	13.12976
hour_17.wknd_or_h	-822.4925	12.05394	hour_14.wknd_or_h	-790.18353	12.99151
hour_18.wknd_or_h	-620.38341	11.51246	hour_15.wknd_or_h	-806.04282	12.74982
hour_19.wknd_or_h	-624.72917	10.83627	hour_16.wknd_or_h	-752.1147	12.40136
hour_20.wknd_or_h	-467.68507	10.00511	hour_17.wknd_or_h	-782.13449	11.94033
hour_21.wknd_or_h	-296.20895	8.96994	hour_18.wknd_or_h	-586.84646	11.36079
hour_22.wknd_or_h	-18.183	7.65279	hour_19.wknd_or_h	-599.48927	10.65069
hour_23.wknd_or_h	87.33024	5.84776	hour_20.wknd_or_h	-450.81702	9.79057
-	-	-	hour_21.wknd_or_h	-285.45787	8.74897
-	-	-	hour_22.wknd_or_h	-16.39575	7.47063
-	-	-	hour_23.wknd_or_h	84.24174	5.81483
In-sample set: Jan 2018 to May 2020 ($n = 21162$)			In-sample set: Jan 2018 to May 2020 ($n = 21162$)		
In-sample RMSE: 60.06			In-sample RMSE: 60.92		