Studies on Modeling Approaches for Analyzing Factors of Deviations Between Electricity Demand and its Procurement Plan

電力需要と調達計画の乖離に影響を与える因子解析のための モデリング手法に関する研究

February, 2023

Nanae KANEKO
金子　奈々恵

Studies on Modeling Approaches for Analyzing Factors of Deviations Between Electricity Demand and its Procurement Plan

電力需要と調達計画の乖離に影響を与える因子解析のための モデリング手法に関する研究

February, 2023

Waseda University Graduate School of Advanced Science and Engineering

Department of Advanced Science and Engineering, Research on Electrical Engineering and Bioscience A

Nanae KANEKO
金子　奈々恵

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction
1.1 Background of recent electricity demand
1.1.1 Usage of electricity demand estimations in power system planning and operations

The long- and medium-term planning of the electric power procurement based on the modeling of electricity demand, namely, the difference between total demand and aggregated supply of distributed variable renewable energy sources, is crucial for electric utility operators to achieve a stable electricity supplement, such as developing power sources and power grids, and securing reserve capacity [1-1]. The electricity demand modeling that parametrically specify the relationship between the electricity demand and factors estimates the behavior of demand as a basis for deciding the power procurement plans; the modeled electricity demand are used to allocate electricity. For example, power system operators, such as the Organization for Cross-regional Coordination of Transmission Operators (OCCTO) in Japan, constantly monitor demand, and they estimate the maximum demand over the next 10 years based on demand modeling using economic indicators such as gross domestic product (GDP). This is utilized in planning the development of power plants. In addition, such system operators estimate the demand one year or several days ahead based on the economic indicators and weather condition; it is used to evaluate the power system supply capacity to ensure a stable supply power with respect to long- and medium-term demand changes. The required capacity margin is between 8%–10% with respect to the maximum electricity demand in case of severe weather conditions [1-2]. Meanwhile, with the recent large-scale penetration of renewable energy sources, promotion of electricity conservation by electricity utilities and the deregulation of the electricity market, the change in the electricity demand and generation are uncertain and difficult to forecast. In addition, the trend in the electricity demand has been changing owing to the various factor such as extreme weather, economic level, and recent developments in society; especially, the consumer behavior of electricity usage and conservation are affected by the Great East Japan Earthquake of 2011 and the global COVID-19 pandemic [1-3, 1-4]. Accordingly, Fig. 1-1 shows the trend of maximum electricity demand in Japan between 1980–2022. The figure indicates that the electricity demand had increased monotonically until 2001; in contrast, the demand had decrease or remained approximately constant since around 2010, according to the economic shocks and natural disasters. This approximately constant trend has been especially observed in various regions owing to the growing worldwide interest in energy conservation and renewable energy. Under this situation, the demand changes that cannot be estimated using economic indicators and weather information, which have traditionally been focused on, are an open problem for system operators.

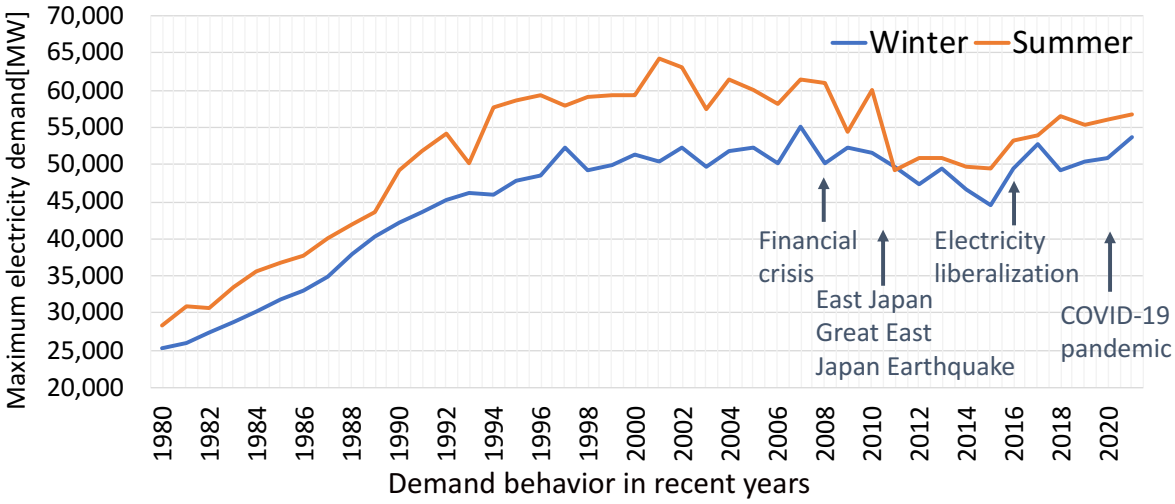


Fig. 1-1 Maximum electricity demand between 1980–2022.

1.1.2 Importance of hourly electricity demand estimations

Since the estimation of maximum electricity demand is becoming increasingly difficult, the hourly demand estimation is of interest to system operators. Recent changes in the power/social system have affected the load curve of electricity demand; the behavior of the demand time series has changed. The refinement of demand modeling by estimating hourly electricity demand is discussed as an attractive idea to improve the accuracy of demand estimation. In addition, the estimated hourly demand estimation can be used to assess fuel availability planning and evaluation. Although conventional maximum power demand estimations represent only the power capacity [W] required at a given moment in time, the area of the hourly demand curve represents the amount of electricity [Wh] required in a given period of time. A framework for constantly monitoring and evaluating hourly demand is being introduced to assess the risk to electricity supply in the event of power generation curtailment during fuel procurement price increases or power plant problems.

1.2 Model-based analysis of the factors affecting the recent changes in electricity demand

1.2.1 Current procedures of the demand modeling

Traditionally, electric utility providers have analyzed the electricity demand based on the electricity demand modeling framework and constructed appropriate demand models reflecting the results of such analysis; they have been modeling the electricity demand based on macro-frames and statistical modeling approaches using the few factors that have been screened according to the prior knowledge of experts, namely, weather information and key industry economic indicators, such as gross domestic product (GDP) and stock prices [1-5, 1-6]. Figure 1-2 shows an overview of the conventional demand modeling procedure conducted by electricity utility providers [1-7]. In the conventional procedure, the electric utility capacity that determines the electricity procurement plan describes the statistical structure between the demand and factors based on modeling, which focuses on certain factors. For example, the electricity demand is estimated with respect to the component of base demand estimated based on historical demand trends and long-term economic growth, and the component of the seasonal changes derived based on parameters in demand modeling. Note that the one-to-one correspondence scheme regarding the impact of changing each factor on electricity demand is crucial in the conventional demand modeling procedures. This parameter-based description enables utility providers to attain an evidence-based estimation of reasonable demand considering various scenarios with respect to each factor. For instance, providers realize the reasonable demand assumed under various situations where the demand could change significantly, such as extreme heat, while quantitatively identifying changes in demand in response to temperature changes, based on certain parameters. On the other hand, recently, as social and power systems have become more complex, the correlation between the factors and demand has declined significantly [1-3]; the description of the electricity demand is difficult using few factors, which have been used ad hoc base on the prior knowledge. This prominent issue causes significant deviations between actual electricity demand and its procumbent plan based on the modeled demand, i.e. *demand deviation*, as depicted in Fig. 1-3 (a) ; it may lead a serious risks in the stable electricity supplements.

In recent years, numerous kinds of variable have been observed and available as open data in various fields such as finance and industry. Under these situation, the usage of the large number of available observed variables has been expected for demand modeling. It has been discussed by both domestic and foreign electric utility providers, to find important factors that can describe the demand, which changes according to such as recent changes in consumer habit and economic levels, and the energy conservation trends, and to fix conventional modeling procedures, as depicted in Fig. 1-3 (b). However, the relationships between demand and the large number of variables are opaque and uncertain, and discussing the relationships among these variables is still scarce stage; the development of approaches for analyzing the relationships among variables are important to identify the important factors.

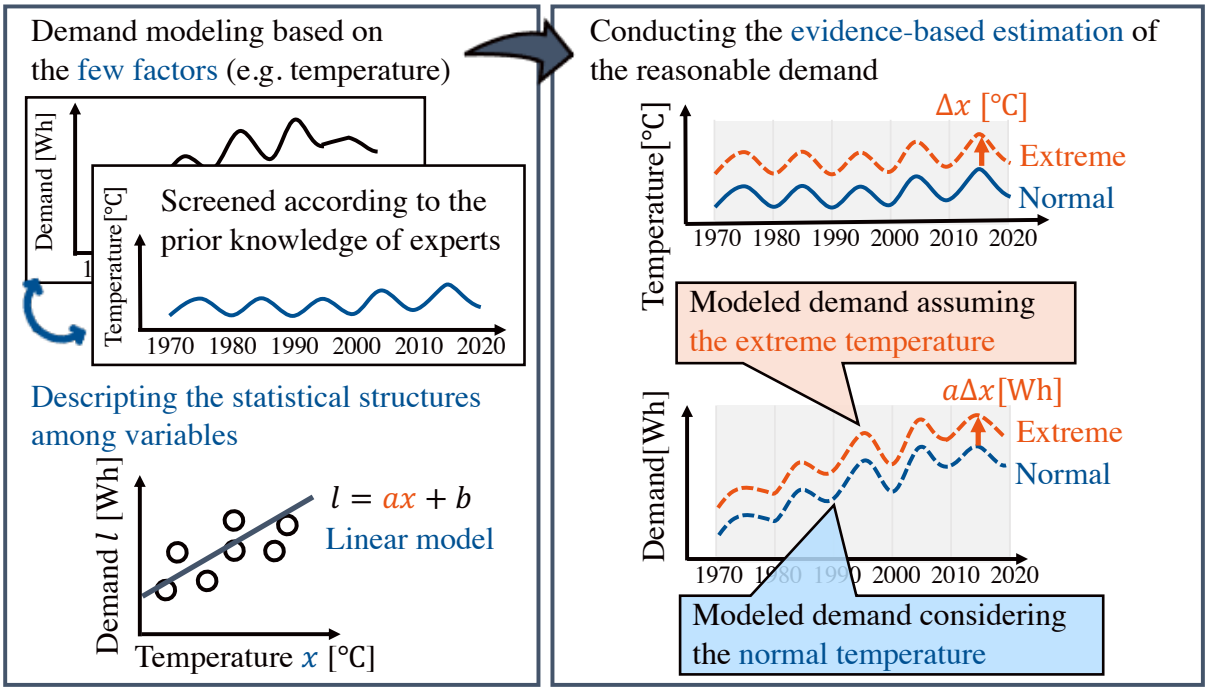
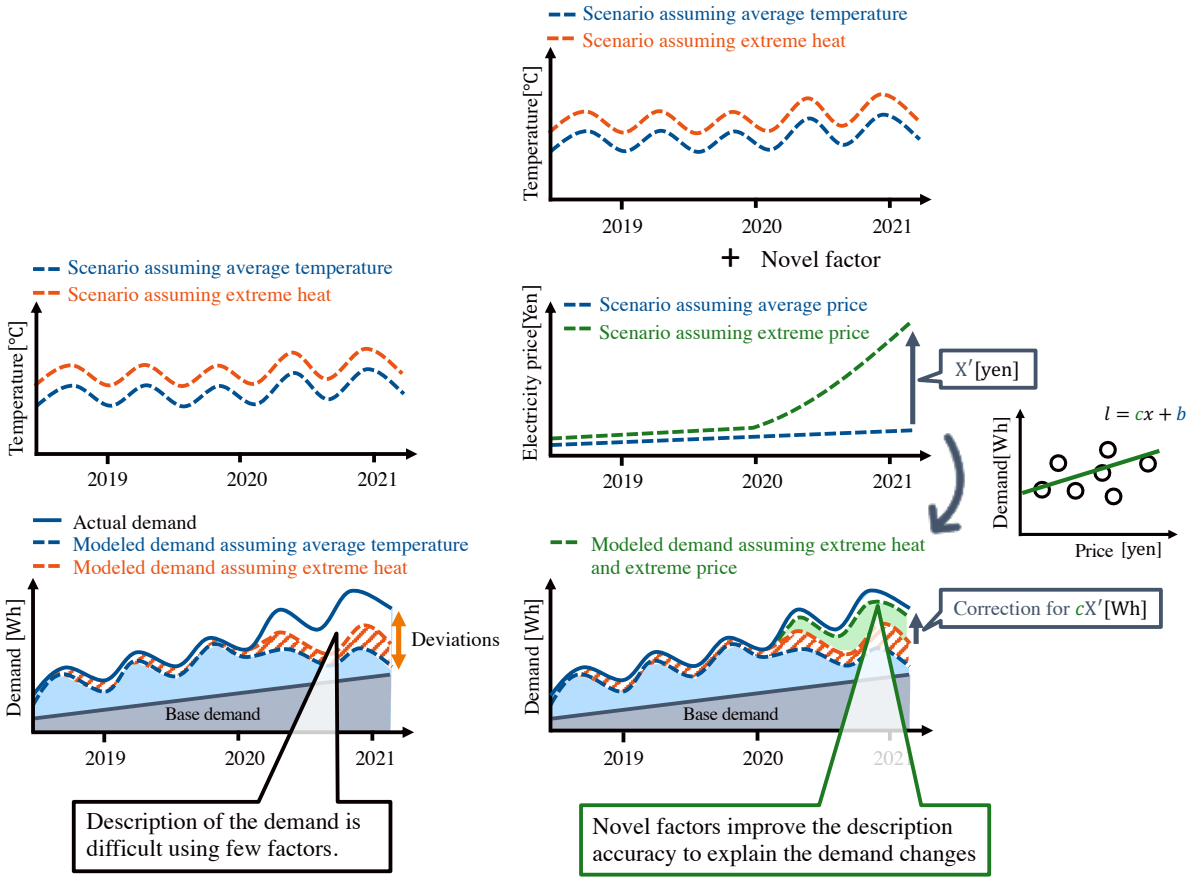


Fig. 1-2 Overview of the conventional demand modeling procedure.



(a) Before (b) After

Fig. 1-3 Main problem in the current demand modeling procedure.

1.2.2 Difficulties of the analysis of factors affecting the deviations between the actual demand and its procurement plan

This thesis addresses the analysis of factors affecting the demand deviation from three perspectives: factors affecting the behavior of electricity demand, factors of deviations between the actual demand and its estimation in long- and medium-term demand modeling, and factors of occurrence of imbalance, i.e., the deviations between the actual demand and its procurement plan in electricity market.

Firstly, analyzing the factors affecting the behavior of electricity demand is a research topic that has traditionally been actively conducted using the demand modeling framework [1-8, 1-9, 1-10]. Several variables, such as GDP, stock prices, weather information, and major industry economic indicators which affect the electricity demand, such as the steel industry, have been mainly discussed in previous studies as factors for explaining electricity demand based on the assumption of linearity to the demand. Additionally, when modifying the demand modeling procedures to describe the recent electricity demand changes, it is important to identify additional factors useful in explaining demand among a large number of observed variable, considering the linearity/nonlinearity to the demand.

Second, analyzing factors of the deviations between the actual demand and its estimation in long- and medium-term demand modeling, and the components of the impact of each factor on deviations is important under the recent situations where the structure of the social system has become complex; because the deviation of procurement plan form the actual electricity has been increasing, the essential factors that explain this deviation have also attracted research interest [1-11, 1-12]. The electric utility providers and previous studies empirically determined that the demand deviations changes depending on various environmental factors that affect the propensity to consume electricity, such as economic conditions and consumer behavior habits. Meanwhile, no reasonable factors have been identified that statistically represent the deviations. Here, the electricity demand was modeled based on long- and medium-term scenarios of important factors, where the demand deviations consist of two components with different characteristics: (a) deviation caused by the factors that behave differently from the scenario and (b) deviation caused by factors that change according to the scenario. Especially, demand modeling appropriate to the situation is particularly important in situations where the power and social system may change rapidly, taking into account the structure of the current and future demand deviations.

Third, the analyzing factors of occurrence of imbalance in electricity market is conducted. In a deregulated electricity market, the amount of electricity procured by the entire grid is determined based on the demand forecasted by each retail utility. Electricity deregulation has been adopted recently; however, it has not been extensively discussed in the literature. The existing literature suggests that grid procurement and electricity demand, especially in markets with numerous actors, are highly uncertain [1-13, 1-14]. Moreover, the deviations between procurement predictions and actual demand, namely, the *imbalance*, could be extremely large, depending on the weather conditions and small changes in the decision-making processes of retail operators. In such situations, assessing and ensuring the necessary reserve capacity for any large imbalances that may occur in the future power is important in system management based on a power system planner [1-15]. This technology is expected to clarify the relationship between the occurrence of imbalance and main factors using statistical models; research on building statistical models that explain imbalances is extremely scarce. Based on the demand

modeling framework, electric utility providers need to analyze the sensitivity of changes in each important factor to extremely large imbalances to develop stable grid planning.

Several approaches have been proposed for demand modeling. In traditional long- and medium-term demand modeling, which are approaches based on constructing linear models, focusing mainly on a limited number of variables that have been emphasized according to the prior knowledge of economic experts, such as annual and monthly data, as factors explaining demand variation, has been widely utilized considering the interpretability of results [1-5, 1-6].

On the other hand, to cope with demand modeling whose structure has changed recently, machine learning approaches that systematically handle linearity/nonlinearity among variables while considering numerous observed variables as candidate factors are also expected to improve the analysis accuracy. Furthermore, the characteristics of changes in electricity demand vary with seasonal conditions, such as month and daytime. The situation-dependent modeling framework identifies the factors that effectively describe demand characteristics whose behavior varies depending on seasonal conditions.

The first technical difficulty in demand modeling when analyzing the factors that explain the demand is selecting the necessary factors to describe demand changes while appropriately identifying linearity/nonlinearity among the variables for numerous observed variables. When focusing on several variables, identifying candidate factors, including redundant variables for explaining the demand, and selecting important variables while excluding redundant variables in demand modeling, is important to improve model interpretability and descriptive accuracy. Furthermore, the frameworks that can deal with the highly nonlinear relationships among variables unnecessarily complicate data-centric decision-making.

The second difficulty is to minimize the number of factors needed to describe seasonal variations when constructing seasonal demand models in situation-dependent modeling. While modeling methods based on machine learning are usually highly accurate in terms of the explanatory power of the constructed model, a small difference in the conditions under which the sample is collected can significantly alter the factor analysis results [1-16]. Owing to this characteristic, when building seasonal models using data collected at different times of the day, the variables selected may be entirely different and the number of important factors that describe the seasonal variations may be excessively large. This worsens the interpretability of the assumed results and unnecessarily complicates the variable observation and collection tasks conducted by the electric utility providers. Accordingly, it is important to identify the essential variables such that the number of factors is minimized.

This study proposes several approaches to analyze the factors crucial for describing the electricity demand while considering these technical difficulties.

1.3 Contributions

The objective of this research is to proposed approaches for identifying the important variables to describe the deviation between the long- and medium-term demand deviations. Furthermore, we quantitatively evaluate the validity of each proposed method based on numerical experiments.

(i) Proposal of a sparse modeling approach for identifying the dominant factors affecting situation-dependent hourly electricity demand

In this chapter, an approach is proposed to analyze the factors that affect the dynamic characteristics of the hourly electricity demand. Conventionally, such demand analysis has been conducted by targeting a limited number of explanatory variables that have been screened according to the prior knowledge of experts. The identification of essential explanatory variables through data-centric analysis, with a focus on variables that co-occur with demand, has long been recognized as important; however, discussion has been limited because it is difficult to describe plausible statistical relationships among the many possible explanatory variables with only a limited number of historical data samples available. This study focuses on the dynamics in hourly electricity demand and approach to identify annually important variables by constructing situation-dependent models. These models are based on the dataset consisting of demand and multiple explanatory variables that co-occur in the target time slices. The class of Partially Linear Additive Model (PLAM) is an attractive framework in demand modeling; PLAM is a white-box model that is expected to ensure the interpretability as linear model that can additively describe the effects of individual variables, and also improve the descriptive accuracy by partially introducing nonlinearity among the factors as one of the nonlinear models. Furthermore, the powerful concept of sparse modeling is applied to handle the large number of possible explanatory variables used in the situation-dependent modeling process. In particular, this study discusses inconsistency of the selected variables when the statistical models are constructed focusing on different data subsets; when the model is trained based on a dataset focusing on a specific time period, the selected variables may be significantly different from those resulting from a dataset focusing on another time period. We propose to derive a limited number of annually dominant variables by enumerating suboptimal models for each situation, and by selecting, as much as possible, essential variables that are commonly and consistently used for all situations. The proposed scheme was applied to a real-world demand dataset in Japan and discussed in the context of representation errors and interpretability. The results show that the proposed method is an effective approach for representing the situation-dependent impact of variables on demand.

(ii) Proposal of a model-based analysis approach for identifying impacts of dominant factors affecting the electricity deviations during COVID-19

In this chapter, focusing on the COVID-19 pandemic, the proposed sparse partially linear additive modeling framework is evaluated to verify that it adequately identifies the key factors affecting demand changes caused by the pandemic. An approach is proposed to identify the dominant factors that explain the demand deviation focusing on the electricity demand under the pandemic of COVID-19. In recent years, the economic conditions and consumer lifestyle have drastically changed under COVID-19 pandemic. Because of these environment changes, the behavior of electricity residual demand, which is defined as the difference between the total electricity consumption in domestic customers and the consumption from renewable energy sources, has changed drastically. Under these situation, the electricity the demand deviation which is derived before the pandemic has become seriously large; it has been a world-wise problem. The electricity

demand deviation changes depending on the several factors such as weather, economic condition and the consumers lifestyle. Analyzing the impact of important factors, which have changed their behavior significantly due to the pandemic, on the demand deviation enables stakeholders to construct a electricity business strategy. During the pandemic of COVID-19, several studies have been conducted to understand the characteristics of the recent electricity demand deviation. These studies suggested that the recent demand deviation has been related to several significant socially events such as the rapid spread of infection and lockdown. On the other hand, the discussion about the impact of specific factors such as the economic conditions and the consumer lifestyle on the demand has been limited because the impact of the pandemic on the economy and lifestyles has been wide-ranging, and it is difficult to analyze the plausible statistical relationships among the many possible explanatory variables. In this study, we focus on the hourly electricity demand and a large number of variables that co-occur in the target time slices, and propose an approach to identify the impact of important factors on the electricity demand deviation caused by the pandemic. The concept of the Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) model provides a scheme for identify the factors whose behavior has changed significantly during the pandemic. Additionally, PLAM describes the relationships between the target demand and explanatory variables while considering the linearity/nonlinearity among variables; the concept is an attractive approach to analyze the additive contribution of each explanatory variables for describing the electricity demand deviation. As whole proposed procedure, we achieve to identify the important variables that vary drastically during the pandemic and accordingly evaluate the impact of these variables on the electricity demand deviation. The proposed approach was applied to a real-world demand dataset collected in Germany. The results show that the proposed method is an effective approach for representing the impact of variables on the demand deviation.

(iii) Proposal of a sensitivity analysis approach for identifying the impact of factors on extreme imbalance between procurement plans and actual demand

In this chapter, an approach is proposed for analyzing the important factors affecting extremely large deviations of electricity procurement from actual demand in the electricity market. In a liberalized power system, operators control the power system to compensate for imbalances, which are the differences between the scheduled power procurement in electricity markets and the actual power supply. Interpreting the influence of factors that affect extremely large imbalances (i.e., extreme imbalance events) enables power system operators to implement appropriate system operation plans. Conventionally, such imbalance analysis has focused only on a limited number of factors and described variation in imbalance by utilizing highly interpretable statistical models, assuming that the probability of imbalance events varies monotonically with changes in those factors. In particular, sensitivity analysis of such models can be a powerful tool to support the decision-making of system operators. However, when dealing with the increasingly complex behavior of markets involving many actors, a flexible statistical analysis framework is required to identify informative factors among the various observable quantities, and to describe nonmonotonic relationships when essentially necessary. This study focuses on the statistical behavior of the odds ratio of the extreme imbalance events by concentrating on an inherently large number of explanatory variables. We propose a model-based approach using a class of partially

linear additive models and a variable selection method to reveal the statistical relationships between the relevant variables and extreme imbalance events. The framework further provides an analysis scheme to determine the response sensitivity of relevant variables to the odds ratio of extreme imbalance events. The usefulness of the framework was demonstrated by applying the approach to a real-world dataset collected in the Japanese electricity market system. The results show that the proposed approach based on partially linear additive models works well to describe the extreme imbalance events; the constructed models derive the interpretable sensitivity curves to clarify the impact of informative variables to extreme events, while identifying monotonicity/nonmonotonicity among variables.

1.4 Organizations of thesis

The rest of the thesis is organized as follows. Chapter 2 reviews relevant works on the long- and medium- term electricity demand modeling, and provide the technical difficulties to conduct the electricity demand modeling. Furthermore, the several key ideas for demand modeling are gives considering the technical difficulties. Chapter 3 provides an approach to identify the annually dominant variables in the situation-dependent modeling of the electricity demand based on the sparse PLAM technique. Chapter 4 provides an approach to identify the important variables to describe the demand during the pandemic, and the impact of important variables on the electricity deviation. Chapter 5 provides an approach to analyze the sensitivity of the influence of factors on extreme imbalance based on PLAMs. Finally, in Chapter 6, concluding remarks and the furfure business opportunity are provided.

Reference

[1-1] A. Singh, M. Pratap, A. Das, P. Sharma and K. Gupta “Regulatory Framework for Long-Term Demand Forecasting and Power Procurement Planning.” *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3440507.

[1-2] Ministry of Economy, Trade and Industry, "Approach to securing the regulatory power required for general transmission and distribution projects (Discussion at the System Design WG of the Advisory Committee for Natural Resources and Energy)(in Japanese)", 2015, [Online]. Available: https://www.emsc.meti.go.jp/activity/emsc_electricity/pdf/003_03_02.pdf. [Accessed 31 October 2022].

[1-3] K. Honjo, H. Shiraki and S. Ashina, “Dynamic Linear Modeling of Monthly Electricity Demand in Japan: Time Variation of Electricity Conservation effect.(Research Article).” *PLoS ONE*, vol. 13, no. 4, 2018, pp. 1-23 , doi:10.1371/journal.pone.0196331.

[1-4] L. Huang, Q. Liao, R. Qiu, Y. Liang, and Y. Long, “Prediction-based analysis on power consumption gap under long-term emergency: A case in China under COVID-19,” *Appl. Energy*, vol. 283, p. 116339, Feb. 2021, doi: 10.1016/J.APENERGY.2020.116339.

[1-5] R. Abdel-Aal and A. Al-Garni, “Forecasting Monthly Electric Energy Consumption in Eastern Saudi Arabia Using Univariate Time-Series Analysis.” Energy, vol. 22, no. 11, 1997, pp. 1059–69, doi:10.1016/S0360-5442(97)00032-7.

[1-6] H. Ching-Lai, S. Watson and S. Majithia, “Analyzing the Impact of Weather Variables on Monthly Electricity Demand.” IEEE Transactions on Power Systems, vol. 20, no. 4, 2005, pp. 2078–85, doi:10.1109/TPWRS.2005.857397.

[1-7] Organization for Cross-regional Coordination of Transmission Operators, “Demand Assumption Procedure (in Japanese)”, 2022, [Online]. https://www.occto.or.jp/juyousoutei/2022/files/220401_jyuyousouteiyouryou.pdf. [Accessed 31 October 2022].

[1-8] Z. Zhang, D. Gong and J. Ma, “A Study on the Electric Power Load of Beijing and Its Relationships with Meteorological Factors During Summer and Winter.” Meteorological Applications, vol. 21, no. 2, 2014, pp. 141–48, doi:10.1002/met.1313.

[1-9] M. Auffhammer, P. Baylis, C. Hausman and M. Auffhammer, “Climate Change Is Projected to Have Severe Impacts on the Frequency and Intensity of Peak Electricity Demand Across the United States.” Proc. the National Academy of Sciences of the United States of America, vol. 114, no. 8, 2017, pp. 1886–91, doi:10.1073/pnas.1613193114.

[1-10] D. Vu, K. Muttaqi and A. Agalgaonkar, “A Variance Inflation Factor and Backward Elimination Based Robust Regression Model for Forecasting Monthly Electricity Demand Using Climatic Variables.” Applied Energy, vol. 140, 2015, pp. 385–94, doi:10.1016/j.apenergy.2014.12.011.

[1-11] F. Alasali, K. Nusair, L. Alhmoud, and E. Zarour, “Impact of the COVID-19 Pandemic on Electricity Demand and Load Forecasting,” *Sustain. 2021, Vol. 13, Page 1435*, vol. 13, no. 3, p. 1435, Jan. 2021, doi: 10.3390/SU13031435.

[1-12] M. Sabbir Hossain, S. Ahmed, and M. Jamal Uddin, “Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model,” *Sci. Total Environ.*, vol. 761, p. 143315, 2021, doi: 10.1016/j.scitotenv.2020.143315.

[1-13] M. Bueno-Lorenzo, M. Á. Moreno, and J. Usaola, “Analysis of the imbalance price scheme in the Spanish electricity market: A wind power test case,” *Energy policy*, vol. 62, pp. 1010–1019, 2013, doi: 10.1016/j.enpol.2013.08.039.
[1-14] F. Lisi and E. Edoli, “Analyzing and Forecasting Zonal Imbalance Signs in the Italian Electricity Market,” *The Energy journal (Cambridge, Mass.)*, vol. 39, no. 5, p. 1, 2018, doi: 10.5547/01956574.39.5.flis.
[1-15] Ministry of Economy, Agency for Natural Resources and Energy, " Institutional Environment Improvement to Ensure an Efficient and Stable Balance of Electricity Supply (in Japanese)", 2018, [Online]. Available:https://www.meti.go.jp/shingikai/enecho/denryoku_gas/denryoku_gas/pdf/007_06_00.pdf [Accessed 31 October 2022].
[1-16] S. Hara and T. Maehara "Enumerate Lasso Solutions for Feature Selection." Proc. *AAAI*, 2017, pp. 1985–91.

Chapter 2
Approaches for electricity demand modeling analysis

2.1 Related studies for electricity demand modeling

Several researchers and real-world system operators have attempted to clarify the relevant factors influencing these demand variations based on the electricity demand modeling. Table 2-1 summarizes the previous works discussing these demand dynamics. In the previous studies focusing on the long-span temporal demand transition such as monthly demand and yearly demand, the effects of atmospheric trends, economic indicators such as gross domestic product (GDP), and public and national holidays have been thoroughly analyzed [2-1, 2-3]; these factors are mainly observed by day, month and year. On the other hand, in the studies focusing on the short-span seasonality in daily load curves, the effects of hourly variable such as weather condition were also primarily emphasized [2-8, 2-9]. These studies suggest that the dominant factors influencing the temporal change in electric power demand may differ depending on the situation and target dynamics.

Analyzations based on a limited number of variables, derived from the empirical prior knowledge of experts, is convenient for quickly interpreting these target demand effects. However, in a situation where the power demand trend is changing and the factors affecting power systems are becoming increasingly complex, it may be inappropriate to explain demand variations using only limited variables. For instance, reference [2-10] has suggested that energy conservation would depend on various types of technical, demographic, climatic, economic, and psychological factors. Further, reference [2-7] focused on residential energy conservation after the Great East Japan Earthquake and determined that peak demand was reduced by a factor corresponding to electricity prices. Reference [2-12] also found that the electricity demand under the COVID-19 pandemic changes depending on the domestic infection levels.

Previous studies have also discussed assumptions regarding the description of relationships between target demand and possible influential factors. Al-Garni et al. [2-1] assumed linear relationships between variables. Although this assumption is so simple that it may not generalize to any given dataset, statistical models based on this linearity assumption have been widely utilized for analyzing the main effects of each variable on demand. In recent years, there have been works explicitly describing nonlinearity in the relationships among variables; e.g. Huang et al. [2-8] proposed a nonlinear modeling approach for describing target demand.

In this study, we propose an analyzation approach for identifying the dominant influential factors which affect the demand deviations, considering various situations characterized by the seasonality, while also considering the linearity and nonlinearity in the variable relationships. To account for the diversity of all possible factors, we focus on hundreds of explanatory variables, which can be categorized into attribute groups such as weather, stock price, calendar, industry activity index, and GDP data. The relationship between electricity demand and these variables has not been thoroughly discussed in the prior electricity demand studies, so this work fills a necessary gap and improves the generalizability and applicability of such demand studies.

Table 2-1 Representation of existing studies discussing demand variations.

Study	Target dynamics	Factors	Linear/ Nonlinear
Al-Garni et al. (1994) [2-1]	Monthly electricity consumption	Weather, population	Linear
Abdel-Aal et al. (1997) [2-2]	Monthly domestic electric energy consumption	Weather, GDP, population	Linear
Ching-Lai et al. (2005) [2-3]	Average monthly demand	Weather, GDP	Linear
Zhang et al. (2014) [2-4], Auffhammer et al. (2017) [2-5]	Peak daily demand	Weather	Linear
Vu et al. (2015) [2-6]	Peak monthly demand	Weather	Linear
Matsukawa (2016) [2-7]	Daily peak demand	Weather, electricity prices	Linear/ Nonlinear
Huang et al. (2016) [2-8], Kumar et al. (2018) [2-9]	Hourly demand	Weather	Nonlinear
Honjo et al. (2018) [2-10]	Monthly peak demand	Weather, the economic production index, GDP, electricity price indices, real wage index, population	Linear
Zhang et al. (2019) [2-11]	Peak daily demand	Weather, Income	Nonlinear

2.2 Several approaches of electricity demand modeling

In this chapter, several promising approaches are provided to identify the important factors in electricity demand modeling.

2.2.1 Partially linear additive model

Utilizing linear models in power demand prediction has been well-studied. For example, Ahmed et al. [2-13] have applied a linear model to modeling the electric power demand and verified its effectiveness. The explanatory variables used in their research were annual and monthly data, such as population, monthly mean of temperature, humidity, and solar radiation. Although the assumption of linearity is considerably simple, linear models have been widely used for analyzing the main effect of each variable on the actual demand, as well as modeling long- and medium-term power demand. However, these linear models may be insufficient to describe the recent complex changes that have occurred in the electricity demand.

Nonlinear models have also been used in the demand forecast to achieve an accurate result by focusing on nonlinear main effects and nonlinear interactions among variables. Neural networks [2-14] and random forests [2-15] are popular nonlinear models that

have been studied in the demand forecast field. Although nonlinear models have been adopted to derive accurate forecast results by focusing on complicated interactions among the explanatory variables, the derived model containing complex interactions among variables is unsuitable for interpreting the individual variable contribution. A class of nonlinear models called the partially linear additive model [2-16] is a promising class that considers the advantages of both linear and nonlinear models. This model class explicitly but partially describes the nonlinear main effect of the explanatory variables, which essentially corresponds to the nonlinearity in the demand, while excluding the complicated effect of the interactions among variables. The linear regression models, which are widely used in the existing electricity modeling schemes, belong to this class that describes the linear relationships between the target demand and all factors [2-1]. The additive regression model is another model in this class, which can describe more complex relationships between variables and demand by assuming that all variables are nonlinear. To adequately describe the PLAM model, appropriately selecting the important variables used in the modeling and assuming the linearity/nonlinearity of each variable are crucial. Several approaches have been proposed for identifying such relationships among variables in statistical modeling. For example, a variable selection approach based on the wrapper method [2-17] is commonly used when performing multiple regression modeling. This method iteratively tests several subsets of explanatory variables and selects the appropriate variable subset according to modeling accuracy. Furthermore, the stepwise forward/backward selection approach [2-18] has been frequently adopted for regression model-based analysis [2-19]. Notably, within this framework, it is possible to identify whether the dominant variables have a linear or nonlinear relationship with demand by running the selection process among the explanatory variable candidates, which consists of both original and nonlinearly transformed variables [2-20]. Given the several explanatory variables, one difficulty with this approach is a combinatorial explosion of the possible variable subsets considered during evaluation. Particularly, when we focus on simultaneously identifying the linearity or nonlinearity between the variables, the number of models, which are composed of possible combinations of candidate variables, is so enormous that identifying dominant variables by deriving appropriate variable subsets in this manner becomes a difficult task. In addition, the multiple regression analysis does not accurately assess the importance of each variable when many explanatory variables, relative to the number of samples, are considered.

2.2.2 Sparse modeling for identification of dominant variables

Sparse partially liner additive model

The embedded method [2-21] selects important variables while simultaneously constructing the models. The sparse regression models [2-22], which are typically based on the least absolute shrinkage and selection operator (LASSO) [2-23], are categorized within this method class. Such methods have attracted significant attention in the machine learning domain for their ability to select a limited number of informative variables from numerous candidates to determine a target value. This approach performs well in general, even for ill-defined problems using a limited number of samples with numerous variables [2-24, 2-25]. Sparse PLAM is a framework that extends the variable selection scheme adopted in LASSO; it is a promising application of sparse modeling techniques for selecting a limited number of informative variables among numerous candidate variables while establishing plausible linear or nonlinear relationships.

Enumerate sparse partially liner additive model

The sparse modeling technique adopted in the sparse PLAM is flexible and computationally efficient when selecting informative variables. Moreover, it identifies linearity or nonlinearity in the relationships between explanatory variables and target demand. Meanwhile, the selected variables may change considerably when we construct seasonal models using data collected over different time periods. This implies that all selected variables in the models are necessary to describe annual demand behavior, and the number of important variables for seasonal variation expression can be excessively large. Accordingly, the interpretation of the derived factors gets unnecessarily complex.

The scheme based on the PLAM is enumerated to construct situation-dependent models using a limited number of variables that are commonly and consistently used, as much as possible. This method improves the interpretability of the annually dominant variables used to describe the electricity demand behavior. In practice, when we construct a sparse regression model using numerous explanatory variables, some variables might make similar contributions to the model. In this case, the variables selected for explaining the seasonal variation can be replaced by other variables while preserving the same description accuracy. The enumerated scheme does not focus only on a single global optimal model under each situation but also enumerates a set of plausible models consisting of nearly optimal models. The variables used only in rare seasonal situations tend to be replaced by substitutable variables used in other models; accordingly, the variables used in models tend to overlap.

2.2.3 Relevant-redundant measure based variable selection approach

The filter method based on the relevant-redundancy measure is another attractive approach for selecting the informative variables while identifying the linearity/nonlinearity among variables. The method selects important variables based on scores assigned with respect to various statistical tests conducted regarding the relevance between the target demand and explanatory variables. Note that the variable selection based on the sparse modeling assumes nonlinearities for all explanatory variables before model construction; consequently, it is a difficult task to decide the proper nonlinearity given the numerous explanatory variables. The variable selection based on the filter method can be performed as a preprocessing step before model construction; it is implemented to refine the informative variables from numerous variables. In such a selection method, the parameters used to describe the nonlinearity among variables are selected more flexibly and efficiently than sparse modeling. In addition, the linearity and nonlinearity among variables can be identified using an appropriate framework for comparing the difference between the scores that evaluate the nonlinear and linear relevance of variables with respect to electricity demand.

Reference

[2-1] A. Al-Garni, S. Zubair and J. Nizami, "A Regression Model for Electric-Energy-Consumption Forecasting in Eastern Saudi Arabia." *Energy*, vol. 19, no. 10, 1994, pp. 1043–49, doi:10.1016/0360-5442(94)90092-2.

[2-2] R. Abdel-Aal and A. Al-Garni, "Forecasting Monthly Electric Energy Consumption in Eastern Saudi Arabia Using Univariate Time-Series Analysis." Energy, vol. 22, no. 11, 1997, pp. 1059–69, doi:10.1016/S0360-5442(97)00032-7.

[2-3] H. Ching-Lai, S. Watson, and S. Majithia, "Analyzing the Impact of Weather Variables on Monthly Electricity Demand." IEEE Transactions on Power Systems, vol. 20, no. 4, 2005, pp. 2078–85, doi:10.1109/TPWRS.2005.857397.

[2-4] Z. Zhang, D. Gong and J. Ma, "A Study on the Electric Power Load of Beijing and Its Relationships with Meteorological Factors During Summer and Winter." Meteorological Applications, vol. 21, no. 2, 2014, pp. 141–48, doi:10.1002/met.1313.

[2-5] M. Auffhammer, P. Baylis, C. Hausman and M. Auffhammer, "Climate Change Is Projected to Have Severe Impacts on the Frequency and Intensity of Peak Electricity Demand Across the United States." Proc. the National Academy of Sciences of the United States of America, vol. 114, no. 8, 2017, pp. 1886–91, doi:10.1073/pnas.1613193114.

[2-6] D. Vu, K. Muttaqi and A. Agalgaonkar, "A Variance Inflation Factor and Backward Elimination Based Robust Regression Model for Forecasting Monthly Electricity Demand Using Climatic Variables." Applied Energy, vol. 140, 2015, pp. 385–94, doi:10.1016/j.apenergy.2014.12.011.

[2-7] I. Matsukawa, "Consumer Energy Conservation Behavior After Fukushima." 1st ed., Springer Singapore, 2016, doi:10.1007/978-981-10-1097-2.

[2-8] N. Huang, G. Lu and D. Xu, "A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest." Energies, vol. 9, no. 10, 2016, p. 767, doi:10.3390/en9100767.

[2-9] V. Kumar and P. Dixit, "Artificial Neural Network Model for Hourly Peak Load Forecast." International Journal of Energy Economics and Policy, vol. 8, no. 5, 2018, pp. 155–60.

[2-10] K. Honjo, H. Shiraki, and S. Ashina, "Dynamic Linear Modeling of Monthly Electricity Demand in Japan: Time Variation of Electricity Conservation effect.(Research Article)." PLoS ONE, vol. 13, no. 4, 2018, pp. 1-23 , doi:10.1371/journal.pone.0196331.

[2-11] T. Fujimi and S. Chang, "Adaptation to Electricity Crisis: Businesses in the 2011 Great East Japan Triple Disaster." Energy Policy, vol. 68, 2014, pp. 447–57, doi:10.1016/j.enpol.2013.12.019.

[2-12] L. Huang, Q. Liao, R. Qiu, Y. Liang and Y. Long, "Prediction-based analysis on power consumption gap under long-term emergency: A case in China under COVID-19," *Appl. Energy*, vol. 283, p. 116339, Feb. 2021, doi: 10.1016/J.APENERGY.2020.116339.

[2-13] A. AI-Garni, Z. Syed and S and S. Nizami, "A Regression Model for Electric-Energy-Consumption Forecasting in Eastern Saudi Arabia", *Energy*, vol. 19, no. 10, 1994,pp. 1043-1049

[2-14] K. Murat and U. Ergun "Neural Network Approach with Teaching–Learning-Based Optimization for Modeling and Forecasting Long-Term Electric Energy Demand in Turkey", *Neural Computing and Applications*, vol 28, 2017, pp. 737-747

[2-15] W. Zhi-jun “Electric Power Forecasting in Inner Mongolia by Random Forest”, *2012 3rd International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012)*, Chapter 66, 2013, pp. 665-674
[2-16] L. Yin, B. Jacob, C. Rich and G. Johannes, “Sparse Partially Linear Additive Models”, *Journal of Computational and Graphical Statistics*, vol. 25, no.4, 2016, pp.1126-1140
[2-17] R. Kohavi and H. George, "Wrappers for feature subset selection." *Artificial intelligence,* vol. 97, no.1-2, 1997, pp. 273-324.
[2-18] J. Reunanen, “Overfitting in Making Comparisons Between Variable Selection Methods.” *Journal of Machine Learning Research*, vol. 3, 2003, pp. 1371–82.
[2-19] R. Miller, L. Golab and C. Rosenberg, “Modelling Weather Effects for Impact Analysis of Residential Time-of-Use Electricity Pricing.” Energy Policy, vol. 105, 2017, pp. 534–46, doi:10.1016/j.enpol.2017.03.015.
[2-20] Y. Fujimoto, S. Murakami, N. Kaneko, H. Fuchikami, T. Hattori and Y. Hayashi, “Machine Learning Approach for Graphical Model-Based Analysis of Energy-Aware Growth Control in Plant Factories.” IEEE Access, vol. 7, 2019, pp. 32183–96, doi:10.1109/ACCESS.2019.2903830.
[2-21] A. Teixeira, “Classification and regression tree.” Revue Des Maladies Respiratoires, vol. 21, no. 6, 2004, pp. 1174–76.
[2-22] I. Rish and G. Genady, “Sparse modeling: theory, algorithms, and applications. ” CRC press, 2014.
[2-23] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso.” Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, 1996, pp. 267–288, doi:10.1111/j.2517-6161.1996.tb02080.x.
[2-24] S. Caston, “Forecasting Medium-Term Electricity Demand in a South African Electric Power Supply System.” Journal of Energy in Southern Africa, vol. 28, no. 4, 2017, pp. 54–67, doi:10.17159/2413-3051/2017/v28i4a2428.
[2-25] G. Huebner, D. Shipworth, I. Hamilton, Z. Chalabi and T. Oreszczyn, “Understanding Electricity Consumption: A Comparative Contribution of Building Factors, Socio-Demographics, Appliances, Behaviours and Attitudes.” Applied Energy, vol. 177, 2016, pp. 692–702, doi:10.1016/j.apenergy.2016.04.075.

Chapter 3
Sparse modeling approach for identifying the dominant factors affecting the situation-dependent hourly electricity demand: A Japanese case study

Symbols

y	Index for indicating a specific year
m	Index for indicating a specific month
d	Index for indicating a specific day
h	Index for indicating a specific hour
l_{ymdh}	Hourly demand observed at hour h in day d in month m of year y
I	Number of explanatory variables
$\boldsymbol{x}_{ymdh}$	$= \left(x_{ymdh}^1, \dots, x_{ymdh}^I\right)$. Explanatory variables observed at hour h in day d in month m of year y
$\mathcal{L}$	$\subseteq \mathcal{S}$. Index subset indicating the variables related to the target demand linearly
$\mathcal{N}$	$\subseteq \mathcal{S}$. Index subset indicating the variables related to the target demand nonlinearly
$\boldsymbol{\beta}$	Set of coefficient parameters for explanatory variables
$\phi_j^k(.)$	Cubic spline transformation function
K	Number of bases in cubic spline transformation functions
$\boldsymbol{\tau}_j$	Vector of coefficient parameters for transformation functions $\phi_j^1\left(x^j\right), \dots, \phi_j^K\left(x^j\right)$
λ	Positive constant penalty for regularization
$\mathcal{S}$	$= \{1, \dots, P\}$. Index set of explanatory variables
$\mathcal{S}_{mh}^{(i)}$	Index subset of the selected variables corresponding to the non-zero components
$\boldsymbol{\mathcal{S}}_{mh}$	Set of variable index subsets enumerated in Algorithm 1
I_{mh}	Number of enumerated models under the given situation (m, h)
$s_{mh}^{j,(i_{mh})}$	Binary dummy variables introduced for a binary linear programming problem
$d_{mh}^{i_{mh}}$	Binary dummy variables introduced for a binary linear programming problem

Nomenclature

GDP	Gross domestic product
PLAMs	Partially linear additive models
LASSO	Least absolute shrinkage and selection operator
RMSE	Root mean squared errors

3.1 Introduction

Recently, electric power system structures have undergone changes because of the large-scale penetration of renewable energy sources and energy-saving trends. Although electricity demand is known to be affected by weather and economic factors [3-1], the recent rapid transition of power systems has broadened and diversified the factors that affect the dynamic characteristics of electricity net-loads, i.e. the difference between total demand and aggregate supply of distributed variable renewable energy sources [3-2, 3-3]. Especially in Japan, the demand behavior changed drastically after the Great East Japan Earthquake in 2011 [3-4, 3-5]. Analyzing the relationships between demand and observable variables that are considered to affect consumer behavior plays a key role for stakeholders to construct an electricity business strategy [3-6]

Various studies have been conducted on modeling to analyze the statistical relationships between variables that explain such demand [3-7, 3-8, 3-9]. For example, Ching-Lai et al. [3-7] focused on the monthly peak demand, which changes depending on influences such as weather and economic conditions, and discussed the factors that affect those dynamics. These early works suggest to us possible, and useful, analyzation approaches for describing the variations of the target demand characteristics. Most of these studies initially focused on a limited number of variables, e.g. meteorological observations and gross domestic product (GDP), to describe the demand variation. Additionally, they specifically focused on relatively large-scale dynamics, such as the peak and average load and its annual or monthly seasonality. Meanwhile, in recent years it has been recognized that demand can be influenced by additional factors including power-saving trends, electricity prices, and residential income [3-6, 3-1, 3-9]. Further, it has been proposed that factors affecting the hourly demand dynamically change depending on seasonal situations [3-10, 3-11]. Ultimately, these studies suggest that the existence and impact of various factors possibly affecting demand may change more than previously assumed.

In this study, we focused on the dynamics of hourly demand targeting, with relatively higher time resolution than most of previous works. We propose an approach for constructing statistical models that utilizes the co-occurrence dataset consisting of hourly demand and a large number of explanatory variables corresponding to the target time period. The proposed approach provides a scheme for identifying the important variables for each time period by comparing the relative contribution of variables in the models, focused on specific seasonal situations. To identify the important situation-dependent variables, while considering a large number of possible explanatory variables, we utilize a powerful regression scheme that follows the sparse modeling concept [3-12]. This concept has been well-established in the machine learning community and has produced remarkable achievements for many real-world ill-defined problems [3-13, 3-14], such that the relationship among numerous variables can be discussed based on a limited number of samples. In particular, we focus on an attractive class of sparse regression models, sparse partially linear additive models (sparse PLAMs) [3-15, 3-16], which select a limited number of important variables among a large number of possible explanatory variables, and simultaneously selects the linearity/nonlinearity in the plausible relationships between the selected variables and the target demand. A possible difficulty in this type of analyzation approach, based on situation-dependent demand modeling, will be inconsistency of the selected variables under various situations in the interpretation process. When the model is trained based on a dataset with a specific time period focus, the selected variables may be significantly different from those resulting

from a dataset focusing on another time period. In identifying a limited number of dominant variables for describing the demand behavior, in consideration of various situations, a methodology is required to discuss the substitutability of selected variables with others, with respect to description accuracy. In this study, we adopt a scheme to enumerate multiple plausible sparse PLAMs composed of different variables that achieve a similar description accuracy [3-17] for each situation. Additionally, we propose a procedure for identifying minimum and dominant variable subsets required for modeling annual demand behavior, considering its situational variation. The major contributions of this study are as follows:

1. We focus on many possible variables that have not been discussed sufficiently in previous studies for situation-dependent demand modeling.
2. We propose an hourly demand modeling framework based on a sparse modeling technique called sparse PLAMs.
3. We propose a procedure for identifying a limited number of annually dominant variables using the situation-dependent modeling results by adopting a model enumeration technique.
4. We apply the framework to a real-world dataset and discuss the dominant factors that affect the situation-dependent hourly demand.

The rest of the chapter is organized as follows. Section 3.2 provides several characteristics of the electric power demand targeted in this study and reviews other relevant works. Section 3.3 provides a basic outline of the situation-dependent modeling of the hourly demand based on the sparse modeling technique. Section 3.4 is devoted to explaining the proposed identification scheme of annually dominant variables using the situation-dependent modeling results of hourly demand. Section 3.5 shows the evaluation results of the proposed framework by using a real-world dataset and provides some observations about the identified variables. Finally, in Section 3.6, concluding remarks are provided.

3.2 Identifying the dominant factors characterizing the situation-dependent behavior in electric power demand

The electricity net-load varies greatly according to various factors. In particular, the large-scale penetration of renewable energy sources and energy-saving trends could drastically change net-load behavior. Figure 3-1 shows the transition of hourly electric power demand observed in the Tokyo, Japan geographic area from April 1, 2005 to March 31, 2016. As shown in the figure, the electric power demand has explicit annual seasonality, i.e. the demand tends to increase during summer and winter and decrease in the spring and autumn. Meanwhile, the figure also implies that the load curve could be affected by various time-dependent factors, such as weekdays and holidays, weather conditions, and socio-economic factors, causing annual peak demand to vary significantly. Particularly, the annual peak demand behavior has changed drastically since the Great East Japan Earthquake in 2011; distributed renewable power sources have increasingly penetrated, and energy-saving tendencies seem to have been encouraged based on situational changes.

Figure 3-2 shows average daily load curves constructed by focusing on certain months from Fig. 3-1. This figure indicates that hourly demand has a large variance during the daytime and also shows that the timing and magnitude of the peak demands can differ depending on situational factors.

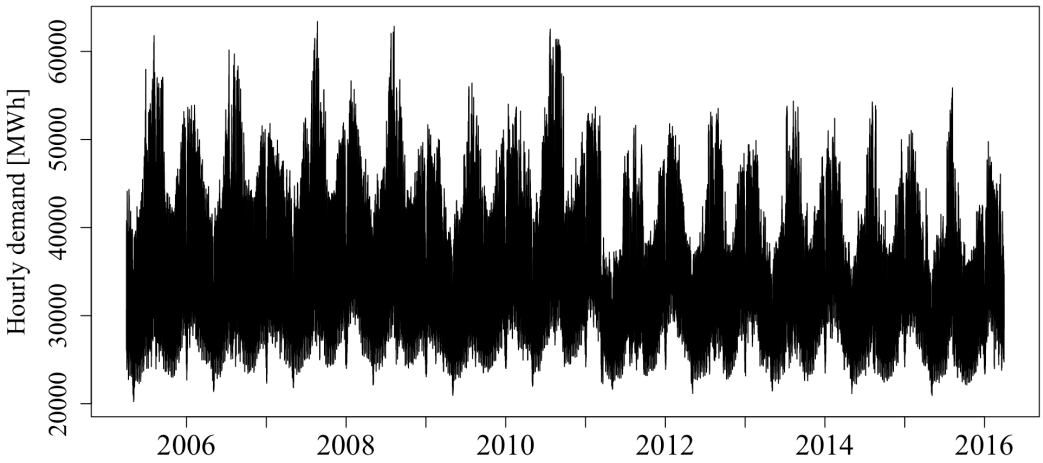


Fig. 3-1 Hourly electricity demand plot for the Tokyo, Japan geographic area for a time span from April 1, 2005 through March 31, 2016.

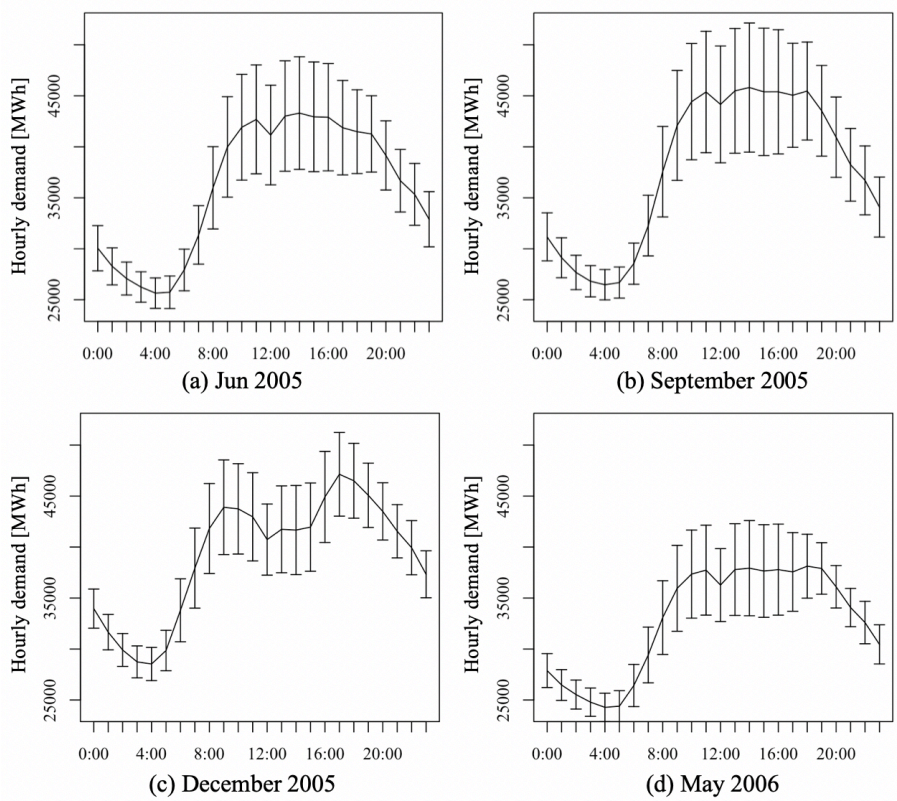


Fig. 3-2 Examples of the average hourly demand, constructed from the same data shown in Fig. 3-1.

Figure 3-3 provides examples of the relationships between demand and other variables observed simultaneously. This figure suggests that the relationships among variables could be explicitly nonlinear.

In this study, we focus on hourly demand and propose an analyzation approach for identifying the dominant influential factors, considering various situations characterized by the time periods and seasonality, while also identifying the linearity or nonlinearity in the variable relationships. To account for the diversity of all possible factors, we focus on hundreds of explanatory variables, which can be categorized into attribute groups such as weather, interest rate stock price, calendar, and GDP data (see Appendix for the details of the variables utilized in this study). The relationship between electricity demand and these variables has not been thoroughly discussed in the prior electricity demand studies, so this work fills a necessary gap and improves the generalizability and applicability of such demand studies. In the next section, we introduce the idea of sparse PLAMs, which is a very promising application of sparse modeling techniques for selecting a limited number of informative variables among a large number of candidate variables, while also selecting plausible linear or nonlinear relationships.

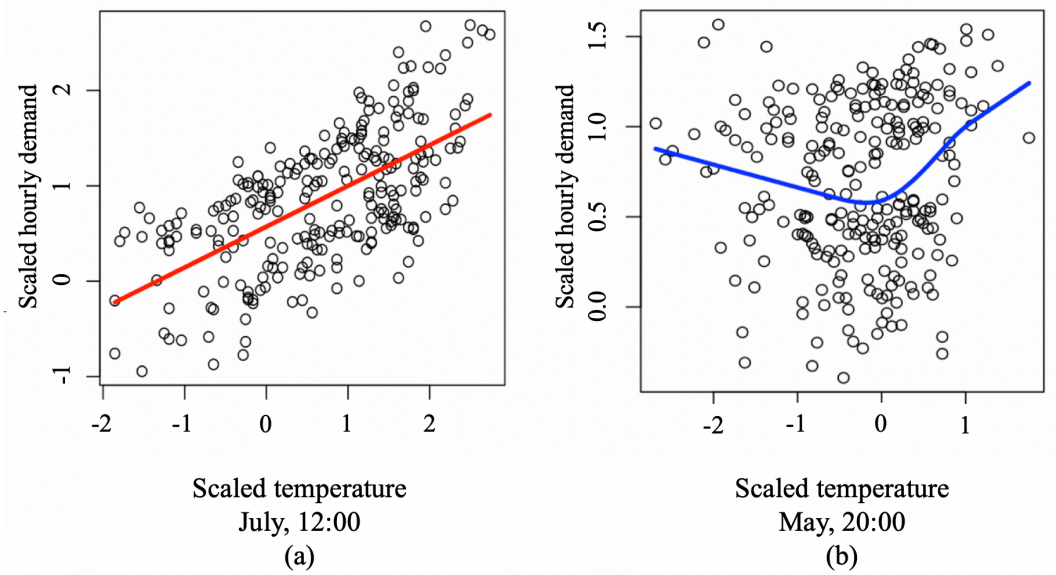


Fig. 3-3 Examples of relationships between target demand and temperature in (a) July and (b) May. The solid lines show the curves derived by sparse PLAM introduced in Section 3.3.

3.3 Basic framework for selecting the informative variables

3.3.1 Sparse partially linear additive models (Sparse PLAMs)

Let $\{(\boldsymbol{x}_{ymdh}, l_{ymdh})\}$ be a set of pairs containing l_{ymdh}, which is an hourly demand observed at hour h in day d in month m of year y, and $\boldsymbol{x}_{ymdh} = (x^1_{ymdh}, \dots, x^I_{ymdh})$, which is a vector of I variables observed at the corresponding hour, day, minute, and year[1]. We also let $\mathcal{L}$ and $\mathcal{N}$ be the index subsets of $\mathcal{S} = \{1, \dots, I\}$ indicating the variables related to the target demand linearly and nonlinearly, respectively. The partially linear additive model [3-18] is a formulation for describing target demand with a large number of variables containing partially linear and nonlinear relationships, and is defined as follows:

$$l_{ymdh} \cong f(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}) = \beta_0 + \sum_{j \in \mathcal{L}} \beta_j x^j_{ymdh} + \sum_{j \in \mathcal{N}} \sum_k \tau_{j,k} \phi^k_j(x^j_{ymdh}) , \quad (3.1)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_P), \{\boldsymbol{\tau}_j = \{\tau_{j,1}, \dots, \tau_{j,K}\}\}\}$ is a set of the model coefficient parameters; $\boldsymbol{\beta}$ is the set of coefficient parameters for explanatory variables $\boldsymbol{x}$; and $\boldsymbol{\tau}_j$ is the vector of coefficient parameters for transformation functions $\phi^1_j(x^j), \dots, \phi^K_j(x^j)$. In this study, the following cubic spline transformation functions with K bases, which are supposed to work well with PLAM in previous studies[2] [3-18], are used:

$$[\phi^k_j(x^j_{ymdh}); k = 1, .., K] = \left[x^j_{ymdh}, \left(x^j_{ymdh} - x^j_{(1)}\right)^3_+, \dots, \left(x^j_{ymdh} - x^j_{(K-1)}\right)^3_+\right], \quad (3.2)$$

$$(z)_+ = \begin{cases} 0 \ (z < 0) \\ z \ (z \geq 0) \end{cases}, \quad (3.3)$$

where $x^j_{(1)}, \dots, x^j_{(K-1)}$ are knots of the spline chosen from quantiles in the sample set. The parameter is estimated on the basis of minimizing the squared error loss under the given subsets $\mathcal{L}$ and $\mathcal{N}$, as follows:

$$\hat{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{y,m,d,h} \left(l_{ymdh} - f(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta})\right)^2 . \quad (3.4)$$

The model shown in Eq. (3.1) is a flexible approach for representation of the linear and nonlinear relationships between explanatory variables and the target demand. However, when the number of explanatory variables I is large, instability in the estimation results, caused by multicollinearity [3] [3-19], will be a nonnegligible problem. Further, the combinatorial explosion of possible candidates is another potential issue in selecting the appropriate index subsets $\mathcal{L} \subseteq \mathcal{S}$ and $\mathcal{N} \subseteq \mathcal{S}$, since they have to be selected by constructing and evaluating the models for all the possible mutually exclusive index subsets in $\mathcal{S}$.

---

[1] Through this paper, we assume that x^p_{ymdh} and l_{ymdh} are standardized to have zero mean and unit variance.

[2] The method of nonlinear transformation utilizing cubic spline changes the model description accuracy depending on the symmetry of the explanatory variables (see Fig. 3-A1 in the Appendix for the skewness of each variable). In future works, the model description accuracy may be enhanced through processing of the explanatory variables to improve symmetry and implementation of alternative nonlinear transformation techniques such as quadratic or higher quartic spline transformations, or other nonlinear transformations [3-21]. These methods can be comparatively evaluated through statistical validation.

[3] The variance inflation factor (VIF) [3-22] of each focused explanatory variable in this study, which measures collinearity with other explanatory variables shows that the explanatory variables exhibit high levels of collinearity with other variables (see Fig. 3-A2 in the Appendix for the detail of the result of VIF).

The sparse PLAM is a promising technique for alleviating these limitations. In the sparse PLAM scheme, the formulation given in Eq. (3.1) is reformulated as follows:

$$l_{ymdh} \cong f(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}) = \beta_0 + \sum_{j \in \mathcal{S}} \left((\beta_j + \tau_{j,1}) x^j_{ymdh} + \sum_{k=2}^{K} \tau_{j,k} \phi^k_j (x^j_{ymdh}) \right). \quad (3.5)$$

In this case, the coefficient parameter $\boldsymbol{\theta} = \{\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p), \{\boldsymbol{\tau}_j = \{\tau_{j,1}, \dots, \tau_{j,K}\}\}\}$ is estimated by introducing a sparse modeling technique known as the overlap group LASSO [3-20] and minimizing the following objective function,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} F(\boldsymbol{\theta}; \mathcal{S}, \lambda) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{y,m,d,h} \left(l_{ymdh} - f(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}) \right)^2 + \lambda \sum_{j \in \mathcal{S}} (|\beta_j| + \|\boldsymbol{\tau}_j\|_2), \quad (3.6)$$

where λ is a positive regularization constant for the penalty and $\|\boldsymbol{\tau}\|_2 = \sqrt{\sum_{k=1}^{K} \tau_k^2}$ is the L2-norm of the vector $\boldsymbol{\tau}$. Components of the minimizer $\widehat{\boldsymbol{\theta}}$ in Eq. (3.6) tend to be zero for reducing absolute penalty and the L2-norm penalty, while reducing the squared error loss. Redundant variables tend to be zero so that a subset of explanatory variables, composed of multiple informative variables, are expected for selection. This type of regularization penalty also alleviates the risk of estimation instability caused by the multicollinearity. In particular, the regularization introduced in Eq. (3.6) encourages the estimation results of the sparse PLAM to achieve one of the following situations:

- $\beta_j \neq 0, \tau_{j,k} = 0$ $(\forall k)$: variable x^j has a linear relationship to the demand;
- $\tau_{j,k} \neq 0$ $(\exists k)$: variable x^j has a nonlinear relationship to the demand;
- $\beta_j = 0, \tau_{j,k} = 0$ $(\forall k)$: variable x^j has no relationship with the demand.

These sparse PLAM characteristics efficiently derive the subset of informative explanatory variables, while simultaneously identifying linear and nonlinear relationships between these variables and the demand.

3.3.2 Situation-dependent modeling for electric power demand

Factors that affect the demand might vary depending on seasonal situations. Constructing a model using data observed under a certain seasonal situation while identifying informative explanatories and comparing with models constructed under other situations will be useful for discussing changes in factors according to situations. A naive approach for selecting situation-dependent informative variables can be achieved by modeling a sparse PLAM based on the data subset containing the targeting situation. Figure 3-4 shows the overview of the basic approach for selecting situation-dependent informative variables; the informative variables are derived for each situation in this approach.

For example, if we focus on the specific month m for revealing annual seasonality, we may derive informative variables by modeling a sparse PLAM based on the data subset collected in month m as follows:

$$\widehat{\boldsymbol{\theta}}_m = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} F_m(\boldsymbol{\theta}; \mathcal{S}, \lambda)$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{y,d,h} \left(l_{ymdh} - f\left(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}\right) \right)^2 + \lambda \sum_{j \in \mathcal{S}} \left(|\beta_j| + \left\| \boldsymbol{\tau}_j \right\|_2 \right) . \quad (3.7)$$

The derived parameter $\widehat{\boldsymbol{\theta}}_m$ in Eq. (3.7) contains non-zero coefficients for informative explanatory variables that describe a specific seasonal situation (i.e., month m). In the same manner, if we focus on daily and annual seasonality simultaneously by focusing on samples collected in month m and hour h in the dataset, informative variables under this situation can be derived as follows:

$$\widehat{\boldsymbol{\theta}}_{mh} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} F_{mh}(\boldsymbol{\theta}; \mathcal{S}, \lambda)$$
$$= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{y,d} \left(l_{ymdh} - f\left(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}\right) \right)^2 + \lambda \sum_{j \in \mathcal{S}} \left(|\beta_j| + \left\| \boldsymbol{\tau}_j \right\|_2 \right) . \quad (3.8)$$

Non-zero coefficients contained in the derived parameter $\widehat{\boldsymbol{\theta}}_{mh}$ in Eq. (3.8) suggest informative explanatory variables under the specific seasonal situation focusing on month m and hour h. Note that the parameters involve information of the selected variables within the given data set, so that the regression model-based approaches introduced in the previous subsection achieve selection of relatively informative variables under the conditional data subset, focused on the time of interest. We call the approach that builds individual models focusing on specific seasonal situations, such as Eqs. (3.7) and (3.8), *situation-dependent modeling*.

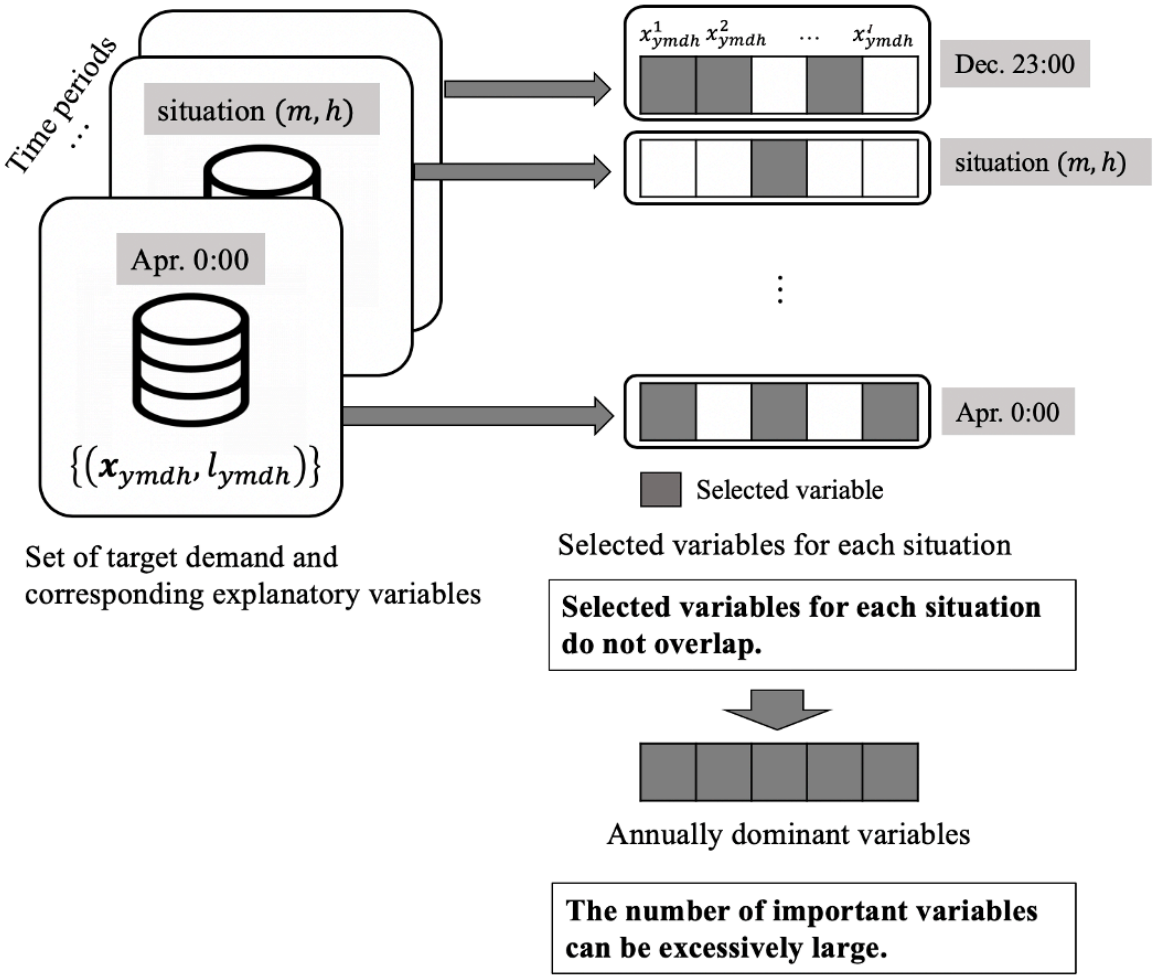


Fig. 3-4 Overview of basic framework for selection of informative variables.

3.3.3 Issues with variable selection based on situation-dependent modeling

The sparse modeling technique adopted in the sparse PLAM enjoys flexible and computationally efficient selection of informative variables and identifies linearity or nonlinearity in the relationships between explanatory variables and the target demand. Meanwhile, the situation-dependent modeling approach has the limitation that selection results can vary significantly depending on the data collected in a given situation of interest [3-23]. In our context, if we focus on samples collected in month m and hour h within the dataset, the sparse PLAM is derived as a global minimizer of Eq. (3.8), but this type of approach can overlook potentially relevant features. Due to this characteristic, the selected variables may change entirely when we construct seasonal models using data collected in different time periods. This implies that all selected variables in the models are necessary to describe annual demand behavior, and the number of important variables for seasonal variation expression can be excessively large as shown in Fig. 3-4 schematically. This will complicate the interpretation of the derived factors unnecessarily.

In recent years, some studies have been conducted on the open difficulty on reproducibility of the variable selection results when constructing sparse models from data extracted from the same statistical population. For example, adaptive LASSO [3-24] has been proposed to realize the oracle property that guarantees the stochastic consistency of variables selected under certain data condition. Random LASSO [3-25] has been proposed under another idea and quantifies the importance of informative variables in terms of the consistency of the variables selected under slightly different datasets. However, these approaches focus on the performance of variable selection on datasets holding the same statistical properties; the other approach is required to discuss the essential differences in the variables selected from datasets collected in different seasonal situations where the statistical properties may differ.

In practice, when we construct a sparse regression model using a large number of explanatory variables, some variables might make very similar contributions to the model. In this case, the variables selected for explaining the seasonal variation can be replaced by other variables while preserving the same description accuracy. Considering this, in the next section we propose an approach to construct situation-dependent models using a limited number of variables that are commonly and consistently used, as much as possible, to improve the interpretability of the annually dominant variables in describing the electricity demand behavior.

3.4 Identifying the annually dominant variables based on enumerate sparse PLAMs

3.4.1 Enumeration and selection approach

Now, we propose an approach for identification of a limited number of annually dominant variables according to the situation-dependent modeling approach. Figure 3-5 shows the overview of the proposed approach.

The proposed approach does not focus only on a single global optimal model under each situation but also enumerates a set of plausible models consisting of nearly optimal models (Step 1 in Fig. 3-5). The proposed approach contributes to identify the minimum number of essential variables required to describe annual demand behavior (Step 2 in Fig. 3-5). Consequently, variables used only in rare seasonal situations tend to be replaced by substitutable variables used in other models, and the variables used in models tend to overlap. The approach is expected to identify a limited number of essential and dominant variables for describing the situation-dependent behavior of annual demands and allow us to analyze situational changes in the informative variables. In the remainder of this section, we focus on the situation-dependent models shown in Eq. (3.8), given by the pair of $m \in \{1, \dots, M\}$ and $h \in \{1, \dots, H\}$ for description.

Our proposed framework is composed of the following two steps:

Step 1: Enumerate the plausible candidate models by focusing on the given situation (m, h), using the set of target demand and corresponding explanatory variables, i.e. $\{(\boldsymbol{x}_{ymdh}, l_{ymdh}); \forall y, d\}$. This step is repeated for all pairs $\{(m, h)\}$.

Step 2: Select a representative model from the candidates enumerated in Step 1 for each situation (m, h), such that the cardinality of the union of informative variables used in the extracted models is minimized; the extracted models provide a set of annually dominant variables.

In the following subsections, the details of each step are explained.

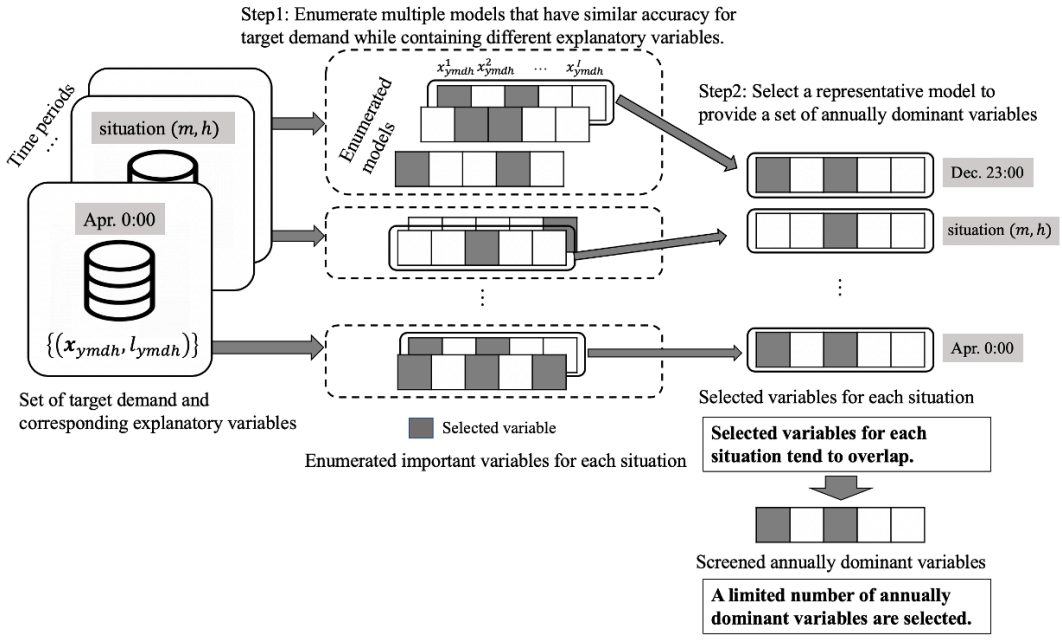


Fig. 3-5 Overview of the proposed framework for selection of informative variables.

3.4.2 Enumeration of situation-dependent models based on sparse PLAMs

In Step 1, we apply the enumeration framework for sparse PLAMs, which has been motivated by the enumerate LASSO [3-17]. The framework utilizes the following property of the sparse PLAMs, i.e. for given $\mathcal{S}' \subset \mathcal{S}$,

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{mh} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, F_{mh}(\boldsymbol{\theta}; \mathcal{S}', \lambda) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, F_{mh}(\boldsymbol{\theta}; \mathcal{S}, \lambda)\,, \end{aligned} \tag{3.9}$$

holds if $\left|\hat{\beta}_j\right| = 0$ and $\left\|\hat{\boldsymbol{\tau}}_j\right\|_2 = 0$ hold for all $j \in \mathcal{S} \setminus \mathcal{S}'$ (see reference [3-26] for detail). The scheme proposed in reference [3-17] enumerates a set of suboptimal models by efficiently searching in ascending order of the objective function values shown in Eq. (3.8) while replacing the candidate variables selected in the optimal model. Algorithm 1 shows the enumeration step.

This step serves to enumerate multiple models that have similar accuracy for target demand while containing different explanatory variables. A positive parameter ε in line 8 of Algorithm 3-1 controls the suboptimality of the enumerated results.
We let $\mathcal{S}_{mh}^{(i)} \subseteq \mathcal{S}$ be the index subset of the selected variables corresponding to the non-zero components in parameter $\hat{\boldsymbol{\theta}}_{mh}^{(i)}$ of the enumerated sparse PLAMs, and $\boldsymbol{\mathcal{S}}_{mh} = \{\mathcal{S}_{mh}^{(1)}, \mathcal{S}_{mh}^{(2)}, \dots, \mathcal{S}_{mh}^{(I_{mh})}\}$ stand for the set of variable index subsets enumerated in Algorithm 1 where I_{mh} is the number of enumerated models under the given situation (m, h). By applying Algorithm 1 to all (m, h) situations in Step 1, we can obtain an index set of informative variable subsets, $\{\boldsymbol{\mathcal{S}}_{mh}; \forall m, h\}$.

3.4.3 Selection for identification of annually dominant variables

In Step 2, we focus on all the sets of variable index subsets $\{\boldsymbol{\mathcal{S}}_{mh}; \forall m, h\}$ enumerated for each situation and select the essential variables composed of a minimal set required to represent annual demand behavior, that is,

$$\hat{\mathcal{S}} = \min_{\{i_{mh} \in \{1,\dots,I_{mh}\}; \forall m,h\}} \left| \bigcup_{m,h} \mathcal{S}_{mh}^{(i_{mh})} \right|. \tag{3.10}$$

The procedure evaluates enumerated models from the viewpoint of consistency in selected variables under various situations and selects situation-dependent models represented by variables commonly adopted among the models as much as possible; these variables suggest the minimal dominant variables for description of demand. The minimizer of Eq. (3.10) can be found by evaluating $\prod_{m,h} I_{mh}$ possible combinations, but this brute force approach is computationally expensive when the number of candidates I_{mh} is large for each (m, h). In this study, the minimization problem Eq. (3.10) is reformulated as a binary linear programming problem [3-27]. Let $s_{mh}^{j,(i_{mh})}$ be the binary value given as:

$$s_{mh}^{j,(i_{mh})} = \begin{cases} 1, & j \in \mathcal{S}_{mh}^{(i_{mh})} \\ 0, & j \notin \mathcal{S}_{mh}^{(i_{mh})}, \end{cases} \tag{3.11}$$

and $d_{mh}^{i_{mh}} \in \{0,1\}$ $(i_{mh} \in \{1, \dots, I_{mh}\}, \forall m, h)$ be the binary variable indicating the utilization of the index subset $\mathcal{S}_{mh}^{i_{mh}}$. Then, the minimization problem given in Eq. (3.10) is reformulated as follows using auxiliary variables $v_j \in \{0,1\}$ $(\forall j \in \mathcal{S})$:

$$\min \sum_j v_j, \tag{3.12}$$

subject to the following constraints:

$$v_j \leq \sum_{m,h} \sum_{i_{mh} \in \{1,\ldots,I_{mh}\}} d_{mh}^{i_{mh}} s_{mh}^{j,(i_{mh})} \leq MHv_j \ (\forall j \in \mathcal{S}), \tag{3.13}$$

$$\sum_{i_{mh} \in \{1,\ldots,I_{mh}\}} d_{mh}^{i_{mh}} = 1 \ (\forall m, h) . \tag{3.14}$$

The optimizer of the variables $\{\hat{d}_{mh}^{i_{mh}}\}$ indicates the selected candidate $\hat{\imath}_{mh}$ such that $\hat{d}_{mh}^{i_{mh}} = 1$ holds in the enumerated models, so that the annually dominant variables are given as:

$$\hat{\mathcal{S}} = \left\{ j;\ s_{mh}^{j,(\hat{\imath}_{mh})} = 1, \exists m, h \right\}. \tag{3.15}$$

As a whole procedure, suboptimal models for representing situation-dependent demand behavior under each situation are enumerated. Further, identification of the minimal dominant variables $\hat{\mathcal{S}}$, necessary for representing annual demand behavior, is realized.

Algorithm 3-1: Enumeration of suboptimal sparse PLAMs for the situation (m, h).

Input: (m, h).
Set target $\mathcal{C} \leftarrow \mathcal{S}$ and discovered variable indices $\mathcal{D} \leftarrow \emptyset$.
$\widehat{\boldsymbol{\theta}}_{mh} \leftarrow \mathrm{argmin}_{\boldsymbol{\theta}} F_{mh}(\theta; \mathcal{C}, \lambda)$.
Initialize min-heap $\mathcal{T} \leftarrow \left\{\left(\widehat{\boldsymbol{\theta}}_{mh}, \mathcal{C}, \mathcal{D}\right)\right\}$ with key $F_{mh}(\widehat{\boldsymbol{\theta}}_{mh}; \mathcal{C}, \lambda)$.
for $i = 1,2, \ldots$ **do**
  Extract $\left(\widehat{\boldsymbol{\theta}}, \mathcal{C}, \mathcal{D}\right)$ from heap $\mathcal{T}$.
  Store a pair $\left(\widehat{\boldsymbol{\theta}}_{mh}^{(i)}, F_{mh}^{(i)}\right) \leftarrow \left(\widehat{\boldsymbol{\theta}}, F_{mh}(\widehat{\boldsymbol{\theta}}; \mathcal{C}, \lambda)\right)$.
  if $(F_{mh}^{(i)} - F_{mh}^{(1)})/F_{mh}^{(1)} < \varepsilon$ **then**
    for $j \in \{j; \left|\hat{\beta}_j\right| > 0$ or $\left\|\hat{\boldsymbol{\tau}}_j\right\|_2 > 0, j \notin \mathcal{D}\}$ **do**
      $\widehat{\boldsymbol{\theta}}_{mh} \leftarrow \mathrm{argmin}_{\boldsymbol{\theta}} F_{mh}(\boldsymbol{\theta}; \mathcal{C} \setminus \{j\}, \lambda)$.
      Insert $\left(\widehat{\boldsymbol{\theta}}_{mh}, \mathcal{C} \setminus \{j\}, \mathcal{D}\right)$ to $\mathcal{T}$ with key $F_{mh}(\boldsymbol{\theta}; \mathcal{C} \setminus \{j\}, \lambda)$.
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{j\}$.
    end for
  else
    break
  end if
end for
Output: $\{\left(\widehat{\boldsymbol{\theta}}_{mh}^{(i)}, F_{mh}^{(i)}\right)\}$.

3.5 Numerical simulation

3.5.1 Simulation setup

In this section, situation-dependent models targeting hourly electricity demand were constructed based on the dataset collected in Tokyo, Japan according to the proposed approach. Actual hourly demand data set collected from April 2005 to March 2013 was used to train the models so that it could estimate hourly demand from April 2013 to March 2016 for evaluation. We utilized 274 explanatory variables belonging to typical categories such as weather, interest rate stock price, calendar, and GDP data, as shown in Table 3-1, with full details included in the Appendix. These hourly demand and explanatory variables are standardized to have zero means and unit variances. In Eq. (3.2), we used parameter $K = 3$ for nonlinear transformation. The parameter λ in Eq. (3.6) is determined by cross validation [3-28] using the training data. In Step 2 of the proposed method, we used the parameter $\varepsilon = 0.003$, which controls the number for bootstrap sampling.

We focus on four models shown in Table 3-2 to describe the hourly demand. Conventional model focuses on annual seasonality by targeting the data subset collected in month m while constructing linear models. Model 1 also focuses on annual seasonality by targeting the data subset collected in month m based on Eq. (3.7); in particular, explanatory variables, i.e. the weather information, that have been used alone in some typical studies (e.g. [3-42]), are utilized. Meanwhile, Model 2 focuses on a large number of variables that have been used to explain the demand curve in several recent works (e.g. [3-6]). Model 3 focuses on daily seasonality by targeting the data subset collected in month m and hour h based on Eq. (3.8) which expresses different seasonality with Models Conventional, 1 and 2; note that Model 3 is a naive extension of the recent works [3-43, 3-44] modeling nonlinearity based on the sparse modeling technique [3-45, 3-46] to the situation-dependent model. Model 4 focuses on daily seasonality which is the same as Model 3 while implementing the proposed procedure described in Section 3.4 for derivation of a limited number of annually dominant variables. We compare these models from the viewpoints of the description accuracy and the characteristics of the selected variables, and evaluate the effectiveness of the proposed method implemented in Model 4.

Table 3-1 Categories of explanatory variables used.

Attribute: Number of the variables	
Weather [3-29]: v1-10	Hourly
Interest rate and stock price [3-30, 3-31, 3-32, 3-33]: v11-16	Daily
Indices of Tertiary Industry Activity [3-34]: v17-204	Monthly
Indices of Industrial Production [3-35]: v205	
GDP [3-36]: v206-223	
Producer Price Index/ Consumer Price Index [3-37, 3-38]: v224-254	
Electric Appliance Installation [3-39, 3-40]: v255-262	
Number of internet searches (power related words) [3-41]: v263-273	
Calendar data (holiday/weekday; binary dummy): v274	Daily

∗All explanatory variables are standardized to have zero means and unit variances.

Table 3-2. Condition of constructed models.

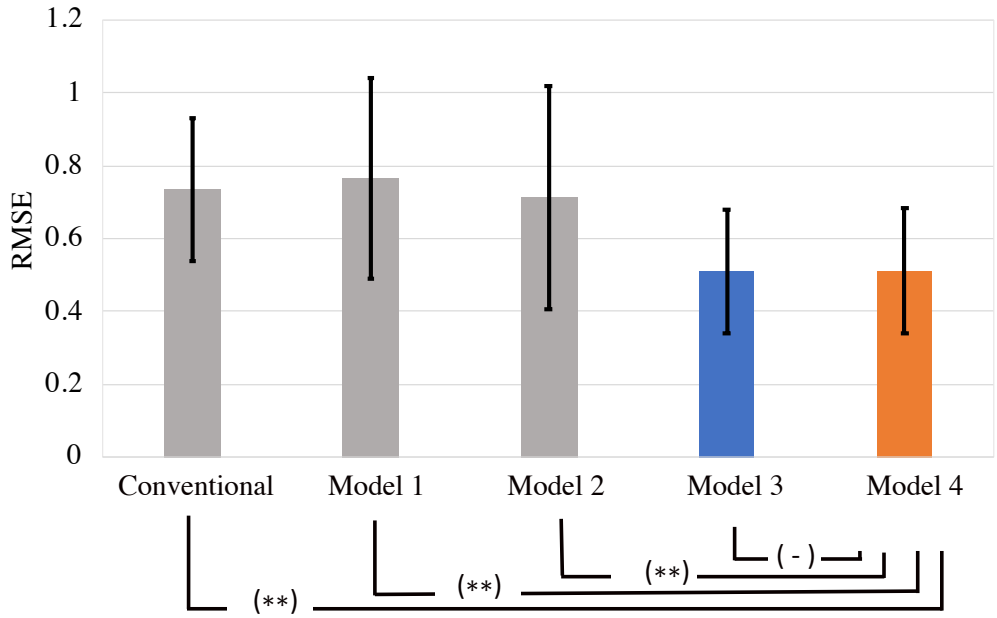
Models	Target variables		Note
Conventional	Data subset collected in month m	$\mathcal{S} = \{1, \dots, 10\}$ (Weather variables)	Utilize linear models
Model 1	Data subset collected in month m	$\mathcal{S} = \{1, \dots, 10\}$ (Weather variables)	Utilize sparse PLAMs w/o applying enumeration and selection process introduced in Section 3.4
Model 2		$\mathcal{S} = \{1, \dots, 274\}$ (All variables)	
Model 3	Dataset collected in month m and hour h	$\mathcal{S} = \{1, \dots, 274\}$ (All variables)	
Model 4		$\mathcal{S} = \{1, \dots, 274\}$ (All variables)	Proposed approach applying the process introduced in Section 3.4

3.5.2 Description accuracy of constructed models

First, we derived the situation-dependent models of hourly demand using the naive and the proposed approaches and compared them from the viewpoint of description accuracy. The motivation for this evaluation is to show that the description accuracy of the proposed method (Model 4) is not significantly worse than the naive extension of the existing modeling schemes. Figure 3-6 shows the average and the standard deviation of the root mean squared errors (RMSE) derived for each time period. The figure also shows the results of the Wilcoxon signed-rank test [3-47, 3-48] for the evaluation of significant differences in error distribution; the results were derived according to the procedure for sequentially rejective multiple test [3-49]. The results demonstrate that Model 2 achieves a lower description error than Conventional and Model 1, implying the existence of informative variables other than the weather variables. The results also show that Models

3 and 4 achieve significantly lower description errors than Conventional, Models1 and 2. We stress that Model 4 shows almost the same performance as Model 3 from the viewpoint of description error; the result of Wilcoxon signed-rank test does not reveal the significant difference between description errors of Models 3 and 4. These results suggest that the situation-dependent models that use a dominant variables subset identified by the proposed method have almost equivalent representation power to the simple situation-dependent modeling approach that naively utilizes the sparse PLAMs.

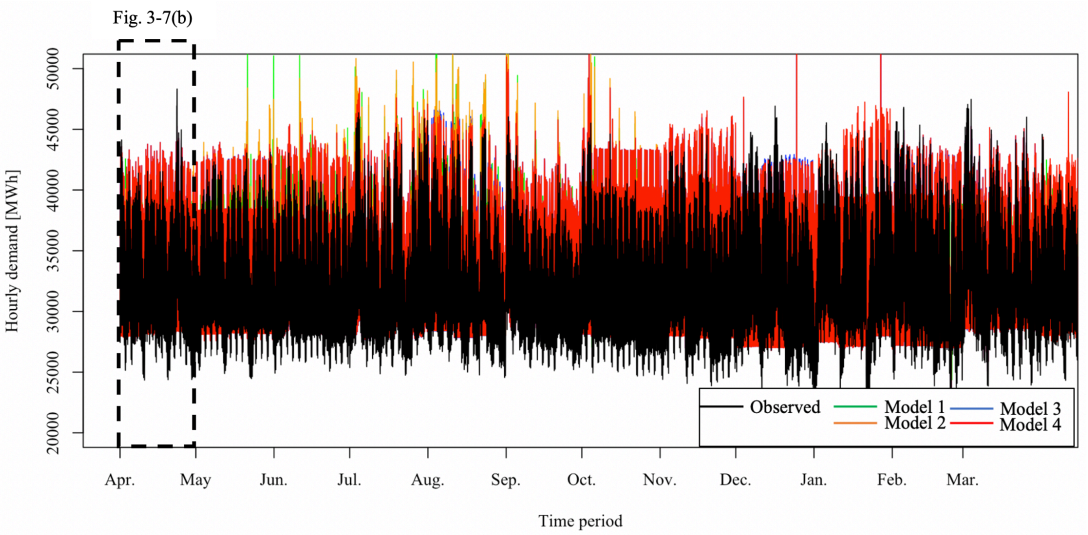
Figure 3-7 shows examples of hourly demand curve focusing on several subsequences in the evaluation period of April 2013. These results show that the hourly demand curves described by Models 1 and 2 are similar and demand curves described by Models 3 and 4 are similar. In particular, the curves represented by Models 3 and 4 capture the transition of actual demand dynamics relatively well. Figure 3-8 shows the corresponding daily description error of the time period shown in Fig. 3-7 (a) measured by RMSE. The results suggest that Models 3 and 4 work well to describe daily variations like on- and off-peaks of demand. This case study shows the usefulness of the modeling scheme focusing on detailed situations, and the appropriateness of the proposed method (Model 4) compared to the naive extension of existing modeling schemes that has been attracting attention in recent years.



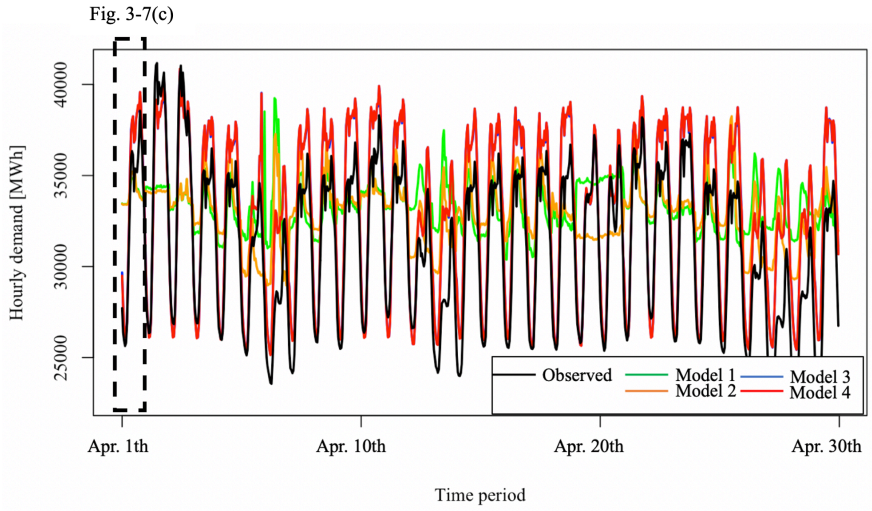
** represents the test result does not accept the null hypothesis under significance levels, $\alpha = 0.025$

\- represents the test result accepts the null hypothesis under significance levels, $\alpha = 0.025$

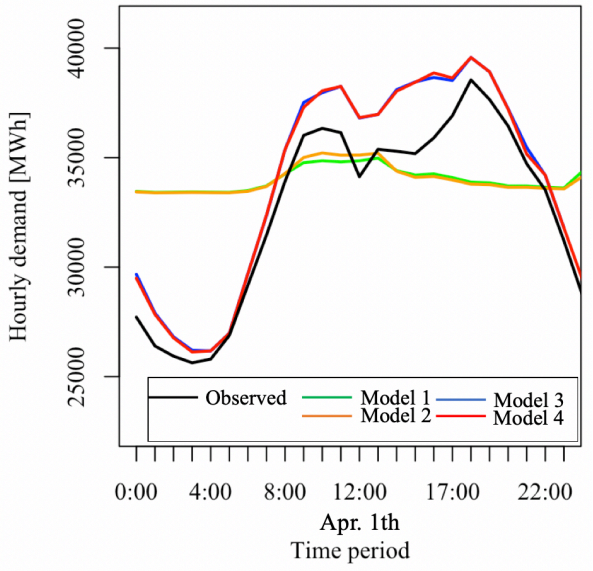
Fig. 3-6 Average of the situation-wise RMSEs and results of Wilcoxon signed-rank test for evaluation of models described in Table 3-2. Each error bar shows standard deviation of the RMSEs derived under various $(\boldsymbol{m}, \boldsymbol{h})$.



(a) Example of hourly demand curves of the evaluation period 2013



(b) Example of hourly demand curves of the evaluation period April 2013



(c) Example of hourly demand curves of the evaluation period April 1st in 2013

Fig. 3-7 Examples of hourly demand curves during the evaluation period of April 2013.

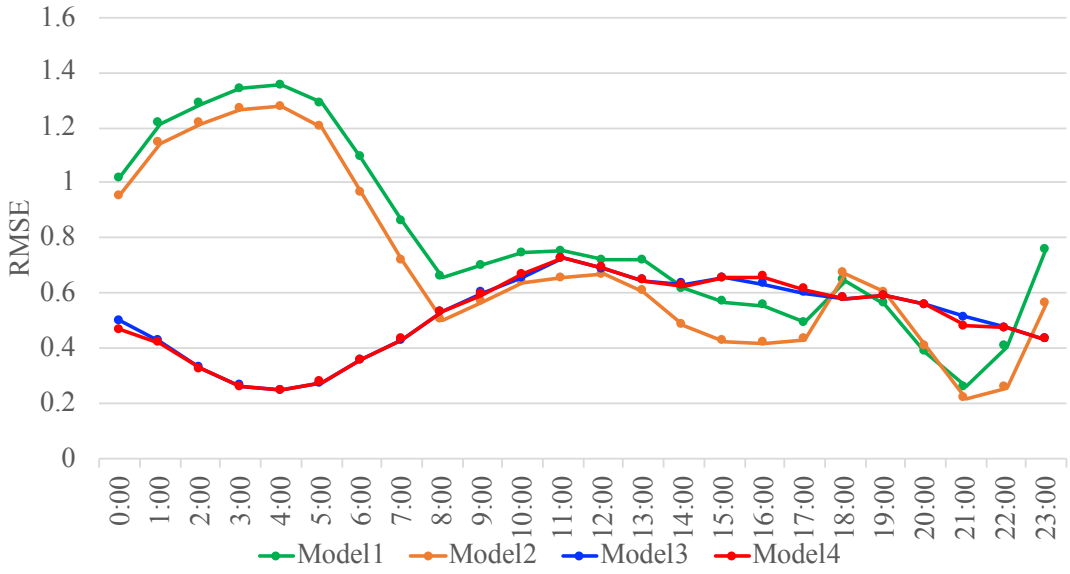


Fig. 3-8 Hourly RMSEs during the evaluation period of April 2013.

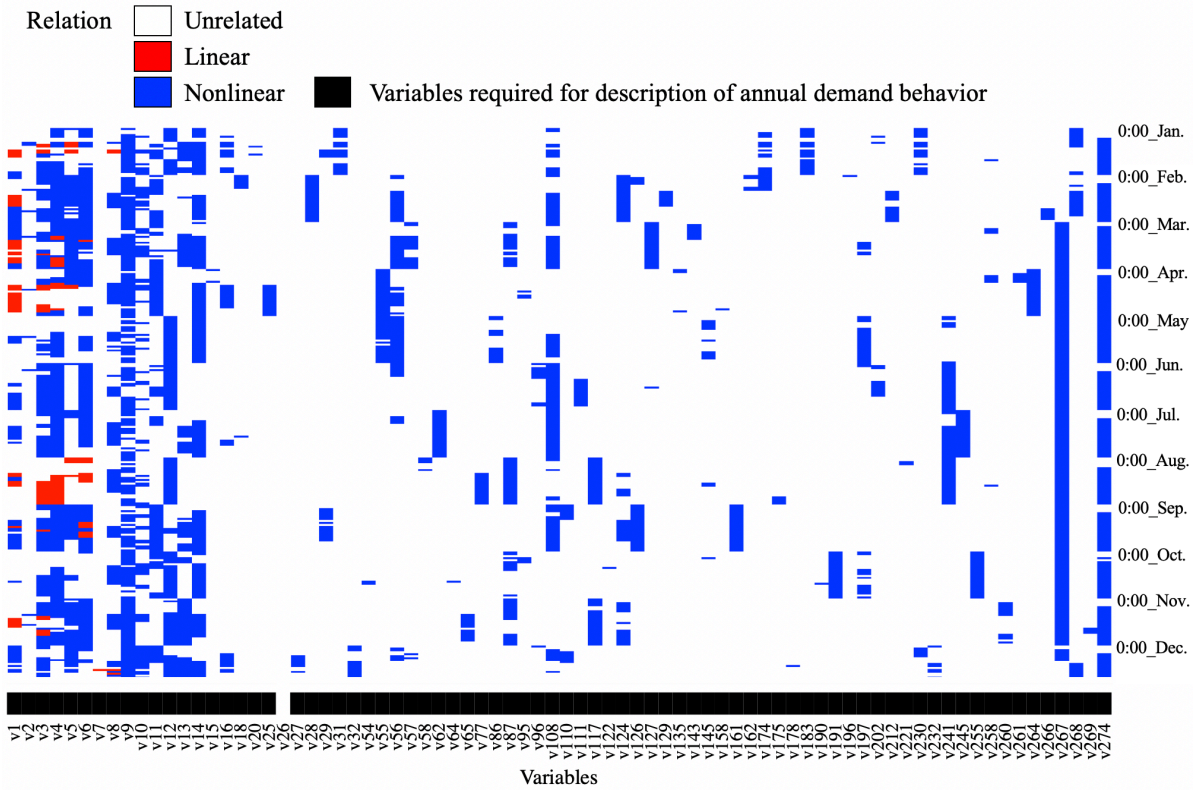
3.5.3 Discussion of selected variables

Figures 3-9 (a) and (b) show the most significant variables selected in Models 3 and 4, respectively. These results, derived based on the situation-dependent modeling approach shown in Eq. (3.8), successfully show the difference in variables selected under various seasonal situations given by pairs of month m and hour h. The variables selected by Model 4 differed in most of the time periods from those selected by Model 3, suggesting that some variables selected in Model 3 were replaced in Model 4. The results shown in Fig. 3-9 indicate that the number of variables required to describe annual demand behavior, using situation-dependent models, is reduced to 77 and 60 among the 274 explanatory variables by Models 3 and 4, respectively. Here, we focus on variables v15 (i.e. *stock price: Nikkei share average*), and v135 (i.e. *index of tertiary industry activity: apparel retail trade*) which are selected only in few seasonal situations by Model 3.

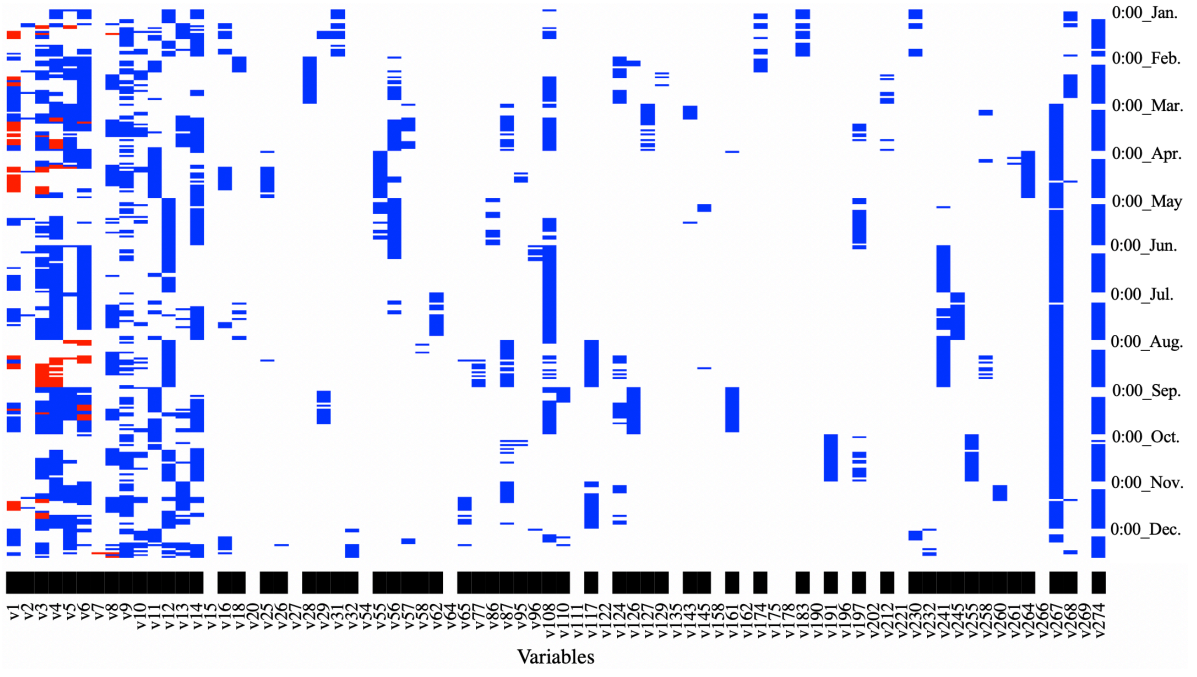
Figures 3-10 (a) and (b) show the variables selected at 0:00 in April by Models 3 and 4, respectively; the results suggest that v15 and v135 are not selected in Model 4. Figure 3-11 shows the selected variables in four models, respectively, enumerated according to the process introduced in Section 3.2 in the corresponding time period.

The results shown in Fig. 3-11 suggests that v15 is removed at the second model ($i_{mh} = 2$) and v9 (i.e., *precipitation*) is removed at the third model ($i_{mh} = 3$). Then, at the fourth model ($i_{mh} = 4$), v135 is removed while v9, v14 (i.e., *interest rate: yield on long gilt*), v25 (i.e., *index of tertiary industry activity: private sector broad casting*), and v161 (i.e. *index of tertiary industry activity: fast food restaurant/ restaurant service industry*) are selected. Model 4 adopted the variables appeared in the fourth row in Fig. 3-11 for this situation by comprehensively utilizing the enumeration results of the models derived under other situations. Consequently, v15 and v135 appeared in Model 3 at 0:00 in April are replaced by v14, v25 and v161 in Model 4. Note that v15 and v135 are removed in all the seasonal situations in Model 4. These results suggest that the proposed approach successfully screens annually dominant variables necessary for situation-dependent models to

describe seasonal changes in demand curves by focusing on variables that can be used interchangeably under each situation.



(a)



(b)

Fig. 3-9 Variables selected in situation-dependent modeling process by (a) Model 3 and (b) Model 4. The x-axis represents the selected variables, and the y-axis represents the transition of the situation ($\boldsymbol{m}$, $\boldsymbol{h}$).

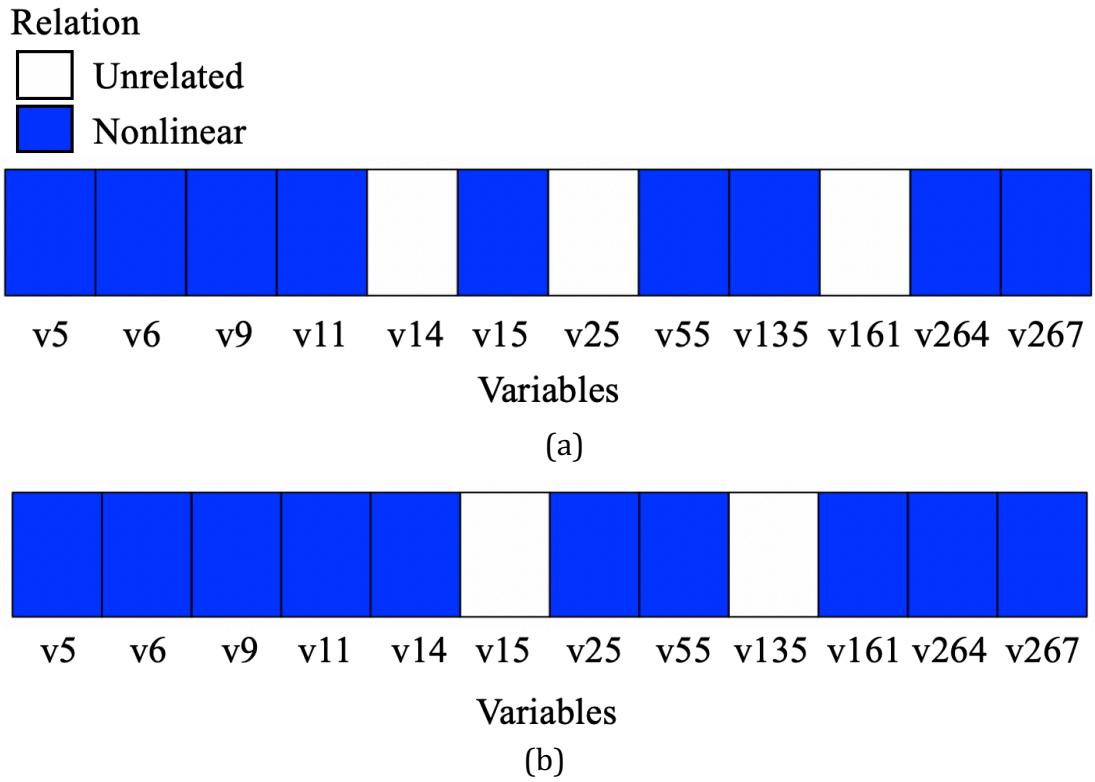


Fig. 3-10 Variables selected in situation-dependent modeling process by (a) Model 3 and (b) Model 4 at 0:00 in April.

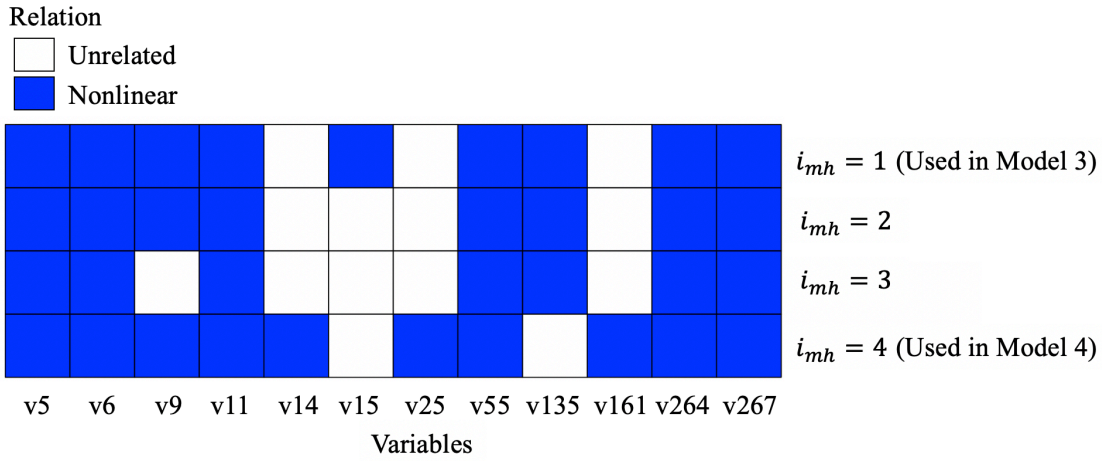
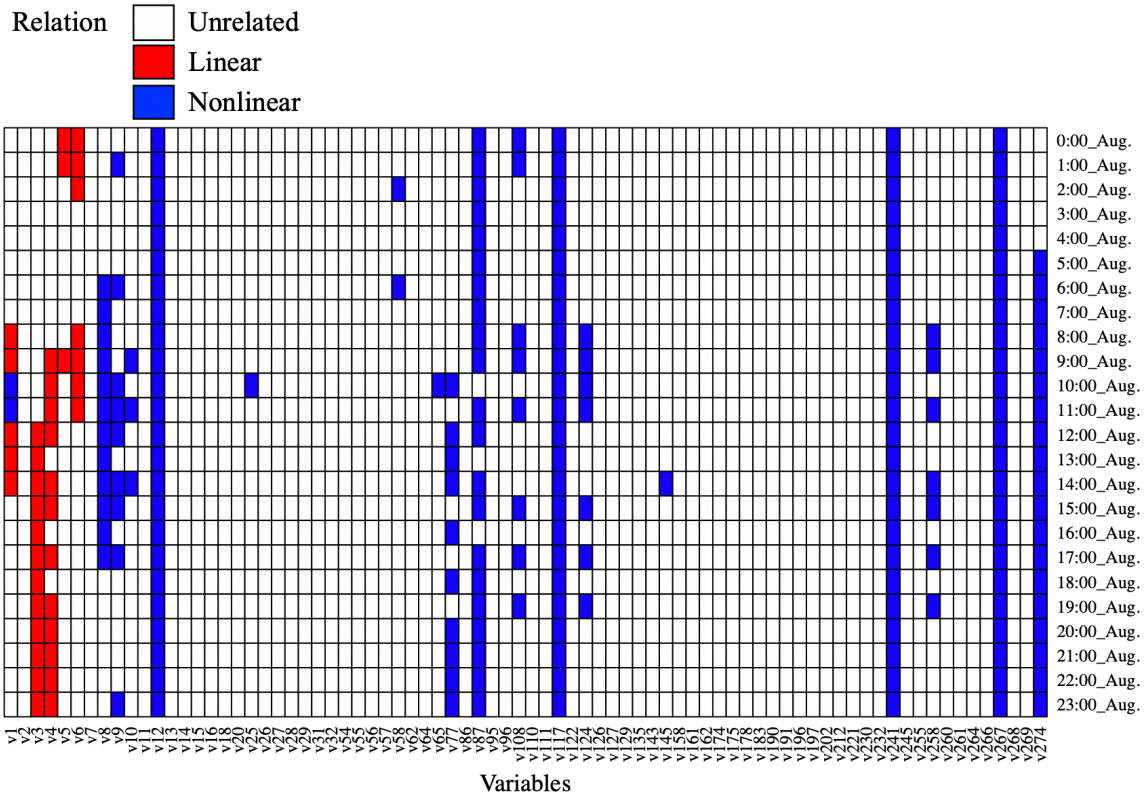


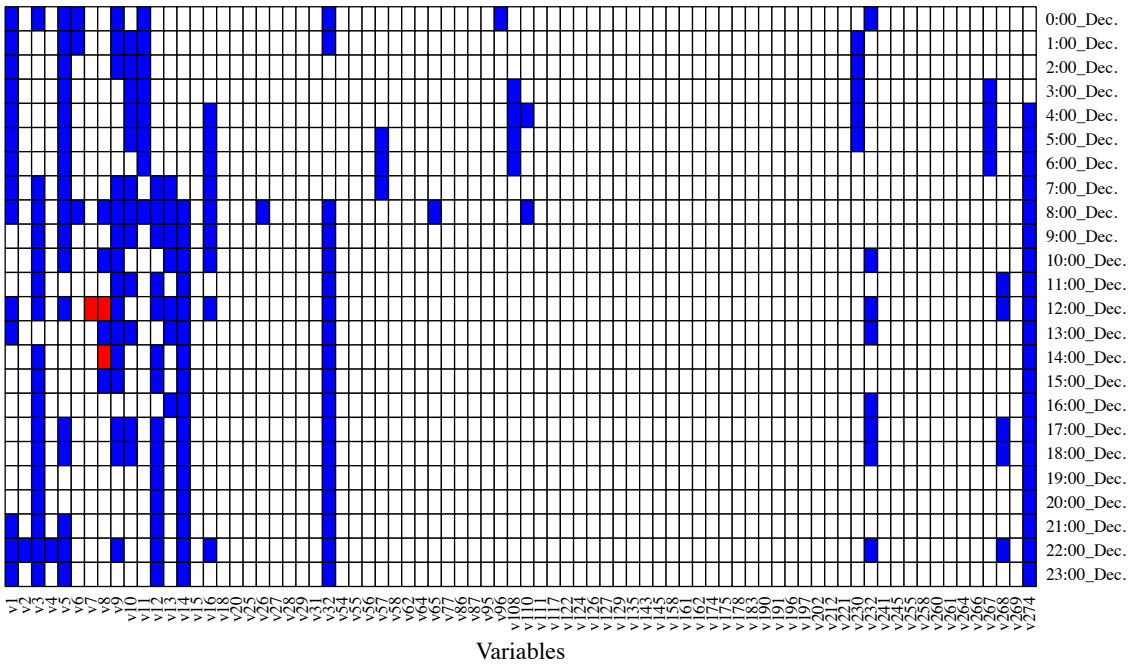
Fig. 3-11 Enumerated sets of variables under $\boldsymbol{m = 4}$ indicating "April" and $\boldsymbol{h = 0}$ indicating "0:00". The model in the first row with ($\boldsymbol{i_{mh} = 1}$) is the optimal model and the other models are enumerated suboptimal models.

Next, we focus in detail on several specific months shown in Fig. 3-9 (b). Figures 3-12 (a) and (b) show the variables selected for each time period in August and December, respectively. Clearly, slightly different variables are selected for each time period in August and December, and most of the linear relationships observed in August become nonlinear in December. For example, although temperature tends to have a linear relationship with demand in August, it tends to have a nonlinear relationship in December; the result suggests the improving the statistical model by assuming nonlinearity between temperature and demand in the winter demand estimation. Additionally, the results show the relevance of v241 (i.e. *Producer price index of transportation equipment*) in August, and the relevance of v32 (i.e. *Indices of tertiary industry activity about motion picture, video and television program distribution*) in December; the result suggests that the necessary to consider additional seasonal factors in statistical modeling according to the seasonal situations. Figure 3-13 shows the behavior of the producer price index of transportation equipment, GDP and hourly electricity demand in August. The figure shows that the producer price index of transportation equipment tends to more adequately explain variations in demand in August.

Figures 3-14 show scatterplots between the scaled demand and the scaled explanatory variables of v1 (i.e., *temperature*) and v267 (i.e., *the number of internet searches for energy saving*), respectively. The data correspond to those variables in Fig. 3-12 are plotted for August in Fig. 3-14 (a) and (c), and for December in Fig. 3-14 (b) and (d). The plots shown in Figs. 3-14 (a) and (b) indicate that the demand tends to be linear and increase monotonically with temperature in August, while the demand decreases nonlinearly according to the temperature in December. This can be interpreted as a typical example of how variable influence changes due to seasonality, derived from the results of situation-dependent modeling. On the other hand, the plots shown in Fig. 3-14 (c) and (d) indicate a nonlinear relationship between the demand and the number of internet searches for all time periods in August, though no such relationship can be seen in December. This indicates a seasonal pattern, that energy saving trends among consumers are more influential in summer than in winter [3-6]. The results in Fig. 3-14 (c) and (d) suggest that consumer interest around energy saving affects the seasonal dynamic demand characteristics. These results show that the proposed method is an effective approach for representing the situation-dependent impact of variables on demand, while also identifying a limited number of dominant variables necessary to explain annual demand behavior.



(a)



(b)

Fig. 3-12 Variables selected for (a) August and (b) December in Model 4.

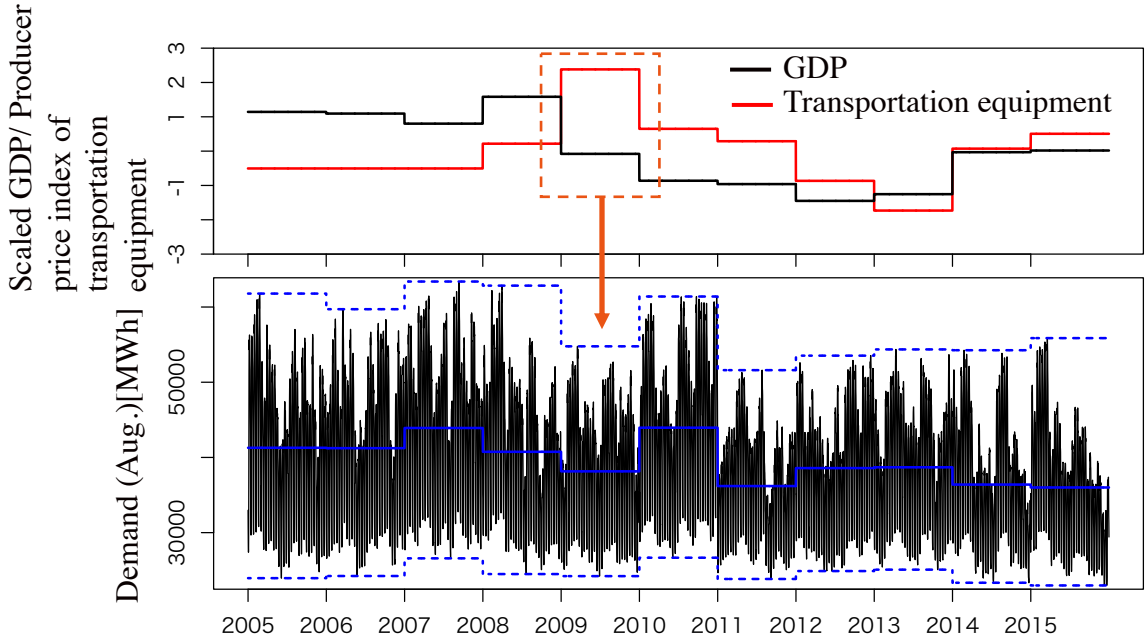


Fig. 3-13 Behavior of v32 (i.e., *Producer price index of transportation equipment*), v212 (i.e., *GDP about changes in inventories*) and hourly electricity demand in August.

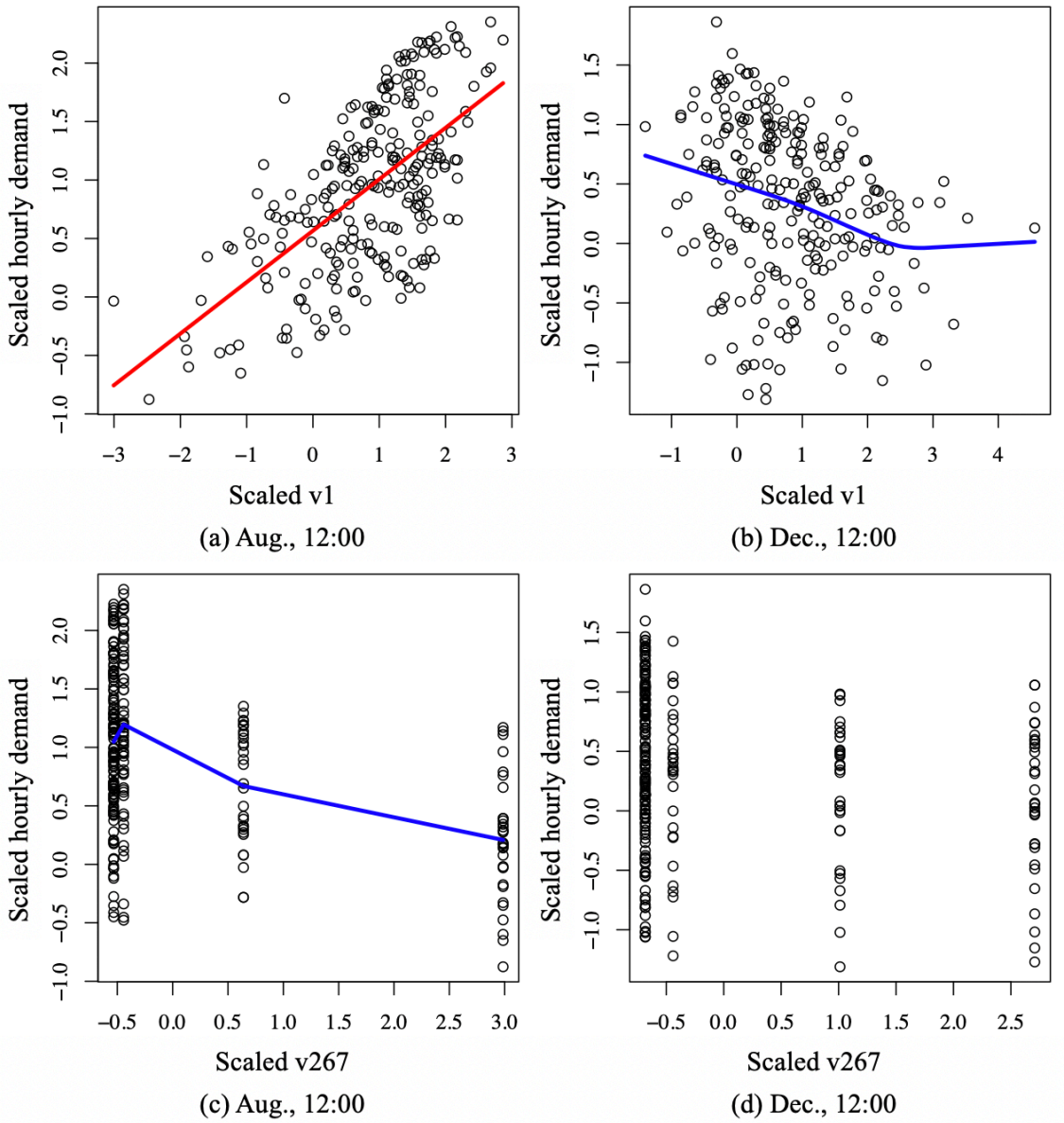


Fig. 3-14 Relationships between the target demand and v1 (*temperature*) and v267 (*the number of internet searches for energy saving*) in (a, c) August and (b, d) December.

3.6 Concluding remarks in the chapter

In this paper, we discussed an approach for identifying annually dominant explanatory variables by constructing situation-dependent models to describe hourly demand behavior. We focused on a sparse modeling technique, specifically sparse partially linear additive models (sparse PLAMs), to discuss the relationship between demand and a large number of variables that may experience seasonal changes. The proposed approach, based on enumerate sparse PLAM and the linear programming technique, successfully identifies a limited number of dominant variables based on the situation-dependent demand modeling. The derived variables are expected to be commonly and consistently related to the demand under various situations and represent situational changes in relation to demand. The key findings in this manuscript are described as follows:

1. The idea of situation-dependent modeling using many possible variables works well to describe the electricity load curve.
2. Hourly demand modeling framework based on a sparse modeling technique achieves to select linear/nonlinear relationships flexibly and the different informative variable set under various seasonal situations.
3. The proposed enumeration and selection approach contributes to identify a limited number of annually dominant variables based on situation-dependent modeling results.

Reference

[3-1] L. Hunt, G. Judge, and Y. Ninomiya, "Underlying Trends and Seasonality in UK Energy Demand: a Sectoral Analysis." *Energy Economics*, vol. 25, no. 1, 2003, pp. 93–118, doi:10.1016/S0140-9883(02)00072-5.

[3-2] S. Cho, K. Tanaka, J. Wu, R. Robert, and T. Kim "Effects of Nuclear Power Plant Shutdowns on Electricity Consumption and Greenhouse Gas Emissions after the Tohoku Earthquake." *Energy Economics*, vol. 55, 2016, pp. 223–33, doi:10.1016/j.eneco.2016.01.014.

[3-3] S. Khuntia, J. Rueda, and M. van Der Meijden "Forecasting the Load of Electrical Power Systems in Mid- and Long-Term Horizons: a Review." *IET Generation Transmission & Distribution*, vol. 10, no. 16, 2016, pp. 3971–77, doi:10.1049/iet-gtd.2016.0340.

[3-4] T. Fujimi and S. Chang "Adaptation to Electricity Crisis: Businesses in the 2011 Great East Japan Triple Disaster." *Energy Policy*, vol. 68, 2014, pp. 447–57, doi:10.1016/j.enpol.2013.12.019.

[3-5] O. Kimura and K, Nishio "Responding to Electricity Shortfalls Electricity-Saving Activities of Households and Firms in Japan after Fukushima." *Economics of Energy & Environmental Policy*, vol. 5, no. 1, 2016, pp. 51–71, doi:10.5547/2160-5890.5.1.okim.

[3-6] K. Honjo, H. Shiraki, and S. Ashina "Dynamic Linear Modeling of Monthly Electricity Demand in Japan: Time Variation of Electricity Conservation effect.(Research Article)." *PLoS ONE*, vol. 13, no. 4, 2018, pp. 1-23 , doi:10.1371/journal.pone.0196331.

[3-7] H. Ching-Lai, S. Watson, and S. Majithia "Analyzing the Impact of Weather Variables on Monthly Electricity Demand." *IEEE Transactions on Power Systems*, vol. 20, no. 4, 2005, pp. 2078–85, doi:10.1109/TPWRS.2005.857397.

[3-8] S. Zhang, D. Gong, and J. Ma "A Study on the Electric Power Load of Beijing and Its Relationships with Meteorological Factors During Summer and Winter." *Meteorological Applications*, vol. 21, no. 2, 2014, pp. 141–48, doi:10.1002/met.1313.

[3-9] D. Vu, K. Muttaqi and A. Agalgaonkar "A Variance Inflation Factor and Backward Elimination Based Robust Regression Model for Forecasting Monthly Electricity Demand Using Climatic Variables." *Applied Energy*, vol. 140, 2015, pp. 385–94, doi:10.1016/j.apenergy.2014.12.011.

[3-10] V. Dordonnat, S. Koopman, M. Ooms, A. Dessertaine, and J. Collet "An Hourly Periodic State Space Model for Modelling French National Electricity Load." *International Journal of Forecasting*, vol. 24, no. 4, 2008, pp. 566–87, doi:10.1016/j.ijforecast.2008.08.010.

[3-11] Y. Chang, C. Kim, J. Miller, J. Park, and S. Park "Time-Varying Long-Run Income and Output Elasticities of Electricity Demand with an Application to Korea." *Energy Economics*, vol. 46, 2014, pp. 334–47, doi:10.1016/j.eneco.2014.10.003.

[3-12] P. Filzmoser, M. Gschwandtner, and V. Todorov "Review of Sparse Methods in Regression and Classification with Application to Chemometrics." *Journal of Chemometrics*, vol. 26, no. 3-4, 2012, pp. 42–51, doi:10.1002/cem.1418.

[3-13] D. Moeller, and Springerlink. "Mathematical and Computational Modeling and Simulation: Fundamentals and Case Studies." *Springer Berlin Heidelberg*, 2004, p. 408, doi:10.1007/978-3-642-18709-4.

[3-14] W. Lawless, T. Castelao, and J. Ballas “Virtual Knowledge: Bistable Reality and the Solution of Ill-Defined Problems.” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 1, IEEE, 2000, pp. 119–24, doi:10.1109/5326.827482.
[3-15] Y. Lou, J. Bien, R. Caruana and J. Gehrke “Sparse Partially Linear Additive Models.” *Journal of Computational and Graphical Statistics*, vol. 25, no. 4, 2016, pp. 1126–40, doi:10.1080/10618600.2015.1089775.
[3-16] A. Chouldechova and T. Hastie “Generalized Additive Model Selection.” *arXiv.org*, 2015, http://search.proquest.com/docview/2082690272/.
[3-17] S. Hara and T. Maehara "Enumerate Lasso Solutions for Feature Selection." Proc. *AAAI*, 2017, pp. 1985–91.
[3-18] R. Engle, C. Granger, J. Rice and A. Weiss “Semiparametric Estimates of the Relation Between Weather and Electricity Sales.” Journal of the American Statistical Association, vol. 81, no. 394, 1986, pp. 310–320, doi:10.1080/01621459.1986.10478274.
[3-19] A. Alin “Multicollinearity.” Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 3, 2010, pp. 370–374, doi:10.1002/wics.84.
[3-20] G. Obozinski, L. Jacob and J. Vert “Group Lasso with Overlaps: The Latent Group Lasso Approach.” arXiv.org, Cornell University Library, 2011, http://search.proquest.com/docview/2086889347/.
[3-21] D. W. Marquardt, “Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation,” Technometrics, vol. 12, no. 3, p. 591, Aug. 1970, doi: 10.2307/1267205.
[3-22] M. Egerstedt, Control theoretic splines optimal control, statistics, and path planning, Course Book. Princeton: Princeton University Press, 2010.
[3-23] S. Mullainathan and J. Spiess “Machine Learning: An Applied Econometric Approach.” Journal of Economic Perspectives, vol. 31, no. 2, 2017, pp. 87–106, doi:10.1257/jep.31.2.87.
[3-24] H. Zou, "The Adaptive Lasso and Its Oracle Properties," Journal of the American Statistical Association, vol. 101, no. 476, 2006, p. 1418–1429, https://doi.org/10.1198/016214506000000735
[3-25] S. Wang, S. Rosset and J. Zhu "Random Lasso," arXiv.org, vol. 51, no. 1, 2011, p. 468–485, https://doi.org/10.1214/10-AOAS377
[3-26] Y. Fujimoto, S. Murakami, N. Kaneko, H. Fuchikami, T. Hattori and Y. Hayashi “Machine Learning Approach for Graphical Model-Based Analysis of Energy-Aware Growth Control in Plant Factories.” IEEE Access, vol. 7, 2019, pp. 32183–96, doi:10.1109/ACCESS.2019.2903830.
[3-27] A. Chung “Linear Programming. ” C. E. Merrill Books, 1963.
[3-28] R. Picard and R. Cook “Cross-Validation of Regression Models.” Journal of the American Statistical Association, vol. 79, no. 387, 1984, pp. 575–83, doi:10.1080/01621459.1984.10478083.
[3-29] "Search for Past Weather Data (in Japanese)," Meteorological Agency, [Online]. Available: http://www.data.jma.go.jp/obd/stats/etrn/index.php. [Accessed 1 September 2019].
[3-30] "Nikkei Stock Average (in Japanese)," Nikkei Indexes, [Online]. Available: https://indexes.nikkei.co.jp/nkave. [Accessed 1 September 2019].
[3-31] "Call market related statistics (in Japanese)," Bank of Japan, [Online]. Available: https://www3.boj.or.jp/market/jp/menu_m.htm. [Accessed 1 September 2019].

[3-32] "JBA TIBOR rate history (in Japanese)," JBA TIBOR Administration, [Online]. Available: http://www.jbatibor.or.jp/rate/log.html. [Accessed 1 September 2019].
[3-33] "Government interest rate information (in Japanese)," Ministry of Finance, [Online]. Available: https://www.mof.go.jp/jgbs/reference/interest_rate/. [Accessed 1 September 2019].
[3-34] "Analysis of All Industrial Activities (in Japanese)," Ministry of Economy, Trade and Industry, [Online]. Available: http://www.meti.go.jp/statistics/tyo/sanzi/result-1.html. [Accessed 15 November 2019].
[3-35] "Industrial index (in Japanese)," Ministry of Economy, Trade and Industry, [Online]. Available: http://www.meti.go.jp/statistics/tyo/iip/index.html. [Accessed 15 November 2019].
[3-36] "National Accounts Calculation (in Japanese)," Cabinet Office, Government of Japan, [Online]. Available: https://www.esri.cao.go.jp/en/sna/menu.html. [Accessed 15 November 2019].
[3-37] "Consumer Price Index (in Japanese)," Statistic Japan, [Online]. Available: https://www.stat.go.jp/data/cpi/index.html. [Accessed 15 November 2019].
[3-38] "Statistics (in Japan)," Bank of Japan, [Online]. Available: http://www.boj.or.jp/statistics/pi/cgpi_release/index.htm/. [Accessed 1 September 2019].
[3-39] "List of statistical tables (Ministry of Economy, Trade and Industry production dynamic statistics) (in Japanese)," Ministry of Economy, Trade and Industry, [Online]. Available: https://www.meti.go.jp/statistics/tyo/seidou/result/ichiran/08_seidou.html. [Accessed 15 November 2019].
[3-40] "Consumer Confidence Survey (in Japanese)," Cabinet Office, [Online]. Available: https://www.esri.cao.go.jp/jp/stat/shouhi/shouhi.html. [Accessed 25 November 2019].
[3-41] "Google Trend," Google, [Online]. Available: https://trends.google.co.jp/trends/?geo=JP. [Accessed 15 November 2019].
[3-42] A. Al-Garni, S. Zubair and J. Nizami "A Regression Model for Electric-Energy-Consumption Forecasting in Eastern Saudi Arabia." Energy, vol. 19, no. 10, 1994, pp. 1043–49, doi:10.1016/0360-5442(94)90092-2.
[3-43] S. Caston "Forecasting Medium-Term Electricity Demand in a South African Electric Power Supply System." Journal of Energy in Southern Africa, vol. 28, no. 4, 2017, pp. 54–67, doi:10.17159/2413-3051/2017/v28i4a2428.
[3-44] G. Huebner, D. Shipworth, I. Hamilton, Z. Chalabi and T. Oreszczyn "Understanding Electricity Consumption: A Comparative Contribution of Building Factors, Socio-Demographics, Appliances, Behaviours and Attitudes." Applied Energy, vol. 177, 2016, pp. 692–702, doi:10.1016/j.apenergy.2016.04.075.
[3-45] I. Rish and G. Genady "Sparse modeling: theory, algorithms, and applications. " CRC press, 2014.
[3-46] R. Tibshirani "Regression Shrinkage and Selection Via the Lasso." Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, 1996, pp. 267–288, doi:10.1111/j.2517-6161.1996.tb02080.x.
[3-47] Z. Zhang, W. Hong and J. Li "Electric Load Forecasting by Hybrid Self-recurrent Support Vector Regression Model with Variational Mode Decomposition and

Improved Cuckoo Search Algorithm." IEEE Access, vol. 8, 2020, pp. 14642-14658, doi: 10.1109/ACCESS.2020.2966712.
[3-48] Y. Dong, Z. Zhang and H. Wei-Chiang "A Hybrid Seasonal Mechanism with a Chaotic Cuckoo Search Algorithm with a Support Vector Regression Model for Electric Load Forecasting." Energies, vol. 11, no. 4, 2018, doi: https://doi.org/10.3390/en11041009.
[3-49] S. Holm "A Simple Sequentially Rejective Multiple Test Procedure." Scandinavian Journal of Statistics, vol. 6, no. 2, 1979, pp. 65–70

Appendix

A data set utilized in this study is composed of 274 variables.

Table 3-A1 Details of all variables utilized in this study.

Category	Index	Item	Source
Weather	v1	Temperature	Japan Meteorological Agency
	v2	Temperature fluctuation	
	v3	Highest temperature	
	v4	Lowest temperature	
	v5	Maximum temperature on the previous day	
	v6	Lowest temperature on the previous day	
	v7	Solar radiation	
	v8	Sunshine hours	
	v9	Precipitation	
	v10	Wind speed	
Interest rate	v11	Secured overnight call rate	Bank of Japan
	v12	Unsecured overnight call rate	
	v13	Tokyo interbank offered rate	JBA TIBOR Administration
	v14	Yield on long gilts	Ministry of Finance
Stock price	v15	Nikkei share average	Nikkei Indexes-Nikkei 225 Official Site-
	v16	Tokyo stock market (spot price)	Bank of Japan
Indices of tertiary industry activity	v17	Production and distribution of gas	Ministry of Economy, Trade and Industry
	v18	Heat supply	
	v19	Collection, purification and distribution of water, and sewage collection, processing and disposal	
	v20	Fixed telecommunications	
	v21	Long-distance telecommunications	
	v22	Services incidental to internet	
	v23	Mobile telecommunications	
	v24	Public broadcasting, except cablecasting	
	v25	Private-sector broadcasting, except cablecasting	
	v26	Computer programming and other software services	
	v27	Custom software services	
	v28	Head offices primarily engaged in managerial operations	
	v29	Miscellaneous data processing and information services	
	v30	Fish aquaculture	
	v31	Motion picture, video and television program distribution	
	v32	Motion picture and video production, except television program and animation production	
	v33	Recording and disk production	
	v34	Radio program production	
	v35	Newspaper publishers	
	v36	Publishers, except newspapers	
	v37	Publishers (weekly magazine)	
	v38	Publishers (monthly magazine)	
	v39	Publishers (book)	
	v40	Railway passenger transportation	
	v41	Railway passenger transportation, jr	
	v42	Railway passenger transportation	
	v43	Rail freight forwarding	
	v44	Bus business	
	v45	Taxi business	
	v46	Common motor trucking, except special group cargo motor trucking	
	v47	Home delivery	
	v48	Water transport	
	v49	Water freight transportation	
	v50	Ocean freight shipping	

Category	Index	Item	Source
Indices of tertiary industry activity	v51	Coastal cargo shipping	Ministry of Economy, Trade and Industry
	v52	Air passenger transportation	
	v53	International air passenger transportation	
	v54	Domestic air passenger transportation	
	v55	International air freight forwarding	
	v56	Domestic air freight transportation	
	v57	Ordinary warehouse	
	v58	Refrigerated warehousing	
	v59	Port transport	
	v60	Packing and crating	
	v61	Transport facilities services	
	v62	Fixed facilities for road transport	
	v63	Airports and air fields, heliports	
	v64	Postal services, including mail delivery	
	v65	Road passenger transport	
	v66	Road freight transport	
	v67	Establishments engaged in administrative or ancillary economic activities	
	v68	Lode gold ore and silver ore mining	
	v69	Iron ore mining	
	v70	Coal mining, including cleaning and grading	
	v71	Lignite mining	
	v72	Crude petroleum production	
	v73	Natural gas production	
	v74	Crude petroleum and natural gas production	
	v75	Granite and related rock and stone quarrying	
	v76	Liparite and related rock and stone quarrying	
	v77	Andesite and related rock and stone quarrying	
	v78	Miscellaneous stone quarrying, sand, gravel and cobble-stone pits	
	v79	Fire clay mining	
	v80	Pyrophyllite (agalmatolite) mining	
	v81	Miscellaneous wholesale trade, n.e.c.	
	v82	Financial intermediary	
	v83	Financial settlement	
	v84	Bill exchange height	
	v85	Boj current account settlement amount	
	v86	Total silver system transaction volume	
	v87	Foreign exchange yen settlement exchange volume	
	v88	Building work, except wooden building work	
	v89	Credit card businesses	
	v90	Sales credit	
	v91	Consumer finance business, credit card	
	v92	Wooden building work	
	v93	Issue	
	v94	Distribution	
	v95	Life insurance institutions	
	v96	Non-life insurance institutions	
	v97	Industrial machinery leasing	
	v98	Medical machine leasing	
	v99	Machine tool leasing	
	v100	Commercial service lease	

Category	Index	Item	Source
Indices of tertiary industry activity	v101	Industrial machinery leasing	Ministry of Economy, Trade and Industry
	v102	Office machine leasing	
	v103	Information-Related equipment leasing	
	v104	Other office machine leasing	
	v105	Other leases	
	v106	Civil engineering and construction equipment rental	
	v107	Information-Related equipment rental	
	v108	Music, video software rental	
	v109	Other rentals	
	v110	Car leasing	
	v111	Car rental	
	v112	Carpentry work	
	v113	Scaffolding work	
	v114	Lawyers' offices	
	v115	Patent attorneys' offices	
	v116	Earth work and concrete work	
	v117	Certified public accountants and certified tax accountants' offices	
	v118	Newspaper advertisement	
	v119	Magazine advertising	
	v120	Tv advertising	
	v121	Radio advertising	
	v122	Other advertising	
	v123	Traffic advertisement	
	v124	Outdoor advertising	
	v125	Insert. direct mail	
	v126	Construction consultant	
	v127	Surveying services	
	v128	Geological survey	
	v129	Tile work	
	v130	Engineering	
	v131	Compound service	
	v132	Waste disposal business	
	v133	Livestock products	
	v134	Guard services	
	v135	Retail trade (woven fabrics, apparel, apparel accessories and notions)	
	v136	Retail trade (food and beverage)	
	v137	Fuller's earth (Japanese acid clay) mining	
	v138	Diatomaceous earth mining	
	v139	Fuel stores	
	v140	Miscellaneous retail trade, n.e.c.	
	v141	General civil engineering work and building work	
	v142	Real estate agencies	
	v143	Sales agents of buildings and houses	
	v144	And land sub dividers and developers	
	v145	Detached house buying and selling	
	v146	Condominium sales	
	v147	Land sub dividers and developers	
	v148	Real estate agents and brokers	
	v149	Establishments engaged in administrative or ancillary economic activities (69 real estate lessors and managers)	
	v150	Automobile parking	

Category	Index	Item	Source
Indices of tertiary industry activity	v151	Piping work, except water-well drilling work	Ministry of Economy, Trade and Industry
	v152	Air conditioning and heating equipment installation work	
	v153	Dental clinics	
	v154	Home care service	
	v155	Facility care service	
	v156	Hotels	
	v157	Hotels	
	v158	Eating places, except specialty restaurants	
	v159	Drinking houses and beer halls	
	v160	Coffee shops	
	v161	Fast food restaurant and restaurant service industry	
	v162	Flooring work	
	v163	Barbershops	
	v164	Hair-Dressing and beauty salon	
	v165	Public bathhouses	
	v166	Miscellaneous construction work by specialist contractor	
	v167	Domestic travel	
	v168	Overseas trip	
	v169	Foreigner trip	
	v170	Funeral services	
	v171	Wedding ceremony hall services	
	v172	Film developing and finishing	
	v173	Cinemas	
	v174	Professional sports.	
	v175	Boxing	
	v176	Professional baseball	
	v177	Football	
	v178	Golf	
	v179	Bicycle race track operations	
	v180	Horse race track operations	
	v181	Motorcar and motorboat race track operations	
	v182	Racetrack	
	v183	Golf courses	
	v184	Golf driving ranges	
	v185	Bowling alleys	
	v186	Fitness centers	
	v187	Theme parks	
	v188	Pachinko parlors	
	v189	Supplementary tutorial schools	
	v190	Foreign language instructions	
	v191	General automobile maintenance services	
	v192	Personal service	
	v193	Business service	
	v194	Broad sense personal service	
	v195	Broad non-selective service for individuals	
	v196	Broad sense and personal service	
	v197	Broad-sense business service	
	v198	Tourism related products	
	v199	Wholesale trade. retail	
	v200	Real estate business	

Category	Index	Item	Source
Indices of tertiary industry activity	v201	Academic research, specialization, technical services	Ministry of Economy, Trade and Industry
	v202	Accommodation business.	
	v203	Life-related service industry. entertainment	
	v204	Other service businesses, excluding public affairs, etc.	
	v205	Indices of industrial production	
Gross domestic product (real terms)	v206	Private final consumption expenditure	Cabinet office
	v207	Private residential investment	
	v208	Private non-residential investment	
	v209	Private changes in inventories	
	v210	Government final consumption expenditure	
	v211	Gross domestic fixed capital formation	
	v212	Changes in inventories	
	v213	Exports of goods and services	
	v214	Imports of goods and services	
	v215	Private final consumption expenditure	
Gross domestic product	v216	Private residential investment	
	v217	Private non-residential investment	
	v218	Private changes in inventories	
	v219	Government final consumption expenditure	
	v220	Gross domestic fixed capital formation	
	v221	Changes in inventories	
	v222	Exports of goods and services	
	v223	Imports of goods and services	
Producer price index	v224	Food and drink	Bank of Japan
	v225	Fiber products	
	v226	Wood, wood products	
	v227	Pulp and paper products	
	v228	Chemical products	
	v229	Oil and coal products	
	v230	Plastic products	
	v231	Ceramic industry.	
	v232	Steel	
	v233	Non-ferrous metal	
	v234	Metal products	
	v235	General equipment	
	v236	Production equipment	
	v237	Commercial equipment	
	v238	Electronic components and devices	
	v239	Electrical equipment	
	v240	Information communication equipment	
	v241	Transportation equipment	
	v242	Other industrial products	
	v243	Total average. agriculture, forestry and fishery	
	v244	Total average. mineral products	
	v245	Total average. scrap	
Consumer price index	v246	City gas	Ministry of Internal Affairs and Communications
	v247	Propane gas	
	v248	Other utility expenses	
	v249	Kerosene	
	v250	Water	

Category	Index	Item	Source
Consumer price index	v251	Sewerage charges	Ministry of Internal Affairs and Communications
Producer price index	v252	City gas	Bank of Japan
	v253	Electric power	
Consumer price index	v254	Electric power	Ministry of Internal Affairs and Communications
Electric appliance installation	v255	Home air conditioner (number of units owned)	Cabinet Office
	v256	Commercial air conditioner (number of units owned)	
	v257	Electric refrigerator (in stock)	Ministry of Economy, Trade and Industry
	v258	Electric refrigerator (number of shipments)	
	v259	Electric refrigerator (number of sales)	
	v260	Electric refrigerator (sales amount)	
	v261	Electric refrigerator (number of production)	
	v262	Electric refrigerator (production value)	
Number of internet searches (Power related words)	v263	省エネ ("Energy saving" in Japanese)	Google Trends
	v264	省エネ対策 ("Energy saving measures" in Japanese)	
	v265	省エネルギー ("Energy saving" in Japanese)	
	v266	省エネルギー対策 ("Energy saving measures" in Japanese)	
	v267	節電 ("Energy saving" in Japanese)	
	v268	節電対策 ("Energy saving measures" in Japanese)	
	v269	ピークカット ("Peak cut" in Japanese)	
	v270	温暖化 ("Global warming" in Japanese)	
	v271	温暖化対策 ("Global warming measures" in Japanese)	
	v272	地球温暖化 ("Global warming" in Japanese)	
	v273	地球温暖化対策 ("Global warming measures" in Japanese)	
Calendar	v274	Week day or holiday	

The result of the skewness of each explanatory variables shows that the several variables possess highly skewed distributions. In future works, the model description accuracy may be enhanced by processing of the explanatory variables to alleviate the distorted distribution.

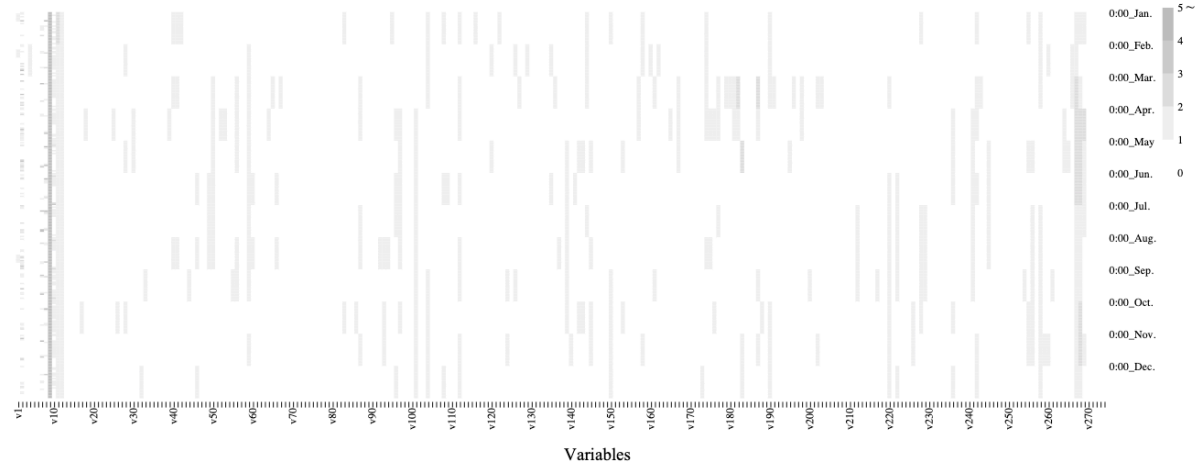


Fig. 3-A1 Skewness of each explanatory variables. The x-axis represents the explanatory variables, and the y-axis represents the transition of the situation ($\boldsymbol{m}$, $\boldsymbol{h}$).

The collinearity of each explanatory variable with other variables is examined using the variance inflation factor (VIF). The results show that the explanatory variables exhibit high levels of collinearity with other variables.

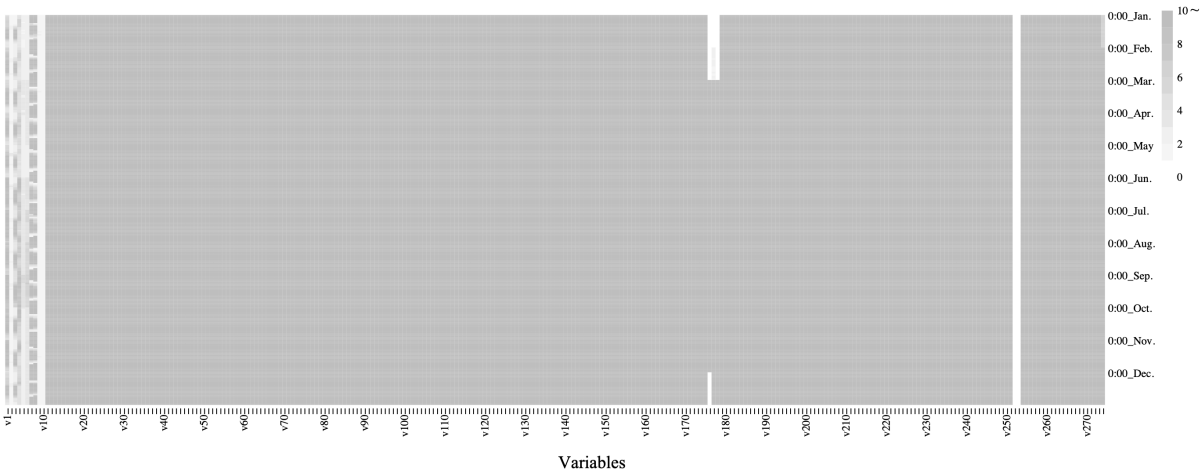


Fig. 3-A2 Collinearity of each explanatory variable with other variables. The x-axis represents the selected variables, and the y-axis represents the transition of the situation ($\boldsymbol{m}$, $\boldsymbol{h}$).

Chapter 4
Model-based analysis for identifying impacts of factors affecting the electricity deviations during COVID-19: A German case study

Symbols

$\mathcal{T}$	Training period of hourly data
T	Number of samples of data per hour
$\mathcal{D}$	Training periods of daily data
D	Number of samples of data per day
$\mathcal{M}$	Training period of monthly data
M	Number of samples of data per month
$\mathcal{Y}$	Training period of yearly data
Y	Number of samples of data per year
y	Index for indicating a specific year
m	Index for indicating a specific month
d	Index for indicating a specific day
h	Index for indicating a specific hour
x_d^j	Explanatory variables observed in day d
x_m^j	Explanatory variables observed in month m
x_y^j	Explanatory variables observed in year y
$h_a(d)$	Indicate functions that derive the dummy variables depending on the month
$h_b(d)$	Indicate functions that derive the dummy variables depending on the day of week
$\boldsymbol{\kappa}$	A set of coefficient parameters of the autoregressive component
$\boldsymbol{\mu}$	A set of coefficient parameters of the moving average component
$\boldsymbol{\nu}, \boldsymbol{\xi}$	Sets of coefficient parameters of exogenous variables
l_{ymdh}	Hourly demand observed at hour h in day d in month m of year y
I	Number of explanatory variables
$\boldsymbol{x}_{ymdh}$	$= \left(x_{ymdh}^1, \dots, x_{ymdh}^I\right)$. Explanatory variables observed at hour h in day d in month m of year y
$\mathcal{S}$	$= \{1, \dots, I\}$. Index set of explanatory variables
$\mathcal{L}$	$\subseteq \mathcal{S}$. Index subset indicating the variables related to the target demand linearly

$\mathcal{N}$	$\subseteq \mathcal{S}$. Index subset indicating the variables related to the target demand nonlinearly
$\boldsymbol{\beta}$	Set of coefficient parameters for explanatory variables
$\phi_j^k(.)$	Cubic spline transformation function
K	Number of bases in cubic spline transformation functions
$\boldsymbol{\tau}_j$	Vector of coefficient parameters for transformation functions $\phi_j^1(x^j), \dots, \phi_j^K(x^j)$
λ	Positive constant penalty for regularization
I_{mh}	Number of enumerated models under the given situation (m, h)
$\mathcal{S}_{mh}^{(i)}$	Index subset of the selected variables corresponding to the non-zero components
$F_{mh}^{(i)}$	Squared-error loss under the variable set $\mathcal{S}_{mh}^{(i)}$
D_{ymdh}	Electricity demand deviation at hour h in day d in month m of year y
$\hat{\mathcal{S}}_{mh}^{deviance}$	Index of the key variables deviating from the scenarios in each seasonal period
$\hat{\mathcal{S}}_{mh}^{expected}$	Index of variables that varies within the range of the expected scenario

Nomenclature

GDP	Gross domestic product
ARIMAX	Autoregressive integrated moving average with an exogenous variable
PLAMs	Partially linear additive models
MICP	Harmonized Indices of Consumer Prices
RMSE	Root mean squared errors

4.1 Introduction

After the COVID-19 pandemic, the economic conditions and lifestyles of consumers have changed, which translates into impacts on their electricity consumption patterns [4-1]. Furthermore, the recent modifications in electric power systems, such as the large-scale penetration of renewable energy and the growing interest in energy efficiency, have aggravated the impact of these factors on the electricity residual demand—defined as the difference between the total electricity consumption of domestic customers and the consumption from renewable energy sources. As such, these variations in electricity demand constitute a global issue, and understanding the impact of factors on electricity demand is essential for the stakeholders of the electricity business to construct an effective strategy in terms of energy facilities planning and reserve power procurement [4-2]. Traditionally, numerous modeling studies have been conducted to analyze the long-term temporal demand, which suggested several attractive approaches to identify the statistical relationships between variables describing the electricity consumption behavior [4-3]–[4-5]. A majority of the early research in this field focused on a limited number of variables that were selected based on prior empirical knowledge. For instance, Ching et al. [4-3] focused on the relationship between monthly electricity demand and factors such as weather and gross domestic product (GDP). As reported, in a situation in which the electric power system is subject to modifications, demand can be influenced by additional factors, including power-saving trends, electricity, or prices of daily life necessities [4-6]. During the COVID-19 pandemic, electricity demand varied drastically depending on several factors that were affected by reduced economic activity and lockdown. In particular, the demand deviation between the actual demand and the forecasted demand derived before the pandemic has caused a serious issue, which causes risks in decision-making based on electricity utilities such as reserve planning and facility design. For instance, Huang et al. [4-1] found that electricity demand in China fluctuated by up to ~12% compared to the planned demand expected prior to the pandemic. Moreover, their findings indicated a relationship between the tendency of demand fluctuations and the pandemic stage. In addition, Chen et al. [4-7] revealed the dependence of electricity demand on consumer mobility data. These studies confirmed the drastically widening deviation in electricity demand caused by fluctuations in several factors, such as economic conditions and consumer lifestyle, during the pandemic. Although the deviation occurring under emergency pandemic situations may be highly uncertain in the long term, electricity utilities must identify the trend of the electricity demand deviation during the pandemic, and accordingly identify the key factors of these deviations to forecast future electricity demand. Nonetheless, the impact of essential variables on electricity demand during COVID-19 has only been briefly investigated, and the respective literature discussing the analysis from an academic perspective is still scarce.

In this study, we focused on the hourly electricity residual demand, constituted by the electricity demand excluding the supply from renewable sources. We proposed an approach to construct a model-based analysis to characterize the influence of relevant factors on the electricity deviation caused during COVID-19. The model of autoregressive integrated moving average with an exogenous variable (ARIMAX) [4-8] provides a scheme for forecasting the long-term scenario and behavior of each exogenous variable. To identify the impact of such variables on the electricity deviation, we utilized a class of partially linear additive models (PLAMs) [4-9], [4-10], which considerably aided in identifying the additive contribution of each variable to describe the target value

considering the linearity/nonlinearity in variables [4-11]. In the proposed approach, we identified the variable exhibiting deviations from the expected scenario and evaluated its impact on demand. The scenario forecasted an electricity demand assuming no pandemic, as it was supposedly derived based on long-term forecasting using datasets created before the pandemic. The deviation of the variables from the scenario may cause an unexpectedly large demand deviation. Overall, we identified the impact of essential variables on the pandemic-influenced demand deviation. The analysis supported the use of electrical utilities to identify variables involving uncertainties under recent emergencies, and their variations should be intensively analyzed to achieve highly accurate demand forecasts.
The major contributions of this paper are stated as follows:

1. We set out to analyze the impact of key variables on the demand deviation, considering several possible variables that have not been adequately discussed in prior studies on demand deviation modeling.
2. We utilize ARIMAX to identify explanatory variables exhibiting deviations during COVID-19 from the long-term variable scenario estimated using pre-pandemic data.
3. We apply the enumerated sparse PLAMs for demand modeling to identify the essential variables and additive contributions for model construction.
4. We propose a procedure to identify the impact of each key variable deviating from the long-term scenario during the pandemic on the demand deviation.
5. We apply the framework to a real-world dataset and discuss the informative factors influencing hourly demand.

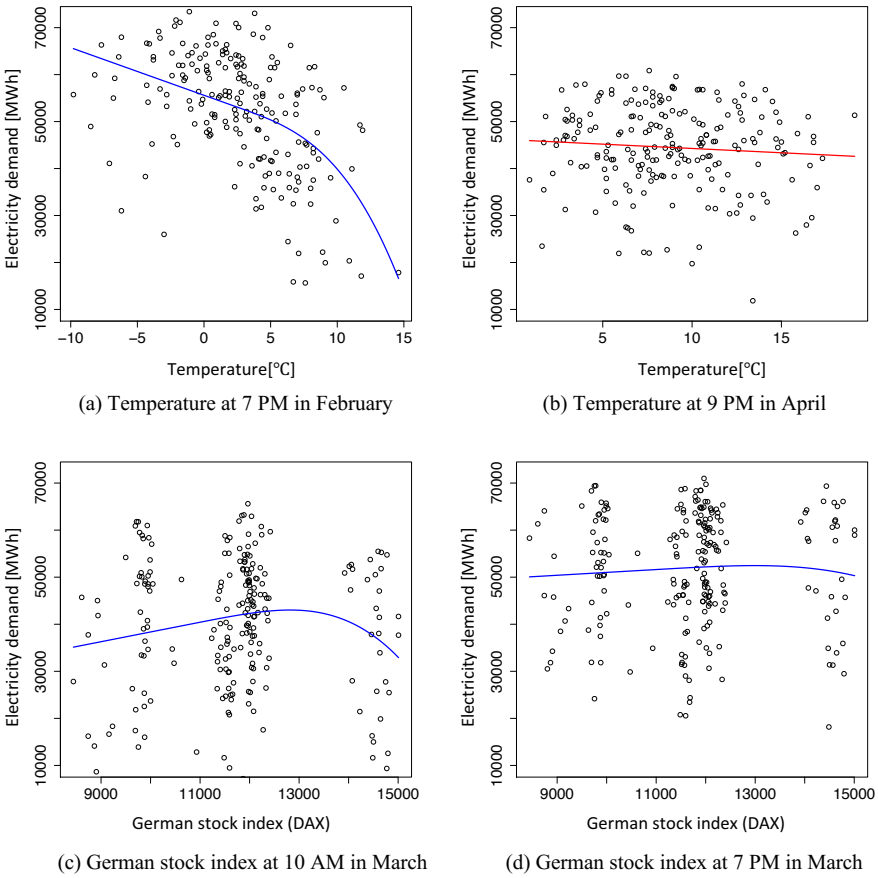
The remainder of this paper is organized as follows. The characteristics of the targeted electricity demand and several explanatory variables, including a review of relevant studies and an overview of the proposed approach, are explained in Section 4.2. The proposed approach to identify the key variables describing demand during the pandemic along with their impact on the electricity deviation is presented in Section 4.3. In Section 4.4, the evaluation results of the proposed framework applied to a real-world dataset in Germany—a nation that was severely affected by the pandemic and equipped with the largest extent of renewable energy resources—are presented along with certain observations regarding the influence of the identified variables on the electricity demand deviation. Finally, the conclusions of this study are discussed in Section 4.5.

4.2 Deviation in electricity demand in Germany

4.2.1 Behavior of electricity demand in Germany

The electricity demand depends on various factors, such as weather conditions, installed capacity of renewable energy generation, economic conditions, and electricity conservation behavior. Herein, we focus on the hourly electricity demand in Germany and analyze the behavior of electricity demand and several explanatory variables affecting it. Figure 4-1 provides examples of the relationship between the electricity demand and other explanatory variables observed simultaneously in certain seasonal scenarios; the solid lines represent curves derived by the sparse PLAM introduced in Section 4.3. This figure indicates the variations in electricity demand depending on the behavior of additional factors. Moreover, the relationships between the variables are not linear in all cases and may vary with seasonal conditions. Thus, we focus on the long-term behavior

of the hourly electricity demand curve. Since 2020, political interventions such as lockdowns and shutdowns during the COVID-19 pandemic have significantly impacted economic conditions, prices, and electricity consumption behavior, resulting in considerable variations in electricity demand. In addition to the hourly electricity demand curve, the yearly and monthly average demand from 2015 to 2021 are illustrated in Fig. 4-2. As observed, the annual electricity demand in Germany has decreased, potentially because of the large-scale domestic penetration of renewable energy sources. Although the average electricity demand decreased in 2020, it observably increased in 2021 after the pandemic. As illustrated in Fig. 4-3, we focused on the variations in the monthly average electricity demand from 2015 to 2021. In Fig. 4-4, the variations in the monthly demand prevailing through 2020 and 2021 are compared with those occurring in 2019. The findings portrayed in these figures reveal that the electricity demand tended to decrease in 2020; the demand decreased by up to 27.8% in April 2020. In contrast, the demand tended to increase in 2021, and in March 2021, it increased significantly by 21.9% compared with the same period in 2019.



(a) Temperature at 7 PM in February

(b) Temperature at 9 PM in April

(c) German stock index at 10 AM in March

(d) German stock index at 7 PM in March

Fig. 4-1 Examples of relationships between electricity demand and several variables across seasonal conditions. Solid lines represent curves derived by sparse PLAM introduced in Section 4.3.

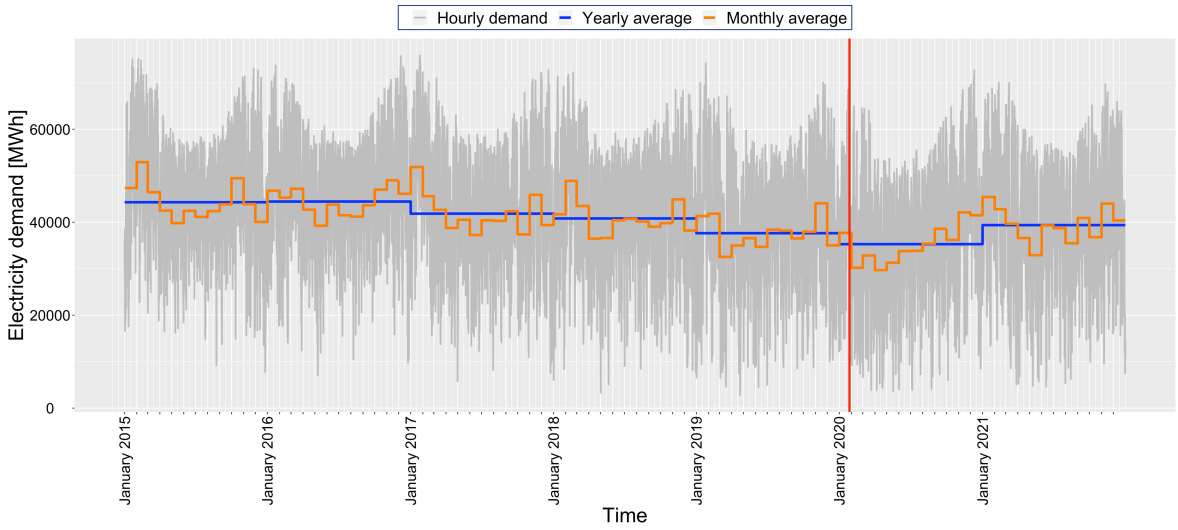


Fig. 4-2 Hourly electricity demand. Red line marks the day on which COVID-19 was first detected.

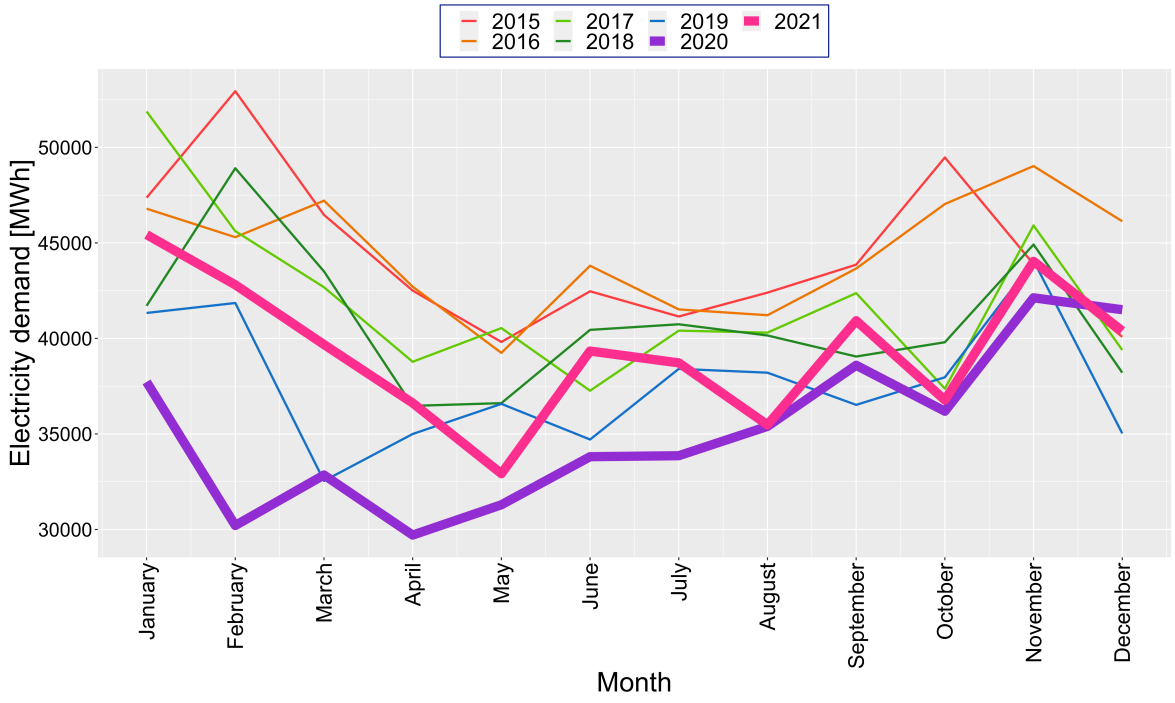


Fig. 4-3 Monthly electricity demand for each year (2015–2021).

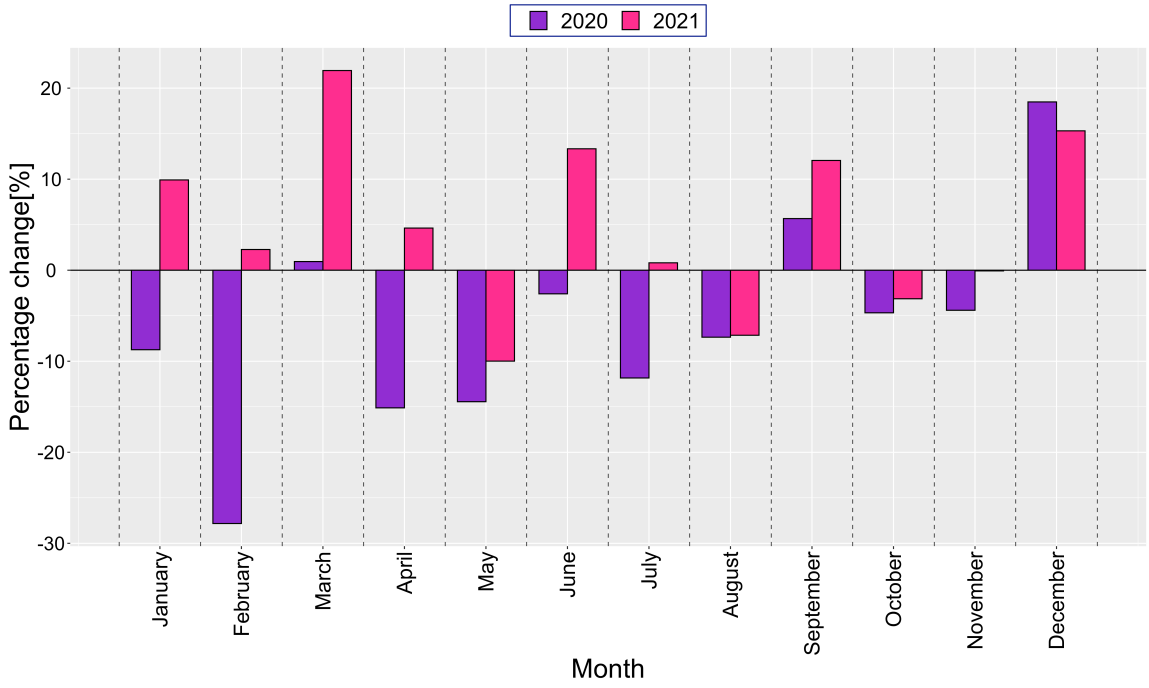


Fig. 4-4 Monthly variations in electricity demand from 2020 to 2021 compared with the same period in 2019.

Furthermore, the average daily demand curves were constructed by focusing on certain months from 2015 to 2016, as depicted in Fig. 4-5. The hourly demand indicates that the timing and magnitude of the peak demands vary across seasons throughout the year. The electricity demand in Germany tends to decrease during the daytime. Specifically, the daytime electricity demand decreased drastically in 2020 and 2021, implying that the variations in daily demand curves caused by the pandemic may vary across seasons and time slots.

In Fig. 4-6, we provide examples of the long-term behavior of the variables influencing electricity demand between 2015 and 2021. The production index of the coal mining industry[4-12] is presented in Fig. 4-6 (a), wherein certain variables remained in the state over the entire span. In contrast, the trends of several variables varied during the pandemic. The behavior of the German stock index [4-13] exemplified in Fig. 4-6 (b) indicates the domestic economic condition of Germany, which drastically deteriorated during the pandemic. The Harmonized Indices of Consumer Prices (HICP) of electricity [4-14] are plotted in Fig. 4-6 (c), which measures the variations in the prices of consumer goods and services acquired by households. Moreover, the Google Trends [4-15] presented in Fig. 4-6 (d) indicate the trends in consumer interest in electricity problems. In these figures, the filled range indicates the variations expected in 2020–2021 compared to the assumptions derived from the data trends through 2019. As observed, the behaviors of several variables altered drastically during the pandemic; the restrictions imposed for preventing infection transmission and the lockdown during the pandemic significantly impacted the economic and consumer lifestyle conditions.

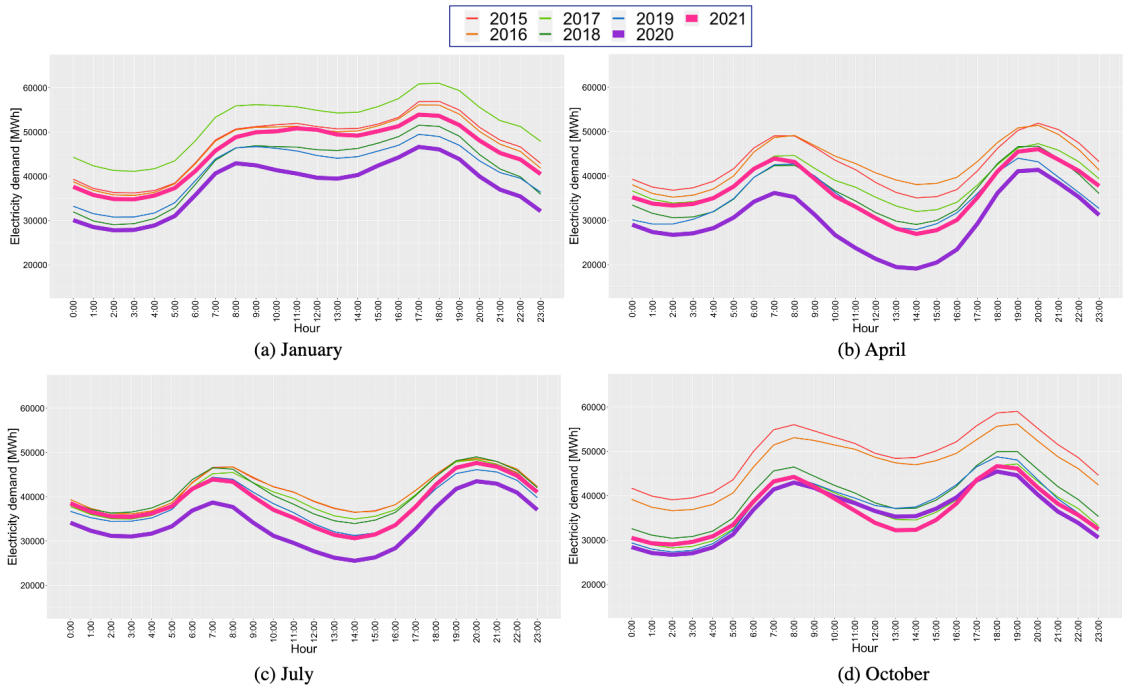
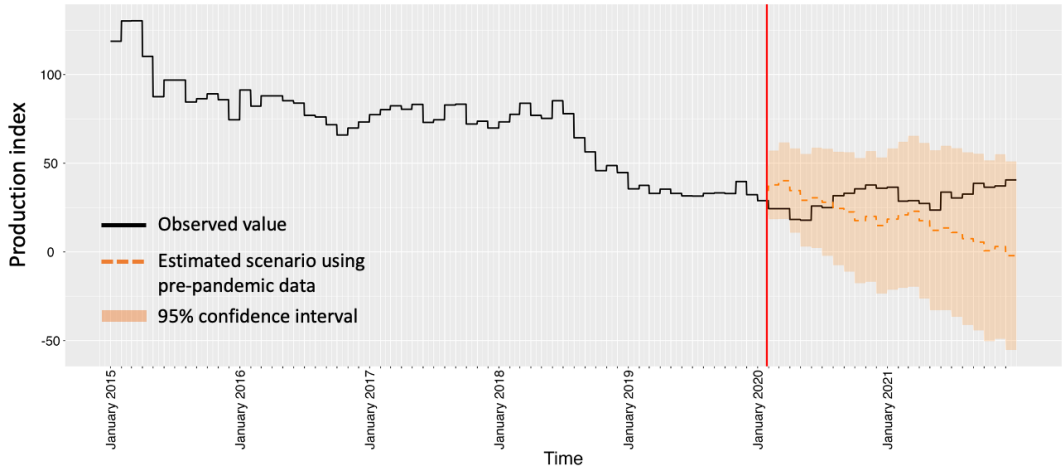
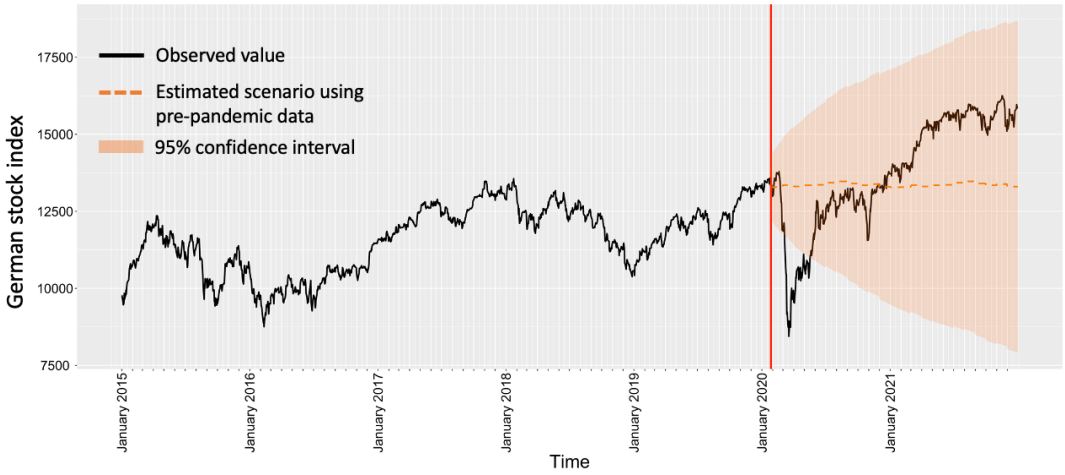


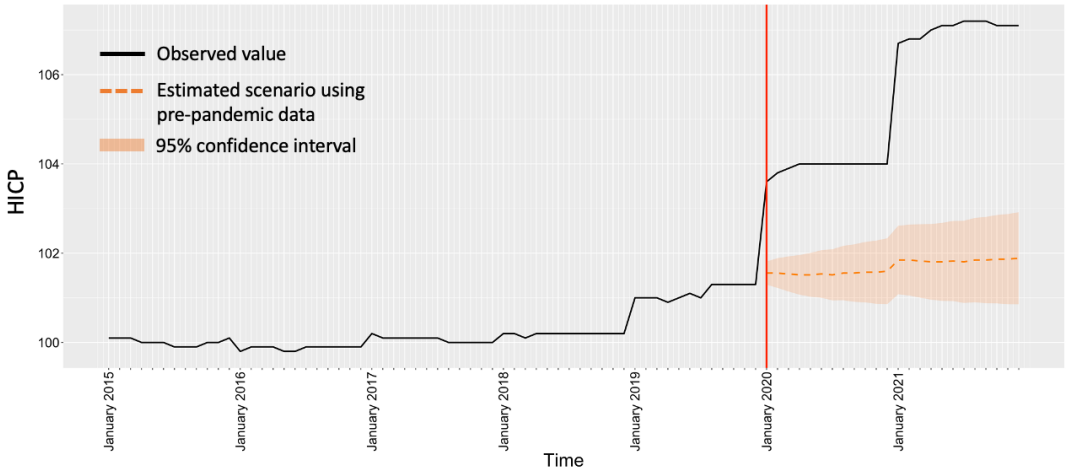
Fig. 4-5 Examples of average hourly demand.



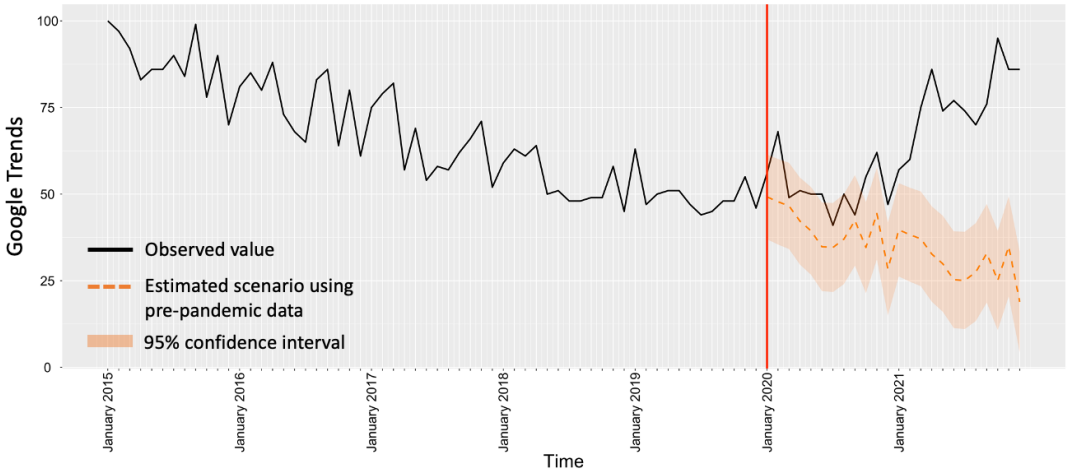
(a) Coal mining



(b) German stock index (DAX)



(c) HICP (Refuse collection)



(d) Google Trends of "energy efficiency"

Fig. 4-6 Behavior of explanatory variables across scenarios before, during, and after the pandemic. The black line indicates the observed variable. The orange line and shaded region indicate the variable scenario and the range of the 95% confidence interval, respectively, which were estimated using pre-pandemic data based on ARIMAX introduced in Section 4.3.

4.2.2 Analysis of deviations in electricity demand during COVID-19 pandemic

This study focuses on the electricity demand deviation during the pandemic and proposes an approach to identify the impact of variable fluctuations caused by the pandemic on electricity demand. In particular, we considered hundreds of explanatory variables, which can be categorized into attribute groups such as weather, stock price, prices of commodities, and calendar data (see the Appendix for the details of the variables considered).

An overview of the proposed approach is presented in Fig. 4-7. The proposed approach aims to select the key variables that deviate from the scenario estimated using pre-pandemic data and to identify the impact of these variables on the demand deviation. First, we derive the scenarios of the long-term behavior of each explanatory variable from 2020 to 2021 considering the dataset acquired before the pandemic. Thereafter, we constructed statistical models to describe the situation-dependent electricity demand. Finally, we forecast the electricity demand based on variable scenarios and observed variables using statistical models to identify the amount of the demand deviation during the pandemic. Notably, the demand deviation is defined as the difference between the model-based demand derived from the set of these scenarios and the model-based demand derived from the realizations of these variables. Additionally, the demand deviation can be classified into two components: the deviance-oriented deviation caused by the fluctuations of variables acting differently from scenarios during the pandemic and the expected deviation caused by variables acting according to the expected variable scenarios. In particular, we identify the behavior of the deviance-oriented deviation and analyze the additive contribution of each variable to identify its impact on demand. The analysis was conducted considering various situations characterized based on yearly, monthly, and daily time periods.

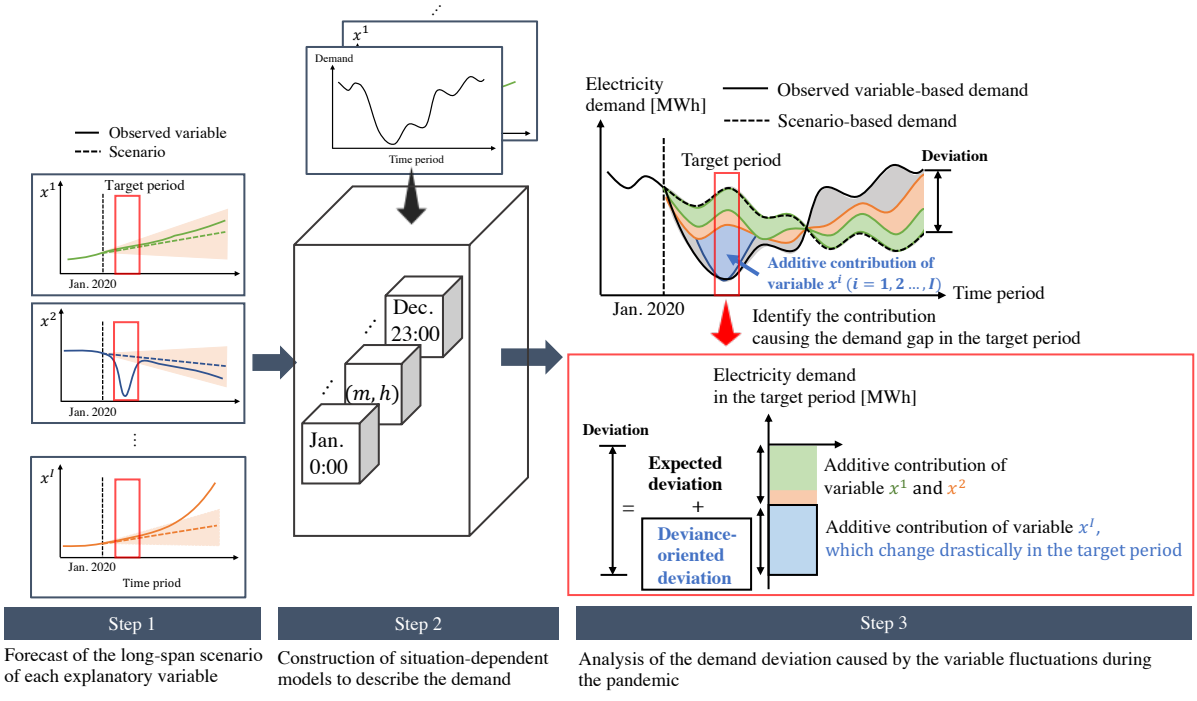


Fig. 4-7 Overview of electricity demand deviation. Scenario of variables and constructed models derived using enumerate sparse PLAMs and ARIMAX (Section 4.3).

The proposed approach comprises the following three steps: forecasting the scenario of each explanatory variable using pre-pandemic data, constructing a statistical model based on explanatory variables, and determining the impact of a key explanatory variable on the pandemic-influenced demand deviation based on the deviations between the scenarios and realizations for each variable. In Step 1, the scenarios of the variables were derived to discuss their behavior during the pandemic. The ARIMAX model provides an approach to estimate the long-term behavior of time-series explanatory data [4-16]–[4-18]. In general, ARIMA is a popular approach for describing the relationships between target demand and historical demand. ARIMAX enables a model of demand behavior by considering the relationships between the target demand and exogenous explanatory variables. To construct the demand models in Step 2, PLAM [4-10] forms an attractive class to describe a target variable, and the fundamental effect of the explanatory variables is discussed, excluding the complicated effect of the interactions between variables. PLAM constructs a statistical model by incorporating the advantages of linear and nonlinear models; it applies a flexible and interpretable approach that can develop statistical models and describe the demand, thereby enabling either a linear or nonlinear effect on the target demand. The linear regression model is one of this class that can describe the linear relationships between the target demand and possible influential factors, and it has been widely used for analyzing the impact of variables on demand owing to the simplicity of its assumptions [4-4]. However, recent studies have explicitly discussed the nonlinear relationships between variables. The additive regression model is involved in PLAMs, and the model can improve the description accuracy by considering the nonlinear relationships between the target demand and each explanatory variable [4-19]. Model selection, i.e., the selection of explanatory variables that describe the demand and aid in the identification of the linearity/nonlinearity between the variables, is essential for constructing an appropriate model.

The sparse enumeration technique constructs models by selecting the essential importance and identifying the linearity/nonlinearity between variables. The sparse modeling technique aids in the automatic identification of plausible linear or nonlinear relationships between the variables, and it selects the key variables among numerous explanatory variables. Kaneko et al. [4-6] introduced the enumerate sparse PLAM, which can identify the limited number of annually dominant variables. This study focuses on enumerate sparse PLAMs and proposes an approach to identify the deviance-oriented deviation and expected deviation, including the additive contributions of each variable to describe the demand deviation.

4.2.3 Related researches on the electricity demand model construction

Several researchers and real-world system operators have attempted to develop models for the analysis of variables relevant to electricity demand. In most existing studies, a limited number of variables, such as economic indicators (e.g. GDP) and public and national holidays, have been derived from the empirical knowledge of experts and analyzed from the perspective of demand-influencing factors [4-3], [4-4]. However, as power systems are complex, an analysis considering numerous variables such as technical, demographic, climatic, economic, and psychological factors [4-2] revealed that the dominant factors influencing the temporal variations in electric power demand may vary with the situation and target dynamics [4-6]. Previous studies have discussed the assumptions regarding the description of the relationships between target demand and possible influential factors. Although Al-Garni et al. [4-4] simply assumed linear

relationships between variables, which might not be generalized to any given dataset, statistical models based on such linearity assumptions have been widely utilized for analyzing the predominant effects of each variable on demand. Recently, studies have explicitly reported on the nonlinear relationships between variables. For instance, Huang et al. [4-20] proposed a nonparametric and nonlinear modeling approach for describing target demand.

These studies provide attractive approaches for forecasting long-term electricity demand behavior. However, in the recent emergency situation during the pandemic, the trend of electricity demand has drastically varied, and the demand has seriously deviated from the expected scenario. Therefore, several studies have been conducted to analyze the electricity demand deviation and the factors impacting recent demand during the pandemic [4-7], [4-21]–[4-25]. For instance, Huang et al. [4-1] focused on the monthly electricity demand in China, analyzed the deviation between the actual demand and expected demand, and suggested that the demand behavior tends to vary depending on the number of confirmed infection cases. Ceylan et al. [4-22] focused on daily electricity demand and discussed the relationship between the implementation of lockdown and demand. These studies reported that the demand varies based on the infection situation. Furthermore, Chen et al. [4-7] proposed a daily demand modeling approach considering mobility data as a variable that reflects the condition of economic activity. These studies reported that drastic variations in economic activity and consumer lifestyle may significantly impact electricity demand during the pandemic. However, the specific factors affecting the pandemic-derived demand deviation have not been thoroughly analyzed in prior related studies.

In particular, the proposed approach focuses on numerous variables affecting the electricity demand deviation during the pandemic and aims to identify key variables describing the deviation, which were selected differently in each seasonal condition based on the linearity/nonlinearity between the variables. Overall, the presented approach aims to identify the key variables that fluctuated considerably during the pandemic and impacted the demand deviation. This will support electric utilities in discussing recent demand behavior and forecasting future demand.

4.3 Analyzing the deviation in electricity demand

This section details the proposed statistical modeling approach (Fig. 4-7) to identify the impact of drastic variable fluctuations on the electricity demand deviation. In Step 1, we introduce the ARIMAX model to estimate the behavior of explanatory variables during the pandemic using pre-pandemic data. In Step 2, we target the electricity demand and corresponding explanatory variables to construct PLAMs for identifying the key variables that describe the hourly electricity demand. In particular, the statistical models and key variables are derived for each seasonal scenario. Based on the constructed PLAMs, the scenario- and observed variable-based demands are forecasted in Step 3 by using the projected scenarios and observations of explanatory variables, respectively.

4.3.1 Approach for forecasting scenario of explanatory variables

For each variable that is observed to have fluctuated during the pandemic, the long-term scenario is forecasted in Step 1, and their deviation from the projected scenario is discussed. An overview of the estimation of the variable scenario is illustrated in Fig. 4-8. Based on the ARIMAX models, we estimate the long-term variable scenarios using the

dataset before the pandemic, that is, a forecasted variable and its confidence interval. Specifically, we focus on the deviation between the scenarios and observed variables to identify variables that deviate from the estimated scenarios during the pandemic. Notably, the variables targeted for scenario forecasting, which are apparently affected by the pandemic, are observed for each day, month, or year (refer to the Appendix for details on the observation granularity of each explanatory variable). An overview of the observation granularity for each variable is illustrated in Fig. 4-9. Let $\mathcal{T} = (1, \dots, T)$ denote a training period and T indicate the number of data samples per hour. The training period for the acquired daily data, e.g., stock price, is defined as $\mathcal{D} = \{1, \dots, D\}$, where D denotes the number of samples. Similarly, the training periods of monthly data, e.g., the index of production in service, are defined as $\mathcal{M} = \{1, \dots, M\}$,, and the training periods of the yearly data, for example, the amount of power generation, are defined as $\mathcal{Y} = \{1, \dots, Y\}$.

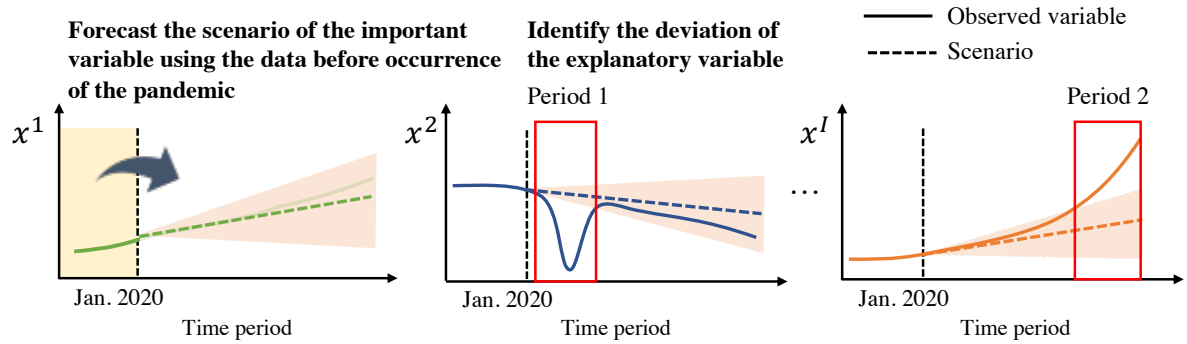
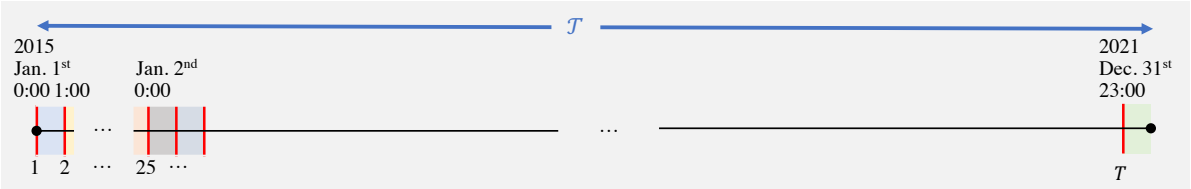
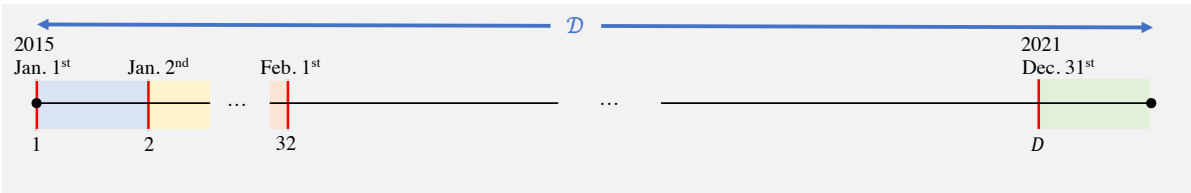


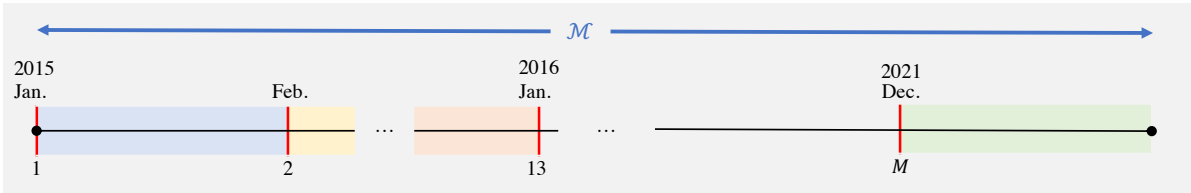
Fig. 4-8 Overview of estimation of long-term scenario of each explanatory variable.



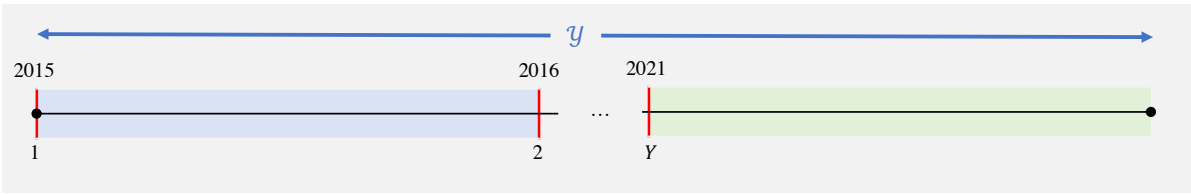
(a) Hourly data



(b) Daily data



(c) Monthly data



(d) Yearly data

Fig. 4-9 Overview of observation granularity for each variable.

If we focus on the daily variable x_d^j $(d \in \mathcal{D})$, the ARIMAX (P, O, Q) model represents a formulation for describing the target variable with the historical variable and several exogenous variables indicating the information of the month and day of the week, defined as follows:

$$x^{*j}_d = \sum_{p=1}^{P} \kappa_p \, x^{*j}_{d-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{d-q} + \sum_{a \in \mathcal{A}} \nu_a \, h_a(d) + \sum_{b \in \mathcal{B}} \xi_b \, h_b(d) + \varepsilon_d, \quad (4.1)$$

$$x^{*j}_d = \Delta^O x^j_d, \quad (4.2)$$

$$\varepsilon_d \sim N(0, \sigma_d^2), \quad (4.3)$$

where $h_a(d)$ and $h_b(d)$ indicate functions that derive the dummy variables depending on the month $\mathcal{A} = (January, \dots, December)$ and the day of week $\mathcal{B} = (Monday, \dots, Sunday)$ in the target period as follows:

$$h_a(d) = \begin{cases} 1 \text{ if the target period } d \text{ is in the target month } a \in \mathcal{A} \\ 0 \text{ if the target period } d \text{ is not in the target month } a \in \mathcal{A} \end{cases}, \quad (4.4)$$

$$h_b(d) = \begin{cases} 1 \text{ if the target period } d \text{ is in the target day of the week } b \in \mathcal{B} \\ 0 \text{ if the target period } d \text{ is not in the target day of the week } b \in \mathcal{B} \end{cases}. \quad (4.5)$$

Additionally, $\Delta^O x^j_d$ denotes the O-order differentiating time series, σ_d^2 denotes the variance of white noise ε_d, and $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_P)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_Q)$, $\boldsymbol{\nu} = (\nu_{January}, \dots, \nu_{December})$, and $\boldsymbol{\xi} = (\xi_{Monday}, \dots, \xi_{Sunday})$ represent the sets of model coefficient parameters; $\boldsymbol{\kappa}$ indicates a set of coefficient parameters of the autoregressive component; $\boldsymbol{\mu}$ denotes a set of coefficient parameters of the moving average component; and $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ represent sets of coefficient parameters of exogenous variables.

Here, the parameters $\boldsymbol{\kappa}$, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, and $\boldsymbol{\xi}$ are estimated based on maximum likelihood estimation (MLE), which is similar to the least-squares estimates as follows [4-26]:

$$\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\xi}}{\operatorname{argmin}} \sum_{d \in \mathcal{D}} \left(x^{*j}_d - \sum_{p=1}^{P} \kappa_p \, x^{*j}_{d-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{d-q} + \sum_{a \in \mathcal{A}} \nu_a \, h_a(d) + \sum_{b \in \mathcal{B}} \xi_b \, h_b(d) \right)^2, \quad (4.6)$$

where σ_d is derived from the standard deviation of the difference between the right- and left-hand sides of Eq. (1) based on the estimated parameters, expressed as follows:

$$\sigma_d = \mathrm{sd}\left(x^{*j}_d - \sum_{p=1}^{P} \kappa_p \, x^{*j}_{d-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{d-q} + \sum_{a \in \mathcal{A}} \nu_a \, h_a(d) + \sum_{b \in \mathcal{B}} \xi_b \, h_b(d) \right). \quad (4.7)$$

The value of $\hat{x}^j_d$ $(d > D)$ can be predicted based on the estimated parameters as a long-term scenario during the pandemic, and the 95% confidence interval can be determined based on the difference between $\hat{x}^j_{d,upper}$ and $\hat{x}^j_{d,lower}$ [4-27]. The confidence intervals are estimated as follows:

$$\hat{x}^j_{d,lower} \leq \hat{x}^j_d \leq \hat{x}^j_{d,upper}, \quad (4.8)$$

$$\hat{x}^j_{d,lower} = \hat{x}^j_d - 1.96\sigma_d, \quad (4.9)$$

$$\hat{x}^j_{d,upper} = \hat{x}^j_d + 1.96\sigma_d. \quad (4.10)$$

The deviating variables are identified based on the deviation of the actual variable x^j_d

from the range of the confidence interval as follows:

- $\hat{x}^j_{d,lower} \leq x^j_d \leq \hat{x}^j_{d,upper}$: variable j varies across the range of scenarios.
- $\hat{x}^j_{d,lower} > x^j_d \cup \hat{x}^j_{d,upper} < x^j_d$: Variable j deviates from the scenario during the pandemic.

Similarly, the ARIMAX (P, O, Q) model for describing the monthly variable x^j_m $(m \in \mathcal{M})$ is defined as follows:

$$x^{*j}_m = \sum_{p=1}^{P} \kappa_p \, x^{*j}_{m-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{m-q} + \sum_{a \in \mathcal{A}} \nu_a \, h_a(m) + \varepsilon_m, \tag{4.11}$$

$$\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\xi}}{\operatorname{argmin}} \sum_{m \in \mathcal{M}} \left(x^{*j}_m - \sum_{p=1}^{P} \kappa_p \, x^{*j}_{m-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{m-q} + \sum_{a \in \mathcal{A}} \nu_a \, h_a(m) \right)^2. \tag{4.12}$$

The scenarios during the pandemic are estimated based on $\hat{x}^j_m$ $(m > M)$ as follows:

$$\hat{x}^j_{m,lower} \leq \hat{x}^j_m \leq \hat{x}^j_{m,upper}, \tag{4.13}$$

$$\hat{x}^j_{m,lower} = \hat{x}^j_m - 1.96\sigma_m, \tag{4.14}$$

$$\hat{x}^j_{m,upper} = \hat{x}^j_m + 1.96\sigma_m. \tag{4.15}$$

The ARIMAX (P, O, Q) model for describing the yearly variable x^j_y $(y \in \mathcal{Y})$ is defined as follows:

$$x^{*j}_y = \sum_{p=1}^{P} \kappa_p \, x^{*j}_{y-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{y-q} + \varepsilon_y, \tag{4.16}$$

$$\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\xi}}{\operatorname{argmin}} \sum_{y \in \mathcal{Y}} \left(x^{*j}_y - \sum_{p=1}^{P} \kappa_p \, x^{*j}_{y-p} + \sum_{q=1}^{Q} \mu_q \, \varepsilon_{y-q} \right)^2. \tag{4.17}$$

The scenarios during the pandemic are further estimated based on $\hat{x}^j_y$ $(y > Y)$ as follows:

$$\hat{x}^j_{y,lower} \leq \hat{x}^j_y \leq \hat{x}^j_{y,upper}, \tag{4.18}$$

$$\hat{x}^j_{y,lower} = \hat{x}^j_y - 1.96\sigma_y, \tag{4.19}$$

$$\hat{x}^j_{y,upper} = \hat{x}^j_y + 1.96\sigma_y. \tag{4.20}$$

Figure 4-8 indicates that variable x^1 fluctuates within the range of the confidence interval, and x^2 and x^I fluctuate unexpectedly. In particular, the figure shows that variable x^2 deviates from the confidence interval in Period 1, and variable x^I deviates from the confidence interval in Period 2.

4.3.2 Situation-dependent modeling based on partially linear additive models

In Step 2, we construct a situation-dependent model to describe the hourly electricity demand. An overview of the development of PLAMs is presented in Fig. 4-10. Primarily, we focus on the set of target demand and the corresponding explanatory variables for each period to construct the models that can describe the behavior of hourly demand during the target period. In principle, situation-dependent modeling provides an approach to describe the hourly electricity demand, which can be affected by various time-dependent factors.

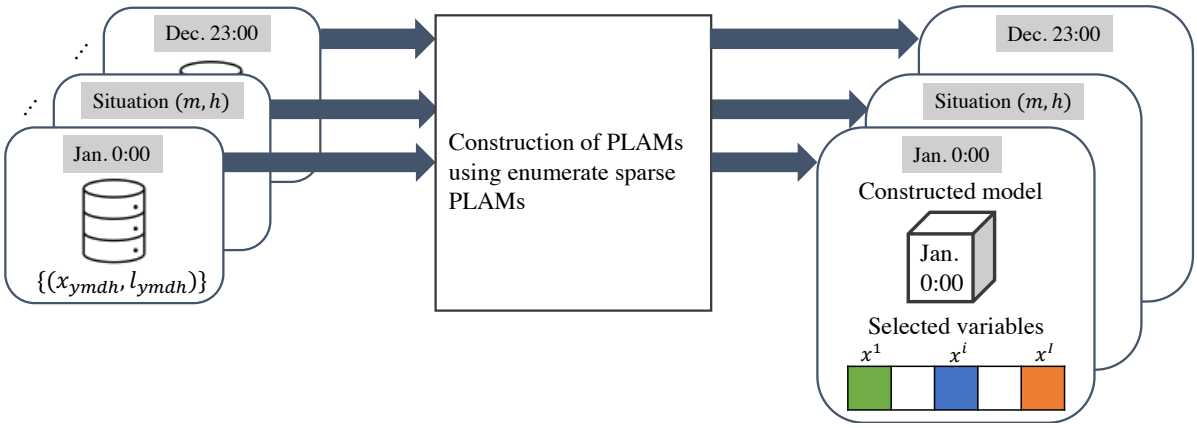


Fig. 4-10 Overview of the situation-dependent modeling.

Let $\mathcal{S} = \{1, \dots, I\}$ denote an index subset of the explanatory variable; $\{(\boldsymbol{x}_{ymdh}, l_{ymdh})\}$ represents a set of pairs containing l_{ymdh}, which reflects the hourly demand observed at hour h of day d in month m of year y; and $\boldsymbol{x}_{ymdh} = (x^1_{ymdh}, \dots, x^I_{ymdh})$ is a vector of I variables observed at the corresponding hour, day, month, and year. Overall, the daily, monthly, and yearly variables were interpolated according to the nearest neighbor method. In addition, we let $\mathcal{L} \subseteq \mathcal{S}$ and $\mathcal{N} \subseteq \mathcal{S}$ denote the index subsets of $\mathcal{S} = \{1, \dots, I\}$ to indicate the variables linearly and nonlinearly related to the target demand, respectively. PLAM [4-6] constitutes a formulation for describing the target demand with numerous variables containing partially linear and nonlinear relationships, defined as follows:

$$\begin{aligned} l_{ymdh} &\cong f(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}) \\ &= \beta_0 + \sum_{j \in \mathcal{S}} \left((\beta_j + \tau_{j,1}) x^j_{ymdh} + \sum_{k=2}^{K} \tau_{j,k} \phi^k_j (x^j_{ymdh}) \right), \end{aligned} \tag{4.21}$$

where $\boldsymbol{\theta} = \left\{ \boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_I), \left\{ \boldsymbol{\tau}_j = \{\tau_{j,1}, \dots, \tau_{j,K}\} \right\} \right\}$ represents a set of model coefficient parameters, $\boldsymbol{\beta}$ denotes the set of coefficient parameters for explanatory variables $\boldsymbol{x}$, and $\boldsymbol{\tau}_j$ indicates the vector of coefficient parameters for transformation functions $\phi^1_j(x^j), \dots, \phi^K_j(x^j)$. Moreover, we use the following cubic spline transformation functions, which are supposed to work well with PLAM in previous studies [4-6], with K bases:

$$\left[\phi^k_j (x^j_{ymdh}); k = 1, .., K \right] = \left[x^j_{ymdh}, \left(x^j_{ymdh} - x^j_{(1)} \right)^3_+, \dots, \left(x^j_{ymdh} - x^j_{(K-1)} \right)^3_+ \right], \tag{4.22}$$

$$(z)_+ = \begin{cases} 0 \ (z < 0) \\ z \ (z \geq 0) \end{cases}, \tag{4.23}$$

where $x^j_{(1)}, \dots, x^j_{(K-1)}$ represent the knots of the spline selected from quantiles in the sample set.

The sparse modeling technique provides an approach to select flexible and computationally efficient informative variables and identifies the linearity/nonlinearity in relationships between the explanatory variables and target demand. For a specific month m and hour h, the parameter can be estimated based on minimizing the squared error loss F_{mh} under the given subsets $\mathcal{L}$ and $\mathcal{N}$, as follows:

$$\widehat{\boldsymbol{\theta}}_{mh} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} F_{mh}(\boldsymbol{\theta}; \mathcal{S}, \lambda)$$
$$= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{y,d} \left(l_{ymdh} - f\left(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}\right) \right)^2 + \lambda \sum_{j \in \mathcal{S}} \left(|\beta_j| + \|\boldsymbol{\tau}_j\|_2 \right), \quad (4.24)$$

where λ denotes a positive regularization constant for the penalty, and $\|\boldsymbol{\tau}\|_2 = \sqrt{\sum_{k=1}^{K} \tau_k^2}$ represents the L2-norm of the vector $\boldsymbol{\tau}$. The components of the minimizer $\widehat{\boldsymbol{\theta}}$ in Eq. (4.24) tend to zero to reduce the absolute and L2-norm penalties, in addition to decreasing the squared error loss. Furthermore, the redundant variables tend to zero, such that a subset of explanatory variables composed of multiple informative variables can be expected to be selected [4-10]. The regularization result in Eq. (4.24) derives the informative variables under specific seasonal situations and identifies the linearity/nonlinearity between the variables as follows:

- $\beta_j \neq 0, \tau_{j,k} = 0\ (\forall k)$: Variable x^j exhibits a linear relationship with demand.
- $\tau_{j,k} \neq 0\ (\exists k)$: Variable x^j exhibits a nonlinear relationship with demand.
- $\beta_j = 0, \tau_{j,k} = 0\ (\forall k)$: Variable x^j exhibits no relationship with demand.

To improve the interpretability of the annually dominant variables for accurately describing the hourly demand, the enumerate sparse modeling technique [4-28] forms an attractive approach to identify the limited number of variables that are commonly and consistently used in situation-dependent modeling. In the enumeration scheme, the plausible candidate models are enumerated by focusing on the given situation (m, h), and a representative model is selected from the candidates to minimize the union of explanatory variables commonly used in models constructed for individual situations $\{(m, h)\}$. Let $\left(\mathcal{S}_{mh}^{(1)}, F_{mh}^{(1)}\right)$ denote a set of pairs containing the index set of variables selected by the sparse PLAM expressed in Eq. (4.24) and the squared-error loss under the variable set $\mathcal{S}_{mh}^{(1)}$, and let $\left\{\left(\mathcal{S}_{mh}^{(2)}, \ldots, \mathcal{S}_{mh}^{(I_{mh})}\right), \left(F_{mh}^{(2)}, \ldots, F_{mh}^{(I_{mh})}\right)\right\}$ denote the set of variable index subsets and the squared error loss enumerated in the enumerate sparse PLAM scheme [4-6], [4-28]. These sets of enumerated variables $\left(\mathcal{S}_{mh}^{(2)}, \ldots, \mathcal{S}_{mh}^{(I_{mh})}\right)$ contain almost similar information to describe the electricity demand, and the sets of the squared-error loss $\left(F_{mh}^{(2)}, \ldots, F_{mh}^{(I_{mh})}\right)$ exhibit only minute differences. Thus, the squared error loss in each enumerated model satisfies the following conditions.

$$0 < \frac{F_{mh}^{(i)} - F_{mh}^{(1)}}{F_{mh}^{(1)}} < \varepsilon \ \left(\forall i \in (2, \ldots, I_{mh})\right), \quad (4.25)$$

where the positive parameter ε controls the suboptimality of the enumerated results. In particular, the situation-dependent key variables are selected as follows:

$$\hat{\mathcal{S}} = \min_{\{i_{mh} \in \{1, \ldots, I_{mh}\}; \forall m,h\}} \left| \bigcup_{m,h} \mathcal{S}_{mh}^{(i_{mh})} \right|. \quad (4.26)$$

Overall, statistical models are developed to describe the hourly electricity demand behavior for each seasonal situation, and accordingly, a limited number of variables were identified. The variables selected for each seasonal situation are described in $\hat{\mathcal{S}}_{mh}$.

4.3.3 Evaluation of additive contribution to demand deviation

The demand deviation between the scenario-based demand and observed variable-based forecasted demand is analyzed in Step 3. An overview of the analysis for identifying the impact of key variables on the demand deviation is presented in Fig. 4-11. The electricity demand deviation, defined as the difference between the scenario-based demand and observed variable-based demand, is derived, and the additive contribution of each explanatory variable is identified to represent the deviance-oriented deviation. Here, the additive contributions of each variable fluctuating drastically during the target period are derived for various seasonal periods.

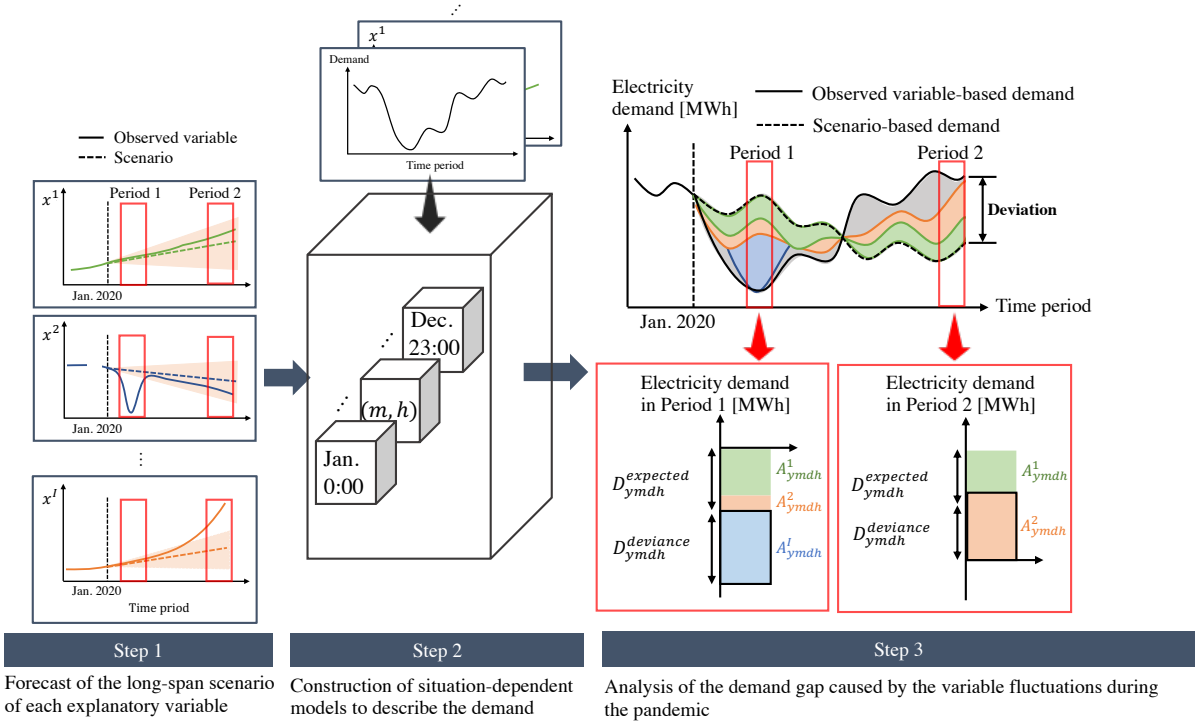


Fig. 4-11 Overview of analysis for identifying impacts of key variables on demand deviation.

Let D_{ymdh} denote the electricity demand deviation at hour h on day d in month m of year y. The scenario-based demand, observed variable-based forecasted demand, and demand deviation between these demands are described based on the constructed model described in Step 2 (Eq. (21)) as follows:

$$\hat{l}_{ymdh}^{scenario} = f(\hat{\boldsymbol{x}}_{ymdh}; \boldsymbol{\theta}), \tag{4.27}$$

$$\hat{l}_{ymdh}^{observed} = f(\boldsymbol{x}_{ymdh}; \boldsymbol{\theta}), \tag{4.28}$$

$$D_{ymdh} = \hat{l}_{ymdh}^{scenario} - \hat{l}_{ymdh}^{observed}$$

$$= \sum_{j \in \hat{\mathcal{S}}_{mh}} \left((\beta_j + \tau_{j,1})(\hat{x}_{ymdh}^{j} - x^{*j}_{ymdh}) + \sum_{k=2}^{K} \tau_{j,k} \left(\phi_j^k(\hat{x}_{ymdh}^{j}) - \phi_j^k(x_{ymdh}^{j}) \right) \right), \tag{4.29}$$

where $\hat{\mathcal{S}}_{mh} \subseteq \mathcal{S}$ denotes the index subset selected by the enumerate sparse PLAM for each seasonal situation in Step 2. Thus, the additive contribution of variable j can be described as follows:

$$A_{ymdh}^{j} = (\beta_j + \tau_{j,1})(\hat{x}_{ymdh}^{j} - \hat{x}_{ymdh}^{j}) + \sum_{k=2}^{K} \tau_{j,k} \left(\phi_j^k(\hat{x}_{ymdh}^{j}) - \phi_j^k(\hat{x}_{ymdh}^{j}) \right), \tag{4.30}$$

where the demand deviation is classified into two types of deviations: *the deviance-oriented deviation* caused by the variables acting differently from scenarios, and *the expected deviation* caused by the variables that varied in the range of expected scenarios during the pandemic. These deviance-oriented demand deviations and expected deviations are described based on additive contributions as follows:

$$D_{ymdh} = D_{ymdh}^{expected} + D_{ymdh}^{deviance}, \tag{4.31}$$

$$D_{ymdh}^{deviance} = \sum_{j \in \hat{\mathcal{S}}_{ymdh}^{deviance}} A_{ymdh}^{j}, \tag{4.32}$$

$$D_{ymdh}^{expected} = \sum_{j \in \hat{\mathcal{S}}_{ymdh}^{expected}} A_{ymdh}^{j}, \tag{4.33}$$

$$\hat{\mathcal{S}}_{mh}^{deviance} \cup \hat{\mathcal{S}}_{mh}^{expected} = \hat{\mathcal{S}}_{mh}, \tag{4.34}$$

$$\hat{\mathcal{S}}_{mh}^{deviance} \cap \hat{\mathcal{S}}_{mh}^{expected} = \emptyset, \tag{4.35}$$

where $\hat{\mathcal{S}}_{mh}^{deviance}$ represents the index of the key variables deviating from the scenarios in each seasonal period, and $\hat{\mathcal{S}}_{mh}^{expected}$ denotes the variable that varies within the range of the expected scenario. These are identified based on the difference between the variable scenario and the actual variable value described in Step 1. In Fig. 4-11, variable x^I deviates from the scenario in Period 1 and causes a deviance-oriented deviation. In Period 2, the variable x^2 causes a deviance-oriented deviation. In all targeted periods, variable x^1 varies across all scenarios and can be deemed a key variable that causes the expected deviation.

4.4 Case study

4.4.1 Simulation setup

In this section, the demand deviation during the pandemic was analyzed based on a dataset acquired in Germany. We utilized the actual hourly electricity demand from January 2015 to December 2021, and the 179 explanatory variables associated with typical categories, such as weather, interest rate, stock price, calendar, and GDP data, are described in Table 4-1. To construct the sparse PLAMs, we utilized parameter K in Eq. (4.21) for nonlinear transformation and parameter λ in Eq. (4.24) for the sparse scheme. In particular, these were determined using a one-day walk forward validation [4-29] from January 2018 to December 2021, where the statistical models were reconstructed per day and evaluated against the forecast of the subsequent day. In the enumerate scheme in Step 2 of the proposed method, we used parameter $\varepsilon = 0.005$, which controlled the number of enumerated models.

To describe the hourly demand, we focused on the four models listed in Table 4-2. As defined, LM focuses on a limited number of explanatory variables, e.g., weather, the German stock index, and holiday/weekday data, which have been used singularly in certain typical studies [4-3], [4-4], considering a linear relationship between them. Similar to LM, AM1 focuses on a limited number of variables, considering nonlinearity between the variables. In contrast, AM2 focuses on numerous variables (179 variables) explaining the curve in several recent studies [4-2], [4-6] considering nonlinearity between the variables. In addition, PLAM focuses on numerous variables to implement the procedure described in Section 4.3 and derives a limited number of annually dominant variables. These models were compared according to their description accuracy.

Table 4-1 Categories of explanatory variables used.

Attribute: Number of the variables	
Generation [4-30]: v1-12	Yearly
Weather [4-31]: v12-17	Hourly
Stock price [4-13]: v18	Daily
Indices of production in industry [4-12]: v19-47	Monthly
Indices of production in service [4-32]: v48-70	
Indices of production in construction [4-33]: v71-72	
GDP [4-34]: v73	
Producer price index [4-14]: v74-167	
Number of internet searches (power related words) [4-15]: v168-171	
Calendar data (holiday/weekday/day of week; binary dummy): v172-179	Daily

∗All explanatory variables are standardized to have zero means and unit variances.

Table 4-2 Condition of constructed models.

Item	Model	Note
LM	Linear Model	Utilize the limited number of explanatory variables, i.e., weather, German stock index, holiday/weekday dummy, considering linearity between variables
AM1	Additive Model	Utilize the limited number of explanatory variables, i.e., weather, German stock index, holiday/weekday dummy, considering nonlinearity between variables
AM2	Additive Model	Utilize all variables, assuming nonlinearity between variables
PLAM	Partially Linear Additive Model	Utilize all explanatory variables, assuming linearity/nonlinearity between variables

4.4.2 Description accuracy of constructed models

First, the models listed in Table 4-2 were derived based on the dataset. Subsequently, they were compared to determine the description accuracy. The average root–mean–squared error (RMSE) derived for each period is listed in Table 4-3. The results of the Nemenyi test [4-35] evaluating the significant variations in the description rank of each model are presented in Fig. 4-12. The results demonstrated that AM1 and AM2 achieved a lower description accuracy than LM, implying that the appropriate assumptions of linearity/nonlinearity between the variables and the selection of explanatory variables should improve the description accuracy. Interestingly, PLAM achieved a higher description accuracy than naive modeling approaches, signifying that the proposed approach can accurately select the key variables and identify the linearity/nonlinearity between these variables.

Table 4-3 RMSE of models described in Table 4-2.

Model	RMSE [MWh]
LM	9133.1
AM1	22777.0
AM2	11073.7
PLAM	6804.2

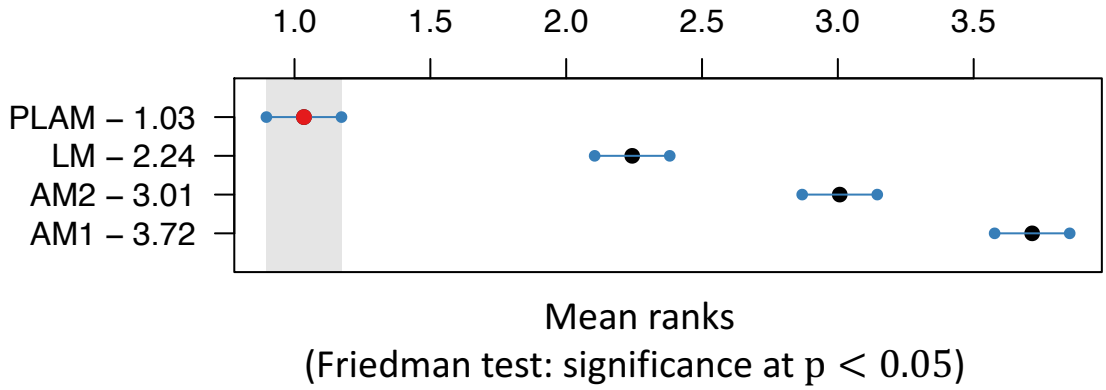
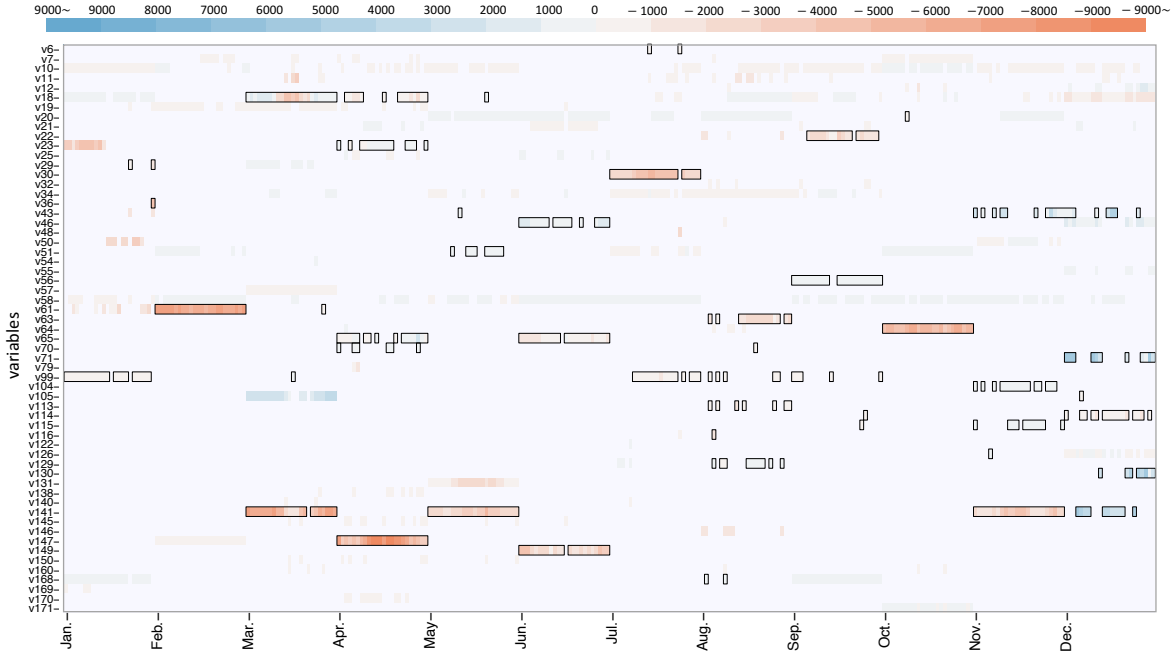


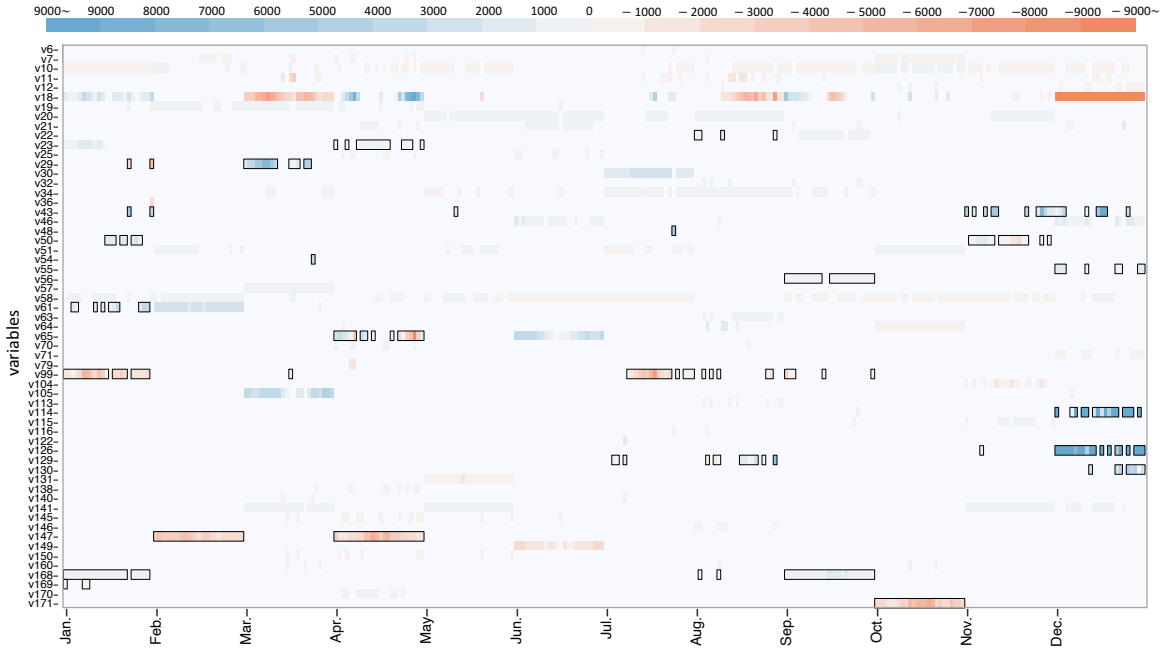
Fig. 4-12 Nemenyi test results evaluating rank of description accuracy.

4.4.3 Discussing the selected variables

The proposed approach selected key variables in each seasonal situation. Among 179 explanatory variables, the proposed approach selected 72 variables (all variables selected for each seasonal situation are described in Appendix). The results of the additive contributions of each explanatory variable influencing the electricity demand deviation in each seasonal period of 2020 and 2021 are presented in Fig. 4-13. These results were derived for each key variable, excluding weather and day-of-week information that remains unaffected by the pandemic. These results unveil the differences in additive contributions for various seasonal situations given by sets of year y, month m, and hour h. The results indicated a large negative demand deviation in 2020 and an increasing positive deviation in 2021, which is generally consistent with the trend in long-term demand fluctuations presented in Fig. 4-3. In addition, we emphasize the selected variables that caused the expected and deviance-oriented deviations. The index and each item name of the key variables affecting the demand deviation during the pandemic are presented in Table 4-4 and Table 4-5, respectively. These variables were selected from 2020 to 2021. The expected index of the variable $\hat{S}_{mh}^{expected}$ varying according to the scenario during the target period is listed in Table 4-4. The results indicated that several variables, e.g., amount of power generation (v7, 10, 11, 12), manufacturing of basic pharmaceutical products and pharmaceutical preparations (v34), and consumer price of passenger transport by road (v131), were considered crucial, even if they did not fluctuate during the pandemic. In contrast, the index of the variable $\hat{S}_{mh}^{deviance}$ that deviated from the scenario during the target period is presented in Table 4-5. The results signified that the key variables included the production of land transport and transport via pipelines (v50), postal and courier activities (v54), consumer price of domestic services and household services (v116), and passenger transport *via* railways (v130), which affected the deviance-oriented deviation. These variables may vary depending on the lockdown and the increasing time spent at home during the pandemic, which may cause an unexpectedly large demand deviation. Moreover, the characteristics of these deviations may vary across seasons. For instance, in March 2020, the German stock index (v18) and the consumer price of recording media (v141) deviated from the scenario and substantially impacted the demand deviation during the target period. However, in June 2021, the deviation can be explained based on the variables that varied according to the scenario.



(a) 2020



(b) 2021

Fig. 4-13 Results of additive contribution of each key variable that affected electricity demand deviation. Black box represents variables deviating from the scenario during target period.

Table 4-4 Key variables fluctuating with scenarios.

Category	Index	Item
Generation	v7	Nuclear
	v10	Fossil gas
	v11	Hydropumped storage
	v12	Other conventional generation[1]
Production in industry	v19	Mining of coal and lignite
	v21	Other mining and quarrying[2]
	v25	Manufacture of tobacco products
	v32	Manufacture of coke and refined petroleum products
	v34	Manufacture of basic pharmaceutical products and pharmaceutical preparations
Production in service	v57	Programming and broadcasting activities
	v58	Telecommunications
Consumer price	v79	Fruit
	v122	Paramedical services
	v131	Passenger transport by road
	v138	Equipment for reception, recording, and reproduction of sound and picture
	v140	Information processing equipment
	v145	Maintenance and repair of other major durables for recreation and culture
	v146	Games, toys, and hobbies
	v150	Veterinary and other services for pets
	v160	Electrical appliances for personal care
Google trend	v170	renewable energy

[1]This class includes the amount of conventional power generation but does not include the generation from nuclear, lignite, hard coal, fossil gas, and hydropumped storage.
[2]This class includes the mining and quarrying of various minerals and materials: abrasive materials, asbestos, siliceous fossil meals, natural graphite, steatite (talc), and feldspar.

Table 4-5 Key variables that deviated drastically from projected scenarios.

Category	Index	Item
Generation	v6	Other renewables[3]
Stock index	v18	German Stock Index
Production in industry	v20	Extraction of crude petroleum and natural gas
	v22	Mining support service activities
	v23	Manufacture of food products
	v29	Manufacture of wood and products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
	v30	Manufacture of paper and paper products
	v36	Manufacture of other nonmetallic mineral products[4]
	v43	Manufacture of other transport equipment[5]
	v46	Repair and installation of machinery and equipment
Production in service	v48	Wholesale and retail trade and repair of motor vehicles and motorcycles
	v50	Land transport and transport via pipelines
	v51	Water transport
	v54	Postal and courier activities
	v55	Publishing activities
	v56	Motion picture, video and television programmed production, sound recording and music publishing activities
	v61	Legal and accounting activities
	v63	Advertising and market research
	v64	Other professional, scientific and technical activities[6]
	v65	Rental and leasing activities
	v70	Office administrative, office support and other business support activities
Production in construction	v71	Buildings
Consumer price	v99	Refuse collection
	v104	Liquid fuels
	v105	Solid fuels
	v113	Major tools and equipment
	v114	Small tools and miscellaneous accessories
	v115	Nondurable household goods
	v116	Domestic services and household services
	v126	Spare parts and accessories for personal transport equipment
	v129	Other services in respect of personal transport equipment[7]
	v130	Passenger transport by railway
	v141	Recording media
	v147	Equipment for sport, camping and open-air recreation
	v149	Pets and related products
Google Trends	v168	“energy efficiency”
	v169	“global warming”
	v171	“solar energy”

[3]This class includes the amount of power generated by renewable energy resources but does not include the generation from biomass, hydropower, wind offshore, wind onshore, and photovoltaics.
[4]This class includes manufacturing activities related to a single substance of mineral origin. This class includes the manufacturing of glass and glass products (e.g., flat glass, hollow glass, fibers, and technical glassware), ceramic products, tiles and baked clay products, and cement and plaster from raw materials to finished articles.
[5]This class includes manufacturing of transportation equipment such as ship-building and boat manufacturing, manufacturing of railroad-rolling stock, locomotives, air, and spacecraft, and related components.
[6]This class includes diverse service activities generally delivered to commercial clients. This includes activities for which more advanced professional, scientific, and technical skill levels are required, but it does not include ongoing, routine business functions that are typically of short duration.
[7]This class includes hiring of garages or parking spaces that do not provide parking related to dwellings.

4.4.4 Analyzing the deviation caused by pandemic

We focus on the result of the pandemic deviation for seasonal situations. Fig. 4-14 shows the results of the monthly average of the expected deviation and deviance-oriented deviation. The result suggests that in April 2020, when the largest deviation occurred, approximately 95% of the deviation was caused by the deviance-oriented demand deviation. Additionally, the results show that the characteristics of the demand structure changes in winter of 2020; the deviance-oriented deviation drastically decreases in 2021. This result indicates that the components of the deviations are different depending on the seasonal situation. We also focus on an example of the electricity deviation and the additive contributions of each key variable in certain seasonal periods. Examples of the hourly averages of the actual, observed variable-based, and scenario-based demand are presented in Fig. 4-15. The results suggest that the variable-based demand estimated by the proposed approach successfully reproduced the actual demand. In Fig. 4-16, an example of the additive contributions of the key variables is presented for the same period considered in Fig. 4-15, where the additive contributions constitute the demand deviation, i.e., the difference between the observed variable- and scenario-based demands. As depicted in Fig. 4-16 (a), the manufacturing of food products (v23) and the production of land transport and transport via pipelines (v50) in January 2020 significantly impacted the demand deviation. These variables did not deviate significantly in the target period, and the results suggest that the demand deviation in January 2020 may not have been caused by the pandemic. In particular, COVID-19 infections in Germany started at the end of January, and thus, the pandemic did not affect this demand in January. As depicted in Fig. 4-16 (b), the impact caused by the variables deviating from the scenario during the pandemic drastically increased in February 2020 compared to January 2020. For instance, the production of legal and accounting activities (v61) created a considerable impact on the demand deviation, because the legal and accounting sector was seriously affected by the pandemic and experienced a business downturn with restrictions [4-36]. As portrayed in Fig. 4-16 (c), the manufacturing of paper and paper products (v30) remained vital throughout the entire time period in July 2021. Moreover, the consumer price of refuse collection (v99) was drastically impacted during the daytime hours (6:00 to 17:00), implying the significant impact of variable deviations only during the daytime hours in the target month. This result indicated that the additive contribution of the key variable caused by its deviations completely altered its behavior according to season and time. These seasonal and temporal situations were periods of idiosyncratic variations that were strongly influenced by pandemic-influenced economic conditions and consumer behavior. To create an accurate forecast of the long-term behavior of electricity demand, the key variables and the behavior of each important variable in such seasonal conditions must be appropriately characterized for a deep understanding. Their behavior should be appropriately assumed with respect to economic conditions and consumer interest.

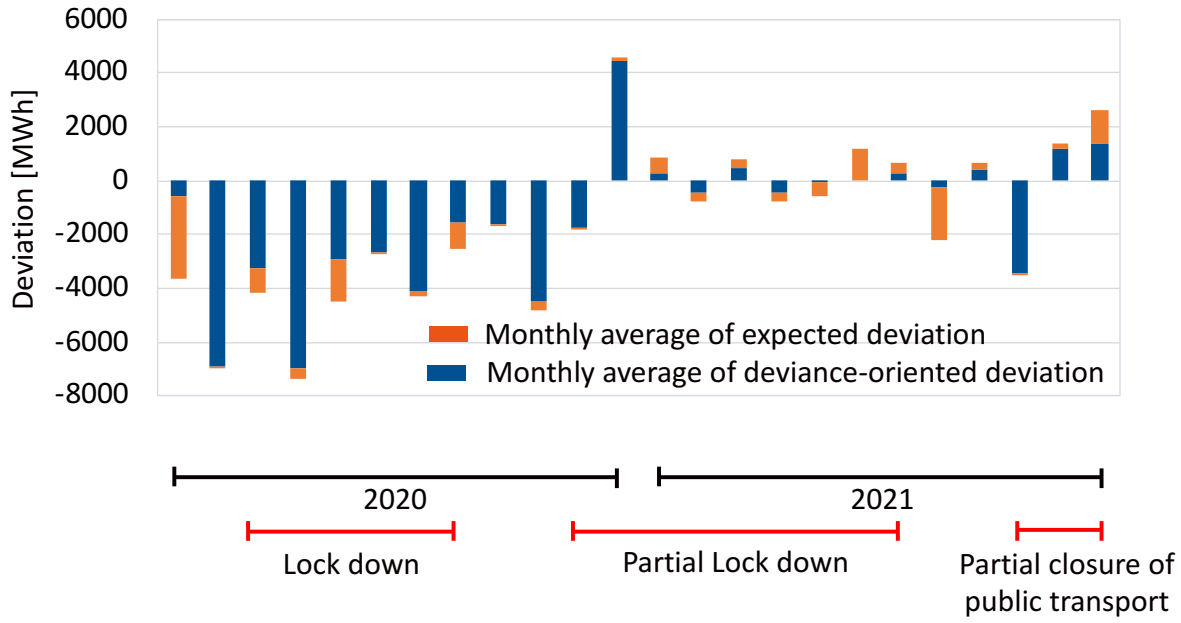
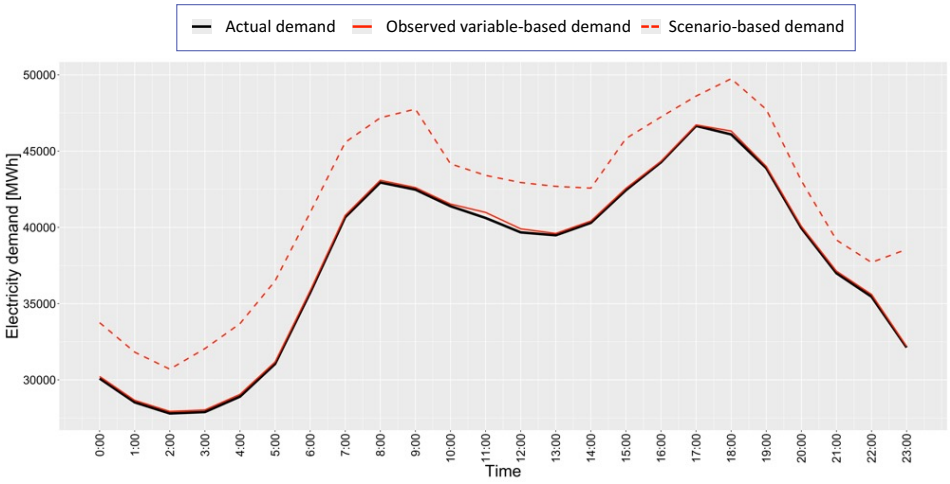
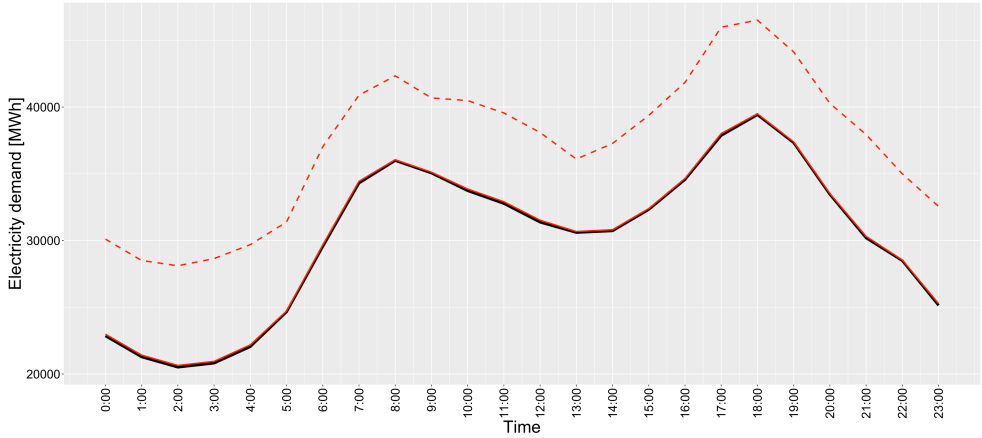


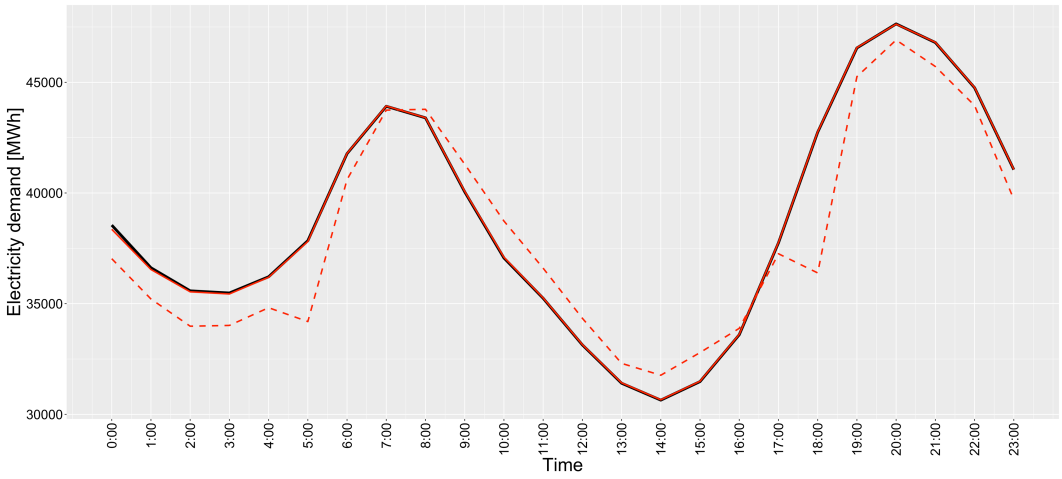
Fig. 4-14 Results of the monthly average of expected deviation and deviance-oriented deviation.



(a) January 2020

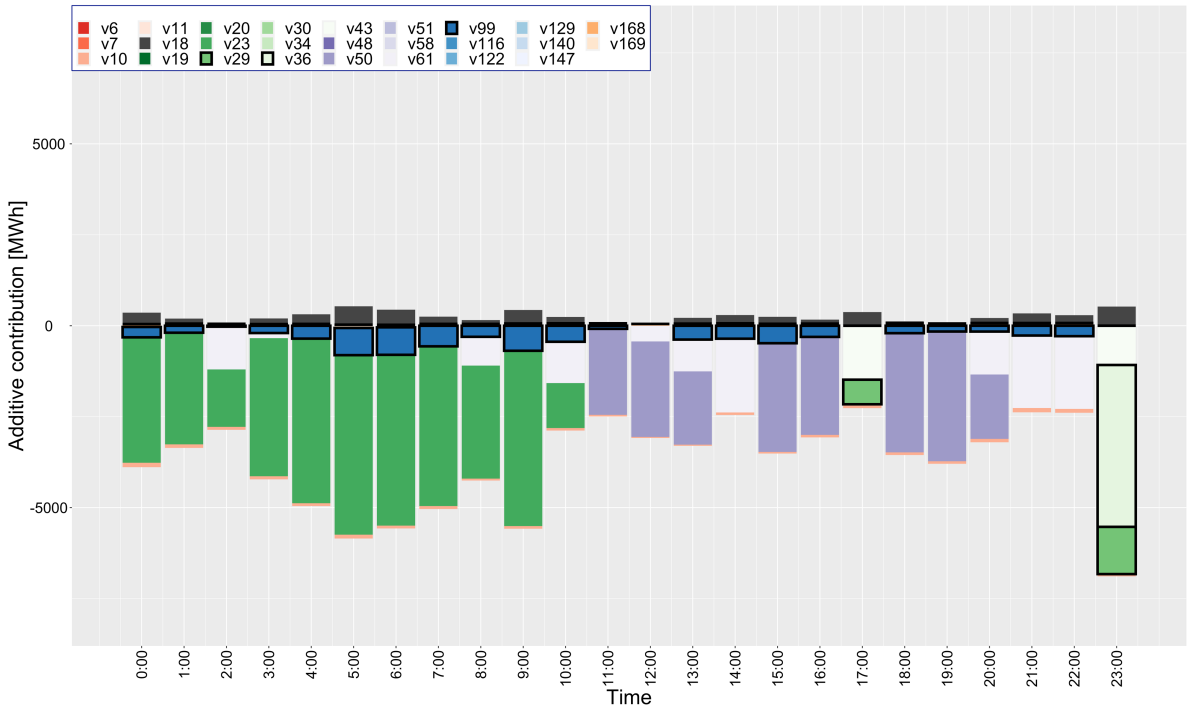


(b) February 2020

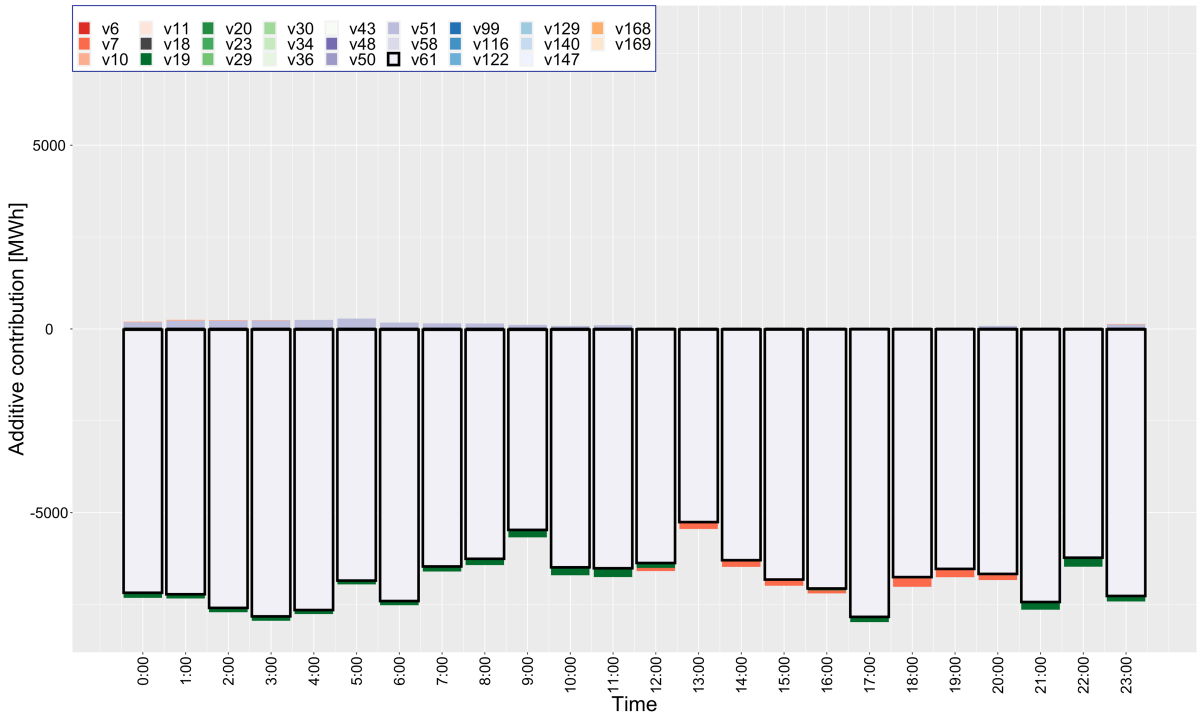


(c) July 2021

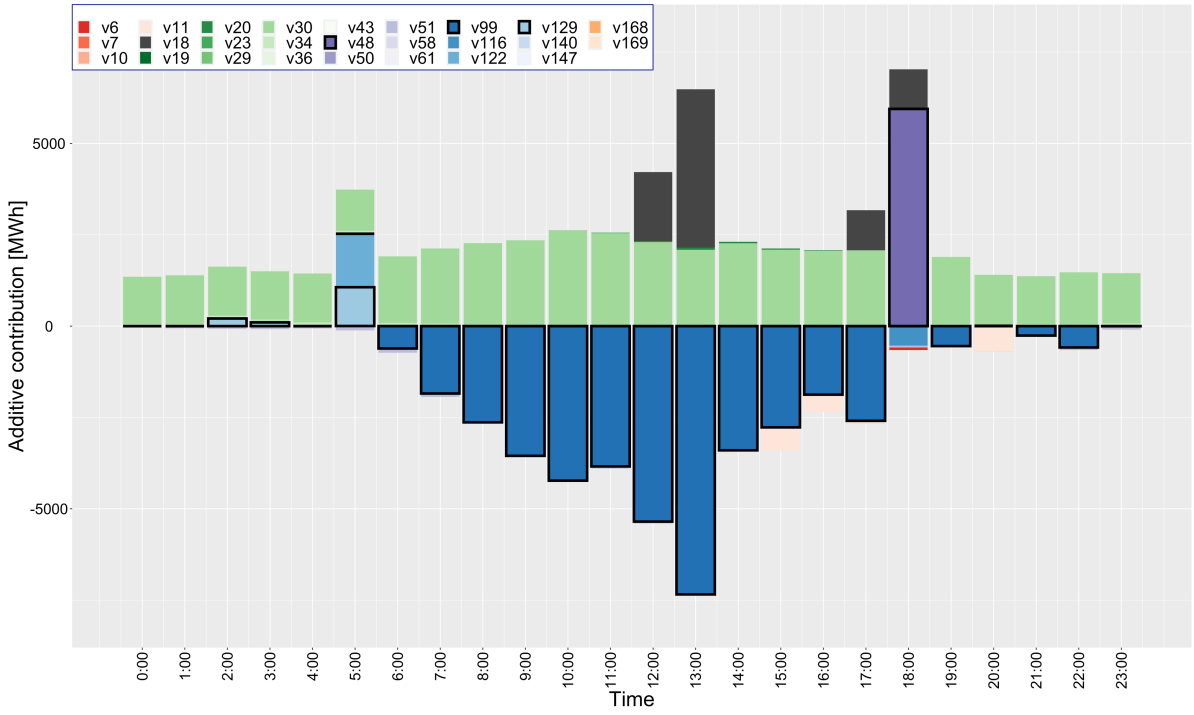
Fig. 4-15 Examples of hourly average of actual, observed variable-based and scenario-based demand period.



(a) January 2020



(b) February 2020



(c) July 2021

Fig. 4-16 Examples of additive contributions of key variables (same period as that in Fig. 4-14). The black box represents variables deviating from the scenario projected for the target duration.

4.4.5 Analyzing the factors affecting the deviation caused by the pandemic

Finally, we focus on the trend of additive contribution of each key variable that affected deviance-oriented deviation A^{j}_{ymdh} ($\forall j \in \hat{S}^{deviance}_{mh}$), and extract the time trend of the major patterns of contributions that constitute the deviance-oriented deviation by using principal component analysis (PCA) [4-37]. Figure 4-17 shows the five major patterns of contributions that constitute the deviance-oriented deviation during the target period. The figure shows that there exist several deviations patterns caused by the pandemic imposed government regulation of consumer behavior. For example, the major patterns pc1 and pc2 in Fig. 4-17 strongly appears related to the implementation of the partial closure of public transport, pc2 strongly appears related to the lock down, and pc5 strongly appears in the early stages of the pandemic. Additionally, we discuss the key factors that affect the major patterns; Fig. 4-18 shows the contributions of key factors to each major pattern. The result shows that the impact of the consumer price of small tools and miscellaneous accessories (v114) and spare parts and accessories for personal transport equipment (v126), are dominant in pc1 and pc2; this result suggests that demand of personal mobile mobility may affect the deviance-oriented deviation under the partial closure of public transport. The impact of the consumer price of equipment for sports, camping and open-air recreation (v147) is dominant in pc3, which suggests that outdoor equipment demand mainly affects the demand deviation under the lockdown. The impact of the production of legal and accounting activities (v61) is dominant in the major pattern pc5, which suggests that the demand deviation may change according to the several business and government activities under the pandemic. These results indicate that the analysis using PCA worked effectively in extracting the various patterns of the deviance-oriented demand deviation caused by the several events under the pandemic and identifying the key factors of each pattern.

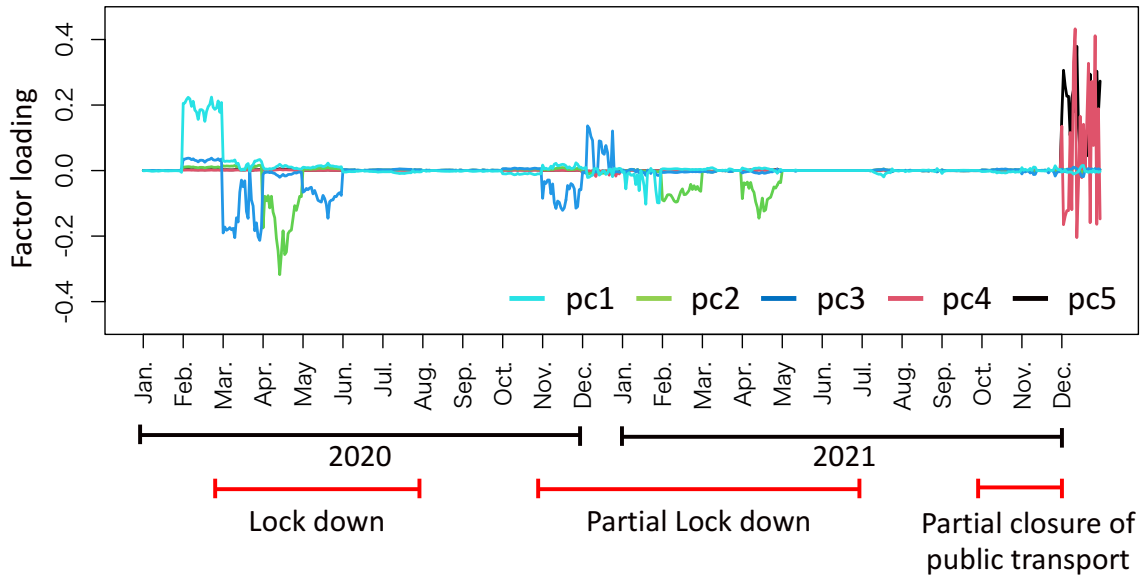


Fig. 4-17 Five major patterns of contributions that constitute the deviance-oriented deviation.

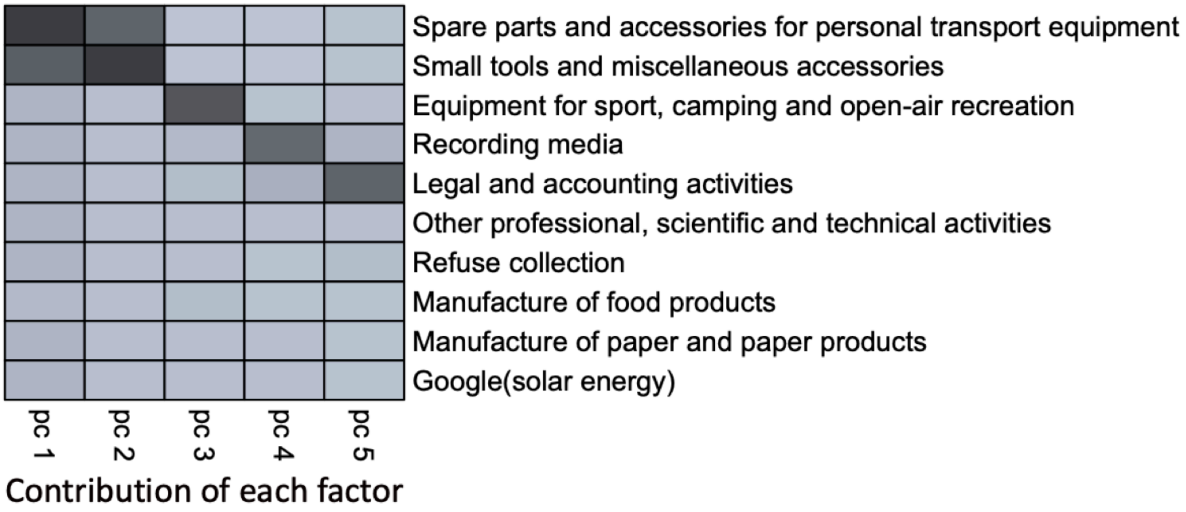


Fig. 4-18 Contributions of key factors of each major pattern.

4.5 Concluding remarks in the chapter

In this study, we proposed an approach to identify the impact of key variables on the electricity demand deviation caused by the COVID-19 pandemic based on data from Germany. We developed ARIMAX models to determine the deviation of each explanatory variable from the estimated scenarios. Furthermore, we employed PLAMs to construct demand models for selecting the key variables and describing demand behavior. The proposed approach based on ARIMAX and PLAMs successfully selected the key variables that were potentially affected by the variations in economic conditions and consumer behavior during the pandemic and identified the impact of such variables on the demand deviation for each seasonal situation. The major findings of this study are stated as follows:

1. The ARIMAX model contributes to forecasting the long-term scenario of explanatory variables during a pandemic.
2. The variable selection scheme based on the sparse enumeration technique reveals the key variables to describe the electricity demand.
3. The modeling approach using PLAMs describes the relationships between the electricity demand and key variables under various seasonal conditions.
4. The proposed approach identifies the impact of each important variable on the demand deviation caused by COVID-19.

Reference

[4-1] L. Huang, Q. Liao, R. Qiu, Y. Liang, and Y. Long, "Prediction-based analysis on power consumption gap under long-term emergency: A case in China under COVID-19," Appl. Energy, vol. 283, p. 116339, Feb. 2021, doi: 10.1016/J.APENERGY.2020.116339.

[4-2] K. Honjo, H. Shiraki, and S. Ashina, "Dynamic linear modeling of monthly electricity demand in Japan: Time variation of electricity conservation effect," PLoS One, vol. 13, no. 4, pp. e0196331–e0196331, 2018, doi: 10.1371/journal.pone.0196331.

[4-3] C. L. Hor, S. J. Watson, and S. Majithia, "Analyzing the impact of weather variables on monthly electricity demand," IEEE Trans. Power Syst., vol. 20, no. 4, pp. 2078–2085, Nov. 2005, doi: 10.1109/TPWRS.2005.857397.

[4-4] A. Z. Al-Garni, S. M. Zubair, and J. S. Nizami, "A regression model for electric-energy-consumption forecasting in Eastern Saudi Arabia," Energy (Oxford), vol. 19, no. 10, pp. 1043–1049, 1994, doi: 10.1016/0360-5442(94)90092-2.

[4-5] C. Zhang, H. Liao, and Z. Mi, "Climate impacts: temperature and electricity consumption," Nat. Hazards, vol. 99, no. 3, pp. 1259–1275, Dec. 2019, doi: 10.1007/S11069-019-03653-W/FIGURES/5.

[4-6] N. Kaneko, Y. Fujimoto, S. Kabe, M. Hayashida, and Y. Hayashi, "Sparse modeling approach for identifying the dominant factors affecting situation-dependent hourly electricity demand," Appl. Energy, vol. 265, no. December 2019, p. 114752, 2020, doi: 10.1016/j.apenergy.2020.114752.

[4-7] Y. Chen, W. Yang, and B. Zhang, "Using Mobility for Electrical Load Forecasting During the COVID-19 Pandemic." [Online]. Available: https://github.com/chennnnnyize/Load-Forecasting-During-COVID-19.

[4-8] A. Jalalkamali, M. Moradi, and N. Moradi, "Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index," Int. J. Environ. Sci. Technol., vol. 12, no. 4, pp. 1201–1210, Apr. 2015, doi: 10.1007/S13762-014-0717-6/FIGURES/7.

[4-9] T. J. Hastie, "Generalized Additive Models," Stat. Model. S, pp. 249–307, Nov. 2017, doi: 10.1201/9780203738535-7.

[4-10] Y. Lou, J. Bien, R. Caruana, and J. Gehrke, "Sparse Partially Linear Additive Models," J. Comput. Graph. Stat., vol. 25, no. 4, pp. 1126–1140, 2016, doi: 10.1080/10618600.2015.1089775.

[4-11] Y. Fujimoto, S. Murakami, N. Kaneko, H. Fuchikami, T. Hattori, and Y. Hayashi, "Machine learning approach for graphical model-based analysis of energy-aware growth control in plant factories," IEEE Access, vol. 7, pp. 32183–32196, 2019, doi: 10.1109/ACCESS.2019.2903830.

[4-12] "Statistics | Eurostat|Production in industry - monthly data." https://ec.europa.eu/eurostat/databrowser/view/STS_INPR_M__custom_2059092/default/table?lang=en (accessed Jan. 01, 2023).

[4-13] "DAX 30 Index - 27 Year Historical Chart | MacroTrends." https://www.macrotrends.net/2595/dax-30-index-germany-historical-chart-data (accessed Jan. 01, 2023).

[4-14] "Statistics | Eurostat|HICP - monthly data (index)." https://ec.europa.eu/eurostat/databrowser/view/prc_hicp_midx/default/table?lang=en (accessed Jan. 01, 2023).

[4-15] "Google Trends." https://trends.google.co.jp/trends/?geo=JP (accessed Jan. 01, 2023).

[4-16] F. Alasali, K. Nusair, L. Alhmoud, and E. Zarour, "Impact of the COVID-19 Pandemic on Electricity Demand and Load Forecasting," Sustain. 2021, Vol. 13, Page 1435, vol. 13, no. 3, p. 1435, Jan. 2021, doi: 10.3390/SU13031435.

[4-17] M. Mahmudimanesh, M. Mirzaee, A. Dehghan, and A. Bahrampour, "Forecasts of cardiac and respiratory mortality in Tehran, Iran, using ARIMAX and CNN-LSTM models," Environ. Sci. Pollut. Res., vol. 29, no. 19, pp. 28469–28479, Apr. 2022, doi: 10.1007/S11356-021-18205-8/TABLES/2.

[4-18] M. Sabbir Hossain, S. Ahmed, and M. Jamal Uddin, "Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model," Sci. Total Environ., vol. 761, p. 143315, 2021, doi: 10.1016/j.scitotenv.2020.143315.

[4-19] M. Narajewski and F. Ziel, "Ensemble forecasting for intraday electricity prices: Simulating trajectories," Appl. Energy, vol. 279, p. 115801, 2020, doi: 10.1016/j.apenergy.2020.115801.

[4-20] N. Huang, G. Lu, and D. Xu, "A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest," Energies (Basel), vol. 9, no. 10, p. 767, 2016, doi: 10.3390/en9100767.

[4-21] H. M. Alhajeri, A. Almutairi, A. Alenezi, and F. Alshammari, "Energy Demand in the

State of Kuwait During the Covid-19 Pandemic: Technical, Economic, and Environmental Perspectives.," Energies, vol. 13, no. 17, pp. 1cn-1cn, Sep. 2020, doi: 10.3390/EN13174370.

[4-22] Z. Ceylan, "The impact of COVID-19 on the electricity demand: a case study for Turkey," Int. J. Energy Res., vol. 45, no. 9, pp. 13022–13039, Jul. 2021, doi: 10.1002/ER.6631.

[4-23] T. Kanitkar, "The COVID-19 lockdown in India: Impacts on the economy and the power sector," Glob. Transitions, vol. 2, pp. 150–156, Jan. 2020, doi: 10.1016/J.GLT.2020.07.005.

[4-24] M. Malec, G. Kinelski, and M. Czarnecka, "The Impact of COVID-19 on Electricity Demand Profiles: A Case Study of Selected Business Clients in Poland," Energies 2021, Vol. 14, Page 5332, vol. 14, no. 17, p. 5332, Aug. 2021, doi: 10.3390/EN14175332.

[4-25] H. Lu, X. Ma, and M. Ma, "A hybrid multi-objective optimizer-based model for daily electricity demand prediction considering COVID-19.," Energy (Oxf)., vol. 219, p. 119568, Mar. 2021, doi: 10.1016/j.energy.2020.119568.

[4-26] Rob J. Hyndman and George Athanasopoulos, Forecasting : principles and practice : [pbk.], 2nd ed. OTexts, 2018.

[4-27] D. Altman, D. Machin, T. Bryant, and M. Gardner, "Statistics with Confidence : Confidence Intervals and Statistical Guidelines.," p. 254, 2013.

[4-28] S. Hara and T. Maehara, "Enumerate Lasso Solutions for Feature Selection," Proc. AAAI Conf. Artif. Intell., vol. 31, no. 1, 2017, doi: 10.1609/aaai.v31i1.10793.

[4-29] R. Pardo, "Walk-Forward Analysis," Eval. Optim. Trading Strateg., pp. 237–261, Sep. 2015, doi: 10.1002/9781119196969.CH11.

[4-30] "SMARD | Download market data." https://www.smard.de/en/downloadcenter/download-market-data/?downloadAttributes=%7B%22selectedCategory%22:1,%22selectedSubCategory%22:false,%22selectedRegion%22:false,%22selectedFileType%22:false%7D (accessed Jan. 01, 2023).

[4-31] "Climate Data Center." https://cdc.dwd.de/portal/ (accessed Jan. 01, 2023).

[4-32] "Statistics| Eurostat| Production in services-monthly data." https://ec.europa.eu/eurostat/databrowser/view/sts_sepr_m/default/table?lang=en (accessed Jan. 01, 2023).

[4-33] “Statistics | Eurostat|Production in construction - monthly data.” https://ec.europa.eu/eurostat/databrowser/view/sts_copr_m/default/table?lang=en (accessed Jan. 01, 2023).

[4-34] “Real Gross Domestic Product for Germany (CLVMNACSCAB1GQDE) | FRED | St. Louis Fed.” https://fred.stlouisfed.org/series/CLVMNACSCAB1GQDE (accessed Jan. 01, 2023).

[4-35] P. Nemenyi, “Distribution-free multiple comparisons,” PhD thesis, Princet. Univ. 1963., 1963, Accessed: Sep. 05, 2022. [Online]. Available: https://www.proquest.com/docview/302256074/previewPDF/9AD23167E3524469PQ/1?accountid=14891.

[4-36] William and Hoke, “Law and Accounting Firms vs. COVID-19 - Bird & Bird,” TAX NOTES Int. April 13, 2020, Accessed: Sep. 05, 2022. [Online]. Available: https://www.twobirds.com/en/insights/2020/global/law-and-accounting-firms-vs-covid-19.

[4-37] “Principal Components Analysis,” SAGE Res. Methods Found., 2020, doi: 10.4135/9781526421036878174.

Appendix

PLAM selected important variables while identifying the linearity/nonlinearity among variables.

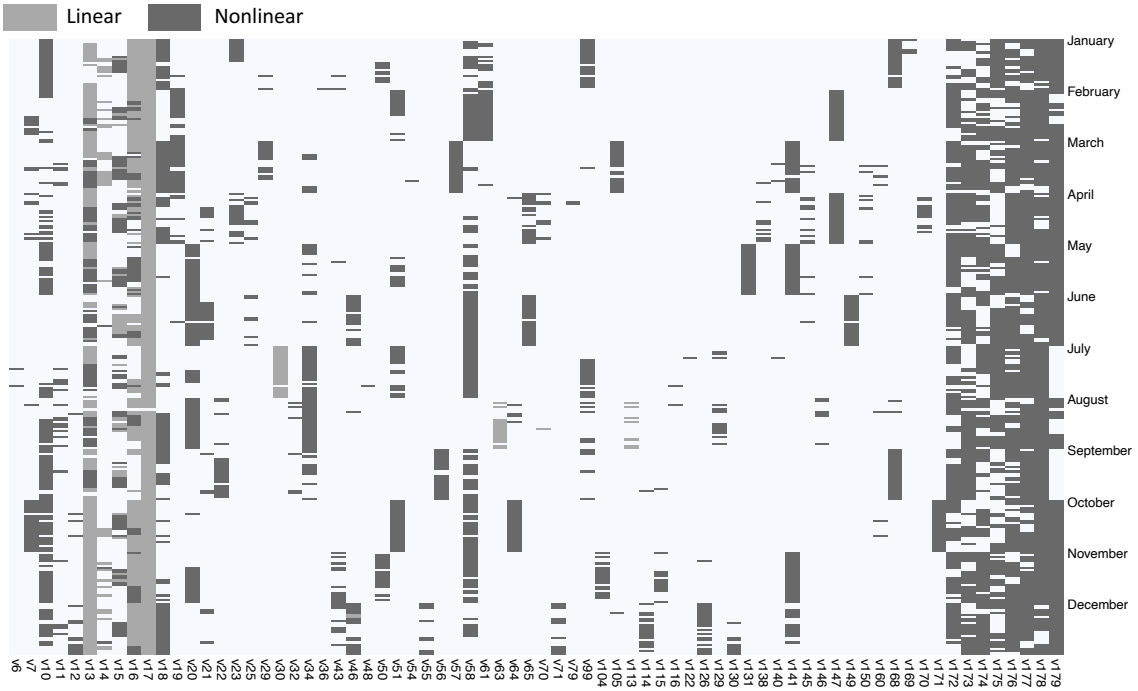


Fig. 4-A1 Selected variables derived by using PLAMs. The x-axis represents the selected variables, and the y-axis represents the transition of the situation ($\boldsymbol{y}$, $\boldsymbol{m}$, $\boldsymbol{h}$).

A data set utilized in this study is composed of 179 variables.

Table 4-A1 Details of all variables utilized in this study.

Category	Index	Item	Granularity
Generation	v1	Biomass	Yearly
	v2	Hydropower	
	v3	Wind offshore	
	v4	Wind onshore	
	v5	Photovoltaics	
	v6	Other renewable	
	v7	Nuclear	
	v8	Lignite	
	v9	Hard coal	
	v10	Fossil gas	
	v11	Hydro pumped storage	
	v12	Other conventional	
Weather	v13	Humidity	Hourly
	v14	Precipitation	
	v15	Sun shine duration	
	v16	Temperature	
	v17	Wind speed	
Stock index	v18	German Stock Index	Daily
Production in industry	v19	Mining of coal and lignite	Monthly
	v20	Extraction of crude petroleum and natural gas	
	v21	Other mining and quarrying	
	v22	Mining support service activities	
	v23	Manufacture of food products	
	v24	Manufacture of beverages	
	v25	Manufacture of tobacco products	
	v26	Manufacture of textiles	
	v27	Manufacture of wearing apparel	
	v28	Manufacture of leather and related products	
	v29	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	
	v30	Manufacture of paper and paper products	
	v31	Printing and reproduction of recorded media	
	v32	Manufacture of coke and refined petroleum products	
	v33	Manufacture of chemicals and chemical products	
	v34	Manufacture of basic pharmaceutical products and pharmaceutical preparations	
	v35	Manufacture of rubber and plastic products	
	v36	Manufacture of other non-metallic mineral products	
	v37	Manufacture of basic metals	
	v38	Manufacture of fabricated metal products, except machinery and equipment	
	v39	Manufacture of computer, electronic and optical products	
	v40	Manufacture of electrical equipment	
	v41	Manufacture of machinery and equipment n.e.c.	
	v42	Manufacture of motor vehicles, trailers and semi-trailers	
	v43	Manufacture of other transport equipment	
	v44	Manufacture of furniture	
	v45	Other manufacturing	
	v46	Repair and installation of machinery and equipment	
	v47	Electricity, gas, steam and air conditioning supply	
Production in service	v48	Wholesale and retail trade and repair of motor vehicles and motorcycles	Monthly
	v49	Wholesale trade, except of motor vehicles and motorcycles	
	v50	Land transport and transport via pipelines	

Category	Index	Item	Granularity
Production in service	v51	Water transport	Monthly
	v52	Air transport	
	v53	Warehousing and support activities for transportation	
	v54	Postal and courier activities	
	v55	Publishing activities	
	v56	Motion picture, video and television program production, sound recording and music publishing activities	
	v57	Programming and broadcasting activities	
	v58	Telecommunications	
	v59	Computer programming, consultancy and related activities	
	v60	Information service activities	
	v61	Legal and accounting activities	
	v62	Architectural and engineering activities; technical testing and analysis	
	v63	Advertising and market research	
	v64	Other professional, scientific and technical activities	
	v65	Rental and leasing activities	
	v66	Employment activities	
	v67	Travel agency, tour operator and other reservation service and related activities	
	v68	Security and investigation activities	
	v69	Services to buildings and landscape activities	
	v70	Office administrative, office support and other business support activities	
Production in construction	v71	Buildings	Monthly
	v72	Civil engineering works	
GDP	v73	Gross Domestic Product(GDP)	Monthly
Consumer price	v74	Bread and cereals	Monthly
	v75	Meat	
	v76	Fish and seafood	
	v77	Milk, cheese and eggs	
	v78	Oils and fats	
	v79	Fruit	
	v80	Vegetables	
	v81	Sugar, jam, honey, chocolate and confectionery	
	v82	Food products n.e.c.	
	v83	Coffee, tea and cocoa	
	v84	Mineral waters, soft drinks, fruit and vegetable juices	
	v85	Spirits	
	v86	Wine	
	v87	Beer	
	v88	Clothing materials	
	v89	Garments	
	v90	Other articles of clothing and clothing accessories	
	v91	Cleaning, repair and hire of clothing	
	v92	Shoes and other footwear	
	v93	Repair and hire of footwear	
	v94	Actual rentals paid by tenants	
	v95	Other actual rentals	
	v96	Materials for the maintenance and repair of the dwelling	
	v97	Services for the maintenance and repair of the dwelling	
	v98	Water supply	
	v99	Refuse collection	
	v100	Sewerage collection	

Category	Index	Item	Granularity
Consumer price	v101	Other services relating to the dwelling n.e.c.	Monthly
	v102	Electricity	
	v103	Gas	
	v104	Liquid fuels	
	v105	Solid fuels	
	v106	Heat energy	
	v107	Furniture and furnishings	
	v108	Carpets and other floor coverings	
	v109	Repair of furniture, furnishings and floor coverings	
	v110	Major household appliances whether electric or not	
	v111	Small electric household appliances	
	v112	Repair of household appliances	
	v113	Major tools and equipment	
	v114	Small tools and miscellaneous accessories	
	v115	Non-durable household goods	
	v116	Domestic services and household services	
	v117	Pharmaceutical products	
	v118	Other medical products	
	v119	Therapeutic appliances and equipment	
	v120	Medical services	
	v121	Dental services	
	v122	Paramedical services	
	v123	Motor cars	
	v124	Motor cycles	
	v125	Bicycles	
	v126	Spare parts and accessories for personal transport equipment	
	v127	Fuels and lubricants for personal transport equipment	
	v128	Maintenance and repair of personal transport equipment	
	v129	Other services in respect of personal transport equipment	
	v130	Passenger transport by railway	
	v131	Passenger transport by road	
	v132	Passenger transport by air	
	v133	Passenger transport by sea and inland waterway	
	v134	Combined passenger transport	
	v135	Other purchased transport services	
	v136	Telephone and telefax equipment	
	v137	Telephone and telefax services	
	v138	Equipment for the reception, recording and reproduction of sound and picture	
	v139	Photographic and cinematographic equipment and optical instruments	
	v140	Information processing equipment	
	v141	Recording media	
	v142	Repair of audio-visual, photographic and information processing equipment	
	v143	Major durables for outdoor recreation	
	v144	Musical instruments and major durables for indoor recreation	
	v145	Maintenance and repair of other major durables for recreation and culture	
	v146	Games, toys and hobbies	
	v147	Equipment for sport, camping and open-air recreation	
	v148	Gardens, plants and flowers	
	v149	Pets and related products	
	v150	Veterinary and other services for pets	

Category	Index	Item	Granularity
Consumer price	v151	Recreational and sporting services	Monthly
	v152	Cultural services	
	v153	Books	
	v154	Newspapers and periodicals	
	v155	Miscellaneous printed matter	
	v156	Stationery and drawing materials	
	v157	Restaurants, cafes and the like	
	v158	Canteens	
	v159	Hairdressing salons and personal grooming establishments	
	v160	Electrical appliances for personal care	
	v161	Other appliances, articles and products for personal care	
	v162	Jewelry, clocks and watches	
	v163	Other personal effects	
	v164	Insurance connected with the dwelling	
	v165	Insurance connected with health	
	v166	Insurance connected with transport	
	v167	Other insurance	
Google Trends	v168	Energy efficiency	Monthly
	v169	Global warming	
	v170	Renewable energy	
	v171	Solar energy	
Calendar	v172	Saturday(dummy)	Daily
	v173	Sunday(dummy)	
	v174	Monday(dummy)	
	v175	Tuesday(dummy)	
	v176	Wednesday(dummy)	
	v177	Thursday(dummy)	
	v178	Friday(dummy)	
	v179	Holiday(dummy)	

Chapter 5
Sensitivity analysis of factors relevant to the extreme imbalance between procurement plans and actual demand: Case study of the Japanese electricity market

Symbols

t	Index of a specific time slot
P_t	Procured power observed at time t
A_t	Actual power observed at time t
π_t	Imbalance price observed at time t
π_t^{unit}	Imbalance unit price observed at time t
π_t^D	Electricity price observed at time t in day-ahead market
π_t^I	Electricity price observed at time t in intraday market
α_t	Coefficient for derivation of the imbalance pricing set at time t
η	Regional-specific value that adjusts the imbalance price
$\tilde{\pi}_t^D$	Virtual price observed at time t in day-ahead market
s	$\in \{Spring, Summer, Fall, Winter\}$
p	$\in \{Earlymorning, Morning, Afternoon, Midnight, Night\}$
$\mathcal{T}$	Evaluation period
$\mathcal{T}_{sp}$	$\subset \mathcal{T}$. Specific subperiod set representing one of twenty seasonal situations
z	$\in \{shortage, surplus\}$
g_t^Z	Binary variable which indicates the occurrence of extreme imbalance at time t
$\boldsymbol{x}_t$	$= [x_t^1, \dots, x_t^I]$. Explanatory variables observed at time t
I	Number of explanatory variables
$\mathcal{S}$	$= \{1, \dots, I\}$. Index set of explanatory variables
$\mathcal{S}_{sp}^z$	Index set of explanatory variables in target situation
$\boldsymbol{\beta}_{sp}^{zi}$	Vector of coefficient parameters for explanatory variables
$\phi_k^i(\cdot)$	Cubic spline transformation function
$\boldsymbol{K}_t$	Number of bases in cubic spline transformation functions
$\boldsymbol{\tau}_{sp}^{zi}$	Vector of coefficient parameters for transformation functions $\phi_k^i(\cdot)$
$\mathcal{L}_{sp}^z$	Index subset indicating the variables related to the target extreme imbalance event linearly
$\mathcal{N}_{\mathrm{sp}}^{\mathrm{z}}$	Index subset indicating the variables related to the target extreme imbalance event nonlinearly
$I(\cdot)$	Dependency between two variables
ε_{sp}^z	Positive parameter for relative importance constant
B	Positive bootstrap sampling parameter

Nomenclature

DA	Day-ahead market
ID	Intraday market
PLAMs	Partially linear additive models
MIC	Maximal information coefficient
MI	Mutual Information

5.1 Introduction

In liberalized electricity markets, the transmission and distribution of electric power to end-use consumers has been accomplished by two actors: (1) retailers who purchase electricity from spot markets and electric pools and (2) system operators who supply electricity to consumers, businesses, and industries. Electricity is mainly procured from several markets, such as the day-ahead (DA) market and the intraday (ID) market. The DA market is a financial market where participants can procure and sell electricity at day-ahead prices (i.e., buyers and sellers can lock and assign energy prices before the real-time operating day). Meanwhile, the ID market, which is also known as the short-term wholesale power market, is one in which the procurement and selling of power occurs on the same day as the power delivery. Here, the schedules of procurement by the electricity retailers are monitored by the system operators to manage, operate, and control the power system.

Herein, imbalance is defined as the difference between the scheduled power procured in the DA and ID markets and the actual power supply. It should be compensated in the real-time balancing operation by system operators. Monitoring of this imbalance plays a key role in maintaining a modern power system. In recent years, it has been recognized that the structure of the electric power system has changed because of the large-scale penetration of renewable energy sources and energy-saving trends. For example, in Japan, electricity demand has continually decreased over the past decade because of the large-scale penetration of renewable energy sources and energy-saving trends [5-1], [5-2]. Similar changes have been observed in power systems worldwide; they affect the behavior of power procurement and consumption in the market [5-3], [5-4]. In addition, the scheduled power procured in the electricity markets depends on several factors such as market prices and generation; these factors may be influenced by electricity demand, weather and power grid conditions. These factors are also considered to cause imbalance through complex interactions [5-5].

These recent rapid transitions in the power environment have broadened and diversified the factors affecting imbalance events. These various factors may lead to an unexpectedly large imbalance called extreme imbalance [5-6]. These uncertain imbalance events pose a risk for real-time balancing operations. Under this situation, a scheme is required to determine the factors that affect the extreme imbalance event and detect these events for the smooth operation of the system. Although several factors have been reported in recent studies, analyzing the relative relationships between the extreme events and these factors is only now attracting the interest of researchers [5-6].

The imbalance settlement system has a short history, and the literature discussing the system from the academic perspective is still scarce stage [5-6]. For instance, Aïd et al. [5-7] have focused on the electricity demand and discussed a theoretical model to minimize the imbalance. Usaola et al. [5-8] and Bueno-Lorenzo et al. [5-9] have focused on the relationship between the wind energy generation and imbalance, and discussed the optimal bidding plans. Meanwhile, several researchers and real-world system operators have employed statistical and numerical model-based analyses to reveal the statistical relationships among model variables that explain the behavior of imbalance [5-5], [5-6]. The model-based analysis of data provides a useful approach for organizing relationships among various observable variables and discussing the effect of each variable; this type of analysis reveals the impact of the factors to the imbalance behavior, and contributes to the decision making of stakeholders, while understanding the risk of expected imbalance in power system. For example, Goodarzi et al. [5-6] have focused on

the volume of the imbalance, which changes depending on influencing factors such as renewable energy generation, and discussed the impact of the electricity generation change to the imbalance. They conducted a model-based analysis for revealing the statistical relationships of various variables to the imbalance event, assuming monotonically increasing/decreasing trends of the imbalance with respect to the amount of power generation. Moreover, Lisi et al. [5-10] have focused on the statistical regression models to predict the imbalance dynamics, and also suggested the dependency between the imbalance and the factors such as load and historical imbalance.

These early works suggest some of the possible and efficient analytical approaches for describing the variations of the target imbalance characteristics. Most of these studies initially focused on a limited number of explanatory variables screened according to the prior knowledge of experts. Meanwhile, in recent years, it has been recognized that imbalance can be influenced by additional factors, including weather and market conditions, increasing complexity [5-11]. Further, some of these additional factors have a nonmonotonic relationship with the imbalance events [5-10], where the elucidation of such is required to accurately describe the extreme imbalance events. Although an appropriate description of the nonlinear relationships is critical to the performance of the constructed models, in practice, describing all variables as nonlinear models without identifying the essential linearity among variables will greatly complicates the decision-making of system operators. Therefore, a framework is required for constructing such analytical models, finding the minimally necessary nonlinear relationships between variables by identifying essential nonlinearity. Such an approach for analyzing the relevant factors affecting extremely large imbalance would contribute to the understanding of the statistical behavior of imbalance in various modern power systems, accounting for drastically changing structures resulting from liberalization and the penetration of renewable energy worldwide.

In this study, we focus on the extreme shortage and surplus events observed in the Japanese imbalance settlement system for 30-min power sequences consisting of 48 slots per day. A strategy is proposed for the construction of a model-based analysis for the informative influence of the relevant variables on the odds ratio of the extreme event while considering a large number of variables. In particular, we focus on a statistical modeling approach to analyze the sensitivities of these variables to the odds ratio of the extreme imbalance events. For the sensitivity analysis of these relevant variables, we introduce a generalized class of partially linear additive model (PLAM) [5-12], called the partially linear additive logistic model (PLALM). The class of PLAM is an attractive class for organizing the relationships among variables [5-5], [5-13], which explicitly describes the linearity/nonlinearity between variables. The analyzation of the PLAM enables us to determine the main effects of the relevant explanatory variables while considering the linear/nonlinear relationships between variables and eliminate interactions among them [5-11].

The proposed scheme combines two key concepts: (1) a filter method [5-14] based on the relevance-redundancy measure [5-15] for finding the variables nonmonotonically related to extreme imbalance events and (2) a forward step-wise approach to select the appropriate nonlinearity for model description to account for the nonmonotonic relationships of these variables to the extreme imbalance. Although the proposed scheme was created through an examination of the Japanese electricity market system, similar extreme shortage and surplus events occur throughout the world; thus, this scheme can be applied to electricity market systems in a broad variety of contexts.

The major contributions of this study are as follows:

1. We formulate the problem to analyze statistically relevant factors for extreme events under seasonal conditions, which has not been sufficiently discussed in previous studies on power supply imbalance
2. We propose a partially linear additive logistic model-based procedure for discussing the sensitivities to the odds ratio of the extreme event for each relevant variable
3. We propose a procedure for identifying the explanatory variables relevant whilst identifying monotonicity and nonmonotonicity in relationships between these variables and the extreme imbalance event using the statistical relevance-redundancy measure
4. We analyze a real-world dataset using the proposed framework and discussion of the influences of the informative variables on the extreme imbalance.

The sensitivities derived by the proposed scheme suggest the impact of explanatory variables on the extreme imbalance event considering the essentially monotonic/nonmonotonic relationships; the construction of such a non-black-box model to explain the occurrence of imbalance will greatly support the understanding of risk and decision making in real-time balancing operations.

The remainder of this paper is organized as follows. Section 5.2 briefly describes the reviews relevant studies on imbalance analysis. Section 5.3 describes the imbalance settlement system in the Japanese electricity market and the basic characteristics of the extreme events in the Japanese electricity market. Section 5.4 introduces a framework for a PLAM-based response sensitivity analysis. Section 5.5 presents concepts for the identification of the relevant variables, comprising a statistical relevance-redundancy measure to find the important variables and identify intrinsic nonmonotonicity in their relationships in conjunction with a forward step-wise method to select the appropriate nonlinearity in model description. Section 5.6 shows the evaluation results of the proposed framework using a real-world dataset and presents some findings on the identified variables. Finally, Section 5.7 presents the concluding remarks.

5.2 Statistical analysis of factors affecting extreme imbalance

In the liberalized power systems worldwide, dynamics of imbalance have been constantly monitored to manage the grids. Several researchers and real-world system operators have attempted to discuss the relevant factors influencing these electricity imbalance variations in each power system [5-6], [5-9], [5-10]. Bueno-Lorenzo et al. [5-9] focused on the imbalance in the Spanish electricity market and discussed the relationship between the imbalance and wind power generation. Lisi et al. [5-10] focused on the probability of having a shortage/surplus in the Italian electricity market and provided evidence of the influence of renewable resources on imbalance events. Further, Goodarzi et al. [5-6] focused on the quantiles of the imbalance and discussed the influence of renewable resources on each quantile; the results suggested that the influences of factors can vary depending on the volume of imbalances and that extremely large imbalances lead to operation risk in power systems. They suggested that these informative factors change depending on the season. Thus, the analysis of real-world data is important to grasp such extreme imbalances.

Meanwhile, imbalance varies greatly according to various factors in a situation where

electric power system structures have undergone changes and the factors affecting power systems are becoming increasingly complex. In such a situation, an analysis that focuses on a limited number of variables may be insufficient to explain the mechanism of imbalance; therefore, an analysis that considers a large number of possible variables is required. For instance, Honjo et al. [5-16] suggested that energy conservation would depend on various types of climatic and economic factors. Here, each factor may have a different influence on the probability trend of extreme events depending on the change in the variable amount: a monotonic trend increases/decreases depending on the change in the number of variables, and a nonmonotonic trend changes under the specific variable conditions. As the factors affecting imbalance events become more complex depending on the power system changes, grasping such monotonic/nonmonotonic relationships between the factors and imbalance events also becomes important from the viewpoint of improving the performance of describing the imbalance events [5-10].

Sensitivity is an important measure of variable stability in the context of model-based analysis of the influence of factors on extreme imbalance[5-17]. The sensitivity of a variable translates to the variation in the response of the model with respect to a change in the interested variable. The main objective of statistical analysis for extreme imbalance events is the identification of variables and their associated sensitivities to extreme events using the imbalance results and observable variables. One of the traditional approaches for sensitivity analysis of the response model is known to be local sensitivity analysis [5-17] where sensitivities are quantified as the partial derivative of the target function with respect to the relevant variable. The sensitivity analysis of a variable in a nonlinear model reveals the effect of a variable while considering the nonlinearity of other relevant variables. Such variable sensitivities help clearly characterize the variability with respect to a particular target variable in the model, which may be locally high [5-18], [5-19]. In contrast, in the case of a linear model, the sensitivity of a variable is given as a constant value regardless of the change in the variable, making the analysis simpler owing to high interpretability; thus, it has been widely used to identify, prioritize, and screen influential factors to the response variable [5-6], [5-20], [5-21]. PLAMs explicitly identify the variables that can be represented linearly; therefore, they are attractive for interpreting and discussing the effects of relevant variables with extreme imbalance. However, one difficulty in this analysis is finding the relevant linearity/nonlinearity among variables for model construction. In recent years, some procedures have been discussed for dealing with large numbers of variables in statistical analysis. Variable selection approaches are widely used for identifying the relevant variables [5-22], [5-23]; the concept for which aims to emphasize the informative variables by screening the irrelevant and redundant variables and improving the performance accuracy of the response value. For example, the embedded method presented by Hastie [5-24] is a popular framework for finding important variables while constructing the models simultaneously. The idea of sparse regression models [5-25] such as sparse partially linear additive models [5-11] is categorized within this methodology. Such models have attracted attention in the machine learning domain for selecting a limited number of informative variables among a large number of candidate variables while also selecting for plausible linear or nonlinear relationships. However, such methods assume nonlinearity for all explanatory variables in advance of model construction, for which it is clearly a difficult task to decide the proper nonlinearity given the large number of explanatory variables.

In this study, we are particularly interested in the sensitivity analysis of the influence of factors on extreme imbalance based on PLALMs and propose a framework for identifying

the variables relevant for construction of PLALMs. This analysis aims to find the relevant variables while identifying the monotonic/nonmonotonic trend based on the filter method using relevant-redundancy measures and select the appropriate degree of nonlinearity of each nonmonotonic relationship for the model based on the forward step-wise selection approach.

5.3 Analyzing the basic characteristics of extreme events

Observed imbalances are recorded and disclosed according to the scheme of each country's electricity settlement system. Analyzing the behavior of such observed imbalances and the relevant factors affecting them is important for managing modern power systems. In this study, we focus on a coefficient that is disclosed to the public in Japan's settlement system; this coefficient represents the volume of the shortage or surplus imbalance. This section describes the mechanism by which a coefficient represents the volume of imbalance and the basic characteristics of the observed imbalance.

5.3.1 Japanese imbalance settlement system

In April 2016, the Japanese electricity market was liberalized [5-26], which enabled a large number of electricity retailers to enter the wholesale electricity market for trading electricity. In this market system, the electricity retailers trade electricity for 30 min and procure 48 slots of electricity per day. The power system operators must compensate the imbalance between the electricity procured by retailers and actual power demand on the target day. The occurrence of imbalance in the wholesale electric power system is explained in a form of a coefficient in the settlement system between the retailers and the system operators, where the imbalance compensation costs are settled between the electricity retailers and the system operator according to this coefficient. Let π_t be the imbalance price settled for each time slot t, P_t be the procured power, and A_t be the actual power demand handled by the electric power retailers. In the current Japanese electricity market, the imbalance price π_t is given as follows:

$$\pi_t = \pi_t^{\text{unit}} \times |P_t - A_t|, \tag{5.1}$$

where $P_t - A_t$ is the imbalance, which indicates the difference between the procured power and the actual power demand in the market. Then,

$$\pi_t^{\text{unit}} = \alpha_t \times h(\pi_t^D, \pi_t^I) + \eta, \tag{5.2}$$

is the imbalance unit price; here, α_t is the coefficient used for derivation of the imbalance pricing set for the target time slot; π_t^D and π_t^I are the prices obtained in the DA and ID markets, respectively; the function h indicates the weighted average of π_t^D and π_t^I based on the number of bits in the DA and ID markets; and η is a region-specific value that adjusts the price based on the difference in the level of imbalance compensation cost for 10 areas: Hokkaido, Tohoku, Kanto, Chubu, Hokuriku, Kansai, Chugoku, Shikoku, Kyusyu, and Okinawa. The system operator pays the imbalance price to the retailer when $P_t - A_t > 0$ holds; on the other hand, the retailer pays the imbalance price to the system operator when $P_t - A_t < 0$ holds. Figure 5-1 shows an overview of the derivation mechanism of the value α_t; the value $\tilde{\pi}_t^D$ is the virtual price contracted in the DA market, assuming that the imbalance is traded in the market.

Then, the procurement and demand imbalance in the entire power system in each time

slot is accessible to the public as the settlement coefficient α_t. In the current Japanese settlement system, it is defined as follows:

$$\alpha_t = \tilde{\pi}_t^D \,/\, \pi_t^D. \tag{5.3}$$

Here, $\alpha_t > 1$ implies that the procurement was insufficient (shortage imbalance), and $0 < \alpha_t < 1$ implies that the procurement was excessive (surplus imbalance).

This study focuses on the imbalance in the Japanese electricity market system; it analyzes the seasonal characteristics of the behavior of extreme shortage/surplus and discusses an approach for analyzing the factors affecting these extreme events.

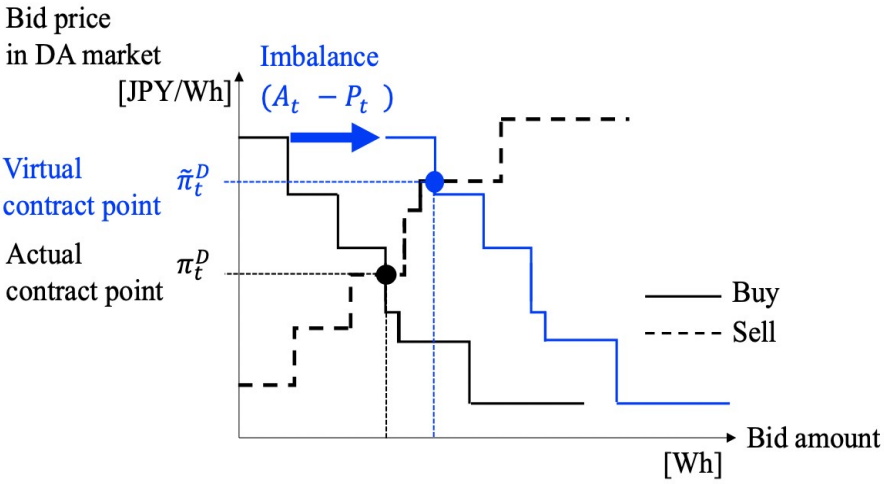


Fig. 5-1 Overview of the derivation mechanism of α_t.

5.3.2 Situation-dependent behavior of extreme imbalance

Let α_t $(t \in \mathcal{T})$ be the settlement coefficient observed in the evaluation period $\mathcal{T}$, which in the present simulation is defined as the period from January 2017 to December 2019, divided into 30-min time slots (48 slots per day). Figure 5-2 shows an example of the transition in the settlement coefficient observed on a certain day. In this figure, $\alpha_t = 1$ indicates that the demand was equivalent to the procurement at that moment; a large value of $|\alpha_t - 1|$ suggests a large deviation between demand and procurement. Here, we define the thresholds $\overline{\alpha}$ and $\underline{\alpha}$ based on the standard deviation of $\{\alpha_t; t \in \mathcal{T}\}$ as follows:

$$\overline{\alpha} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \alpha_t + \sqrt{\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \left(\alpha_t - \frac{1}{|\mathcal{T}|}\sum_{t'\in\mathcal{T}} \alpha_{t'}\right)^2}, \tag{5.4}$$

$$\underline{\alpha} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \alpha_t - \sqrt{\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \left(\alpha_t - \frac{1}{|\mathcal{T}|}\sum_{t'\in\mathcal{T}} \alpha_{t'}\right)^2}. \tag{5.5}$$

For the period $\mathcal{T}$ evaluated in this study, these values are given as $\overline{\alpha} = 1.12$ and $\underline{\alpha} = 0.82$. We categorize the level of imbalance based on these thresholds as follows:

- Extreme shortage: $\alpha_t > \overline{\alpha}$,
- Nearly balanced: $\underline{\alpha} \leq \alpha_t \leq \overline{\alpha}$,
- Extreme surplus: $\alpha_t < \underline{\alpha}$.

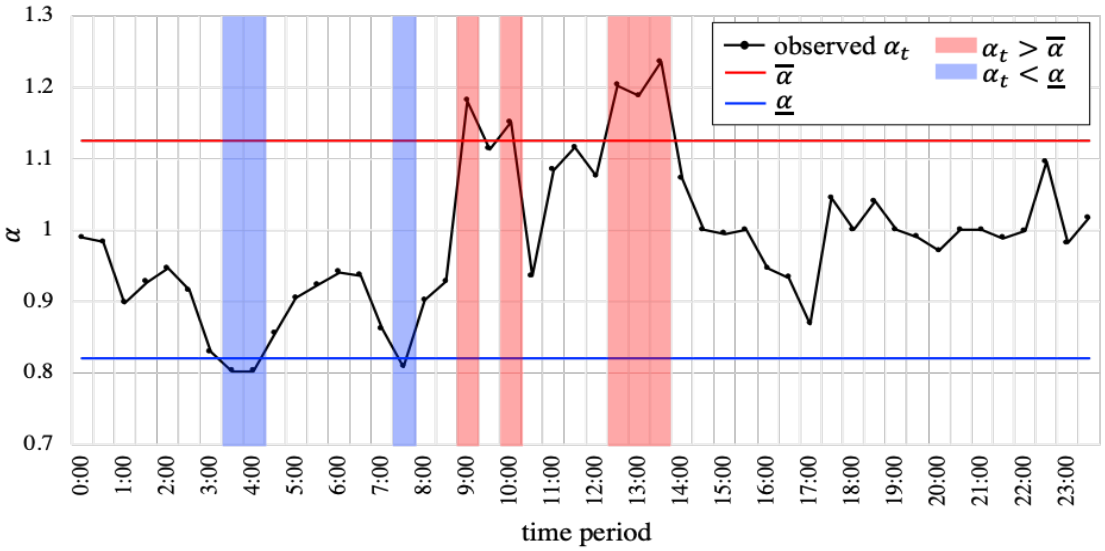
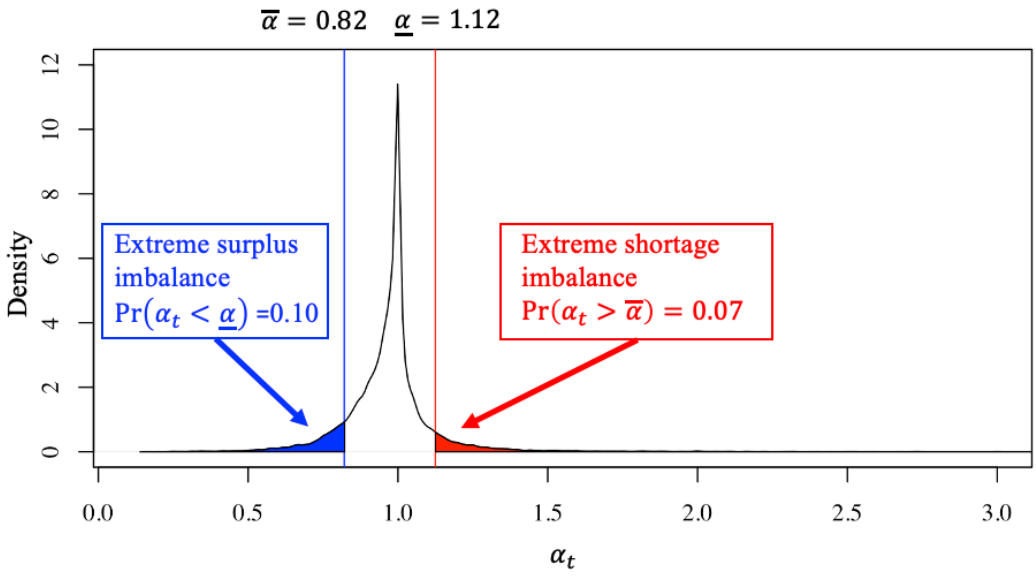


Fig. 5-2 Example of transition of the value $\boldsymbol{\alpha_t}$.

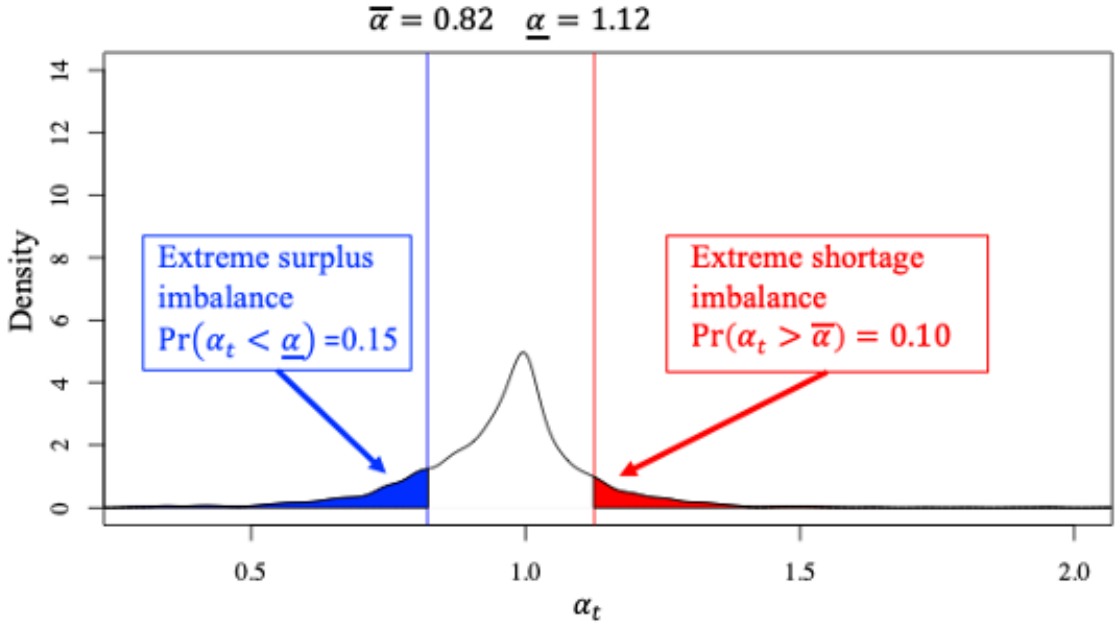
Figure 5-3 shows the probability density of the settlement coefficient α_t in the evaluation period $\mathcal{T}$, and skewness and kurtosis of the probability density; it also shows the probability of the extreme shortage and surplus event $\Pr(\alpha_t > \overline{\alpha})$ and $\Pr(\alpha_t < \underline{\alpha})$. As shown in the figure, α_t is distributed around 1.0, but follows a non-Gaussian distribution with high kurtosis and skewness. Figure 5-3 also shows that $|\alpha_t - 1|$ occasionally becomes significantly large, resulting in α_t exceeding $\overline{\alpha}$ or being less than $\underline{\alpha}$, which suggests an extreme shortage/surplus. Since the impact of imbalance on actual operation varies greatly depending on the situation of each power system, there is no common global definition of *extreme* imbalance.

With regard to the seasonal characteristics of the distribution of the coefficient α_t, we focus on a specific subperiod $\mathcal{T}_{sp} \subset \mathcal{T}$. Here, $\mathcal{T}_{sp}$ is a specific subset representing one of twenty seasonal situations represented by a combination of four seasons $s \in \{Spring, Summer, Fall, Winter\}$ and five time frames $p \in \{Earlymorning, Morning, Afternoon, Midnight, Night\}$ [5-27]. Figure 5-4 shows the examples of the distributions under several time frames in summer and winter. This figure shows that the distribution of α_t has seasonality in which the frequency of the extreme surplus and shortage events appear to differ depending on seasonal factors.



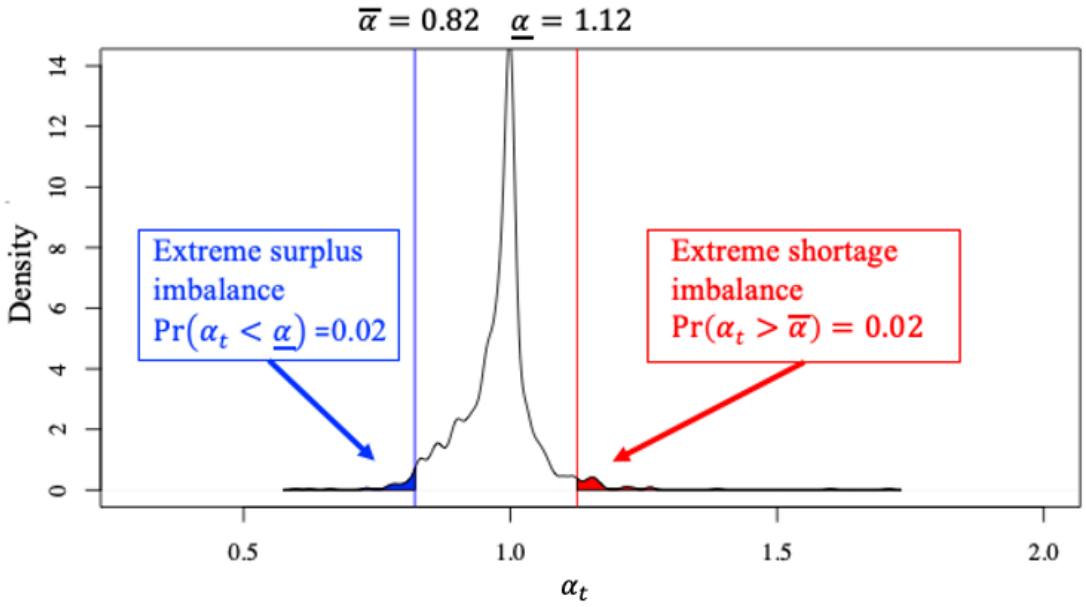
(Skewness: 7.11, Kurtosis: 63.3)

Fig. 5-3 Distribution of the settlement coefficient $\boldsymbol{\alpha_t}$ in the entire period $\boldsymbol{t} \in \boldsymbol{\mathcal{T}}$.



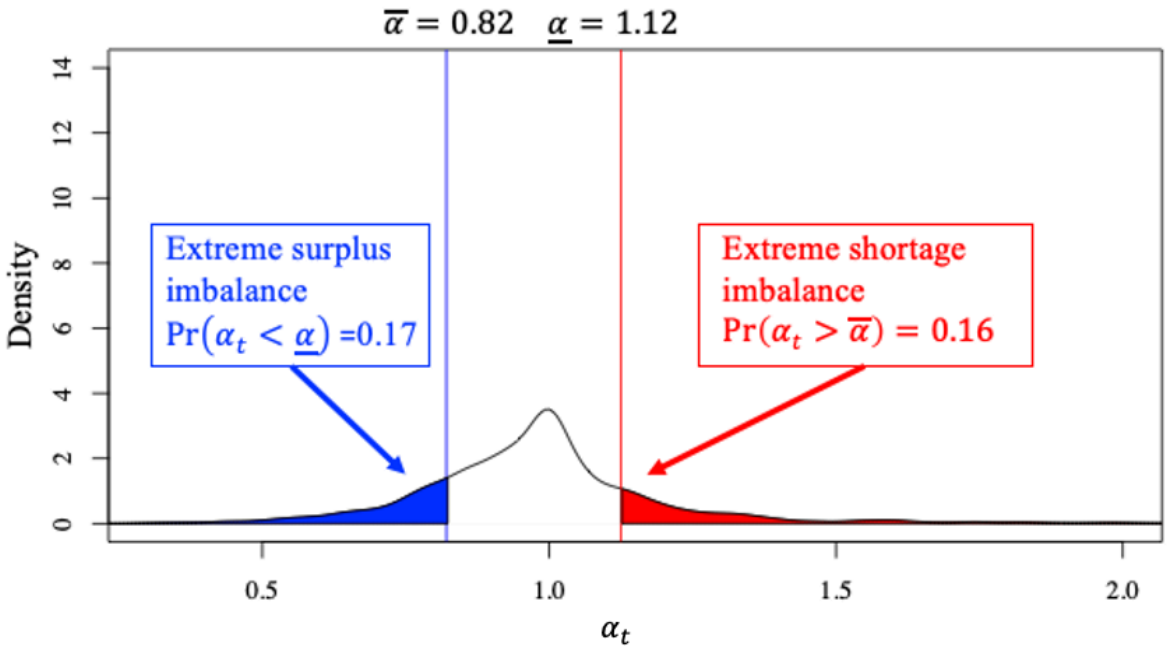
(Skewness: 3.05, Kurtosis: 9.64)

(a) 7:00–11:00 for summer



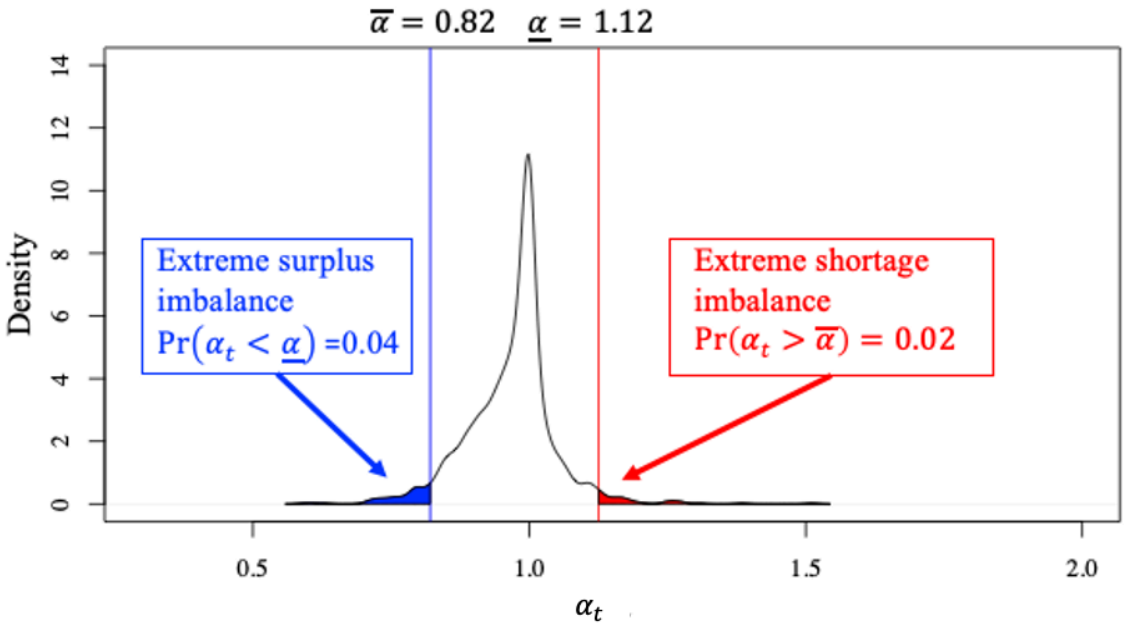
(Skewness: 4.06, Kurtosis: 19.76)

(b) 21:00–0:00 for summer



(Skewness: 3.06, Kurtosis: 9.42)

(c) 7:00–11:00 for winter



(Skewness: 2.96, Kurtosis: 9.47)

(d) 21:00–0:00 for winter

Fig. 5-4 Examples of the distribution of $\boldsymbol{\alpha_t}$ from January 1, 2017, to December 31, 2019.

5.4 Partially linear additive logistic model

We introduce the idea of additive logistic models in this section to analyze the response sensitivity to the odds ratio of the extreme event with respect to the selected variables. The class of additive models can be flexibly formulated for describing the main effect of each variable in the additive form. First, we apply the additive logistic models to describe the effect of the variables to extreme events. Subsequently, the sensitivity to odds ratio is determined by the derivative function of the additive form of the constructed models.

5.4.1 Logistic models for extreme imbalance events

Let $z \in \{shortage, surplus\}$ be an indicator of the target imbalance, and g_t^z be a binary variable defined as

$$g_t^z = \begin{cases} 1 & \text{if } \big((z = shortage) \wedge (\alpha_t > \overline{\alpha})\big) \vee \big((z = surplus) \wedge (\alpha_t < \underline{\alpha})\big) \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

and $\boldsymbol{x}_t = [x_t^1, \dots, x_t^I]$ be a vector of I explanatory variables observed in the corresponding time slot t. We also let $\mathcal{S} = \{1, \dots, I\}$ be the index set of these variables. In this study, we specifically discuss the relevance for each subperiod $\mathcal{T}_{sp}$; therefore, we focus on $(\boldsymbol{g}_{sp}^z, \boldsymbol{x}_{sp})$, where $\boldsymbol{g}_{sp}^z = [g_t^z; t \in \mathcal{T}_{sp}]$ and $\boldsymbol{x}_{sp} = [\boldsymbol{x}_t; t \in \mathcal{T}_{sp}]$.

We derive the probability of an extreme event

$$\Pr(g_t^z | \boldsymbol{x}_t) = c_t^z(\boldsymbol{x}_t), \tag{5.7}$$

as a function of the given explanatory $\boldsymbol{x}_t$ for each seasonal situation. Considering the logit transformation, the odds ratio of the extreme imbalance events can be written as the

following parametric model, which is represented by a set of the model coefficient parameters $\boldsymbol{\theta}$:

$$f(c_t^z(\boldsymbol{x}_t);\boldsymbol{\theta}_{sp}^z) \quad = \log\frac{c_t^z(\boldsymbol{x}_t)}{1-c_t^z(\boldsymbol{x}_t)}. \tag{5.8}$$

We introduce several types of logistic representations by focusing on the probability of extreme events using the filtered variables (for which the details of the function $c_t^z(\boldsymbol{x}_t)$ are described in the following section).

Here, the parameter $\boldsymbol{\theta}$ is estimated based on minimizing the log likelihood as follows:

$$\widehat{\boldsymbol{\theta}}_{sp}^z = \underset{\boldsymbol{\theta}_{sp}^z}{\operatorname{argmax}} \sum_{t\in\mathcal{T}_{sp}} \left(g_t^z \log c_t^z(\boldsymbol{x}_t) + (1-g_t^z)\log\left(1-c_t^z(\boldsymbol{x}_t)\right)\right). \tag{5.9}$$

5.4.1.1 Ordinary logistic model

Logistic model

A naive logistic model can be used to describe the linear main effects of the explanatory variables. In the linear regression-based logistic model, the function $f(c_t^z(\boldsymbol{x}_t);\boldsymbol{\theta}_{sp}^z)$ can be written as follows:

$$f(c_t^z(\boldsymbol{x}_t);\boldsymbol{\theta}_{sp}^z) = \beta_{sp}^{z0} + \sum_{i\in\mathcal{S}_{sp}^z} \beta_{sp}^{zi} x_t^i, \tag{5.10}$$

where $\mathcal{S}_{sp}^z$ is a set of indices of the explanatory variables and $\boldsymbol{\theta}_{sp}^z = \{\beta_{sp}^{z0}, \{\beta_{sp}^{zi}\}\}$ is a set of model coefficient parameters for the variables relevant to the target imbalance probability. Here, the probability of extreme events is defined as follows:

$$c_t^z(\boldsymbol{x}_t) = \frac{\exp(\sum_{i\in\mathcal{S}_{sp}^z} \beta_{sp}^{zi} x_t^i)}{1+\exp(\sum_{i\in\mathcal{S}_{sp}^z} \beta_{sp}^{zi} x_t^i)}. \tag{5.11}$$

The linear logistic regression shown in Eq. (5.10) describes the linear contribution of the explanatory variables to the extreme event in the logit form. Given its simplicity, statistical models based on this linearity assumption have been widely utilized for analyzing the main effects of each variable on the response variables [5-28]. Note that the appropriateness of the representation expressed in Eq. (5.10) is largely attributed to the linearity assumption in the relationships between the variable and the extreme event. When the variables have a linear relationship with the extreme imbalance, the constructed model will satisfactorily describe the extreme imbalance event. Meanwhile, if the assumption of the linear relationship of the variables is violated, the model could be unsuitable.

Additive logistic model

A class of the additive models [5-29] that is a natural extension of the logistic model describes the nonlinear main effects of the explanatory variables while excluding the complicated effect of the interactions among the variables. Here, we let $\mathcal{N}_{sp}^z = \mathcal{S}_{\mathrm{sp}}^z$ be the index subsets indicating the variables related to the target extreme events nonlinearly. The additive logistic model for the function $f(c_t^z(\boldsymbol{x}_t);\boldsymbol{\theta}_{sp}^z)$ is described by introducing the basis functions $\{\phi_k^i(\cdot)\}$ as follows:

$$f(c_t^z(\boldsymbol{x}_t);\boldsymbol{\theta}_{sp}^z) = \beta_{sp}^{z0} + \sum_{i\in\mathcal{N}_{sp}^z} \sum_k \tau_{spk}^{zi} \phi_k^i(x_t^i), \tag{5.12}$$

where $\boldsymbol{\theta}_{sp}^z = \left\{\beta_{sp}^{z0}, \boldsymbol{\tau}_{sp}^{zi} = \left\{\tau_{spk}^{zi}\right\}\right\}$ is a set of model coefficient parameters used to describe the nonlinearities. We may describe the cubic spline bases using the vector of the number of bases for each variable $\boldsymbol{K}_t = \{K_t^i\}$ as follows:

$$[\phi_k^i(x_t^i); k = 1, \dots, K_t^i] = \left[x_t^i, \left(x_t^i - x_{t(1)}^i\right)_+^3, \dots, \left(x_t^i - x_{t(K_t^i-1)}^i\right)_+^3\right], \tag{5.13}$$

where

$$(z)_+ = \begin{cases} 0 \ (z < 0) \\ z \ (z \geqq 0) \end{cases}, \tag{5.14}$$

and $x_{t(1)}^i, \dots, x_{t(K-1)}^i$ are the knots of the spline chosen from the quantiles in the sample set. Here, the probability of extreme events is defined as follows:

$$c_t^z(\boldsymbol{x}_t) = \frac{\exp(\sum_{i \in \mathcal{S}_{sp}^z} \sum_k \tau_{spk}^{zi} \phi_k^i(x_t^i))}{1+\exp(\sum_{i \in \mathcal{S}_{sp}^z} \sum_k \tau_{spk}^{zi} \phi_k^i(x_t^i))}. \tag{5.15}$$

We specify the appropriate nonlinearity to the extreme imbalance for each explanatory variable by the number of bases in the cubic spline expressed in Eq. (5.12). Such an additive logistic model assumes that the effect varies according to the range of the variable space and provides a flexible representation of the nonlinear relationships between explanatory variables and the extreme imbalances. However, given the large number of explanatory variables, there is a combinatorial explosion of the possible number of bases for each variable $\boldsymbol{K}_t$; it can thus be a quite difficult task to decide the appropriate number of bases for each of the explanatory variables. Further, the analyses, which focus on the differences in the effects according to the range of the variable space for all variables, unnecessarily complicates the interpretation of informative characteristics of the influence of the variables. Considering this, we follow by introducing partially linear additive models to develop the description of the minimally necessary nonlinearity for effectively analyzing the impact of each explanatory variable on the extreme imbalance.

5.4.1.2 Partially linear additive logistic model

The partial linear additive model [5-12] is a formulation for describing target extreme imbalance events with a large number of variables having partially linear and nonlinear relationships. Here, we let $\mathcal{L}_{sp}^z$ and $\mathcal{N}_{sp}^z$ be the index subsets of $\mathcal{S}_{\mathrm{sp}}^{\mathrm{z}}$ indicating the variables related to the target extreme events linearly and nonlinearly, respectively. The PLALM for the function $f(c_t^z(\boldsymbol{x}_t); \boldsymbol{\theta}_{sp}^z)$ is given as follows:

$$f(c_t^z(\boldsymbol{x}_t); \boldsymbol{\theta}_{sp}^z) = \beta_{sp}^{z0} + \sum_{j \in \mathcal{L}_{sp}^z} \beta_{sp}^{zi} x_t^i + \sum_{i \in \mathcal{N}_{sp}^z} \sum_k \tau_{spk}^{zi} \phi_k^i(x_t^i), \tag{5.16}$$

$$[\phi_k^i(x_t^i); k = 1, \dots, K_t^i] = \left[x_t^i, \left(x_t^i - x_{t(1)}^i\right)_+^3, \dots, \left(x_t^i - x_{t(K_t^i-1)}^i\right)_+^3\right], \tag{5.17}$$

where $\boldsymbol{\theta}_{sp}^z = \left\{\beta_{sp}^{z0}, \{\beta_{sp}^{zi}\}, \boldsymbol{\tau}_{sp}^{zi} = \{\tau_{spk}^{zi}\}\right\}$ is a set of the model coefficient parameters. Here, the probability of extreme events is defined as follows:

$$c_t^z(\boldsymbol{x}_t) = \frac{\exp\left(\sum_{j \in \mathcal{L}_{sp}^z} \beta_{sp}^{zi} x_t^i + \sum_{i \in \mathcal{N}_{sp}^z} \sum_k \tau_{spk}^{zi} \phi_k^i(x_t^i)\right)}{1+\exp\left(\sum_{j \in \mathcal{L}_{sp}^z} \beta_{sp}^{zi} x_t^i + \sum_{i \in \mathcal{N}_{sp}^z} \sum_k \tau_{spk}^{zi} \phi_k^i(x_t^i)\right)}. \tag{5.18}$$

PLALM expressed in Eq. (5.16) provides a representation of both linear and nonlinear relationships between the explanatory variables and extreme imbalance events.

5.4.2 Response sensitivity of relevant variables

The odds ratio described by the logistic models introduced in this study is represented in the following additive form:

$$\log\frac{\Pr(g^z=1|\mathbf{x}_t)}{\Pr(g^z=0|\mathbf{x}_t)} = \log\frac{\Pr(g^z=1|\mathbf{x}_t)}{1-\Pr(g^z=1|\mathbf{x}_t)} = \sum_i f^i(x_t^i), \tag{5.19}$$

where $f^i(x_t^i)$ indicates the main effect of x_t^i on the odds ratio.

Given the linear assumption on the variable x_t^i, the main effect expressed in Eq. (5.19) affects the odds ratio uniformly in the domain of the explanatory variable. Meanwhile, assuming the nonlinear relationship for x_t^i, the corresponding variable affects the odds ratio in various ways. To grasp the effect of the explanatory variable on the odds ratio of extreme events, we focus on the derivatives of the main effect $f^i(x_t^i)$, which represents the response sensitivity $f'^i(x_t^i)$ [5-30] of x_t^i:

$$f'^i(x_t^i) = \begin{cases} \beta_{sp}^{zi} \ (i \in \mathcal{L}_{sp}^z) \\ \sum_k \tau_{spk}^{zi} \phi_k'^i(x_t^i) \ (i \in \mathcal{N}_{sp}^z), \end{cases} \tag{5.20}$$

where

$$\begin{aligned} &[\phi_k'^i(x_t^i); k = 1, \dots, K_t^i] \\ &= [1, 3(x_t^i - x_{t(1)}^i)_+^2, \dots, 3(x_t^i - x_{t(K_t^i-1)}^i)_+^2]. \end{aligned} \tag{5.21}$$

By focusing on Eq. (5.20) under the derived model, the response sensitivity of the relevant explanatory variables to the odds ratio of extreme imbalance can be understood even with nonlinear relationships.

5.4.3 Issues encountered when constructing partially linear additive models

PLALMs and the model-based response sensitivity analysis facilitates flexible and interpretable discussion of the influence of the informative variables on the extreme imbalance. Here, the difficulty of such statistical modeling is the appropriate finding of effective explanatory variables, identifying the linear/nonlinear relationships with the extreme events, and deciding the appropriate degree of nonlinearity to describe PLALMs. First, considering the large number of variables, several variables may be irrelevant and redundant; the variable selection scheme is thus a key element of model construction. Second, it is important to identify the linearity/nonlinearity for description of PLALMs for each of the selected variables. Finally, a decision on the number of bases for nonlinear transformation for the variables, which have nonlinear relationships with extreme events, is necessary to achieve the appropriate description of the logistic models. Considering this, we propose the following approach to identify the variables relevant for partially linear additive logistic modeling.

5.5 Identifying variables relevant to extreme imbalances

5.5.1 Selecting variables relevant to extreme imbalance events

It is necessary to determine the prospective variables that have a significant influence on the odds ratio of the target imbalance event while identifying the monotonic/nonmonotonic relationships between the variable and the extreme event. Further, the number of bases for nonlinear transformation need to be selected to describe the nonlinearities for each variable that have nonmonotonic relationships to the events. Figure 5-5 shows the overview of the proposed framework to identify the variables relevant for the construction of PLALMs. We introduce two types of relevant-redundancy measures to quantify the monotonic/nonmonotonic relevance between the variables and the observed extreme event. First, the filtering method, based on the relevant-redundancy measure, contributes to the identification of relevant variables from a large number of variables. Second, we provide an algorithm to find the monotonicity/nonmonotonicity trend of their relationships with the extreme event in which, from several relevant variables selected, the minimal necessary variables with nonmonotonic relationships are identified. Finally, we propose a forward step-wise approach for selecting the number of bases for nonlinear transformation that indicates the complexity of the nonmonotonic description in the model construction for each variable that has a nonmonotonic relationship with the extreme events. Consequently, we select effective models for construction of the model-based sensitivity analysis.

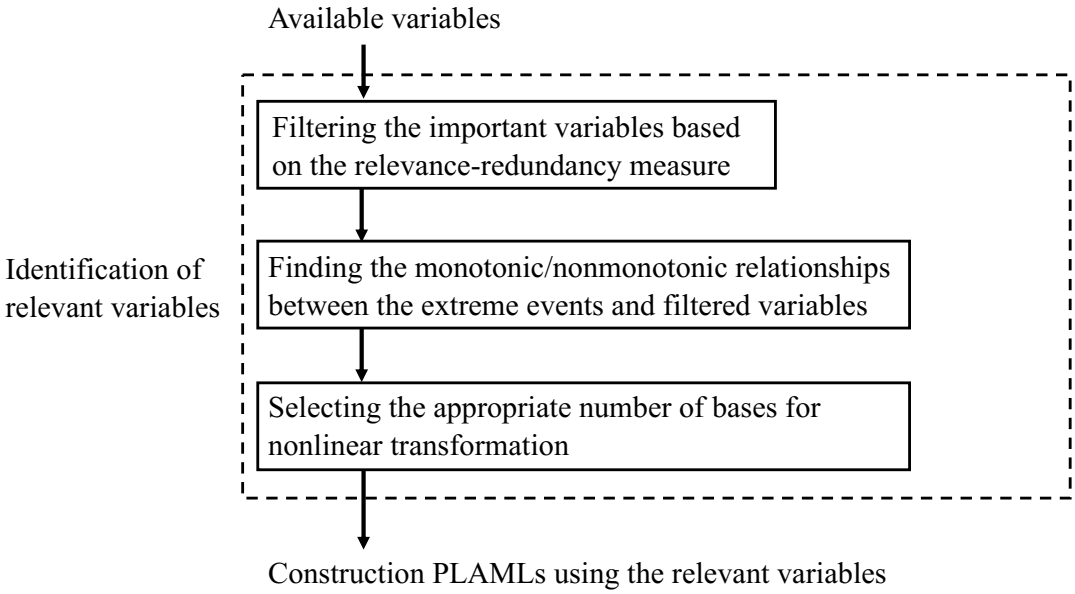


Fig. 5-5 Overview of the proposed framework for identifying the relevant variables.

5.5.1.1 Filtering method based on the relevance-redundancy measure

The identification scheme proposed aims to identify the subset of informative variables that are most relevant and least redundant in explaining extreme events while considering monotonic/nonmonotonic relationships. Variable filtering is adopted to extract the subset of relevant variables $\{\boldsymbol{x}_{sp}^{i} = [x_t^i; t \in \mathcal{T}_{sp}]; i \in \mathcal{S}'(\subset \mathcal{S})\}$ by using a criterion that measures the goodness of a variable [5-31]. Sun et al. [5-32] proposed an attractive concept called the relevance-redundancy measure, which is a criterion that quantifies the relative importance of $\boldsymbol{x}_{sp}^{i}$ to $\boldsymbol{g}_{sp}^{z}$ in comparison with a subset $\mathcal{S}' \subset \mathcal{S}\backslash\{i\}$, which is relevant to $\boldsymbol{g}_{sp}^{z}$ as follows:

$$R_{sp}^{z}\left(\boldsymbol{x}_{sp}^{i}\right) = I\left(\boldsymbol{g}_{sp}^{z}, \boldsymbol{x}_{sp}^{i}\right) - \varepsilon_{sp}^{z} \frac{1}{|\mathcal{S}'|} \sum_{j \in \mathcal{S}'} I\left(\boldsymbol{x}_{sp}^{i}, \boldsymbol{x}_{sp}^{j}\right), \tag{5.22}$$

where $I(\cdot)$ represents the dependency between two variables and ε_{sp}^{z} is a positive parameter. The first term of Eq. (5.22) evaluates the relevance between the extreme imbalance $\boldsymbol{g}_{sp}^{z}$ and variable $\boldsymbol{x}_{sp}^{i}$. The second term evaluates the redundancy of $\boldsymbol{x}_{sp}^{i}$ by quantifying the average relevance between $\boldsymbol{x}_{sp}^{i}$ and other explanatory variables $\boldsymbol{x}_{sp}^{j}(j \in \mathcal{S}')$ relevant to $\boldsymbol{g}_{sp}^{z}$. Equation (22) represents the balance between the relevance and redundancy of $\boldsymbol{x}_{sp}^{i}$, where $R_{sp}^{z}(\boldsymbol{x}_{sp}^{i}) > 0$ implies that $\boldsymbol{x}_{sp}^{i}$ has relatively important information relevant to $\boldsymbol{g}_{sp}^{z}$, which cannot be provided by other variables in $\mathcal{S}'$.

Algorithm 5-1 presents the details of the variable selection procedure implemented in our study using the relevance-redundancy measure. In the first step, the indices of the explanatory variables, $i \in \mathcal{S}$, are sorted in descending order of their dependency on $\boldsymbol{z}_{sp}^{z}$. Then, the informative variables are selected in a greedy manner by sequentially evaluating the relevance-redundancy measure $R_{sp}^{z}(\boldsymbol{x}_{sp}^{i})$ while maintaining the selected variable set $\mathcal{S}_{sp}^{z} \subseteq \mathcal{S}$ through exclusion of the redundant variables based on the approximate Markov blanket condition [5-15], [5-33]. By applying this algorithm to the extreme shortage/surplus in all seasonal situations (s, p), we obtain the index sets of the situation-dependent informative variable subsets, $\{\mathcal{S}_{sp}^{z}; \forall z, s, p\}$.

Algorithm 5-1: Algorithm for variable selection

Input:(z, s, p).
for $i = 1$ to V **do**
  Calculate $I(\boldsymbol{g}_{sp}^{z}, \boldsymbol{x}_{sp}^{i})$.
end for
$\mathcal{S}' \leftarrow$ Order $\mathcal{S}$ in descending $I(\boldsymbol{g}_{sp}^{z}, \boldsymbol{x}_{sp}^{i})$ value.
    #Sort the explanatory variables.
$\mathcal{S}_{sp}^{z} \leftarrow \emptyset$.
while $\mathcal{S}' \neq \emptyset$ **do**
  $i \leftarrow \mathcal{S}'^{(1)}$.
  $\mathcal{S}_{sp}^{z} \leftarrow \mathcal{S}_{sp}^{z} \cup \{i\}$.
  if $I(\boldsymbol{g}_{sp}^{z}, \boldsymbol{x}_{sp}^{i}) - \varepsilon_{sp}^{z} \frac{1}{|\mathcal{S}_{sp}^{z}|} \sum_{j \in \mathcal{S}_{sp}^{z}} I(\boldsymbol{x}_{sp}^{i}, \boldsymbol{x}_{sp}^{j}) > 0$ **then**
    $\mathcal{S}' \leftarrow \mathcal{S}' \setminus \{i\}$.
    $\mathcal{S}_{sp}^{z} \leftarrow \mathcal{S}_{sp}^{z} \cup \{i\}$. #Identify the informative variables
  else
    $\mathcal{S}' \leftarrow \mathcal{S}' \setminus \{i\}$.
  end if
  for j in $\mathcal{S}'$ **do**
    if $I(\boldsymbol{g}_{sp}^{z}, \boldsymbol{x}_{sp}^{j}) - \varepsilon_{sp}^{z} \frac{1}{|\mathcal{S}_{sp}^{z}|} \sum_{i \in \mathcal{S}_{sp}^{z}} I(\boldsymbol{x}_{sp}^{j}, \boldsymbol{x}_{sp}^{i}) < 0$ **then**
      $\mathcal{S}' \leftarrow \mathcal{S}' \setminus \{j\}$ #Remove the redundant variables
    end if
  end for
Output: $\boldsymbol{\mathcal{S}}_{sp}^{z}$.

5.5.1.2 Statistical dependency in relevance-redundancy measure

The relevance-redundancy measure introduced in Eq. (5.22) requires an index of statistical dependency between two variables. In this study, we introduce two types of dependency indices: Pearson's correlation [5-34] and the maximal information coefficient (MIC) [5-15]; the former represents monotonic dependency, while the latter represents the dependency regardless of whether it is monotonic or nonmonotonic, capturing a wide range of statistical relationships including linearity/nonlinearity. In this paper, we introduce a scheme for variable selection considering the nonmonotonic dependency between the variable and the extreme event plus identification of monotonicity/nonmonotonicity based on the linear/nonlinear dependency indices.

Pearson's correlation

Pearson's correlation ρ is one of the most popular criteria to quantify the statistical monotonic dependency between variables [34]. Given two variables $\mathbf{a} = [a_i, i = 1, \dots, N]$ and $\mathbf{b} = [b_i, i = 1, \dots, N]$, the correlation-based dependency index is defined as follows:

$$I(\mathbf{a}, \mathbf{b}) = |\rho(\mathbf{a}, \mathbf{b})| = \left|\frac{\mathrm{cov}(\mathbf{a}, \mathbf{b})}{\sigma_{\mathbf{a}} \sigma_{\mathbf{b}}}\right|, \tag{5.23}$$

where $\mathrm{cov}(\mathbf{a}, \mathbf{b})$ is the covariance between variables $\mathbf{a}$ and $\mathbf{b}$ and $\sigma_{\mathbf{a}}$ and $\sigma_{\mathbf{b}}$ are the

standard deviations of the variables **a** and **b**. Pearson's correlation ρ has the range $0 \leq |\rho(\mathbf{a},\mathbf{b})| \leq 1$, where $|\rho(\mathbf{a},\mathbf{b})| = 1$ suggests that variables **a** and **b** have a monotonic bijective projection function. Consequently, the utilization of the monotonic correlation as a dependency index in Eq. (5.22) is expected to provide a powerful relevance-redundancy measure for variable filtering[4].

Maximal information coefficient (MIC)

The maximal information coefficient (MIC) [5-15] is another statistical criterion to evaluate the dependency between variables. The MIC captures a wide range of statistical relationships between variables and quantifies the nonmonotonic dependency between variables based on mutual information (MI) [5-35].

$$MI(\mathbf{a},\mathbf{b}) = \sum_{a\in\mathbf{a}} \sum_{b\in\mathbf{b}} p(\mathrm{a},\mathrm{b}) \log \frac{p(\mathrm{a},\mathrm{b})}{p(\mathrm{a})p(\mathrm{b})}, \tag{5.24}$$

where $p(\mathrm{a},\mathrm{b})$ is the joint probability density and $p(\mathrm{a})$ and $p(\mathrm{b})$ are the marginal probability densities of the variables a and b. In the derivation of the MIC, the originally continuous variables a and b are discretized to ensure that the MI is evaluated using categorical probability densities. The MIC is defined as scaled maximal mutual information under various approaches to discretization of the variables. Let G be a grid partition that discretizes **a** and **b** into l-bins and m-bins [5-15]. Then, the MIC is defined as follows:

$$\begin{aligned} I(\mathbf{a},\mathbf{b}) &= MIC(\mathbf{a},\mathbf{b}) \\ &= \max_{G} \frac{MI(\mathbf{a},\mathbf{b}|G)}{\log \min\{l,m\}}. \end{aligned} \tag{5.25}$$

The MIC has the range $0 \leq MIC(\mathbf{a},\mathbf{b}) \leq 1$. When variables $(\mathbf{a},\mathbf{b})$ have a strong monotonic relationship, the MIC-based relevance measure as well as Pearson's correlation becomes optimal. Furthermore, when variables $(\mathbf{a},\mathbf{b})$ have a strong nonmonotonic relationship, the MIC-based relevance measure also becomes optimal. Consequently, the variable filtering approach based on the MIC-based relevance-redundancy measure is expected to select the informative variables while considering both monotonic and nonmonotonic variable relationships.

5.5.2 Identifying the monotonic relationships using bootstrap sampling test

In our scheme, the monotonicity/nonmonotonicity in relationships between the extreme events and each informative variable is identified based on a bootstrap sampling test using these linearity/nonlinearity dependency indicates. Ideally, when variables have nonmonotonic relationships to extreme events, the absolute value of the Pearson's correlation and the MIC tend to deviate; otherwise, these values are expected to behave similar. In this study, the monotonic relationships are identified by the test for comparison of statistical characteristics of the MIC/Pearson's values derived by bootstrap sampling. We expect the relevance-redundancy measure introduced in Eq. (5.22) to be useful in filtering the variables from the viewpoint of the given dependency index. To identify monotonicity/nonmonotonicity of the filtered variables with regard to extreme events, we propose the procedure based on the Wilcoxon signed rank test [5-36] using bootstrap sampling [5-37]. This procedure provides a nonparametric test for the

[4] We should stress that the Pearson's correlation may compromise robustness under high departure from a Gaussian distribution. Since we focused on this measure to quantify linearity here, the utilization of other well-known nonparametric measures such as rank correlation will be inappropriate. Further study will be needed to clarify the potential impact of using the Pearson's correlation in general problems.

difference in distribution of the Pearson's correlation and the MIC derived from the bootstrap samples.

Algorithm 5-2 shows the framework of the identification scheme. Here, $\mathcal{L}_{sp}^{z}$ is the index of the variables that have monotonic relationships with the extreme event, and $\mathcal{N}_{sp}^{z}$ is the index of the variables that have nonmonotonic relationships. For the filtered variables $\mathcal{S}_{sp}^{z}$, the bootstrap samples of $[(g_t^z, x_t^i); t \in \mathcal{T}_{sp}]$ are generated to derive a pair of dependency indices $\left([MIC^{i(b)}], [\rho^{i(b)}]\right)$. The Wilcoxon signed rank test is utilized to compare the significant difference between $[MIC^{i(b)}; b = 1, \dots, B]$ and $[\rho^{i(b)}; b = 1, \dots, B]$ for identifying the linearity/nonlinearity of the relationship. Let $Median_{MIC}^{i}$ and $Median_{\rho}^{i}$ be the medians of $[MIC^{i(b)}]$ and $[|\rho^{i(b)}|]$, respectively. The Wilcoxon signed-rank test for these bootstrap sample sets suggests the following:

•Variable x_{sp}^{i} tends to be monotonic to the extreme event, if $p.boot_{sp}^{zi} \geq 0.01$ $(Median_{MIC}^{i} = Median_{\rho}^{i})$ holds.

• Variable x_{sp}^{i} tends to be nonmonotonic to the extreme event, if $p.boot_{sp}^{zi} < 0.01$ $(Median_{MIC}^{i} \neq Median_{\rho}^{i})$ holds.

In such a procedure, we aim to identify whether the filtered variables have essential monotonic effects or nonmonotonic effects on the extreme event.

Algorithm 5-2: Algorithm for the identification scheme

Input:(z, s, p).
$\mathcal{L}_{sp}^{z} \leftarrow \emptyset$.
$\mathcal{N}_{sp}^{z} \leftarrow \emptyset$.
for $i \in \mathcal{S}_{sp}^{z}$ **do**
  for $i \in \{1, \dots, B\}$ **do**
    Generate $\left(\boldsymbol{g}_{sp}^{z(b)}, \boldsymbol{x}_{sp}^{i(b)}\right)$ by sampling from $[(\boldsymbol{g}_t^z, \boldsymbol{x}_t^i); t \in \mathcal{T}_{sp}]$.
    # Generate bootstrap samples.
    $MIC^{i(b)} \leftarrow MIC\left(\boldsymbol{g}_{sp}^{z(b)}, \boldsymbol{x}_{sp}^{i(b)}\right)$.
    $\rho^{i(b)} \leftarrow \rho\left(\boldsymbol{g}_{sp}^{z(b)}, \boldsymbol{x}_{sp}^{i(b)}\right)$.
  end for
  $p.boot_{sp}^{zi} \leftarrow$ Test for paired data $\left([MIC^{i(b)}], [|\rho^{i(b)}|]\right)$.
  # Wilcoxon signed-rank test.
  if $p.boot_{sp}^{zi} < 0.01$ **then**
    $\mathcal{N}_{sp}^{z} \leftarrow \mathcal{N}_{sp}^{z} \cup \{i\}$.
  else
    $\mathcal{L}_{sp}^{z} \leftarrow \mathcal{L}_{sp}^{z} \cup \{i\}$.
  end if
end for
Output: $(\mathcal{L}_{sp}^{z}, \mathcal{N}_{sp}^{z})$.

5.5.3 Selecting the number of bases for nonlinear transformation by a forward stepwise algorithm

We propose a forward stepwise algorithm to select the number of bases to describe the nonlinearities for each variable that has a nonmonotonic relationship with extreme events, as expressed in Eq. (5.13) and (5.17). An outline of the implementation is provided in Algorithm 5-3. This may be considered a greedy algorithm for selecting nonlinear bases as it starts to evaluate the initial number of bases and incrementally increases the bases to achieve the best accuracy of the model. Consequently, this algorithm efficiently identifies the appropriate number of bases for nonlinear transformation for each variable.

Algorithm 5-3: Algorithm for the selection of the number of bases for nonlinear transformation

1: **Input**: $\mathcal{N}_{sp}^{z}$.
2: **Initialize** the follwing values as follows:
3: Set $\boldsymbol{K}_0 = \{K^i\} = \{2, \dots, 2\}$ $(i \in \mathcal{N}_{sp}^{z})$.
#Index of the number of bases for variables that have nonmonotonic relationships with extreme events.
4: Construct models with $\boldsymbol{K}_0$ bases and set m_0 #Accuracy of models with $\boldsymbol{K}_0$ bases.
5: **Update** the index set $\boldsymbol{K}_d$ and the accuracy m_d $(1 < d < D)$ as follows:
6: **while** $(m_d > m_{d-1})$
7: **for** $i \in \{\mathcal{N}_{sp}^{z}\}$ **do**
8: $K^i \leftarrow K^i + 1$ #Update the index set $\boldsymbol{K}_d$.
9: Evaluate the model description accuracy m_d^i with $\boldsymbol{K}_d$ bases.
10: **if** $m_d^i > m_{d-1}$ **then**
11: $m_d \leftarrow m_d^i$.
12: **end if**
13: $K^i \leftarrow K^i - 1$.
14: **end for**
15: **end while**
16: **Output**: $(\boldsymbol{K}_{d-1})$.

5.6 Case study

5.6.1 Simulation setup

Informative variables for extreme events under various seasonal situations were derived based on the dataset collected in the Japanese settlement system. Table 5-1 lists the variables used in this evaluation, comprising 214 explanatory variables belonging to typical categories, such as weather, calendar, capacity of the interconnection line, market price, and power supply configuration. All explanatory variables were standardized to have zero mean and unit variance. Data were collected from January 2017 to December 2019. Actual extreme imbalance data set collected from January 2017 to December 2018 was used to train the models so that it could estimate extreme imbalance events from January 2019 to December 2019 for evaluation. In the proposed variable selection scheme, we tuned the parameters ε_{sp}^{z} based on cross-validation [5-38], and the parameter $B = 10$ which controls the number for bootstrap sampling.

In this evaluation, we focus on the five models listed in Table 5-2 to describe the probability of occurrence of the target extreme event. The first model (LM (Limited)) is a logistic model that uses a limited number of explanatory variables (i.e., the area price and renewable energy power). Note that this model is a naive extension of the models in recent studies [5-6 , 5-7]. The second model is a logistic model (LM (All)) similar to the first model, but utilizes a large number of variables that may affect extreme events. The third model (LM) is a logistic model that uses the variable filter method based on the Pearson's correlation -based relevance-redundancy measure; the informative variables are based on the assumption of linearity. The fourth model (ALM) focuses on the additive logistic model, which uses the variable filter method based on the MIC-based relevance-redundancy measure and the selection of bases for nonlinear transformation; the informative variables are based on the assumption of nonlinearity. The fifth model (PLALM) describes the events using PLALMs, identifying the linearity/nonlinearity to the odds ratio of the events by implementing the procedure described in Section 5.5 for the identification of the relevant variables. We compared these models from the viewpoints of description accuracy and sensitivities of the relevant variables.

Table 5-1 Categories of variables.

Categories	Attributes	Variables	Source
Weather [a]	Temperature, precipitation, wind speed, humidity, solar radiation, and daylight hours	v1-v60	[5-39]
Capacity of interconnection line [b]	Free capacity (forward/reverse) and planar capacity	v61-v93	[5-40]
Electricity price [c]	Market price (DA)	v94-v103	[5-41]
Electricity supply [a]	Regional total, Nuclear, Thermal, Hydro, Geothermal, Biomass, Solar, Solar suppression, Wind, Wind suppression, Pumped storage	v104-v213	[5-40]
Calendar data	Holiday/weekday; binary dummy	v214	

[a] Variables are corrected in each Japanese region area: Hokkaido, Tohoku, Kanto, Chubu, Hokuriku, Kansai, Chugoku, Shikoku, Kyusyu, and Okinawa.

[b] Variables are corrected in each Japanese interconnection line from Hokkaido to Tohoku, from Tohoku to Kanto, from Kanto to Chubu, from Chubu to Kansai, from Chubu to Hokuriku, from Hokuriku to Kansai, from Kansai to Chugoku, from Kansai to Shikoku, from Chugoku to Shikoku, from Chugoku to Kyusyu and from Chubu and Kansai to Hokuriku.

[c] Variables are corrected in each Japanese region area: Hokkaido, Tohoku, Kanto, Chubu, Hokuriku, Kansai, Chugoku, Shikoku, and Kyusyu.

Table 5-2 Models used in the simulation.

Model	Variables	Type	Note
LM (Limited)	Area price, Solar power, Wind power	Logistic model	–
LM (All)	All variables (214 variables)	Logistic model	–
LM		Logistic model	Utilize the Pearson's correlation -based filter method
ALM		Additive logistic model	Utilize the MIC-based filter method and the forward step-wise scheme for nonlinear transformation
PLALM		Partially linear additive logistic model	Proposed approach applying the process introduced in Section 5.5

5.6.2 Description accuracy of constructed models

First, we derived the models listed in Table 5-2 based on the dataset and compared them from the viewpoint of description accuracy determined by the F-measure between the estimates. The estimated imbalance event $\hat{g}_t^z$ during the evaluation period is a binary value defined as follows:

$$\hat{g}_t^z = \begin{cases} 1 & \text{if } c_t^z(\boldsymbol{x}_t; \boldsymbol{\beta}_{sp}^z, \boldsymbol{\tau}_{sp}^z) \geq 0.5 \\ 0 & \text{otherwise.} \end{cases}. \tag{5.26}$$

Here, the F-measure is defined as follows:

$$Precision = \frac{\sum_{t \in \mathcal{T}_{sp}} g_t^z \hat{g}_t^z}{\sum_{t \in \mathcal{T}_{sp}} \hat{g}_t^z} \tag{5.27}$$

$$Recall = \frac{\sum_{t \in \mathcal{T}_{sp}} g_t^z \hat{g}_t^z}{\sum_{t \in \mathcal{T}_{sp}} g_t^z} \tag{5.28}$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{5.29}$$

The F-measure was scaled such that $0 \leq F \leq 1$, where $F = 1$ for a model implies that it achieves high description accuracy.

Figure 5-6 shows the average F-measures derived for each model. The F-measure quantifies the representation power for the test data. The figure shows that the results derived by the statistical modeling achieve higher accuracy than persistence results. The figure demonstrates that models constructed using a large number of variables achieve higher description accuracy than those constructed using a limited number of variables, implying the existence of informative variables other than the price and renewable power variables used in previous studies. These results also show that LM, ALM, and PLALM using the selection process achieved higher description accuracies than other models. Thus, the variable selection scheme tends to improve the description accuracy. Moreover, the PLALM achieves accuracy with the highest representation power, suggesting the existence of essential monotonic relationships with the extreme events. The proposed scheme for identifying monotonic/nonmonotonic variables works well to clarify selection of the minimally necessary nonmonotonic relationships amongst variables. Figure 5-7 shows the average and standard deviation of the F-measure derived for each time period targeting extreme shortage and extreme surplus power supply, respectively. These results suggested that PLALMs achieve high description accuracy specifically for the case of extreme surplus.

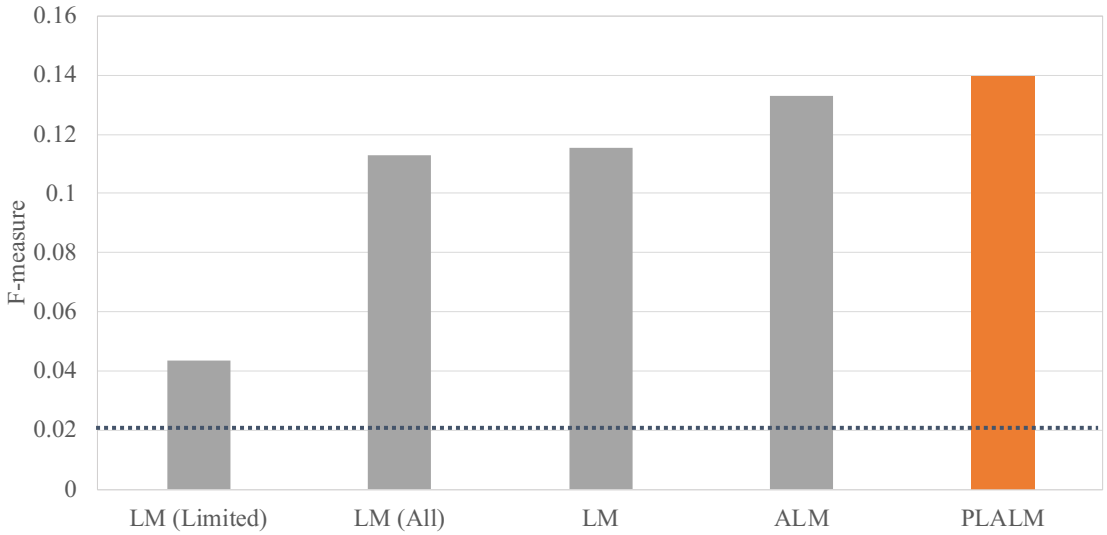
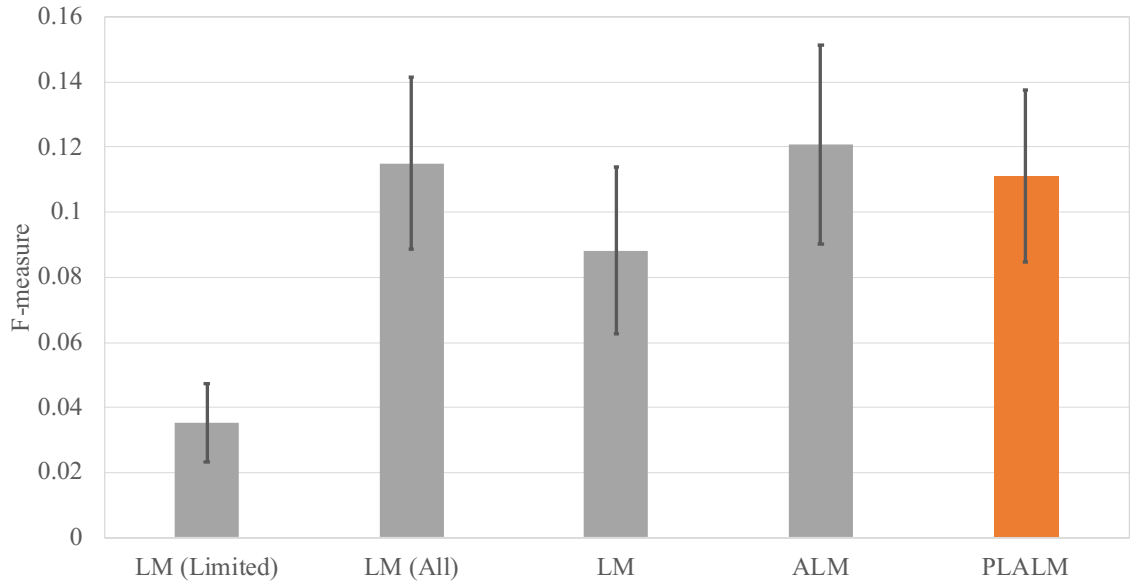
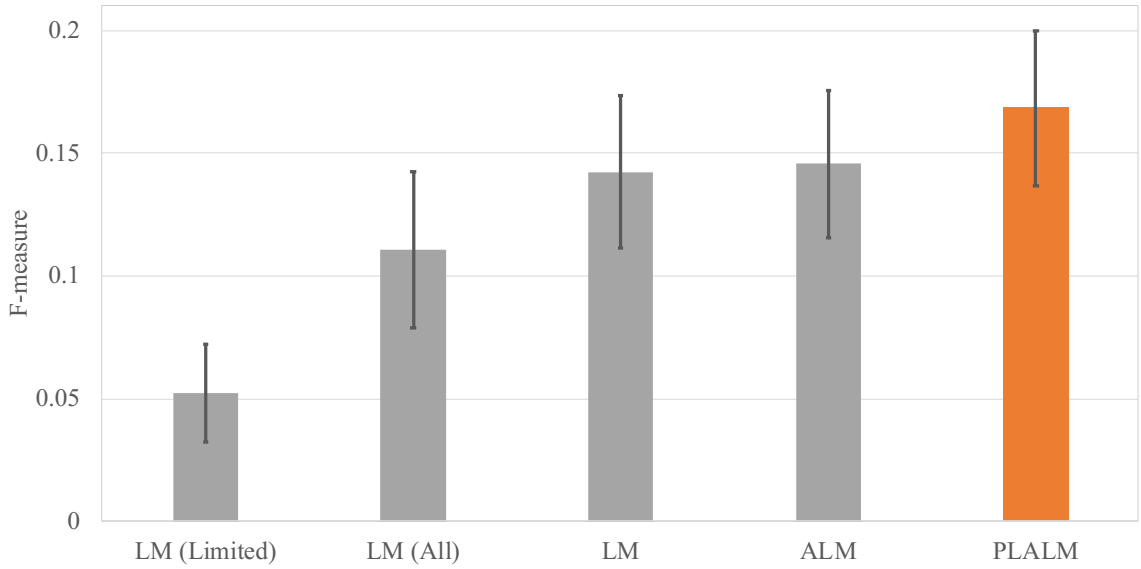


Fig. 5-6 Average F-measures for the evaluation of the constructed models. The dotted line indicates the accuracy used persistence results that are calculated by the value in the previous day.



(a) Extreme shortage



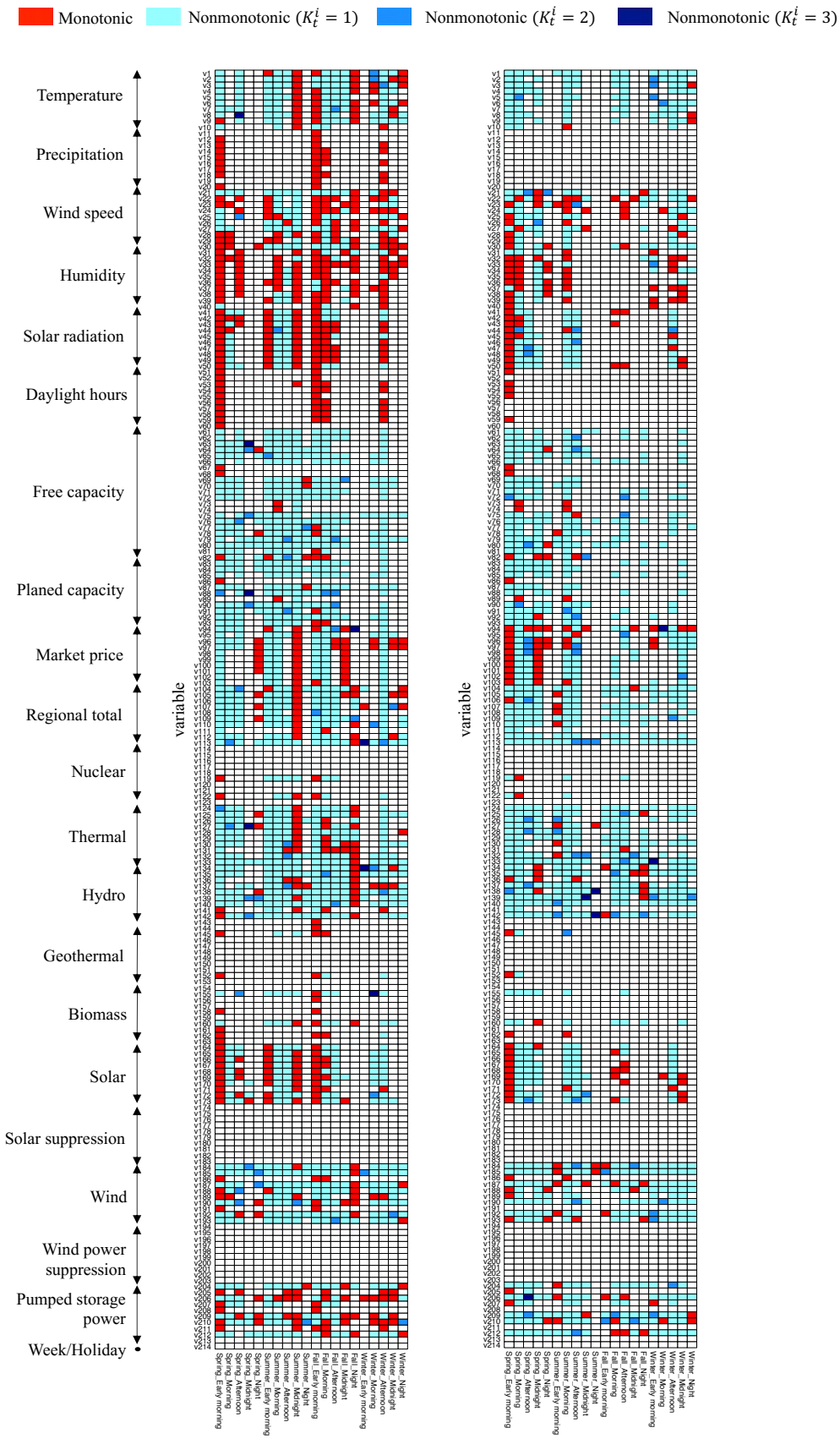
(b) Extreme surplus

Fig. 5-7 Average F-measures for the evaluation of constructed models targeting: (a) extreme shortage and (b) extreme surplus. Error bars show the standard deviation of the F-measure derived under $\boldsymbol{\mathcal{T}_{sp}}$ $(\forall \boldsymbol{s}, \boldsymbol{p})$.

5.6.3 Discussion regarding the selected variables

The results derived using PLALMs in the selection process to identify monotonic/nonmonotonic relationships of the variables are shown in Figure 5-8. Figures (a) and (b) show the most significant variables selected for describing extreme shortage and surplus, respectively, as well as the monotonic/nonmonotonic relationships to the odds ratio of extreme events with bases K_t^i for nonlinear transformation. These results suggest that the informative variables for the extreme shortage/surplus have different characteristics in various seasonal situations. For extreme shortage events, the temperature, wind speed, and humidity tend to be important in all seasons. Other variables such as precipitation, solar radiation, and daylight hours tend to be important in both early morning and morning, although solar radiation in summer is important at all time frames except at night. These results suggest that weather information affect various events such as people's lifestyles and the amount of renewable energy generation, which may affect the occurrences of extreme imbalance. Further, many of the weather variables have monotonic relationships with the extreme events. Meanwhile, for extreme surplus events, the results suggest that the important weather variables tend to be less significant when compared to shortage events. The temperature, wind speed, humidity, and solar radiation are generally selected throughout the year. The capacity of the interconnection line tends to be important for all seasons targeting extreme shortage, but only spring and summer when targeting extreme surplus. The electricity price is selected at all seasons targeting shortage and surplus events.

Focusing on the electricity supply variables, the results suggest that large-scale power generation, such as thermal power, hydropower, and pumped storage power, and distributed power generation such as solar and wind power are mainly selected at several time frames targeting shortage/surplus events. The results also suggest that variables that were not previously considered in several attributes have been selected to describe the extreme shortage/surplus, including minor renewable resources such as geothermal power and biomass. Further, several variables such as system price and the electricity supply from hydropower in Hokkaido needed a large number of bases for nonlinear transformation, suggesting that such variables have complex nonmonotonic relationships with the extreme events. Figure 5-9 shows the frequency with which individual variables were adopted in the models constructed under 20 different situations. The tendency for variables to be informative for extreme shortage and surplus is similar, but the results suggest that extreme shortage imbalance tends to be affected by the additional factors such as precipitation and daylights hours.



(a) Extreme shortage (a) Extreme surplus

Fig. 5-8 Selected variables derived by using PLALMs. The x-axis represents the selected variables, and the y-axis represents the transition of the situation $(\boldsymbol{s}, \boldsymbol{p})$.

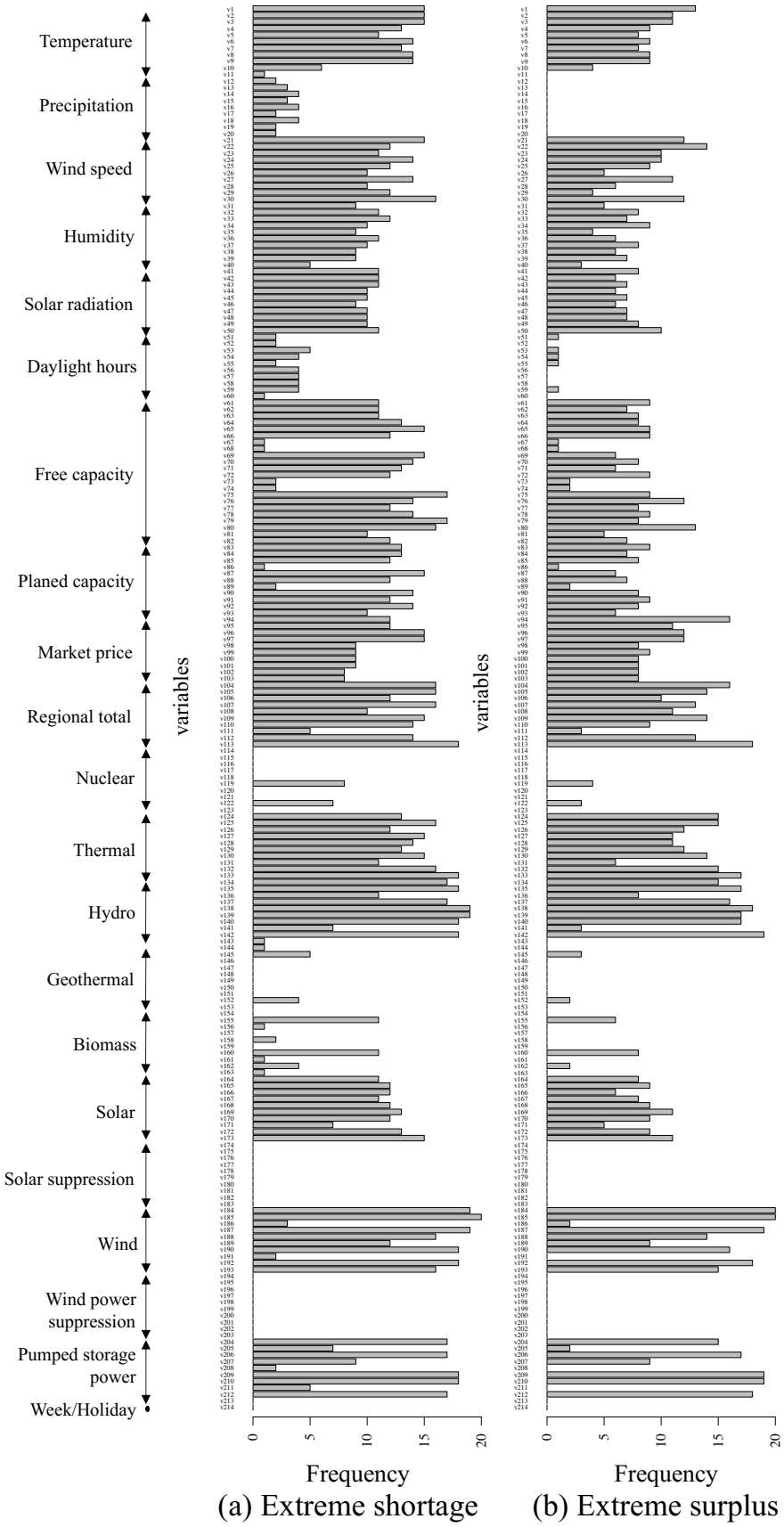


Fig. 5-9 Frequency with which individual variables were adopted in the models constructed under 20 different situations.

Here, we focus on variables the capacity of the interconnection line from Chugoku to Kyusyu (i.e., v79) and the wind speed in Kanto (i.e., v23), which are selected during early summer morning for describing extreme surplus, and the capacity of the interconnection line from Kanto to Chubu (i.e., v65) which are eliminated during the same situation. Figure 5-10 shows two components appeared in Eq. (5.22) representing relevance and redundancy of the MIC-based relevance-redundancy measure of each variable. The figure suggests that v79 clearly has a high relevance with extreme events. The figure also shows that v23 is an informative variable for describing extreme events with low redundancy, although the relevance is not significantly high. On the other hand, v65 is a redundant variable that is strongly related to other variables. The result suggests that the proposed filter method selected the informative variables for describing extreme events, while screening the irrelevant and redundant variables. We also discuss the monotonicity/nonmonotonicity focusing on the capacity of the interconnection line from Chugoku to Kyusyu (v79) and the wind speed in Kanto (v23) in Fig. 5-11. Figure 5-11 (a) and (b) respectively shows the distribution of the relevance measure derived by the evaluation of the Pearson's correlation and MIC for the bootstrap sample sets. Figure 5-11 (a) shows that MIC tends to be larger than the Pearson's correlation; it suggests that v79 has nonmonotonic relationship with extreme surplus. Meanwhile, Figure 5-11 (b) shows the Pearson's correlation tend to be large, and the distribution of MIC is roughly encompassed by the distribution of Pearson's correlation, indicating that the v23 has essential monotonic relationship with extreme events. These results suggest that the proposed scheme based on the bootstrap sampling works well to identify monotonicity/nonmonotonicity among variables.

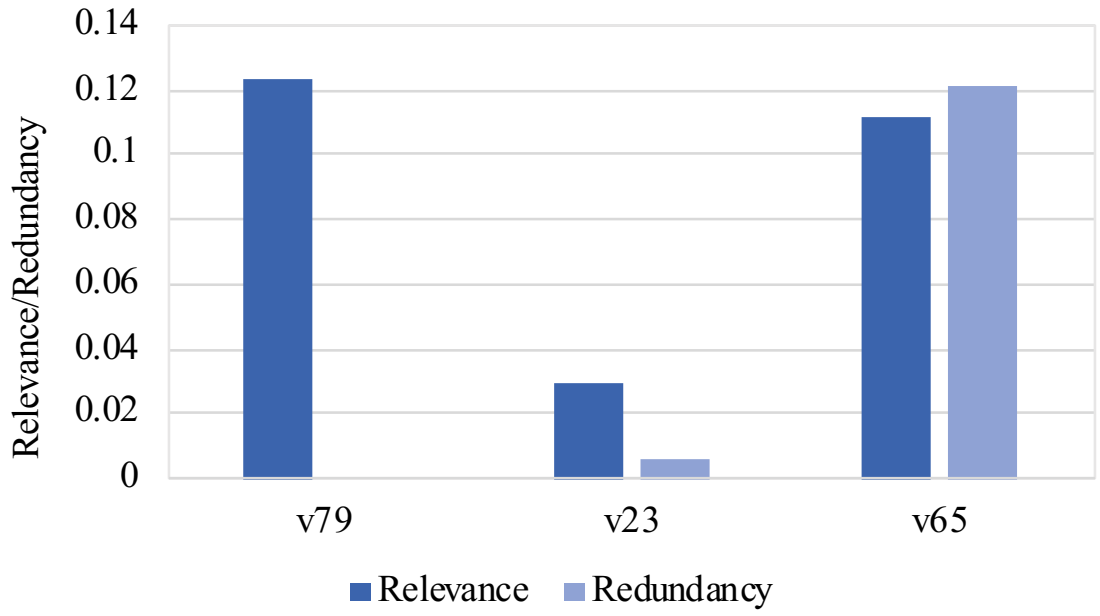
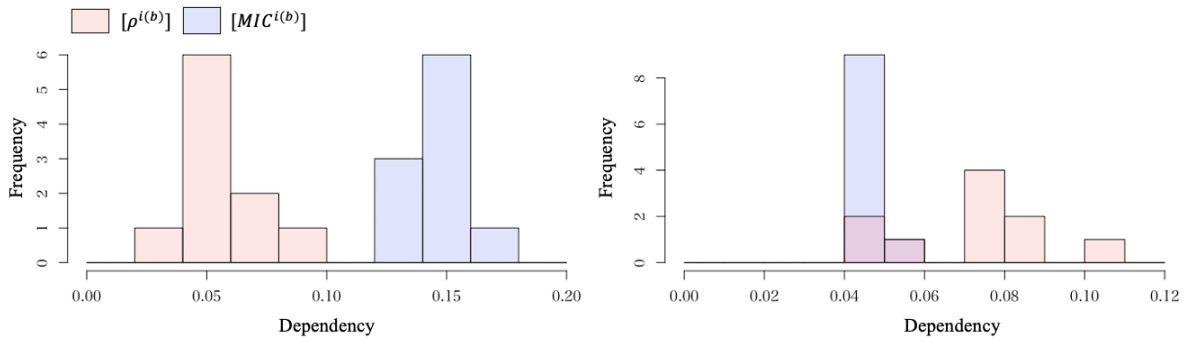


Fig. 5-10 Relevance and redundancy of the MIC-based relevance-redundancy measure given in Eq. (5.22) for observed explanatory variables: capacity of the interconnection line from Chugoku to Kyusyu (v79), wind speed in Kanto (v23), and capacity of the interconnection line from Kanto to Chubu (v65).



(a) v79
(Capacity of the interconnection line from Chugoku to Kyusyu)

(b) v23
(Wind speed in Kanto)

Fig. 5-11 Distribution of the relevance measure derived by the evaluation of the Pearson's correlation and MIC for the bootstrap sample sets of observed explanatory variables: capacity of the interconnection line from Chugoku to Kyusyu (v79) and wind speed in Kanto (v23).

As an of example of the applicability of our PLALM scheme, we focus on several informative variables and analyze their influence on an extreme shortage event. Specifically, we focus on the regional total of the electricity supply in Chubu (i.e., v107), electricity supply from hydropower in Hokkaido (i.e., v134), and that in Tohoku (i.e., v135) during an early winter morning. Figure 5-12 shows scatterplots of the extreme shortage versus the observed explanatory variables: (a) regional total of the electricity supply in Chubu, (b) the electricity supply from hydropower in Hokkaido, and (c) the electricity supply from hydropower in Tohoku. The solid line shows the probability trends of the imbalance targeting the observed explanatory variables. Figure 5-13 show the main effect of each variable on the odds ratio of the extreme shortage, which is described as $f^i(x_t^i)$ in Eq. (5.19), corresponding to Fig. 5-12. These results show that the proposed scheme works well for identifying the monotonic/nonmonotonic relationships of the variables. The results suggest that the contributions of each target variable to the extreme events are different; the extreme shortage monotonically increases with the regional total of electricity supply in Chubu, and nonmonotonically changes depending on the electricity supply from hydropower in Hokkaido and Tohoku in the target seasonal situation. Further, the electricity supply from hydropower in Hokkaido and Tohoku requires a different number of bases K_t^i for describing the main effect, as expressed in Eq. (5.17): the hydropower system in Tohoku requires two bases, while that in Hokkaido requires four bases. These results suggest that the electricity supply from hydropower in Hokkaido may have a frequent change in sensitivity to the odds ratio of extreme shortage events in variable space. Figure 5-14 shows the sensitivity to the odds ratio of the extreme event defined in Eq. (5.20) corresponding to Figs. 5-12 and Fig. 5-13. These figures indicate the sensitivities to the extreme event with respect to the variables: the regional total of electricity supply in Chubu shows a positive immutable variable, while that from hydropower in Hokkaido and Tohoku shows a nonmonotonic relationship with the extreme event. The sensitivity increases in the negative direction for large values of the hydropower variables in both of these regions. The hydropower in Hokkaido also shows locally positive sensitivity in the variable space.

As a final example of the performance of the PLALM scheme, we focus on certain explanatory variables observed in several seasonal situations and discuss the response sensitivity to the odds ratio of extreme shortage and surplus events. Figure 5-15 shows the sensitivities of the odds ratio of extreme shortage and surplus with respect to the electricity supply from wind power in Hokkaido (i.e., v184) in several seasonal situations. These results show that the absolute value of sensitivity tends to increase based on the increase in the amount of wind power generation; thus, the sensitivity to the odds ratio of the extreme events increases with the share of wind power generation in the market. In contrast, the sensitivity to the odds ratio of the extreme shortage with respect to wind power generation in Hokkaido tends to be negative and nearly constant for small values such that the probability of the occurrence of shortage events decreases monotonically with the change of the variable. To further clarify this, the sensitivity at early morning in winter rapidly becomes large for large values around 200 MWh. Meanwhile, the sensitivity to the odds ratio of the extreme surplus with respect to the wind power generation in Hokkaido tends to be negatively large for larger values. Specifically, the result shows that the sensitivity at midnight in fall has a large positive sensitivity around 150 MWh; these results may be influenced by uncertainties in power demand and planning by sudden wind power outages during wind speed. Overall, these results demonstrate that our proposed PLALM approach successfully accounts for the characteristics of seasonal sensitivity and identifies locations in the variable space where the model presents the highest variability with respect to the target variable. These analysis help stakeholders to identify situations in which the probability of an extreme imbalance event is likely to increases/decreases locally, and play key role in decision making and power system management.

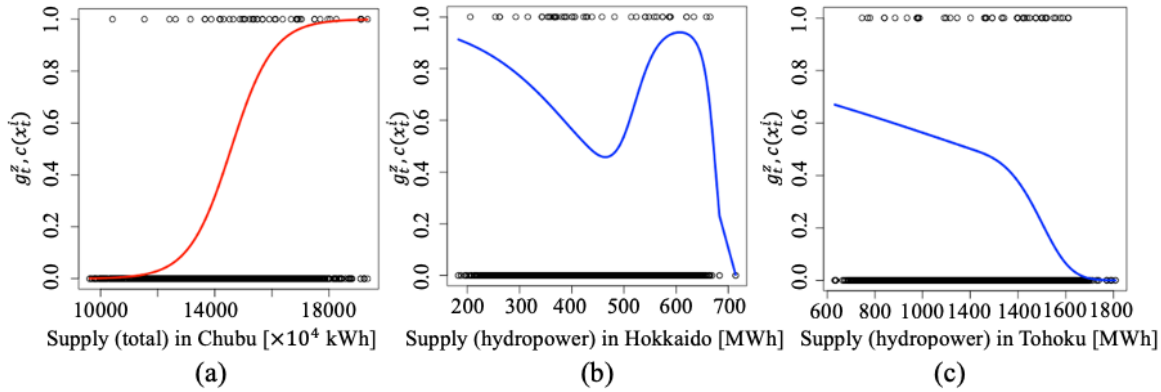


Fig. 5-12 Relationship between the target extreme shortage and observed explanatory variables: (a) regional total of the electricity supply in Chubu (v107), (b) electricity supply from hydropower in Hokkaido (v134), and (c) electricity supply from hydropower in Tohoku (v135). The scatterplots show the relationships between observed variables and extreme events, and the solid lines show the the probability trends of the imbalance targeting the observed explanatory variables.

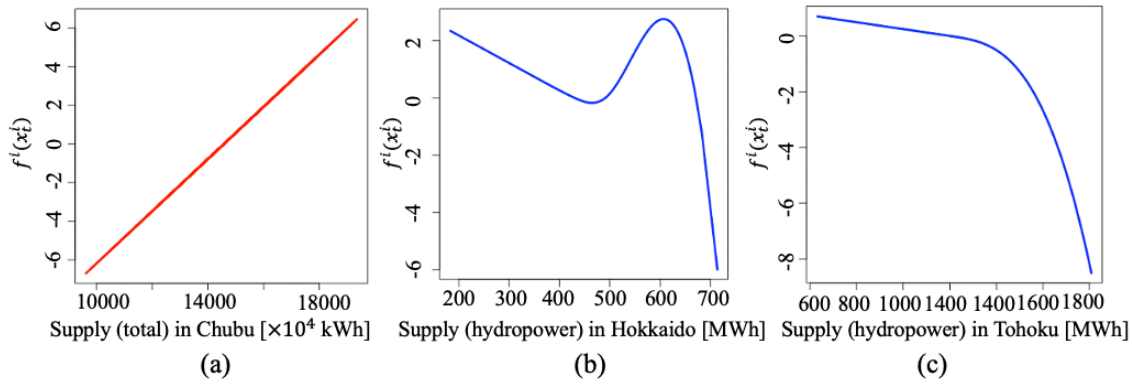


Fig. 5-13 Main effects of (a) the regional total of the electricity shortage in Chubu (v107), (b) electricity supply from hydropower in Hokkaido (v134), and (c) electricity supply from hydropower in Tohoku (v135).

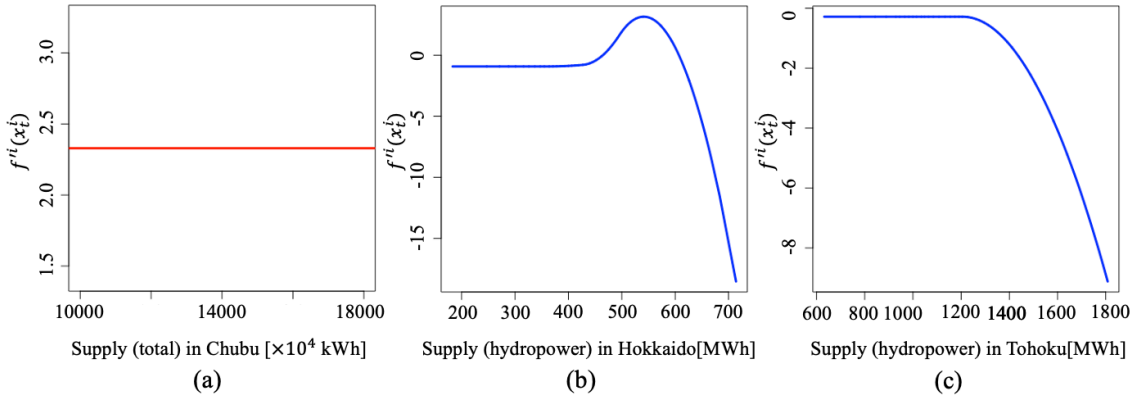


Fig. 5-14 Sensitivities of the extreme shortage with respect to (a) the regional total of electricity supply in Chubu (v107), (b) electricity supply from hydropower in Hokkaido (v134), and (c) electricity supply from hydropower in Tohoku (v135).

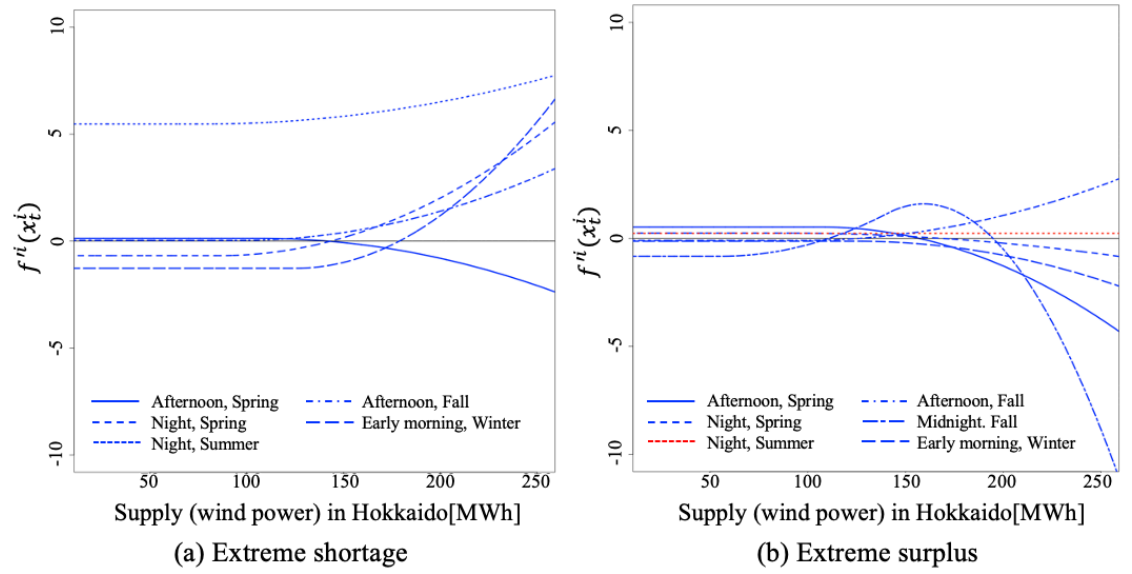


Fig. 5-15 Sensitivities of (a) the extreme shortage and (b) extreme surplus with respect to the electricity supply from wind power in Hokkaido (v184).

5.7 Concluding remarks in the chapter

In this study, we discussed the characteristics of extreme shortage/surplus events in various seasonal situations and elucidated the influence of relevant variables on the odds ratio of extreme events. We focus on the model-based analysis scheme using the partially linear additive models (PLALMs) to determine the response sensitivity of relevant factors to the odds ratio of extreme imbalance events. Furthermore, we developed a scheme to determine the prospective variables that have a significant influence on the odds ratio of the target imbalance event, including a relevance-redundancy measure to find variables relevant to the extreme imbalance event while identifying the monotonic/nonmonotonic relationships between the variable and the extreme event, combined with a forward step-wise selection scheme for the number of bases needed for nonlinear transformation of each variable that has a nonmonotonic relationship with extreme events. The proposed approach successfully identifies the seasonal sensitivity of the relevant variables while identifying the monotonic and nonmonotonic relationships between the variables and the extreme event. Our case study results using real-world data suggest the following key findings:

1. The identification scheme for the variables relevant to extreme imbalances using a relevance-redundancy measure and a forward step-wise selection scheme achieves to identify the informative variables required to describe the extreme event.
2. The model-based approach using PLALMs achieves to describe the impact of each informative variable on the extreme event flexibly under various seasonal situations.
3. The proposed sensitivity analysis reveals the interpretable influence of individual relevant variables while considering the monotonic/nonmonotonic relationships between the variable and the extreme event.
4. The proposed sensitivity analysis for factors affecting the extreme imbalance contributes to the understanding of the statistical behavior of imbalance in many countries where electricity is deregulated.

Reference

[5-1] T. Fujimi and S. E. Chang, "Adaptation to electricity crisis: Businesses in the 2011 Great East Japan triple disaster," *Energy policy*, vol. 68, pp. 447–457, 2014, doi: 10.1016/j.enpol.2013.12.019.

[5-2] O. Kimura and K. I. Nishio, "Responding to electricity shortfalls," *Economics of energy & environmental policy*, vol. 5, no. 1, pp. 51–72, 2016, doi: 10.5547/2160-5890.5.1.okim.

[5-3] P. Drobinski, M. Mougeot, D. Picard, R. Plougonven, and P. Tankov, *Renewable Energy: Forecasting and Risk Management Paris, France, June 7-9, 2017*, 1st ed. 2018. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-99052-1.

[5-4] D. Chattopadhyay, "Modelling renewable energy impact on the electricity market in India," *Renewable & sustainable energy reviews*, vol. 31, pp. 9–22, 2014, doi: 10.1016/j.rser.2013.11.035.

[5-5] N. Kaneko, Y. Fujimoto, and Y. Hayashi, "Graphical Modeling for Analysis of Hourly Electricity Demand and Market Price," *International Conference on the European Energy Market, EEM*, vol. 2020-September, Sep. 2020, doi: 10.1109/EEM49802.2020.9221986.

[5-6] S. Goodarzi, H. N. Perera, and D. Bunn, "The impact of renewable energy forecast errors on imbalance volumes and electricity spot prices," *Energy Policy*, vol. 134, p. 110827, 2019, doi: 10.1016/j.enpol.2019.06.035.

[5-7] R. Aïd, P. Gruet, and H. Pham, "An optimal trading problem in intraday electricity markets", doi: 10.1007/s11579-015-0150-8.

[5-8] J. Usaola and M. Á. Moreno, "Optimal bidding of wind energy in intraday markets," *2009 6th International Conference on the European Energy Market, EEM 2009*, 2009, doi: 10.1109/EEM.2009.5207179.

[5-9] M. Bueno-Lorenzo, M. Á. Moreno, and J. Usaola, "Analysis of the imbalance price scheme in the Spanish electricity market: A wind power test case," *Energy policy*, vol. 62, pp. 1010–1019, 2013, doi: 10.1016/j.enpol.2013.08.039.

[5-10] F. Lisi and E. Edoli, "Analyzing and Forecasting Zonal Imbalance Signs in the Italian Electricity Market," *The Energy journal (Cambridge, Mass.)*, vol. 39, no. 5, p. 1, 2018, doi: 10.5547/01956574.39.5.flis.

[5-11] N. Kaneko, Y. Fujimoto, S. Kabe, M. Hayashida, and Y. Hayashi, "Sparse modeling approach for identifying the dominant factors affecting situation-dependent hourly electricity demand," *Applied Energy*, vol. 265, no. December 2019, p. 114752, 2020, doi: 10.1016/j.apenergy.2020.114752.

[5-12] R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss, "Semiparametric Estimates of the Relation between Weather and Electricity Sales," *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 310–320, 1986, doi: 10.1080/01621459.1986.10478274.

[5-13] Y. Fujimoto, S. Murakami, N. Kaneko, H. Fuchikami, T. Hattori, and Y. Hayashi, "Machine learning approach for graphical model-based analysis of energy-aware growth control in plant factories," *IEEE Access*, vol. 7, pp. 32183–32196, 2019, doi: 10.1109/ACCESS.2019.2903830.

[5-14] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "Filter Approach Feature Selection Methods to Support Multi-label Learning Based on ReliefF and Information Gain," in *Advances in Artificial Intelligence - SBIA 2012*, Berlin,

Heidelberg: Springer Berlin Heidelberg, pp. 72–81. doi: 10.1007/978-3-642-34459-6_8.
[5-15] G. Sun, J. Li, J. Dai, Z. Song, and F. Lang, "Feature selection for IoT based on maximal information coefficient," *Future generation computer systems*, vol. 89, pp. 606–616, 2018, doi: 10.1016/j.future.2018.05.060.
[5-16] K. Honjo, H. Shiraki, and S. Ashina, "Dynamic linear modeling of monthly electricity demand in Japan: Time variation of electricity conservation effect," *PloS one*, vol. 13, no. 4, pp. e0196331–e0196331, 2018, doi: 10.1371/journal.pone.0196331.
[5-17] S. Razavi and H. V. Gupta, "What do we mean by sensitivity analysis? The need for comprehensive characterization of 'global' sensitivity in Earth and Environmental systems models," *Water resources research*, vol. 51, no. 5, pp. 3070–3092, 2015, doi: 10.1002/2014WR016527.
[5-18] J. Nossent, P. Elsen, and W. Bauwens, "Sobol' sensitivity analysis of a complex environmental model," *Environmental modelling & software : with environment data news*, vol. 26, no. 12, pp. 1515–1525, 2011, doi: 10.1016/j.envsoft.2011.08.010.
[5-19] R. Lidén and J. Harlin, "Analysis of conceptual rainfall–runoff modelling performance in different climates," *Journal of hydrology (Amsterdam)*, vol. 238, no. 3, pp. 231–247, 2000, doi: 10.1016/S0022-1694(00)00330-9.
[5-20] H. LeTreut, *Climate Sensitivity to Radiative Perturbations Physical Mechanisms and Their Validation*, 1st ed. 1996. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996. doi: 10.1007/978-3-642-61053-0.
[5-21] M. K. Muleta and J. W. Nicklow, "Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model," *Journal of hydrology (Amsterdam)*, vol. 306, no. 1, pp. 127–145, 2005, doi: 10.1016/j.jhydrol.2004.09.005.
[5-22] N. Huang, G. Lu, and D. Xu, "A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest," *Energies (Basel)*, vol. 9, no. 10, p. 767, 2016, doi: 10.3390/en9100767.
[5-23] R. Miller, L. Golab, and C. Rosenberg, "Modelling weather effects for impact analysis of residential time-of-use electricity pricing," *Energy Policy*, vol. 105, pp. 534–546, 2017, doi: 10.1016/j.enpol.2017.03.015.
[5-24] T. Hastie, *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*, 2nd ed. 2009. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.
[5-25] I. Rish and G. Grabarnik, *Sparse Modeling*. Baton Rouge: CRC Press, 2015. doi: 10.1201/b17758.
[5-26] A. J. Chapman and K. Itaoka, "Energy transition to a future low-carbon energy society in Japan's liberalizing electricity market: Precedents, policies and factors of successful transition," 2017, doi: 10.1016/j.rser.2017.06.011.
[5-27] K. Muralitharan, R. Sakthivel, and Y. Shi, "Multiobjective optimization technique for demand side management with load balancing approach in smart grid," *Neurocomputing (Amsterdam)*, vol. 177, pp. 110–119, 2016, doi: 10.1016/j.neucom.2015.11.015.
[5-28] A. Z. Al-Garni, S. M. Zubair, and J. S. Nizami, "A regression model for electric-energy-consumption forecasting in Eastern Saudi Arabia," *Energy (Oxford)*, vol. 19, no. 10, pp. 1043–1049, 1994, doi: 10.1016/0360-5442(94)90092-2.

[5-29] T. W. Yee, *Vector Generalized Linear and Additive Models With an Implementation in R*, 1st ed. 2015. New York, NY: Springer New York, 2015. doi: 10.1007/978-1-4939-2818-7.
[5-30] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *34th International Conference on Machine Learning, ICML 2017*, 2017, vol. 4, pp. 2976–2987.
[5-31] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005, doi: 10.1109/TPAMI.2005.159.
[5-32] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.
[5-33] "Toward optimal feature selection | Proceedings of the Thirteenth International Conference on International Conference on Machine Learning." https://dl.acm.org/doi/10.5555/3091696.3091731 (accessed May 18, 2021).
[5-34] M. A. Hall, *Correlation-based feature selection for discrete and numeric class machine learning*, vol. 8. Hamilton, N.Z: Dept. of Computer Science, University of Waikato, 2000.
[5-35] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. Mcvean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, "Detecting Novel Associations in Large Data Sets," *Science (American Association for the Advancement of Science)*, vol. 334, no. 6062, pp. 1518–1524, 2011, doi: 10.1126/science.1205438.
[5-36] A. Mackridge and P. Rowe, *A Practical Approach to Using Statistics in Health Research*. Newark: John Wiley & Sons, Incorporated, 2018.
[5-37] B. Efron, "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," *Biometrika*, vol. 68, no. 3, p. 589, 1981, doi: 10.2307/2335441.
[5-38] R. Kohavi., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 2020. Accessed: May 18, 2021. [Online]. Available: https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection
[5-39] Meteorological Agency, "Searchforpastweatherdata(in Japanese)."
[5-40] Organization for Cross-regional Coordination of Transmission Operators, "System information service (in Japanese)," 2020.
[5-41] Japan Electric Power Exchange, "Trading information (in Japanese)," 2020.

Chapter 6
Conclusions
6.1 Conclusions of research

In this thesis, analytical approaches to identify the important factors that affect the deviations between the actual demand and its procumbent plans were proposed. Furthermore, the effectiveness of the proposed approaches was demonstrated by applying them to real-world datasets collected.

In chapter 2, a survey of previous studies regarding the electricity demand was conducted. Moreover, the technical issues were determined based on the demand modeling approach to analyze the factors that affect the electricity demand. In addition, the key techniques for the demand modeling were summarized to solve the technical issues.

In chapter 3, a situation-dependent modeling approach based on the partially linear additive model (PLAM) and sparse modeling was proposed to identify annually important variables that affect the electricity demand. The proposed approach applies the sparse modeling concept to PLAM to select the factors that affect demand changes while identifying linearity/nonlinearity among variables from numerous variables, which are focused on as candidate factors affecting the demand. Furthermore, the enumerate technique was applied to sparse PLAM to identify the limited number of variables that are commonly and consistently used in situation-dependent modeling. The enumerate technique is a promising method used to enumerate a subset group of variables with the similar level of importance to describe the demand instead of selecting a unique optimal important variable, which is done in conventional methods. In a case study focusing on the demand before and after the Great East Japan Earthquake in Japan, as well as numerous co-occurring observed variables, the proposed approach identifies the limited number of factors (60 out of 274 variable items are focused on as candidate factors) necessary to explain situation-dependent hourly electricity demand. Moreover, this approach reduces the average error in demand modeling by about 33 % compared to the conventional methods.

In chapter 4, focusing on the pandemic of COVID-19, the proposed sparse partially linear additive modeling framework is evaluated to verify that it adequately identifies the key factors affecting demand changes caused by the pandemic. An approach based on autoregressive integrated moving average with exogenous variable (ARIMAX) model and PLAM was proposed to analyze the factors affecting the deviations between the demand deviations, focusing on the electricity demand trend in Germany during the COVID-19 pandemic. Modeling long- and medium-term scenarios for each factor based on ARIMAX enables identifying factors that deviate from the scenario trends during the pandemic. Furthermore, the proposed approach, which applied the sparse PLAM scheme proposed in chapter 3, clarifies the limited number of factors necessary to explain the seasonal changes in electricity demand. Moreover, it identifies the additive contributions to the demand deviations based on each factor that deviated differently from the scenario trend during the pandemic. In the case study focusing on the electricity demand in Germany and numerous co-occurring observed variables, the proposed approach identified the monthly deviations caused by the COVID-19 pandemic. For instance, during minimum monthly demand, we confirmed that approximately 95% of the deviations were caused by factors that evolved differently from the scenario trend. Furthermore, the

characteristics of the key factors affecting the demand deviations were analyzed, focusing on the differences in the additive contributions and seasonality for each factor.

In chapter 5, a sensitivity analysis approach based on PLAM was proposed to analyze the impact of changes in factors on the occurrence of extremely large imbalances. The differences between the scheduled power procurement in electricity markets and actual power supply were determined by focusing on the Japanese electricity market. In the proposed approach, the scheme based on the relevant-redundancy measure for detecting the variables nonmonotonically related to extreme imbalance events and forward stepwise approach to select the appropriate nonlinearity for model description were introduced to flexibly describe the impact of each variable on the imbalance that occurs when constructing the PLAM. In the case study focusing on the Japanese electricity market and numerous observed variables, the results indicated that the proposed approach improves the modeling accuracy regarding the occurrences of extremely large imbalances by approximately three times compared to the conventional method. The proposed approach is useful for analyzing factors that are important for explaining extremely large imbalances. In addition, the differences among the impacts related to the changes of each important factor on the imbalance occurrences are discussed based on the results.

6.2 Future research

Firstly, in this thesis, the important variables that affect the deviations between electricity demand and its procurement plan are analyzed. These variables may uncertainly change depending on the conditions of the economy and living environment, and cause unexpectedly large demand deviations. The approach for forecasting the electricity demand while properly considering the uncertainty in these variables is necessary to accurately forecast the demand trend after the pandemic.

Secondly, in this thesis, the data collected in Japan and Germany are utilized as application targets of the proposed modeling scheme to evaluate it. The important factors affecting the deviations between electricity demand and its procurement plan may change depending on the government policy and other significant socially events, such as economic shock, disaster and pandemic. Analyzing the data from multiple points of view, including other geographic areas and periods, will be continued. Accordingly, each result should be compared.

Finally, in this thesis, the factors that had a direct impact on the target demand are analyzed. However, another interesting approach is constructing a hierarchical structure for factors that may indirectly affect demand. Analysis based on such statistical graph structures will provide another informative aspect of the dominant factors affecting the electricity demand. This will be investigated in future research.

6.3 Business opportunity

The first opportunity is the utilization for monitoring and evaluation techniques by power system operators to assess the power capacity [W] and the amount of electricity [Wh] required at a certain time. In recent years, the enhancement of the accuracy of hourly power demand estimations, which serve as the foundation for such monitoring and evaluation, has been actively discussed as a pressing issue that must be addressed expeditiously by operators. The proposed approach for analyzing the factors affecting the hourly electricity demand will be essential in developing a framework for reliable and highly accurate demand modeling.

The second opportunity is to use to evaluate energy conservation trends in the future. These trends, specifically those associated with the reduction of carbon dioxide emissions, are garnering global attention. The proposed approaches contribute to analyzing dynamic characteristic features of structure in the electricity consumption; the results are expected to be deployed to analyze and forecast future energy and electricity conservation trends, as well as the impact of these trends on the carbon dioxide emissions. In addition, by clarifying the key factors that affect the energy and electricity conservation trends, the proposed method can be utilized to realize policies to reduce carbon dioxide emissions.

The third opportunity is to use the proposed approaches to contribute to effective industrial use of big data. With the recent increase in the number of observed data items, the effective use of big data is attracting attention for their potential applications in various rea-world problems. The method proposed in this study focuses on the relationship between the target variable and numerous observed data, and is a highly interpretable method that can identify only statistically significant relationships. Moreover, it is expected to be applied to other multiple problems. For example, a procedure using the sparse PLAM and numerous observed variables, including the uncontrollable environmental factors and controllable machine parameters, is proposed to create optimal operational plans in plant factories. Industrial applications of such big data and machine learning approaches are expected to be used in various fields for efficient production planning and profit enhancement.

Acknowledgements

I would like to offer my deepest gratitude to my supervisor Prof. Yasuhiro Hayashi for providing me the opportunities to grow as a researcher through getting involved in various research projects. I also would like to extend my deepest gratitude to my mentor Associate Prof. Yu Fujimoto for his consistent support, encouragement, and instruction of an attitude towards difficulties.

I express my sincere gratitude to Prof. Toru Asahi, Prof. Noboru Murata, Prof. Hideo Ishii at Waseda University, and Associate Prof. Ryoichi Hara at Hokkaido University for being referees in my thesis defense.

I am grateful to Dr. Satoshi Kabe, Mr. Motonari Hayashida and Dr. Tomohiro Inoue at Central Research Institute of Electric Power Industry, Mr. Tomoyuki Hirai and Mr. Shota Tajima at Daigas Group, Dr. Wataru Hirohashi at Advanced Collaborative Research Organization for Smart Society, and Prof. Hans-Arno Jacobsen at University of Toronto for their discussions with my research.

I express special thanks to PhD students of Hayashi laboratory, Dr. Hiroshi Kikusato, Dr. Satoru Akagi, Dr. Yuji Takenobu, Dr. Akihisa Kaneko, Dr. Kohei Murakami, Dr. Ryu Ando, Mr. Tatsuki Okuno, Mr. Ryosuke Shikuma, Mr. Yujiro Tanno, Mr. Fumiaki Osaki and Mr. Koto Watanabe for supporting my laboratory life.

I thank all students and staff of Hayashi laboratory and my colleagues at the Leading Graduate Program in Science and Engineering, Power Energy Professionals Program and Open Innovation Ecosystem W-SPRING Program in Waseda University for their assistance.

I gratefully thankful to my family, my mother Rie Kaneko, father Natsuki Kaneko, sister Yurie Kaneko, grandmother Sakiko Kaneko and grandfather Michiho Kaneko.

February 2023
Nanae Kaneko

List of Research Achievements

* indicates the research directly related with this thesis.

Journal Papers

1. *Nanae Kaneko, Yu Fujimoto, Yasuhiro Hayashi, “Sensitivity analysis of factors relevant to extreme imbalance between procurement plans and actual demand: Case study of the Japanese electricity market”, Applied energy, vol. 313, p. 118616–, May 2022.
2. *Nanae Kaneko, Yu Fujimoto, Satoshi Kabe, Motonari Hayashida, Yasuhiro Hayashi, “Sparse modeling approach for identifying the dominant factors affecting situation-dependent hourly electricity demand”, Applied Energy, vol. 265, p. 114752-, May 2020.
3. Yu Fujimoto, Saya Murakami, Nanae Kaneko, Hideki Fuchikami, Toshirou Hattori, Yasuhiro Hayashi, “Machine Learning Approach for Graphical Model-Based Analysis of Energy-Aware Growth Control in Plant Factories”, IEEE access, vol. 7, p. 32183–, March 2019.

Proceedings

1. Nanae Kaneko, Yu Fujimoto ,Yasuhiro Hayashi, “Graphical Modeling for Analysis of Hourly Electricity Demand and Market Price”, 17th International Conference on the European Energy Market, September 2020.
2. Nanae Kaneko, Yu Fujimoto , “Analysis of Energy Procurement Balance Focusing on Imbalance Settlement Scheme in Japanese Electricity Market”, 25th International Conference on Electrical Engineering, July 2019.
3. Nanae Kaneko, Yu Fujimoto ,Yasuhiro Hayashi, “Toward Data-Driven Identification of Essential Factors Causing Seasonal Change in Daily Electricity Demand Curves”, 2019 IEEE Milan PowerTech, June 2019.
4. Nanae Kaneko, Yu Fujimoto ,Yasuhiro Hayashi, “Nonlinear variable transformation for construction of long-term demand forecast models”, 24th International Conference on Electrical Engineering, June 2018.

Conference Papers (in Japanese)

1. 金子奈々恵，藤本悠，広橋亘，　平井友之，田嶋翔太，林泰弘，「実需要家における運用を想定した家庭用燃料電池と蓄電池による一次調整力調達のためのアグリゲーション手法の評価」，令和 4 年 電気学会電力・エネルギー部門大会，福井，2022 年 9 月．
2. 金子奈々恵，井上 智弘，「スパースモデルを用いた JEPX スポット市場価格の予測および要因分析」，第 38 回エネルギーシステム・経済・環境コンファレンス，オンライン，2022 年 1 月．
3. 金子奈々恵，藤本悠，広橋亘，　平井友之，田嶋翔太，林泰弘，「一次調整力提供のための家庭用燃料電池と蓄電池のアグリゲーション手法」，令和 3 年電気学会 全国大会，オンライン，2021 年 3 月．

4. 金子奈々恵，藤本悠，林泰弘，「電力負荷推移の要因分析を目的とした統計的グラフ構造の解析―日本とドイツの比較―」，令和 2 年 電気学会電力・エネルギー部門大会，オンライン，2020 年 9 月.
5. 金子奈々恵，藤本悠，林泰弘，「時別電力負荷推移の要因分析を目的とした統計的グラフ構造解析の検討」，令和元年電気学会電力・エネルギー部門大会，広島，2019 年 9 月.
6. 金子奈々恵，藤本悠，「インバランス料金の算定機構に着目した変数間の関係性の把握に基づく需給状況の解析」，平成 31 年電気学会 全国大会，北海道，2019 年 3 月
7. 金子奈々恵，藤本悠，林泰弘，「中期先電力需要予測における変数属性の要因分析を目的としたグループ正則化の検討」，平成 30 年 電気学会電力・エネルギー部門大会，徳島，2018 年 9 月.
8. 金子奈々恵，藤本悠，「インバランス料金制度における供給力不足時間帯の予測に関する一検討」，平成 30 年 電気学会 全国大会，福岡，2018 年 3 月.
9. 金子奈々恵，藤本悠，「説明変数の時別選択に基づくインバランス料金単価予測手法の検討」，平成 29 年 電気学会電力・エネルギー部門大会，東京，2017 年 9 月
10. 須藤慧，金子奈々恵，藤本悠，林泰弘，「インバランス料金単価予測手法の検討」，平成 29 年 電気学会 全国大会，富山，2017 年 3 月

Awards

1. 平成 29 年電気学会全国大会優秀論文発表賞, 2017 年 3 月 15 日.
2. 平成 29 年電気学会 YPC 優秀文発表賞, 2017 年 9 月 5 日.
3. 平成 29 年電気学会 優秀論文発表 A 賞, 2017 年 9 月 5 日.
4. IEEE PES Japan Joint Chapter Student Best Paper Award, 2019 年 1 月 15 日.
5. 令和 2 年電気学会 電力・エネルギー部門大会 Younge engineer Oral Competition(YOC)奨励賞, 2020 年 9 月 11 日.
6. 令和 3 年電気学会全国大会優秀論文発表賞, 2021 年 3 月 10 日.
7. 令和 4 年電気学会 電力・エネルギー部門大会 Younge engineer Oral Competition(YOC)奨励賞, 2022 年 9 月 9 日.