Thesis for the Doctor of Philosophy

Two Models for Electricity Demand using Keyword Search Volume and Panel Artificial Neural Network

Sungjun Park

Graduate School of Hanyang University

August 2019

Thesis for the Doctor of Philosophy

Two Models for Electricity Demand using Keyword Search Volume and Panel Artificial Neural Network

Thesis Supervisor: Jinsoo Kim

A Thesis submitted to the graduate school of
Hanyang University in partial fulfillment of the requirements for the
degree of Doctor of Philosophy

Sungjun Park

August 2019

Department of Earth Resources and Environmental Engineering

Graduate School of Hanyang University

This thesis, written by Sungjun Park,
has been approved as a thesis for the Doctor of Philosophy.

August 2019

Committee Chairman: YOON KYUNG KIM (Signature)

Committee member: WON MO SUNG (Signature)

Committee member: Sunjoon Kim (Signature)

Committee member: Joongmoo Byun (Signature)

Committee member: Jinsoo Kim (Signature)

Graduate School of Hanyang University

List of Contents

List of Table

List of Figure

Abstract

Two Models for Electricity Demand using Keyword Search Volume and Panel Artificial Neural Network

Sungjun Park
Dept. of Earth Resources and Environmental Engineering
The Graduate School
Hanyang University

Big data analysis and machine learning are rising analytical tools in data analysis. Big data is an area that collects and maintains a huge amount of raw data for field-specific data analysis. Machine learning is the main analytical tool for handling such data. This study investigates the applicability of keyword search volume, and develops an ANN (Artificial Neural Network) model using panel data to analyze electricity demand and forecast prices. There is no analysis using keyword search volume in econometrics, especially energy economics. Therefore, this study intends to build a new electricity demand model. In addition, since there is no model building study that applies panel data, this study constructs a novel panel ANN model. This study consists of two essays: panel analysis model development and panel ANN model development.

In the first essay, this analysis derives the relationship between US household electricity consumption and renewable energy. For this purpose, keyword search volume is used to present new influential factors in analyzing economic indicators. The model considers three keywords related to electricity consumption: “renewable,” “weather forecast,” and “temperature.” Furthermore, there has been no way to quantify household renewable energy consumption, no studies have analyzed the correlation between renewable energy and US household electricity consumption. Such consumption is difficult to estimate and it is more difficult to grasp than other major sectors including commerce and industry because of issues related to personal information collection and the cost of measurement. This study therefore analyzes the correlation with household electricity consumption by constructing a model including interest in renewable energy using keyword search volume.

The model, which analyzes the impact of these keywords is constructed using three regression equations based on the static energy demand model, and analyze the impact of these keywords. In the household sector, although a variety of renewable energy is used, it is difficult to derive the economic implications of such use as it is not converted into a quantifiable value. Therefore, this study uses the search keyword “renewable” to estimate the impact of renewable energy. “Weather forecast” and “temperature” were also selected as Internet search keywords. These keywords are used because temperature is one of the important factors in determining household electricity consumption.

As a result, all the variables are stationary and the Hausman test indicates that the fixed effects estimation is more robust than the random effects estimation. In the case of the model using the keyword “renewable” as an explanatory variable, all the variables except the price variable are statistically significant at the 1% level; this search term has a negative correlation with household electricity consumption. Household electricity consumption decreases by 16.017 million kWh for every one unit increase in the keywords search using “renewable.” “Temperature” also has a negative coefficient, which is similar to heating degree days.

The correlation between the two variables, which intuitively appear to be unrelated, could have significant meaning. When one searches for “renewable” in the context of their

household, they probably have a clear purpose. In the event that excessive electricity is consumed or electricity bills are high, households will search for alternatives to reduce electricity consumption. In the case of households equipped with renewable energy facilities, the power consumption will decrease in proportion to the capacity, and the results of the estimation can be seen.

This study finds that the correlation coefficient of the “renewable” variable is the highest, and the “temperature” variable also has a significant correlation with household electricity consumption. The “renewable” keyword has a large negative correlation with household electricity consumption, which can be estimated as being a result of the growing interest in renewable energy. Although the electricity consumption patterns of households are influenced by many variables, this study suggests that interest in renewable energy should also be included as a major factor influencing such consumption.

In the second essay, this study predicts electricity price using ANN, which have already been used as tools for prediction in various fields. In general, ANN have been used for short-term forecasting in many economic analysis studies. On the other hand, as the forecast point increases, the accuracy of prediction decreases sharply. The forecasting accuracy in long-term forecasting is greater than that of short-term forecasting in the same dataset. Therefore, this study uses panel data to compensate for the decline in ANN forecasting accuracy in long-term forecasts in the same dataset.

The panel data contains information that time series data does not have. It has trend information of time series data as well as state or country characteristics. However, there are very few studies in economics that have used panel data for prediction using ANN. Existing studies use panel data without differentiating between entities in the model structure. The panel ANN studies did not differentiate between state and national data or

have independent learning such as the pooled OLS method.

Therefore, this study constructs a panel ANN structure using the advantages of panel data and analyzes its accuracy according to the change of forecasting periods. The model intends to improve the accuracy of predicted values by learning the unobserved heterogeneity contained in panel data from each state. The analysis is conducted on the assumption that it would be possible to learn not only time series information but also country or state information.

The panel analysis removes the cross-sectional dependence in the unobserved heterogeneity of the panel data. Unlike panel analysis, this study constructs a model structure to learn the unobserved heterogeneity of such data. The learning is conducted separately for each state, and two or three hidden layers are inserted. After 6, 12, 18 and 24 months forecasting, total RMSE and MAPE are estimated and the optimal model is selected.

For empirical analysis, this study uses panel data of US electricity prices by state. Natural gas prices are also predicted for additional model verification. For the electricity price forecasting model, the accuracy of the result using time series data in 6 and 12 months forecasts is higher than using panel data. On the other hand, the results of 18 and 24 months indicate that the results of panel data are much better. In the case of natural gas Citygate prices, the results of the model using time series data for only 6-month predictions are better while other predictions show that the panel data model has high accuracy. A noteworthy point is that panel data models tend to be more accurate as the forecast period increases. Although the timing of improvement in accuracy differs, both models show an improvement of the panel data forecasting model in long-term predictions.

According to the results, when estimating a small number of predicted values, the trend of the time-series data greatly influences the result and a time-series model produces better

predictions. On the other hand, the longer the forecast period, the better the panel data model that learns from unobserved heterogeneity of the states rather than from the trends. Since weights are updated without affecting each layer, it can be said that the model learns by considering the heterogeneity of each state. In comparison to a time series model in which only the trend is learned, the panel data model utilizes more information to improve accuracy by learning the trends and heterogeneity of each state.

In this study, electricity consumption is analyzed using panel data and electricity price prediction is performed. The electricity consumption analysis suggests a new approach based on the model considered in household electricity consumption literature that incorporates data drawn from keyword search volume. This study used keyword search volume as a substitute variable to analyze the phenomenon that was impossible to explain due to the lack of quantitative data. This study shows that variables that have not been used hitherto, as they are not quantifiable or statistically significant, can be analyzed through keyword search volume.

In the electricity price forecasting analysis, a novel panel ANN model is proposed to compensate for the decrease in forecasting accuracy when the forecasting period increases Panel ANN is a model that can be applied from day-to-day and hourly forecasts to long-term trends of several years depending on the type of panel data. In analyzing the long-term trends, a neural network model that can replace the large-scale simulation models such as NEMS (National Energy Modeling System) and WEM (World Energy Model) can also be constructed. Therefore, this model can be applied in various fields ranging from the hourly price forecast of the next day's electricity market to the long-term trend of CO_2 emissions.

Chapter 1. Introduction

1.1 Research question

Big data analysis and machine learning are two major focuses of data science. Big data represents an area that collects and maintains enormous amounts of raw data for domain-specific data analysis. Many technology-based companies are currently driving data collection and maintenance to become a core business. They are investing in product development to solve monitoring, experimentation, data analysis, simulation and other knowledge and business requirements. Social media organizations are also constantly generating very large amounts of data. Machine learning is the main analytical tool used to process such big data, and various models are being studied and developed through artificial neural network (ANN) or deep learning.

Extracting meaningful patterns from large amounts of input data for decision making, forecasting, and other reasoning is a key challenge in big data analysis. In addition to analyzing and processing large amounts of data, it is also important to extract data that has not been used in the past and apply it to research. That is, when researching with big data, creative and innovative data analysis and management are needed. Machine learning is also very useful as a tool for big data analysis. In particular, problems such as rapid information search and differential modeling can be solved more effectively through ANN, which are a type of machine learning. However, there is no analysis using keyword search volume in econometrics, especially in the field of energy economics. Therefore, this study builds a novel energy demand model including keyword search volume. In addition, in the case of the ANN model applying the panel data, a novel panel ANN model is constructed through this study because there is no model construction case.

First, this study uses Google Trends to examine the applicability of panel analysis for keyword search volume. This study analyzes the relationship between household electricity consumption and keyword search volume using panel analysis. The motivation of electricity consumption panel analysis is twofold. The first is to derive the relationship between renewable energy and electricity consumption. The second is to confirm the availability of keyword search volume to "predict the present." Following the Paris Agreement in 2015, the United States has been forced to reduce CO_2 emissions, which has raised interest in renewable energy. However, there is no way to express this interest quantitatively in traditional economics. While US industrial electricity use can be restricted and managed at the national level, the degree of renewable energy in household electricity consumption is unknown, making it difficult to estimate. In addition, it is difficult to grasp than other major sectors including commerce and industry because of issues related to personal information collection and the cost of measurement [1]. Therefore, this study uses Internet queries to build a model that includes interest in renewable energy and analyze its correlation with household electricity consumption.

Google has one of the largest search engines in the Internet search market. Google offers search services around the world, Google Trends shows the search frequency of a keyword based on all Google searches conducted globally and in real time. As Google Trends can use information that is yet to be announced, its data can predict the present. "Predict the present" refers to analyzing the current situation, which has not yet been officially announced and cannot be otherwise analyzed. Google Trends analysis has the more favorable capability to "predict the present" rather than "predict the future" [2]. In this study, the analysis is conducted under the assumption that Google Trends data could be used as an explanatory variable in the econometric analysis. Therefore, if search frequency is more influential than existing explanatory variables, researchers should consider constructing a model based on the frequency of Internet-based searches.

Secondly, this study proposes a novel panel ANN model structured to fit the panel data. This study uses panel data to compensate for the decline in ANN forecasting accuracy in long-term forecasts. In general, ANN have been used for short-term forecasting in many economic analysis studies and show a high level of accuracy [3-8]. On the other hand, as the forecast point increases, the accuracy of prediction decreases sharply. Regardless of whether long-term prediction or short-term prediction, if forecast point increases, the prediction accuracy decreases. Therefore, under the same data set, the long-term forecasting accuracy becomes more inaccurate than the short-term prediction.

This study focuses on the forecast point because we want to analyze the change in forecasting accuracy with the increased number of predictions as the forecast period changes from short- to long-term in the same dataset. The aim of this study is to compensate for the decline in forecasting accuracy of long-term predictions that are based on an economic forecasting model using time-series data. With an increase in the number of predicted points, the total forecasting accuracy decreases. As with all forecasting models, the predictions of the farther future will decline in accuracy, which will reduce the total forecasting accuracy.[1]

The panel data includes information that does not exist in the time series data. It includes not only trends in time series data, but also state or national characteristics. ANN shows high accuracy with one or two point forecasts only with time series data, but as mentioned above, more information is needed to improve the accuracy when the forecast point is increased. Therefore, this study tries to improve forecasting accuracy by using panel data.

[1] In the case of large-scale simulation models such as NEMS (National Energy Modeling System) and WEM (World Energy Model), the difference between forecasts by scenario increases.

However, this model does not consider that prediction values are used repeatedly for the next forecast. Using prediction values that cannot determine the accuracy in neural network learning will cause the model to lose credibility.

For model verification, electricity demand analysis and electricity price forecasting are performed. The electricity industry is the foundation of the national economy, and electricity supply, demand, and price are major considerations in all industrial sectors and also have a significant impact on living standard of residents [9-12]. In recent years, one of the most important commodities in the people's life are electricity. In terms of the national policy, the electricity market is one of the most important determinants and has sensitive [13]. Especially, the US electricity market is one of the largest in the world and its installed capacity and electricity consumption are also the world's largest. Therefore, this study attempts to analyze the US electricity market in terms of electricity consumption and prices. A more detailed description of the methodology is provided in Chapter 2, and consumption analysis and price forecasting are provided in chapters 3 and 4, respectively. Finally, the overall summary and conclusions are provided in the last chapter.

1.2 Literature review

1.2.1. Panel analysis of electricity demand

Many researchers have studied of the temperature, price, and income elasticity of household electricity consumption in various ways. Some studies emphasize the importance of temperature variables in electricity consumption analysis. Hekkenberg *et al.* [14] analyze electricity demand in the Netherlands from 1970 to 2007. They find that, since 1970, electricity demand has been increasingly dependent on temperature differences; thus, it is appropriate to set the temperature factor as a variable in demand analysis. Bessec and Fouquau [15] analyze the correlation between European Union (EU) electricity demand and temperature. Based on panel data from 15 European countries over 20 years, they use a panel threshold regression model to calculate the results. Emphasizing that temperature is an important factor in determining European electricity consumption, they show that temperature sensitivity to electricity consumption is increasing.

Many panel studies have focused on the price and income elasticity of household electricity consumption. Paul *et al.* [16] use monthly average prices and electricity demand data by state from 1990 to 2006 and apply the partial adjustment model including monthly heating degree days (HDD) and cooling degree days (CDD). The coefficient value of annual average prices is -0.13 in the short term and -0.36 in the long term, confirming that electricity demand is price elastic. Paul *et al.* [16] argue that prices are exogenous in the demand equation. Alberini *et al.* [17] analyze residential electricity and gas consumption in the United States from 1997 to 2007. They find that the change in demand due to electricity prices is small in the short term (with short-run price elasticities ranging between -0.08 and -0.15) and suggest a price increase to reduce energy consumption from a long-term perspective. Alberini and Filippini [18] analyze household electricity demand for 48 states in the United States from 1995 to 2007. They conduct panel analysis to overcome external validity limitations and, unlike Alberini *et al.* [17], use household-level data. The price elasticity of electricity demand is estimated to range from -0.860 to -0.667. Further, the price elasticity of electricity demand decreases with income but its effect is minimal. These results are in stark contrast with other research and government figures. Sun [19] analyzes household electricity consumption in 48 US states from 1995 to 2010. He raises the issue of the endogeneity of electricity prices and uses the bias-corrected least squares dummy variable (LSDV) and generalized method of moments (GMM) methods to address this problem. Salari and Javid [20] analyze household electricity consumption from 2005 to 2013 in 48 states of the United States. Static and dynamic panel estimation models are used to analyze electricity consumption. Price elasticity is estimated to be -0.076 and income elasticity is estimated to be 0.052 in the short term. The results show that both HDD and CDD have a greater impact on household electricity consumption than prices and building age in both panel estimation models. They also show that price and income elasticities have a large correlation only in the long term.

Studies of household electricity consumption have also been conducted outside the United States. Filippini [21] empirically analyzes household electricity consumption for 22 cities in Switzerland from 2000 to 2006. In the long run, he finds that electricity prices are able to adjust consumption patterns by comparing peak and off-peak electricity demand elasticities. However, he also shows that electricity prices have no significant correlation with electricity consumption from a short-term perspective. Azevedo *et al.* [22] estimate the price and income elasticities of the United States and EU and, find that household electricity prices are inelastic. When United States and EU panel data are analyzed together, the price elasticity is estimated to be -0.18, -0.21 compared with -0.20, -0.21 for EU countries and -0.21, -0.25 for the United States. Wiesmann *et al.* [23] focus on the impact of residential characteristics on electricity consumption in Portugal. They find that the direct effect of income on electricity consumption is low, and appears to be even lower if relevant control variables are included.

These studies yield significantly different coefficients because of differences in the underlying methodology, variables, and period. In addition, all examine household electricity consumption trends from a long-term perspective, and therefore mostly use annual data. Hence, the purpose of this study is to explore the short-term relationship, leading us to exclude household income from the analysis and use monthly data.

1.2.2. Keyword search volume

Despite the wide variety of online information available, it has not been used in traditional econometrics. In particular, although this information is provided continuously and in real time, it has not been used in economics, even for short-term analysis. To address these issues, Choi and Varian [2] apply Google Trends to traditional econometrics and suggest the applicability of Google Trends data for analyzing automobile sales, unemployment claims, travel destination planning, and consumer confidence. Using simple seasonal AR(1) models with Google Trends variables improves forecast accuracy by 5-20% thereby, encouraging its use in various fields. As a result, research has used Google Trends to analyze household behavior such as commodity consumption activity in the labor and housing markets [24].

Keyword search volume is also used to analyze social phenomena. Typically, many studies are analyzing unemployment rates. Askitas and Zimmermann [25] reveal a strong correlation between unemployment rates and search keywords from 2004 to 2009, using monthly German data. As the explanatory variables, "unemployment office or agency", "unemployment rate", "personnel consultant", "most popular job search engines in Germany" were used. They argue that Google Trends data are an appropriate input for policymaking. In traditional econometric economics, it is impossible to consider the explanatory variables that can immediately reflect policy changes. Therefore, Internet activity data are useful for forecasting complex and rapidly changing trends. D'Amuri and Marcucci [26] use Google Trends to predict the US unemployment rate. In particular, the results of their Google Trends analysis yield better predictions than state-level or expert survey forecasts. Fondeur and Karamé [27] predict the unemployment rate in France by using data from Google queries on Internet keywords. It estimates the unemployed population aged 15-24 and shows an improvement in RMSE of 17.5%.

The use of keyword search volume in the medical field is also increasing. Cooper *et al.* [28] find that search activity for certain cancers matches the expected incidence for 2001-2003. Eysenbach [29] finds a high correlation between epidemiological data and the number of clicks on the search results for flu-related keywords in the Canadian flu season during 2004 and 2005. Polgreen *et al.* [30] show that the search volume for influenza-related searches is correlated with the number reported continuously over 2004-2008. Ginsberg *et al.* [31] and Doornik [32] use Google search data on influenza virus surveillance. Based on this methodology, Google Flu Trends has predicted the incidence of flu in real time in many countries. Hulth *et al.* [33] estimate similar results in a study of the search keywords submitted to the Swedish medical website.

US household electricity consumption is usually announced with a one-quarter delay. Moreover, some statistics may be published a year later. This fact is rarely highlighted in traditional econometrics. From this perspective, research has been conducted on the possibility of using Google Trends. McCarthy [34] and Gunn III and Lester [35] argue that the presentation of suicide-related indicators in public health statistics is too late to affect social factors. McCarthy [34] analyze the correlation between the suicide rate and Google Trends keyword searches (“depression”, “suicide” and “teen suicide”) and suggesting that self-injury and suicide can be predicted. Gunn III and Lester [35] also analyze the association between suicide rates and Google Trends suicide searches (“commit suicide”, ‘‘how to suicide’’ and ‘‘suicide prevention’’) and concludes that the search volume of three keywords in the US 50 states is positively correlated. The results show that we can monitor the trend of suicide rates more quickly than the central government’s presentation of suicide statistics. Sueki [36] analyzes the changes in Google search volume for suicide and depression in Japan. He suggests that searches for “depression” could alert public health officials to an impending rise in suicide rates. Therefore, these studies show the advantage of using keyword search volume for real-time information gathering and to overcome analysis errors resulting from data release delays

In chapter 3, this study uses keyword search volume to analyze US household electricity consumption. No study has thus far analyzed the relationship between electricity consumption and keyword search volume except Park and Kim [37]. Although Salahuddin *et al.* [38] showed that there is a significant positive relationship between Internet usage, electricity consumption, and CO_2 emissions in the long run, there is no analysis using Internet search terms. Given that keyword selection is an essential part of Google Trends data analysis, the theoretical assumptions are explained in Section 3.3.

1.2.3. Artificial Neural network of electricity price forecast

Various attempts at predictive analysis through ANN have been made in the field of electricity prices. Wang and Ramsay [39] conducted electricity prices forecasts for public holidays and weekends using front-end processors (FEP) and ANN. They mentioned that the trend of SMP (System Marginal Cost), electric power demand, generation capacity and tender participants are factors affecting current SMP. However, considering the periodic characteristics of SMP due to changes in demand, generation capacity and tender participants were excluded from the variables. To estimate the SMP of each holiday, the settlement period index, estimated electricity demand, ID flag and historical SMP were used as variables. According to the predictions, each mean absolute percentage error (MAPE) was calculated as 9.40% on Saturday, 8.93% on Sunday and 12.19% on public holidays. Saturdays had a higher number of errors compared to Sunday and, due to lack of information, this was concluded to be a result of the influence of weekdays and holidays.

Yao *et al.* [40] implemented SMP prediction based on wavelet transform and ANN. Through the wavelet, the SMP is decomposed into several details related to the low-frequency approximation portion and the high-frequency portion. After decomposing, the author estimated the approximate SMP using ANN learned from low frequency and electricity load data. Finally, short-term SMP prediction was performed by summing up the estimated approximate part and the weighted detail part. For empirical analysis, this study predicts the SMP of the UK electricity market from January to March 1997. The proposed wavelet transform and ANN method show good prediction results. On the other hand, SMPs generally rely heavily on generator bidding patterns, which depend on available power generation and transmission, system power reserve (SPR), and system potential demand (SPD). Therefore, to predict SMP, SPR and SPD should be included. However, this study suggests that only SPD data is used due to lack of SPR data, and in future, SPR should be included.

Zhang Li *et al.* [41] propose using a forecasting method of market clearing prices (MCP) with high volatility through an ANN for effective bidding strategies by power companies or independent power producers. MCP is difficult to predict because of the many uncertainties which interact with the bidding strategy in complex ways. This means a predictive value of key explanatory variables is required. Therefore, they compares and analyzes the effects of uncertainty measures and predictions on the predictive distribution of ANN through continuous and discontinuous networks. For the empirical analysis, the author performed MCP prediction of New England, and compared these to the existing predictive value. The data used for the prediction of MCP are the predicted load, actual load, expected MCP and actual MCP, and a total of more than 50 factors were used for MCP prediction. They implemented a prediction and confidence interval estimation method using a continuous-type ANN to improve the prediction of MCP in the New England. In addition, the differentiated network with MLP and radial based neural network prediction method was applied, offering an efficient ANN to calculate the exact prediction and confidence region.

Jau-Jia and Luh [42] used the "committee machine" consisting of multiple networks, and made forecasts for the New England electricity market price. In the case of a single neural network, when an improper ANN was applied, there is a chance of errors in the relationships that can be inferred from the input and output data. Therefore, the committee machine applied in order to mitigate these errors. They show that the committee machine is more advantageous than a single network through two empirical analyses. In the first empirical analysis, standard deviation of radial basis function (RBF) and MLP predictions are used as weights of the committee machine and this shows improved forecasting accuracy. The second empirical analysis uses the committee machine to predict the average peak-hour MCP in the New England electricity market. Committee machine prediction results show that the MAPE improves by 1.66% and 2.53%, respectively, compared with RBF and MLP. When multiple neural networks are applied, the new weighting method can be used to improve prediction performance.

Rodriguez and Anders [43] conducted a study to forecast the MCP combined with ANN and fuzzy logic for the electricity market in the Ontario, United States. Three scenarios are set up for analysis. One where only demand is considered, another scenario in which generation capacity decreases sharply, and another scenario in which both demand and generation capacity decrease sharply. Four situations were set for each scenario in order to measure each MAPE. When fuzzy logic was also applied, scenarios were added. In the case of applying fuzzy logic, MAPE decreased significantly compared to the ANN which did not consider it. It is analyzed that the conventional Independent Market Operator (IMO) prediction has an average error of 55.21%, whereas the ANN using fuzzy logic has a value of -1.47%, which does not reflect sudden loss of power supply or power failure. It also shows that the predictive power of the existing model is reduced even at the point where demand increases sharply.

Gonzalez *et al.* [44] performed electricity price forecasting for the Spanish electricity market through an input / output hidden Markov model (IOHMM), which is an unsteady time series model using ANN. In order to optimize them, the Expectation-Maximization (EM) algorithm is used. In this study, two neural networks were studied using conditional distribution function to learn about market conditions and market price respectively. For the empirical analysis, the hourly electricity price from January to September 2001 in the Spanish power spot market was used. For the input variables, hourly power generation and demand data, and hourly power data with constant time lag were used. In addition, only physical variables related to power demand and available resources were considered to clearly distinguish and separate market conditions. On the other hand, the lagged variables of the electricity price, fuel price, the sum of the supply functions, and the stake of the power company were considered as additional variables, but these did not affect the forecasting power. As a result, MAPE was 15.83%, which shows good prediction results in terms of accuracy as well as dynamic information on the market.

Lee *et al.* [45] proposed a prediction technique of SMP using back propagation ANN. This study consists of two approaches for input data, time axis and day axis, and applied ANN using patterns derived from both methods. The data pattern of the time axis approach consists of the vector of the weekly SMP for a specific time recorded in chronological order. The data pattern of the day axis approach consists of the vector of hourly SMP of the specific week. The time axis approach reflects the current market trends, and the day axis approach reflects hourly, daily, and seasonal characteristics. The application and comparison of the two approaches reflected the rapid real time changes, daily and seasonal characteristics and the spot market, and the improved SMP prediction results were derived. The proposed method can be applied to real power market data for short-term price forecasting, and electric power market participants can use it as a tool for optimal strategy establishment.

Gareta *et al.* [46] conducted power price forecasting using the Gray Box Model, which is the midpoint between the Black Box Model in which the relationship between input and output nodes is unknown and the White Box Model in which the relationship is expressed through equations. Twenty-four output nodes are set up to predict the hourly electricity price, and the input variables were sorted first because the problem of overfitting may occur when inputting low correlation variables. All input and output variables are normalized to have a value between -1 and 1. 85 % of prediction results showed an error of less than 0.01€ and 75% of the results were less than 0.75€. In the case of long-term prediction using the predicted value as an input variable, 50 to 60% of the results had a prediction error of 0.05€ or less.

Georgilakis [47] conducted power price forecasting for the California power market using back propagation ANN and verified the predictive power. Based on the predicted power load, the author predicted 24 hour MCP with a “persistence method” for comparison with ANN. In the process of predicting the power load, neural networks were set up with 24 nodes inputting the load for the past 24 hours, 24 nodes inputting the 24 hour load a week ago, and 24 output nodes. Experimental results show that the MAPE of the predicted value is lowest when past MCP, past load, and predicted load are used as the input parameters of ANN. In conclusion, this study suggests that California power price forecasting should be based on data that do not tolerate price shocks, as predicted by ANN, compared to the previously used “persistence method”.

ANN used in the electricity price prediction can be classified as a feedforward network in which there is no direct connection between the output layer and the input layer, and a recurrent network in which inter-layer circulation exists. It can also be classified according to the number of nodes in the output layer. ANN with a single output node is used to forecast various type of prices such as electricity price [4, 44, 45, 48, 49], price at peak load [43, 50, 51], average base load price [3]. ANN, including multiple output nodes, are generally composed of 24 or 48 nodes, and predict the total price of the next day [52]. In many studies, electricity price forecasts are analyzed using various neural network structures. The neural network models and algorithms used for analysis and the input variables are presented in Table 1.1.

Table 1.1. Input variables and predict period of each artificial neural network models

Research	NN model	Learning algorithm	Input variables	Predict period
Wang and Ramsay [39]	MLP	BP	Historical electricity prices, Forecast load, Settlement period, ID flag	60 days
Rodriguez and Anders [43]	MLP, fuzzy MLP	BP, LM	Forecast load, Generation outages, capacity excess/shortfall, imports/exports	1 day & 30 days
Gareta et al. [46]	MLP	BP	Historical electricity prices, day type, month	2 days
Georgilakis [47]	MLP	BP	Historical electricity prices, Forecast load, Historical load	2 weeks
Szkuta et al. [52]	MLP	BP	Historical electricity prices, Forecast load, Forecast reserves, Settlement period, day type, month , holiday code, Xmas code, clock change	1 week
Wang and Ramsay [53]	MLP	BP	Historical electricity prices, Forecast load, capacity excess/shortfall, Settlement period, day type	1 week
Gao et al. [54]	MLP	BP	Historical electricity prices, Forecast load, imports/exports, Past MCQ(market-clearing quantity), fuel price, weather, Settlement period, day type, season	1 month

Table 1.1. Input variables and predict period of each artificial neural network models (continued)

Research	NN model	Learning algorithm	Input variables	Predict period
Mandal *et al.* [4]	MLP	BP	Historical electricity prices, Forecast load, Historical load, temperature, Settlement period, day type	1 week, 1month
Yamin *et al.* [3]	MLP	BP	Historical electricity prices, Forecast load, Historical load, Historical reserves, Forecast reserves, Settlement period, day type, line status, congestion index, line limits	1 week
Hu et al. [55]	MLP	BP	Historical electricity prices, Forecast load, MRR	1 week
Lora et al. [56]	MLP	BP	Historical electricity prices	3 month

1.3 Research framework

This section describes the research framework for panel analysis include keyword search volume and panel ANN. A more detailed description of the model structure is given in Chapters 3 and 4, respectively.

1.3.1. Panel analysis of electricity demand using keyword search volume

This study analyzes the correlation between interest in US renewable energy and US household electricity consumption by using keyword search volume. The model constructs based on the static energy demand model and uses HDD, CDD as temperature variables. In addition, electricity price variables and keyword search volume set as explanatory variables. This model considers three keywords related to electricity consumption: “renewable,” “weather forecast,” and “temperature.” The “weather forecast” and “temperature” keywords are selected to assess whether the weather variables represented by HDD and CDD could be replaced by Google Trends data. On the other hand, the “renewable” keyword is chosen to ascertain the impact of renewable energy on electricity consumption. The Figure 1.1 shows the flow of panel analysis using keyword search volume. A more detailed description is provided in Chapter 3

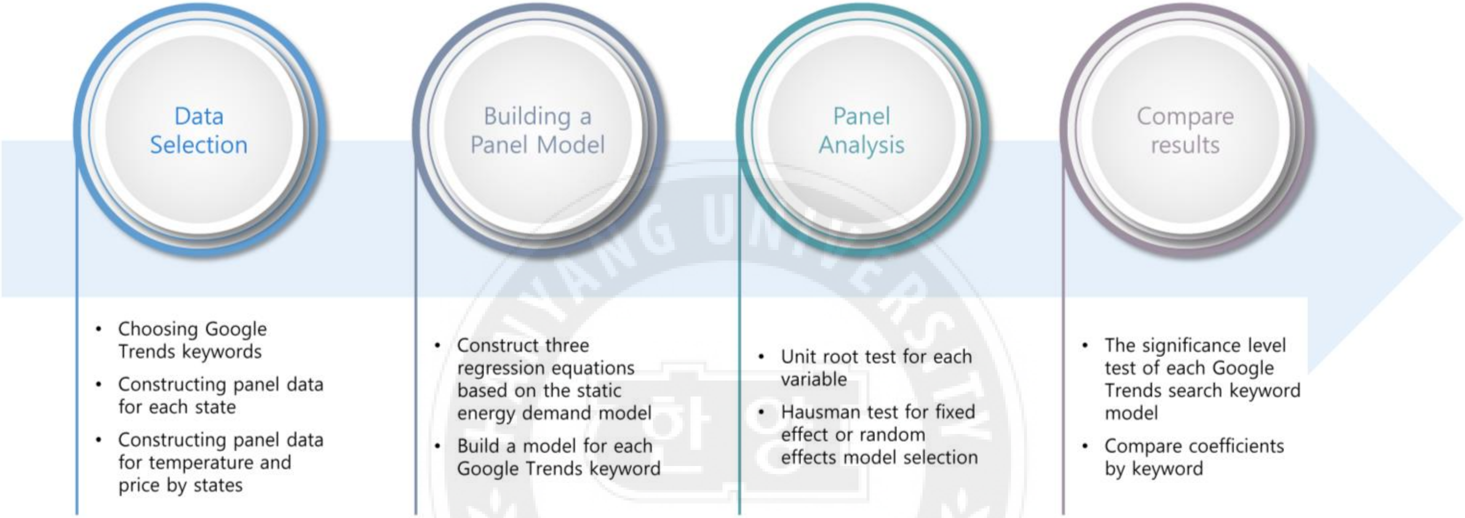
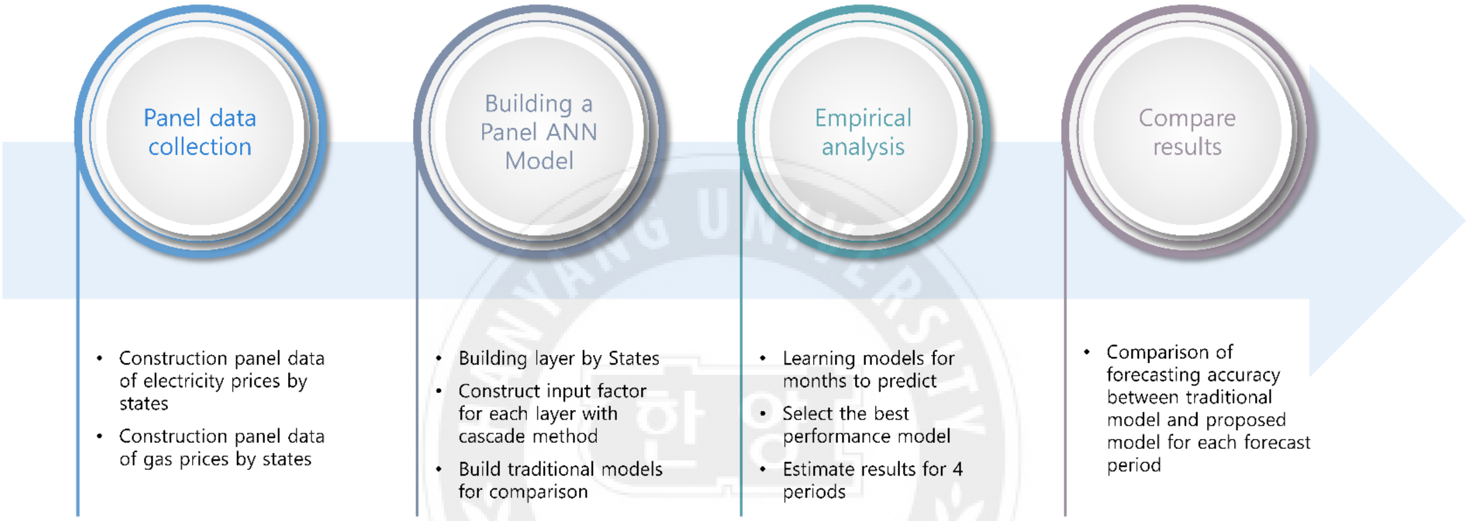

Data Selection
Building a Panel Model
Panel Analysis
Compare results
Choosing Google Trends keywords
Constructing panel data for each state
Constructing panel data for temperature and price by states
Construct three regression equations based on the static energy demand model
Build a model for each Google Trends keyword
Unit root test for each variable
Hausman test for fixed effect or random effects model selection
The significance level test of each Google Trends search keyword model
Compare coefficients by keyword

Figure 1.1. Core flow of panel analysis using keyword search volume

1.3.2. Panel artificial neural network model for electricity price forecasting

This study constructs panel ANN structure using the advantages of panel data and analyze the accuracy according to the change of forecasting periods. The model intend to improve the accuracy of the predicted values by learning the unobserved heterogeneity contained in each states of the panel data. This ANN model is based on the assumption that not only time series information but also each state information can be learned. Therefore, unlike the panel analysis, the model structure is constructed in order to learn the unobserved heterogeneity of the panel data. In order to verify the model, empirical analysis is conducted using the panel data of the US electricity prices and gas prices by states. The Figure 1.2 shows the flow of panel ANN model for electricity and gas price forecasting. A more detailed description is provided in Chapter 4.



Panel data collection

- Construction panel data of electricity prices by states
- Construction panel data of gas prices by states

Building a Panel ANN Model

- Building layer by States
- Construct input factor for each layer with cascade method
- Build traditional models for comparison

Empirical analysis

- Learning models for months to predict
- Select the best performance model
- Estimate results for 4 periods

Compare results

- Comparison of forecasting accuracy between traditional model and proposed model for each forecast period

Figure 1.2. Core flow of panel ANN

Chapter 2. Theoretical background

2.1 Panel analysis

Panel analysis is conducted to identify social phenomena in various fields such as economics, social sciences, etc. [37, 57]. Panel analysis is a type of longitudinal study that repeatedly observes the same subject to study changes over a period of time. In an econometric analysis, the panel data has multidimensional information. Panel data has an advantage that additional information can be obtained because it includes time-series data as well as cross-sectional data information.

Time series data is a time-sequential record of the phenomenon or characteristics of a particular object. Cross-sectional data, on the other hand, is a collection of phenomena or characteristics of several individuals at a particular time. That is, time series data has several observation points for a specific object, whereas cross sectional data has several objects observed at a specific point in time. The panel data is a combination of the time series data and the cross sectional data. Panel data refers to data collected over many years for fixed entities such as individuals, companies, and countries. Therefore, the panel data is a record of the phenomenon or characteristics of several objects by observation points. As shown in the Figure 2.1, the time series data can only be analyzed by comparing the yearly variation through annual comparisons, but panel data can identify the differences between the cohorts.

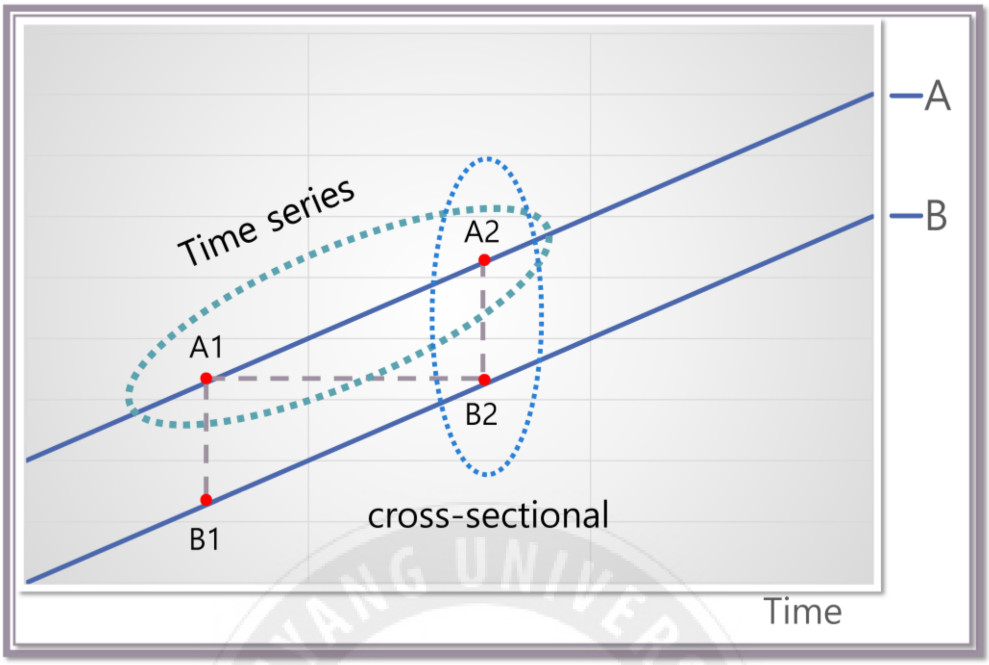


Figure 2.1. Difference between time series data set and panel data set

The panel data is distinct from the pooled cross-sectional data. It should be differentiated from a simple combination of cross-sectional and time-series data. The key to panel data is to fix observation groups. In pooled cross-sectional data, different entities are surveyed at each time point, rather than repeatedly observing the same entity. Panel data, on the other hand, basically repeatedly surveys the same entity over time.

Since the cross-sectional data is an examination of several entities at a specific time, only static relationships between variables can be estimated. On the other hand, the dynamic relationship can be estimated in the panel data because the individual is repeatedly observed. In addition, the unobserved heterogeneity of entities can be considered in the model. If the characteristics of these panel entities are excluded from regression analysis, the omitted variable bias may occur. In the regression model using panel data, model misspecification can be reduced because individual or national unobserved heterogeneity can be reflected.

As mentioned above, panel data analysis can control unobserved heterogeneity by controlling the differences between countries or states. Additionally, it is possible to mitigate the endogeneity problem due to an omitted variable bias. In other words, data from one individual over many years can be used to control unobserved heterogeneity of individuals who do not change over time. This is an advantage of the fixed effect model or the within estimator. On the other hand, attenuation bias due to measurement error can be reduced. This is an advantage of the between estimator. Additionally, using a between estimator can control the effects of economic fluctuations [58, 59].

2.1.1. Panel unit root test

Before the panel analysis, we have to determine whether the variables contain panel unit roots. The unit root test is performed to check whether each variable in the panel model is time series stationary. When the analysis is performed using nonstationary time series data, spurious regression occurs, meaning that there is a high correlation even though each variable is unrelated. The panel unit root test methods are Levin–Lin–Chu test [60], Harris–Tzavalis test [61], Hadri Lagrange multiplier stationarity test [62], Im–Pesaran–Shin(IPS) test [63], Fisher-type tests [64, 65], Breitung test [66] and so on.

2.1.1.1. Im–Pesaran–Shin test

IPS sets the regression equation for the Augmented Dickey-Fuller (ADF) test as follows.

$$\Delta y_{it} = \phi_i y_{i,t-1} + z_{it}^{'}\gamma_i + \varepsilon_{it} \qquad (2.1)$$

When the panel data satisfies the null hypothesis, the IPS determines that the data has a unit root and concludes that it is a nonstationary time series data. On the other hand, if the null hypothesis is rejected, it can be concluded that some cross-sectional data show a stationary time-series distribution.

The IPS test obtains each t-value from the cross-sectional data through an ADF t-test. Then, the unit root test is conducted by calculating the average for all i as follows:

$$\bar{t}_{N,T} = \frac{1}{N}\sum_{i=1}^{T}\overline{t_{i,T}} \qquad (2.2)$$

As the IPS test method can have various lag lengths for each cross-sectional data point, it is less required than other unit root test methods. Therefore, a more realistic conclusion can be drawn [67].

2.1.1.2. Fisher-type tests

The Fisher-type test uses an average statistic such as the IPS method, but performs the panel unit root test by using the p-value in a meta-analysis. This test is the most commonly used unit root test and is known to have the highest power [67]. The Fisher-ADF has been developed by Maddala and Wu [64], and the Fisher-PP by Choi [65]. Fisher-ADF uses the p-value of the statistic obtained from the ADF unit root test while Fisher-PP uses the p-value of the individual cross-sectional data as follows:

$$P = -2\sum_{i=1}^{N} \ln(p_i) \tag{2.3}$$

2.1.1.3.Breitung test.

The Breitung test is a unit root test proposed by Breitung [66]. The study of large-scale Monte Carlo simulation by Hlouskova and Wagner [68] revealed that the Breitung test has the highest power and smallest distortions among the generation panel unit root test. The regression model for the Breitung test is set as follows:

$$y_{it} = u_i + \beta_i t + \sum_{k=1}^{p+1} \alpha_{ik} x_{i,t-k} + \varepsilon_{it} \tag{2.4}$$

The null hypothesis is that the process is difference stationary.

$$H_0 : \rho_i = \sum_{k=1}^{p+1} \alpha_{ik} - 1 = 0 \tag{2.5}$$

The alternative is that the y_{it} is trend stationary, that is $\rho_i < 0$ for all i. To construct unbiased test statistic, Breitung [66] performed transformed vectors as follows:

$$Y_i^* = AY_i = [y_{i1}^*, \cdots, y_{iT}^*]' \tag{2.6}$$

$$X_i^* = BX_i = [x_{i1}^*, \cdots, x_{iT}^*]' \tag{2.7}$$

Under this assumption, following statistic has a standard normal distribution.

$$\lambda_{UB} = \frac{\sum_{i=1}^{N} \sigma_i^{-2} Y_i^{*}\,' X_i^{*}}{\sqrt{\sum_{i=1}^{N} \sigma_i^{-2} X_i^{*'} A' A X_i^{*}}} \tag{2.8}$$

2.1.2. Panel OLS

When estimating static energy demand models by panel analysis, it is common to explain unobserved heterogeneity by using fixed or random effects. The following linear regression model considers the heterogeneity of the panel data [69]:

$$y_{it} = \alpha + \beta x_{it} + u_i + e_{it} \tag{2.9}$$

In the fixed effect model, the error term u_i of the above equation is regarded as the parameter to be estimated. On the contrary, assuming u_i as a random variable would make it the random effect model.

In the case of the random effect model, the first-order autocorrelation problem of the error term is occurred using OLS estimation and cannot obtain an efficient estimator. Therefore, in order to obtain the efficient estimator in the random effect model, we need to use a generalized least squares (GLS) model that solves the autocorrelation problem. In addition, if the $COV(x_{it}, u_i) = 0$ assumption that implies the exogenous of explanatory variable is not established, then the random effect model estimator cannot be a consistent estimator.

When choosing between fixed and random effect models, the primary criterion is the inference of u_i, which refers to the heterogeneity of the panel data. If the panel data are derived from random sampling of the population, then the error term, u_i can be assumed to follow a probability distribution. In terms of econometric theory, the choice between the two models is determined based on whether the $COV(x_{it}, u_i) = 0$ assumption is established. This is called the Hausman test, which sets the null and alternative hypotheses as follows [70]:

$$H_0 : COV(x_{it}, u_i) = 0 \tag{2.10}$$

$$H_1 : COV(x_{it}, u_i) \neq 0 \tag{2.11}$$

If the null hypothesis is rejected (not rejected), then the random (fixed) effect model is more efficient. Therefore, this study conducts a panel analysis on the unobserved heterogeneity by using the Hausman test.

2.2 Keyword search volume

As interest in Big Data has increased recently, research using Big Data has been actively conducted [71-75]. Big data is not a simply collection of large amounts of data. Big Data contains a huge amount of data that has accumulated information that has not been utilized before, and also includes the concept of analyzing and utilizing it. When the information that has not been regarded as data such as internet search, SNS activity and smartphone movement path is collected, the size of the data exceeds the imagination. Amazon, Google, and Facebook analyze search and log information to provide customized services to customers. The Korean government is also seeking to utilize the Big Data through the public data portal and to utilize it efficiently.[2]

Google is now the most advanced company on Big Data and accounts for 90.91% of the global search market as of August 2018.[3] Google Trends shows the search frequency of a keyword based on all Google searches conducted globally and in real time. As Google Trends can use information that is yet to be announced, its data can predict the present. "Predict the present" refers to analyzing the current situation, which has not yet been officially announced and cannot be otherwise analyzed. However, traditional economic and business models rely on statistics gathered form government data, annual/quarterly reports, and financial statements. These economic statistics are published at least a quarter or a year later and are made available with a significant delay. In particular, GDP is often used as a major factor in economic analysis; quarterly GDP in the US is officially estimated about one month after the end of the reference quarter. Previous studies have used now-casting to obtain an "early estimate" before official figures are published [76]. In other words, they

[2] https://www.data.go.kr

[3] http://gs.statcounter.com/search-engine-market-share

estimates the target variables using previously announced official figures. For instance, when estimating GDP, we can use expenditure components related to individual consumption or production, which are available at a monthly frequency. However, as mentioned earlier, Google Trends allow us to collect and analyze data immediately.

Google Trends analysis has the more favorable capability to “predict the present” rather than “predict the future” [2]. In fact, recent research using Google Trends can predict the flu a few days before it actually happens [31]. Google Trends has also been used for predictions in the US and South Korean presidential elections, as well as trend analysis on Britain's Brexit. In this study, the analysis is conducted under the assumption that Google Trends data could be used as an explanatory variable in the econometric analysis. Therefore, if search frequency is more influential than existing explanatory variables, researchers should consider constructing a model based on the frequency of Internet-based searches.

Google Trends graphs trends for each keyword when selecting keywords. It also provides search volume for states and cities as well as countries, regional trends can also collected. Google Trends can track trends by selecting a date range starting in 2004 and providing search volume on a weekly basis. The search figure is provided as a relative search frequency by fixing the maximum value to 100, not the actual search volume. The Figure 2.2 shows the interest over time of each keyword trends and Figure 2.3 shows the interest by subregion of "renewable" keywords. However, in order to collect panel data by the Google trends, it is necessary to download each keyword trend of each state and reconstruct it.

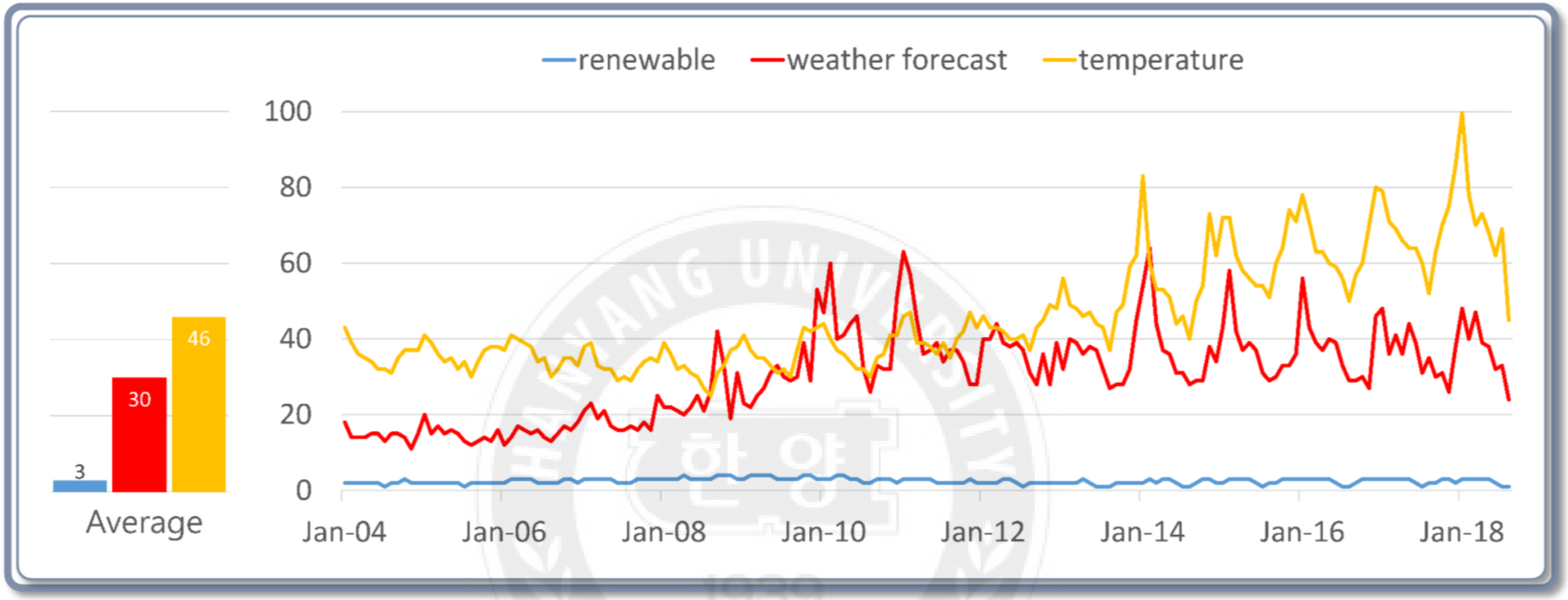


Figure 2.2. Interest over time of each keyword in Google Trends

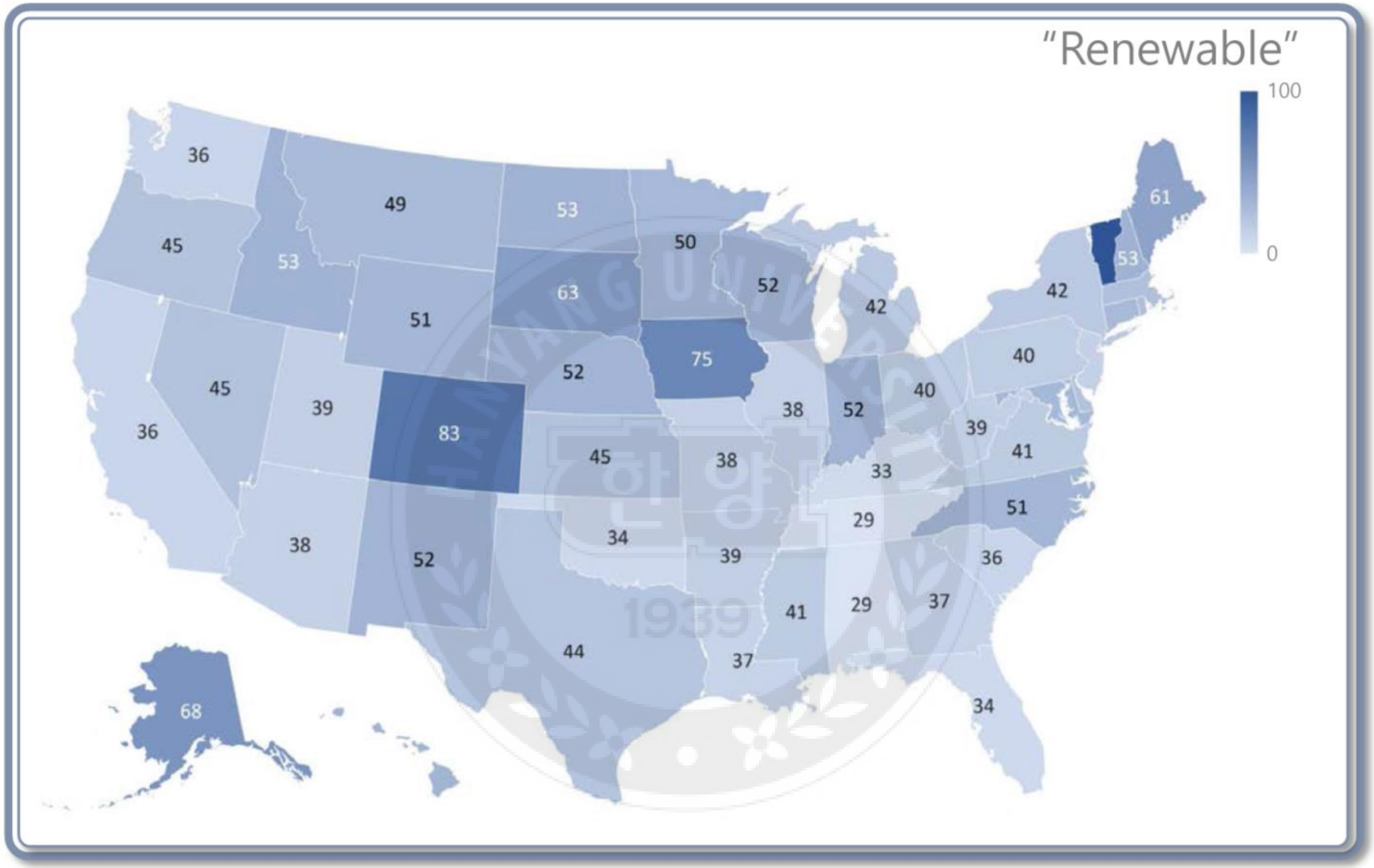


Figure 2.3. Interest by subregion of "renewable" keywords in Google Trends

2.3 Artificial neural network

ANN is a type of machine learning algorithm that learns the given information and identifies a nonlinear relationship with a target value [77]. In addition, because it has self-learning ability, it is applied to various fields such as finance, accounting and marketing. It is also possible to process complexity and non-linearity data and it is possible to analyze it regardless of qualitative and quantitative variables [78]. In other words, even if the econometric model based on the theoretical background does not exist or is complex, it is possible to construct a prediction model through ANN.

ANN have already been used as tools for prediction in various fields [39, 43, 46, 47, 52, 79, 80]. In particular, computer hardware and algorithms for analysis of ANN have been developed in recent years. Moreover, the volume of ANN-based research has been increasing over the last six years [81]. However, ANN analysis is avoided in economics due to the problem of interpretation of results, called the black box issue [82]. In fact, energy consumption or price prediction is very difficult due to internal complexity and irregularities in various interaction factors [83]. In econometrics, a prediction model has been set up using only historical data or a forecasting model using explanatory variables in a multifactor-influenced forecasting method. However, these models are difficult to apply in the real world because they have to assume model types in advance, and therefore the use of ANN that do not need to assume model types is increasing [84]. Additionally, the accuracy and usability of ANN forecasting has increased in recent years. It is also useful for analyzing phenomena that have the complexity and interdependence of multiple elements [85].

McCulloch and Pitts [86] proposed a computational model for neural networks based on mathematics and algorithms that mimic human brains. Rosenblatt [87] proposed a concept called perceptron and constructed a basic ANN that sets and outputs weights through learning. In the 1980s, it was re-examined as MLP. Rumelhart and McClelland [88] proposed a backpropagation algorithm and various studies have been made.

The ANN, which had been stagnated for 20 years, was revitalized due to the back propagation algorithm. By combining the back propagation algorithm with the MLP including the hidden layer, various problems such as the linear separation problem of the conventional perceptron method are solved.

MLP is a neural network in which one or more intermediate layers, i.e., hidden layers, are present between the input layer and the output layer. It has a hierarchical structure as shown in Figure 2.4. It was initially proposed as a feedforward network with no direct connection from the output layer to the input layer. The structure is similar to a single layer perceptron, but improves network capability by making the input and output of the hidden layer and each node nonlinear. In the MLP, as the number of layers increases, the decision section becomes more and more subdivided. That is, the number of regression lines that can separate the zones according to the number of layers increases. In the case of a three-layer structure, any type of zone can be formed in theory. Also, by applying the activation function of each layer as a sigmoid nonlinear function, the decision area can be represented by a gentle curve instead of a straight line, so that the back propagation learning algorithm can be performed.

The learning method of the general MLP is as follows. When input data is presented to each node of the input layer, it is transformed at each node, transferred to the hidden layer, and output to the output layer. This output value is compared with the target value and the connection weight is adjusted in the direction of reducing the error. In supervised learning, input and target value patterns are presented to the network. The network compares the output pattern, which changes as the input pattern given to the input layer propagates to the output layer, with the target value. If they match, the learning stops and if they do not match, the connection weight of the network is adjusted in the direction of reducing the error and the learning proceeds.

Back propagation learning algorithm is the most used algorithm in MLP. In a MLP with one or more hidden layers between the input layer and the output layer, the error back propagates from the output node to the internal node and adjusts the connection weight. Back propagation learning algorithm requires a large amount of learning data until the learning process converges, and it is impossible to correct the learned pattern or to learn further, but it is the most widely used algorithm because it is easy to implement and quick learning is possible [78].

ANN can be classified according to its structure and learning algorithm. The structure is classified according to the connection method consisting of an input/output layer and a hidden layer, and the learning algorithm is classified by adjusting the weight of each layer. The hidden layer is weighted according to the data received from the input layer and transmits it to the output layer [89]. The algorithm provides a way to adjust the weights among nodes in all layers as the neural network learning proceeds [78].

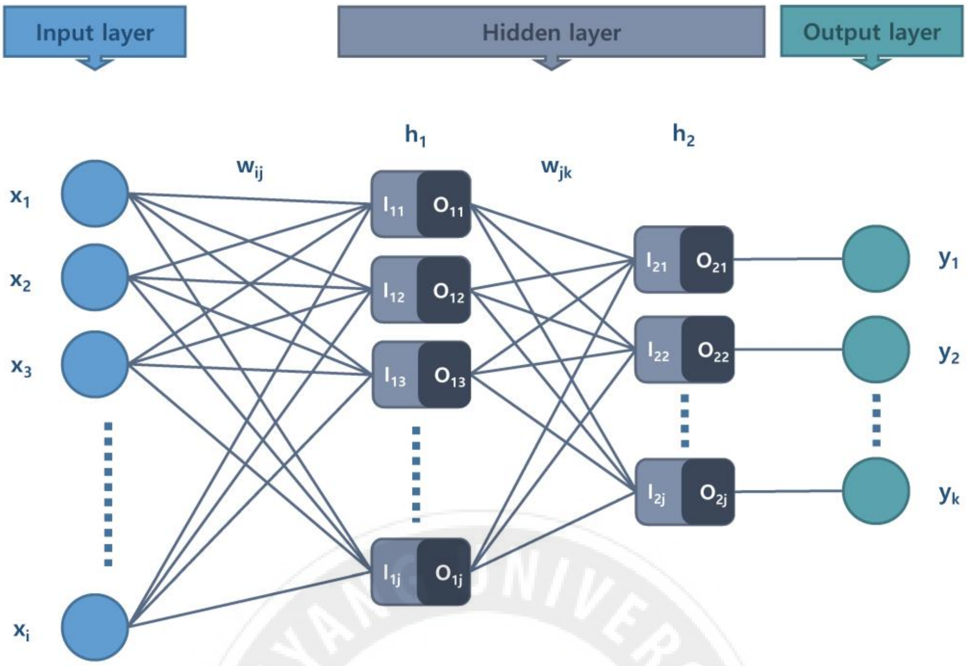


Figure 2.4. Multilayer artificial neural network structure

Figure 2.4 is a multilayer artificial neural network structure. The relationship between input layer, hidden layer, and output layer is shown here.

The input value of the first hidden layer I_{1j} can be computed as:

$$I_{1j} = \sum_i w_{ij} x_i \tag{2.12}$$

x_i is the i-th input value of the input layer, w_{ij} is the weights that connects the i-th input layer and the j-th hidden layer, and I_{1j} is input to the hidden layer through the weight.

The output value of h_{1j} is calculated through the activation function, $f(\cdot)$.

$$O_{1j} = f\left(I_{1j}\right) \tag{2.13}$$

This value is again calculated as a weight between hidden layers, and output through the activation function.

$$O_{2j} = f\left(\sum_j w_{jk} O_{1j}\right) \tag{2.14}$$

Through the above procedure, the total error is calculated as given below.

$$E_{tot} = \frac{1}{k}\sum (y_k - O_{2j})^2 \tag{2.15}$$

After calculating the error of the output value, the weight of each layer is updated by gradient descent as given below.

$$w_{jk}{}^{new} = w_{jk} - \gamma \frac{\partial E_{tot}}{\partial w_{jk}} \tag{2.16}$$

Where γ is the learning rate, and partial derivatives are computed through back propagation [90]. The back propagation algorithm is done through the following procedure, where the bias is omitted for the sake of explanation of the structure. The bias is a value that is set to prevent the situation where the output value of each node becomes zero.

First, back propagation of the weights in the first layer from the output proceeds as follows. w_{jk} can be represented by three partial differential equations by chain rule as given below.

$$\frac{\partial E_{tot}}{\partial w_{jk}} = \frac{\partial E_{tot}}{\partial O_{2j}} \frac{\partial O_{2j}}{\partial I_{2j}} \frac{\partial I_{2j}}{\partial w_{jk}} \tag{2.17}$$

$$\frac{\partial E_{tot}}{\partial O_{2j}} = -(y_k - O_{2j}) \tag{2.18}$$

Assuming $f(\cdot)$ is a sigmoid function, O_{2j} is the result of putting I_{2j} into sigmoid function. The derivative of the sigmoid function is as follows, where the above partial differential equation can be recalculated.

$$\frac{\partial O_{2j}}{\partial I_{2j}} = sigmoid(I_{2j}) \cdot \left(1 - sigmoid(I_{2j})\right) \tag{2.19}$$

Finally, if I_{2j} is differentiated by w_{jk}, O_{1j} is computed. In combination with the above equation, it is possible to determine how w_{jk} affects the total error.

$$\frac{\partial I_{2j}}{\partial w_{jk}} = O_{1j} \tag{2.20}$$

$$\frac{\partial E_{tot}}{\partial w_{jk}} = -(y_k - O_{2j}) \cdot sigmoid(I_{2j}) \cdot \left(1 - sigmoid(I_{2j})\right) \cdot O_{1j} \tag{2.21}$$

The calculated value is substituted into the equation (2.16) to calculate the updated weight value. In this way, each weight is updated to back propagate and iterative learning is performed until the error reaches the target value.

The sigmoid function used in updating the weights is one of the activation functions.

Activation functions are used in various forms in many algorithms of artificial intelligence. Also, it is very important to use an appropriate activation function because the output value varies depending on which activation function is used. The input value of the node is output via the activation function and is determined to be activated in the next step. The most widely used function among the activation functions is shown in Figure 2.5. First, the most basic activation function is a step function, which has the same shape as a staircase, and is calculated as 0 or 1. The sigmoid function has a continuous output value as a nonlinear function. It is one of the most widely used functions because it has a smooth curve that can be differentiated for every x. The last activation function is a rectified linear unit (ReLU) function, which is a linear function. This is one of the functions that have been used recently to solve the problem of gradient vanishing in sigmoid function. In the sigmoid function with a value between 0 and 1, when the layer increases, the weight can be converges to 0, which is called the gradient vanishing problem. To solve this problem, ReLU has been developed and has the advantage that the differentiation is simple. The neural network learning of this study is done using MATLAB program and the preprocessing is done through minmax method.

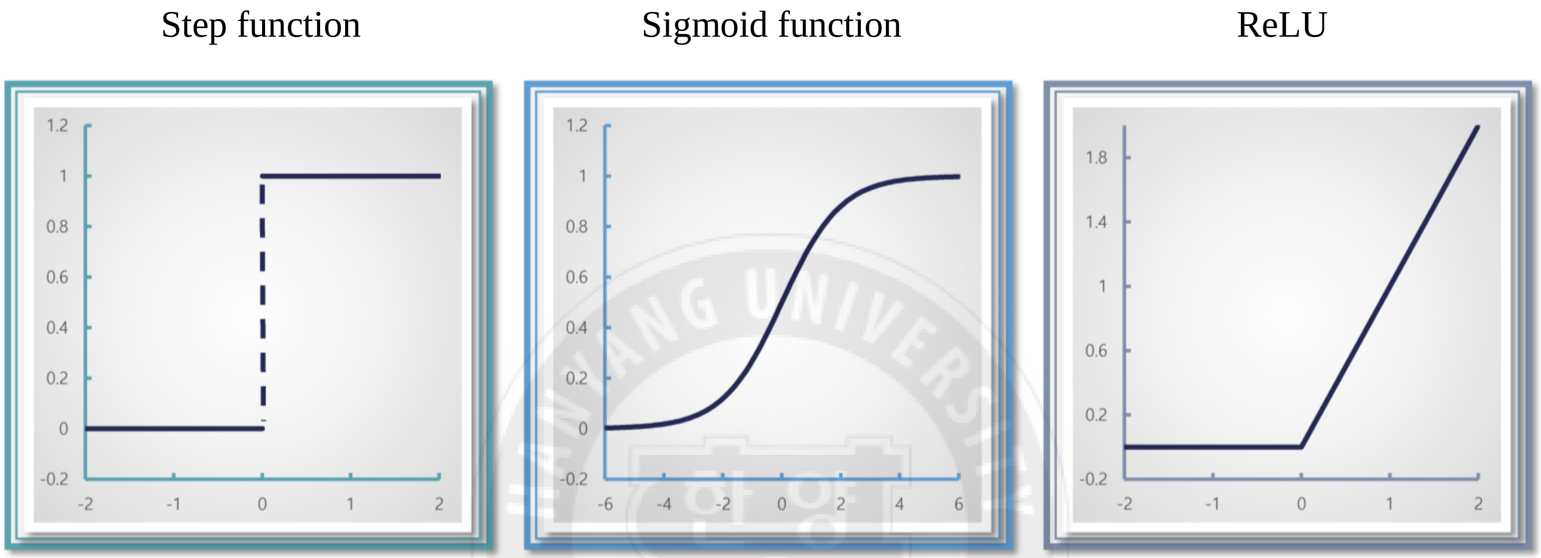


Figure 2.5. Type of activation functions

Chapter 3. Development of panel analysis model

This study analyzes the correlation between interest in US renewable energy and US household electricity consumption by using keyword search volume to “predict the present.” Currently, US household energy consumption accounts for about 22% of total energy consumption, and while electricity consumption comprises about 35%. This significant level of electricity consumption requires detailed analysis over time [91]. This study employs panel methodology to consider the effect across the United States. Studies that have adopted panel methodology to derive the factors that influence household electricity consumption have used the following explanatory variables. Alberini and Filippini [18] set electricity and gas prices, population, income, HDD, and CDD as explanatory variables. Constructing similar explanatory variables, Alberini *et al.* [17] select various domestic characteristics as variables instead of using population variables. Azevedo *et al.* [22] adopt only electricity prices, consumption expenditure, and HDD as explanatory variables; CDD is excluded because of the lack of EU data. Filippini [21] uses peak and off-peak electricity prices together with HDD and CDD to analyze electricity consumption per period. Hence, most electricity demand analyses include HDD and CDD as explanatory variables. In addition, these climate factors play an important role in the electricity market [92]. Therefore, this model sets HDD and CDD as explanatory variables in the household electricity demand analysis and utilize keyword search volume to explore the effect of renewable energy on electricity consumption.

The model considers three keywords related to electricity consumption: “renewable,” “weather forecast,” and “temperature.” The “weather forecast” and “temperature” keywords are selected to assess whether the weather variables represented by HDD and CDD could be replaced by keyword search volume data. On the other hand, the “renewable” keyword is chosen to ascertain the impact of renewable energy on electricity consumption.

3.1 Panel data

This study uses monthly data for all states in the United States (51 units including the 50 states and the District of Columbia) from September 2013 to June 2016. Table 3.1 shows the descriptive statistics of the main variables. This model uses retail electricity sales as a proxy of household electricity consumption. The US Energy Information Administration (EIA) provides the retail sales and average retail prices.[4] HDD and CDD are obtained from the Degree Days website.[5]

Keyword search volume data on the key variables are downloaded from Google.[6] Google provides real-time data on keyword search volumes with significant traffic, namely the search volumes for each month compared with the largest search volume over the selected range. As shown in Table 3.1, the Google Trends data normalizes the reported volumes against the highest value for the respective keyword, which is set to 100. This normalization is very important. As the number of people searching on Google constantly increases, comparisons using raw search values are not possible. It is also impossible to compare across regions because population varies by state or country.

Google Trends data reflects a fixed maximum value and minimum values that change depending on the region or period. Since the data follow normal distribution, the characteristic that the mean and standard deviation change according to the minimum value can be confirmed in Table 3.1. In addition, due to the characteristic that the spike point for each variable varies by period, it is necessary to analyze each variable separately. A spike

[4] https://www.eia.gov/
[5] http://www.degreedays.net/
[6] https://www.google.com/trends/

point refers to a sudden acceleration of search interest in a particular subject as compared to the general search volume. To make direct variable comparisons, all variables must be added during the data collection to extract relative values. In this study, each variable is separately applied to the formula.

Table 3.1. Descriptive statistics of the main variables

Variable	Mean	Std. Dev.	Min	Max
Household electricity consumption (million kWh)	2253.457	2367.554	124.487	17143.150
Electricity price (cents/kWh)	13.125	4.098	7.700	38.270
HDD	171.182	196.030	0.000	942.000
CDD	108.685	125.153	0.000	644.000
“renewable”	36.424	17.720	3.000	100.000
“temperature”	68.254	14.857	28.000	100.000
“weather forecast”	44.557	16.226	12.000	100.000

3.2 Panel analysis model

This study analyzes the relationship between keyword search volume and household electricity demand through a panel analysis. Monthly energy consumption is usually influenced by many external factors [93]. Three regression equations are constructed based on the static energy demand model, as follows:

$$E_{i,t} = \alpha + \beta_1 P_{i,t} + \beta_2 HDD_{i,t} + \beta_3 CDD_{i,t} + \beta_4 RE_{i,t} + \mu_i + e_{i,t} \tag{3.1}$$

$$E_{i,t} = \alpha + \beta_1 P_{i,t} + \beta_2 HDD_{i,t} + \beta_3 CDD_{i,t} + \beta_4 T_{i,t} + \mu_i + e_{i,t} \tag{3.2}$$

$$E_{i,t} = \alpha + \beta_1 P_{i,t} + \beta_2 HDD_{i,t} + \beta_3 CDD_{i,t} + \beta_4 WF_{i,t} + \mu_i + e_{i,t} \tag{3.3}$$

Where $E_{i,t}$ denotes US household electricity consumption, $HDD_{i,t}$ and $CDD_{i,t}$ are HDD and CDD, respectively, i denotes the respective state of the United States, and t denotes the time period. P is the electricity price; RE, T, and WF are the keyword search volume "renewable," "temperature," and "weather forecast," respectively; and μ and e are the error terms.

As explained above, this analysis selects six explanatory variables including three different keyword search volume. Of these keywords, "temperature" has been considered to be the most important variable when electricity is used for heating purposes; temperature sensitivity to electricity demand is increasing, thereby making it an important factor in determining demand [14, 15]. However, this model uses HDD and CDD instead of temperature, because these have non-linear relationships with electricity consumption [15]. These variables are selected to take on the value of the respective capital city. This is not only due to data availability but also because of small temperature variations within the same states and the generally larger populations of the capital cities [94]. The EIA publishes average household electricity prices, which are calculated by dividing the revenue of utilities in the household sector by electricity sales to this sector. However, these data are not reflective of all power suppliers (in the case of retail power companies); therefore, whether average household electricity price is an appropriate variable is questionable, and the endogenous problem of price data has also been raised [18]. Hence, this study analyzes the effect of the price variables by setting them as explanatory variables.

Household income is excluded from this model. As shown in previous studies, the income-electricity consumption relation is inelastic in the short term [17, 20, 23]. Therefore, this study estimates the short-term relationship to observe the effects of keyword search volume, assuming that household income does not change significantly in the short term. Household income data are excluded from the variables and included in the μ error term, which indicates the individual characteristics of the panel that do not change over time. In fact, income and population do not show large fluctuations in the short run.[7] In addition, household income cannot be included in the analysis because of the unavailability of monthly data.

[7] Median income (year): \$54,525 (2013), \$53,718 (2014), \$56,516 (2015). Source: US Census Bureau, https://www.census.gov/data/tables/2016/demo/income-poverty/p60-256.html

3.3 Keyword search volume

The model choses “weather forecast” and “temperature” as keyword search volume. As mentioned in Section 3.2, temperature is one of the important factors in determining household electricity consumption, and analyze the impact of these keywords.

The main analysis keyword “renewable” is selected for two reasons. The first reason is the growing interest in renewable energy [95, 96]. In the industrial sector, this interest has already been reflected in policy matters such as national penalties and subsidies to reduce carbon emissions. This sector is also attempting to reduce energy consumption through renewable energy generation. In the household sector, although a variety of renewable energy is used, it is difficult to derive the economic implications of such use as it is not converted into a quantifiable value. [8] Nevertheless, since household electricity consumption in the United States accounts for 35% of total consumption,[9] it should be considered together with industrial and commercial consumption. Therefore, this model uses keyword search volume to estimate the impact of renewable energy. Keyword search volume data can be used as an explanatory variable since it is able to provide statistical data continuously and in real time. Although the “renewable” keyword may affect household electricity consumption, it has not been considered by previous studies due to the lack of quantified values. Therefore, the model reflects this variable by using keyword search volume.

The second reason is that because of renewable energy, some end users are not only consumers but also producers, which can lead to changes in domestic demand patterns [97].

[8] In the US Census Bureau, only geothermal energy, solar energy, and wood count toward residential renewable energy use and only data from 2014 are collected.

[9] http://www.eia.gov/electricity/

If expenditure on electricity consumption increases, consumers will endeavor to reduce spend. One way of doing so is by using renewable energy. Consumers will try to generate their own electricity by installing various renewable energy facilities. Regardless of whether these electricity generation facilities are efficient, this will reduce electricity consumption in the short term.

Since the “renewable” keyword have a fairly wide range of meanings, and most of the household renewable energy facilities are photovoltaic facilities, it is necessary to further refine the keyword selection. However, this study uses the “renewable” keyword because it has not been long since the data of the Google Trends has been provided and there are missing data by region and time. Also, since the main purpose of this study is the utilization of keyword search volume model construction, it is reasonable to focus more on model construction.

In the multicollinearity problem between the keyword search volume and the temperature variable, if both variables are fully multicollinearity, it is impossible to estimate, but it is still BLUE (best linear unbiased estimation) because it is not in a perfect cointegration relationship[98]. The "temperature" and “weather forecast” keywords are explanatory variables that indicates the number of search frequencies, while the HDD and CDD are explanatory variables that indicate the temperature. It can also reduce the problem of multicollinearity by building panel data. Therefore, this study tries to solve the multicollinearity problem through the construction of panel data.

3.4 Result and discussion

Table 3.2 presents the unit root test results for each variable. If there is a unit root (i.e., both the dependent variable and the independent variable are unstable time series), the regression model would need to be estimated after the first difference. However, all the variables are stationary.

It can be seen from Table 3.2 that all the variables are stationary, which implies that there is no need for a cointegration test. Therefore, the second step is to apply the Hausman test to assess whether the unobserved heterogeneity is explained as a fixed or a random effect. As shown in Table 3.3, the null hypothesis of the Hausman test is rejected at the 1% level. That is, the Hausman test shows that the fixed effect estimation is more robust than the random effect estimation.

Table 3.2. Unit root test results for each variable

Variable	IPS	Fisher-ADF*	Fisher-PP*
Household electricity consumption	-20.735***	-26.330***	-15.491***
Price	-8.293***	-17.047***	-10.079***
HDD	-22.089***	-23.752***	-13.235***
CDD	-23.682***	-21.094***	-17.088***
“renewable”	-16.045***	-19.584***	-21.485***
“temperature”	-18.926***	-18.082***	-21.474***
“weather forecast”	-20.899***	-20.734***	-20.748***

Note: * denotes the inverse normal Z statistic[10]
*** (**) denotes rejection of the null hypothesis at 1% (5%) level

[10] Choi [65] simulation results suggest that the inverse normal Z statistic offers the best trade-off between size and power.

Table 3.3. Hausman test results of each equation

Model	Equation (3.1) Keyword: “renewable”	Equation (3.2) Keyword: “temperature”	Equation (3.3) Keyword: “weather forecast”
Hausman statistic	66.68***	18.96***	29.03***

Note: *** (**) denotes rejection of the null hypothesis at 1% (5%) level

Therefore, based on the results of Hausman test, the fixed effect model is selected for estimation in this study. The estimation results are shown in Tables 3.4-3.6. In all three keyword models, the price variable cannot be rejected at the 10% level.

In the case of the model using the “renewable” keyword as an explanatory variable, all the variables except the price variable are statistically significant at the 1% level (Table 3.4). Both the HDD and the CDD variables have positive coefficients; hence, household electricity consumption increases by 1.850 and 2.522 million kWh when HDD and CDD increase by one unit. On the contrary, the “renewable” variable has a negative coefficient and a much greater effect than HDD and CDD.

In the model using the “temperature” keyword, the HDD and CDD variables also have positive coefficients: household electricity consumption increases by 2.218 and 3.532 million kWh, respectively for every one unit increasing in HDD and CDD (Table 3.5). The “temperature” keyword variable also has a negative coefficient, but it has a similar size of effect as HDD.

For the “weather forecast” variable, the results cannot be rejected at the 10% level (Table 3.6). Thus, the “weather forecast” variable has no explanatory power for residential electricity demand and it is preferable to exclude it from the influence factors.

Table 3.4. Fixed effect results choosing the “renewable” keyword

Household electricity consumption	**Coef.**	***t*-statistic**	$P > \|t\|$
Price	-14.604	-1.11	0.266
HDD	1.850	19.56	0.000***
CDD	2.522	15.21	0.000***
“renewable”	-16.017	-13.68	0.000***
Constant	2437.807	13.08	0.000***

Note: *** (**) denotes statistical significance at the 1% (5%) level

Table 3.5. Fixed effect results choosing the "temperature" keyword

Household electricity consumption	**Coef.**	***t*-statistic**	$P > \|t\|$
Price	-11.498	-0.830	0.406
HDD	2.218	20.990	0.000***
CDD	3.532	22.430	0.000***
"temperature"	-2.542	-2.140	0.032**
Constant	1814.357	9.090	0.000***

Note: *** (**) denotes statistical significance at the 1% (5%) level

Table 3.6. Fixed effect results choosing the "weather forecast" keyword

Household electricity consumption	**Coef.**	***t*-statistic**	**$P > \|t\|$**
Price	-12.727	-0.920	0.359
HDD	2.129	20.890	0.000***
CDD	3.609	23.210	0.000***
"weather forecast"	0 .007	0.010	0.994
Constant	1663.447	8.770	0.000***

Note: *** (**) denotes statistical significance at the 1% (5%) level

The results show that the coefficient of the price variable cannot be rejected at the 10% significance level. A long-term price elasticity on the US household electricity consumption used to be significant [17, 18, 20]. In a short-term, however, it can be inelastic [18, 20]. The insignificant result for the coefficient of the price variable dose thus not violate a common belief on a price elasticity since our model is estimated using monthly data. Note that we do not take a logarithm on variables because of the characteristics of keyword search volume data. In addition, the purpose of this study is to examine the correlation of keyword search volume with household electricity consumption; so, this study focuses on the keyword search volume instead of price.

In the case of the HDD and CDD variables, all previous studies have yielded positive coefficients that are significant at 1% significance level. However, the reason why the coefficients differ across these studies is likely due to the data period (e.g., summer or winter). For example, Salari and Javid [20]use data from 2005 to 2013, leading to HDD estimates of 0.21 and CDD estimates of 0.09. These results differ from the estimated values of this study. Since Salari and Javid [20] use annual data and logarithms are taken of all the variables to determine elasticity, it is unreasonable to compare these coefficients directly. As mentioned earlier, this difference occurs due to the data period. Also, these variations can be explained by the fact that it has not been long since households have started using electricity for heating and cooling.

Regarding keyword search volume, which form the focus of this study, the keyword “temperature” has a negative coefficient of -2.542, which is similar to HDD. In other words, if the frequency of searching for “temperature” in the home increases by one unit, household electricity consumption reduces by 2.542. On the other hand, the effect of “renewable” keyword is large. This search term also has a negative correlation with household electricity consumption. If the search frequency increases by one unit, household electricity consumption decreases by 16.017 million kWh. The correlation between the two variables, which intuitively seem to be unrelated, could have significant meaning. When one searches for "renewable" in the context of their household, they probably have a clear purpose. In the event that excessive electricity is consumed or electricity bills are high, homes will search for alternatives to reduce electricity consumption (e.g., installing renewable appliances). In the case of households equipped with renewable energy facilities, the power consumption will decrease in proportion to the capacity, and the results of the estimation can be seen.

From the results, it can be seen that interest in renewable energy affects electricity demand. Nevertheless, there was no quantitative data on renewable energy for the analysis. As introduced in the Literature review section, there were positive correlations between keyword search volume and social activities such as "depression" and suicide death rate [36], "suicide" and intentional self-injury [34], "breast cancer" and attack rate of breast cancer [28], and "Trucks & SUVs" and motor vehicle and parts sales [2]. This study tries to extend this approach toward an energy economics filed. Consequently, this study could find out a significant influence of keyword search volume on household electricity demand.

Chapter 4. Development of panel ANN model

In the meantime, there are very few studies in economics that have used panel data for prediction by ANN. Existing studies use panel data without differentiating between entities in the model structure [99-101]. This can be seen as not taking advantage of panel data. Pao and Chih [99] conducted ANN forecasting using panel data from high-tech companies in Taiwan for three years. The firm's debt ratio was estimated using eight independent variables as input nodes. Panel data from 207 companies were used, but the model was learned under the same conditions without considering the heterogeneity of each company. In other words, a pair of data sets was transmitted to the network one after another without distinction between companies. Al Shami *et al.* [100] predicted four indexes that represent the competitiveness of a knowledge-based economy. A model was constructed in which the input and the output were set to the same four indexes, and it was learned without discrimination by country. Crane-Droesch [101] constructed an ANN model using the weather and soil variables to predict the yield of corn. Input variables are divided into two categories, and time-invariant data such as soil, latitude, and longitude are configured to affect the output without going through the hidden layer. However, each US state variable and its corresponding time-invariant data are used to calculate the predicted values for each state in this model structure. In other words, the yield of a specific US state is forecasted using the model that learns from the information of all US states. Like other studies, this eventually results in not learning each state independently. It is necessary to compare the prediction accuracy of the models that have been learned with the time series data of a particular state.

The panel ANN studies mentioned above did not differentiate between states and national data or have independent learning such as the pooled OLS method. Pooled OLS is a panel analysis method that is rarely used because it involves very strong assumptions and causes autocorrelation, heteroscedasticity, and endogenous problems in almost all panel data [102]. Therefore, this study constructs a panel ANN structure using the advantages of panel data and analyzes its accuracy according to the change of forecasting periods. This study intends to improve the accuracy of predicted values when the forecast point increases by learning the unobserved heterogeneity contained in panel data from each state.

However, short- or long-term forecasting and the number of predictions need to be differentiated. Short-term forecasting refers to predicting the near future, regardless of the number of predictions. Additionally, long-term forecasting considers only the length of a prediction period, regardless of the number of predictions, and is used in research that is intended to be used several decades later. On the other hand, the number of estimated predictions is not related to the long- or short-term prediction. In particular, machine learning, including ANN, is differentiated by increasing the number of predictions and not by specifying the learning period. This is because ANN model is not designed to distinguish between monthly and yearly time-series data. Therefore, it is necessary to clarify that a large number of predictions are not long-term predictions and that a small number of predictions are not short-term predictions (Figure 4.1)

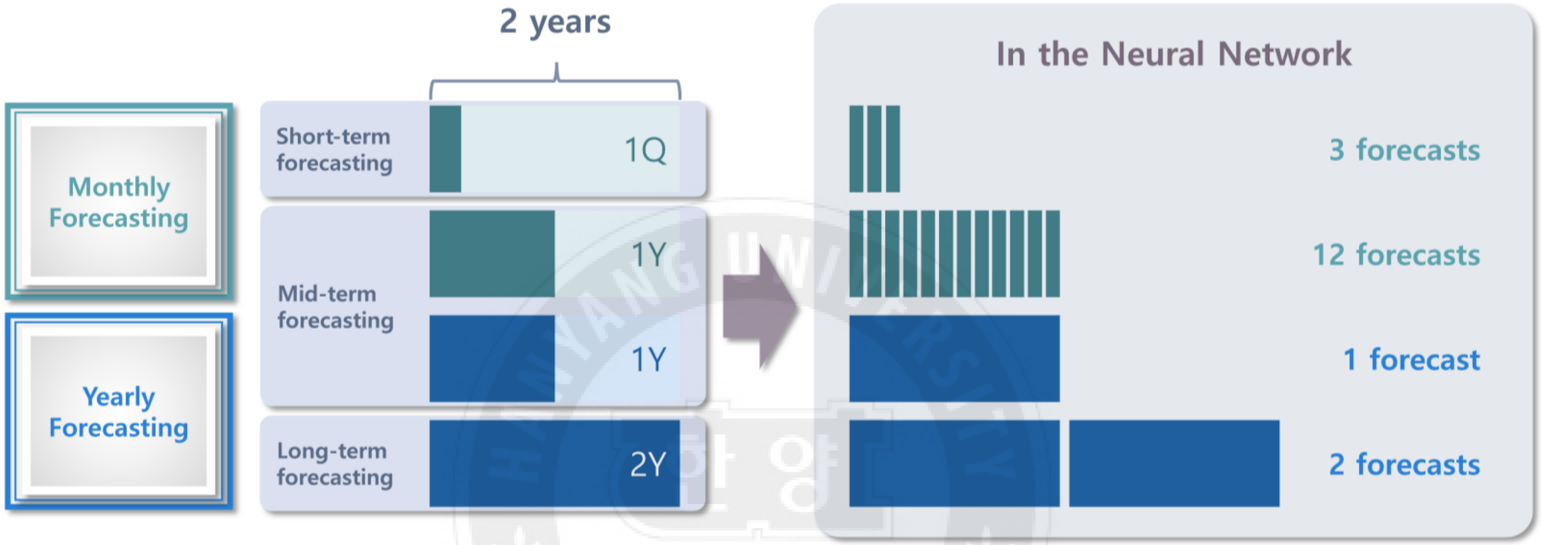


Figure 4.1. Conceptual difference between short-term, long-term prediction and number of predictions

.

4.1 Model development

As mentioned above, panel analysis removes the cross-sectional dependence in the unobserved heterogeneity of the panel data. Unlike panel analysis, a model is constructed structure to learn the unobserved heterogeneity of the panel data. In other words, panel analysis removes the heterogeneity of the panel data, but our model uses this to learn.

As seen in Figure 4.2, each US state consists of different layers and the weights of all layers are added to the last layer. Finally, one output (here, the US average data) is calculated. This is constructed for learning the unobserved heterogeneity of each state. The learned weights of each state are calculated through the last layer. Therefore, the layers of each state are learned without relation and have unique weights.

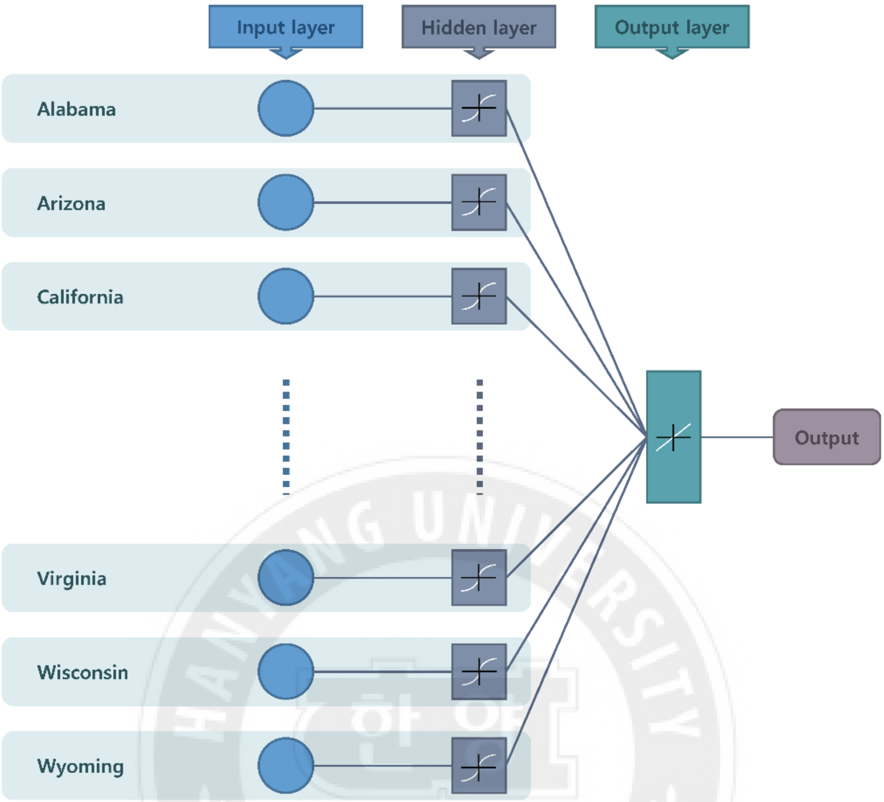


Figure 4.2. Panel ANN structure

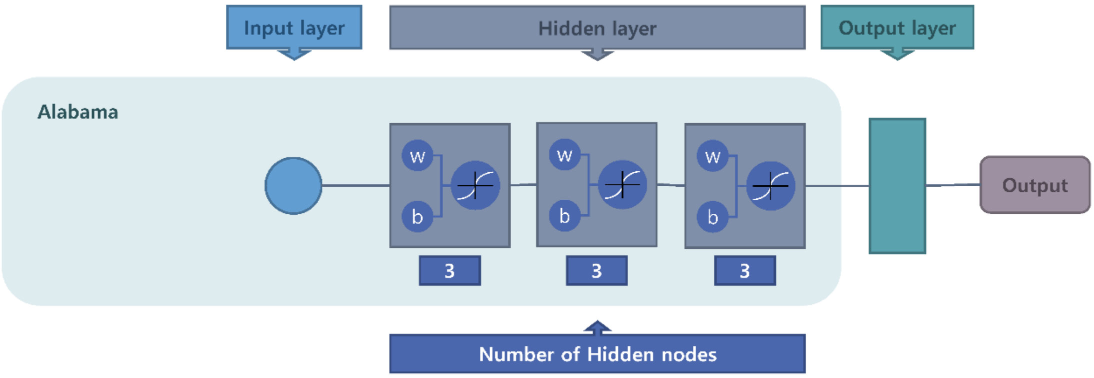


Figure 4.3. Each layer structure of Panel ANN

Looking at the structure of each state, there are two or three layers for each state. (Figure 4.3). If there is only one layer, adjusting weights by the gradient descent method results in adjustment in only one direction, and so two or three layers are inserted and the results of each model are compared. In other words, by increasing the number of layers, the number of selected regression lines is increased to enhance the learning effect. Also, the number of hidden nodes in each layer is adjusted to 2 ~ 3, and the results are compared. According to Demuth *et al.* [103], 2 or 3 hidden nodes were found to be suitable. Simply increasing the hidden nodes is a result of repeatedly drawing the regression line at one point. The structure described in this paragraph applies equally to the existing ANN method, which learns only US average data and not panel data. This is done to compare the panel and the time series data models under the same conditions.

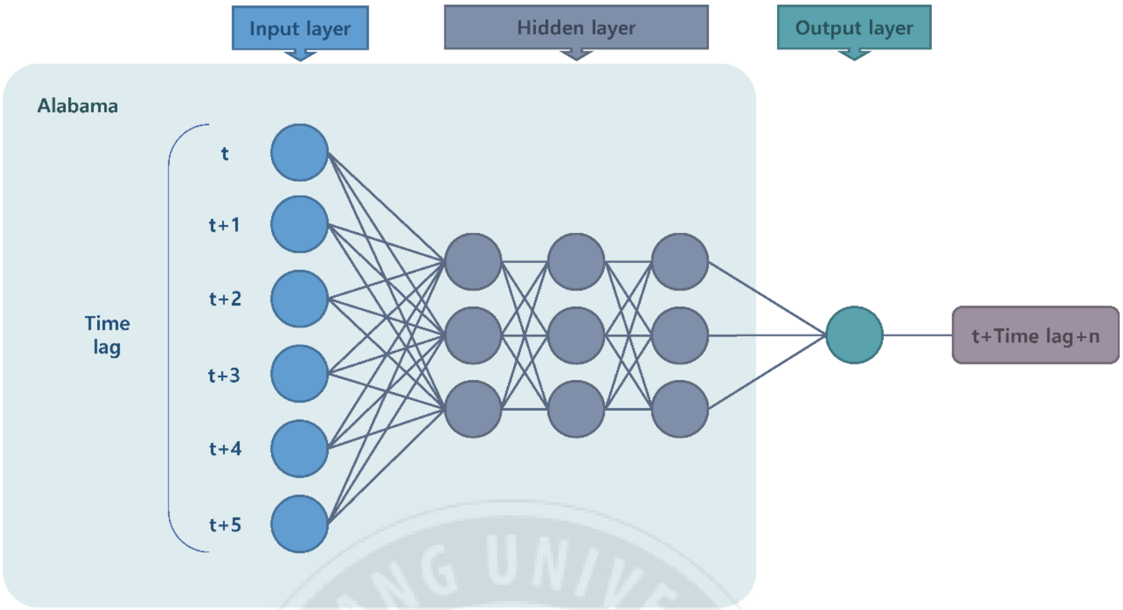


Figure 4.4. Cascade method in each layer structure of Panel ANN

As seen in Figure 4.4, the input for each state is composed by a cascade method. Since the time series data have monthly trends, the model is set the time lag from 1 to 12 to find the most suitable structure [5]. Therefore, 1 ~ 12 input nodes are inserted for each state. As a result, the prediction is carried out by combining the above three structures. In other words, the number of inputs is determined by the number of time lags for each state (time lag multiplied by US states in this empirical study), and the number of outputs is one. The difference from [99-101] is that the panel data set used in this study had a single variable, and is not multivariate. The multivariate input set is replaced by the time lag set in this panel ANN model.

The learning ANN model consists of 2 ~ 3 layers and 2 ~ 3 hidden nodes by states, and finally generates one prediction value. In other words, the optimal learning model is constructed by transforming the hidden layer and hidden node into 2 ~ 3. The learning function is the Levenberg-Marquardt algorithm, and the learning weight is converged using gradient descent in the backpropagation. The data used in this study is for 50 states in the United States. Since there are 50 entities and the amount of data to be inserted at one epoch is large, the number of layers and the activation function are both set to a simple structure. Four periods are forecasted: 6, 12, 18, and 24 months.

One of the problems that can occur when learning a neural network is to include the predicted data for comparison with the learning data and to calculate the most suitable model by adjusting the number of layers and nodes. This learned neural network model will obviously have a small error, but the risk of overfitting is very large, and it is a way to fade the meaning of predictions. Therefore, in this study, the prediction comparison data (6, 12, 18, or 24 monthly data) is removed beforehand, and learning is performed by selecting the validation set and the test set to find the most suitable model. Then, we insert the recent input data into the learning completed model and calculate the predicted value as shown in Figure 4.5.

Additionally, this model does not predict whole forecasting months in one learning but learn every month and calculate the result. As mentioned before, this model did not learn by putting the predicted value into the input data because learning is performed by excluding the predictive comparison data. As seen in Figure 4.5, we set up a model structure that repeats learning to predict each month. The model forecasting by using input data from t-n to t-1 predicts the first month and forecasting t + 1 with the same input value predicts the next month. Values are predicted for 6, 12, 18, or 24 months through this iterative learning, and the total RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) are calculated and compared. This is a method of adopting a kind of ensemble technique. The ensemble method is known as a better technique for generalization by bagging, i.e. randomly dividing the training set, and calculating the error value of the whole after learning [104]. Therefore, the value calculated by the ensemble method is advantageous for generalized model construction compared with the predicted value at 6 months or 24 months at once, and overfitting problem is also solved.

For comparison, electricity price is also predicted by the ANN structure used in earlier studies. This is in the same context as the use of pooled OLS as a comparison for existing panel research. Each state’s price is learning through one hidden layer as shown in Figure 4.6. The difference from the above model is not only in the separation of the hidden layers, but also in the cascade-based input configuration. In other words, a pair of data sets regardless of the state is transmitted to the network one after another [99].

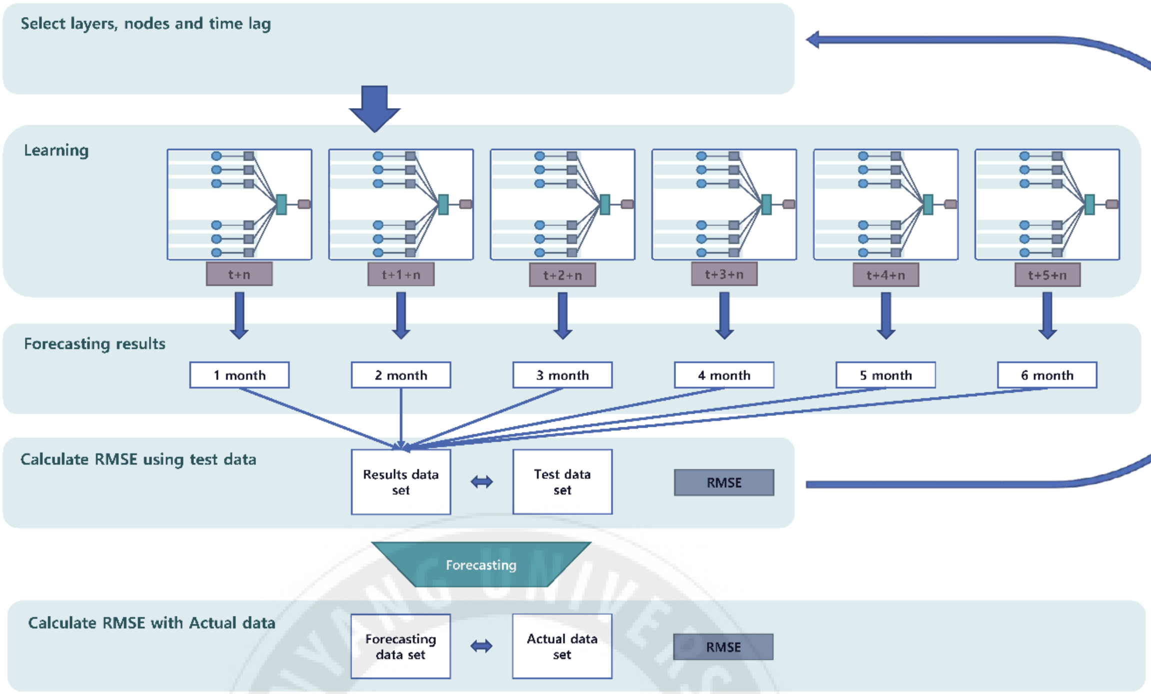


Figure 4.5. Learning and prediction flow of Panel ANN

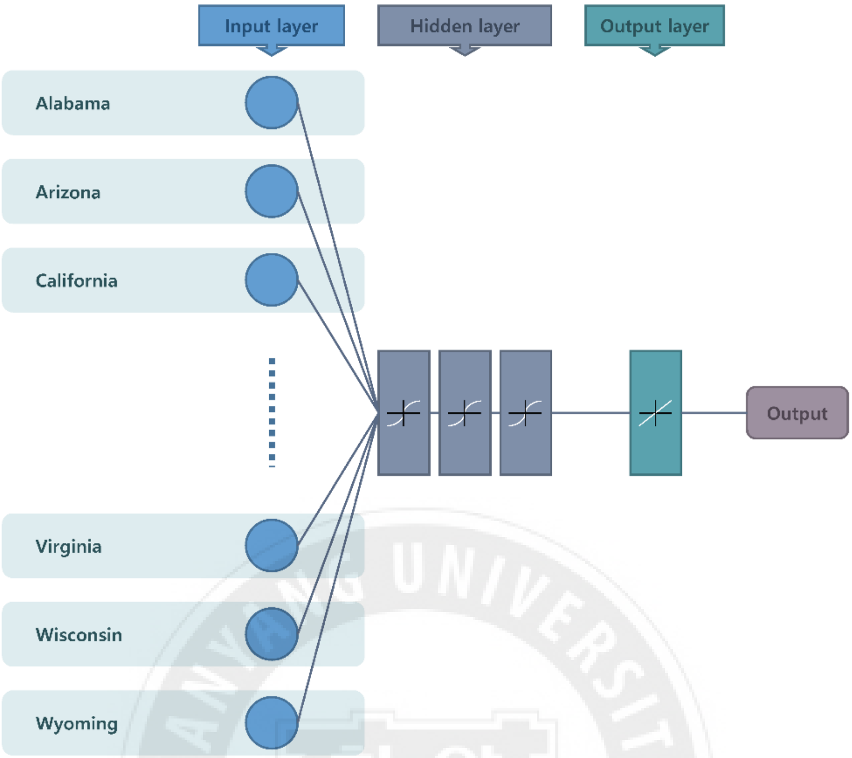


Figure 4.6. Previous panel ANN method

4.2 Panel Data

This study analyzes the impact of panel data on the prediction of ANN. For this purpose, empirical analysis is conducted using panel data of US electricity price by states. For additional model verification, US natural gas city gate price forecasting is also performed. Obtaining panel data for research is difficult, so this model uses gas prices, which are easily available. In case of Korean electricity price, SMP (system marginal cost) is decided by the KPX (Korea power exchange) and it is added to CP (capacity payment). In addition, since KEPCO (Korea electric power corporation) does not publish data for building regional panel data, there is a problem in building data. Therefore, US electricity prices are used to verify the model.

For national panel, it is not possible to obtain monthly data, and even for US states, monthly data is limited. The reason for using monthly data in this study is that the amount of annual data that can be used to learn is limited, because annual data is available only for 50 years. Sufficient data should be available to learn the characteristics of each state using ANN method rather than the econometric analysis method, and therefore this model limits learning data to electricity and gas prices. Electricity and natural gas city gate prices are selected because their aggregation is easier and the monthly figures are regularly announced by the EIA.

To construct a data set suitable for ANN, the data to be used in this study needs some restrictions. Since sufficient data must be available for learning, this analysis excludes annual data and use monthly data. Additionally, national panel data is not used as this is an empirical study for prediction. This is because in the numerical prediction of global GDP, world average of electricity price, gas price, or consumption is not meaningful. Of course, it is necessary to predict global CO_2 emissions such as [105], but this is an analysis of long-term trends, and the purpose is to find the cause and solution of CO_2 emissions by analyzing the long-term trends rather than improving the accuracy of the forecast. Of course, it should also be differentiated from other panel analysis studies. There is no restriction in the case of panel analysis that is not a prediction.

Therefore, for the empirical analysis, this study uses monthly US electricity and natural gas city gate price data published by EIA. The electricity price data from January 2001 to March 2018 are used along with the natural gas city gate data from January 2001 to December 2016. In the case of the latter, data from up to 2016 are used as the rest of the data is missing. Also, time series data (average data of the US) are analyzed and compared with the predicted results. As mentioned in the introduction, ANN show better predictive accuracy than traditional methods. This analysis compares the results of ANN using panel data with the those using time series data to demonstrate the superiority of panel data on ANN.

4.3 Result and discussion

All the models changed the number of hidden layers to 2 ~ 3 and calculated the results. These models try to improve accuracy by increasing the number of regression lines using two to three hidden layers. In the case of the ANN using panel data, the data set consisting of only state-level data (A data set) and the one consisting of state-level data including the US average data (B data set) are separately learned and the results are calculated.

After the ANN structure, i.e., the number of hidden layers, hidden nodes, and the time lag in this study is set up by using an empirical method, the RMSE of the test set is compared and the model is selected [106]. Then, this study forecasts with the selected model and compare results with the actual data that was previously removed from learning.

All models use the most populous Levenberg-Marquardt backpropagation training function. The Bayesian algorithm is also used as an optimization function, but the forecasting accuracy is estimated to be significantly lower. The results are shown in the following Tables 4.1-4.8 and Figure. 7-8.

Table 4.1. The results of electricity price forecast in 6 month

	US average time series data			State level panel data				Previous panel ANN
				Data set A[a]	**Data set B[b]**	**Data set A[a]**	**Data set B[b]**	
Number of Hidden layers	1	2	3	2	2	3	3	2
Number of Hidden nodes	2	2	2	3	2	3	2	2
Time Lag	9	11	8	7	9	10	11	1
RMSE	0.2042	0.0713*	0.2042	0.0759	0.2399	0.1944	0.1896	0.1273
MAPE	1.6015	0.5599	1.5176	0.6291	1.8573	1.6300	1.5935	0.9518

[a] Panel data set consisting of only state level data
[b] Panel data set by states, including US average data
* Best performance

Table 4.2. The results of electricity price forecast in 12 month

	US average time series data			State level panel data				Previous panel ANN
				Data set A[a]	Data set B[b]	Data set A[a]	Data set B[b]	
Number of Hidden layers	1	2	3	2	2	3	3	2
Number of Hidden nodes	3	2	2	2	2	3	2	2
Time Lag	12	11	10	10	6	8	9	1
RMSE	0.1238	0.1001*	0.2213	0.5987	0.1762	0.2118	0.3664	0.3033
MAPE	0.9640	0.8196	1.7729	5.0980	1.4535	1.7660	3.0552	2.1481

[a] Panel data set consisting of only state level data
[b] Panel data set by states, including US average data
* Best performance

Table 4.3. The results of electricity price forecast in 18 month

	US average time series data			State level panel data				Previous panel ANN
				Data set A[a]	Data set B[b]	Data set A[a]	Data set B[b]	
Number of Hidden layers	1	2	3	2	2	3	3	2
Number of Hidden nodes	3	3	2	3	2	2	3	3
Time Lag	12	10	11	4	8	10	11	1
RMSE	0.2701	0.3364	0.2578	0.3405	0.4282	0.2703	0.1850*	0.5395
MAPE	2.3644	2.7000	2.1772	2.8889	3.4557	2.2684	1.3222	4.0501

[a] Panel data set consisting of only state level data
[b] Panel data set by states, including US average data
* Best performance

Table 4.4. The results of electricity price forecast in 24 month

	US average time series data			State level panel data				Previous panel ANN
				Data set A[a]	**Data set B[b]**	**Data set A[a]**	**Data set B[b]**	
Number of Hidden layers	1	2	3	2	2	3	3	2
Number of Hidden nodes	2	2	3	3	3	2	3	2
Time Lag	9	7	9	10	10	10	7	1
RMSE	0.3676	0.3417	0.4148	0.3154	0.3091	0.3707	0.2258*	0.4607
MAPE	2.8458	2.7191	3.4387	2.4132	2.1391	2.9149	1.5235	3.3204

[a] Panel data set consisting of only state level data
[b] Panel data set by states, including US average data
* Best performance

Table 4.5. The results of natural gas citygate price forecast in 6 month

	US average time series data			State level panel data			
				Data set A[a]	Data set B[b]	Data set A[a]	Data set B[b]
Number of Hidden layers	1	2	3	2	2	3	3
Number of Hidden nodes	2	2	2	3	3	3	3
Time Lag	5	2	5	11	7	11	3
RMSE	0.2610*	0.4729	0.4166	0.7436	0.4618	0.6701	1.2513
MAPE	5.3725	9.8035	7.6065	12.8871	10.1927	15.3618	27.7857

[a] Panel data set consisting of only state level data

[b] Panel data set by states, including US average data

* Best performance

Table 4.6. The results of natural gas citygate price forecast in 12 month

	US average time series data			State level panel data			
				Data set A[a]	Data set B[b]	Data set A[a]	Data set B[b]
Number of Hidden layers	1	2	3	2	2	3	3
Number of Hidden nodes	2	3	2	2	2	2	3
Time Lag	10	6	2	6	3	3	6
RMSE	0.9894	0.5974	0.5317	0.4031*	1.7088	0.6236	0.4074
MAPE	24.2706	14.9306	12.6017	6.9528	25.6600	13.0101	8.8944

[a] Panel data set consisting of only state level data
[b] Panel data set by states, including US average data
* Best performance

Table 4.7. The results of natural gas citygate price forecast in 18 month

	US average time series data			State level panel data			
				Data set A[a]	Data set B[b]	Data set A[a]	Data set B[b]
Number of Hidden layers	1	2	3	2	2	3	3
Number of Hidden nodes	2	3	2	3	3	2	2
Time Lag	3	2	3	6	7	3	2
RMSE	1.1414	1.0692	1.4987	1.0749	2.0114	1.2502	0.9548*
MAPE	27.0446	25.5283	36.2998	25.7125	49.5870	30.4815	22.4587

[a] Panel data set consisting of only state level data
[b] Panel data set by states, including US average data
* Best performance

Table 4.8. The results of natural gas citygate price forecast in 24 month

	US average time series data			State level panel data			
				Data set A[a]	Data set B[b]	Data set A[a]	Data set B[b]
Number of Hidden layers	1	2	3	2	2	3	3
Number of Hidden nodes	3	3	2	3	2	3	3
Time Lag	7	1	8	1	4	1	4
RMSE	1.5893	1.6086	1.5712	1.0179	1.3145	0.6143*	1.5977
MAPE	38.2446	39.2987	37.1558	23.6017	32.4418	11.8672	38.7652

[a] Panel data set consisting of only state level data

[b] Panel data set by states, including US average data

* Best performance

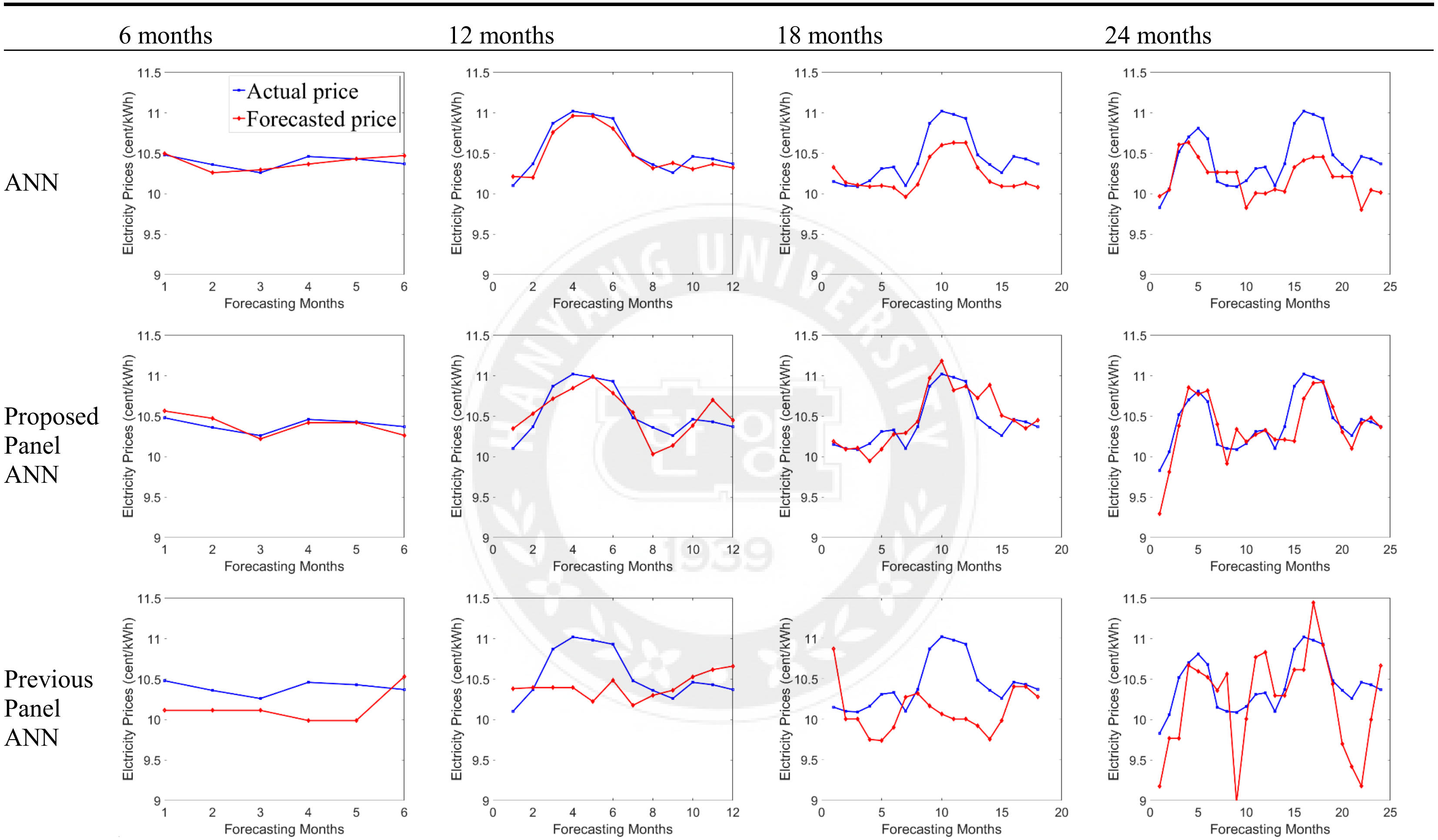


Figure 4.7. The results of electricity price forecast

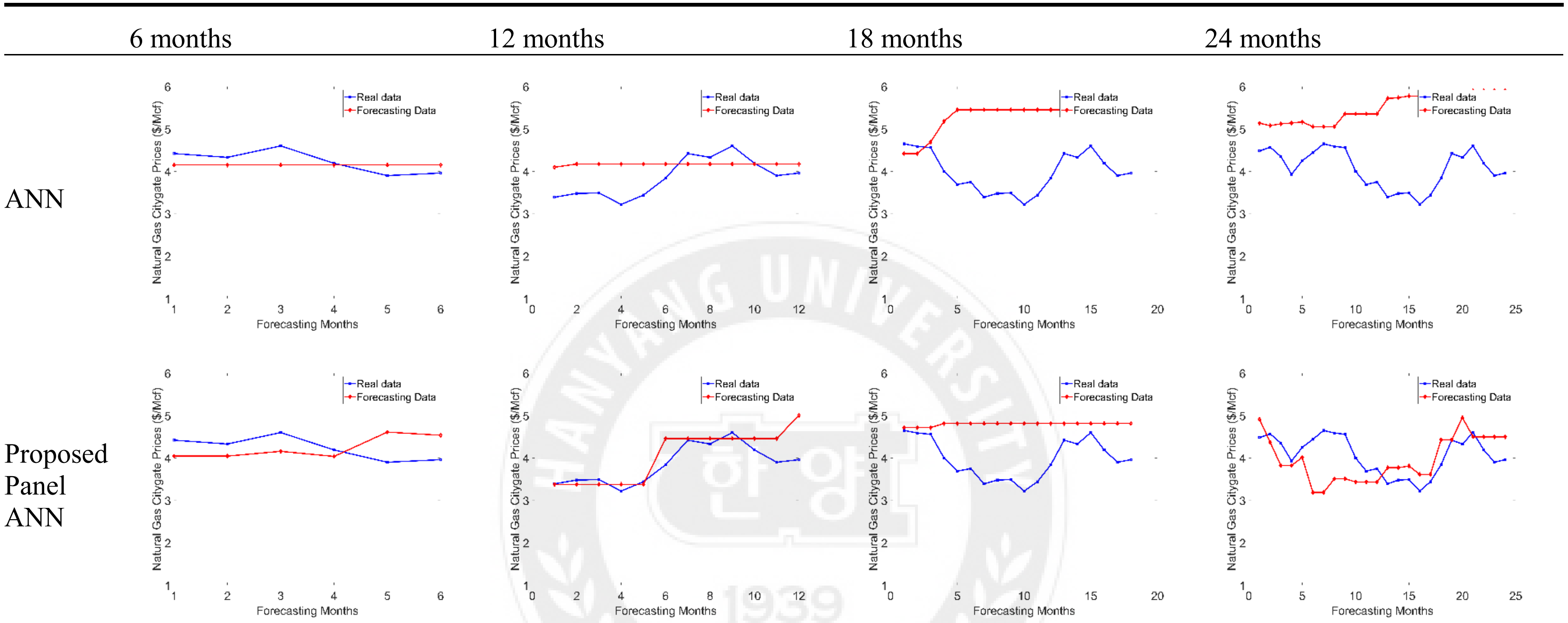


Figure 4.8. The results of natural gas citygate price forecast

In case of the electricity price, the accuracy of the result using time series data in 6 months and 12 months forecast is higher than using panel data. Of the learning models which showed the best results, MAPE is calculated as 0.5599 and 0.6291 respectively for the time series and panel data models in the 6-month forecast analysis. In the 12 months forecast analysis, it is 0.8196 and 1.4535, respectively. On the other hand, the results of 18 and 24 months show that the result of panel data are much better. MAPE for the panel data models is 1.3222 and 1.5235, while in the time series data model, it is 2.1772 and 2.7191. As the forecasting period increases, it is reasonable that MAPE and RMSE will increase. Therefore, rather than comparing the MAPE of each model, we need to focus on the 18 and 24 months estimates in a model using panel data. In case of the previous panel ANN method, it can be confirmed that accuracy is low in all the results.

Natural gas city gate price forecasting is also carried out for additional model verification. In the case of natural gas city gate prices, the results of the model using time series data for only 6 months predictions are better while other predictions show that the panel data model has high accuracy. MAPE is 5.3725 in 6 months forecasting model, and 6.9528, 22.4587 and 11.8672 respectively for the other three models. A noteworthy point is that panel data models tend to be more accurate as the forecast period increases (i.e., the number of predictions increases). Although the timing of improvement in accuracy differs (electricity prices start at 18 months, natural gas city gate prices start at 12 months), both models show an improvement of the panel data forecasting model in long-term predictions.

According to the results, when estimating a small number of predicted values, the trend of the time-series data greatly influences the result and a time-series model produces better predictions. On the other hand, longer the forecast period, better the panel data model that learns from unobserved heterogeneity of the states rather than from the trends. Since weights are updated without affecting each layer, it can be said that the model learns by considering the heterogeneity of each state. In other words, although different learning is performed for each layer, the models share the output layer weight when updated through back propagation because they have the same out-layer and target values. In comparison to a time series model in which only the trend is learned, the panel data model utilizes more information to improve accuracy by learning the trends and heterogeneity of each state.

ANN is an accurate method in the field of forecasting. The results of this empirical study are also superior in smaller number of forecasting. In particular, results of the 6 months forecast show highly accurate results for both electricity and natural gas city gate price. On the other hand, in the case of ANN using panel data, prediction accuracy is lower than that of time series data in smaller number of forecasting, but this improves in larger number of forecasting. This can be seen as an improvement in accuracy by applying heterogeneity of each state in network learning using panel data. In both the empirical studies, panel data results are better than time series results in larger number of forecasting. Therefore, in case of long-term forecasting (i.e., when forecasting period increases), building a model with panel ANN structure as proposed in this study can improve accuracy.

Chapter 5. Conclusions

This study investigates the applicability of the keyword search volume to panel analysis and develops a new model of ANN using panel data. For model validation of panel model and ANN model, this study conducts demand analysis and price forecasting of the electricity market, which is considered as the most important commodity of daily life in recent years. In particular, the analysis focuses on the US electricity market, which is the largest in the world.

In the first essay, this study conducts the panel analysis using one of the Internet search terms, Google Trends, and confirmed the correlation with electricity consumption. Econometric studies using keyword search volume have been analyzed only in terms of the behaviors of households such as consumption activity in labor and housing markets. In particular, since there is no case of using Internet search words in electricity demand model analysis, a new model based on electricity static demand model is proposed.

As a result of the electricity demand panel analysis, I found that the coefficient of the “renewable” variable is statistically significant and that the “temperature” variable is also significantly correlated with residential electricity demand. The “renewable” keyword has a large negative correlation with household electricity consumption, which can be estimated as a result of the growing interest in renewable energy. Although the electricity consumption patterns of households are influenced by many variables, this study suggests that interest in renewable energy should also be included as a major factor influencing electricity consumption. Taken together, our research shows that as searches for “renewable” increases and interest rises, electricity consumption tends to be replaced by renewable energy, thereby reducing total household electricity consumption.

The significance of this study can be divided into two. It takes about a quarter or a year for the official announcement of the electricity consumption statistics, as well as the prices and income statistics that can explain them; this has rarely been mentioned in traditional econometrics. In terms of this issue, the keyword search volume demand model proposed in this study can be useful for predicting the present. As another implication, this study shows that variables that have not been used hitherto, as they are not quantifiable or statistically significant (e.g., interest), can be analyzed through keyword search volume. That is, by using variables that cannot be quantified, the current situation can be predicted and analyzed. In addition, it is easy to solve the issue of data collection, which is the biggest disadvantage of panel analysis. This study indicates that the research can be expanded through keyword search volume.

Undoubtedly, a clearer search keyword could have been used; however, there are limitations on the data provided by state in keyword search volume. If the use of Google search and the cumulative period is increased, it may be possible to adopt clearer and more diverse search keywords for analysis. Further analysis of electricity consumption or expenditure through more diverse keyword search volume may be considered for further study.

In the second essay, ANN, which is a type of machine learning, is proposed as a prediction model suitable for panel data. Furthermore, in the case of the ANN using the panel data, only the analysis on the quantitative aspect of the panel data is performed, and there is no model utilizing the characteristics of the panel data. The main objective of this study is to develop a model that utilizes panel data to increase quantitative aspects of learning data and utilize its characteristics. Panel data includes cross-sectional and time-series characteristics and implies unobserved heterogeneity. Using this characteristic, a new ANN prediction model was constructed and analyzed. Specifically, in the long-term prediction, it is an appropriate method to learn the ANN using the panel data.

This section outlines the proposal for building a panel ANN model. In the hidden layer configuration, it is reasonable to calculate the best result by increasing the number of hidden layers from 1 to 3. From the results, models with multiple hidden layers are selected over models with one hidden layer. Meanwhile, the use of more than 4 hidden layers requires deep learning, which is not recommended with a small volume of data, such as economics data. Deep learning is not necessary when you are looking at learning time and accuracy. Moreover, it cannot build big data needed for deep learning.

In the case of machine learning predictions such as ANN, since the prediction model cannot learn the information about the time intervals of the learning data, it should be considered from the modeling. If the input data and the output data have the same period, it can be said that short-term prediction is performed when the learning data is time-based data and long-term prediction is performed when the data is yearly data. This should be kept in mind while interpreting the results.

Panel ANN is a model that can be applied from day-to-day and hourly forecasts to long-term trends of several years depending on the type of panel data. In analyzing the long-term trends, a neural network model that can replace the large-scale simulation models such as NEMS and WEM can also be constructed. Therefore, this model can be applied in various fields ranging from the hourly price forecast of the next day's electricity market to the long-term trend of CO_2 emissions. Also, if the privatization of electricity is progressed in domestic application, it will be applicable model.

BIBLIOGRAPHY

[1] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable and Sustainable Energy Reviews. 2009;13:1819-35.

[2] Choi H, Varian HAL. Predicting the Present with Google Trends. Economic Record. 2012;88:2-9.

[3] Yamin HY, Shahidehpour SM, Li Z. Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets. International Journal of Electrical Power & Energy Systems. 2004;26:571-81.

[4] Mandal P, Senjyu T, Funabashi T. Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market. Energy Conversion and Management. 2006;47:2128-42.

[5] Park SJ, Kim JS. An Application of Artificial Neural Network in Short-run Natural Gas Price Forecasting. The Korean Society of Mineral and Energy Resources Engineers. 2014;51:761-70.

[6] Park SJ, Kim JS. An Application of Grey Neural Network for Forecasting Short-term Natural Gas Consumption of Korea. The Korean Society of Mineral and Energy Resources Engineers. 2016;53:78-87.

[7] Lin W-M, Gow H-J, Tsai M-T. An enhanced radial basis function network for short-term electricity price forecasting. Applied Energy. 2010;87:3226-34.

[8] Wang D, Luo H, Grunder O, Lin Y, Guo H. Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm. Applied Energy. 2017;190:390-407.

[9] Druce DJ. Modelling the transition from cost-based to bid-based pricing in a deregulated electricity-market. Applied Energy. 2007;84:1210-25.

[10] Walawalkar R, Blumsack S, Apt J, Fernands S. An economic welfare analysis of demand response in the PJM electricity market. Energy Policy. 2008;36:3692-702.

[11] Upton J, Murphy M, Shalloo L, Groot Koerkamp PWG, De Boer IJM. Assessing the impact of changes in the electricity price structure on dairy farm energy costs. Applied Energy. 2015;137:1-8.

[12] Hung M-F, Huang T-H. Dynamic demand for residential electricity in Taiwan under seasonality and increasing-block pricing. Energy Economics. 2015;48:168-77.

[13] Sun M, Li J, Gao C, Han D. Identifying regime shifts in the US electricity market based on price fluctuations. Applied Energy. 2017;194:658-66.

[14] Hekkenberg M, Benders RMJ, Moll HC, Schoot Uiterkamp AJM. Indications for a changing electricity demand pattern: The temperature dependence of electricity demand in the Netherlands. Energy Policy. 2009;37:1542-51.

[15] Bessec M, Fouquau J. The non-linear link between electricity consumption and temperature in Europe: A threshold panel approach. Energy Economics. 2008;30:2705-21.

[16] Paul AC, Myers EC, Palmer KL. A partial adjustment model of US electricity demand by region, season, and sector. 2009.

[17] Alberini A, Gans W, Velez-Lopez D. Residential consumption of gas and electricity in the U.S.: The role of prices and income. Energy Economics. 2011;33:870-81.

[18] Alberini A, Filippini M. Response of residential electricity demand to price: The effect of measurement error. Energy Economics. 2011;33:889-95.

[19] Sun Y. Electricity Prices, Income and Residential Electricity Consumption. 2015.

[20] Salari M, Javid RJ. Residential energy demand in the United States: Analysis using static and dynamic approaches. Energy Policy. 2016;98:637-49.

[21] Filippini M. Short- and long-run time-of-use price elasticities in Swiss residential electricity demand. Energy Policy. 2011;39:5811-7.

[22] Azevedo IML, Morgan MG, Lave L. Residential and Regional Electricity Consumption in the U.S. and EU: How Much Will Higher Prices Reduce CO2 Emissions? The Electricity Journal. 2011;24:21-9.

[23] Wiesmann D, Lima Azevedo I, Ferrão P, Fernández JE. Residential electricity consumption in Portugal: Findings from top-down and bottom-up models. Energy Policy. 2011;39:2772-9.

[24] Smith P. Google's MIDAS Touch: Predicting UK Unemployment with Internet Search Data. Journal of Forecasting. 2016;35:263-84.

[25] Askitas N, Zimmermann KF. Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly. 2009;55:107-20.

[26] D'Amuri F, Marcucci J. 'Google it!'Forecasting the US Unemployment Rate with a Google Job Search Index. FEEM Working Paper. 2010;31.

[27] Fondeur Y, Karamé F. Can Google data help predict French youth unemployment? Economic Modelling. 2013;30:117-25.

[28] Cooper PC, Mallon PK, Leadbetter S, Pollack AL, Peipins AL. Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003. J Med Internet Res. 2005;7:e36.

[29] Eysenbach G. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. AMIA Annual Symposium Proceedings. 2006;2006:244-8.

[30] Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. Clinical Infectious Diseases. 2008;47:1443-8.

[31] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2008;457:1012.

[32] Doornik JA. Improving the timeliness of data on influenza-like illnesses using Google search data. Working paper. 2009.

[33] Hulth A, Rydevik G, Linde A. Web Queries as a Source for Syndromic Surveillance. PLOS ONE. 2009;4:e4378.

[34] McCarthy MJ. Internet monitoring of suicide risk in the population. Journal of Affective Disorders. 2010;122:277-9.

[35] Gunn III JF, Lester D. Using google searches on the internet to monitor suicidal behavior. Journal of Affective Disorders. 2013;148:411-2.

[36] Sueki H. Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: Data from 2004 to 2009 in Japan. Psychiatry and Clinical Neurosciences. 2011;65:392-4.

[37] Park S, Kim J. The effect of interest in renewable energy on US household electricity consumption: An analysis using Google Trends data. Renewable Energy. 2018;127:1004-10.

[38] Salahuddin M, Alam K, Ozturk I. The effects of Internet usage and economic growth on CO2 emissions in OECD countries: A panel investigation. Renewable and Sustainable Energy Reviews. 2016;62:1226-35.

[39] Wang AJ, Ramsay B. A neural network based estimator for electricity spot-pricing with particular reference to weekend and public holidays. Neurocomputing. 1998;23:47-57.

[40] Yao SJ, Song YH, Zhang LZ, Cheng XY. Prediction of System Marginal Prices by Wavelet Transform and Neural Networks. Electric Machines and Power Systems. 2000;28:983-93.

[41] Li Z, Luh PB, Kasiviswanathan K. Energy clearing price prediction and confidence interval estimation with cascaded neural networks. IEEE Transactions on Power Systems. 2003;18:99-105.

[42] Jau-Jia G, Luh PB. Improving market clearing price prediction by using a committee machine of neural networks. IEEE Transactions on Power Systems. 2004;19:1867-76.

[43] Rodriguez CP, Anders GJ. Energy price forecasting in the Ontario competitive power system market. IEEE Transactions on Power Systems. 2004;19:366-74.

[44] Gonzalez AM, Roque AMS, Garcia-Gonzalez J. Modeling and forecasting electricity prices with input/output hidden Markov models. IEEE Transactions on Power Systems. 2005;20:13-24.

[45] Lee JK, Shin JR, Park JB. A system marginal price forecasting method based on an artificial neural network using time and day information. The Transactions of the Korean Institute of Electrical Engineers A. 2005;54:144-51.

[46] Gareta R, Romeo LM, Gil A. Forecasting of electricity prices with neural networks. Energy Conversion and Management. 2006;47:1770-8.

[47] Georgilakis PS. Market Clearing Price Forecasting in Deregulated Electricity Markets Using Adaptively Trained Neural Networks. In: Antoniou G, Potamias G, Spyropoulos C, Plexousakis D, editors. Hellenic Conference on Artificial Intelligence. Heraklion, Crete, Greece: Springer; 2006. p. 56-66.

[48] Hu Z, Yang L, Wang Z, Gan D, Sun W, Wang K. A game-theoretic model for electricity markets with tight capacity constraints. International Journal of Electrical Power & Energy Systems. 2008;30:207-15.

[49] Areekul P, Senju T, Toyama H, Chakraborty S, Yona A, Urasaki N, et al. A New Method for Next-Day Price Forecasting for PJM Electricity Market. International

Journal of Emerging Electric Power Systems. 2010;11:Article 3.

[50] Zhang L, Luh PB. Neural network-based market clearing price prediction and confidence interval estimation with an improved extended Kalman filter method. IEEE Transactions on Power Systems. 2005;20:59-66.

[51] Pao H-T. A Neural Network Approach to m-Daily-Ahead Electricity Price Prediction. In: Wang J, Yi Z, Zurada JM, Lu B-L, Yin H, editors. Advances in Neural Networks - ISNN 2006: Third International Symposium on Neural Networks, Chengdu, China, May 28 - June 1, 2006, Proceedings, Part II. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 1284-9.

[52] Szkuta BR, Sanabria LA, Dillon TS. Electricity price short-term forecasting using artificial neural networks. IEEE Transactions on Power Systems. 1999;14:851-7.

[53] Wang A, Ramsay B. Prediction of system marginal price in the UK Power Pool using neural networks. Neural Networks, 1997, International Conference on: IEEE; 1997. p. 2116-20.

[54] Gao F, Guan X, Cao X-R, Papalexopoulos A. Forecasting power market clearing price and quantity using a neural network method. 2000 Power Engineering Society Summer Meeting. Seattle, USA: IEEE; 2000.

[55] Hu Z, Yu Y, Wang Z, Sun W, Gan D, Han Z. Price forecasting using an integrated approach. Electric Utility Deregulation, Restructuring and Power Technologies, 2004(DRPT 2004) Proceedings of the 2004 IEEE International Conference on: IEEE; 2004. p. 28-31.

[56] Lora AT, Santos JR, Santos JR, Ramos JLM, Exposito AG. Electricity Market Price Forecasting: Neural Networks versus Weighted-Distance k Nearest Neighbours. In: Hameurlain A, Cicchetti R, Traunmüller R, editors. Database and Expert Systems Applications: 13th International Conference, DEXA 2002 Aix-en-Provence, France, September 2–6, 2002 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002. p. 321-30.

[57] Jin T, Kim J. What is better for mitigating carbon emissions – Renewable energy or nuclear energy? A panel data analysis. Renewable and Sustainable Energy Reviews. 2018;91:464-71.

[58] Solon G, Corcoran M, Gordon R, Laren D. A Longitudinal Analysis of Sibling Correlations in Economic Status. The Journal of Human Resources. 1991;26:509-34.

[59] Kim B, Gibson J, Chung C. Using panel data to exactly estimate under-reporting by the self-employed. Department of Economics Working Paper Series 2008.

[60] Levin A, Lin C-F, James Chu C-S. Unit root tests in panel data: asymptotic and finite-sample properties. Journal of Econometrics. 2002;108:1-24.

[61] Harris RDF, Tzavalis E. Inference for unit roots in dynamic panels where the time dimension is fixed. Journal of Econometrics. 1999;91:201-26.

[62] Hadri K. Testing for stationarity in heterogeneous panel data. Econometrics Journal. 2000;3:148-61.

[63] Im KS, Pesaran MH, Shin Y. Testing for unit roots in heterogeneous panels. Journal of

Econometrics. 2003;115:53-74.

[64] Maddala GS, Wu S. A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. Oxford Bulletin of Economics and Statistics. 1999;61:631-52.

[65] Choi I. Unit root tests for panel data. Journal of International Money and Finance. 2001;20:249-72.

[66] Breitung J. The local power of some unit root tests for panel data. Nonstationary panels, panel cointegration, and dynamic panels: Emerald Group Publishing Limited; 2001. p. 161-77.

[67] Kim, Yundae, Jun, Chi H. A New Approach to Unit Root Testing of Panel Data. Korean Institute Of Industrial Engineers; 2010. p. 1462-9.

[68] Hlouskova J, Wagner M. The Performance of Panel Unit Root and Stationarity Tests: Results from a Large Scale Simulation Study. Econometric Reviews. 2006;25:85-116.

[69] Allison PD. Fixed effects regression models: SAGE publications; 2009.

[70] Hausman JA. Specification Tests in Econometrics. Econometrica. 1978;46:1251-71.

[71] Pan B, Yang Y. Forecasting Destination Weekly Hotel Occupancy with Big Data. Journal of Travel Research. 2017;56:957-70.

[72] Sciascia S, Radin M. What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus. International Journal of Medical Informatics. 2017;107:65-9.

[73] Jun S-P, Yoo HS, Choi S. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change. 2018;130:69-87.

[74] Naccarato A, Falorsi S, Loriga S, Pierini A. Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technological Forecasting and Social Change. 2018;130:114-22.

[75] Yu L, Zhao Y, Tang L, Yang Z. Online big data-driven oil consumption forecasting with Google trends. International Journal of Forecasting. 2018.

[76] Banbura M, Giannone D, Reichlin L. Nowcasting with daily data. European Central Bank, Working Paper. 2011.

[77] Li G, Shi J. On comparing three artificial neural networks for wind speed forecasting. Applied Energy. 2010;87:2313-20.

[78] Kim DS. Neural Networks: Theory and Application. Seoul: Jinhan M&B; 2005.

[79] Şenkal O, Kuleli T. Estimation of solar radiation over Turkey using artificial neural network and satellite data. Applied Energy. 2009;86:1222-8.

[80] Khosravi A, Nahavandi S, Creighton D. Quantifying uncertainties of neural network-based electricity price forecasts. Applied Energy. 2013;112:120-9.

[81] Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. arXiv preprint arXiv:170802709. 2017.

[82] Sarle WS. How to measure the importance of inputs? Technical Report, SAS Institute

Inc. 1998.

[83] Tang L, Wang S, He K, Wang S. A novel mode-characteristic-based decomposition ensemble model for nuclear energy consumption forecasting. Annals of Operations Research. 2015;234:111-32.

[84] Zeng Y-R, Zeng Y, Choi B, Wang L. Multifactor-influenced energy consumption forecasting using enhanced back-propagation neural network. Energy. 2017;127:381-96.

[85] Beccali M, Ciulla G, Lo Brano V, Galatioto A, Bonomolo M. Artificial neural network decision support tool for assessment of the energy performance and the refurbishment actions for the non-residential building stock in Southern Italy. Energy. 2017;137:1201-18.

[86] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 1943;5:115-33.

[87] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958;65:386.

[88] Rumelhart DE, McClelland JL. Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations. 1986.

[89] Jin J, Kim J. Forecasting Natural Gas Prices Using Wavelets, Time Series, and Artificial Neural Networks. PLOS ONE. 2015;10:e0142064.

[90] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521:436.

[91] Salari M, Javid RJ. Modeling household energy expenditure in the United States. Renewable and Sustainable Energy Reviews. 2017;69:822-32.

[92] Mulder M, Scholtens B. The impact of renewable energy on electricity prices in the Netherlands. Renewable Energy. 2013;57:94-100.

[93] Zhou K, Yang S. Understanding household energy consumption behavior: The contribution of energy big data analytics. Renewable and Sustainable Energy Reviews. 2016;56:810-9.

[94] Wang Z, Lu M, Wang J-C. Direct rebound effect on urban residential electricity use: An empirical study in China. Renewable and Sustainable Energy Reviews. 2014;30:124-32.

[95] Rouholamini M, Mohammadian M. Energy management of a grid-tied residential-scale hybrid renewable generation system incorporating fuel cell and electrolyzer. Energy and Buildings. 2015;102:406-16.

[96] Rouholamini M, Mohammadian M. Heuristic-based power management of a grid-connected hybrid energy system combined with hydrogen storage. Renewable Energy. 2016;96:354-65.

[97] Soares A, Gomes Á, Antunes CH. Categorization of residential electricity consumption as a basis for the assessment of the impacts of demand response actions. Renewable and Sustainable Energy Reviews. 2014;30:490-503.

[98] Yoo W, Mayberry R, Bae S, Singh K, He Q, Lillard J. A Study of Effects of MultiCollinearity in the Multivariable Analysis2014.

[99] Pao H-T, Chih Y-Y. Comparison of TSCS regression and neural network models for panel data forecasting: debt policy. Neural Computing & Applications. 2006;15:117-23.

[100] Al Shami A, Lotfi A, Lai E, Coleman S. Forecasting Macro-Knowledge Competitiveness; Integrating Panel Data and Computational Intelligence. Nottingham Trent University, Nottingham, United Kingdom. 2011.

[101] Crane-Droesch A. Semiparametric panel data models using neural networks. arXiv preprint arXiv:170206512. 2017.

[102] Baltagi B. Econometric analysis of panel data: John Wiley & Sons; 2008.

[103] Demuth HB, Beale MH, De Jess O, Hagan MT. Neural network design: Martin Hagan; 2014.

[104] Hansen LK, Salamon P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1990;12:993-1001.

[105] Schmalensee R, Stoker TM, Judson RA. World Carbon Dioxide Emissions: 1950–2050. The Review of Economics and Statistics. 1998;80:15-27.

[106] Stathakis D. How many hidden layers and nodes? International Journal of Remote Sensing. 2009;30:2133-47.

Abstract in Korean

빅데이터 분석과 기계 학습은 데이터 분석에 있어서 주요 분석 도구이다. 빅데이터는 분야 별 데이터 분석을 위해 엄청난 양의 원데이터를 수집하고 유지 관리하는 영역이며, 기계 학습은 이런 빅데이터를 처리하기 위한 주요 분석 도구이다. 본 연구는 빅데이터 중 하나인 키워드 검색량의 패널 분석에 있어서 적용 가능성을 알아보고, 패널 데이터를 사용한 인공신경망(Artificial Neural Network, ANN) 모형을 개발하여 전력 소비 분석과 가격 예측을 시행하고자 한다. 계량경제학 분야, 특히 에너지 경제학 분야에서 키워드 검색량을 사용한 분석 사례는 전무하며, 따라서 새로운 전력 수요 모형을 구축하고자 한다. 또한 패널 데이터를 적용시킨 ANN 모형의 경우, 모형 구축 사례가 없기 때문에 본 연구를 통해 새로운 패널 ANN 모형을 구축하였다. 본 연구는 패널 분석 모형 개발과 패널 ANN 모형 개발 두 개의 연구로 구성된다.

먼저 키워드 검색량과 미국 주거용 전력 소비와의 상관 관계를 분석하여 키워드 검색량의 활용성을 알아보고자 한다. 따라서 전력 소비와 관련이 있는 키워드를 고려하여 “Renewable”과 “Weather forecast”, “Temperature”를 인터넷 검색어 키워드로 설정하였다. 그 동안, 주거용 재생 에너지 소비량을 수치화 할 수 있는 방법이 없었기 때문에, 미국의 재생 에너지와 주거용 전력 소비량 간의 상관 관계에 대한 연구는 전무한 실정이다. 주거용 전력 소비는 수준을 예측하기가 어려우며, 개인 정보 수집에 따른 문제, 측정 비용 문제로 인해 상업, 산업 같은 다른 주요 부문에 비해 파악이 힘들다. 따라서 인터넷 검색어를 사용하여 재생 에너지에 대한 관심을 포함한 모형을 구성하여 주거용 전력 소비와의 상관 관계를 분석하고자 한다.

패널 분석을 위해 에너지 수요 모형을 변형하여 모형을 구축하였으며, 3가지 키워드 검색량 키워드에 따른 모형을 분석하여 비교하였다. 모든 변수의 단위근 검정 결과 모두 안정적으로 산출되었으며, 하우스만 검정 하에 고정 효과 모형을 선택하여 분석하였다. “Renewable” 키워드를 변수에 적용한 모형의 경우, 가격 변수를 제외한 모든 변수가 1% 유의수준 하에서

통계적으로 유의미한 것으로 나타났다. 키워드 검색량 “renewable” 변수는 음의 상관관계를 가지며 키워드 검색량이 한 단계 증가할 때 전력 소비량이 16.017 million kWh 감소하는 것으로 나타났다. “Temperature” 키워드 모형의 경우, “Renewable”과 마찬가지로 음의 상관관계를 가지나 HDD 변수와 비슷한 영향을 미치는 것을 볼 수 있다.

직관적으로 아무런 관계가 없을 것 같은 두 변수의 상관관계가 크게 산출된 것은 분명한 의미를 지니고 있을 것이다. 분명 가정에서 “Renewable”을 검색할 때는 분명한 목적을 가지고 있기 마련이다. 전력 소비가 과도하게 이루어 졌거나 요금이 과도하게 부과되었을 때, 가정에서는 전력 소비를 줄이기 위한 대안을 모색할 것이다. 물론 신 재생 기기를 설치한 가정이 많지 않겠지만, 신 재생 기기를 시행한 가정의 경우 그만큼 전력 소비가 감소할 것이며 그 추정 결과가 나타난 것이라 볼 수 있다.

3가지 키워드 검색량을 변수로 설정하며 진행한 결과, “Renewable” 변수의 상관계수가 가장 높게 산출된 것을 볼 수 있으며, “Temperature” 변수도 주거용 전력 소비와 유의미한 상관관계를 가지는 것으로 나타났다. “Renewable” 키워드는 주거용 전력 소비량과 큰 음의 상관관계를 가지며, 이는 최근 증가하고 있는 재생에너지에 대한 관심도에 따른 결과로 추정해 볼 수 있다. 가정의 소비 형태는 많은 변수에 영향을 받지만, 소비 변화에 영향을 주는 주요 인자로 관심도 또한 포함되어야 함을 시사한다.

다음으로는 인공신경망(Artificial Neural Network, ANN)을 이용하여 전력 가격 예측을 시행하였다. ANN은 다양한 분야에서 예측을 위한 툴로 사용되고 있다. 경제학 분석 연구에서는 일반적으로 단기 예측을 위해 사용되어 왔다. 반면, 예측 시점이 증가하게 되면 예측 정확도는 급격하게 하락하게 된다. 같은 데이터 셋 하에서는 장기 예측 시 예측 정확도는 단기 예측 보다 커지기 마련이다. 따라서 본 연구는 같은 데이터 셋 하에서 예측치 개수 증가로 인한 예측 정확도의 하락을 보완하고자 한다.

패널 데이터는 기존 시계열 데이터가 가지고 있지 못한 정보를 포함하고 있다. 시계열 데이터의 추세 정보는 물론이고, 주 혹은 국가 별 특성 또한

가지고 있다. 하지만 그 동안 경제학 분야에서 패널 데이터를 사용하여 ANN으로 예측한 연구는 거의 없다. 기존 연구들도 전부 패널 데이터를 사용할 뿐 국가 혹은 지역을 모형 내에서 구분해주지 않는다. 이는 패널 데이터의 장점을 활용하지 못한 것으로 볼 수 있다. 기존 패널 ANN 연구들은 Pooled OLS 방식과 같이 지역 별, 혹은 국가 별 자료를 구분하지 않거나, 독립적인 학습이 이루어지지 않은 모형을 구축하였다. 따라서 본 연구에서는 패널 데이터가 가지고 있는 특성을 활용하여 패널 ANN 모형을 구축하고 예측 개수의 변화에 따른 정확도를 분석 하고자 한다. 패널 데이터의 각 지역 및 국가가 가지고 있는 특성을 학습시켜 추정된 예측치의 정확도 향상을 도모하고자 한다. 패널 데이터는 관측 불가능한 특성을 포함하고 있으며, 때문에 시계열 정보뿐 아니라 각 국가 혹은 지역 별 정보도 학습할 수 있을 것이란 가정 하에 분석을 진행하였다.

패널 분석은 패널 데이터가 가지고 있는 관찰 불가능한 특성을 고려하여 횡단면 의존성을 제거하게 된다. 반면 본 장의 ANN 모형은 패널 데이터가 가지고 있는 이 관찰 불가능한 특성을 학습시키기 위한 모형으로 설계하였다. 각 주 별로 학습을 분리시켜 시행하였으며, 두 개 혹은 세 개의 은닉층이 주 별로 위치하였다. 모형 학습 후, 6, 12, 18, 24개월을 예측한 뒤 전체 RMSE와 MAPE를 구하여 최적 모형을 산출하였다.

모형 검증을 위해, 미국의 전력 가격의 주 별 패널 데이터를 사용하였다. 또한 추가적인 모형 검증을 위해 천연 가스 가격의 예측도 함께 시행하였다. 전력 가격 예측 모형의 경우, 6, 12개월 예측치는 기존 단일 데이터를 사용한 ANN의 정확도가 더 높게 산출되었다. 반면 18, 24개월의 결과는 패널 데이터를 사용한 결과가 좋은 것을 확인할 수 있었다. 가스 가격의 경우에는 6개월 예측치만 단일 데이터를 사용한 모형의 결과가 좋게 산출되었고 그 이후 예측치는 패널 ANN 모형이 보다 높은 정확도를 산출하였다. 주목할 만한 요소는 예측치가 많을수록, 즉, 예측 기간이 길수록 패널 ANN 모형이 더 높은 정확도를 보인다는 것이다. 기간의 차이는 있지만 두 모형 전부 예측 기간이 길어질 수록 정확도 향상이 이루어짐을 보인다.

결론적으로, 예측 기간이 짧을 때는 시계열 데이터의 추세가 결과값에 큰 영향을 미쳐 단일 시계열 모형이 더 좋은 예측값을 산출해 낸다고 볼 수 있다. 반면, 예측 기간이 길어질수록, 추세 보다는 주 별 특징을 학습한 모형이 더 좋은 결과를 가지게 된다. 각 주 별로 가중치가 서로 영향을 끼치지 않고 갱신되기 때문에, 주 별 특성이 고려된 모형으로 학습되었다고 볼 수 있다. 이로 인해 패널 데이터의 각 주 별 추세나 특성을 학습하여, 더 많은 정보를 활용 가능하게 되고, 추세만을 학습한 단일 시계열 모형보다 정확도 향상을 도모할 수 있게 된다.

본 연구는 패널 데이터를 활용하여 전력 소비량을 분석하고 전력 가격 예측을 시행하였다. 전력 소비량의 패널 분석은 주거용 전력 소비 문헌에서 고려된 모형을 기반으로 키워드 검색량 자료와 접목하여 새로운 모형을 제시하고 있다. 신 재생 설비 설치 용량 뿐 아니라 재생에너지에 대한 관심이 전력 소비량에 영향을 미침에도 불구하고 정량적인 데이터가 없어 설명할 수 없었던 부분을 본 연구에서는 키워드 검색량을 대체 변수로 사용하여 분석하였다. 결론적으로 이 연구는 지금까지 사용되지 않았지만 계량화가 불가능하거나 통계적으로 유의미한 변수가 키워드 검색량을 통해 분석 될 수 있음을 보여준다.

전력 가격 예측 분석에서는 예측치의 개수가 증가할 때 예측 정확도가 하락하는 점을 보완 해주기 위한 새로운 패널 ANN모형을 제안하였다. 패널 ANN은 패널 데이터의 구축 여부에 따라 시간 별 예측에서부터, 장기 추세 예측까지 적용할 수 있는 모형이다. 장기 추세 분석의 경우, NEMS(National Energy Modeling System)나 WEM(World Energy Model)같은 대규모 시뮬레이션 모형을 대체할 수 있는 신경망 모형 또한 구축 가능할 것이다. 따라서, 다음날의 전력 시장의 시간 별 가격 예측에서부터 CO_2 배출량의 장기 추세 예측까지 다양한 분야에서 활용 가능할 것이다.

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

JUNE 05, 2019

Degree : Doctor

Department : DEPARTMENT OF EARTH RESOURCES AND ENVIRONMENTAL ENGINEERING

Thesis Supervisor : Kim, Jinsoo

Name : PARK SUNGJUN 박성준 (Signature)

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2019년06월05일

학위명 : 박사

학과 : 자원환경공학과

지도교수 : 김진수

성명 : 박성준 박성준 (서명)

한양대학교대학원장귀하