A regression model estimating the impact of natural, socio-economic regional characteristics on hourly electricity consumption in Japan

by

Yuki HIRUTA, Lu GAO and Shuichi ASHINA,

Yuki HIRUTA, Research Associate
Center for Social and Environmental Systems Research, National Institute for Environmental Studies
16-2 Onogawa, Tsukuba, Ibaraki 305-8506, JAPAN
Phone: 029-850-2227 / Email: hiruta.yuki@nies.go.jp

Abstract
Given that oil thermal power and pumped storage hydroelectric power are mainly used during periods of peak demand for electricity in Japan, fossil fuels tend to be consumed to larger degrees when there are more fluctuations in electricity demand. Global warming, and a declining and aging population in Japan, mean that the measures undertaken to date (such as peak shifts) are insufficient. Comprehensive measures such as land-use planning should be applied concurrently. We need to develop a method to evaluate the effect of multiple concurrent factors, such as regional climate, and the socio-economic and geographical characteristics of the region, on hourly electricity demand. In this research, we propose a series of methods with which to construct a single regression model that explains hourly electricity consumption based on various regional characteristics such as hourly climatic conditions and the socio-economic and geographical conditions of the region. We seek to overcome three technical problems for this purpose. The first is how to consider the complicated combination of region-specific factors and time-varying factors. Given that Japan has a variety of regions, from subarctic to subtropical, as well as four distinct seasons, electricity consumption is affected by multiple time-varying factors and region-specific factors. The second question is how to select an effective variable from a wide range of possibilities. The third is how to model complex relationships such as the nonlinear relationship between power consumption and factors, and at the same time to construct generalizable models using as few parameters as possible. We propose a unique algorithm using RandomForests to narrow down the small number of effective explanatory variables from more than 6,000 possibilities, including variables for region-specific, time-varying and interaction factors between the region-specific factor and time-varying factor. We apply MARS for the modeling, which rendered it possible to construct a flexible yet highly generalizable model while selecting important variables. We constructed a regression model that explains hourly power consumption nationwide. The constructed model exhibits high quality in terms of both fitness and generalization, but it could still be improved. We experimentally applied the simulation to understand the effect of the spatial distribution of households on hourly power demand using the model we constructed. The proposed method may offer a means to assess the comprehensive measures, including land-use policies, to lower the fluctuation of hourly electricity demand, even there is a room for improvement.

1. Introduction

We propose a series of methods with which to construct a single regression model that explains hourly electricity consumption based on various regional characteristics such as hourly climatic conditions and the socio-economic and geographical conditions of the region.

Day-night and seasonal variations in electricity use are increasing, and this trend is expected to continue in Japan[1]. Given that oil thermal power and pumped-storage hydroelectric power are mainly used during periods of peak demand for electricity in Japan, fossil fuels tend to be consumed to larger degrees when there are increased fluctuations in electricity demand [2]. Owing to global warming, as well as a declining and aging population in Japan, the measures undertaken to date (such as peak shifts) are insufficient. Comprehensive measures such as land-use planning and should be studied concurrently. In order to find solutions to reduce the range of fluctuation in power demand, it is first necessary to clarify the factors that affect the fluctuation in power demand and the marginal effect of the factors. Generally, hourly

fluctuations in power consumption are determined by weather conditions such as temperature, humidity, and wind speed, and human activity level such as the proportion of people working at the time, however, it is hypothesized that the magnitude of the hourly fluctuation in power demand, is scaled by the socio-economic conditions in each region, such as population, gross regional products, and land-use. If we aim to understand the factors that affect power demand in each region, we can simply apply the analysis to consider time-varying factors such as hourly weather data and data on human activity in each hour. If we accept the hypothesis above, however, and aim to find the factors that affect absolute hourly power demand, we need to take region-specific factors into consideration, and time-varying factors at the same time. Consideration of the effect of region-specific factors on hourly power demand is indispensable when studying comprehensive measures including socio-economic conditions and land-use policies.

In the past, many types of research have been conducted in order to understand the relationship between climatic conditions and power demand. Non-linear statistical learning methods were recently applied in the USA to model the complex relationship between power demand and climatic conditions, and were reported as performing well [3][4]. The segmented regression technique was also applied to estimate the relationship between hourly electricity demand and meteorological conditions reference temperatures for cooling- and heating-degree hours was determined, and it reported the significant effect of past temperatures and relative humidity on electricity load [5]. Although there are few studies estimating the scale dependency of the impact of climate difference on energy demand due to region-specific variables, Amato et.al.(2005) considered region-specific factors with a similar hypothesis as in our study [6], but did not analyze hourly data; they explored monthly regional energy demand responses to climate change by assessing temperature-sensitive energy demand. Although there are some studies presenting the sensitivity of the electricity consumption to temperature in Japan [7][8], we were not able to find studies modeling the relationship between power demand and climatic conditions, considering the region-specific factors and multiple climate indicators.

For the purpose of offering a regression model with which to assess the comprehensive measures, including socio-economic conditions and land-use policies, to lower the fluctuation of hourly electricity demand, we propose a series of methods to construct a single regression model that explains hourly electricity consumption considering the effect of region-specific factors as well as time-varying factors. There are three technical problems to overcome, however, in order to build the regression model.

The first challenge is how to build a model that can take both region-specific factors and time-varying factors into account. We classify the factors affecting hourly power demand into the following three groups, from a more quantitative point of view: 1) region-specific factors, 2) time-varying factors, 3) the interaction factors. The region-specific factors in this research are the factors where the difference by location is greater than the change over time in one year; for example, the population of each region, industrial production, building composition, annual average temperature. The time-varying factors are the factors that the change over time in one year and are larger than the difference by location, such as hourly temperature, humidity, and the percentage of sleeping people. The interaction factors are the factors that indicate the synergistic effect of the region-specific factors and time-varying factors. If we accept the hypothesis above, the hourly fluctuation of power demands scaled by the region-specific factors, the absolute fluctuation of power demand, should be represented as the interaction effect between time-varying factors and the region-specific factors, and theoretically, these interaction factors will be the main factors that determine hourly power demand. A specific example for the interaction factor may be the effect such in the region with larger the population, the greater the increase in power demand when the hourly temperature rises by 1 ° C.

The second challenge is how to identify the indicators that accurately explain the hourly power demand, among many possible indicators. There are many factors that may affect power demand. Considering the interaction effects of many combinations of factors, there are innumerable possible index candidates representing the factors that determine hourly power demand. Many of the candidate indicators are correlated, and it is not easy to identify more accurate indicators from correlated indicators. In this research, we construct original algorithms that apply existing machine learning methods repeatedly to select an effective variable from a wide range of possible indicators.

The third challenge is how to model complex relationships such as the nonlinear relationship between power consumption and the determinants, and concurrently to construct generalizable models using as few parameters as possible. The application of MARS：multivariate adaptive regression splines [9] made it possible to construct a flexible and highly generalizable model while selecting important variables. Although there are few cases in which the MARS is applied to estimate hourly power consumption, some studies showed the excellent prediction performance of MARS with short-term power consumption data [10][11]. MARS is therefore considered to be a suitable model for this study.

The regression model built by overcoming these challenges could be a useful tool not only to understand the relationship between climate and power consumption, but with which to discuss various measures including land-use planning and

peak management for lowering the fluctuation of the power demand.

We first prepare as many indicators as possible that express these three types of factors, and select indicators that explain power consumption more accurately by an algorithm using machine learning. Targeting the whole area of Japan, except the Okinawa region, we then construct a single regression model, MARS, which explains hourly power demand using the selected indicator as an explanatory variable. Section 1 has presented the research background, objective, awareness of the problem, and outline of the paper. In Section 2, we describe a method that is configured to overcome the above-mentioned technical problems and construct a target regression model. In Section 3, we describe the process and results of variable selection. In Section 4, we present the constructed model and the model's performance. Section 5 shows an example of a simulation using the constructed model. Section 6 we experimentally applied the simulation to understand the effect of the spatial distribution of households on hourly power demand using the model we constructed. Finally, in Section 7, we summarize the study and presents the remaining issues.

2. Method

We construct a regression model that estimate hourly power consumption in Japan based on various factors including region-specific factors and time-varying factors under the hypothesis that impacts on hourly energy consumption are scale dependent due to region-specific factors.

2.1. Target area

For taking region-specific factors into consideration in Japan, we need analysis covering the large area of Japan which is long from north to south and has diverse nature, climate, and industry. We use the power supply-demand data provided by major electric power companies for the analysis. There are 10 major electric power companies called “general power transmission and distribution business operators” in Japan, hereinafter referred to as EPC. As shown in Figure 1, the 10 EPC’s jurisdictional area covers almost the whole national land of Japan, and the power used in the area is supplied by EPC. This research targets the 9 EPC’s jurisdictional area excluding Okinawa. We excluded Okinawa, because Okinawa is a subtropical island area and too unique in many respects, such as industry, culture, and building style, it was difficult to represent the whole area of Japan including Okinawa by one model.

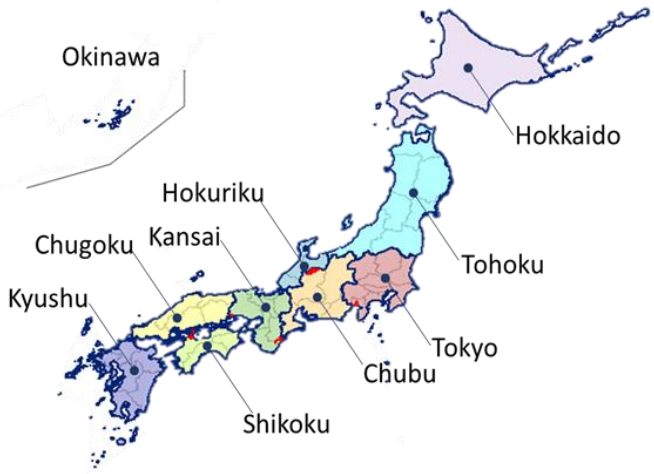

Figure 1 Target Area

2.2. Data source and variables

We used the hourly power supply-demand data which are recorded in each EPC’s jurisdictional area as the explained variable. This is the actual power consumption in each EPC, recorded every hour, for a period from April 1, 2016, to March 31, 2018.

Numerous factors that are known to affect power demand, and it is difficult to identify the specific determinant factors in advance: identifying which indicator more accurately represents those factors is even more difficult. We therefore collected as many indicators as possible that could represent temporal and spatial factors. Table 1 shows a list of the candidate explanatory variables (predictors).

There are 23 candidate predictors for time-varying factors, and 618 for region-specific factors, and variables representing the interaction between region-specific factors and time-varying factors.

The 23 indicators representing time-varying factors were prepared for each power company to correspond with the explained variable. The 618 indicators representing region-specific factors were prepared for each prefecture and aggregated into each power company. The interaction variables are the product of the variables representing time-varying factors and region-specific factors.

The problem arises here of determining which climate data should be representative of the climate for the target area. The jurisdiction area of each EPC is wide, and the climatic conditions in the area are not uniform. Using the average climatic data observed at the stations in each EPC jurisdiction area is not appropriate since the spatial distribution of power consumers is not evenly placed within the jurisdiction area. The climate data of the city with the largest population is used directly as representative data. We do not weight the data by population because, like the other reason-specific factors, the population is one of the factors that explain electricity consumption.

Table 1 Data source for candidate predictors

	Data	Ref.	Example of the indicators	Number of indicator
Power data (Hourly) 1	**Historical energy demand**; FY2016, FY2017 Organization for Cross-regional Coordination of Transmission Operators Japan,	[12]	Energy demand	1
Region-specific data (By prefecture) 168	**Social indicators by prefecture; 2015** Statistics Bureau Management and Coordination Agency Government of Japan, Edited by Japan Statistical Association	[13]	Establishments of each industry, No. of persons in each industry, Urbanization control area, Residential area, Neighboring commercial area, Exclusive industrial area, Out/in flow population, No. of Convenience stores, Floor area of owned houses (per dwelling), etc.	103
	National census 2015; 2015 Statistics Bureau, Ministry of Internal Affairs and Communications Japan,	[14]	Population, No. of households, DID area, DID population, Daytime population, Night-time population, etc. *DID: Density Inhabited District	9
	Environmental statistics; FY2016, FY2017 Ministry of the Environment Government of Japan,	[15]	Gross prefecture production, import tax, Forest area, Natural park area, Cultivated land area, etc.	23
	Energy consumption statistics by prefecture; 2015 Agency for Natural Resources and Energy.	[16]	Estimated consumption of coal, crude oil, natural gas, city gas, renewable and underutilized energy, hydropower for business, nuclear power generation, etc.	9
	Annual Weather data; FY2016, FY2017 (Derived from Historical Weather Data in 2016 and 2017)	[17]	Temperature, Relative humidity, Solar radiation, Wind speed, Rainfall amount, Snowfall amount, etc.	24
Time-varying data (Hourly, by region) 22	**Historical Weather Data; FY2016, FY2017** Japan Meteorological Agency,	[17]	Temperature, Relative humidity, Solar radiation, Wind speed, Rainfall amount, Snowfall amount, etc.	6
	Thermal index; FY2016, FY2017 (Derived from Historical Weather Data in 2016 and 2017)	[17]	WCI(Wind Chill), DI(Discomfort Index)	2
	NHK Data Book 2015 National Time Use Survey 2015; 2015 NHK Broadcasting Culture Research Institute.	[18]	The Proportion of sleeping people, working people, people who stay at home and wake to the population, etc.	4
	Precipitation dummy (Derived from Historical Weather Data in FY2016 and FY2017)	[17]	Rainfall time dummy, Snowfall time dummy	2
	Human activity dummy (Derived from calendars in FY2016 and FY2017)	-	Late night dummy, Active hour dummy, Weekday active hour dummy, Weekend dummy, Public holiday dummy, Long holiday season dummy, etc.	8

2.3. variable selection

If we simply count the possible cases, the candidate variables would reach over 14,855. From such numerous variables, we need to narrow down those that explain power consumption more precisely. We developed the original algorithm for selecting the variables that have a greater impact on power demand. The detail of the variable selection process is shown in Section 3.

2.4. Model development

We construct the MARS using the finally selected predictors. Although there have been many applicable regression models in recent years, MARS is useful in that it is highly generalizable and can represent non-linear relationships through a combination of relatively simple linear piecewise functions. We applied the data from the 2016 fiscal year (from April 1, 2016, to March 31, 2017) for model building, and data from the 2017 fiscal year (from April 1, 2017, to March 31, 2018) for the validation of the models. The details of the modeling process are shown in Section 4.

3. Variable selection

In this section, we propose an algorithm using RandomForests (Breiman 2001)[19] and MARS (Friedman 1991)[9] to narrow down the small number of effective predictors among more than 6,000 possibilities, including variables for region-specific, time-varying and interaction factors. Figure 2 shows the process of selecting the variables.

First, we prepared 618 indicators representing region-specific factors for 46 prefectures, excluding Okinawa, from 47 Japanese prefectures, as well as the corresponding power consumption data. The 618 indicators were narrowed down to 194 indicators using Algorithm-A, as shown in Figure 2. Algorithm-A is an algorithm that excludes meaningless variables to power demand using RF. Many variables are strongly correlated with each other among the 648 candidate predictors. In such a case, it is difficult to perform accurate variable selection simply by applying a single RF and performing variable selection based on the magnitude of the variable importance measure. Algorithm-A makes use of the property of IncMSE, which is one of the variable importance measures that we can obtain concurrently by RF modeling. If the measure (IncMSE) of a variable is negative, it means that the variable poorly explains power demand comparing to the shuffled random variable. In Algorithm-A, we first, build the RF, then, 2) obtain the IncMSE and exclude the variables that have negative IncMSE, and 3) build the RF again with the remaining variables. By repeating operations 2) and 3) until there are no excluded variables (until the number of variables converges) then only the variables that have some meaning for the explained variable remain. In order to combine time-varying variables, the selected region-specific variables for each of the 46 prefectures are then aggregated into 9 EPDs. We add 24 regional climate data, such as the mean temperature of a year to the 194 selected region-specific variables. Figure 3 visualizes the variable selection process by algorithm-A taking Algorithm-A1 as an example.

Second, we prepared 23 time-varying variables and randomly sampled 10,000 observations from 78,813 observations that is the observations excluded missing observations from the 78,840 (364 days×24 hours×9 EPC) observations.

Third, we merge region-specific variables to time-varying variables by the regions (EPC), and generated interaction variables by multiplying region-specific variables and time-varying variables. At this stage, the number of predictors was 6,095.

Fourth, we again applied Algorithm-A to this dataset. This time, the explained variable is the hourly power supply-demand data which is recorded in each EPC's jurisdictional. The number of predictors is then narrowed down to 1,039. In Algorithm-A using RF, as long as it is a variable that explains the explained variables, all variables will be adopted even if they are correlated predictors.

Fifth, we examined highly correlated variables and omit the variables that less likely to explain the power demand from the highly correlated variable group. This is the only human judgment in the procedure. The number of predictors is then narrowed down to 458.

Sixth, we apply Algorithm-B. MARS included important variables in the model and excludes unimportant ones when we built the model. We can select important variables from the correlated predictors using MARS, however, there is still some arbitrariness in the selection when correlated predictors included. In Algorithm-B we thus applied MARS 1000 times to the dataset, and counted the frequency with which each variable is selected. We kept 44 predictors that were frequently (more than 10 times) selected by MARS. At this stage, the number of predictors was 44.

Seventh, we applied Algorithm-B again. This time, we weighted the observation of each region (EPC) by the log of the reciprocal of the mean value of hourly power demand data in each region (EPC). We weighted the observations because the properties of small EPCs tend not to be reflected the model, and it was difficult to weight the observations until the number of predictors was narrowed down to less than about 50.

Finally, we obtained 20 variables that are considered to explain the hourly power consumption accurately. We adopt these 20 variables for the final model

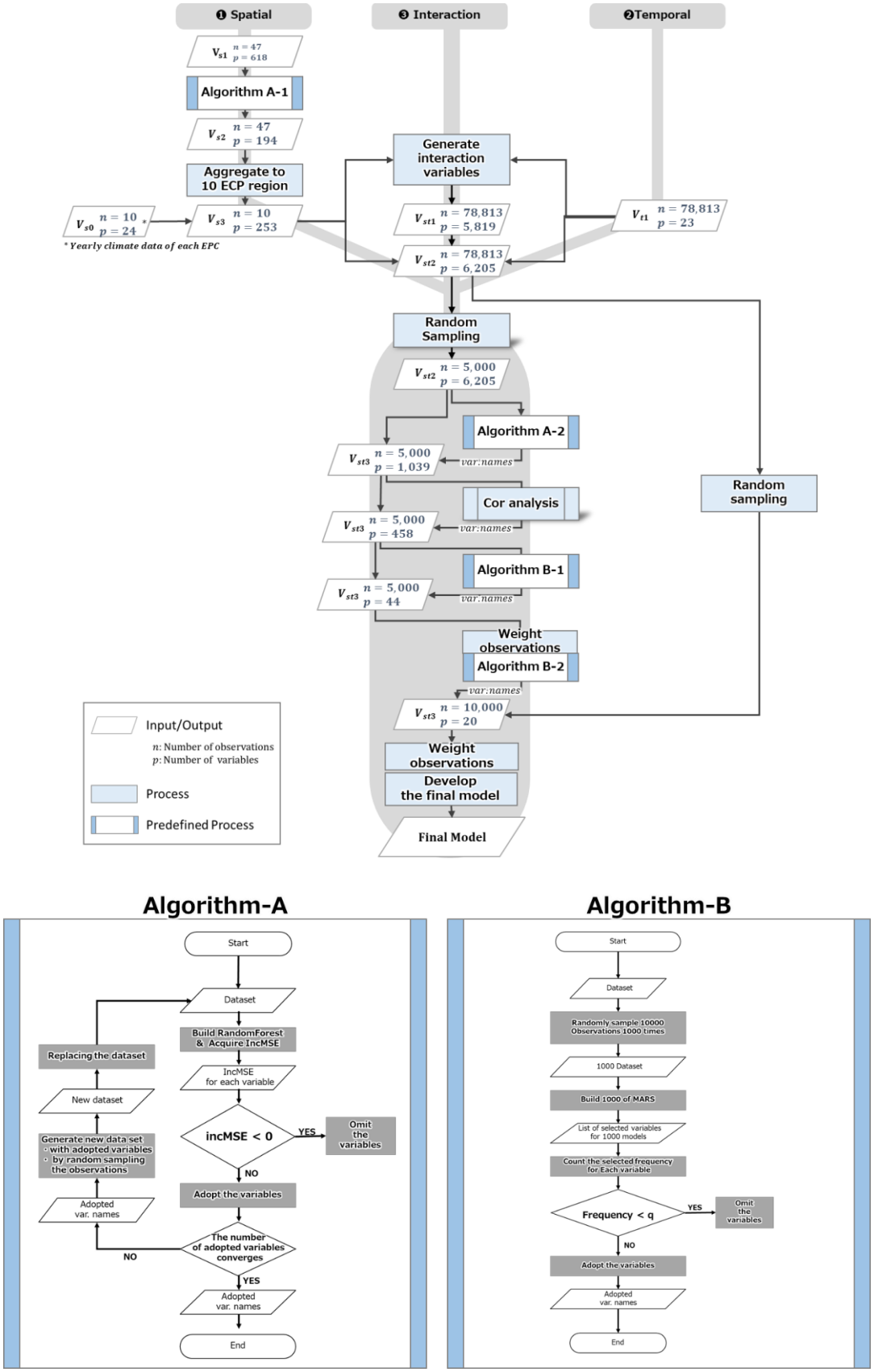

Figure 2 Variable selection procedure

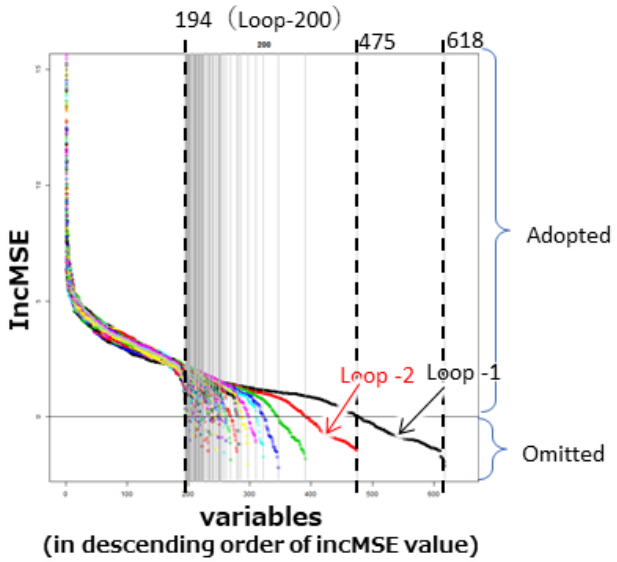

Figure 3 An example of the variable selecting process using algorithm-A (Algorithm-A1)

4. Constructed model

We built the final model using all 78,813 observations for 20 selected variables. Ultimately, one region-specific variable and 9 interaction variables were adopted for the model. Table 2 shows the selected variables and their explanations. The terms and coefficients of the constructed model are shown in

Table 3. As a result, one region-specific predictor and 9 interaction predictors were adopted for the final model from 20 predictors, and 36 (including the intercept) terms were used for the model. The model R^2 and generalized R^2 are 0.990341 and 0.9901697, respectively.

The MARS model was tuned based on the correlation coefficient (r) between observed and predicted power demand using out-of-sample data. In this study, since we used 2 degrees of interaction variables prepared in advance, interaction terms which generated by MARS (form term-18 to term-36 in the

Table 3) played a less important role in the model but they contributed to gain the flexibility of the model.

Table 2 Selected variables and the explanation

Selected variables		Explanations		
❶	Gross	S1	Gross	Gross regional product
❷	Gross_2 * WORK%	S2	Gross_2	Gross regional product of secondary industry
		T1	WORK%	The proportion of working people to the population
❸	Worker_3 * WAKE%	S3	Worker_3	Persons in tertiary industry
		T2	WAKE%	The proportion of people who stay at home and wake to the population
❹	Tax_L * DI	S4	Tax_L	Local taxes [prefecture]
		T3	DI	Discomfort Index generated by temperature and humidity
❺	log(minHUM) * WCI	S5	minHUM	Annual minimum humidity
		T4	WCI	Wind Chill Index generated by temperature and wind speed
❻	FlArea * DI	S6	FlArea	Floor area of owned houses (per dwelling)
		T3	DI	Discomfort Index generated by temperature and humidity
❼	log(HCG) * DI	S7	AHCG	Ratio of households covered by city gas supply system
		T3	DI	Discomfort Index generated by temperature and humidity
❽	Dwell * noWorkD	S8	Dwell	Dwellings occupied by households
		T7	noWorkD	Dummy for not work day
❾	FlArea * TEMP	S6	FlArea	Floor area of owned houses (per dwelling)
		T8	TEMP	Temperature
❿	Dwell * SLEEP%	S8	Dwell	Dwellings occupied by households
		T9	SLEEP%	The proportion of sleeping people to the population

Table 3 The model information

	Selected terms	Coefficients
Term-1	(Intercept)	2,361.615492
Term-2	h(741850-❶)	−0.016226
Term-3	h(❶-741850)	0.016729
Term-4	h(9.50472e+07-❸)	0.000001
Term-5	h(❸-9.50472e+07)	0.000002
Term-6	h(1.0716e+08-❹)	0.000057
Term-7	h(❹-1.0716e+08)	−0.000042
Term-8	h(784.567-❺)	4.790679
Term-9	h(❺-784.567)	1.784924
Term-10	h(❻-1.20769e+10)	0.000000
Term-11	h(❼--49.4723)	−31.543987
Term-12	h(6.5991e+06-❽)	0.000156
Term-13	h(❽-6.5991e+06)	−0.000062
Term-14	h(1.30189e+10-❾)	0.000000
Term-15	h(❾-1.30189e+10)	0.000000
Term-16	h(6.9783e+07-❿)	0.000038
Term-17	h(❿-6.9783e+07)	−0.000024
Term-18	h(741850-❶)*h(❺-804.497)	−0.000004
Term-19	h(741850-❶)*h(804.497-❺)	−0.000003
Term-20	h(❶-741850)*h(❷-6.66484e+07)	-0.00402×10^{-10}
Term-21	h(❶-741850)*h(6.66484e+07-❷)	-0.71476×10^{-10}
Term-22	h(❶-741850)*h(❻-1.43866e+11)	0.00134×10^{-10}
Term-23	h(❶-741850)*h(1.43866e+11-❻)	0.00034×10^{-10}
Term-24	h(184846-❶)*h(6.9783e+07-❿)	0.98111×10^{-10}
Term-25	h(4.14333e+07-❷)*h(6.5991e+06-❽)	-0.01697×10^{-10}
Term-26	h(❷-4.14333e+07)*h(6.5991e+06-❽)	0.00505×10^{-10}
Term-27	h(4.17096e+08-❷)*h(❿-6.9783e+07)	0.00060×10^{-10}
Term-28	h(❷-4.17096e+08)*h(❿-6.9783e+07)	0.00047×10^{-10}
Term-29	h(❸-9.50472e+07)*h(❿-3.02393e+07)	0.00003×10^{-10}
Term-30	h(❸-9.50472e+07)*h(3.02393e+07-❿)	0.00143×10^{-10}
Term-31	h(469.164-❺)*h(❾-1.30189e+10)	1.83491×10^{-10}
Term-32	h(❺-469.164)*h(❾-1.30189e+10)	-0.88180×10^{-10}
Term-33	h(778.874-❺)*h(❿-6.9783e+07)	$-42.76737 \times 10^{-10}$
Term-34	h(❺-778.874)*h(❿-6.9783e+07)	$-16.62343 \times 10^{-10}$
Term-35	h(1.64382e+11-❻)*h(6.9783e+07-❿)	0.00000×10^{-10}
Term-36	h(❻-1.64382e+11)*h(6.9783e+07-❿)	0.00005×10^{-10}

5. Model performance

5. 1. Fitting and generalization ability

Figure 4 compares actual observations to the estimated value by the constructed MARS. The figure on the left hands shows the result using the training data to build the model (in sample result, fitting), and the figure on the right shows the result using the test data which is not used for the model construction (out of sample result, prediction).

The coefficient of determination by linear regression (OLS) is 0.99026 in the in-sample result and 0.98909 in the out-of-sample result. The determination coefficient of the in-sample result is higher, but the determination coefficient of the out-of-sample result is also sufficiently high, so it can be said that the constructed model is a high-quality model in terms of both fitting and generalization. One of the reasons for the high performance of the model is the property of MARS that expresses the complexity of a potential model by applying a locally linear model. The other is the variable selection process: we constructed RF and MARS many times with a dataset which was generated by repeating random sampling in the variable selection algorithm.

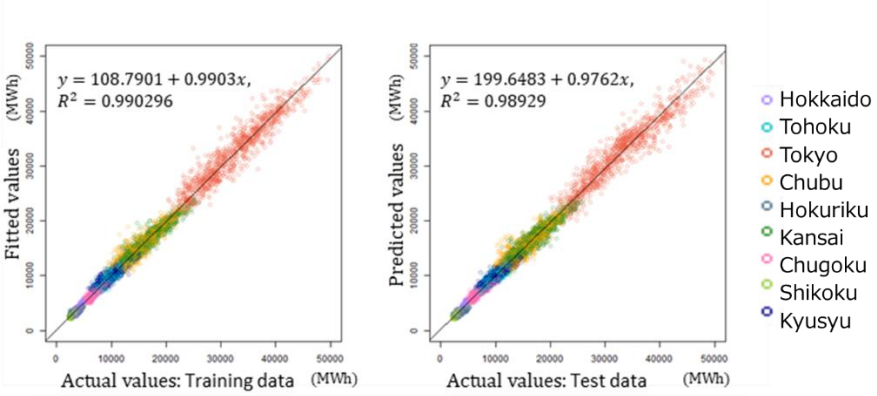

Figure 4 Plot of actual values against predicted values for electricity demand

5. 2. Normality of residuals

Figure 5 (left) shows the Q-Q plot of the constructed model, and Figure 5 (right) compares the density distribution of the residuals and the normal distribution with its mean and variance as parameters. Figure 5 (left) shows that the residuals are fat-tailed, which means compared to the normal distribution there is more data located at the extremes of the distribution, however, being fat-tailed does not necessarily mean that the variance of the residual is large. In Figure 5 (right figure), although the residuals of the constructed model include observed values that deviate greatly from the estimated values, most residuals are concentrated around 0 and become a density function with high kurtosis. As the explained variable of this model is hourly power consumption, frequent outliers are inevitably caused by temporary events. The estimation result of the model is good for most of the observed values, however.

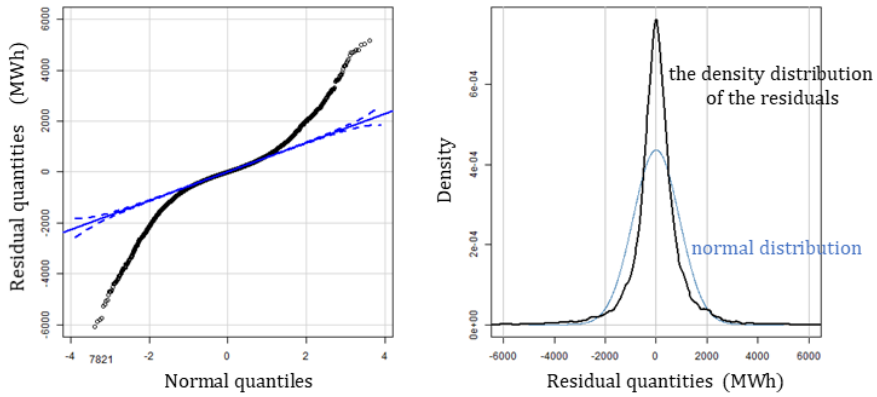

Figure 5 The normal Q–Q plot and the distribution of the residuals of the constructed model

5. 3. Simulation for model evaluation

We examined the relationship between temperature and estimated power consumption in order to evaluate the model. This is because the relationship between temperature and power consumption is often studied as temperature sensitivity to power consumption. A simulation with the conditions shown in Table 4 was performed using the constructed model. For each hour from 1 to 24 in the simulation, the average value of each region was applied to the time-varying variables, but only for temperature, we generated and applied a sequence value from minimum temperature of the region to maximum temperature of the region increasing by 0.01℃. The region-specific variables in each region were adopted without any change. Figure 6 shows the results, with the temperature on the horizontal axis and the power consumption on the vertical axis. The black points represent observed data, and the colored lines represent estimation results from the constructed model. Although in most regions, hourly estimated results seem to cover the distribution range of observed data, in some regions, there is a noticeable difference between the observation and estimation. Estimated results in the region where total power demands are low such as Hokkaido and Shikoku especially clearly did not fit the distribution range of observed data. This suggests that the indicators adopted in the constructed model are not sufficient to express the actual power consumption determinants. There is still room for improving the model by finding more appropriate spatial variables and examining the temporal variables interacting with it. In this study, during the variable selection process, we did not allow MARS to include interaction terms in the algorism-B. However, allowing MARS to select interaction combination, other than we set up interaction variables in advance, may one way to find the better variables and the combination.

Table 4 Settings for the simulation

	Variable name	Setting
S1	Gross	Average value in each region (EPC)
S2	Gross_2	Average value in each region (EPC)
S3	Worker_3	Average value in each region (EPC)
S4	Tax_L	Average value in each region (EPC)
S5	minHUM	Average value in each region (EPC)
S6	FlArea	Average value in each region (EPC)
S7	AHCG	Average value in each region (EPC)
S8	Dwell	Average value in each region (EPC)
S9	FlArea	Average value in each region (EPC)
S10	Dwell	Average value in each region (EPC)
T1	WORK%	Value at each hour
T2	WAKE%	Value at each hour
T3	DI	Derived from sequential TEMP and averaged HUM at each hour in region (EPC)
T4	WCI	Derived from sequential TEMP and averaged WIND at each hour in region (EPC)
T7	noWorkD	0 (week day)
T8	TEMP	Regular sequence value
T9	SLEEP%	Value at each hour

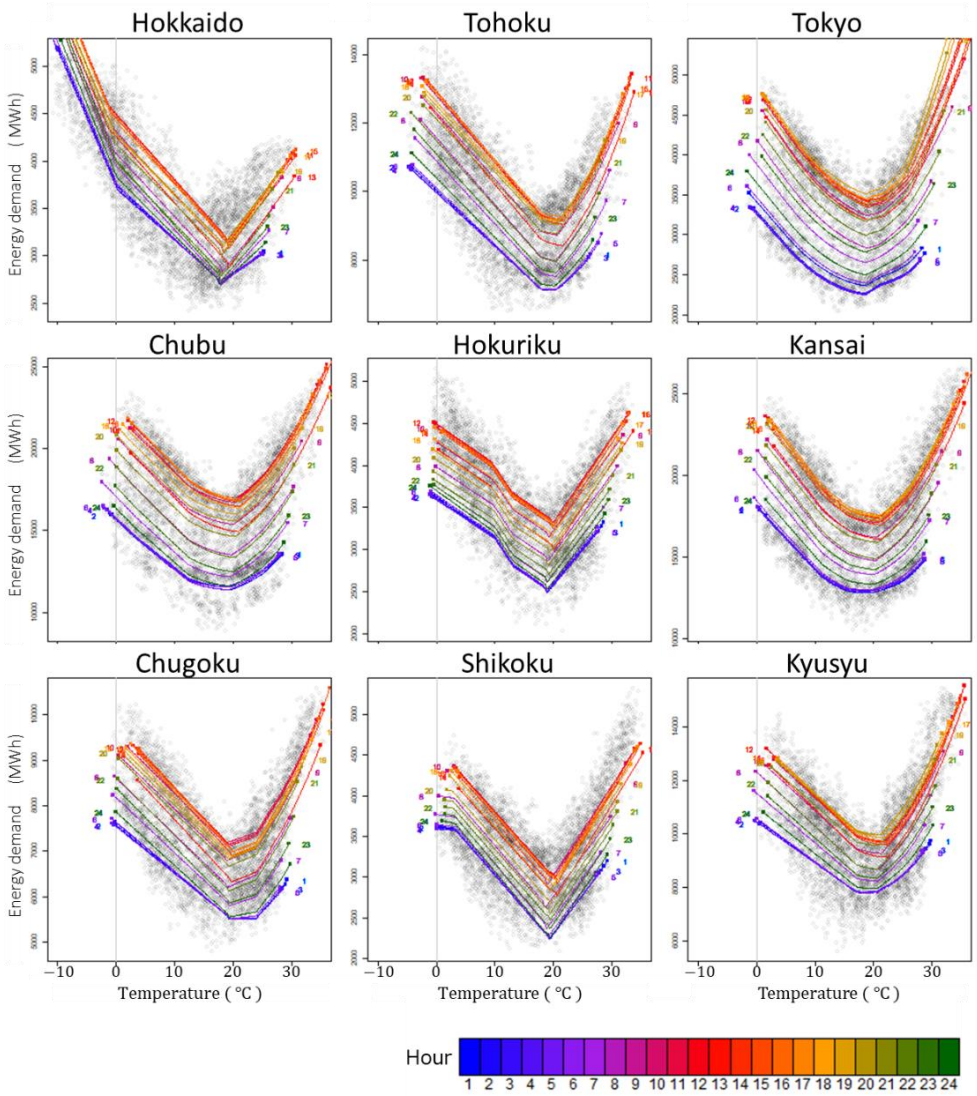

Figure 6 The relationship between temperature and estimated power consumption estimated by the simulation

6. An example of simulation: the effect of spatial distribution of households on hourly power demand

In this section, we simulate the change in the ratio of households in the city gas supplied area using the constructed model. The ratio of households in the city gas supplied area is hereinafter referred to as HCG. HCG is the only indicator among other adopted variables that stands for the spatial distribution of households. We regard the city gas supplied area as an urbanized area, because such costly infrastructure for gas supply would only be developed in an urbanized place where households are concentrated. Using this assumption, a region where HCG is high can be assumed to be a relatively high residential density region because the population of the region is concentrated in the urbanized (city gas supplied) areas. Conversely, in a region where HCG is low, the region can be described as a relatively low residential density region. The population in densely inhabited districts (DID), and the population in an urbanization control area, for example, can be considered an index for the spatial distribution of the population and households, however HCG is a better indicator for representing the effect of household spatial distribution on power consumption, because the careful variable selection in Section 3 means that HCG is one of the variables that explain power consumption. HCG is also the actual urbanized area, rather than an ideally defined urbanized area, because it is where the infrastructure investment for gas supply was actually made. Table 5 shows the settings for the simulation conditions. HCG is given as regular sequences from 0 to 100. The average value in each region was adopted for the other variables.

Figure 7 shows the expected power consumption per household of each EPD estimated by the simulation. The horizontal axis shows regular sequences of HCG, and the vertical axis shows power consumption per household. As mentioned above, since there is still a lot of room for improvement in this model; it is not suitable for discussing numerical values, but we can indicate a general trend. It is assumed that the city gas supply area does not shrink even the population in the region declined, because once the infrastructure constructed there is less likely to withdraw the system. Therefore, this simulation is not for discussing the decreasing scenario of HCG, but for discussing the increasing scenario. For all regions, a similar trend can be seen that the power consumption decreases as HCG increases. In other words, the region with higher residential density is likely to consume less power per household. In addition, we can see the rough trend that in areas where the current HCG is low, the reduction rate of power consumption with rising HCG is high. From the trend, it is assumed that the regions where HCG is currently small have a greater potential to reduce power consumption per household by concentrating the households.

Although the importance of high-density dwelling has been repeatedly emphasized from the viewpoint of low carbonization, there are few studies that present the evidence quantitatively. The model proposed in this study could potentially be a method with which to evaluate the effect of the spatial distribution of households as regards energy consumption. It is not easy to implement measures to change the spatial distribution of households in practice. However, for finding the comprehensive measure to find the solutions to reduce the fluctuation range of the power demand, the refinement and improvement of the constructed model are necessary.

Table 5 Simulation settings for evaluating the effect of HCG

	Variable name	Settings
S1	Gross	Average value in each region (EPC)
S2	Gross_2	Average value in each region (EPC)
S3	Worker_3	Average value in each region (EPC)
S4	Tax_L	Average value in each region (EPC)
S5	minHUM	Average value in each region (EPC)
S6	FlArea	Average value in each region (EPC)
S7	AHCG	Sequence value from 0.000 to 1.000 by 0.001
S8	Dwell	Average value in each region (EPC)
S9	FlArea	Average value in each region (EPC)
S10	Dwell	Average value in each region (EPC)
T1	WORK%	Value at 11 AM
T2	WAKE%	Value at 11 AM
T3	DI	Derived from average TEMP and HUM at 11 AM in region (EPC)
T4	WCI	Derived from average TEMP and HUM at 11 AM in region (EPC)
T7	noWorkD	0 (week day)
T8	TEMP	Average value at 11 AM in region (EPC)
T9	SLEEP%	Value at 11 AM

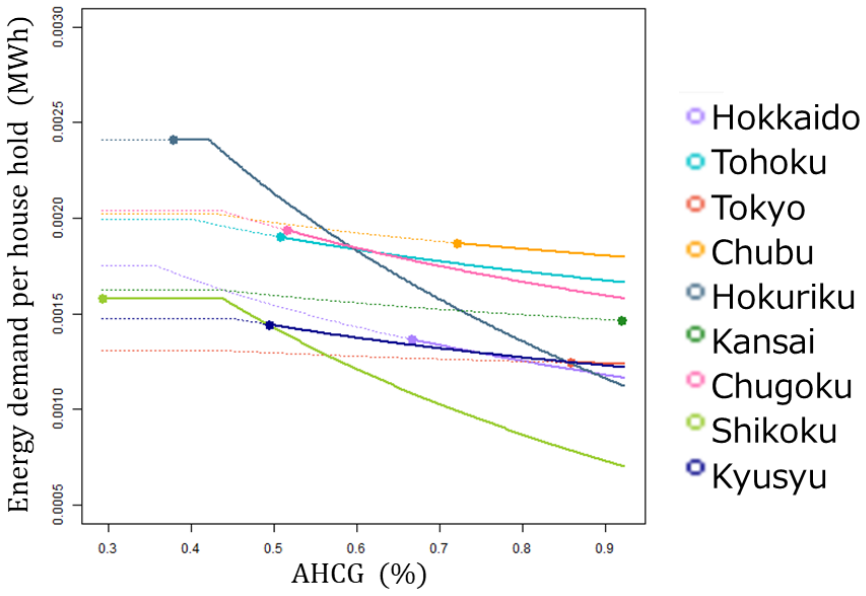

Figure 7 Simulation for evaluating the effect of HCG (Dots are the current HCG in each region)

7. Conclusion

We proposed a series of methods to construct a single regression model that explains hourly electricity consumption based on both region-specific factors and time-varying factors, in order to offer a regression model to assess multiple measures, including socio-economic conditions and land-use policies to lower the fluctuation of hourly electricity demand.

We developed our original algorithms using existing machine learning methods (RF, MARS) to select an effective variable from numerous possible indicators, including region-specific factors, time-varying factors, and their interaction factors. We then modeled the complex relationships between power consumption and multiple factors by applying MARS. Although the performance of the whole model was appropriate from the viewpoint of fit and generalization, there is still room to improve it, because the difference between the power consumption estimated for each temperature and the actual data was obvious in some regions. As we were able to test a general simulation for estimating the effect of the spatial distribution of household on electricity demand, however, further improvement of the constructed model will offer a useful tool, not only for understanding the relationship between climate and power consumption but for discussing various measures such as land-use planning for power management.

In order to build a model that is applicable to practical use, it is necessary to develop a more efficient variable selection method that can deal with multi-dimensional interactions in the early stage of the variable selection process. It is also necessary to discover new variables that can represent the characteristics of the Okinawa region, so as to include the region in the model. It may necessary to consider adopting more complex (flexible) regression models that retain a generalization ability than MARS. Although MARS is excellent in flexibility, generalization, and interpretability, the number of predictors that can be adopted for the model is hard to tune, and it is difficult to incorporate small effects of factors of electricity demand into the model.

There have been many models built for predicting power, or for identifying the determinants of power demand, however, there are few models to evaluate the effect of changes in regional factors such as land use on hourly power demand, taking both region-specific factors and time-varying factors into consideration. Global warming and a declining and aging population in Japan mean that comprehensive measures should be studied. Although there is a room for improvement, this research proposes a means of modeling that may contribute to planning policy.

References

[1] Temporal and seasonal fluctuation of electricity demand (01-09-05-07) - ATOMICA - n.d. https://atomica.jaea.go.jp/data/detail/dat_detail_01-09-05-07.html (accessed March 23, 2019).

[2] Japan TF of EPC of. Load leveling of electricity demand n.d. http://www.fepc.or.jp/enterprise/jigyou/juyou/index.html (accessed March 23, 2019).

[3] Mukherjee S, Nateghi R. Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States. Energy 2017;128:688–700. doi:10.1016/J.ENERGY.2017.04.034.

[4] Mukherjee S, Vineeth CR, Nateghi R. Evaluating regional climate-electricity demand nexus: A composite Bayesian predictive framework. Applied Energy 2019;235:1561–82. doi:10.1016/J.APENERGY.2018.10.119.

[5] Wang Y, Bielicki JM. Acclimation and the response of hourly electricity loads to meteorological variables. Energy 2018;142:473–85. doi:10.1016/J.ENERGY.2017.10.037.

[6] Amato AD, Ruth M, Kirshen P, Horwitz J. Regional Energy Demand Responses To Climate Change: Methodology And Application To The Commonwealth Of Massachusetts. Climatic Change 2005;71:175–201. doi:10.1007/s10584-005-5931-2.

[7] Hirano Y, Gomi K, Nakamura S, Yoshida Y, Narumi D, Fujita T. Analysis of the impact of regional temperature pattern on the energy consumption in the commercial sector in Japan. Energy and Buildings 2017;149:160–70. doi:10.1016/J.ENBUILD.2017.05.054.

[8] Kittaka K, Miyazaki H. Study on the Air Temperature Sensitivity in Wide Area (Special Issue : AIUE2015 : 12th international conference of Asia Institute of urban environment : Survival strategy in the rapidly changing urban environment). Asian Institute of Urban Environment; 2015.

[9] Friedman JH. Multivariate Adaptive Regression Splines. The Annals of Statistics 1991;19:1–67. doi:10.1214/aos/1176347963.

[10] Sigauke C, Chikobvu D. Daily peak electricity load forecasting in South Africa using a multivariate non-parametric regression approach. ORiON; Vol 26, No 2 (2010) 2010.

[11] Al-Musaylh MS, Deo RC, Adamowski JF, Li Y. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Advanced Engineering Informatics 2018;35:1–16. doi:10.1016/J.AEI.2017.11.002.

[12] Organization for Cross-regional Coordination of Transmission Operators J. Historical energy demand n.d. http://occtonet.occto.or.jp/public/dfw/RP11/OCCTO/SD/CC01S042C?fwExtention.pathInfo=CC01S042C&fwExtention.prgbrh=0.

[13] Statistics Bureau Management and Coordination Agency Government of Japan, editor. Social indicators by prefecture. Japan Statistical Association; 2015.

[14] Statistics Bureau, Ministry of Internal Affairs and Communications J. National census 2015. https://www.stat.go.jp/english/index.html.

[15] Ministry of the Environment Government of Japan. Environmental statistics 2016. http://www.env.go.jp/doc/toukei/contents/1shou.html#1shou.

[16] Agency for Natural Resources and Energy. Energy consumption statistics by prefecture n.d. https://www.enecho.meti.go.jp/statistics/energy_consumption/ec002/results.html#headline2.

[17] Japan Meteorological Agency. Historical Weather Data n.d. http://www.data.jma.go.jp/obd/stats/etrn/index.php.

[18] NHK Broadcasting Culture Research Institute. NHK Data Book 2015 National Time Use Survey 2015.

[19] Breiman L. Random forests. Machine Learning 2001;45:5–32. doi:10.1023/A:1010933404324.