Modeling joint distribution of monthly energy uses in individual urban buildings for a year

Jeonghun Song [a], Hoyong Choi [b,*]

[a] *Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland*

[b] *Graduate School of Innovation and Technology Management, College of Business, Korea Advanced Institute of Science and Technology, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea*

* Corresponding author, E-mail: hoyong.choi@kaist.ac.kr

[Abstract]

Monthly energy use in individual buildings is informative data on seasonal energy consumption in urban areas. In the related previous studies based on statistical estimation, mean and variance of energy use for each month have been investigated. However, correlation between energy uses in different months has not been investigated despite of its existence and importance for probabilistic approach. This study provides a regression-based method for modeling a joint probability distribution of monthly electricity and gas uses for a year in individual urban buildings, which reflects correlation between energy uses in different months and between electricity and gas. The mean vector of monthly energy uses is estimated by linear regression models where the explanatory variables are floor area, number of stories, and approval year for use of individual buildings. The covariance matrix of monthly energy uses is estimated using the sample covariance of the residuals of the regression models. Non-constant but increasing covariance (heteroskedasticity) of energy use with increasing floor area has been reflected to ensure realistic magnitude of covariance for a given building size. Based on the estimated mean vector and covariance matrix, a multivariate normal distribution of monthly electricity and gas uses can be established. The multivariate normal distribution can be used for two kinds of tasks which were not able without consideration of correlation – i) sampling vectors of monthly energy uses for a given set of building features, with realistic seasonal patterns and magnitudes of energy use, and ii) data correction like filling in missing values with reasonable values (imputation) and prediction of future values of monthly energy uses in a target building, given correctly measured monthly energy use for some months.

[1. Introduction]

In 2021, the operation of buildings accounted for 30% of global final energy consumption and 27% of total energy sector emissions [1]. Energy saving in building sector is one of the most important activities to alleviate global warming and improve environmental sustainability. One of the key factors for energy saving in building sector is development of estimation methods for energy consumption in individual buildings. Such methods can provide information of building energy performance to decision makers related to energy policy making and energy infrastructure planning.

The methods for estimation of energy use in individual buildings can be separated into two categories – physical methods, and statistical methods [2, 3, 4]. Physical methods adopt detailed physical constraint-based models of building components and external conditions (e.g. detailed construction fabric, detailed shape, lightning and heating, ventilation and air conditioning system, indoor schedule, climate information), then estimate energy use in a target building by simulation tool. Statistical methods adopt regression models which contains energy use record of many individual buildings as the response variable, and features in building register (e.g. floor area, number of stories, category of construction fabric, construction date, etc.) as explanatory variables.

This study belongs to statistical methods, and there are several previous studies on statistical methods for estimation of annual energy use in individual buildings. Many of the studies provide estimation of annual energy use per floor area in the unit of kWh/m^2 year (often called as energy intensity), because the annual energy use of a target building can be estimated as the energy intensity multiplied with its floor area. Some studies have reported constant values of energy intensity of major building uses (e.g. office, retail, hospital, school, etc.) [5, 6, 7]. The other studies have provided linear regression models for estimation of annual energy consumption itself [8, 9] or energy intensity [4, 10, 11] as a function of building features.

This study focuses on 'monthly' energy use in individual buildings, which reflects seasonality of energy use. In general, electricity use in a building is relatively higher in summer due to cooling, while gas use in a building is relatively higher in winter due to heating. Such information of seasonality is helpful for scheduling of fuel supplies, maintenance operation of the utilities and negotiation of contracts between energy companies [12]. Aggregation of monthly energy use of buildings in an urban area enables planning of distributed energy infrastructure and estimation of total capacity of building-integrated energy sources [13]. Also, hourly energy demand pattern

of a building, which is necessary for energy dispatch scheduling, can be estimated from the record of monthly energy use of the building [14, 15, 16].

There are a few previous studies on statistical estimation of monthly energy use in individual buildings, which have been feasible due to availability of open database of monthly energy use in many buildings [15]. The representative studies are as follows.

i) Catalina et al. [17] used linear regression to estimate heating demand in each month for heating period. The dataset has been generated by a dynamic simulation tool for building energy assessment. The explanatory variables are building characteristics (shape factor, transmittance coefficients, window to floor area ratio, etc.) and climate factors (outdoor temperature and global radiation).

ii) Kim et al. [18] used linear regression to estimate electricity use and gas use in each month for a year. The dataset has been obtained from Korean Management System for Building Energy Database. The explanatory variables are floor area, indicator variables for month, building use (neighborhood living or office), subdistrict, number of stories, fabric types of structure and roof.

iii) Xu et al. [19] used two-step k-means clustering to divide the dataset of monthly electricity use in buildings into 16 subsets, then fitted separate normal distribution to each subset. In the first step, the whole dataset has been divided into 4 subsets with respect to magnitude of electricity use. In the second step, each of the 4 subsets has been further divided into 4 subsets with respect to seasonal pattern of electricity use. The dataset has been obtained from smart meter dataset of six cities in Jiangsu province.

The common limitation of the previous studies on statistical estimation of monthly energy use in individual buildings is ignorance of correlation between energy uses in different months or different energy types. In practice, energy uses in different months are expected to be correlated. For example, a building which uses much more electricity in January compared to other buildings with similar size is expected to use much more electricity in February as well. In this sense, positive correlation between electricity uses in January and February is expected. Another example is that gas use for heating in a building depends on the amount of electricity used for electrified heating which is a substitute of gas heating. In this sense, negative correlation between electricity and gas uses in winter is expected.

Considering monthly electricity and gas uses for a year in a building as a 24-dimensional vector, the previous studies have reported information of mean vector and diagonal terms of covariance matrix of the 24-dimensional vector of monthly energy use in individual buildings. However, off-diagonal terms of covariance matrix have not been investigated yet. Information of full covariance matrix including off-diagonal terms enables construction of a 'joint' probability of the vector of

monthly energy use in individual buildings. The joint probability model enables drawing vector samples of monthly energy uses in target buildings given their features, which would be helpful for energy planning for new urban towns with consideration of uncertainty in building energy demand. Also, the joint probability model can enhance data quality, by application to data imputation and prediction which can be done by consideration of correlation in data.

The objective of this study is to provide a statistical method for estimation of 'joint' probability distribution of 'monthly' energy uses for a year in individual urban building. Section 2 presents the dataset used in this study, subset and variable selection for regression, and data pre-processing. Section 3 presents estimation of moment conditions (mean vector and full covariance matrix) of the vector of monthly energy use in individual buildings, based on linear regression models. Section 4 presents the joint probability model and its applications. Section 5 concludes this study with a summary.

[2. Data]

[2.1. Data description]

The following two datasets have been merged and used – i) dataset of monthly electricity and gas use in individual non-residential buildings, provided by Korean Ministry of Land since late 2015; and ii) dataset of building register which includes features of building. Each row of the two datasets corresponds to a single building or multiple buildings corresponding to one address. Each column of the dataset of monthly electricity and gas use is record of electricity use or gas use for one month (in the unit of kWh). The columns of the dataset of building register include address, building use (e.g. office, living neighborhood, hospital, welfare, retail, school, etc), site area, sum of floor area in all stories, number of stories, structure of building and roof, approval date for use, etc.

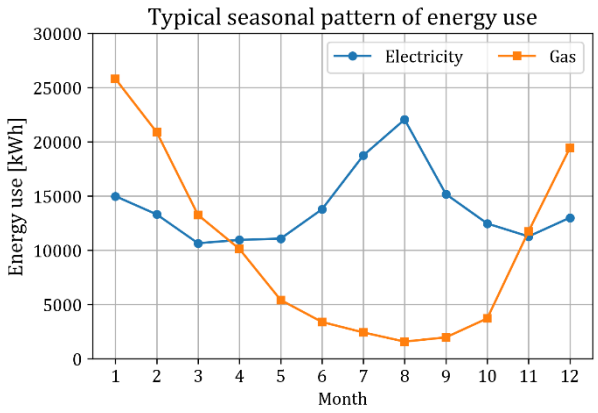

Figure 1. Typical seasonal pattern of monthly energy use in an exemplary building.

Figure 1 shows the typical seasonal pattern of monthly energy use in an exemplary building. The amount of electricity use is relatively higher in summer due to cooling, and relatively lower in spring and fall. The amount of electricity use in winter is usually similar to that in spring or fall. However, in some buildings, it may be as high as that in summer due to recently increasing electrification of heating. The amount of gas use in winter is relatively higher than that in other seasons due to heating. The amount of gas use in seasons other than winter varies much in different buildings depending on building use.

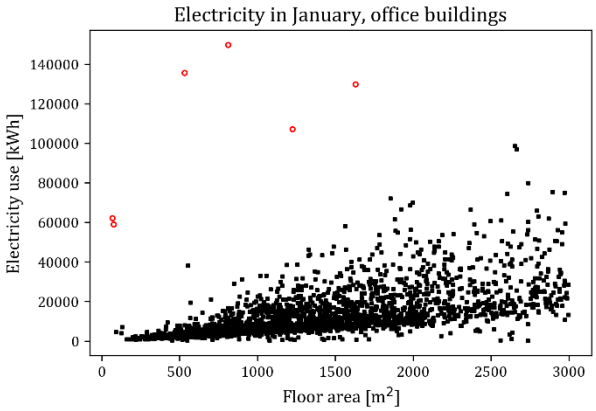

Figure 2. Electricity use for January 2021 in a subset of office buildings in Seoul, for varying floor area (red hollow circles are suspicious to be influential points).

The dataset of monthly energy use in individual buildings has high variance. Figure 2 shows the electricity use in January 2021 in a subset of office buildings in Seoul, for varying floor area. Each data point in Figure 2 corresponds to one office building. The scatterplot presented a somewhat linear relationship, but instead of showing an intensive curve, the dots are more dispersed

towards the end, especially dispersed for higher floor area. This spreading of the points implies two things – i) magnitude of energy use of buildings with similar size can be quite different from building to building, and ii) modern machine learning methods (like neural networks) with low bias but higher variance [20] are not appropriate for this dataset. Rather, traditional linear regression is appropriate for this dataset because linear regression is a method with high bias but lower variance.

[2.2. Data setting]

Among many features in the building register used in this study, the following features may be used to estimate the monthly energy use of individual buildings; floor area, building use, number of stories, approval year for use, category of building structure, and category of roof structure.

The features listed above can be explanatory variables for regression. For example, floor area of individual buildings can be an explanatory variable because the average energy use in individual buildings is expected to increase with increasing building size which is reflected by floor area.

Conversely, the dataset can be divided into subsets with respect to some of the features to make multiple regression models each for one subset. Division into subsets is necessary if the model coefficients are different from subset to subset. For example, the dataset can be divided with respect to building use because energy intensity, which is the coefficient of floor area, has been consistently found to be different for every building use in the previous studies on statistical estimation of energy use in buildings.

[2.2.1. Subsets of the data]

In this study, addition to building use, two additional criteria for subset division have been considered: interval of floor area, and use of gas. The two criteria for division has been selected due to the following reasons.

i) Interval of floor area: Floor area of individual buildings ranges in a very wide interval, from under 100 m^2 to over 100,000 m^2. In Seoul green building standard, the interval of floor area of a building has been divided into four subintervals – under 3,000 m^2, 3,000 m^2 to 10,000 m^2, 10,000 m^2 to 100,000 m^2, and over 100,000 m^2. Different standards of energy performance, management, and renewable energy penetration are applied to each subinterval. Thus, dividing the dataset with respect to the floor area intervals in the standard would make the result of this study practically

available to users in energy policy field. Dividing into clusters obtained by k-means method as in [19] is not considered since it is hard to explain for domain purpose and the optimal classification boundaries can vary for different datasets. Taking log of floor area as in [18] is not considered because the important purpose of this research is to quantify covariance between monthly energy use in different months, not covariance between logged values of monthly energy use.

ii) Use of gas: Some buildings do not use gas, while others use gas. This difference has not been considered as a factor in the previous studies, but it is expected to affect average electricity use in winter because electricity and gas are substitutes for heating in winter. If a building does not use gas and meets its heating demand totally by electric heating, electricity use in winter is expected to be much higher than that in spring or fall. On the contrary, in a building which uses gas for meeting its heating demand, electricity use in winter is expected to be similar to that in spring or fall.

Subsets	Using gas				Not using gas			
	Under 3,000 m²	3,000 m² ~ 10,000 m²	10,000 m²~ 100,000 m²	Over 100,000 m²	Under 3,000 m²	3,000 m² ~ 10,000 m²	10,000 m²~ 100,000 m²	Over 100,000 m²
Office	Subset 1	Subset 2	…					
Neighborhood(I)								
Neighborhood(II)								
Hospital								
Sales								
Welfare								
Lodging								
Education								
Religious								
Cultural								

Table 1. Outline of the division of the building energy dataset into subsets.

Table 1 shows the outline of subset division with respect to the three criteria. Subset division by interval floor area and use of gas will be justified by a statistical test explained in Section 2.2.3, based on linear regression with response variables and explanatory variables explained in Section 2.2.2.

[2.2.2. Response and explanatory variables for regression]

The response variables are electricity and gas use in individual months (for year 2021). For each subset of buildings which use gas, 24 linear regression models are fitted – 12 months multiplied with 2 energy types (electricity and gas). Thus, for a given set of explanatory variables, the mean of electricity or gas use in each month can be estimated separately.

The candidates for explanatory variables are as follows – floor area, number of stories, approval year for use of building (for example, a value 2000 means that the building has been used since year 2000), category of building structure, and category of roof structure. The category of building structure includes ferroconcrete, steel-concrete, steel-frame, brick, cement block, timber, etc. The category of roof structure includes ferroconcrete, slate, tile, etc. Among these candidates, variables to be used for fitting regression models should be determined.

A one-variable regression model including floor area as the only explanatory variable has been considered as the base model. Then, other regression models with additional explanatory variables and interaction terms between floor area and each of the additional explanatory variables have been compared with the base model, in terms of explanatory power (adjusted R^2). The interaction terms can reflect effects of the additional explanatory variables to the intercept and slope of the linear relationship between monthly energy use and floor. For demonstration, the subset of 2,326 office buildings using gas with floor area less than 3,000 m^2 has been selected.

According to the demonstration, number of stories and approval year have been found to enhance explanatory power of the regression model. However, categories of building and roof structures have been found not to enhance explanatory power. Table 2 shows the values of adjusted R^2 for some selected months, corresponding to six cases – i) floor area only (base model); ii) floor area and number of stories; iii) floor area and approval year; iv) floor area and categories of building and roof structures; v) floor area, number of stories, and approval year; vi) all the explanatory variables mentioned above. Compared to the base model, the cases with number of stories or approval year showed greater adjusted R^2. However, the case with categories of structure but without number of stories and approval year showed little improvement in adjusted R^2.

Variables	Case i	Case ii	Case iii	Case iv	Case v	Case vi
Floor area	O	O	O	O	O	O
Number of stories	X	O	X	X	O	O
Approval year	X	X	O	X	O	O
Building structure	X	X	X	O	X	O
Roof structure	X	X	X	O	X	O
Adjusted R^2	Case i	Case ii	Case iii	Case iv	Case v	Case vi
Electricity, January	0.436	0.537	0.563	0.438	0.573	0.576
Electricity, May	0.474	0.526	0.544	0.465	0.540	0.540
Electricity, August	0.525	0.571	0.587	0.524	0.590	0.591
Gas, January	0.348	0.445	0.433	0.353	0.458	0.460
Gas, May	0.240	0.283	0.288	0.247	0.296	0.300

Table 2. Adjusted R^2 of linear regression models each with different response variable (energy use in some selected months) and different set of explanatory variables. O and X denote inclusion and exclusion of the corresponding variable in the regression model, respectively.

Adding number of stories and approval year enhances explanatory power of the model because it makes the model reflect the following aspects – i) heating, ventilation, and air conditioning demand related to surface-volume ratio which is usually higher for tall buildings [21], ii) occupancy rate of the buildings due to business and commercial use which is usually higher for short buildings [18], iii) energy performance of electric appliances and insulation which is usually better for recently built buildings. Meanwhile, categories of building and roof structure could not enhance explanatory power in this study, because most of the buildings belong to one category of building structure and one roof structure. Depending on the building use, about 80~95% of buildings belong to ferroconcrete building and roof. Due to the imbalance of the categorical data, it is hard to estimate the average difference of energy use between different structures, resulting in little enhancement of explanatory power by adding category of structure to the regression models.

Consequently, three features have been adopted as the explanatory variables in this study – floor area, number of stories, approval year. Also, the interactions between floor area and number of stories, and between floor area and approval year have been included. Categories of structure have been excluded because they have little positive impact on explanatory power, and because the categorical variables make the model too complex due to many binary indicator variables.

Although the number of stories is expected to increase with increasing floor area, multicollinearity problem is not expected. For example, the variation inflation factors of floor area, number of stories, and approval year in the model for electricity use in January without interaction terms are 1.447, 2.417, and 1.844, respectively, which are below 5.0 (which is the rule of thumb for potential multicollinearity).

[2.2.3 Statistical test for subset division]

To explain the regression-based statistical test, notations of the data are presented. Denote electricity and gas use in month m in building i as $y_i^{elec,m}$ and $y_i^{gas,m}$, respectively. Then, 12-dimensional column vectors $y_i^{elec} = \left[y_i^{elec,1}, \cdots, y_i^{elec,12}\right]^T$ and $y_i^{gas} = \left[y_i^{gas,1}, \cdots, y_i^{gas,12}\right]^T$ are the record of monthly electricity and gas use for a year, respectively. For the regression model corresponding to electricity use in mth month, the data vector of response variable is $y^{elec,m} = \left[y_1^{elec,m}, \cdots, y_N^{elec,m}\right]^T$ where N is the total number of data points. Also denote x_i^{area}, x_i^{story} and x_i^{year} as floor area, number of stories, and approval year of ith building, respectively. Then, the set of values of explanatory variables for ith data point is a six-dimensional vector $x_i = \left[1, x_i^{area}, x_i^{story}, x_i^{area}x_i^{story}, x_i^{year}, x_i^{area}x_i^{year}\right]^T$ (where 1 is added to estimate the intercept of the model), and the data matrix of explanatory variables is $X = [x_1, \cdots, x_N]^T$. The linear regression model for electricity use in mth month is presented as $y^{elec,m} = X\beta^{elec,m} + \epsilon^{elec,m}$, where $\beta^{elec,m}$ is the model coefficient vector and $\epsilon^{elec,m} = \left[\epsilon_1^{elec,m}, \cdots, \epsilon_N^{elec,m}\right]^T$ is the error vector. The value of $\beta^{elec,m}$ can be estimated as $\hat{\beta}^{elec,m} = (X^TX)^{-1}X^Ty^{elec,m}$ by solving ordinary least squares problem, which aims to minimize the sum of squared errors $(\epsilon^{elec,m})^T\epsilon^{elec,m}$. Using $\hat{\beta}^{elec,m}$, residual vector $\hat{\epsilon}^{elec,m} = y^{elec,m} - X\hat{\beta}^{elec,m}$ and residual sum of squares $SSR^{elec,m} = (\hat{\epsilon}^{elec,m})^T\hat{\epsilon}^{elec,m}$ can also be computed.

Suppose that partitioning $y^{elec,m}$ and X into $\left[\left(y_A^{elec,m}\right)^T \left(y_B^{elec,m}\right)^T\right]^T$ and $[X_A^T\ X_B^T]^T$, respectively, is of interest. If partitioned, two separate regression models $y_A^{elec,m} = X_A\beta_A^{elec,m} + \epsilon_A^{elec,m}$ and $y_B^{elec,m} = X_B\beta_B^{elec,m} + \epsilon_B^{elec,m}$ can be constructed. If the true values of $\beta_A^{elec,m}$ and $\beta_B^{elec,m}$ are the same, the partitioning is meaningless since a single combined regression model $y^{elec,m} = X\beta^{elec,m} + \epsilon^{elec,m}$ would be sufficient to explain the whole data. On the contrary, the partitioning is necessary if the true values of $\beta_A^{elec,m}$ and $\beta_B^{elec,m}$ are different. Thus, the null hypothesis of the test is $\beta_A^{elec,m} = \beta_B^{elec,m}$ while the alternative hypothesis is $\beta_A^{elec,m} \neq \beta_B^{elec,m}$. The null hypothesis can be viewed as a set of equality restrictions to the model coefficients. From this view, $SSR_R^{elec,m}$ is defined as the residual sum of squares of the single combined model, the subscript R means

restricted. In the similar sense, $SSR_U^{elec,m}$ is defined as the sum of the two residual sum of squares of the two models each for one partition, where the subscript U means unrestricted. Then, the test statistic (which approximately follows F distribution under the null hypothesis) can be computed as in Equation 1 [22].

$$\frac{\left(SSR_R^{elec,m} - SSR_U^{elec,m}\right)/r}{SSR_U^{elec,m}/(N-k)} \sim F(r, N-k) \qquad (1)$$

where r is the number of restrictions, and k is the sum of the number of parameters in the separate regression models for each partition. The null hypothesis is rejected if the value of test statistic is over the critical value for a given significance level.

If a building dataset is partitioned based on use of gas, two partitions are made (using gas, not using gas). r and k are 6 and 12, respectively, since $\beta^{elec,m}$ is a six-dimensional vector. If a building dataset is partitioned based on floor area interval, four partitions are made (under 3,000 m^2, 3,000 m^2 to 10,000 m^2, 10,000 m^2 to 100,000 m^2, and over 100,000 m^2). However, for the test in this study, only the first three partitions are considered for the test because the last partition contains only a few or even no buildings depending on the building use. For three partitions, r and k are 12 and 18, respectively.

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Office	276.32	262.38	181.13	77.85	42.95	46.74	51.15	44.08	46.92	33.87	68.44	161.45
Neighborhood(I)	246.41	253.95	135.36	49.44	35.54	34.16	39.63	44.26	41.65	37.45	42.28	111.58
Neighborhood(II)	378.46	414.09	200.75	42.81	25.93	31.07	46.55	63.96	45.32	37.25	35.75	149.85
Hospital	1.04	1.71	1.36	1.42	1.40	1.31	1.20	1.07	1.52	1.86	1.39	1.37
Sales	8.65	7.31	7.89	9.02	8.60	9.56	8.69	8.28	8.03	8.24	8.66	8.35
Welfare	10.04	14.03	11.80	14.31	10.98	10.65	9.75	11.22	11.99	11.49	11.41	11.26
Lodging	64.35	63.19	63.14	55.05	36.58	17.94	7.81	3.30	5.53	10.81	48.40	64.07
Education	16.42	18.94	14.73	8.54	8.28	8.93	9.43	7.73	6.22	5.93	8.84	11.18
Religious	8.62	8.84	8.31	9.38	9.70	8.19	6.12	4.95	4.82	6.64	8.34	8.50
Cultural	2.53	3.01	2.55	2.86	3.06	2.49	2.09	1.79	2.09	2.26	2.85	3.60

Table 3. Test statistics for the hypothesis of dividing subsets with respect to use of gas, computed for the set of each building use with floor area under 3,000 m^2.

Table 3 shows that the null hypothesis of partitioning based on use of gas is rejected for most of the cases, which implies that it is necessary to partition the building dataset based on use of gas. Most of the values of test statistic computed for subsets, each corresponding to one of the buildings uses and floor area under 3,000 m^2, are over the critical value for 1% significance level $F_{0.01}(6, \infty) = 2.803$. The values of test statistic are especially higher for winter, which supports the expected difference in magnitude of electricity use in winter depending on use of gas heating.

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Office	10.97	12.56	8.06	8.40	9.53	9.78	9.46	8.92	8.92	9.02	9.52	12.04
Neighborhood(I)	746.29	617.97	602.20	636.65	735.71	777.12	889.17	619.81	606.53	584.62	636.23	705.95
Neighborhood(II)	537.76	552.68	522.78	468.37	458.24	528.57	574.42	384.83	384.62	423.20	540.94	706.01
Hospital	10.93	9.89	8.62	8.62	8.78	8.03	8.15	6.51	6.36	7.34	10.06	9.84
Sales	8.05	7.84	8.09	8.09	8.18	8.36	8.45	7.96	8.13	8.14	8.22	8.30
Welfare	62.38	76.22	65.22	91.64	86.21	69.55	56.81	63.63	88.15	89.17	74.26	62.49
Lodging	6.81	9.24	12.79	15.78	13.55	11.58	11.08	11.34	13.53	17.02	14.34	10.88
Education	9.80	10.90	8.55	8.20	6.62	6.04	5.99	6.34	6.68	6.27	7.78	9.05
Religious	27.72	22.99	30.39	41.89	44.90	61.37	53.32	33.80	52.23	50.30	38.37	34.37
Cultural	34.58	22.99	36.72	38.41	53.60	46.67	33.23	27.40	14.69	28.56	36.97	37.75

Table 4. Test statistics for the hypothesis of dividing subsets with respect to floor area interval, computed for the set of each building use with gas use.

Table 4 shows that the null hypothesis of partitioning based on floor area interval is rejected for most of the cases, which implies that it is necessary to partition the building dataset based on floor area interval. Most of the values of test statistic computed for subsets, each corresponding to one of the buildings uses and buildings using gas, are over the critical value for 1% significance level $F_{0.01}(12, \infty) = 2.187$.

It is noted that the statistical test has been done using the pre-processed dataset cleaned by the process explained in Section 2.3.

[2.3. Data pre-processing]

There is an issue of data quality of the raw dataset of monthly energy use in individual buildings because there are many abnormal data points which have missing or unrealistic values. In this study, abnormal data points are deleted from the dataset because the number of rows of the total dataset is large enough (order of 10^4). The points with missing numbers, points with abnormal seasonal patterns, and points with abnormal magnitude of energy use have been deleted.

[2.3.1. Data points with missing numbers]

The detailed criteria of deletion are as follows:

i) Any of the 12 values of monthly energy use in the building is missing.

ii) Any of the 3 values of monthly gas use in the building in winter (January, February, and

December) is missing or abnormally low if any of the values of gas use in other months is positive, because it is unusual that a building which uses gas during spring, summer or fall does not use gas or use only a small amount of gas in winter. It is noted that a data point with no record of gas use in all months is regarded as a building not using gas and preserved.

iii) Any of the values of monthly energy use is negative.

iv) Any of the values of explanatory variables is missing.

After applying the criteria to the dataset of buildings in Seoul for year 2021, 79,427 data points have been preserved.

[2.3.2. Data points with abnormal seasonal patterns]

Data points with abnormal seasonal patterns of energy use, which is far different from the exemplary pattern shown in Figure 1, have been deleted. Figure 3 shows examples of the abnormal seasonal patterns of monthly energy use in buildings. The cause of such abnormal patterns may be measurement error, or relatively rapid increasing or decreasing occupants. It is noted that the vertical axis in Figure 3 is the fraction of annual energy use for each month, to investigate only the shape of the seasonal patterns after control of the effect of building size on energy use.

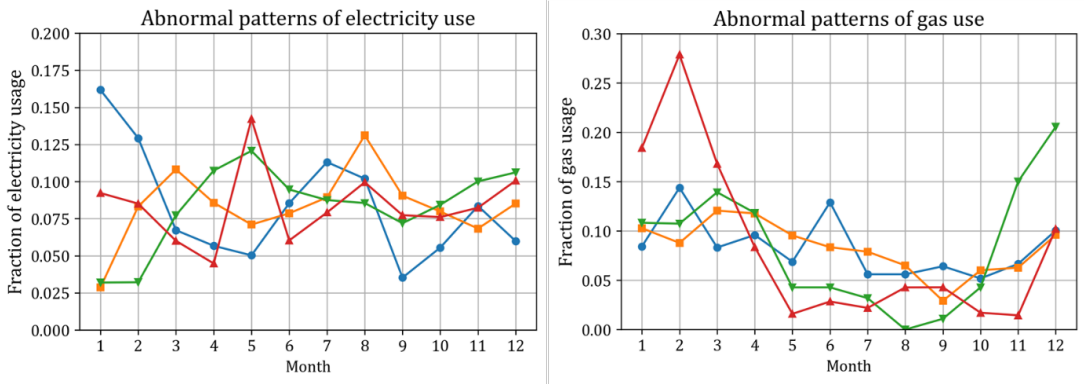

Figure 3. Abnormal seasonal patterns of monthly energy use in individual buildings (Left: electricity, Right: gas).

To apply the method for identification of data points with abnormal seasonal patterns, the dataset of monthly energy use in individual buildings has been transformed into a dataset of portion of annual energy use for each month. Dividing y_i^{elec} with its absolute-value norm $\left|y_i^{elec}\right|_1$,

the obtained vector $\tilde{y}_i^{elec} = y_i^{elec}/|y_i^{elec}|_1$ represents fraction of annual electricity use for each month. $\tilde{y}_i^{gas} = y_i^{gas}/|y_i^{gas}|_1$, which represents faction of annual gas use for each month, can be obtained in the same way. By aggregation of $\tilde{y}_i^{elec}$ and $\tilde{y}_i^{gas}$ of all buildings, new $N \times 12$ data matrices $\tilde{Y}^{elec} = [\tilde{y}_1^{elec}, \cdots \tilde{y}_N^{elec}]^T$ and $\tilde{Y}^{gas} = \left[\tilde{y}_1^{gas}, \cdots, \tilde{y}_N^{gas}\right]^T$, representing the transformed dataset, can be obtained.

A data point with abnormal seasonal pattern of electricity use can be considered as a point which is far from the cluster of points in the 12-dimensional vector space composed of row vectors in $\tilde{Y}^{elec}$. A common approach to find such remote points in the vector space is to compute diagonal elements of the matrix $\tilde{Y}^{elec}\left(\left(\tilde{Y}^{elec}\right)^T\tilde{Y}^{elec}\right)^{-1}\left(\tilde{Y}^{elec}\right)^T$ (often called as hat matrix) [23]. ith diagonal element $\tilde{h}_{ii}$ of the hat matrix can be written as in Equation 2.

$$\tilde{h}_{ii} = \left(\tilde{y}_i^{elec}\right)^T\left(\left(\tilde{Y}^{elec}\right)^T\tilde{Y}^{elec}\right)^{-1}\tilde{y}_i^{elec} \qquad (2)$$

A rule of thumb is to consider ith point as a remote point if $\tilde{h}_{ii}$ is larger than $2k/N$ where k is the dimension of the vector space (12 in this study). The points considered to be remote points following the rule of thumb were found to have abnormal seasonal patterns of electricity use as shown in Figure 3 and deleted from the dataset. The points which have abnormal seasonal patterns of gas use as shown in Figure 3 have been deleted in the same way. After deleting such points, the number of data points has been reduced from 79,427 to 68,135.

[2.3.3. Data points with abnormal magnitude of energy use]

Data points with unusually low or high energy use relative to other buildings with similar size may have a noticeable impact on the model coefficients, resulting in estimates of the coefficients far from its true value. Such points are often called the influential points, and the red hollow circles in Figure 2 are the points that are suspected to be influential points. The cause of such influential points may be measurement error, or unusual type of buildings (for example, energy use records of subway stations were found to be very high relative to the floor area of the station).

A common approach to find influential points is to compute Cook's distance of every ith point, which is a measure of the squared distance between the estimated coefficient vector based on all points and the estimated coefficient vector obtained by deleting ith point [24]. Cook's distance for ith point can be computed as in Equation 3.

$$D_i^{elec,m} = \frac{\left(\hat{\beta}^{elec,m} - \hat{\beta}_{-i}^{elec,m}\right)^T X^T X\left(\hat{\beta}^{elec,m} - \hat{\beta}_{-i}^{elec,m}\right)}{k \cdot MSR^{elec,m}} = \frac{\hat{\epsilon}_i^{elec,m} h_{ii}}{k \cdot MSR^{elec,m}(1-h_{ii})^2} \qquad (3)$$

where $\hat{\beta}_{-i}^{elec,m}$ is the estimates of coefficients obtained by deleting ith point, k is number of coefficients (6 in this study), $MSR^{elec,m} = SSR^{elec,m}/(N-k)$ is the regression mean square of the model containing all points, and h_{ii} is ith diagonal element of $X(X^TX)^{-1}X^T$. It is not required to solve ordinary least squares problem $N+1$ times to obtain Cook's distance of every point. By the term in the right side of equation 2, Cook's distance of every point can be obtained by one computation of $X(X^TX)^{-1}X^T$ and solving ordinary least squares problem once.

Computation of Cook's distance have been applied to each subset of Table 1 since computation of Cook's distance requires regression models which are fitted for each of the subsets separately. For each subset, $D_i^{elec} = \max\left(D_i^{elec,1}, \cdots, D_i^{elec,12}\right)$ is computed for every point. Then, the point which corresponds to the highest value of D_i^{elec} is deleted because at least one of 12 monthly electricity uses in the corresponding building is abnormal in magnitude. This procedure is repeated until a pre-determined number of points are deleted from the dataset. If the subset is the set of buildings using gas, then buildings with abnormal monthly gas use in magnitude are also deleted, following the same procedure. In this study, the number of points to be deleted from each subset by this procedure has been pre-determined as two percent of the data points in the subset.

[3. Estimation of moment conditions]

[3.1. Estimation based on linear regression models]

To establish a joint probability model for monthly energy uses of a certain building given its features (floor area, number of stories, and approval year in this study) covariance between the error terms of two different regression models should be investigated. The linear regression model for electricity use in mth month can be written in pointwise form as Equation 4.

$$y_i^{elec,m} = \beta_0^{elec,m} + \beta_1^{elec,m} x_i^{area} + \beta_2^{elec,m} x_i^{story} + \beta_3^{elec,m} x_i^{area} x_i^{story} + \beta_4^{elec,m} x_i^{year} + \beta_5^{elec,m} x_i^{area} x_i^{year} + \epsilon_i^{elec,m} \quad (4)$$

Then, $\epsilon_i^{elec,1}$ and $\epsilon_i^{elec,2}$ are expected to be positively correlated because a building which uses more electricity in January compared to other buildings with similar size is expected to use more electricity in February compared to other buildings with similar size as well. Meanwhile, $\epsilon_i^{elec,1}$ and $\epsilon_i^{gas,1}$ are expected to be negatively correlated because electricity and gas are substitutes for heating in winter.

A common approach to estimate coefficients of many linear regression models simultaneously considering covariance between error terms of these regression models is Seemingly Unrelated Regression (SUR) [25]. SUR aggregates all of 24 regression models (12 for electricity, and 12 for gas) to make a combined regression model, as shown in matrix form in Equation 5.

$$\begin{bmatrix} y^{elec,1} \\ y^{elec,2} \\ \vdots \\ y^{elec,12} \\ y^{gas,1} \\ \vdots \\ y^{gas,12} \end{bmatrix} = \begin{bmatrix} X^{elec,1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & X^{elec,2} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X^{elec,12} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & X^{gas,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & X^{gas,12} \end{bmatrix} \begin{bmatrix} \beta^{elec,1} \\ \beta^{elec,2} \\ \vdots \\ \beta^{elec,12} \\ \beta^{gas,1} \\ \vdots \\ \beta^{gas,12} \end{bmatrix} + \begin{bmatrix} \epsilon^{elec,1} \\ \epsilon^{elec,2} \\ \vdots \\ \epsilon^{elec,12} \\ \epsilon^{gas,1} \\ \vdots \\ \epsilon^{gas,12} \end{bmatrix} \quad (5)$$

By solving generalized least squares problem for Equation 5, estimates of coefficients with consideration of covariance between error terms can be obtained. However, solving generalized least squares problem is more complicated than solving ordinary least squares problem because the exact structure of covariance is generally not known before solving.

	$\hat{\beta}_0^{elec,m}$	$\hat{\beta}_1^{elec,m}$	$\hat{\beta}_2^{elec,m}$	$\hat{\beta}_3^{elec,m}$	$\hat{\beta}_4^{elec,m}$	$\hat{\beta}_5^{elec,m}$	$\hat{\sigma}^{elec,m}$
JAN	-166973	459.3	-668.8	-0.01	83.66	-0.22	6919
FEB	-112161	408.1	-558.9	-0.03	56.20	-0.20	6464
MAR	-82987	302.0	-400.1	-0.01	41.54	-0.15	4932
APR	-7807	203.3	-285.5	-0.01	3.99	-0.10	4441
MAY	-1387	171.8	-251.4	-0.01	0.78	-0.08	4293
JUN	-3867	197.7	-324.2	-0.01	2.03	-0.09	5059
JUL	-30814	263.3	-434.4	0.00	15.53	-0.13	6114
AUG	-7230	280.5	-454.8	0.02	3.87	-0.13	6966
SEP	32561	195.4	-348.8	-0.03	-16.08	-0.09	5857
OCT	57325	122.4	-242.2	-0.04	-28.56	-0.06	4864
NOV	-16933	203.4	-283.3	-0.02	8.47	-0.10	4661
DEC	-124909	357.1	-491.2	0.02	62.56	-0.17	5534

Table 5. Estimates of coefficients and standard error of the 12 regression models for monthly electricity uses, fitted for the subset of office buildings with floor area under 3,000 m^2 using gas.

	$\hat{\beta}_0^{gas,m}$	$\hat{\beta}_1^{gas,m}$	$\hat{\beta}_2^{gas,m}$	$\hat{\beta}_3^{gas,m}$	$\hat{\beta}_4^{gas,m}$	$\hat{\beta}_5^{gas,m}$	$\hat{\sigma}^{gas,m}$
JAN	456067	-572.0	1145.6	0.33	-225.82	0.29	12966
FEB	357337	-375.8	900.4	0.27	-177.03	0.19	10889
MAR	273736	-273.7	749.6	0.09	-136.17	0.14	8254
APR	199869	-202.7	424.0	0.06	-99.26	0.10	5898
MAY	224357	-207.1	351.9	-0.03	-111.95	0.10	4249
JUN	183681	-153.8	196.0	-0.05	-91.77	0.08	3591
JUL	131569	-67.9	112.0	-0.05	-65.74	0.04	3204
AUG	175755	-139.1	111.9	-0.03	-87.84	0.07	3069
SEP	129230	-73.7	91.3	-0.04	-64.48	0.04	2974
OCT	231163	-236.7	208.8	-0.04	-115.08	0.12	4291
NOV	267671	-309.8	536.8	0.04	-132.65	0.16	6897
DEC	325000	-410.1	693.7	0.22	-160.42	0.21	10102

Table 6. Estimates of coefficients and standard error of the 12 regression models for monthly gas uses, fitted for the subset of office buildings with floor area under 3,000 m^2 using gas.

Fortunately, the generalized least squares estimators of SUR in this study is equivalent to the ordinary least squares estimators of each of the 24 regression models obtained separately, because all 24 models contain the same set of explanatory variables [25]. For example, Tables 5 and 6 shows the estimates of coefficients and standard error of the 24 models for the subset of 2,326 office buildings with floor area under 3,000 m^2 using gas, obtained by solving least squares of each of 24 regression models separately. Given the floor area and number of stories of a certain office building with floor area under 3,000 m^2 using gas, the mean vector of monthly energy use in the building can be determined by the estimates of coefficients. For other subsets, different estimates of coefficients would be obtained.

Covariance and correlation between error terms of different regression models can be estimated by computation of sample covariance matrix and sample correlation matrix of the residuals. Table 5 shows the sample correlation matrix of error terms of the 24 models for the subset of office buildings with floor area under 3,000 m^2 using gas. Table 7(a) shows that error terms corresponding to electricity use in different months are strongly positively correlated, even when the effects of size, height, and age of buildings have been controlled. This result supports the expectation on positive correlation between $\epsilon_i^{elec,1}$ and $\epsilon_i^{elec,2}$. Table 7(b) shows that error terms corresponding to gas use in adjacent different months are also strongly positively correlated. Table 7(c) shows that error terms corresponding to electricity use and gas use in winter are negatively correlated. This result supports the expectation on negative correlation

between $\epsilon_i^{elec,1}$ and $\epsilon_i^{gas,1}$.

(a)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
JAN	1.000	0.967	0.952	0.873	0.820	0.822	0.819	0.797	0.787	0.782	0.851	0.927
FEB	0.967	1.000	0.975	0.893	0.828	0.816	0.808	0.810	0.815	0.800	0.848	0.911
MAR	0.952	0.975	1.000	0.950	0.894	0.875	0.861	0.860	0.866	0.859	0.898	0.931
APR	0.873	0.893	0.950	1.000	0.977	0.950	0.920	0.917	0.929	0.936	0.950	0.915
MAY	0.820	0.828	0.894	0.977	1.000	0.980	0.945	0.928	0.936	0.955	0.959	0.897
JUN	0.822	0.816	0.875	0.950	0.980	1.000	0.979	0.947	0.941	0.954	0.956	0.904
JUL	0.819	0.808	0.861	0.920	0.945	0.979	1.000	0.963	0.940	0.938	0.933	0.897
AUG	0.797	0.810	0.860	0.917	0.928	0.947	0.963	1.000	0.973	0.951	0.920	0.871
SEP	0.787	0.815	0.866	0.929	0.936	0.941	0.940	0.973	1.000	0.977	0.937	0.873
OCT	0.782	0.800	0.859	0.936	0.955	0.954	0.938	0.951	0.977	1.000	0.964	0.887
NOV	0.851	0.848	0.898	0.950	0.959	0.956	0.933	0.920	0.937	0.964	1.000	0.955
DEC	0.927	0.911	0.931	0.915	0.897	0.904	0.897	0.871	0.873	0.887	0.955	1.000
(b)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
JAN	1.000	0.924	0.819	0.718	0.607	0.451	0.319	0.349	0.338	0.516	0.759	0.878
FEB	0.924	1.000	0.894	0.817	0.686	0.527	0.373	0.401	0.371	0.470	0.672	0.792
MAR	0.819	0.894	1.000	0.895	0.780	0.637	0.471	0.502	0.443	0.441	0.564	0.670
APR	0.718	0.817	0.895	1.000	0.912	0.808	0.660	0.661	0.627	0.584	0.611	0.636
MAY	0.607	0.686	0.780	0.912	1.000	0.907	0.756	0.785	0.727	0.665	0.612	0.572
JUN	0.451	0.527	0.637	0.808	0.907	1.000	0.873	0.890	0.852	0.708	0.564	0.476
JUL	0.319	0.373	0.471	0.660	0.756	0.873	1.000	0.730	0.959	0.603	0.525	0.405
AUG	0.349	0.401	0.502	0.661	0.785	0.890	0.730	1.000	0.732	0.767	0.487	0.397
SEP	0.338	0.371	0.443	0.627	0.727	0.852	0.959	0.732	1.000	0.681	0.603	0.466
OCT	0.516	0.470	0.441	0.584	0.665	0.708	0.603	0.767	0.681	1.000	0.815	0.713
NOV	0.759	0.672	0.564	0.611	0.612	0.564	0.525	0.487	0.603	0.815	1.000	0.920
DEC	0.878	0.792	0.670	0.636	0.572	0.476	0.405	0.397	0.466	0.713	0.920	1.000
(c)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
JAN	-0.343	-0.309	-0.226	-0.150	-0.088	0.037	0.123	0.093	0.112	-0.048	-0.209	-0.291
FEB	-0.336	-0.301	-0.218	-0.149	-0.086	0.040	0.124	0.096	0.109	-0.058	-0.213	-0.288
MAR	-0.259	-0.224	-0.145	-0.084	-0.035	0.073	0.149	0.118	0.132	-0.017	-0.152	-0.218
APR	-0.124	-0.097	-0.028	0.014	0.051	0.138	0.191	0.169	0.175	0.053	-0.045	-0.095
MAY	-0.062	-0.037	0.028	0.070	0.107	0.188	0.224	0.206	0.209	0.099	0.013	-0.038
JUN	-0.068	-0.041	0.023	0.068	0.107	0.199	0.236	0.217	0.224	0.108	0.013	-0.039
JUL	-0.056	-0.030	0.035	0.072	0.110	0.207	0.249	0.230	0.233	0.111	0.016	-0.028
AUG	-0.019	0.006	0.063	0.097	0.133	0.228	0.267	0.246	0.253	0.125	0.043	0.007
SEP	-0.038	-0.014	0.044	0.082	0.116	0.207	0.250	0.223	0.239	0.103	0.024	-0.015
OCT	-0.028	-0.006	0.053	0.099	0.137	0.219	0.251	0.230	0.245	0.129	0.047	0.000
NOV	-0.111	-0.083	-0.014	0.035	0.075	0.165	0.210	0.190	0.198	0.075	-0.023	-0.081
DEC	-0.255	-0.223	-0.142	-0.077	-0.019	0.098	0.163	0.137	0.152	0.004	-0.133	-0.210

Table 7. Sample correlation matrix for the residuals of the linear regression models for the subset of office buildings with floor area under 3,000 m^2 using gas ((a): between $\epsilon_i^{elec,p}$ and $\epsilon_i^{elec,q}$, (b): between $\epsilon_i^{gas,p}$ and $\epsilon_i^{gas,q}$, (c): between $\epsilon_i^{elec,p}$ (row) and $\epsilon_i^{gas,q}$ (column), where p and q are month indices).

[3.2. Issues of non-constant covariance (heteroskedasticity)]

In Section 3.1, constant variance and covariance of error terms in each model has been assumed. If this assumption is violated, the estimation of covariance matrix based on as presented in Section 3.1 becomes invalid. Thus, it should be checked whether the variance and covariance are constant with all explanatory variables (homoscedastic) or they vary with at least one varying explanatory variable (heteroskedastic).

[3.2.1. Existence of heteroskedasticity]

Figure 4, the residual plot, shows that variance of monthly energy use is not constant but increasing with increasing floor area. This heteroskedasticity has not been considered for obtaining sample covariance matrix in Section 3.1. Assuming homoskedasticity, the grey regions in Figure 4 represent the bands of $x_i^T \hat{\beta}^{elec,1} \pm 2.58\hat{\sigma}^{elec,1}$ and $x_i^T \hat{\beta}^{gas,1} \pm 2.58\hat{\sigma}^{gas,1}$, where $\hat{\sigma}^{elec,1}$ is the constant standard error of the regression model corresponding to electricity use in January. The band includes the region of large magnitude of residual with small floor area (depicted as dashed triangles), where there are few data points located in that region.

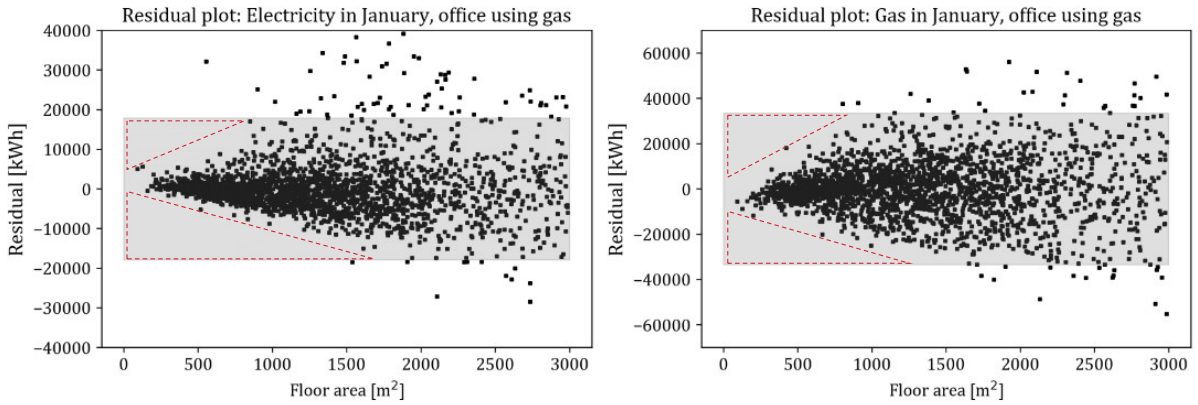

Figure 4. Residual plots for the linear regression model corresponding to energy uses in January, for the subset of office buildings with floor area under 3,000 m² using gas (Left: electricity, Right: gas). The grey areas denote the bands of $x_i^T \hat{\beta}^{elec,1} \pm 2.58\hat{\sigma}^{elec,1}$ and $x_i^T \hat{\beta}^{gas,1} \pm 2.58\hat{\sigma}^{gas,1}$, which capture heteroskedasticity of data poorly. The dashed triangles denote the region that the band includes but actual points are not located.

Thus, variance of energy use in small buildings will be overestimated so that unrealistically small or large amount of energy use can be sampled from the joint probability model based on assumption of constant variance. In contrast, variance of energy use in large buildings will be

underestimated. Despite of such problem, issue of heteroskedasticity has not been considered in the previous studies on statistical estimation of building energy use. The structure of heteroskedasticity should be modeled to correct the estimation of covariance and to make a correct joint probability model.

[3.2.2. Heteroskedasticity modeling]

A common approach to estimate structure of heteroskedasticity in a linear regression model is to make an auxiliary regression model, where the response variable is the squared residual, and the explanatory variables are first and second order terms of explanatory variables which causes heteroskedasticity (floor area in this study) [26]. For the regression model corresponding to electricity use in pth month, the auxiliary regression model can be set up as in Equation 6.

$$\left(\hat{\epsilon}_i^{elec,p}\right)^2 = \alpha_0^{elec,p} + \alpha_1^{elec,p} x_i^{area} + \alpha_2^{elec,p} (x_i^{area})^2 + v_i^{elec,p} \quad (6)$$

where $v_i^{elec,p}$ is the error term of the auxiliary model. By estimation of the coefficients of the auxiliary model, variance can be estimated as a function of x_i^{area}, as in Equation 7.

$$(\hat{\sigma}^{elec,p})^2 = \hat{\alpha}_0^{elec,p} + \hat{\alpha}_1^{elec,p} x_i^{area} + \hat{\alpha}_2^{elec,p} (x_i^{area})^2 \quad (7)$$

where $(\hat{\sigma}^{elec,p})^2$ denotes the estimate of error variance, and $\hat{\alpha}_0^{elec,p}$, $\hat{\alpha}_1^{elec,p}$, $\hat{\alpha}_2^{elec,p}$ denote the estimate of coefficients of the auxiliary model. $(\hat{\sigma}^{gas,p})^2$ as the function of floor area can also be obtained in the same way.

The explained approach can be extended to estimate the heteroskedasticity structure of covariance between error terms of two different regression models as a function of the explanatory variables [27]. For the two regression models corresponding to electricity uses in pth and qth months, the auxiliary regression model can be set up as in Equations 8 and 9.

$$\hat{\epsilon}_i^{elec,p} \hat{\epsilon}_i^{elec,q} = \alpha_0^{elec,(p,q)} + \alpha_1^{elec,(p,q)} x_i^{area} + \alpha_2^{elec,(p,q)} (x_i^{area})^2 + v_i^{elec,(p,q)} \quad (8)$$

$$\hat{\sigma}_{(p,q)}^{e,e} = \hat{\alpha}_0^{elec,(p,q)} + \hat{\alpha}_1^{elec,(p,q)} x_i^{area} + \hat{\alpha}_2^{elec,(p,q)} (x_i^{area})^2 \quad (9)$$

where $\hat{\sigma}_{(p,p)}^{e,e} = (\hat{\sigma}^{elec,p})^2$ and e in the superscript denotes electricity. $\hat{\sigma}_{(p,q)}^{g,g}$ and $\hat{\sigma}_{(p,q)}^{e,g}$ can also be obtained in the same way (where g in the superscript denotes gas).

However, estimate of covariance by Equation 9 may produce unrealistic value of covariance, such as negative variance and negative correlation between error terms of regression models corresponding to electricity use in January and February. For the subset of office buildings with floor area under 3,000 m^2 using gas, the variance of $\epsilon_i^{elec,1}$ and covariance between $\epsilon_i^{elec,1}$ and

$\epsilon_i^{elec,2}$ have been estimated as $\hat{\sigma}_{(1,1)}^{e,e} = -39278300 + 75969x_i^{area} - 5.99(x_i^{area})^2$ and $\hat{\sigma}_{(1,2)}^{e,e} = -35607600 + 70291x_i^{area} - 6.37(x_i^{area})^2$, respectively. Both estimates become negative if x_i^{area} is lower than about 500 m^2, which are practically positive but incorrectly estimated.

To prevent unrealistic estimation of covariance by change of sign, Equations 8 and 9 have been modified to contain only the second order term of floor area in the right side, as in Equations 10 and 11.

$$\hat{\epsilon}_i^{elec,p}\hat{\epsilon}_i^{elec,q} = \alpha_{(p,q)}^{e,e}(x_i^{area})^2 + v_{(p,q),i}^{e,e} \quad (10)$$

$$\hat{\sigma}_{(p,q)}^{e,e} = \hat{\alpha}_{(p,q)}^{e,e}(x_i^{area})^2 \quad (11)$$

The estimate of covariance matrix, constructed by aggregation of all estimates of covariance computed by Equation 11, is generally not positive semidefinite. However, a covariance matrix must be positive semidefinite by its properties. Thus, a positive semidefinite matrix nearest to the estimate of covariance matrix should be computed to be used as the covariance of the joint probability model of monthly energy uses.

The nearest positive semidefinite matrix can be obtained by eigen-decomposition. Denote the estimate of covariance matrix obtained by Equation 11 as $\hat{\Sigma}$. $\hat{\Sigma}$ is generally not positive semidefinite, but it is real-valued and symmetric. Thus, it can be decomposed as $\hat{\Sigma} = VDV^T$ where V is a square matrix containing eigenvectors of $\hat{\Sigma}$ as its columns, and D is a diagonal matrix containing eigenvalues of $\hat{\Sigma}$ as its diagonal elements. Defining a new matrix D_+ which is obtained by replacing negative elements of D with zeros, the nearest positive semidefinite matrix $\hat{\Sigma}_+$ can be computed as $\hat{\Sigma}_+ = VD_+V^T$. Then, $\hat{\Sigma}_+$ is used as the covariance matrix of the joint probability model for monthly energy uses. Table 8 shows the values of elements in $\hat{\Sigma}_+$ for unit floor area, for the subset of office buildings with floor area under 3,000 m^2 using gas. The covariance matrix of the vector of monthly energy uses for a certain office building under 3,000 m^2 using gas can be obtained by multiplication of square of its floor area with the elements in Table 7. For other subsets, different estimates of covariance matrix would be obtained.

(a)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
JAN	20.09	17.95	13.43	10.79	9.75	11.64	14.04	15.64	12.98	10.65	11.25	14.90
FEB	17.95	17.16	12.68	10.22	9.13	10.75	12.87	14.86	12.54	10.16	10.43	13.55
MAR	13.43	12.68	9.91	8.28	7.51	8.77	10.45	11.99	10.12	8.29	8.40	10.53
APR	10.79	10.22	8.28	7.66	7.19	8.33	9.77	11.24	9.52	7.93	7.79	9.03
MAY	9.75	9.13	7.51	7.19	7.08	8.25	9.64	10.93	9.22	7.79	7.56	8.50
JUN	11.64	10.75	8.77	8.33	8.25	10.03	11.93	13.30	11.08	9.24	8.96	10.19
JUL	14.04	12.87	10.45	9.77	9.64	11.93	14.88	16.48	13.46	11.00	10.57	12.24
AUG	15.64	14.86	11.99	11.24	10.93	13.30	16.48	19.81	16.13	12.91	12.00	13.62
SEP	12.98	12.54	10.12	9.52	9.22	11.08	13.46	16.13	13.89	11.11	10.27	11.48
OCT	10.65	10.16	8.29	7.93	7.79	9.24	11.00	12.91	11.11	9.40	8.72	9.63
NOV	11.25	10.43	8.40	7.79	7.56	8.96	10.57	12.00	10.27	8.72	8.77	10.07
DEC	14.90	13.55	10.53	9.03	8.50	10.19	12.24	13.62	11.48	9.63	10.07	12.78
(b)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
JAN	67.19	52.25	34.19	21.11	12.65	7.77	4.75	4.92	4.79	10.62	26.61	45.54
FEB	52.25	48.07	32.40	20.95	12.41	8.01	4.90	5.07	4.53	8.10	19.43	34.08
MAR	34.19	32.40	27.17	17.25	10.65	7.24	4.61	4.71	4.01	5.58	11.95	21.26
APR	21.11	20.95	17.25	13.78	8.84	6.57	4.69	4.47	4.12	5.34	9.09	13.99
MAY	12.65	12.41	10.65	8.84	6.91	5.21	3.77	3.79	3.38	4.37	6.47	8.91
JUN	7.77	8.01	7.24	6.57	5.21	4.87	3.71	3.64	3.38	3.98	5.06	6.21
JUL	4.75	4.90	4.61	4.69	3.77	3.71	3.86	2.51	3.44	2.92	4.23	4.75
AUG	4.92	5.07	4.71	4.47	3.79	3.64	2.51	3.54	2.38	3.73	3.62	4.25
SEP	4.79	4.53	4.01	4.12	3.38	3.38	3.44	2.38	3.36	3.18	4.63	5.23
OCT	10.62	8.10	5.58	5.34	4.37	3.98	2.92	3.73	3.18	6.88	9.02	11.62
NOV	26.61	19.43	11.95	9.09	6.47	5.06	4.23	3.62	4.63	9.02	18.54	25.27
DEC	45.54	34.08	21.26	13.99	8.91	6.21	4.75	4.25	5.23	11.62	25.27	40.70
(c)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
JAN	-14.72	-11.42	-6.88	-3.58	-1.83	-0.23	0.57	0.31	0.52	-1.17	-4.91	-10.06
FEB	-13.15	-10.23	-6.17	-3.24	-1.61	-0.11	0.61	0.38	0.53	-1.09	-4.49	-9.03
MAR	-7.91	-6.01	-3.43	-1.71	-0.82	0.13	0.58	0.40	0.49	-0.51	-2.56	-5.42
APR	-3.69	-2.67	-1.20	-0.45	-0.08	0.51	0.73	0.61	0.64	0.06	-0.95	-2.49
MAY	-2.05	-1.31	-0.25	0.22	0.37	0.83	0.90	0.81	0.80	0.38	-0.26	-1.40
JUN	-2.69	-1.69	-0.44	0.19	0.39	1.02	1.11	1.01	0.99	0.47	-0.39	-1.75
JUL	-2.98	-1.80	-0.42	0.14	0.39	1.21	1.37	1.22	1.19	0.49	-0.50	-1.90
AUG	-2.20	-1.15	-0.04	0.49	0.66	1.60	1.73	1.51	1.57	0.75	-0.03	-1.19
SEP	-2.57	-1.56	-0.41	0.23	0.43	1.20	1.31	1.18	1.22	0.48	-0.37	-1.62
OCT	-1.63	-0.94	-0.01	0.52	0.62	1.13	1.13	1.06	1.07	0.65	0.06	-0.99
NOV	-3.76	-2.59	-1.05	-0.21	0.12	0.74	0.87	0.79	0.79	0.20	-0.86	-2.61
DEC	-8.94	-6.76	-3.73	-1.82	-0.79	0.32	0.71	0.55	0.64	-0.47	-2.77	-6.15

Table 8. Estimates of coefficients of squared floor area for estimation of covariance as a function of floor area ((a): $\hat{\alpha}^{e,e}_{(p,q)+}$, (b): $\hat{\alpha}^{g,g}_{(p,q)+}$, (c): $\hat{\alpha}^{e,g}_{(p,q)+}$). + in the subscript emphasizes that the resulting covariance matrix is positive semidefinite.

Figure 5 shows that the estimates of covariance from $\hat{\Sigma}_+$ represents the heteroskedasticity of the data well. Adding a subscript $+$ which emphasizes that the covariance matrix is positive semidefinite, the modified bands $x_i^T\hat{\beta}^{elec,1} \pm 2.58\hat{\alpha}^{e,e}_{(1,1)+}(x_i^{area})^2$ and $x_i^T\hat{\beta}^{gas,1} \pm 2.58\hat{\alpha}^{g,g}_{(1,1)+}(x_i^{area})^2$ (depicted as grey areas) capture the increasing variance well while they do not contain regions where no data point is located.

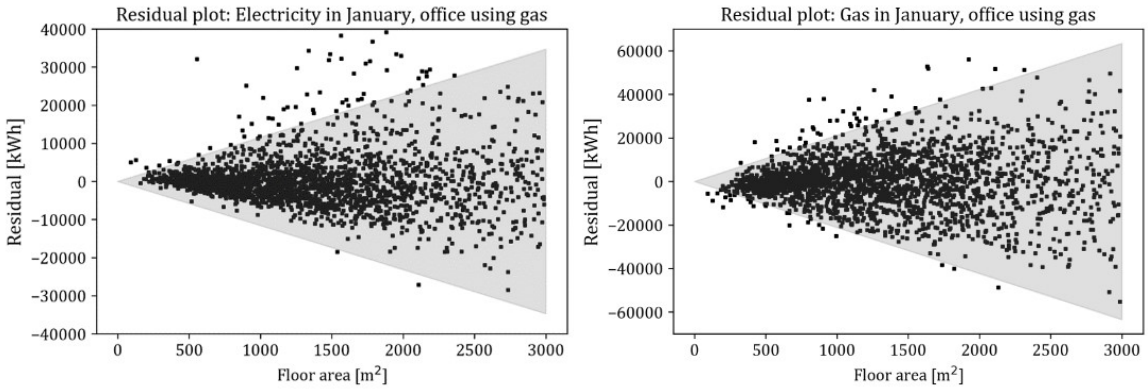

Figure 5. Residual plots for the linear regression model corresponding to energy uses in January, for the subset of office buildings with floor area under 3,000 m² using gas (Left: electricity, Right: gas). The grey areas denote the modified band of $x_i^T\hat{\beta}^{elec,1} \pm 2.58\hat{\alpha}^{e,e}_{(1,1)}(x_i^{area})^2$ and $x_i^T\hat{\beta}^{gas,1} \pm 2.58\hat{\alpha}^{g,g}_{(1,1)}(x_i^{area})^2$, which capture heteroskedasticity of data well.

[4. Joint probability model]

[4.1. Multivariate normal distribution of monthly energy usage]

A multivariate normal distribution for monthly electricity and gas uses for a year can be defined based on the mean vector and covariance matrix of monthly energy uses in a building obtained by the procedure presented in Section 3, conditional on the features of the building (floor area, number of stories, and approval year of the building), as Equation 12.

$$\begin{bmatrix} y_i^{elec,1} \\ y_i^{elec,2} \\ \vdots \\ y_i^{elec,12} \\ y_i^{gas,1} \\ \vdots \\ y_i^{gas,12} \end{bmatrix} \sim MVN\left(\begin{bmatrix} x_i^T\hat{\beta}^{elec,1} \\ x_i^T\hat{\beta}^{elec,2} \\ \vdots \\ x_i^T\hat{\beta}^{elec,12} \\ x_i^T\hat{\beta}^{gas,1} \\ \vdots \\ x_i^T\hat{\beta}^{gas,12} \end{bmatrix}, \begin{bmatrix} \hat{\alpha}^{e,e}_{(1,1)+} & \hat{\alpha}^{e,e}_{(1,2)+} & \cdots & \hat{\alpha}^{e,e}_{(1,12)+} & \hat{\alpha}^{e,g}_{(1,1)+} & \cdots & \hat{\alpha}^{e,g}_{(1,12)+} \\ \hat{\alpha}^{e,e}_{(2,1)+} & \hat{\alpha}^{e,e}_{(2,2)+} & \cdots & \hat{\alpha}^{e,e}_{(2,2)+} & \hat{\alpha}^{e,g}_{(2,1)+} & \cdots & \hat{\alpha}^{e,g}_{(2,12)+} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}^{e,e}_{(12,1)+} & \hat{\alpha}^{e,e}_{(12,2)+} & \cdots & \hat{\alpha}^{e,e}_{(12,12)+} & \hat{\alpha}^{e,g}_{(12,1)+} & \cdots & \hat{\alpha}^{e,g}_{(12,12)+} \\ \hat{\alpha}^{g,e}_{(1,1)+} & \hat{\alpha}^{g,e}_{(1,2)+} & \cdots & \hat{\alpha}^{g,e}_{(1,12)+} & \hat{\alpha}^{g,g}_{(1,1)+} & \cdots & \hat{\alpha}^{g,g}_{(1,12)+} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}^{g,e}_{(12,1)+} & \hat{\alpha}^{g,e}_{(12,2)+} & \cdots & \hat{\alpha}^{g,e}_{(12,12)+} & \hat{\alpha}^{g,g}_{(12,1)+} & \cdots & \hat{\alpha}^{g,g}_{(12,12)+} \end{bmatrix} (x_i^{area})^2 \right) \quad (12)$$

where MVN is the abbreviation of multivariate normal. The covariance matrix of the distribution contains $(x_i^{area})^2$, meaning reflection of heteroskedasticity. There are two advantages of multivariate normal distribution – i) it is one of the simplest multivariate distributions for model construction, interpretation, maintenance, and sampling; and ii) it provides reasonable fit to near-symmetric data with high variance, which is the case of this study (Figure 6).

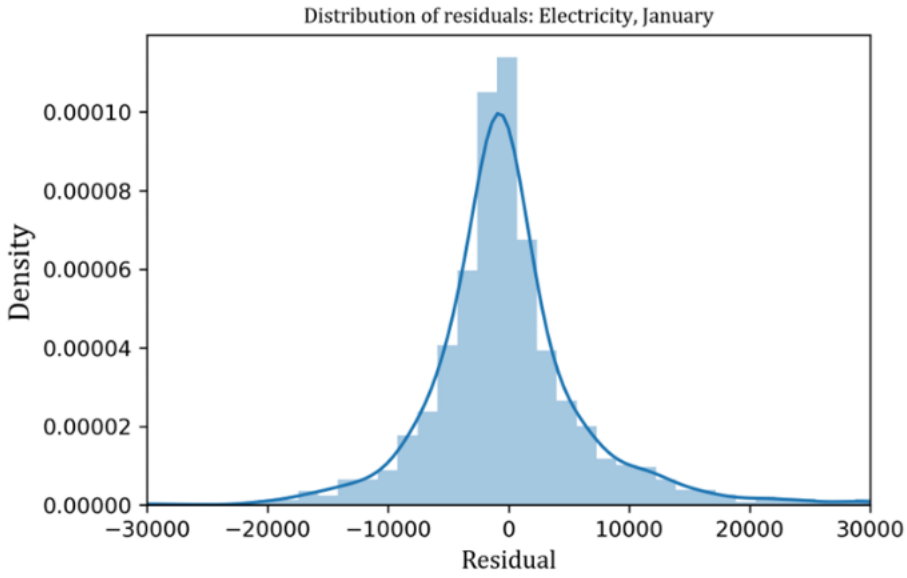

Figure 6. Empirical distribution of residuals from the linear regression model corresponding to electricity use in January, for the subset of office buildings with floor area under 3,000 m^2 using gas. The distribution is bell-shaped and the mode of the distribution is close to zero, which means that approximate normal distribution is applicable to this data.

Figure 7 shows some samples of monthly energy use for one year drawn from the multivariate normal distribution fitted for the subset of office buildings using gas, conditional on floor area 1,500 m^2, seven stories, approved for use in 2000, which show reasonable seasonal patterns of energy use. The key to success in reflection of seasonality in monthly energy use is consideration of covariance between energy uses in different months or different energy types, which was not considered in previous studies. If the covariance is ignored, then samples drawn from the distribution which assumes independency of energy uses in different months or different energy types will show unrealistic seasonal patterns. Figure 8 shows some samples drawn from a different distribution with modified covariance matrix where its off-diagonal elements were replaced with zero. The samples show unrealistic seasonal patterns. Meanwhile, the magnitudes of energy use of the samples in Figure 7 are different to each other due to inevitable high variance nature of the data.

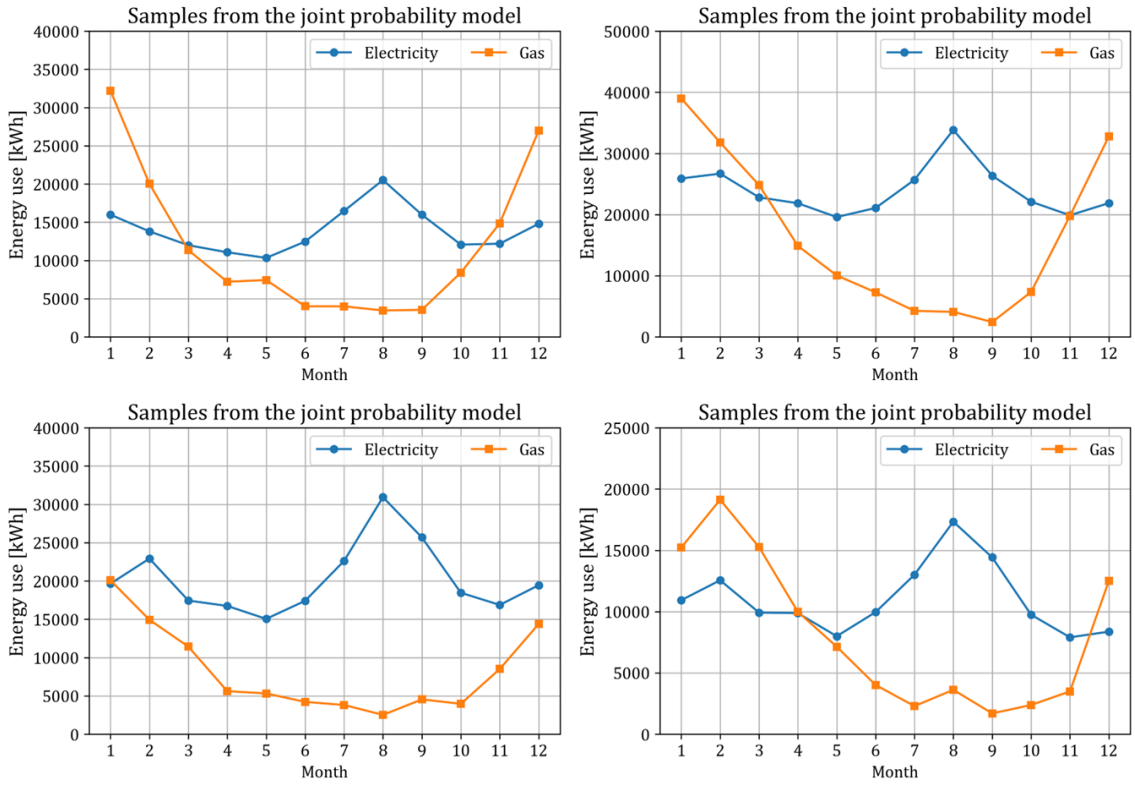

Figure 7. Four samples of monthly energy use for one year drawn from the multivariate normal distribution fitted for the subset of office buildings using gas, conditional on floor area 1,500 m^2, seven stories, approved for use in 2000. The samples show realistic seasonal patterns.

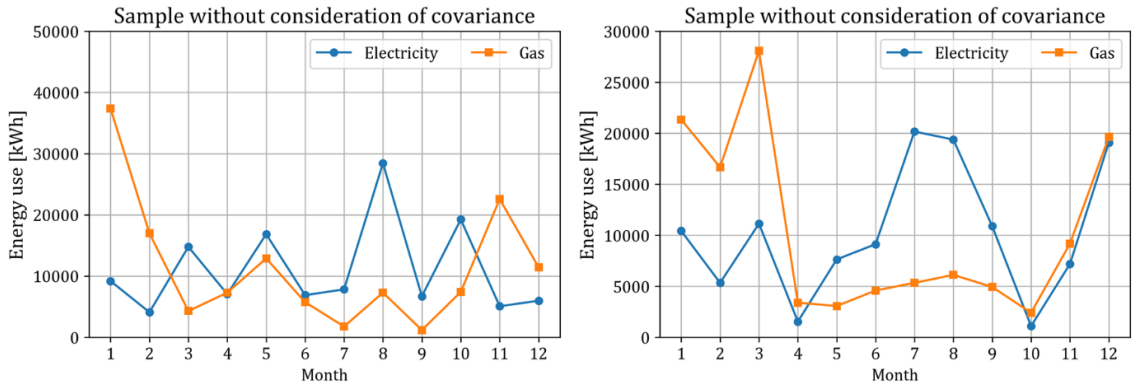

Figure 8. Two samples of monthly energy use for one year drawn from an alternative multivariate normal distribution with modified covariance matrix where its off-diagonal elements were replaced with zero. The samples show volatile and unrealistic seasonal patterns.

To obtain reasonable samples, a post-processing is required because a number of samples may show unrealistic seasonal patterns. Denote the monthly electricity use for a year in a sample

drawn from the multivariate normal distribution as y_0^{elec}. Then, dividing it into its absolute value norm as $\tilde{y}_0^{elec} = y_0^{elec}/|y_0^{elec}|_1$, the quantity $(\tilde{y}_0^{elec})^T \left(\left(\tilde{Y}^{elec} \right)^T \tilde{Y}^{elec} \right)^{-1} \tilde{y}_0^{elec}$ can be computed (as similarly done in Section 2.3.2), where $\tilde{Y}^{elec}$ multiplied with $\tilde{y}_0^{elec}$ is the matrix composed of data preserved after pre-processing in Section 2.3.2. Samples with the quantity over a threshold ($2k/N$ in this study, but it can be adjusted by the user) are deleted. In a numerical experiment for the case of office building using gas with floor area 1,500 m^2 of floor area and seven stories, about 61% of initially drawn samples are preserved after the post-processing. On the contrary, when the post-processing is applied to the samples from the different distribution with ignorance of covariance between error terms for different months, none of the samples are preserved.

[4.2. Application to data correction]

In practice, some values in the record of monthly energy use in a building may be missing or incorrect. Figure 9 shows screenshots of some rows in the database of monthly energy use, which have missing or abnormally low values. If there is a method of filling the missing values or replacing unusual values with reasonable alternative values, it would help enhance data quality of the energy use record. However, models in previous studies with ignorance of covariance or simplified time-series models cannot be used for such task of data correction.

elec_202101	elec_202102	elec_202103	elec_202104	elec_202105	elec_202106	elec_202107	elec_202108	elec_202109	elec_202110	elec_202111	elec_202112
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
3477.0	4495.0	2361.0	762.0	NULL	190.0	246.0	969.0	393.0	1274.0	1352.0	2996.0

gas_202101	gas_202102	gas_202103	gas_202104	gas_202105	gas_202106	gas_202107	gas_202108	gas_202109	gas_202110	gas_202111	gas_202112
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
7714.0	5436.0	8215.0	699.0	NULL	NULL	NULL	NULL	NULL	NULL	NULL	5356.0

gas_202101	gas_202102	gas_202103	gas_202104	gas_202105	gas_202106	gas_202107	gas_202108	gas_202109	gas_202110	gas_202111	gas_202112
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
321.0	5002.0	NULL	4271.0	NULL	NULL	259.0	153.0	165.0	765.0	4969.0	5656.0

Figure 9. Screenshots of some rows in the dataset of monthly energy use, containing missing or abnormally low values.

The joint probability model introduced in Section 4.1 can be used for data correction, based on the conditional multivariate normal distribution where the energy use in month with correct values in the record are assumed to be fixed. For a random vector variable $z = [z_1^T, z_2^T]^T$ following multivariate normal distribution where z_2 has been fixed to be a, the conditional multivariate normal distribution of z_1 can be expressed as Equation 13.

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \Rightarrow P(z_1|z_2 = a) = MVN(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \quad (13)$$

If z is monthly electricity use for a year in a target building where electricity use values for some months z_2 are correct to be a but the values for the other months z_1 are missing or incorrect, the parameters μ_1, μ_2, Σ_{11}, Σ_{12}, Σ_{21}, Σ_{22} become the electricity part of the mean and covariance the joint probability model in Equation 12. The mean of the conditional multivariate normal distribution $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$ can be used as the alternative values for filling the missing values or replacing the incorrect values.

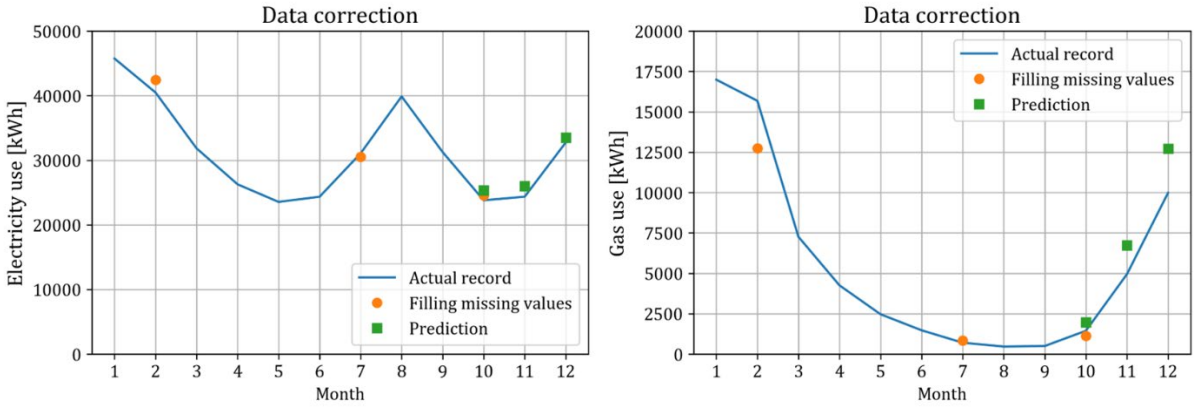

Figure 10. Actual monthly energy use in an exemplary building (connected curve) and estimation of the energy use by the conditional multivariate normal distribution (circles and squares), where the estimation for each group of different marker types has been computed based on assumption of missing values in the corresponding months (Left: electricity, Right: gas).

Figure 10 shows that the mean of the conditional multivariate normal distribution in Equation 13 produces reasonable alternative values. The curve denotes the actual recorded monthly energy use in the exemplary building with known floor area, number of stories and approval year. The circles denote the estimation of energy uses equal to the mean of the conditional multivariate normal distribution, assumed that the energy use record of the months corresponding to the circles (February, July, and October) are missing while record of the other months are available. The squares have the similar meaning as circles (assumed missing values in October, November, and December). The case of squares can be viewed as prediction of monthly energy use since the values of last three months are assumed to be missing and estimated given the energy use of preceding months. For electricity, the estimated values are quite close to the actual record. For gas, although the estimated values are a little deviated from the actual record due to the high variance of gas data, the new data generated by replacing the estimated values shows realistic seasonal pattern.

[5. Summary]

This study provides a statistical method to model the 'joint' probability distribution of 'monthly' electricity and gas uses for a year in individual urban buildings, conditional on the feature of the buildings. The process has been summarized as below:

i) Pre-process the database of monthly energy use and building features. Data points with missing values, or abnormal seasonal pattern of monthly energy use, or abnormal magnitude of energy use have been deleted. Points with abnormal seasonal pattern have been identified by a method which quantifies remoteness of each point from the cluster of the points applied to a transformed dataset. Points with abnormal magnitude of energy use have been identified by computation of Cook's distance.

ii) For each subset of database (divided with respect to building use, floor area interval, use of gas), fit individual linear regression models. The response variable of each regression model is electricity or gas use in each month of buildings. In this study, the selected explanatory variables are floor area, number of stories, and approval year for use of buildings. Obtain the estimates of coefficients and residuals of the regression models.

iii) Establish auxiliary regression models to estimate the covariance of the errors as an increasing function of increasing floor area (in other words, estimate the structure of heteroskedasticity in the data). The response variable is multiplication of two residuals, each from regression models corresponding to the same or different months or energy types. The only explanatory variable is the square of floor area (no intercept). Transform the obtained estimate of covariance matrix into its nearest positive semidefinite matrix.

iv) Define a multivariate normal distribution conditional on the features of a building, where its mean vector is computed based on the estimates of coefficients obtained in ii) and its covariance matrix is computed based on the estimates of covariance matrix obtained in iii).

The joint probability model can be used to generate samples of monthly energy uses for a year in a target building, with realistic seasonal pattern and magnitude. Also, the joint probability model can be used to fill missing values or replace incorrect values of monthly energy use in a building with reasonable estimations, given that some correct values of monthly energy use are recorded in that building. The key to success of the provided model is the consideration of covariance between monthly energy uses, which exists even after controlling the effects of building size, height, and age.

References

[1] IEA (2022), Buildings, IEA, Paris https://www.iea.org/reports/buildings, License: CC BY 4.0

[2] Li, Z., Han, Y., & Xu, P. (2014). Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Applied Energy*, 124, 325-334.

[3] Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering*, 6(1), 1-20.

[4] Ciulla, G., & D'Amico, A. (2019). Building energy performance forecasting: A multiple linear regression approach. *Applied Energy*, 253, 113500.

[5] Turiel, I., Craig, P., Levine, M., McMahon, J., McCollister, G., Hesterberg, B., & Robinson, M. (1987). Estimation of energy intensity by end-use for commercial buildings. *Energy*, 12(6), 435-446.

[6] Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and buildings*, 40(3), 394-398.

[7] Zhong, X., Hu, M., Deetman, S., Rodrigues, J. F., Lin, H. X., Tukker, A., & Behrens, P. (2021). The evolution and future perspectives of energy intensity in the global building sector 1971–2060. *Journal of Cleaner Production*, 305, 127098.

[8] Olofsson, T., Andersson, S., & Sjögren, J. U. (2009). Building energy parameter investigations based on multivariate analysis. *Energy and Buildings*, 41(1), 71-80.

[9] Howard, B., Parshall, L., Thompson, J., Hammer, S., Dickinson, J., & Modi, V. (2012). Spatial distribution of urban building energy consumption by end use. *Energy and Buildings*, 45, 141-151.

[10] Andrews, C. J., & Krogmann, U. (2009). Technology diffusion and energy intensity in US commercial buildings. *Energy Policy*, 37(2), 541-553.

[11] Hsu, D. (2015). Identifying key variables and interactions in statistical models of building energy consumption using regularization. *Energy*, 83, 144-155.

[12] Apadula, F., Bassini, A., Elli, A., & Scapin, S. (2012). Relationships between meteorological variables and monthly electricity demand. *Applied Energy*, 98, 346-356.

[13] Song, J., & Song, S. J. (2020). A framework for analyzing city-wide impact of building-integrated renewable energy. *Applied Energy*, 276, 115489.

[14] Smith, A., Fumo, N., Luck, R., & Mago, P. J. (2011). Robustness of a methodology for estimating hourly energy consumption of buildings using monthly utility bills. *Energy and Buildings*, 43(4), 779-786.

[15] Pagliarini, G., & Rainieri, S. (2012). Restoration of the building hourly space heating and cooling loads from the monthly energy consumption. *Energy and buildings*, 49, 348-355.

[16] Lamagna, M., Nastasi, B., Groppi, D., Nezhad, M. M., & Garcia, D. A. (2020, December). Hourly energy profile determination technique from monthly energy bills. In *Building Simulation* (Vol. 13, No. 6, pp. 1235-1248). *Tsinghua University Press.*

[17] Catalina, T., Virgone, J., & Blanco, E. (2008). Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy and buildings*, 40(10), 1825-1832.

[18] Kim, MK., Kim, BS., & Kim, JA. (2014). Development of a standard model for energy consumption in residential and commercial buildings in Seoul. *City of Seoul*, ISBN: 9791156212942 93530.

[19] Xu, J., Kang, X., Chen, Z., Yan, D., Guo, S., Jin, Y., ... & Jia, R. (2021, February). Clustering-based probability distribution model for monthly residential building electricity consumption analysis. In *Building Simulation* (Vol. 14, No. 1, pp. 149-164). *Tsinghua University Press.*

[20] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). *New York: springer*.

[21] Araji, M. T. (2019, August). Surface-to-volume ratio: How building geometry impacts solar energy production and heat gain through envelopes. In *IOP Conference Series: Earth and Environmental Science* (Vol. 323, No. 1, p. 012034). IOP Publishing.

[22] Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 591-605.

[23] Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1), 17-22.

[24] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15-18.

[25] Davidson, R., & MacKinnon, J. G. (1993). Estimation and inference in econometrics (Vol. 63). *New York: Oxford*.

[26] Amemiya, T., & AMEMIYA, T. A. (1985). Advanced econometrics. *Harvard university press*.

[27] Mandy, D. M., & Martins-Filho, C. (1993). Seemingly unrelated regressions under additive heteroscedasticity: Theory and share equation applications. *Journal of Econometrics*, 58(3), 315-346.