Applied Energy Symposium 2021: Low carbon cities and urban energy systems August 24-27, 2021, Matsue, Japan
Paper ID: 61

Correlation Analysis of Demographic Data and Power System's Load Profile in a Smart City

September 4-8, 2021, Matsue, Japan

Yufan Qiu[1], Xinwei Shen [2]

1 University of California, Berkeley, USA

2 Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China

ABSTRACT

As energy and environmental issues become more and more serious, we need to further improve the overall energy efficiency of modern cities. The operation of the urban energy system is affected by the characteristics of human social behavior, which increases the complexity of the problem. Therefore, cities must find how to conduct a comprehensive analysis of the energy use behavior of social residents from the perspective of smart cities and energy Internet, and then carry out effective guidance and management, assist the transformation and upgrading of the urban energy system, reasonable planning, and realize the improvement of urban comprehensive energy efficiency, cleanliness, and low carbon. The social-physical behavior model of residents formed by the corresponding population mobility attributes and energy attributes can reflect economic conditions and the behavior habits of various groups, thereby assisting the investment decision-making of urban infrastructure construction, and providing a basis for urban and power grid planning.

Keywords: smart cities, urban energy system, social-physical model, population mobility attributes, energy efficiency, correlation analysis, urban and power grid planning, urban infrastructure construction

NONMENCLATURE

Abbreviations

MSE	mean squared error
R^2	R squared

Symbols

l	power load
n	number of customers
c_l	ratio of low consumption people
c_m	ratio of medium consumption people
c_h	ratio of high consumption people
f	percentage of female
m	percentage of male
a_1	people less than 18/total population
a_2	people of 18-30/total population
a_3	people of 31-45/total population
a_4	people of 46-60/total population
a_5	people over 60/total population

1. INTRODUCTION

The content of this chapter mainly includes using population movement data as a predictor variable to estimate the power load of the corresponding substation, and on this basis, considering the potential impact of the population data of the neighboring area of the typical area on the prediction results.

In urban areas, as the social population distribution becomes more obvious, it is more appropriate to use social population distribution data for consumer electricity consumption research such as using a stepwise selection method to study the influencing factors of population clustering characteristics on residents' electricity consumption [1]. However, more existing research focuses on how to mine consumers' social demographic information from the existing massive smart meter data and does not involve the impact of population flow characteristics on changes in power load.

Selection and peer-review under responsibility of the scientific committee of CUE2021

Furthermore, in recent years, the topic of energy internet, and smart cities has attracted numerous scholars and people in industries [2]. The research and application have focused on multi-energy management systems, Urban Internet of Things technology, and smart building system construction [3].

By combining the residential mobile phone big data provided by the China Mobile Big Data Center and the substation power compliance data, the prediction of the power load of the substation based on the population movement data is carried out, reflecting the value of the integration of mobile data and power data. The article is based on the data with "user tags" provided by China Mobile, which indicates the basic structure of the demographic data of this project and deepens the understanding of the "blocks" mentioned in the follow-up research. Population data used for this article are statistics based on the unit of "block", and the division of the block is based on the power supply range of the substations in Zhuhai City and the administrative area of Zhuhai City. In addition, the power load data of each block during 2018 and 2021 is available for establishing a prediction model.

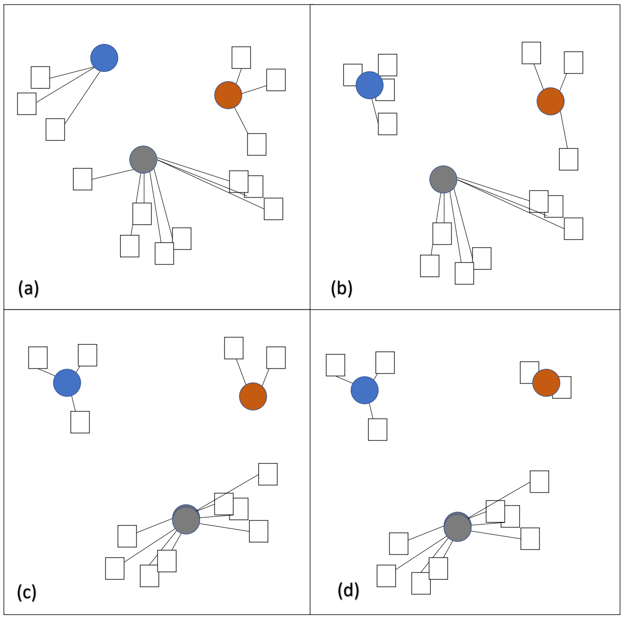
2. CLASSIFICTAION OF REGIONS BASED ON THE SOCIO-DEMOGRAPHIC DATA

2.1 K-means clustering

The k-means clustering algorithm is an iterative solution clustering analysis algorithm [4], which is also illustrated in graph 1. Specifically, it can be described as: the known observation set $(x_1, x_2, x_3, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering must divide these n observations into k sets (k⩽n), so that the sum of squares within the group (WCSS within-cluster sum of squares) is the smallest. In other words, its goal is to find the cluster $\boldsymbol{S}_i$ that satisfies the following formula:

$$arg \min_{\boldsymbol{S}} \sum_{i=1}^{k} \sum_{x \epsilon S_I} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

Among them, $\boldsymbol{\mu}_i$ is the mean value of all points in $\boldsymbol{S}_i$.



Graph 1. k-means algorithms

The steps of k-means algorithms are:

1. Select k initial clustering samples as the initial cluster centers $a = a_1, a_2, \ldots, a_k$;
2. For each sample x_i in the data set, calculate its distance to the k cluster centers and assign them to the cluster corresponding to the cluster center with the smallest distance;
3. For each category a_j, recalculate its cluster center $a_j = \frac{1}{|C_i|} \sum_{x \in C_i} \boldsymbol{x}$ (that is, the centroid of all samples belonging to the category);
4. Repeat the above two steps 2 and 3 until a certain stopping condition is reached (number of iterations, minimum error change, etc.)

Based on k-means clustering and characteristics of each group, four typical types of regions are defined, which are industrial areas, mixed commercial and residential areas, traffic hotspots, and residential areas.

2.2 Correlation coefficient analysis

In statistics, the Pearson correlation coefficient, also known as Pearson product-moment correlation coefficient [5] (Pearson product-moment correlation coefficient, referred to as PPMCC or PCCs), is used to measure two variables The correlation between X and Y (linear correlation), whose value is between -1 and 1. The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables:

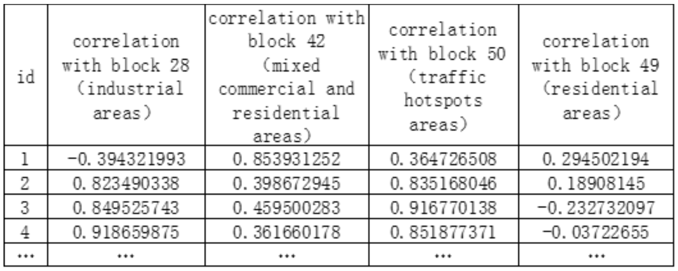
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}.$$

In general, people think that the coefficient between 0.8-1.0 represents an extremely strong correlation, 0.6-0.8 shows a strong correlation, 0.4-0.6 represents a moderate correlation, 0.2-0.4 signs a weak correlation, and 0.0-0.2 means a very weak correlation or no correlation.

2.3 Identification of the types of regional residents based on the social-demographic curve

The first step is to excavate the 24-hour population curve of 60 plots in Zhuhai City based on the k-means clustering method, then select four types of typical blocks through the selection of typical buildings on the map: industrial areas, commercial and residential areas, residential areas, and traffic hotspots areas. Then, all blocks in Zhuhai city are classified by the Pearson correlation coefficient on the basis of population data and residents' attributes, which further support power grid planning.

Considering the incompleteness of the corresponding relationship between the current substations and population blocks, the selected blocks and substations correspond to the four blocks that are relatively clear and have typical functional characteristics as model samples. According to the four different block functions (industrial areas, mixed commercial and residential areas, traffic hotspots, and residential areas), four prediction models were output and their accuracy was verified. On the basis of the discrimination results of the types of regional residents based on the socio-demographic curves, the Gaoxin substation and its corresponding No.28 block, the Hongshan substation and its corresponding No.42 block, and the Kouan substation and its corresponding blocks were finally selected. Block No.50 is used as a model sample of a typical industrial park, a typical mixed commercial and residential area, and a typical traffic hotspot. Because the Beishan substation corresponding to the No.23 block as a typical residential block, which may have a certain impact on the construction of the prediction model, so the No. 49 block with the highest correlation with the No. 23 block and the Gongbei substation corresponding to this block are selected as the model samples of the typical residential area. Part of the classification results is shown below:



id	correlation with block 28 （industrial areas）	correlation with block 42 （mixed commercial and residential areas）	correlation with block 50 （traffic hotspots areas）	correlation with block 49 （residential areas）
1	-0.394321993	0.853931252	0.364726508	0.294502194
2	0.823490338	0.398672945	0.835168046	0.18908145
3	0.849525743	0.459500283	0.916770138	-0.232732097
4	0.918659875	0.361660178	0.851877371	-0.03722655
…	…	…	…	…

Table 1. Part of the typical region types classification results

3. POWER LOAD FORECASTING BASED ON POPULATION MOVEMENT DATA

3.1 Data processing

The data used to train the regression model is the data of real-time passenger flow in No.28, No.42, No.49, and No.50 blocks in January and March 2019, as well as the power load data of Gaoxin substation, Hongshan substation, Gongbei substation, and kouan substation. In terms of testing, we use the real-time passenger flow data of No.28, No.42, No.49, and No.50 blocks in January and March 2020 from Zhuhai Mobile, as well as the power load data of Gaoxin substation, Hongshan substation, Gongbei substation, and kouan substation. MSE and R^2 of models are used as model measurement indicators to compare and analyze the prediction models. The specific formula of MSE is:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2$$

where y is test set data. Comparing MSE compares the models by comparing the average distance between the points in the test data set and the model. The smaller the value, the more accurate the model. The formula of R^2 is $R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$, which represents the proportion of all the dependent variables that can be explained by the independent variables through the regression relationship, where $SS_{residual}$ is the residual sum of squares and SS_{total} is the total sum of squares. The larger the R^2, the higher the accuracy of the model.

Data preprocessing [6] refers to dealing with data before the actual processing, that is, the necessary processing such as auditing, screening, and sorting before the collected data is classified or grouped. The data in real world is generally incomplete, inconsistent, and dirty data, which is hard to be mined, or the mining results are not satisfactory. Before constructing the forecasting model, pre-processing of data is needed, which is to convert power load data in the unit of every minute into power load data in the unit of the hour, to make to consistent with population data. At the same time, the missing data is deleted, and the model is constructed with effective one-to-one correspondence of load data and population data sets.

3.2 Correlation analysis of power load and population data attributes

Correlation coefficients of power load and population data attributes are calculated.

In Gaoxin substation, the power load is positively correlated to number of customers, ratio of high consumption people, percentage of female, ratio of people between 31 and 45 to total population, but negatively correlated to ratio of low consumption people, ratio of medium consumption people, percentage of male, ratio of people less than 18 to total population, ratio of people between 18 and 30 to total population, ratio of people between 31 and 45 to total population, ratio of people between 46 and 60 to total population.

In Hongshan substation, the power load is positively correlated to number of customers, ratio of low consumption people, ratio of medium consumption people, percentage of female, ratio of people less than 18 to total population, ratio of people between 18 and 30 to total population, ratio of people over 60 to total population, but negatively correlated to ratio of high consumption people, percentage of male, ratio of people between 31 and 45 to total population, ratio of people between 46 and 60 to total population.

In Gongbei substation, the power load is positively correlated to ratio of low consumption people, ratio of people between 31 and 45 to total population, ratio of people between 46 and 60 to total population, ratio of people over 60 to total population, but negatively correlated to number of customers, ratio of medium consumption people, ratio of high consumption people, percentage of female, ratio of people less than 18 to total population, ratio of people between 18 and 30 to total population.

In Kouan substation, the power load is positively correlated to number of customers, ratio of high consumption people, percentage of female, ratio of people between 18 and 30 to total population, ratio of people between 31 and 45 to total population, ratio of people between 46 and 60 to total population, but negatively correlated to ratio of low consumption people, ratio of medium consumption people, percentage of male, ratio of people less than 18 to total population, ratio of people over 60 to total population.

3.3 Regression modelling

The regression model is a predictive modeling technique, mainly used to study the relationship between the dependent variable (target) and independent variable (predictor). This technique is usually applied to predictive analysis and to discover causal relationships between variables [7].

This chapter will deal with linear regression and nonlinear regression. Linear regression uses the best fitting straight line (that is, the regression line) to establish a relationship between the dependent variable (Y) and one or more independent variables (X). The equation of linear regression is as follows:

$$Y = a + b \cdot X + e$$

Where a represents the intercept, b represents the slope of the line, and e is the error term. This equation can predict the value of the target variable based on the given predictor variable (X).

Non-linear regression is a regression in which the regression function has a non-linear structure with respect to the unknown regression coefficients. In regression analysis, it is often encountered that the relationship between variables is not linear, but a certain non-linear relationship. The variables in this data set basically conform to linear relationships, so this chapter mainly focuses on the construction and analysis of linear regression models. In the subsequent model fitting and model performance comparison, nonlinear models will be considered at the same time.

The steps of linear regression in this paper are summarized as below:

i. Make a scatter plot between each independent variable and the dependent variable, then observe the trend between the variables. Then focus on observing the scatter plot from the following three aspects to obtain key information: one is to judge whether there is a clue trend; the other is to judge whether the correlation is linear or curvilinear; the third is to observe whether there is a strong influence point that deviates from the trend.

ii. By observing the distribution of the data, judge the normality of the variable and judge whether the value of the independent variable is extreme.

iii. After screening the independent variables, a linear regression model is initially constructed.

iv. Use Residual error analysis to determine whether the residuals are independent and whether the residual distribution is normal. Then determine whether there is an influence point, and whether there is multicollinearity. Based on this, judge whether a nonlinear regression model is a better choice, and then fit a more appropriate model by using higher-order terms or logarithmic forms of the variables through, for example, the transformation of variables.

3.4 Results

The dependent variable in the models is power load, while all population attributes are independent variables. Regression models for four typical types of blocks are performed, and the best prediction model for each block is selected based on comparison of MSE and R^2.

R^2, the coefficient of determination, is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well observed outcomes are replicated by the model. In general, the higher the R-squared, the better the model fits your data. MSE, the mean squared error indicates the average squared difference between the estimated values and the actual value. It is used to measure the quality of the estimator.

An interesting finding is that, for some models, including cubic terms improves the model fitting. A plausible explanation is that the previous model may not have enough model complexity. Including multiple terms (multiple model special cases) increases the complexity of the model and better fits the model. Take the Hongshan substation as an example. Based on power load data and population movement data in March of the Honshan substation, the prediction is shown below: $l{\sim}n + c_l + c_m + m + f + a_1 + a_2 + a_3 + a_4 + a_5$

where $\mathrm{R}^2 = 0.4487, \mathrm{MSE} = 929.0284$.

From the scatter plots of the variables, it seems that c_l is highly correlated with other variables in this model, so quadratic and cubic terms of c_l are considered:

$$l{\sim}n + c_l + c_m + m + f + a_1 + a_2 + a_3 + a_4 + a_5 + (c_l)^2 + (c_l)^3$$

where $R^2 = 0.7209, MSE = 709.4715$.

It shows that the inclusion of cubic terms at this time improves the model fit. The explanation is that the previous model may not have enough model complexity. Including multiple terms increases the complexity of the model and better fits the model.

3.5 Discussion

The selection of data can affect the accuracy of prediction model. We will explore the impact of the following three situations on the accuracy of the four typical block prediction models:

1) Only the population data (from Zhuhai Mobile Company) and power load data of the four typical blocks in February 2019 is used as training data, while population data (from Zhuhai Mobile) and power load data in March 2019 are used for testing to analyze and select the prediction model.

2) Only the population data (from Zhuhai Mobile Company) and power load data of the four typical blocks in April 2019 is used as training data, while population data (from Zhuhai Mobile) and power load data in March 2019 are used for testing to analyze and select the prediction model.

3) Combine population data (from Zhuhai Mobile Company) and power load data of the four typical blocks in April 2019 and population data (from Zhuhai Mobile Company) and power load data of the four typical blocks in February 2019 to train models, while population data (from Zhuhai Mobile) and power load data in March 2019 are used for testing to analyze and select the prediction model. The analysis of four different typical blocks was performed.

Take the typical traffic hotspots area (No.50 block) where the Kouan substation located in, as an example.

In scenario 1), the prediction model is:

$$l{\sim}n + c_l + c_m + f + a_1 + a_2 + +a_4$$

where $R^2 = 0.8541,\ MSE = 4.6173$

In scenario 2), the prediction model is:

$$l{\sim}n + c_l + f + m + a_1 + a_2 + a_3 + a_4$$

where $R^2 = 0.7699,\ MSE = 8.1162$

In scenario 3), the prediction model is:

$$l{\sim}n + c_l + c_m + m + a_1 + a_2 + a_3 + a_4$$

where $R^2 = 0.8822,\ MSE = 2.1529$

This paper uses R^2 and MSE for evaluation and validation. From the above results, it can be found that in the March 2019 load forecast of the port substation, the training data was more accurate in February of the same year, and the model obtained by combining the February and April data as the training data performed relatively best.

3.6 Conclusion

For most blocks and substations, multiple linear regression can better fit the prediction model, but for some specific blocks such as high-tech substations, including multiple items (special cases of multiple models) can improve the model fit and improve the accuracy of the model The possible reason is that the complexity of the previous model is not enough, and the complexity of the model is improved by including multiple items, so as to better fit the model. It is possible to initially verify the feasibility of optimizing the performance of the model by introducing high-order terms of variables for specific blocks and substations.

ACKNOWLEDGEMENT

This work is supported by the Science and Technology Project of China Southern Power Grid (ZBKJXM20180209).

REFERENCE

[1]A.Kavousian, R. Rajagopal, and M. Fischer, “Determinants of res?idential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants’ behavior,” Energy, vol. 55, pp. 184–194, Jun. 2013.

[2] Carli R, Dotoli M, Pellegrino R. A hierarchical decision-making strategy for the energy management of smart cities[J]. IEEE Transactions on Automation Science and Engineering, 2016, 14(2): 505-523.

[3]Hongbin Sun, Qinglai Guo, Boming Zhang, Wenchuan Wu, Bin Wang, Xinwei Shen, and Jianhui Wang, Integrated Energy Management System: Concept, design, and demonstration in China [J]. IEEE Electrification Magazine, May 2018.

[4]Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. The Computer Journal 13(2):156-163.

[5] "The human disease network", Albert Barabasi et al., Plos.org.

[6] M.S. Chen, J. Han, P.S. Yu (1996) "Data mining: an overview from a database perspective ". Knowledge and data Engineering, IEEE Transactions on 8 (6), 866–883.

[7]Jichuan Wang, Zhigang Guo (2001), "logistic regression modelling: methods and application", higher education publish.