A Multivariate Analysis of the Residential Features that Dictate Energy Efficiency

Leah Canter, Cristian Mojica, Razmig Zeitounian

STAT 520

May 11, 2023

Executive Summary
Efficient energy usage is a crucial aspect of building design since it ensures that a building is safe and comfortable for those who may stay within it. We are interested in looking at which factors best predict heating and cooling load in a building. This paper looks at three different methods for understanding the relationship between heating and cooling load and other building factors. These methods are principal component analysis (PCA), factor analysis, and multivariate regression analysis. PCA found four principal components. These components captured important information about building features, such as overall size and shape of the building, the size of the wall area, amount and distribution of glazing, and orientation. For factor analysis, three factors were found to capture underlying information among the building features. Namely, they are a contrast between home area and volume, as well as primary influence from wall area, and glazing. Finally, an overall test of effect by multivariate regression suggests that increased relative compactness or surface area on average predicts an decrease in heating load and cooling load. On the other hand, increased overall height and glazing area on average predicts increase in heating and cooling load.

Data Introduction
Efficient energy usage is a crucial aspect of building design, both in terms of financial and practical considerations. Given that buildings serve as spaces for living, work, leisure, and social interaction, it is imperative that they prioritize energy efficiency to conserve natural resources. Therefore, planning for efficient energy consumption within these facilities should be a primary consideration in building design.

The dataset comes from the UC Irvine Machine Learning Repository, and is known there as the "Energy efficiency Data Set" (UCI Machine Learning Repository, 2012). It has ten attributes (eight predictors and two responses) and 768 observations. This data comes from variations on 12 different building shapes simulated in Ecotect, which is a building performance simulation software that aids in "the simulation, analysis and optimization of high performance buildings and systems" (Ziger | Snead Architects, 2008). The features used for prediction include relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. These predictors are used to explain the two response variables, heating load and cooling load. This data set contains no missing cells.

Methods
Principal Component Analysis

One of the interests in our study was to use PCA to simplify our predictors while still retaining any trends in the data. By doing this, we obtain a streamlined summary of our variables (Johnson & Wichern, 2019). The number of principal components will be selected based on the

cumulative proportion of variation explained by the data being greater than our threshold value of 80 percent.

Factor Analysis

Another interest in our study was to see what underlying factors may explain the relationships in the predictors. For this, we used factor analysis, which groups variables based on their correlations. In particular, it uses domain knowledge to group variables by underlying constructs (Johnson & Wichern, 2019). The factor analysis method used in this paper is based on the above PCA. We will utilize a minimum eigenvalue criterion of one to determine how many factors to keep for an interpretable explanation of the latent variables on building design (Johnson & Wichern, 2019).

Multivariate Regression

The last method used is multivariate regression, which will be performed to understand heating and cooling load as they are simultaneously affected by the predicting features of the buildings. We performed a test of the overall effect of the predictor variables onto heating load and cooling load. Predictor variables to include in the model should not be linear combinations of other variables (Johnson & Wichern, 2019). Multivariate regression models must meet a few assumptions. These include checking assumptions on being multivariate normal and the homogeneity of the variance-covariance matrix of the random error vector (Johnson & Wichern, 2019). All statistical analysis was conducted in SAS.

Results

Principal Component Analysis

Using our threshold of 80 percent cumulative variation, we chose four principal components that together explain 89.45 percent of the variation in the predictor data. The principal components are represented in Table 1 below.

Variable	**PC 1**	**PC 2**	**PC 3**	**PC 4**
Relative compactness	0.4960	-0.2447	0	0
Surface area	-0.5017	0.2315	0	0
Wall area	0.0325	0.8943	0	0
Roof area	-0.5050	-0.2061	0	0
Overall height	0.4926	0.2104	0	0
Orientation	0	0	0	1
Glazing area	0	0	0.7071	0
Glazing area distribution	0	0	0.7071	0

Table 1: Principal Component Eigenvectors

The first principal component has the highest positive loadings on relative compactness and building height, and negative loadings on surface area and roof area. As such, this principal component may measure the overall building shape and size. Buildings that are taller or more compact tend to have higher values, while buildings with a larger surface or roof area tend to have lower values, all relative to their volume. The second principal component has the highest positive loading on the wall area. Additionally, there is also a contrast between the other positive load of surface area and wall area compared to the negative loading of relative compactness and roof area. This principal component may measure the size of the wall area, with larger wall areas leading to larger values. The third principal component measures glazing area and glazing area distribution of the building. Both are positive, so a higher glazing area corresponds with a higher value of this principal component. The fourth principal component only includes orientation, with no influence from any other variables. This makes sense since that variable has no correlation to the other predictor variables.

Factor Analysis

We ran a factor analysis and selected three factors since their eigenvalues were larger than one. 76.83 percent of the total variation in our eight predictor variables can be explained by the VARIMAX rotation factors retained. The rotated factors are represented below:

Variable	**Factor 1**	**Factor 2**	**Factor 3**
Relative compactness	-0.9651	-0.2318	0
Surface area	0.9756	0.2166	0
Wall area	-0.0203	0.9975	0
Roof area	0.9611	-0.2705	0
Overall height	-0.9441	0.2745	0
Orientation	0	0	0
Glazing area	0	0	0.7788
Glazing area distribution	0	0	0.7788

Table 2: Factor Analysis

These results are similar to our principal components. This makes sense since the factor analysis was based on PCA. The main difference from our first principal component is that the first factor has the values' signs reversed. So, there is a positive combination of surface area and roof area contrasted against a negative combination of relative compactness and overall height. This linear

combination could look at the contrast between home area and volume. The second factor is dominated by wall area and slightly influenced by a contrast between surface area and overall height against relative compactness and roof area. The third factor is determined entirely by glazing.

Multivariate Regression

We regressed heating and cooling loads onto each predictor except roof area (since it is a linear combination of other predictors), orientation, and glazing area distribution. The data was found to have a multivariate random error vector, and the homogeneity of the variance-covariance matrix of the random error vector was present. Our assumptions were therefore met, and we were able to conduct a valid test of overall effect. Running this test returns, among other test statistics, Wilk's lambda, which suggests here that about 94% of the variation of the two response variables can be explained by the five predictors. The overall test suggests that there are strong linear associations between the five predictors and two responses.

Statistics	**Value**	**F Value**	**Numerator DF**	**Denominator DF**	**P-Value**
Wilks' Lambda	0.0560	490.83	10	1522	< 0.001
Pillal's Trace	1.0892	182.27	10	1524	< 0.001
Hotelling-Lawley Trace	14.2570	1084.01	10	1138.8	< 0.001
Roy's Greatest Root	14.0728	2144.69	5	762	< 0.001

Table 3: Multivariate Statistics and F Approximations

The resulting regression models are shown here, where each term in both models is statistically significant:

$$\widehat{heating\_load} = 4.073 - 2.894 * (relative\_compactness) - 0.003 * (surface\_area) + 0.005 * (wall\_area) + 0.523 * (overall\_height) + 2.251 * (glazing\_area)$$

$$\widehat{cooling\_load} = 8.540 - 4.944 * (relative\_compactness) - 0.006 * (surface\_area) + 0.004 * (wall\_area) + 0.463 * (overall\_height) + 1.514 * (glazing\_area)$$

Each unit increase in relative compactness or surface area leads to an expected decrease in heating load and cooling load, and each unit increase in overall height and glazing area leads to an expected increase in heating and cooling load.

Reference

1. Johnson, R. A., & Wichern, D. W. (2019). Applied Multivariate Statistical Analysis (6th ed.). Pearson.
2. UCI Machine Learning Repository. (2012). Energy Efficiency Data Set. UCI Machine Learning Repository. Retrieved May 10, 2023, from https://archive.ics.uci.edu/ml/datasets/Energy+efficiency#
3. Ziger | Snead Architects. "ECOTECT (Building Performance Simulation Software)." 5 August 2008. Accessed 26 April 2023. https://www.zigersnead.com/current/blog/post/ecotect-building-performance-simulation-software/