Optimizing Summer Housing for Electricity Conservation

Hyungwoo Kim [*1], Qiuwei Shou [†1], and Ozan Yolasigmaz [‡1]

[1]Department of Mathematics, Indiana University

April 24, 2015

Acknowledgements

We thank Professor Kevin Pilgrim for his guidance throughout the project, Peggy D. Maschino, Andrea Moore, and Ian D. Yarbrough of IU Physical Plant for bearing with our numerous requests for data and explanations about their work, and Ian G. Anson of Indiana Statistical Consulting Center for his advice on regression analysis.

Contents

[*]hk.29@umail.iu.edu

[†]qiuwshou@umail.iu.edu

[‡]oyolasig@indiana.edu

1 Project Description

With more than 40 buildings used as dormitories by students, visitors, and summer school attendants, Indiana University's Residential Programs and Services (RPS) is a complex system with high level of electricity usage, especially during the hot and humid summer days in Bloomington where electricity is used to cool down buildings in various ways. During the fall and spring semesters, dormitories always have close to full occupancy, so there is not a different way we can place the students. However, the current system to determine where all these transient clientele will reside during the summer might not be the most efficient one. Therefore our aim is to analyze the system and have models describing it, so that in the future this choice can be done more effectively.

To explain the problem, we need understanding on buildings and their characteristics as well as a survey of the summer visitors and their needs.

There are around 20 buildings in the system where some of these buildings are actually a collection of smaller buildings. Being built in the period from mid-1900's to 2013, these buildings vary widely with different characteristics; for example, some building provide food services, and have offices and classrooms, but others do not. Also, buildings operate with different cooling systems.

Location matters the most for the summer visitors, as they will be interested the most in staying in a building that is closest to what they need to do, than us saving money, unfortunately. There are all sorts of visitors during the summer including

summer sports schools, a number of conferences and conventions, summer classes. Depending on events they will be attending, they might prefer living at one dorm to another.

Meanwhile, the monthly electricity bill has two main contributors:

- Total electricity usage: This is self-explanatory, all the electricity used in the campus (units of kWh).
- Peak demand: In the whole billing period of a month, the total energy requested by the whole campus from the electricity provider (units of kW) is recorded half-hourly.

For this project, for simplicity, the only portion that we will investigate is the total electricity usage. Although the bill is monthly, we have daily usage data from all of the buildings and their smaller parts, and since the occupancy can change on a daily basis, we will consider the daily electricity usage of each of the buildings. Thus, our models will be designed to predict the daily usage.

In addition to building characteristics, we found that effects of occupancy, temperature and humidity are more significant than the others.

All these factors might be affecting the total electricity usage in various ways, and these are what we will base our simulations on since they are the biggest contributors, and our goal is to be able to create models for each of the buildings that can predict how much electricity will be used given all of these factors have a certain value.

With flood of information and influencing factors, we took the following steps to accomplish the goals of this project. By analyzing the data in hand, we obtained models that give the expected electricity usage of each building depending on occupancy, temperature, humidity, and other factors which we can call to be 'noise.' Then, using the models for each building, we suggest a simulation that can tell the RPS and Physical Plant staffs how much money can be saved from the electricity bill if they allocate the visitors in various ways.

2 Types of Data Used

The data used in this study are mostly provided by the Indiana University's Physical Plant. A wide arrange of data was available to us including electrical usage, building complex, square foot area of the building, number of occupants, number of maximum beds, cooling system in use. We had to face difficulty choosing which data would be

more relevant for our purpose. Additionally, we decided to use weather data, which was drawn directly from Indiana University Building Systems Weather Page[1].

2.1 Panel Nature of the Data

The term "panel data" refers to multi-dimensional data frequently involving measurements over time. Our data contain observations of multiple phenomena (electrical usage, the number of occupancy, temperature, etc.) by multiple entities (different building complex) over multiple time periods (date). Therefore, for our purpose of study, one data point will include multiple phenomena by a certain building complex on certain date.

2.2 Categorical Variables

These variables determine number of data points we have because one data point can only include one building complex on one date. For our case, with 98 date points and 8 different building complex, total data points will be 784, which is 98 times 8.

- **Date** From May 11, 2014 to August 14, 2014. Recorded daily. 98 days are in this period.
- **Building Names / Building ID**: Only 8 buildings that had average occupancy above 10% over the entire summer period and had complete data were selected. Other buildings with insignificant occupancy are excluded because of their outlier nature. Building ID number was assigned to each building for analysis purpose and it is as follows: Ashton (1), Briscoe (2), Forest (3), Foster (4), McNutt (5), Rose (6), Teter (7), and Willkie (8).

2.3 Independent Variables

After carefully looking at wide range of data, we determined the main causes for the electricity usage. Following factors will serve as independent variables in our model:

- **Number of occupants:** The more people we have, the more electricity will be used. Our goal is to determine how exactly the usage and occupancy related are, as occupancy of a building is the only factor we can control.

[1] *http://electron.electronics.indiana.edu/weather/* accessed on April 24, 2015.

- **Daily average temperature:** The hotter it is, the more electricity will be used for cooling. Instead of using raw temperature data, these values are normalized in the following way to be used in our analysis: if μ is the average of daily average temperatures for the summer and σ is the standard deviation, then a temperature of x degrees Fahrenheit is normalized as $\frac{x-\mu}{\sigma}$.

- **Daily average humidity:** Since Bloomington summers are very humid, it is going to affect when visitors want to turn on the cooling. These values are normalized in the same way as the temperature data.

- **Food services:** Some buildings have food service offered to residents and nonresidents, changing from a small shop serving coffee and sandwiches to full sized industrial kitchens and dining halls. The significance of the food service is represented by a decimal value between 0 and 1 where 0 is for no food service and 1 is for the kitchen/cafeteria with the highest electricity usage.

- **Chiller / Central loop / Window units:** There are three different kinds of cooling systems in use and they operate very differently from one another. Each variable will take 0 or 1 value.

  - *Central loop:* If a building is on the central loop, which is a loop carrying the water chilled at one of the plants through the whole campus, then it means that we can clearly identify how much of the electricity used in the building is used for cooling, and how much is used for everything else. A building on the central loop can be 'cycled down' from the IU Physical Plant meaning that it will not be cooled down, or cool down the building to a predetermined temperature. (Briscoe, Foster, McNutt, Rose, Teter)
  - *Chillers:* If a building has a chiller on top of the building, then it means that there is a central air conditioning that is separate from the central loop. These buildings are very large in general so that they have their own chillers. (Forest, Willkie)
  - *Window units:* This is the more traditional window-unit air conditioning system. (Ashton)

2.4 Dependent Variables

The aim of our model is to be able to predict the daily total electricity usage, and it is affected by the above set of independent variables.

- **Total electricity usage including chilled water usage:** daily electricity used for a building in units of kWh. If a building is on the central loop, the chilled water usage will be included to the total electrical usage with necessary conversion.

3 Status Report

3.1 Observations

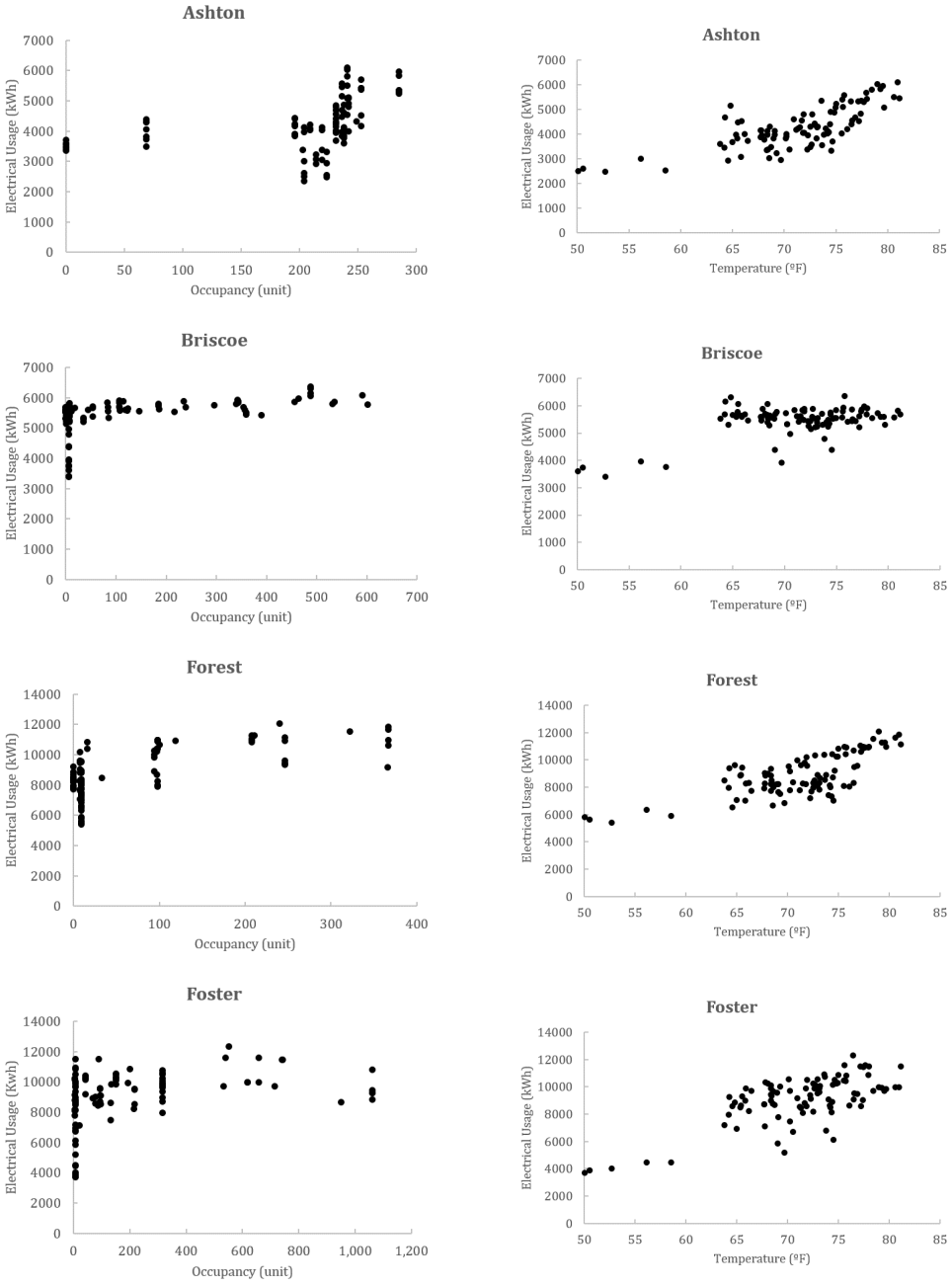
Occupancy and temperature are two major variables in our model. While other variables stay constant most of the time, these two variables are constantly changing throughout the year. This makes more difficult to predict the exact electrical usage over the summer. In order to explore the relationship, we scatter plotted usage, temperature, and occupancy. Based on scatterplots (Figure 1), we were able to observe following:

- Usage vs. Temperature
  - Distributions appear to be similar across the buildings
  - Temperature has very strong and direct effect to the usage
- Usage vs. Occupancy
  - Distributions are also similar across the buildings
  - Occupancy affects the usage, but linkage is weaker than temperature
  - Occupancy distribution is not completely linear

3.2 Mathematical Model

3.2.1 Models Considered

Our data is composed of individual data points which represent the total electricity usage of any of the buildings in one day (recall by definition that this includes chilled water usage), and groups (in our case, buildings) that these data points belong to. First of all, we decided to use a regression model, as the overall behavior of electricity usage was what we intended to model. In order to find a model fitting our system, we considered the following options:



(a) Usage vs. occupancy

(b) Usage vs. temperature

Figure 1: Scatterplots describing our observations

- One model fitting all the buildings regardless of their characteristics. This would be the easiest to come up with, however it would have been too simple a model for this project.

- Three different models, one for each different cooling system in place. This model makes sense to consider as different cooling systems imply different correlations of electricity usage with the independent variables, however, further characteristics of buildings from size to food service makes this an unfeasible option.

- Different models for each of the buildings. We decided to go with this model as the buildings are too different to be grouped with the others.

From our observations, we can see how certain independent variables are affecting the electricity usage in different ways.

- *Occupancy vs. usage* The correlation is not linear, so higher degree terms depending on occupancy were required. An exponential model proved inaccurate, so we have included a second degree term of occupancy.

- *Building-specific dependencies* Some variables like temperature, humidity, and occupancy are affecting different buildings in very different ways. For example, a slight change in temperature is not affecting Rose as much as it affects McNutt.

- *Relation between independent variables* Increases in temperature and number of occupants increases the electricity usage. However, the fact that they are both increasing also manifests itself through an extra change that would not have been there if only the temperature or only the number of occupants, so we have to take into consideration a term involving product of these variables.

- *Zero-occupancy levels* This changes from building to building, so our model has to respect this change as well.

All of these directed our attention to multilevel (hierarchical) regression models.

3.2.2 Multilevel Regression

A different model for each building would imply that we would only have 98 data points for each building, which is not enough to have meaningful analysis. However, multilevel regression is a preferred way of simulation for systems with individual data

points (the total usage in a building in one day) and groups that these individuals belong to (each building in our system, with different characteristics)[1]. By running many regressions at once group-wise and individual-wise, we arrive at a system that allows us to do partial-pooling, meaning that we put enough emphasis on these building characteristics, whereas with regular regression models there is a risk of putting too much or too little importance on these differences. We decided to choose this model as it naturally fits the type of data we have.

3.3 Analysis

With multilevel regression and the above observations in mind, we came up with the following model.

Let $i = 1, 2, \cdots, 8$ represent the building ID. Then, for each of the buildings, the functions predicting a day's total usage of a given building is as follows:

$$E_i = \beta_i + \alpha_1 O_i^2 + \alpha_2 O_i T_i + \alpha_3 H_i + \alpha_4 F_i + \alpha_5 C_i + \alpha_6 W_i + \alpha_7^i O_i + \alpha_8^i T_i \qquad (1)$$

where E is total electricity usage, O is number of occupants, T is average daily temperature, H is average daily humidity, F is the food service coefficient between 0 and 1 (0 if there is no food service, 1 for the largest kitchen and cafeteria, between 0 and 1 for the rest), C is the chiller coefficient (1 if the building is cooled down by a rooftop chiller, 0 otherwise), W is the window unit coefficient (1 if the building is cooled down by window units, 0 otherwise).

We implemented this model using the statistical software package 'R' to represent each of the buildings in our system:

$$lmer(E \sim O^2 + O \times T + H + F + C + W + (1 + O + T|B)) \qquad (2)$$

where B is the building ID. The function 'lmer' is an R function from 'lme4' package that gives a multilevel regression model as a result. The first part $O^2 + O \times T + H + F + C + W$ of this represents the variables for which we are looking for a coefficient that can be used for each of the buildings (fixed effect), and the second part $(1 + O + T|B)$ represents the variables for which we want a different coefficient (random effect), depending on the building ID, that is, changing from building to building.

To make sure that this is the model fitting the best to our system, among many other natural candidates, we manually came up with various possible models by either adding higher order terms of any of the variables appearing in the above model, or by

making the change between fixed effect to random effect or vice versa for a variable. We concluded the above model is the best among these by using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). For these criteria, the lower the score the better-fitting the model, and our model had the lowest of both of these scores among 50+ models we have considered, which is strong evidence in favor of our model.

The coefficients α_j^i and intercepts β_i are as follows:

Building ID - Name	β_i	α_7^i	α_8^i
1 - Ashton	8380.078	4.394767	864.4818
2 - Briscoe	5234.912	2.803321	431.3076
3 - Forest	9507.044	6.950052	996.8407
4 - Foster	11239.886	4.088328	1294.2844
5 - McNutt	16543.092	5.483103	1858.1290
6 - Rose Avenue	6200.875	2.662327	579.3874
7 - Teter	9456.513	3.817447	1114.8543
8 - Willkie	10228.163	4.077273	1092.7284

and $\alpha_1 = -0.002836833$, $\alpha_2 = -1.121572$, $\alpha_3 = 77.83511$, $\alpha_4 = -3329.76$, $\alpha_5 = 2282.977$, $\alpha_6 = -4890.686$.

We need to explain some of these values. The fact that α_5 is positive and α_6 is negative means that relative to central loop buildings[2], either the buildings with rooftop chillers are larger or less efficient. Similarly, either the buildings with window units are smaller or more efficient. From our data, we see that this is mainly happening due to building sizes.

The coefficient for occupancy squared, α_1, is negative, which implies that even though the electricity usage increases with occupancy, it is not linear and it reaches a plateau as occupancy gets really large. Similarly, the coefficient for occupancy times temperature, α_2, is negative, which can be explained by similar reasoning. The coefficient for food service, α_4, is negative as well, and looking at our data we see that this happens because there are small buildings with food service and there are large buildings with very little food service. There are also some large buildings with large value of food service variable, however, this is not enough to bring up this coefficient to positive. Rest of the coefficient values are intuitive.

[2]You might have noticed that the central loop as a cooling system does not appear among our variables in this model. This is due to the analysis being simpler without it as including chillers and window units as variables implies that if a building has 0 in both of those, it is assumed that the building is on the central loop. There are also issues of collinearity between these variables and the food service variable, so it is best to avoid this.

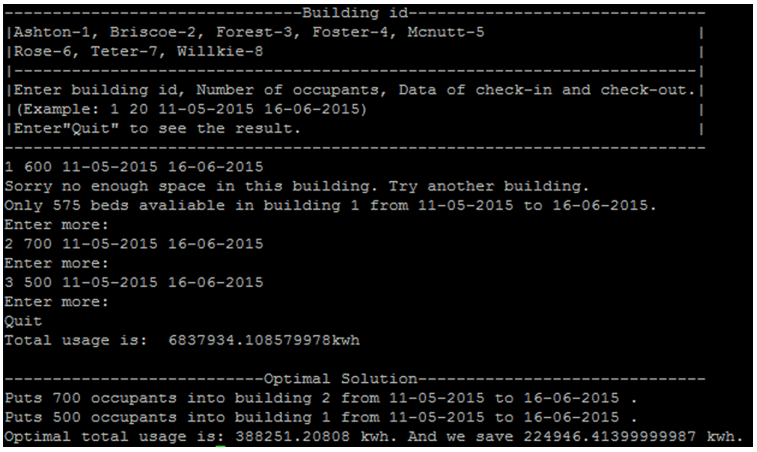


Figure 2: The program 1) asks for number of occupants, which building they would ideally be placed in, and the dates of their stay, 2) computes the electricity usage with respect to this distribution 3) finds the optimal distribution of the visitors 4) computes the electricity usage with respect to this distribution 5) computes how much electricity and money can be saved by using the optimal distribution instead of the preferred one.

Since the type cooling system in place and the type of food service are inherent characteristics of buildings, we can modify β_i to include those values as well. However, when doing the regression analysis, separating these is important, as we see much of the error can be explained by considering these factors. So, from now on, we can consider the following functions:

$$E_i = \widehat{\beta}_i + \alpha_1 O_i^2 + \alpha_2 O_i T_i + \alpha_3 H_i + \alpha_7^i O_i + \alpha_8^i T_i$$

where $\widehat{\beta}_i$ are the modified β_i. Here $\widehat{\beta}_i$ is a constant depending on the building, and the electricity usage E_i of a building is a polynomial function in independent variables occupancy, temperature, and humidity.

3.4 End Product

For the Physical Plant and RPS staff to be able to use our findings, we wrote a small program in the programming language 'Ruby', whose screenshot is above (Figure 2).

This program allows the user to enter the amount of people distributed and the duration of their stay, and the choice of the dormitory these people would be located. Then, it computes the predicted total electricity usage resulting from this arrange-

ment. It also finds out an ideal distribution of these people among the buildings in our system, computes the predicted total electricity usage for this option as well, and gives a comparison of two and how much electricity would be saved if they were to be placed in this ideal distribution instead of the requested distribution. Then, the staff can decide if the choice is worth it or not. While doing so, each group of visitors is kept together, so if there are groups which can be separated, but staying during the same time period, then it would be better to enter these groups individually, so that the program can separate them, if necessary.

3.5 Findings

Running a regression analysis without any variables allows us to compare the total residual (error) with the residual left unexplained by our model, which can be done by running the following in R:

$$lmer(1 + (1|B)).$$

Running ANOVA on this lone model, we see that the total variance is around 2192322 whereas unexplained variance in our model is 679400. Then, by computing

$$\frac{total - unexplained}{total} = \frac{2192322 - 679400}{2192322} = 69\%$$

we see that our model can account for 70% of the variance in the system, and this number can be increased by looking at larger data sets and adding other factors into consideration, which we did not do in this project.

By comparing the coefficients of different buildings for temperature, we can see which building has a smaller increment in total electricity usage with increasing temperature. Such a building must have a smaller temperature coefficient, and the ordered list of temperature coefficients is as follows:

Building ID - Name	Temperature coefficient
2 - Briscoe	431.3076
6 - Rose Avenue	579.3874
1 - Ashton	864.4818
3 - Forest	996.8407
8 - Willkie	1092.7284
7 - Teter	1114.8543
4 - Foster	1294.2844
5 - McNutt	1858.1290

So, the most efficient building seems to be Briscoe, whereas the least efficient is McNutt Quad. However, it should be noted that this computation is for zero occupancy values only, and due to the nature of $T \times O$ term in equation (1), these rankings might change when the occupancy levels of the buildings are higher. Also, building sizes affect these values greatly, so instead of looking at these raw values, we recommend using our program so that all factors are present in the computation.

We ran our program with the data from Summer 2014, to see how well it is for predicting usage by comparing it to the actual data, and to see how much can be saved by using our optimization. We have only ran this analysis for 8 of the buildings in our system, so the other buildings do not have an effect on these values. The total electricity usage in these 8 buildings last summer was 7.035.756 kWh, and our program predicts it to be 7.750.062 kWh, which is reasonably close since our program does not take into the changes in temperature but uses the average over the summer, as weather patterns are unpredictable for the coming summers from this early on, yet we can still rely on averages. With the optimization of our model, the electricity usage drops to 6.224.735 kWh, which, when converted to dollars, corresponds to a savings of $63,357.

3.6 Tools

For cleaning our data and having simple histograms and analyses, we used Excel. It was also the format in which we received the data, so it was an obvious first choice.

To analyze our data and fit into different types of regression models, we used R. It is also useful for getting plots and some of them are included in our analysis.

For writing the end-product program, we used Ruby.

4 Work Remaining and Limitations

We were able to create models to explore the electricity usage behaviors of different buildings, but while doing so, sometimes due to lack of data, and sometimes due to choices we needed to make, we did not always get the best possible results. Here we would like to explain the challenges we have experienced, and suggest possible solutions to these.

We have only worked with data coming from the summer of 2014. Having more data available from previous years can help making better predictions and reduce the error in our work. In the future, this can be solved by more closely tracking this data.

We had complete electricity usage (including chilled water equivalences if applicable) and occupancy data for 8 dormitory complexes out of 15 possible. Some complexes had unreliable data due to limitations in metering, some complexes did not have any residents during the summer of 2014, and some complexes were not open to visitors during the summer. Therefore, we were not able to include all the complexes in our model making it an imperfect model, as we have no clue about the behavior of these buildings. Having data from multiple years can help resolve this issue as well.

There were also challenges due to the theoretical nature of the goals we intended to pursue. A theoretical zero-occupancy electricity usage value for all the buildings was one of the first things we wanted to find out, but even when the central loop is cycled-down and there are no residents in a building, there are still a lot of factors which cannot be exhaustively listed here contributing to the total usage. For a more accurate model, either the definition of this theoretical minimum should be done differently, or there should be data collection for all these minor events going on in a building.

Our choice to analyze the system on a daily basis, instead of hourly, has also limited the capabilities of our models, as occupant-behavior inside a building is very different at different times of the day. However, this decision was made for simplicity and we don't think the analysis we have made resulted in inaccurate conclusions because of this choice.

The model we have decided to use is also a limiting factor. For describing the behavior of systems with groups (in our case, different buildings), the multilevel regression works great, however, it can easily be interpreted for causal inference. If we look at our model, for example, the coefficients for variables such as food, chillers, and window units, have to be interpreted correctly to see the power of this model. For this to not be a problem, we tried our best in the documentation of our findings. There are also limitations due to using inferential statistics as a means of explaining the problem. Wherever such techniques are used, one must also analyze the data summaries to reach at meaningful conclusions.

5 Future Directions

The models we have created can be improved in the future if the same type of analysis is used on larger data sets collected throughout years, as multilevel regression is arguably the best tool to analyze data sets that require partial pooling. However, it can also be taken to several different directions:

- By adding the apartment complexes in the RPS system, the model can be used for more general decision-making problems.
- Other time intervals, such as winter, fall, spring breaks, and even school semesters, can be implemented into the system for more accurate models, and any type of planning that requires the allocation of residents into the dormitories and apartments in the campus.
- More factors affecting a building's electricity usage can be accounted for in a model like ours to reduce the amount of variation and residual error that cannot be explained by our model.
- Weather patterns can be implemented into our program to have better predictions.

6 Conclusion

With all the limitations and challenges we faced, we were able to come up with a model that can predict the electricity usage accurately on a daily basis, and from this model we see that if more efficient choices were made in the summer of 2014, IU could have saved approximately $63,000 in electricity. During the research portion of this project, we visited many ideas for a possible model, and we believe we made an inform decision by choosing a multilevel regression model. Although the system can be improved virtually in many ways, this project as a core is both accurate and useful for future summers, and we were able to come up a simple program which can be further developed that can help the IU Physical Plant and RPS Staff in the summers to come.

References

[1] Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, New York: Cambridge University Press, 2007. Print.