

Spring 5-27-2015

## Using Spatiotemporal Methods to Fill Gaps In Energy Usage Interval Data

Kristin K. Graves  
*CUNY Hunter College*

[How does access to this work benefit you? Let us know!](#)

More information about this work at: [https://academicworks.cuny.edu/hc\\_sas\\_etds/3](https://academicworks.cuny.edu/hc_sas_etds/3)

Discover additional works at: <https://academicworks.cuny.edu>

---

This work is made publicly available by the City University of New York (CUNY).  
Contact: [AcademicWorks@cuny.edu](mailto:AcademicWorks@cuny.edu)

Using Spatiotemporal Methods to Fill Gaps  
In Energy Usage Interval Data

By

Kristin Kate Graves

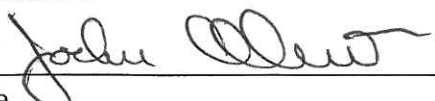
Submitted in partial fulfillment  
of the requirements for the degree of  
Master of Arts  
Hunter College of the City of New York


2015

05/01/2015  
Date

05/01/2015  
Date

Thesis Sponsor:

  
Signature  
Dr. Jochen Albrecht

  
Signature of Second Reader  
Dr. Carson J. Q. Farmer

## Table of Contents

1.	Introduction .....	1
1.1	Statement of the Problem .....	2
1.2	Purpose of the Study .....	6
2.	Literature Review .....	9
2.1	Notation .....	9
2.2	Literature Overview .....	10
2.3	Load Research Literature .....	10
2.4	Spatial Literature .....	21
2.4.1	The Effect of Neighbors on Energy Use .....	22
2.4.2	Spatial Statistics Literature .....	29
2.5	Temporal Lag Literature .....	40
2.6	Spatiotemporal Literature .....	43
2.6.1	Spatiotemporal Statistics .....	43
2.6.2	Spatiotemporal Graphics .....	48
2.6.3	Spatiotemporal Modeling .....	52
2.6.4	Spatiotemporal Literature Summary .....	55
2.7	Literature on Evaluating Success .....	56
2.7.1	Selecting Customers .....	56
2.7.2	Conducting the Evaluation .....	57
2.8	Literature Review Summary .....	60
3.	Methods and Data .....	62
3.1	Research Design .....	62
3.2	Data and Variables .....	62
3.3	Selection of the Sample Data .....	78
3.4	Overview of Data Analysis Procedures .....	79
3.5	Expected Findings .....	80
4.	Exploratory Data Analysis .....	81
4.1	Correlation Coefficients .....	82
4.1.1	Correlation Coefficients for Residential Customers .....	82
4.1.2	Correlation Coefficients for Business Customers .....	84
4.1.3	Correlation Summary .....	87
4.2	Spatial Exploratory Data Analysis .....	87
4.2.1	Inverse Distance Weighted Maps .....	89
4.2.2	Moran's I and Geary's C .....	97
4.2.3	Semivariograms .....	102
4.2.4	Spatial Exploratory Data Analysis Summary .....	106
4.3	Temporal Exploratory Data Analysis .....	107
4.3.1	Temporal Exploratory Data Analysis for Residential Customers .....	107
4.3.2	Temporal Exploratory Data Analysis for Business Customers .....	110
4.3.3	Temporal Exploratory Data Analysis Summary .....	112
4.4	Spatiotemporal Exploratory Data Analysis .....	113
4.4.1	Line Graphs .....	113
4.4.2	Spatiotemporal Correlograms .....	115

4.4.3	Spatiotemporal Semivariograms .....	117
4.4.4	Spatiotemporal Exploratory Data Analysis Summary.....	128
4.5	Summary of Exploratory Data Analysis.....	129
5.	Drawing the Sample .....	131
6.	Results of the Analysis .....	134
6.1	Discussion of Gap Filling Methods.....	134
6.2	Discussion of Evaluation Statistics .....	138
6.3	Gap Filling for the System Peak Hour.....	140
6.4	Gap Filling for Customer Peak Hour .....	143
6.5	Gap Filling for 1 Hour .....	145
6.6	Gap Filling for 3 Hours .....	147
6.7	Gap Filling for 12 Hours.....	149
6.8	Gap Filling for Customer Peak Day.....	151
6.9	Gap Filling for 24 Hours.....	153
6.10	Gap Filling for 7 Days.....	155
6.11	Gap Filling for 1 Month .....	157
6.12	Gap Filling for 3 Months .....	159
6.13	Gap Filling for 6 Months .....	161
6.14	Summary.....	163
7.	Summary of Findings and Call for Further Research .....	169
7.1	Summary of Findings.....	169
7.2	Suggestions for Future Research .....	169
7.2.1	Additional Data Transformations.....	170
7.2.2	Exploratory Data Analysis.....	170
7.2.3	Statistical Sampling.....	170
7.2.4	Additional Evaluation Methods for Gap Filling.....	171
7.2.5	Optimize Each Gap-Filling Method .....	171
7.2.6	Alternative Gap-Filling Methods.....	171
7.2.7	Solve Computer-Related Issues.....	172
7.3	Overall Summary.....	173
Appendix A:	Summary Statistics for Each Variable Used In the Analysis.....	175
Appendix B:	Correlation Coefficients for Data Groups.....	196
Appendix C:	Summary Ranges for Moran's I and Geary's C Statistics.....	217
Appendix D:	Customer Sampling for Gap-Filling.....	232
References	.....	243

## List of Figures

Figure 1.1: Example Energy Print Showing Missing Data Intervals.....	3
Figure 1.2: Example Raw Data Showing Missing Data Intervals .....	4
Figure 1.3: Conceptual Model of Spatiotemporal Energy Usage Gap Filling.....	8
Figure 2.1: Example of Missing Data to Be Filled Using Linear Interpolation.....	11
Figure 2.2: Example of Relationship Between Temperature and Energy Usage.....	15
Figure 2.3: Van Raaij's Behavioral Model of Residential Energy Use .....	25
Figure 2.4: Example Inverse Distance Weighted Map .....	31
Figure 2.5: Example Semivariogram Plot .....	35
Figure 2.6: Example Correlogram Using Sample Autocorrelation Function.....	42
Figure 2.7: Structure of the Knox Index .....	44
Figure 2.8: Example of Spatiotemporal Semivariogram.....	48
Figure 2.9: Example of Space-in-Time or Heat Map .....	49
Figure 2.10: Example of a Line Graph .....	51
Figure 3.1: Gap Lengths for Residential Customers .....	64
Figure 3.2: Gap Lengths for Business Customers .....	65
Figure 3.3: Distribution of Interval Data Gaps for Residential Customers .....	66
Figure 3.4: Distribution of Interval Data Gaps for Business Customers .....	67
Figure 4.1: Residential Customer Locations.....	88
Figure 4.2: Business Customer Locations.....	89
Figure 4.3: Residential IDW Map for June 9 at Hour 17 .....	91
Figure 4.4: Residential IDW Map for January 12 at Hour 16 .....	92
Figure 4.5: Residential IDW Map for January 13 at Hour 16 .....	93
Figure 4.6: Business IDW Map for July 22 at Hour 16 .....	94
Figure 4.7: Business IDW Map for July 21 at Hour 18 .....	95
Figure 4.8: Business IDW Map for June 9 at Hour 17.....	96
Figure 4.9: Residential Semivariogram for July 22, Hour 16 .....	103
Figure 4.10: Residential Semivariogram for January 12, Hour 18.....	104
Figure 4.11: Business Semivariogram for June 9, Hour 17 .....	105
Figure 4.12: Business Semivariogram for July 12, Hour 17 .....	106
Figure 4.13: Residential Correlogram, Raw Energy Usage Interval Data, No Stratification .....	108
Figure 4.14: Residential Correlogram, Raw Energy Usage Interval Data, Stratum 5 .....	109
Figure 4.15: Business Correlogram, Energy Usage Interval Data as Percent of Prior Hour's Interval, Overall .....	111
Figure 4.16: Line Graphs for Two Random Residential Customers and Their Neighbors .....	114
Figure 4.17: Line Graphs for Two Random Business Customers and their Neighbors .....	115
Figure 4.18: Spatiotemporal Correlogram for Residential Customers .....	118
Figure 4.19: Spatiotemporal Correlograms for Additional Residential Customers.....	118
Figure 4.20: Spatiotemporal Correlogram for Business Customers .....	119
Figure 4.21: Spatiotemporal Correlograms for Additional Business Customers.....	119
Figure 4.22: Two-Dimensional Spatiotemporal Semivariogram for Residential Customers .....	121
Figure 4.23: Three-Dimensional Spatiotemporal Semivariogram for Residential Customers .....	122

Figure 4.24: Two-Dimensional Spatiotemporal Semivariogram for Residential Customers, With Boundaries.....	123
Figure 4.25: Three-Dimensional Spatiotemporal Semivariogram for Residential Customers, With Boundaries.....	124
Figure 4.26: Two-Dimensional Spatiotemporal Semivariogram for Business Customers.....	125
Figure 4.27: Three-Dimensional Spatiotemporal Semivariogram for Business Customers .....	126
Figure 4.28: Two-Dimensional Spatiotemporal Semivariogram for Business Customers, With Boundaries.....	127
Figure 4.29: Three-Dimensional Spatiotemporal Semivariogram for Business Customers, With Boundaries.....	128

## List of Tables

Table 1.1: Number of Missing Intervals .....	4
Table 3.1. Overview of Basic Variables Used in the Analysis.....	68
Table 3.1. Overview of Basic Variables Used in the Analysis.....	70
Table 3.2. Overview of Hourly Interval Energy Usage Data Used in the Analysis .....	71
Table 3.3. Overview of Billing Determinants Used in the Analysis.....	72
Table 3.4. Overview of Weather Data Used in the Analysis.....	73
Table 3.5. Overview of Building-Specific Data Used in the Analysis.....	74
Table 3.6. Overview of Customer Demographics from Census Tract Data Used in the Analysis.....	75
Table 3.7. Overview of Building Demographics from Census Tract Data Used in the Analysis .....	77
Table 4.1: Independent Variables Correlated with Raw Energy Usage Intervals (k1-k24) for Residential Customers .....	83
Table 4.2: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of the Daily Maximum Value (pctd1-pctd24) for Residential Customers .....	84
Table 4.3: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Monthly Energy Use in kWh (pctm1-pctm24) for Residential Customers .....	84
Table 4.4: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Annual Energy Use in kWh (pcta1-pcta24) for Residential Customers .....	84
Table 4.5: Independent Variables Correlated with Raw Energy Usage Intervals (k1-k24) for Business Customers .....	85
Table 4.6: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of the Daily Maximum Value (pctd1-pctd24) for Business Customers .....	86
Table 4.7: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Monthly Energy Use in kWh (pctm1-pctm24) for Business Customers .....	86
Table 4.8: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Annual Energy Use in kWh (pcta1-pcta24) for Business Customers .....	86
Table 4.9: Moran's I Ranges for Dependent Variables for Residential Customers.....	99
Table 4.10: Geary's C Ranges for Dependent Variables for Residential Customers.....	100
Table 4.11: Moran's I Ranges for Dependent Variables for Business Customers .....	100
Table 4.12: Geary's C Ranges for Dependent Variables for Business Customers .....	101
Table 4.13: Average Correlation of Temporal Lags for Residential Customers .....	110
Table 4.14: Average Correlation of Temporal Lags for Business Customers .....	112
Table 5.1: Hourly Interval Energy Usage Data Gaps Used in the Research -- Residential..	132
Table 5.2: Hourly Interval Energy Usage Data Gaps Used in the Research -- Business.....	133
Table 6.1: Analysis Results for System Peak Hour -- Residential.....	142
Table 6.2: Analysis Results for System Peak Hour -- Business.....	142
Table 6.3: Analysis Results for Customer Peak Hour -- Residential.....	144
Table 6.4: Analysis Results for Customer Peak Hour -- Business.....	144
Table 6.5: Analysis Results for One Hour -- Residential .....	146
Table 6.6: Analysis Results for One Hour -- Business.....	146
Table 6.7: Analysis Results for Three Hours -- Residential .....	148
Table 6.8: Analysis Results for Three Hours -- Business .....	148
Table 6.9: Analysis Results for Twelve Hours -- Residential .....	150

Table 6.10: Analysis Results for Twelve Hours -- Business.....	150
Table 6.11: Analysis Results for Customer Peak Day -- Residential.....	152
Table 6.12: Analysis Results for Customer Peak Day -- Business .....	152
Table 6.13: Analysis Results for Twenty-Four Hours -- Residential.....	154
Table 6.14: Analysis Results for Twenty-Four Hours -- Business.....	154
Table 6.15: Analysis Results for Seven Days -- Residential .....	156
Table 6.16: Analysis Results for Seven Days -- Business.....	156
Table 6.17: Analysis Results for One Month -- Residential.....	158
Table 6.18: Analysis Results for One Month -- Business.....	158
Table 6.19: Analysis Results for Three Months -- Residential.....	160
Table 6.20: Analysis Results for Three Months -- Business.....	160
Table 6.21: Analysis Results for Six Months -- Residential.....	162
Table 6.22: Analysis Results for Six Months -- Business.....	162
Table 6.23: Rating of Gap-Filling Methods by Gap Length -- Residential.....	164
Table 6.24: Rating of Gap-Filling Methods by Gap Length -- Business.....	164
Table 6.25: Rating of Gap-Filling Methods by Evaluation Statistic -- Residential.....	165
Table 6.26: Rating of Gap-Filling Methods by Evaluation Statistic -- Business.....	165
Table 6.27: Best Evaluation Results for Each Gap Length -- Residential.....	167
Table 6.28: Best Evaluation Results for Each Gap Length -- Business .....	167
Table A.1: Summary Statistics for Residential Analysis Variables .....	175
Table A.2: Summary Statistics for Business Analysis Variables .....	186
Table B.1: Correlation Coefficient Ranges for Residential Customer Data Groups - Raw Energy Usage Intervals (k1-k24).....	197
Table B.2: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Daily Maximum Value (pctd1-pctd24).....	199
Table B.3: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Monthly Billed kWh (pctm1-pctm24) .....	201
Table B.4: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Annual Billed kWh (pcta1-pcta24) .....	203
Table B.5: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Prior Interval (delt1-delt24).....	205
Table B.6: Correlation Coefficient Ranges for Business Customer Data Groups - Raw Energy Usage Intervals (k1-k24) .....	207
Table B.7: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Daily Maximum Value (pctd1-pctd24).....	209
Table B.8: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Monthly Billed kW (pctm1-pctm24).....	211
Table B.9: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Annual Maximum Billed kW (pcta1-pcta24) .....	213
Table B.10: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Prior Interval (delt1-delt24).....	215
Table C.1: Moran's I and Geary's C Ranges for Residential Customer Data Groups.....	218
Table C.2: Moran's I and Geary's C Ranges for Business Customer Data Groups .....	225
Table D.1: Residential Customer Sample for Gap-Filling of System Peak Hour .....	232
Table D.2: Business Customer Sample for Gap-Filling of System Peak Hour .....	232
Table D.3: Residential Customer Sample for Gap-Filling of Customer Peak Hour .....	233



Table D.4: Business Customer Sample for Gap-Filling of Customer Peak Hour .....	233
Table D.5: Residential Customer Sample for Gap-Filling of One Hour .....	234
Table D.6: Business Customer Sample for Gap-Filling of One Hour .....	234
Table D.7: Residential Customer Sample for Gap-Filling of Three Hours .....	235
Table D.8: Business Customer Sample for Gap-Filling of Three Hours .....	235
Table D.9: Residential Customer Sample for Gap-Filling of 12 Hours .....	236
Table D.10: Business Customer Sample for Gap-Filling of 12 Hours .....	236
Table D.11: Residential Customer Sample for Gap-Filling of Customer Peak Day .....	237
Table D.12: Business Customer Sample for Gap-Filling of Customer Peak Day .....	237
Table D.13: Residential Customer Sample for Gap-Filling of 24 Hours .....	238
Table D.14: Business Customer Sample for Gap-Filling of 24 Hours .....	238
Table D.15: Residential Customer Sample for Gap-Filling of 7 Days .....	239
Table D.16: Business Customer Sample for Gap-Filling of 7 Days .....	239
Table D.17: Residential Customer Sample for Gap-Filling of One Month .....	240
Table D.18: Business Customer Sample for Gap-Filling of One Month .....	240
Table D.19: Residential Customer Sample for Gap-Filling of Three Months .....	241
Table D.20: Business Customer Sample for Gap-Filling of Three Months .....	241
Table D.21: Residential Customer Sample for Gap-Filling of Six Months .....	242
Table D.22: Business Customer Sample for Gap-Filling of Six Months .....	242

# 1. Introduction

Researchers analyzing spatiotemporal or panel data, which varies both in location and over time, often find that their data has holes or gaps. This thesis explores alternative methods for filling these gaps and also suggests a set of techniques for evaluating those gap-filling methods to determine which works best. The specific focus of this research is hourly interval energy usage data, as collected by an energy provider. Therefore, this thesis will also enable a better understanding of temporal and spatial patterns of energy use, and the interaction of temporal and spatial factors on that use.

Energy providers (electric, gas, steam, and water) need to understand the usage patterns of their customers so that they can ensure that there is sufficient supply to meet customer demand, avoid overloading their delivery infrastructure, and correctly allocate the cost of providing service to each of several customer classes. To accomplish these goals, most energy providers collect detailed usage data for a random sample of their customers that can be extrapolated to various customer populations. If the collected data are incomplete or otherwise unavailable, it can have a negative impact on the reliability of the extrapolated results.

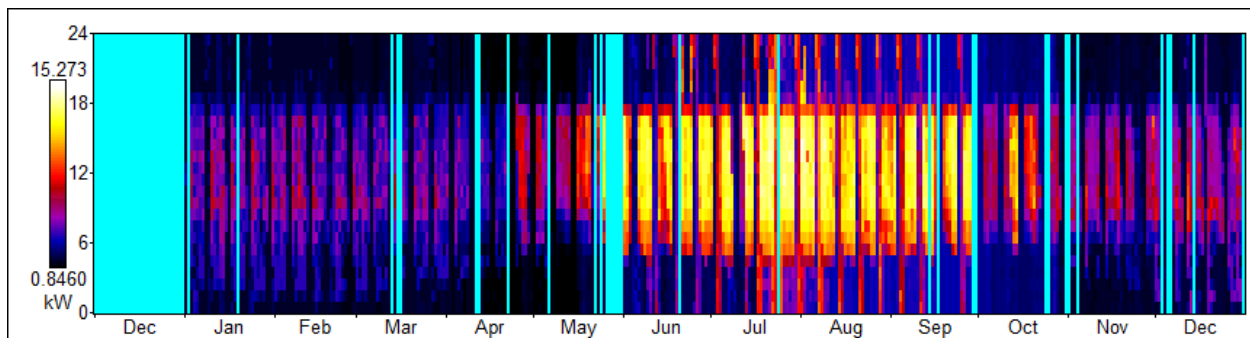
Energy providers employ a diverse assortment of temporal methods to fill gaps in a customer's usage pattern, typically making use of temperature and the customer's historical usage (AEIC 2001; KEMA 2011; Lim and Yao 2012; Mathis 2007; McMenamin and Monforte 1997; Mondragon 2009; Schiermeyer 2006; Smith and Hanna 2008). Although spatial methods are regularly used to fill data gaps in other contexts (e.g., health data as in Auchincloss 2007 or weather data as in Dirks 1998), they have not been applied to energy usage data. Furthermore, while spatiotemporal methods for data exploration and analysis are available, they have generally not been used to fill gaps in data (Andrienko et al 2010; Christakos, Bogaert, and Serre 2002), although recent work includes the gap-filling of air pollution data at semiweekly intervals (Lindström et al 2011).

## **1.1 Statement of the Problem**

The process by which detailed customer usage data are collected, reviewed, extrapolated, and analyzed is referred to as Load Research. A major input to the Load Research process is a stream of energy usage data in hourly intervals for a calendar year for each member of a statistical sample of energy customers. This thesis focuses on one of the many steps of the Load Research process: the efforts to fill gaps in the detailed hourly interval energy usage data.

Data gaps can be created by defective equipment, data communication failures, data storage problems, and human error (AEIC 2001 p. 6-7). It is desirable to accurately fill the data gaps so that as many sampled customers as possible are available for extrapolation, thus preserving the integrity of the randomly selected sample and achieving as high a precision as possible for the extrapolated results. If the holes cannot be filled, then the sample size can be severely degraded, reducing the reliability of the results.

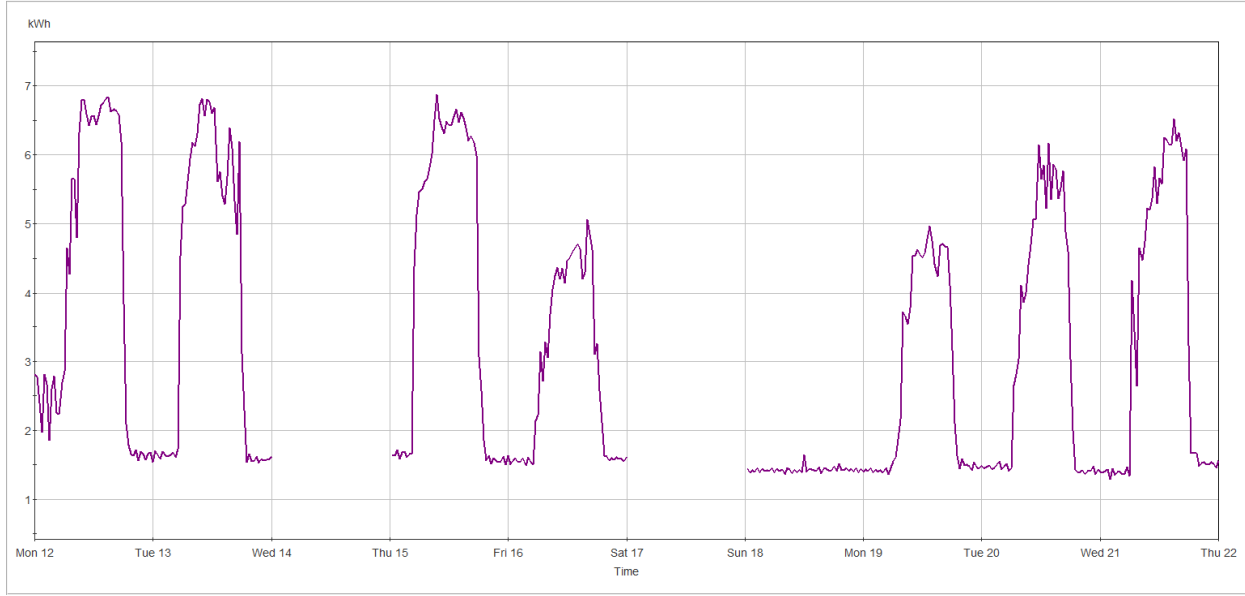
Figure 1.1 uses EnergyProbe™ software to illustrate the problem of missing hourly interval energy usage data for a sample customer. The horizontal axis shows the days of the year from December 1 of one year through to December 31 of the following year, and the vertical axis shows the hours of the day from 1 to 24. The colors represent the quantity of electricity used during each hour of the year, as seen in the color key at the left side. EnergyProbe™ uses dark colors to represent times of low usage, yellow and white to represent times of highest usage, and bright blue to indicate missing data.



**Figure 1.1: Example Energy Print Showing Missing Data Intervals**

Figure 1.2 provides another way of viewing missing data, also illustrated by EnergyProbe™. In this view, hours for ten consecutive days are shown on the horizontal axis and the kWh value of each hourly interval energy usage value is shown on the vertical axis. Missing data appear as gaps in the line.

Table 1.1 provides an estimate of the percentage of total intervals that are missing in an illustrative year.



**Figure 1.2: Example Raw Data Showing Missing Data Intervals**

**Table 1.1: Number of Missing Intervals**

Customer Class	Number of Customers with Interval Data	Possible Number of Hourly Interval Energy Usage Values (number of customers times 8,760)	Number of Missing Intervals	Percent of Intervals That Are Missing
Residential	911	7,980,360	493,031	6.2%
Business	364	3,188,640	518,833	16.3%

Energy providers have developed several methods to fill these holes in usage data. One problem that energy providers face, however, is that the filled data tend to be less "peaky" and more smoothed than actual customer usage data. Because one of the primary uses of the data is to estimate the customer contributions to system peak usage, it is critical that the filled data represent the missing peaks and valleys of usage as accurately as possible (Richardson et al 2010, p. 1884). Load Researchers want to neither smooth existing peaks nor create new ones.

Another issue to keep in mind is that the set of explanatory variables is limited to those that are readily available to energy providers. Energy providers have billing systems

that contain a complete list of their customers, known as a "data frame" in the parlance of statistics. Although the billing systems are comprehensive in terms of covering all customers, their available data elements are generally limited to those that directly serve customer billing or other energy provider purposes. Most often, few if any additional data elements are available for the Load Research sample of customers. Therefore, to be readily usable by energy providers, the methods to be tested must be usable with a limited set of data elements. This thesis research also includes makes use of geodemographic data that are not readily available to energy providers, as a way of assessing the benefits of obtaining these types of data.

For all customers, the readily-available data elements are: customer identification number, customer billing service class (determines customer's rate), monthly billed kiloWatt-hour usage (total usage during bill period), monthly billed kiloWatts (maximum usage during bill period, available for larger customers only), latitude, and longitude. Additionally, for the sampled Load Research customers the kiloWatt-hour (kWh) usage during each time interval is known. For the Load Research sample customers, additional geodemographic data elements have been collected. For the customer's block and lot, data captured are lot footprint, number of buildings and floors, year built, building area, total number of units and of residential units, floor-area ratio, and building area per unit and per residential unit. For the customer's census tract, data captured are median household income; percent of the population that are people of various ages, of various races, and of various nationalities; average household size, percent of households with electric heat, and percent of householders with various levels of education.

## 1.2 Purpose of the Study

In a narrow sense, Load Research departments at energy providers expend a significant amount of time manually reviewing and editing customer hourly interval energy usage data. If a gap-filling method can be found that is accurate and can be automated, it could reduce this workload and allow Load Research analysts to focus their attention on other problems. Additionally, hourly interval energy usage data are an input to the rates charged to customers of energy providers; therefore, having more accurate hourly interval energy usage data will also benefit energy customers by ensuring that their rates are calculated using more accurate estimates of their energy usage when actual measurements are not available.

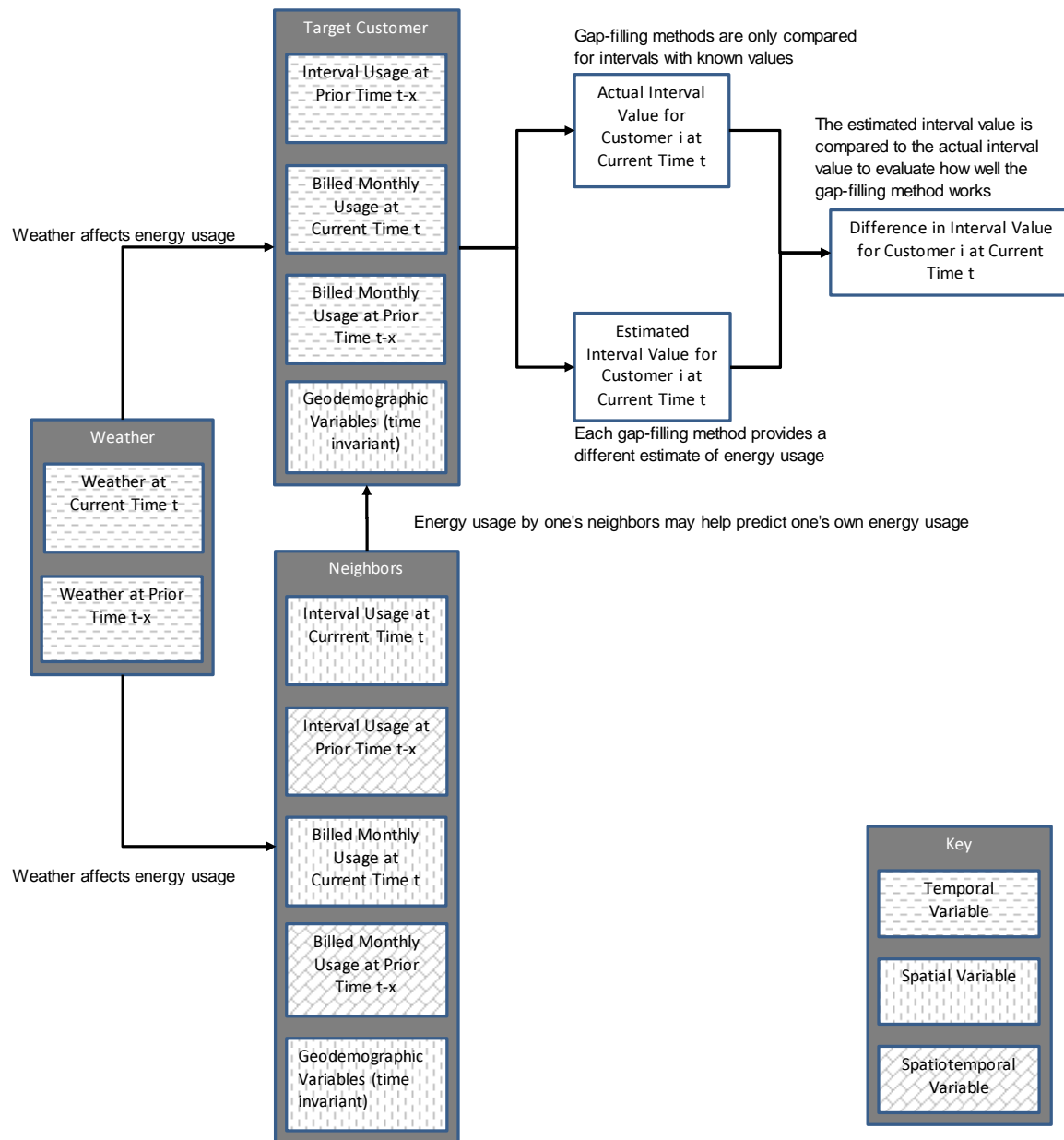
More generally, because energy providers typically focus solely on temporal factors, the information garnered here about spatial factors can prove beneficial. Additionally, any researcher trying to fill gaps in spatiotemporal data could potentially benefit from this study and its comparisons of both gap filling methods and techniques to evaluate the accuracy of those methods.

To reach these goals, the following objectives will be met:

- Artificially create hourly interval energy usage data gaps for individual energy customers.
- Fill the gaps using a variety of methods, including temporal, spatial, and spatiotemporal models.
- Compare the filled data values resulting from each method to the actual values to determine which method differs least from the actual values.

Figure 1.3 provides a conceptual model of the planned research. Weather plays an important role in energy usage, affecting both the target customers (the customers whose data gaps are being analyzed) and their neighbors. In a temporal analysis, only the customer's own data, which varies over time, is considered; therefore, current and historical weather values and historical usage by the target customer are viewed as the driving forces behind the target customer's energy usage during a particular time interval. In a spatial analysis, the only factors considered are those occurring in different locations but simultaneously; therefore the driving forces are the concurrent and historical usage by one's neighbors along with the geodemographic data elements. In a spatiotemporal analysis, both locational and temporal influences are considered; therefore the potential driving forces are concurrent and historical values for weather, current and historical energy usage by one's neighbors, historical usage by the target customer, and the geodemographic variables. After the data gaps are filled, the filled values are compared to the actual values to measure how well the various gap-filling methods perform.





**Figure 1.3. Conceptual Model of Spatiotemporal Energy Usage Gap Filling**

## 2. Literature Review

### 2.1 Notation

O'Sullivan and Unwin (2010, p. 374) describe a set of standardized notation for spatial analysis. In this thesis, their notation is expanded to include the temporal dimension and other factors that are important during different aspects of this analysis:

$x$  = Easting or longitudinal geographic coordinate of the customer's location.

$y$  = Northing or latitudinal geographic coordinate of the customer's location.

$t$  = Temporal identity of the customer's hourly interval energy usage.

$z$  = Numerical value of the customer's hourly interval energy usage.

$n$  = Number of customer locations.

$k$  = Number of customer locations in a spatial neighborhood.

$d$  = Distance between customer locations.

$w$  = Strength or weight of interaction between customer locations.

$s$  = An arbitrary (x,y) location.

$cdh$  = Cooling degree hour, a measure of the need for air conditioning.

$hdh$  = Heating degree hour, a measure of the need for space heating.

$h$  = Hour of the day.

$dow$  = Day of the week.

Although elevation is often included in geographic analysis, in this thesis it is not accurately measurable and is therefore not considered. Energy providers are unlikely to have reliable data on the floor or building story on which their customers are located; further complicating the matter is that a single customer may occupy multiple floors of one building or multiple buildings with different numbers of floors.

Regarding the temporal dimension, patterns in energy usage can be distinguished at multiple levels, including hours of the day, days of the week, and seasonally, with weather as a key component (AEIC 2012, p. 1). Therefore, hourly interval energy usage data is not typically viewed as a continuous stream, but rather as combinations of multiple temporal patterns.

## **2.2 Literature Overview**

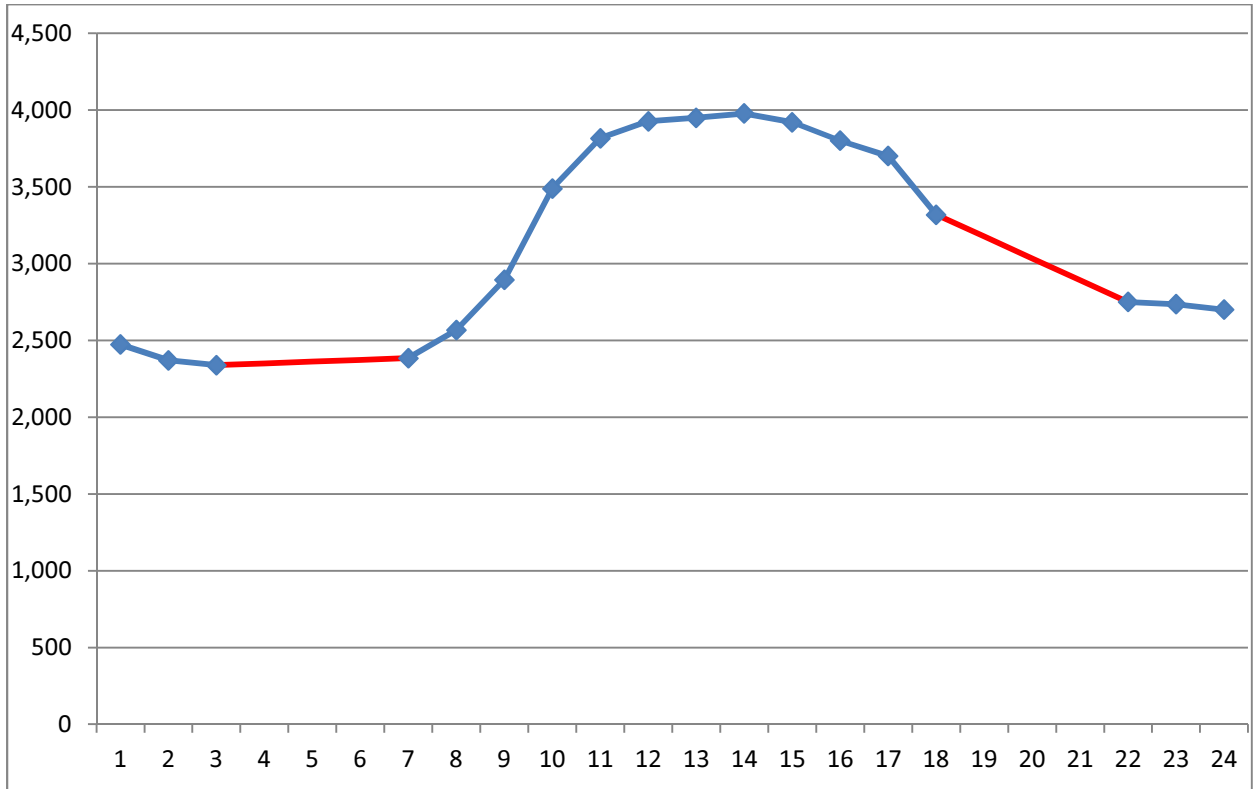
Five bodies of literature are examined herein to understand and address the problem of filling gaps in hourly interval energy usage data: existing gap-filling methods used in Load Research, spatial analysis methods that show promise for addressing gaps in geographic data, methods that can be used to find appropriate temporal lags in time-series data, spatiotemporal methods available for data that simultaneously have both a spatial and temporal nature, and literature on how to evaluate gap-filling success or failure.

## **2.3 Load Research Literature**

In Load Research literature, three techniques are typically proposed for filling gaps in interval data: interpolation, similar day(s), and regression estimates. All methods are

specific to a particular customer location, in that only the customer's own existing data are used to fill its missing data.

The simplest method is to fill a missing data value via linear interpolation between the two available data values that temporally surround the missing interval(s), as illustrated in Figure 2.1. Here, the blue line with the data markers represents the existing data while the red line represents the interpolated values.



**Figure 2.1: Example of Missing Data to Be Filled Using Linear Interpolation**

In mathematical terms, usage for customer for a particular missing time interval is interpolated from the values of the two surrounding intervals for the same customer and day:

$$z_{x,y,dow,h} = \frac{h-(h-1)}{(h+1)-(h-1)} \times (z_{x,y,dow,h+1} - z_{x,y,dow,h-1})$$

where  $z_{x,y,dow,h}$  is the hourly interval energy usage for the customer located at longitude  $x$  and latitude  $y$ , on day of the week  $dow$  at the missing hour  $h$ . The term  $h-1$  designates the non-missing hour prior to the missing hour  $h$  while  $h+1$  is the non-missing hour following hour  $h$ . Therefore,  $z_{x,y,dow,h+1}$  is the hourly interval energy usage for the customer on the same day  $dow$  but at a subsequent hour  $h+1$ , and  $z_{x,y,dow,h-1}$  is the hourly interval energy usage at a prior hour  $h-1$ .

Mathis et al (2007) reports that data gaps of two hours or less can be successfully filled via linear interpolation. Mondragon (2009) also suggests the use of linear interpolation, but does not provide a time-frame limit for its use. For longer gaps, Mathis et al (2007) suggests finding up to three similar days, averaging their load shapes, filling the gap using the averaged values, and then adjusting up or down to consumption totals. Mathis reported that the methodology for longer gaps was not always successful, primarily because of the difficulty in finding like-days and the need to exclude aberrant days with highly fluctuating loads from inclusion as like-days.

In mathematical terms, the average day method is as follows:

$$z_{x,y,dow,h} = \frac{\sum_{t=1}^T z_{x,y,dow+t,h}}{T}$$

where  $dow$  is the day on which hourly interval energy usage data are missing,  $dow+t$  is a day of the same day-type as  $dow$  but on which hourly interval energy usage data are not missing, and  $T$  is the number of days that are averaged.

Mondragon (2009) compares two similar-day methods. In an experiment, he compares the use of a five-day average of similar day-types (i.e., weekdays and weekends/holidays) to the use of the prior day load shape (i.e., use the load shape of the previous similar day). Mondragon notes that when the temperature is stable, the five-day

average can be better than the previous day method, but his overall recommendation is to use the previous day's load shape (presumably a similar day) because it provides the best estimate under most circumstances.

A similar day method is also recommended by the Load Research Committee of the Association of Edison Illuminating Companies, publisher of the *Load Research Manual*. The current edition suggests that analysts should simply substitute similar days or hours from the customer's own history (AEIC 2001, p. 6-8), but offers no specifics. In Load Research practice, a similar day is often considered to be another day that is the same day of the week, although Saturday and Sunday are sometimes considered to be similar to each other, as are the days of Tuesday, Wednesday, and Thursday.

In mathematical terms, the similar day method for replacing a missing interval is the following:

$$Z_{x,y,dow,h} = Z_{x,y,dow+t,h}$$

where  $dow+t$  is a prior or subsequent day of the same day-type as  $dow$  but on which hourly interval energy usage data are not missing.

The three techniques described above (interpolation, similar day, and average days) require extensive manual review of and interaction with the data. Although these techniques may provide adequate to good results, they are simply not practical in situations where there is any significant quantity of data to be reviewed. When a single utility in California is installing more than 9 million Smart Meters (Wood 2010, p. 6), any technique that requires manual intervention for each customer would be prohibitively

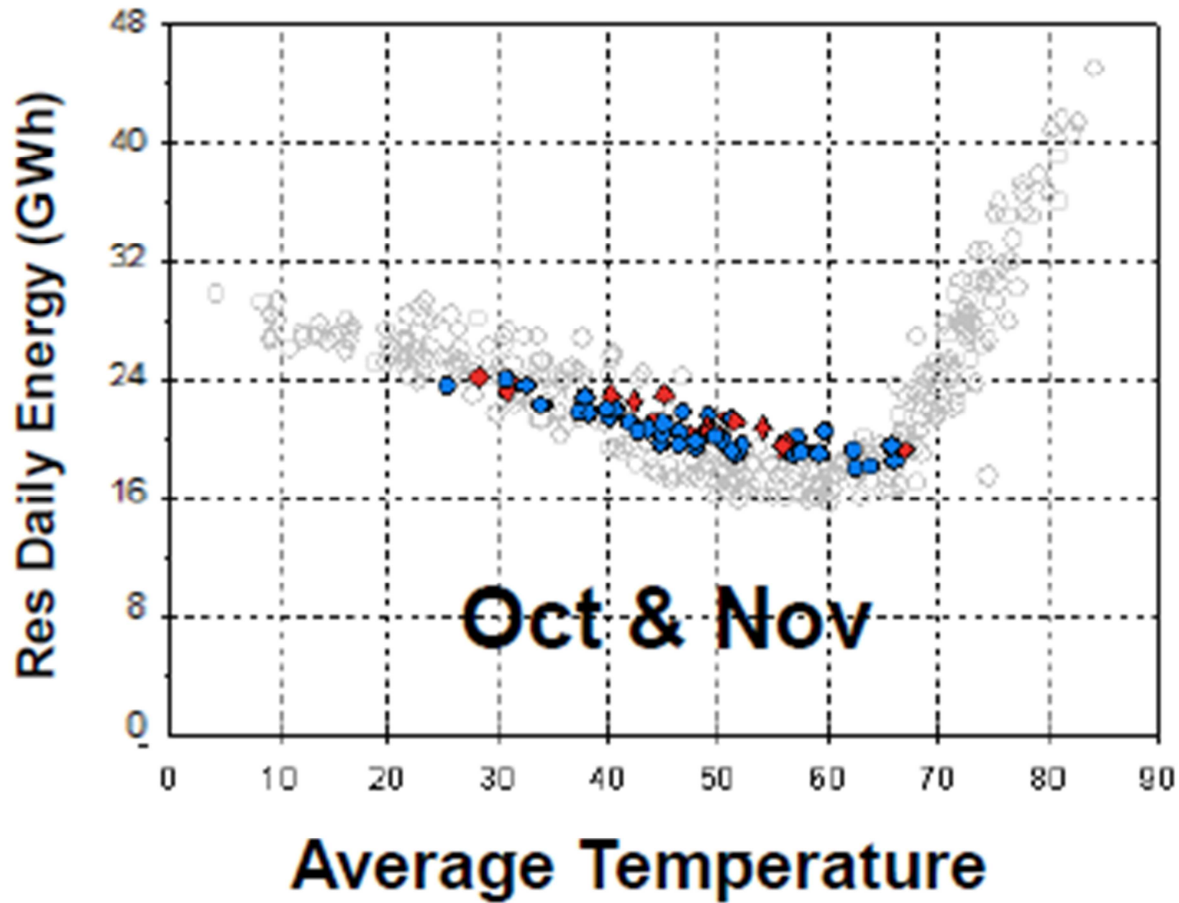
time-consuming to implement<sup>1</sup>. Therefore, these methods are not viable for the types of applications explored in this research.

The fourth method for gap filling explicitly includes the idea that energy use is a function of temperature, and that energy use generally increases as the temperature gets colder or hotter, but that there is some middle temperature range in which energy use is minimized. In other words, a graph of energy use as a function of temperature is approximately V- or U-shaped, perhaps with a flat horizontal area between the two sloped sides of the V. Such a relationship is illustrated in Figure 2.2, below, in which each gray dot represents one day, and the October and November days are highlighted in red and blue.

Such methods were initially developed either to calculate energy savings from conservation measures or to provide weather-normalized estimates of energy usage data. One of the first models developed in this way, for heating load, was the Princeton Scorekeeping Method (PRISM; Fels 1986). PRISM assumes that there is a linear relationship (i.e., a constant heating slope) between changes in daily temperature and corresponding changes in daily energy use. Additionally, PRISM allows the reference temperature for heating, above which the premise follows its baseline consumption level, to be estimated as part of the model (Fels 1986, p. 8).

---

<sup>1</sup> As mentioned in Chapter 1, data gaps have several causes including defective equipment, data communication failures, data storage problems, and human error, all of which can occur with smart meters. In a perfect world there would be no need for data interpolation, but smart meters do not solve all of these problems.



**Figure 2.2: Example of Relationship Between Temperature and Energy Usage<sup>2</sup>**

In mathematical terms, a customer's missing daily usage is statistically calculated by estimating the following regression equation:

$$z_{x,y,dow} = a_{x,y,dow} + b_{x,y,dow} \times degree\ day_{dow} + e_{x,y,dow}$$

where  $a_{x,y,dow}$  is the baseline daily kWh consumption level for the customer on days of day-type  $dow$ ,  $b_{x,y,dow}$  is the slope of the relationship between the *degree day* weather measure and the customer's hourly interval energy usage level on days of day-type  $dow$ , and  $degree\ day_{dow}$  is a weather measure calculated by taking the absolute value of the difference between the average temperature on a day and a reference temperature. In a

<sup>2</sup> Figure taken from McMenamin 2011a.



heating season (the portion of the year during which a customer would typically use space heating), *degree day* only takes a non-zero value if the average actual temperature for that day is below the reference temperature. In a cooling season (the portion of the year during which a customer would typically use air conditioning), *degree day* only takes a non-zero value if the average actual temperature for that day is above the reference temperature. Finally, in the above equation,  $e_{x,y,dow}$  is the error term.

The use of degree day temperatures as a weather measure is common in energy usage studies because it converts the U- or V-shaped relationship between temperature and energy usage (as discussed above) into a linear relationship. If the temperature is very low, the degree day value will be high because it is well below the heating reference temperature. As the temperature gets slightly warmer, the degree day value falls because it is closer to the heating reference temperature. At some point the temperature is above the heating reference temperature and the degree day value is zero. When the temperature climbs further, above the cooling day value, the degree day value again takes on a positive value. As the temperature climbs even further, the degree day value continues to climb.

Although PRISM was designed to measure monthly energy savings for heat conservation measures, it has been expanded and adapted by KEMA, Inc. (2011) as a procedure for filling gaps in interval consumption data taking both heating and cooling loads into account. KEMA (2011, p. 293-301) uses a particular implementation in which each interval of each day-type has its own regression equation. For example, the 52 intervals that end at 10:00 a.m. on Tuesdays are analyzed in a linear regression in which the independent variable is the weather during the interval ending at 10:00 a.m. on

Tuesdays; Tuesdays at 11:00 a.m. make use of a separate linear regression, and so on. In mathematical form, the KEMA method is described thusly:

$$z_{x,y,dow,h} = a_{x,y,dow,h} + b_{x,y,dow,h} \times hdh_{dow,h} + c_{x,y,dow,h} \times cdh_{dow,h} + e_{x,y,dow,h}$$

where  $a_{x,y,dow,h}$  is the baseline hourly interval energy usage level for the customer on days of day-type  $dow$  at hour  $h$ ,  $b_{x,y,dow,h}$  is the slope of the relationship between the heating degree hour ( $hdh$ ) weather measure and the customer's consumption level on days of day-type  $dow$  at hour  $h$ , and  $hdh_{dow,h}$  is a weather measure calculated by taking the absolute difference between the temperature on day  $dow$  at hour  $h$  and a reference heating temperature. The variable  $hdh_{dow,h}$  only takes a non-zero value if the actual temperature is below the reference heating temperature. The variable  $c_{x,y,dow,h}$  is the slope of the relationship between the cooling degree hour ( $cdh$ ) weather measure and the customer's consumption level on days of day-type  $dow$  at hour  $h$ , while  $cdh_{dow,h}$  is a weather measure calculated by taking the absolute difference between the temperature on day  $dow$  at hour  $h$  and a reference cooling temperature. The variable  $cdh_{dow,h}$  only takes a non-zero value if the actual temperature is above the reference cooling temperature. The error term is  $e_{x,y,dow,h}$  for the customer on days of day-type  $dow$  at hour  $h$ .

The KEMA method is widely used by Load Researchers, although it frequently provides results that are not satisfactory upon close inspection (extraordinarily high peaks can result, especially when there is too much missing data).<sup>3</sup> However, the KEMA method will be included in this research as a baseline against which other methods can be compared.

---

<sup>3</sup> The KEMA method also assumes that a weekly time lag is the best predictor, an issue that I will explore later in Section 4.3 of this thesis.

Smith and Hanna (2008, p. 24) opt to model hourly loads using a single regression equation with a series of dummy variables for each hour, rather than conducting separate regression equations for each interval as proposed by KEMA. Additional dummy variables are included to distinguish different day-types and months:

$$z_{x,y,dow,h} = a_{x,y,dow,h} + b_{x,y,dow,h} \times hdh_{dow,h} + c_{x,y,dow,h} \times cdh_{dow,h} + \sum_{t=1}^{T-1} (d_t \times dow \text{ dummy}) \\ + \sum_{m=1}^{M-1} (m_m \times monthly \text{ dummy}) + \sum_{r=1}^{R-1} (h_r \times hourly \text{ dummy}) + e_{x,y,dow,h}$$

where  $d_t$  is the slope of the relationship between the day-type dummy variable *dow dummy* and the customer's consumption level on days of day-type *dow* at hour *h*,  $m_m$  is the slope of the relationship between the monthly dummy variable and the customer's consumption level on days of day-type *dow* at hour *h*, and  $h_r$  is the slope of the relationship between the hourly dummy variable and the customer's consumption level on days of day-type *dow* at hour *h*.

Schiermeyer (2006, p. 6) includes all of the above variables, in addition to a three-day weighted average of heating and/or cooling build-up and hours of daylight.

Related to the above-described temperature-sensitivity techniques is McMenamin's (2011a) suggestion of the use of a linear spline function with HDD and CDD. A linear spline is a joining of two distinct linear regressions at a point referred to as a knot; the two regression lines join at an angle, generally not that of 180°. Mathematically, a spline is a system of equations:

$$z_{x,y,dow,h} \text{ from cooling} = a_{x,y,dow,h} + b_{x,y,dow,h} \times cdh_{dow,h} + e_{x,y,dow,h}$$

$$z_{x,y,dow,h} \text{ from heating} = a_{x,y,dow,h} + b_{x,y,dow,h} \times hdh_{dow,h} + e_{x,y,dow,h}$$

such that when  $hdh=cdh$ , then:

$$z_{x,y,dow,h} \text{ from cooling} = z_{x,y,dow,h} \text{ from heating}$$

Like Smith and Hanna (2008), McMenamin suggests that the temperature response function differs during different seasons or months, thus necessitating the inclusion of dummy variables for different seasons or months into the above equations.

A modification of the temperature-related methods is to identify which customers have temperature-sensitive usage patterns and to implement the temperature-related methods to those customers. Smith and Hanna, for example, use the Spearman Rank-Order Correlation Coefficient,  $r_s$ , to divide customer service classes (rate classifications) into high, medium, or low weather-sensitivity (2008, p. 14). In this formulation, the "low" weather-sensitivity customers (those with an  $r_s$  measure of less than 0.4) utilize a non-temperature-based gap filling method. The "medium" ( $r_s \geq 0.4$  and  $r_s < 0.7$ ) and "high" ( $r_s > 0.7$ ) sensitivity customers are analyzed separately. The cut-points between weather sensitivity levels are arbitrarily chosen by the authors (Smith and Hanna 2008, p. 15).

Raish also looks at weather sensitivity as a factor in analyzing daily energy use, but focuses on the regression coefficient,  $R^2$ , to separate weather-sensitive loads from non-weather sensitive loads. Specifically, Raish considers an  $R^2$  greater than or equal to 0.6 to be weather-sensitive (2007, p. 49), while a non-temperature-based gap filling method is used for lower  $R^2$  values. Raish provides no explanation for his 0.6  $R^2$  cut-off for weather sensitivity.

From the above set of dummy-variable methods, this research will include Smith and Hanna's single regression with dummy variables. This method structurally differs from the KEMA method, but is easy to automate and can handle any length of data gap.

A fifth approach is suggested by McMenamin and Monforte (1997, p. 145), who use an artificial neural network model to develop hourly load shapes. A neural network is flexible in that it can model situations where there is an unknown non-linear relationship between the variables (Francis 2001, p. 255). Neural networks incorporate one or more "hidden layers," each with multiple nodes, all of which modify the data and pass it on to the next layer. The model trains itself, increasing or decreasing the strength of nodes and layers as it learns to predict the dependent variable (Francis 2001, p. 258). Although they can readily handle large amounts of data and complex relationships, neural networks are often criticized because they do not provide specific information about the functional form of the final relationship between the independent variables and the dependent variable, thus the term "hidden layer." The general form of a neural network is the following:

$$z_{-}(x, y, dow, h) = f(X, B) + e_{-}(x, y, dow, h)$$

where  $X$  is a vector of explanatory variables for the customer and  $B$  is a vector of parameters for the customer.

McMenamin and Monforte found that neural networks provided good results in forecasting hourly load data using the following input data elements (1997, p. 154):

- Weather variables: coincident temperature, daily high, daily low, cumulative temperature, temperature gradient, humidity, wind speed, and cloud cover.
- Calendar variables: day of the week, month/season, holidays, days near holidays, sunrise, and sunset.
- Lagged loads: the previous day's morning usage, the previous day's afternoon usage, the same hour's usage from the previous day, and the same hour's usage from two days prior.

In a more recent study, Lim and Yao also used a neural network model to develop hourly load shapes. Lim and Yao's list of input variables (2012, p. 24) includes many factors that are not available to most energy providers about their individual customers (e.g., number of adults and children in the home, household income, and appliance ownership) along with variables that *are* typically available (e.g., temperature, tenure, location). Regardless of what variables are used, the neural network method provides a unique methodology, can be readily automated, and can handle energy usage gaps of any length. It should be noted that neural networks are particularly susceptible to large differences in the values of dependent and independent variables (McCaffrey 2014). For this reason, input data to the neural network are typically normalized by subtracting the minimum value and then dividing by the range of observed values, i.e., the maximum minus the minimum (Francis 2001, p. 263).

From the Load Research literature, then, there are three techniques that bear sufficient merit and distinction from each other to incorporate into this analysis. These are the KEMA method that uses a separate regression for each customer-hour, the Smith and Hanna method that uses dummy variables for each customer-hour, and McMenamin's neural network method.

## **2.4 Spatial Literature**

Spatial methods are based on Tobler's first law of geography, which states that "everything is related to everything else, but near things are more related than distant things" (Tobler 1970, p. 236). In other words, one's characteristics are related to those of one's neighbors. But does energy usage follow Tobler's law?

Although energy providers have implemented numerous methods to fill gaps in energy usage data, the geographic proximity of each customer to its neighbors is almost universally ignored during the process. However, there are proximity-related concepts that may relate to energy usage levels and energy consumption patterns.

#### **2.4.1 The Effect of Neighbors on Energy Use**

Researchers in the fields of geography, sociology, and psychology have studied the notions of neighborhood effects, localization economies, segregation, and social capital, all of which link those who live or work in close proximity. Energy use has been found to be one of the linked behaviors.

Hayes et al (1977, p. 425) note that there are two dimensions to energy use behaviors: levels of energy consumption and patterns of energy consumption. Overall levels of energy consumption, such as yearly or monthly variation for the same customer or from one customer to another, are generally related to factors such as weather, building construction, occupancy, equipment, and appliance stock (Ritchie et al 1981, p. 237). The pattern of energy consumption refers to usage at different times of the day or days of the week, and is based primarily on changes in building occupancy and usage of equipment within the building.

Kaplan et al discuss “neighborhood effects” (2009, p. 279), in which neighbors and neighborhoods exert peer pressure on the behavior of the local residents. Additionally, neighborhoods can provide role models that engage in various activities, thus helping to socialize others into those activities and values (p. 280). Although the discussion in Kaplan is focused on employment-related behaviors, there is little reason to think that social pressure would only affect this single aspect of behavior.

The concept of agglomeration economies stems from Alfred Weber's 1909 theory of the location of industries. There are two types of agglomeration economies in Weber's framework, both involving "agglomeration" (the clustering of activities) and "economies" (savings that stem from the clusters, Weber 1929):

- Urbanization economies are a clustering of unlike or dissimilar industries that results from access to larger labor markets, access to financial markets, and the sharing of urban infrastructure (Hartshorn 1992, p. 121; Kaplan et al 2009, p. 170).
- Localization economies are the clustering of similar kinds of industry that results from access to a specialized workforce, an array of producer services supported by the manufacturing complex, the ability for the firms to cooperate and communicate, and the ability to have greater levels of specialization (Hartshorn 1992, p. 122; Kaplan et al 2009, p. 170).

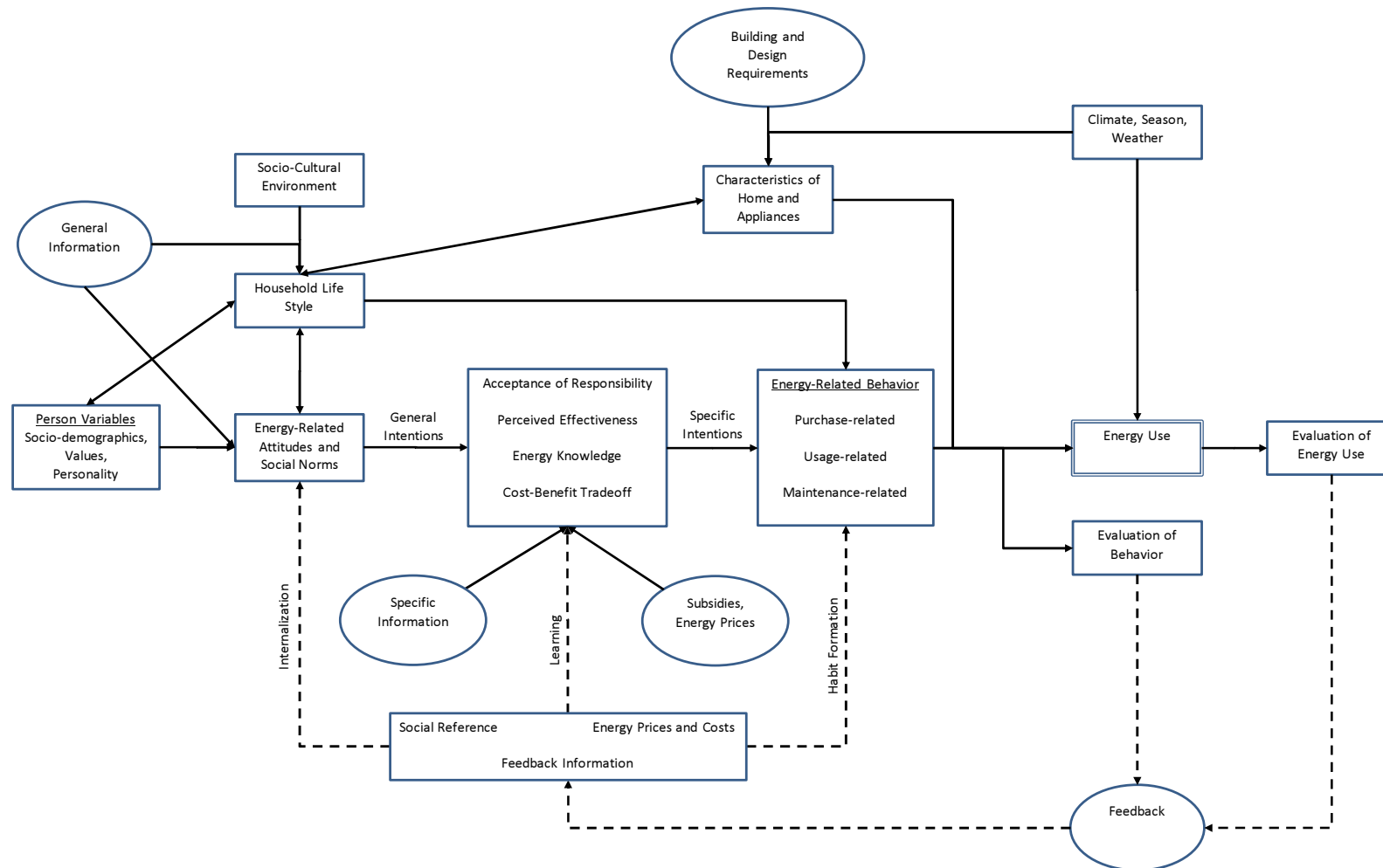
Localization economies result in the clustering of like businesses, just as segregation results in the clustering of like individuals within a neighborhood. In the case of residences, three factors affect segregation: economics, discrimination, and preferences (Kaplan et al 2009, p. 265). Segregation can be externally imposed or self-selected (or a combination of the two), but the end result is that there are nearly uniform residential subareas in cities (Hartshorn 1992, p. 251). Within these uniform neighborhoods there is less variation in social practices than exists between different neighborhoods.

Stern notes that "energy use is not a behavior but an outcome of behavior" (1992, p. 1226). Van Raaij et al propose a behavioral model of residential energy use (1983, p. 121), shown in Figure 2.3, in which one's social-cultural environment, energy-related attitudes, and social norms affect the household lifestyle. Household lifestyle, in turn, feeds into the



characteristics of one's home and appliances, both of which directly impact energy-related behavior and use. Van Raaij et al present two primary types of energy behavior: usage-related behavior (which would impact both energy consumption level and the pattern of energy use), and purchase-related behavior (which would most impact energy consumption level).

Although much of the available literature on energy use is targeted at energy conservation, conservation is complementary to energy use and certainly affects both the level and pattern of energy usage. Energy conservation became a topic of serious interest following the energy crisis in the late 1970s. During the 1980s and early 1990s, a plethora of government programs encouraged energy conservation, which led to numerous academic studies of the issue.



**Figure 2.3: Van Raaij's Behavioral Model of Residential Energy Use**

Foster et al, looking at motivations for energy conservation, state that individuals will change their behavior through peer pressure to align with the ideas or beliefs because of a desire to belong (2010, p. 2). In an article about dissemination of energy-related information, Dennis et al note that people "rely on their peers to determine which innovations are ... desirable" (1990, p. 1110). These findings expand the neighborhood effects to energy use and conservation, and incorporate neighborhood effects into energy usage decision-making.

In determining what energy-using appliances one purchases and how they are used, for example, the receipt of information has been found to be an important factor. Yates and Aronson note that face-to-face interactions have more impact than more concrete information, "even when the more vivid information is less representative" (1983, p. 437). Generally, people can be swayed by the report of even a single individual (Yates and Aronson 1983, p. 438). They tend to purchase equipment that they can observe or try out, and base much of their performance on the behavior of their peer group (Dennis et al 1990, p. 1111).

Essentially, neighbors may be "keeping up with the Joneses" in terms of what energy-using appliances they are buying and how these appliances are used. Conversely, neighbors make-do with less in poorer neighborhoods, even to the point of unsafe conditions (Shai 2006, p. 150). As Manski notes, "persons in the same group tend to behave similarly" (2000, p. 23). These findings are in support of psychology's social diffusion theory, in which the examples of others are more effective than advertising (Yates et al 1983, p. 439). Social diffusion theory, with its emphasis on behavior modification through example, is psychology's equivalent to geography's neighborhood effects.

Localization economies and residential segregation result in businesses or residences, respectively, with similar attributes locating near to one another spatially. When this happens, there are two overarching ways in which energy usage levels and patterns can be affected: similarities in structures, and similarities in denizens (residents or workers).

Similarities in structures mean that the buildings in a neighborhood may be similar to each other. Homes within a neighborhood may have been built around the same time, and thus be of similar size, built of similar materials, and have similar insulation. Apartment buildings typically have vertical “lines” of apartments in which all units in the line have the same size and layout. Stern notes, for example, that “builders, developers, and building owners usually select the appliances, furnaces, and insulation levels in new buildings” (1992, p. 1225). Housing stock is important in energy use due to factors such as insulation, the efficiency of heating and cooling systems, and the number and kind of household appliances (Van Raaij et al 1983, p. 127). Chetty et al (2008) note that important factors include the size of the building (p. 245), the size and layout of the space (p. 248), and the age of the building (p. 243). To the extent that, for example, households reside in an apartment complex, a residential development, or even in a neighborhood where the homes are of a similar vintage, the homes and appliances will be reasonably similar to those of their neighbors. Similarly, if businesses are located in an office complex or industrial park, the buildings and equipment are likely to be similar to those of nearby businesses. Peschiera et al note, however, that even in identical buildings, there can be large differences in energy consumption (2010, p. 1329); these differences are larger than

those associated with energy conservation, and are instead attributable to differences in behavior by the building occupants (p. 1330).

Similarity in denizens refers to the residents and workers in the neighborhood. Ritchie et al note that "the homogeneity of the neighborhood reduced potential variance due to [similarity in] dwelling differences and family demographics" (1981, p. 234). Waldfogel agrees, stating that, to the extent that preferences relate to characteristics such as race, income, age, and ethnicity, then these preferences "will be stimulated by concentration of like individuals" (2010, p. 181). Van Raaij et al (1983, p. 127) state that energy-related life style and habits are formed at least partly from the family's composition and income. Additionally, Van Raaij et al refer to the fact that these home characteristics are in turn matched to the residents who have purchased them (p. 127). Under localization economies, similar businesses are located near to one another. In all these cases, one's energy consumption level and pattern can be expected to be somewhat similar those of one's neighbors.

According to Putnam, social capital "refers to features of social organization such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefit" (1995, p. 67). The idea is that life is easier if there is a good stock of social capital because it results in trust, reciprocity, communication, and collaboration (p. 67).

Numerous authors have noted the impact of social groups on both energy consumption levels and patterns. Van Raaij et al argue that both behavior and lifestyle are affected by one's "network of social contacts" (1983, p. 137), including friends, neighbors, and colleagues; these contacts impact both the dissemination of information and social comparisons (p. 137). Van Raaij et al (1983, p. 127), for example, note that energy-related

life style and habits are partly determined by the family's hobbies, club membership, magazine subscriptions, and the like. In an article about social networks and their ability to affect energy conservation, Mankoff et al discuss the impact that membership groups, including churches, can have in motivating change (2007, p. 3).

Stern adds that "group membership matters" (1992, p. 1229), and that energy-usage information disseminated by community groups is perceived as being of higher quality due to the groups' credibility with the consumers (p. 1228). Van Raaij et al agree, reporting that the diffusion of information is stronger in neighborhoods with many social contacts and active organizations (1983, p. 138), that is, in neighborhoods with stronger social capital.

It is apparent from the above discussion that data related to customer premises and neighborhoods may be of relevance in explaining energy usage. Additionally, weather is seen as a key factor. In this thesis, then, geodemographic variables and weather data will be included as potential explanatory variables.

#### **2.4.2 Spatial Statistics Literature**

Following Tobler's Law and supported by the research mentioned above, it is appropriate to look to geography's spatial methods to fill gaps in hourly interval energy usage data. Spatial data are typically classified into one of five types: point, line, areal, network, and raster. For analysis of customer hourly interval energy usage data, it is appropriate to treat each customer as a point location. Although the footprint of a customer's specific location may well be related to energy usage, energy providers typically do not have this information at hand. Because of this, there is no easy method by which customers can be identified as anything other than a point location. Similarly, although

customers are spread over the landscape, their uneven distribution throughout an energy provider's service territory is not conducive to raster representation. The spatial literature discussed here, then, is limited to that appropriate to point data.

#### **2.4.2.1 Inverse Distance Weighted Maps**

If Tobler's Law holds for hourly energy usage, then one's energy usage at any given hour should be more related to that of one's nearer neighbors than to that of those farther away. Inverse distance weighted (IDW) maps are one way to visually assess whether or not this is true, by producing a continuous contour-like surface from irregularly-spaced data.

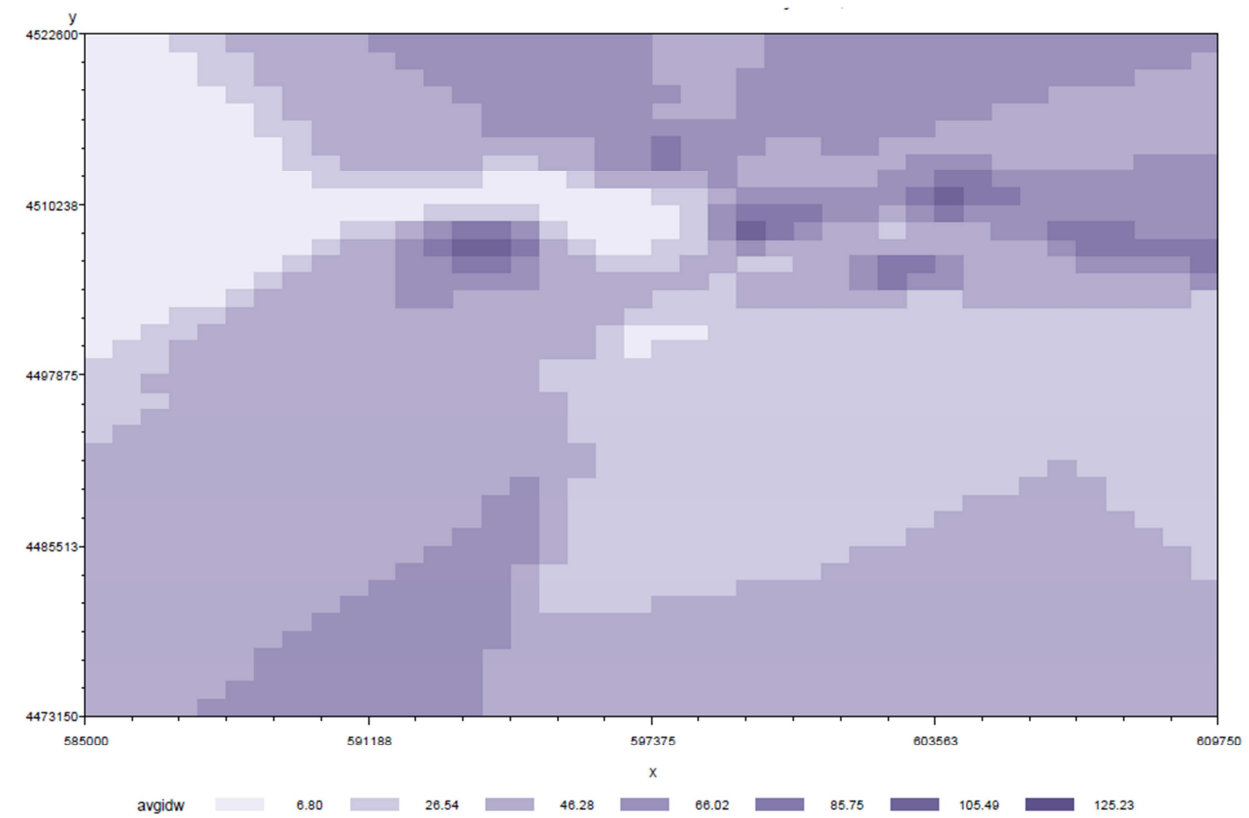
The procedure for IDW maps is, for each target location, to first select a set of neighbors, typically either all neighbors within a finite radius or the a finite number of the  $n$  nearest neighbors (usually somewhere between 4 and 10 neighbors), or some combination of the two (Shepard 1968, p. 519). The Cartesian distance is then calculated between the target location  $z$  and each of the selected neighbors; because Cartesian distances are desired, UTM coordinates (rather than longitude and latitude) should be used to ensure that the distances are consistently calculated. The following equation shows the calculation of the weighted average hourly interval energy usage value at a target location  $s$ , calculated from  $k$  neighbors:

$$z_{x_s, y_s, dow, h} = \sum_{k=1}^K z_{x_k, y_k, dow, h} * \frac{\frac{1}{\sqrt{(x_s - x_k)^2 + (y_s - y_k)^2}}}{\sum_{k=1}^K \frac{1}{\sqrt{(x_s - x_k)^2 + (y_s - y_k)^2}}}$$

Once the neighbor-weighted values have been calculated for each of the  $s$  target locations, the values are presented in a map. A sample IDW map is provided in Figure 2.4. Here, the

darker areas are regions of higher hourly interval energy usage in a particular day-hour combination, and the lighter areas are regions of lower hourly interval energy usage. If there were no regional patterns in hourly interval energy usage during the hour, the entire map would be a single shade.

IDW maps can be helpful in exploratory analyses, to assess whether or not there is spatiality in the data set. However, because IDW maps contain no temporal element, separate maps need to be created for each day-hour combination.



**Figure 2.4: Example Inverse Distance Weighted Map**

#### **2.4.2.2 Spatial Autocorrelation and Semivariance**

Autocorrelation is the correlation of any variable with itself, generally measured through the dimension of time. Spatial autocorrelation measures the correlation of a



variable with itself in space, and refers to the fact that attribute data from nearby locations are more likely to be similar than data from more distant locations (O'Sullivan and Unwin 2010, p. 199). If Tobler's Law does not hold for a particular attribute, then that attribute will be distributed randomly over space rather than being clustered. Positive spatial autocorrelation means that similar values occur near to one another, and negative spatial autocorrelation is when dissimilar values are near one another (O'Sullivan and Unwin 2010, p. 207).

If spatial autocorrelation is found, then spatial analytic techniques are appropriate. Two common tests for spatial autocorrelation are used: Moran's I and Geary's C. Moran's I is calculated using cross-products of deviations from the global mean:

$$I = \frac{n}{\sum_{s=1}^n (z_s - \bar{z})^2} \times \frac{\sum_{s=1}^n \sum_{j=1}^n w_{sj} (z_s - \bar{z})(z_j - \bar{z})}{\sum_{s=1}^n \sum_{j=1}^n w_{sj}}$$

where  $\frac{n}{\sum_{s=1}^n (z_s - \bar{z})^2}$  is the number of observations divided by the overall variance for the dataset. This term ensures that  $I$  is not large merely because the values and variability are large (O'Sullivan and Unwin 2010, p. 206). The covariance term is  $\sum_{s=1}^n \sum_{j=1}^n w_{sj} (z_s - \bar{z})(z_j - \bar{z})$ , in which  $s$  and  $j$  refer to different spatial zones, and  $z$  is the datum value. We find the products of the differences in each zone from the overall mean value of  $z$ , as a way of determining if the differences co-vary. If  $z_s$  and  $z_j$  are both larger or both smaller than the mean, then their product is positive. If one is larger and one is smaller than the mean, however, then their product is negative. The sum of the products will depend on how close the zonal values are to the overall mean. The products are weighted by  $w_{sj}$ , which is a spatial weights matrix that takes a higher value when the zones are spatially near one

another (O'Sullivan and Unwin 2010, p. 205). The term  $\sum_{s=1}^n \sum_{j=1}^n w_{sj}$  is the sum of the spatial weights (O'Sullivan and Unwin 2010, p. 206).

Moran's I takes a positive value if the data are positively spatially autocorrelated and takes a negative value if there is negative spatial autocorrelation. A value close to zero indicates no spatial autocorrelation, but if the Moran's I value is outside the range of -0.3 and 0.3 it is an indication that the data have a relatively strong autocorrelation (O'Sullivan and Unwin 2010, p. 206). O'Sullivan and Unwin note that Moran's I is "effectively the correlation coefficient for the relationship between the attribute values and the local mean attribute values" (p. 208).

Geary's C is an alternative to Moran's I, but makes use of deviations of one observation from another, thus it is more sensitive to local differences:

$$C = \frac{n-1}{\sum_{s=1}^n (z_s - \bar{z})^2} \times \frac{\sum_{s=1}^n \sum_{j=1}^n w_{sj} \times (z_s - z_j)^2}{2 \times \sum_{s=1}^n \sum_{j=1}^n w_{sj}}$$

where  $\frac{n-1}{\sum_{s=1}^n (z_s - \bar{z})^2}$  is the number of observations (minus one) divided by the overall variance for the dataset. This term ensures that  $C$  is not large merely because the values and variability are large (O'Sullivan and Unwin 2010, p. 211). The term  $\sum_{s=1}^n \sum_{j=1}^n w_{sj} \times (z_s - z_j)^2$  is the squared difference in  $y$  between the areas under consideration, weighted by the distance between the areas. The value is larger when there are large differences between nearby observations (O'Sullivan and Unwin 2010, p. 211). The term  $2 \times \sum_{s=1}^n \sum_{j=1}^n w_{sj}$  is the normalizing factor for the combined spatial weights (O'Sullivan and Unwin 2010, p. 211).

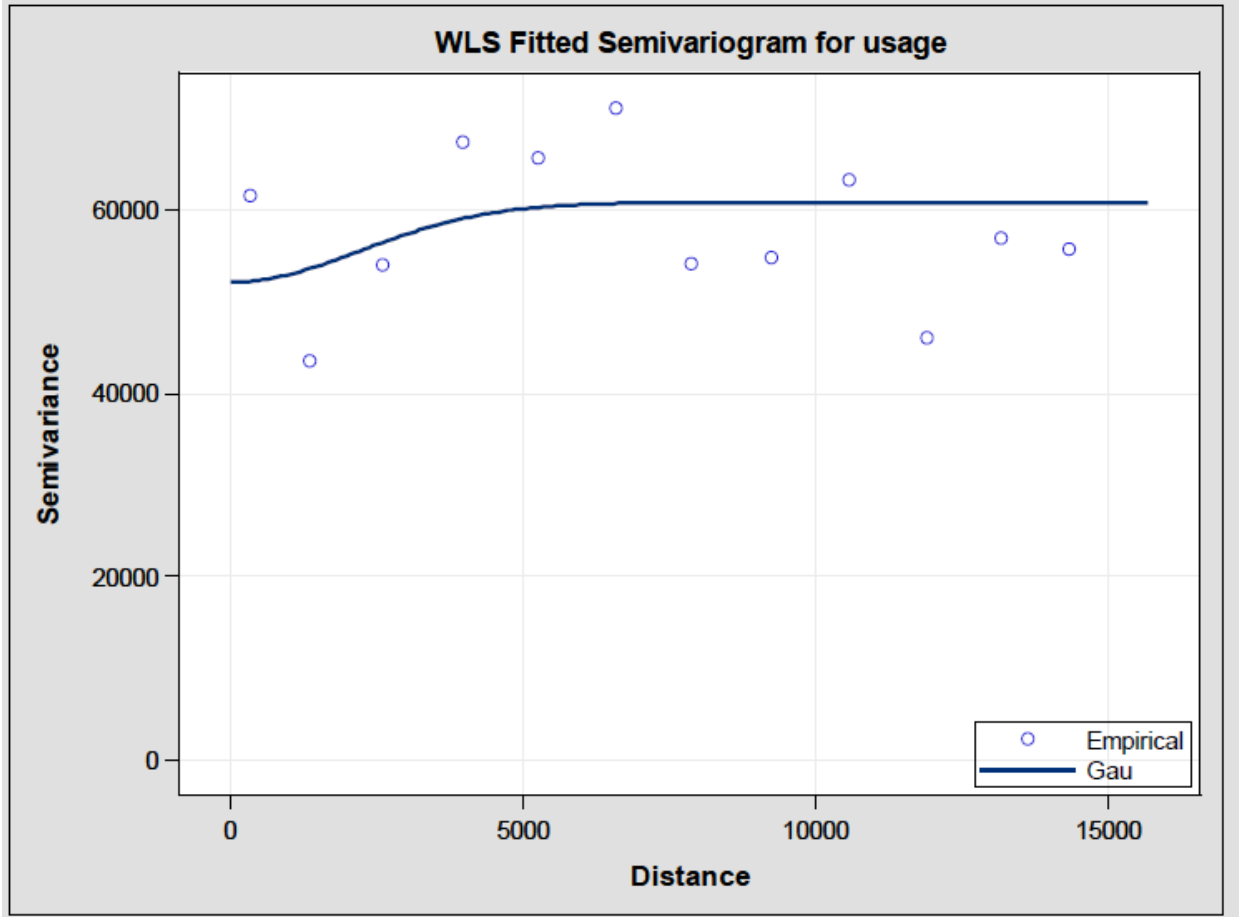
A Geary's C value of one indicates no spatial autocorrelation. Values greater than or equal to 0 but less than one indicate positive spatial autocorrelation. Values greater than one indicate negative spatial autocorrelation.

Another common autocorrelation statistic is semivariance, which mathematically describes how the variance of the variable of interest changes with distance (O'Sullivan and Unwin 2010, p. 295):

$$\hat{\gamma}(d) = \frac{1}{2 \times n(d)} \times \sum_{d_{ij}=d}^n (z_i - z_j)^2$$

where  $\hat{\gamma}(d)$  represents the semivariance for distance  $d$ , where  $d$  is a distance range,  $\gamma$  is the standard symbol for semivariance, and the "hat" indicates that we are making an estimate. The division by two stems from the original development of the idea by Georges Matheron (O'Sullivan and Unwin 2010, p. 293). The term  $\frac{1}{n(d)}$  is the inverse of the number of observations of distance  $d$  from one another (O'Sullivan and Unwin 2010, p. 293). The term  $\sum_{d_{ij}=d}^n (z_i - z_j)^2$  is the sum of squared differences in the attribute  $z$  between all pairs of points at distance  $d$  from one another (O'Sullivan and Unwin 2010, p. 293).

The semivariance statistic  $\gamma$  is most often plotted as a continuous function of the distance between observations, describing the way that the variance of the field changes with distance (O'Sullivan and Unwin 201, p. 295). Figure 2.5 provides an example of such a plot, known as a semivariogram. The distance between pairs of observations are shown on the horizontal axis, and the semivariance statistic is on the vertical axis.



**Figure 2.5: Example Semivariogram Plot**

After a semivariogram model has been developed, the results can then be used to interpolate values of the dependent variable for all locations, whether measured or not. The interpolation uses a weighted sum of local values, taking into account both the spatial relationship of each target location to the observed points and the relationship between the observed points as described by the semivariogram (O'Sullivan and Unwin 2010, p. 302). Based on the semivariance results, the weight assigned to each observed point decreases with the distance of that point from the target location, as in the IDW maps described above.

Methods used to detect spatial autocorrelation, such as IDW maps, Moran's I, Geary's C, and semivariograms, provide (at best) only a broad-brush interpolation methodology on their own. They do, however, provide a means by which the potential value and relevance of other spatial methods can be readily assessed. Therefore, testing for spatial autocorrelation is a logical first step in any spatial analysis. Because Moran's I and Geary's C can detect spatial autocorrelation at different spatial realms, it is appropriate to use both methods in this research. IDW maps are a visual tool that can provide insight into the existence of spatial autocorrelation in hourly interval energy usage data.

#### **2.4.2.3      *Geographically Weighted Regression***

The classical multiple regression model has the following assumptions: there is a linear relationship between each of the explanatory variables and the dependent variable, no exact linear relationship exists among any combination of the explanatory variables, the error term (the residuals left over after all known relationships have been accounted for in the regression) has an expected value of zero and constant variance for all observations, errors corresponding to different observations are uncorrelated, and the error variable has a normal distribution (Pindyck and Rubinfeld 1976, p. 55). With spatial and spatiotemporal data, however, it is expected that the nearby locations will have similar data values, that the model's residuals will vary by location, and that the variance in the residuals will vary by location (Charlton and Fotheringham 2009, p. 3). This phenomenon can have two effects. First, the spatial nature of the model residuals can cause the model parameters to be inefficient, meaning that the standard errors of the parameters are artificially inflated. This, in turn, makes the parameters seem to be less significant than they really are. Second, to the extent that the dependent variable in one observation is

influenced by the explanatory variables in other observations, the estimated parameter values can be both biased, meaning that the estimated parameters are either too high or too low, and inefficient.

If spatial structure, such as spatial autocorrelation, exists in model residuals, then it ought to be accounted for. Geographically weighted regression (GWR) accounts for spatial structure by allowing the model to vary spatially. Specifically, the regression coefficients are allowed to vary from place to place. Not only do the variables themselves change from one location to another, but the relationships between the variables are also allowed to change (O'Sullivan and Unwin 2010, p. 228). In its simplest form, the GWR concept involves partitioning the dataset into a number of regions and estimating a local regression model for each region. Specifically, GWR constructs a local model at every location in the study area, with all data points included in each local model. Each data point is assigned a spatial weight based on its proximity to the modeled location, and the weights are included in the model via weighted linear regression (O'Sullivan and Unwin 2010, p 228).

Charlton and Fotheringham (2009, p. 7) discuss the calculation of the weights, concluding that the kernel (or weighting scheme) should have a Gaussian-like shape, meaning that nearby points should be given a higher weight. They conclude that the bandwidth of the weighting scheme, that is, the distance at which neighboring points are no longer included, is more important than the specific function through which the weights decline with distance (2009, p. 7). For irregularly-spaced locations, such as those in this study, Charlton and Fotheringham propose that the bandwidth of the kernel be allowed to vary such that the bandwidth is increased when points are sparser and decreased when points are denser. Specifically, they suggest the use of an adaptive bandwidth be used in

which the same number of neighboring points is used for each estimation (Charlton and Fotheringham, 2009, p. 7).

The typical procedure for GWR analysis is to start with OLS analysis to find a properly-specified model (Rosenshein et al 2011, p. 45). The same model variables are then run through GWR. Akaike's Information Criterion (AIC) is a useful measure to compare the OLS and GWR versions of the model when they have the same dependent variables.

#### **2.4.2.4      *Spatial Regression***

An alternative to GWR is spatial regression, in which the spatial dependence of the variables is included in the model. This can be done in three ways: by addressing spatial correlation in the dependent variable (spatial lag dependence models), by addressing spatial correlation in the error term (spatial error dependence models), or by a combination of the two, known as higher order models (Anselin 2006, p. 5-12). In a spatial lag model, the dependent variable of, say, energy use at a particular location during a particular interval is a function not only of that location's characteristics but also of its neighbors' energy use during that interval. So, for spatial lag models, a spatial lag variable is included as a dependent variable. The spatial lag variable is a function of the dependent variable observed at other locations, typically a function that is very general and non-linear. The spatial lag variable is simplified by applying a spatial weights matrix that eliminates the values of all but the nearest neighbors (p. 6). Anselin compares this spatial lag variable to an autoregressive term in a time-series model (p. 6). If spatial lags exist but are not modeled, Anselin notes that the modeled equation will have omitted variable error,

and will thus result in OLS coefficient estimates that are biased and inconsistent (Anselin 2006, p. 13).

In a spatial error model, spatial autocorrelation affects the covariance structure of the random disturbance term. For example, energy use by a particular location during a particular interval is a function not only of that location's characteristics but also of its neighbors' characteristics. The rationale behind spatial error models is that the unmodeled spatial effects incorporate multiple units of observation, thus resulting in spatially correlated errors. Anselin (2006, p. 9-12) details a variety of possible structures for the error variance-covariance matrix. If spatially correlated errors are ignored, the coefficient estimates will be unbiased but inefficient (Anselin 2006, p. 13).

As to the choice of a spatial lag versus spatial error model, Anselin suggests the use of the Lagrange Multiplier test applied to the residuals from an OLS regression (Anselin 2006, p. 18). Anselin describes his Lagrange Multiplier test statistic as "essentially the square of Moran's I" (Anselin 2006, p. 19). The choice between a spatial lag and a spatial error model is made by choosing the model with the larger value of the Lagrange Multiplier test statistic (Anselin 2006, p. 22).

For spatial lag models, a common approach is to incorporate the correct spatial lag into the OLS equation, usually through the use of instrumental variables. For spatial error models, OLS can still be applied if the estimated errors are adjusted to take the error correlation into account.

#### **2.4.2.5      *Spatial Statistics Literature Summary***

For this thesis, the focus is on predictive accuracy rather than on the more traditional statistical methods of hypothesis testing or confidence intervals. Specifically,



this thesis will select the "best" results based on the bias and accuracy of the predicted values as compared to the actual values (see Section 2.7, below). As mentioned previously, classical OLS regression makes numerous assumptions about the underlying data; in this thesis, however, the validity of these assumptions is less important because the statistical tests associated with OLS regression are not a key factor in the analysis. Still, Lagrange Multiplier tests will be conducted following OLS regressions. If a spatial lag model is indicated, a subsequent model will be tested that incorporates a spatially-lagged exogenous variable. If a spatial error model is indicated, however, no additional modeling is indicated.

From the spatial literature, inverse distance weighted maps and semivariograms will be used for exploratory analysis, as will the Geary' C and Moran's I statistics. Geographically weighted regression and spatial regression will be used to fill energy usage data gaps.

## **2.5 Temporal Lag Literature**

As discussed in the prior sections, the presence of spatial autocorrelation causes the assumptions of the classical multiple regression model to fail, affecting both the standard errors associated with the model parameters and their bias. Similarly, the presence of temporal autocorrelation in time-series data, in which the value of the dependent variable at one point in time is dependent on the value of the dependent variable at another point in time, also results in a failure to meet the assumptions of the classical multiple regression model.

For time series data, a stochastic model makes the assumption that each value in the series has been drawn randomly from a probability distribution (Pindyck and Rubinfeld 1976, p. 431). An underlying stationary process indicates that the characteristics and

properties of the stochastic process are not varying over time. Time series data meets this assumption if there is no growth trend or other mid-series glitch in the data.

In practice, however, few time series data sets are stationary. Non-stationary time series can be made into stationary series by differencing the series one or more times (Pindyck and Rubinfeld 1976, p. 439). In other words, if the data series  $z_t$  is non-stationary because the value of each observation  $z_i$  is dependent on the value of its predecessor  $z_{i-1}$ , then the data series  $z_t$  can be converted into a stationary data series of the differences between  $z_i$  and  $z_{i-1}$ :

$$\Delta z_t = z_t - z_{t-1}, \text{ where } \Delta z_t \text{ is a stationary series}$$

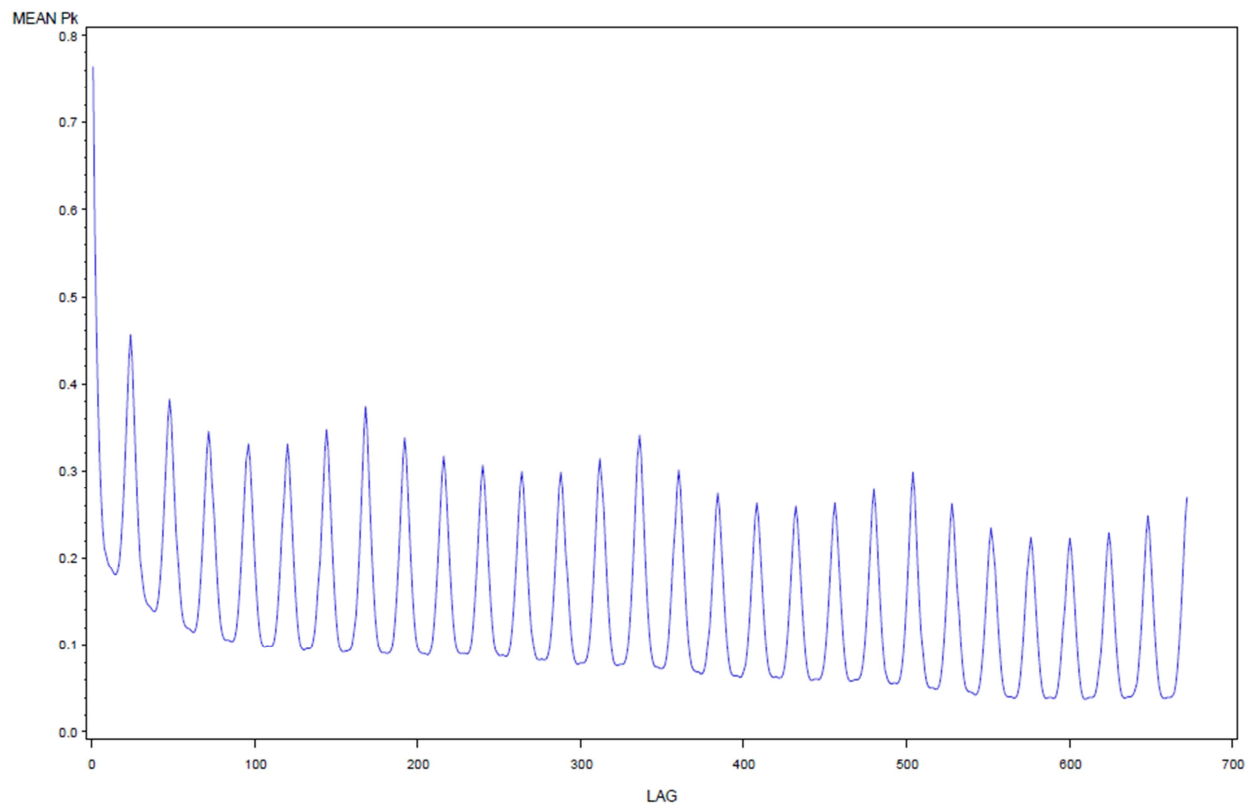
The key to converting a non-stationary series into a stationary series is to identify the appropriate lags and number of lags needed to correctly convert the series. The conversion is done by estimating the sample autocorrelation function (Pindyck and Rubinfeld 1976, p. 437):

$$\hat{p}_h = \frac{(\sum_{t=1}^{T-h} (z_t - \bar{z}) \times (z_{t+h} - \bar{z}))}{\sum_{t=1}^T (z_t - \bar{z})^2}$$

where  $\hat{p}_h$  is the sample autocorrelation function for a data series for a lag of time  $h$ ,  $(z_t - \bar{z})$  is the difference between observation  $z_t$  and the mean value of the series,  $(z_{t+h} - \bar{z})$  is the difference between observation  $z_{t+h}$  and the mean value of the series,  $T$  is the number of observations in the series, and  $h$  is the lag being studied.

It should be noted that the formula for  $\hat{p}_h$ , above, is the same as the formula for a sample regression coefficient,  $b$ , of a line going through the origin, between the dependent value of time series  $z_t$  and its lagged independent version  $z_{t+h}$ .

The result of the sample autocorrelation function calculation is a series of  $\hat{p}_h$  values, one for each lag length being studied in the data series. To determine the appropriate number of lags and lag lengths, one normally views a plot of the sample autocorrelation function, known as a correlogram, and selects the lag values associated with largest values of the sample autocorrelation function (Pindyck and Rubinfeld 1976, p. 441). Figure 2.6 is an example of a correlogram in which significant lags are seen at each 24-hour period, as well as at 168 hour (one week) periods.



**Figure 2.6: Example Correlogram Using Sample Autocorrelation Function**

This thesis will determine the "best" results based on a comparison of actual values with their predictions. The prediction methods traditionally used in Load Research, as discussed in Section 2.3 above, are based on time-series data and incorporate specific, pre-

defined lags. Identifying appropriate lags, however, is an important aspect of any temporal analysis.

## 2.6 Spatiotemporal Literature

Analytic methods for the analysis of space-time datasets are "underexplored" (Goodchild 2009, p. 470). In geographic analyses, Andrienko et al argue that "spatiotemporal data pose serious challenges to analysts" (2010, p. 1), because both space and time are complicated concepts to incorporate in an analysis. For exploratory analysis, a number of spatiotemporal statistics have been developed with the goal of determining if the locations are randomly distributed within both space and time. Graphical techniques take the statistics a step further by incorporating locational attributes and providing additional information to the analyst. Finally, analytic approaches allow both spatial and temporal locations and distances to be incorporated and provide a method for interpolation of missing values. Each of these approaches will be discussed in the following sections.

### 2.6.1 Spatiotemporal Statistics

Knox and Bartlett (1964) developed an index that looks at all possible data pairs and divides them into one of four categories based on the distance between the pairs in terms of space and time, where the definition of "close" is left to the user:

- Both the spatial distance and the time interval are "close."
- The spatial distance is "close" but the time interval is not "close."
- The spatial distance is not "close" but the time interval is "close."
- Neither the spatial distance nor the time interval is "close."

The four categories are illustrated in Figure 2.7, below, where  $O_1, O_2, O_3$ , and  $O_4$  are the number of data pairs falling into each category, and  $S_1, S_2, S_3$ , and  $S_4$  are the row and column totals.

	Close in Time	Not Close in Time	
Close in Distance	$O_1$	$O_2$	$O_1 + O_2 = S_1$
Not Close in Distance	$O_3$	$O_4$	$O_3 + O_4 = S_2$
	$O_1 + O_3 = S_3$	$O_2 + O_4 = S_4$	

**Figure 2.7: Structure of the Knox Index**

The goal of the Knox Index is to determine if there is clustering in space and time, or if the data are randomly distributed. Therefore the observed data distribution must be compared to an expected distribution. The expected number of data pairs in each cell is calculated as the product of the row total and column total for that cell, divided by the total number of observations. So, for example, the expected number of observations in cell 1, called  $E_1$ , is calculated as follows:

$$E_1 = \frac{S_1 \times S_3}{N}$$

The observed number of data pairs in each category is then compared to the expected number of pairs in each category, as follows:

$$\chi^2 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i}$$

Although the Knox Index is measured with a Chi-squared statistic, the observations are not independent and therefore the standard probability test associated with Chi-squared does not hold (Ned Levine 2004, p. 9.5). A primary problem with the Knox Index is that the resultant numeric value can be made to vary considerably simply by adjusting the critical cut-off points that distinguish "close" and not "close" in space distance and time interval (Ned Levine 2004, p. 9.8).

The Mantel Index adjusts the scale of spatial distances and temporal intervals by taking their reciprocals, thus collapsing great distances and spreading out near ones (Mantel 1967, p. 212). Critical spatial distance and time intervals still need to be specified, however, so that the differences from the critical values can be calculated. Using the mean value as a critical value, as done in the CrimeStat model (Ned Levine 2004, p. 9.8), the Mantel Index is calculated as follows:

$$r = \frac{1}{N-1} \times \sum_{i=1}^N \sum_{j=1}^N \frac{D_{ij} - \bar{D}}{S_D} \times \frac{T_{ij} - \bar{T}}{S_T}$$

where  $N$  is the number of observations,  $i$  and  $j$  represent the indices for the two observations in each data pair,  $D_{ij}$  and  $T_{ij}$  are the reciprocals of the individual spatial distances and temporal intervals between each  $ij$  data pair,  $\bar{D}$  and  $\bar{T}$  are the mean values of the reciprocal spatial distances and temporal intervals, and  $S_D$  and  $S_T$  are the standard deviations of the  $D_{ij}$  and  $T_{ij}$  values, respectively.

The Mantel Index is a correlation coefficient, so can be influenced by extreme values in spatial and/or temporal distance. Additionally, the correlation coefficients tend to be smaller than traditional correlation coefficients, thus making them somewhat non-intuitive to interpret (Ned Levine 2004, p. 9.11).

Jacquez developed a pair of  $k$ -nearest neighbor test using a pair of statistics,  $J_k$  and  $\Delta J_k$  (Jacquez 1996):

$$J_k = \sum_{i=1}^N \sum_{j=1}^N n_{ijk}^d n_{ijk}^t$$

where  $N$  is the number of data points,  $k$  is the number of nearest neighbors considered,  $n_{ijk}^d$  takes a value of one if data point  $j$  is as near or nearer to the target data point  $i$  in terms of spatial distance than the  $k$ th nearest neighbor and a value of zero otherwise, and  $n_{ijk}^t$  takes a value of one if data point  $j$  is as near or nearer to the target data point  $i$  in terms of temporal distance than the  $k$ th nearest neighbor and a value of zero otherwise.

Jacquez's second test statistic is calculated as:

$$\Delta J_k = J_k - J_{k-1}$$

where  $J_{k-1}$  is simply  $J_k$  calculated using a smaller value of  $k$ . Although the  $J_k$  are not independent, the  $\Delta J_k$  are independent of each other. The significance of the statistics is assessed via a Monte Carlo simulation (Jacquez 1996, p. 1939).

An advantage of the Jacquez approach is that there is no need to define critical spatial distances or temporal intervals. On the other hand, the analyst still needs to determine an appropriate cut-off for  $k$ , the number of nearest neighbors to consider (Malizia and Mack 2012, p. 7).

Kulldorff (1997) developed a space-time scan statistic that uses cylindrical windows in which the base is spatial and either circular or elliptical in shape, and the height is temporal. The cylindrical window moves through time and space to cover the entire study

region with all possible combinations of time and space, attempting to identify clusters.<sup>4</sup> A likelihood function is then calculated, using any of a number of theoretical models, that compares the number of points found within each cylindrical window to its expected value (Kulldorff 2014, p. 7).

A universal issue with all of the above spatiotemporal statistics is that they only look at the spatial and temporal distribution of the data, but they do not consider the value of any attribute associated with the data locations. An exception is the spatiotemporal semivariance. Earlier, in Section 2.4.2.2, the semivariance statistic was introduced. An extension, referred to as spatiotemporal semivariance, allows for the incorporation of temporal data, as follows:

$$\hat{\gamma}(d, t) = \frac{1}{2 \times n(d, t)} \times \sum_{d_{ij}=d}^n \sum_{t_{ij}=t}^n \{(z_{di} - z_{dj})^2 \times (z_{ti} - z_{tj})^2\}$$

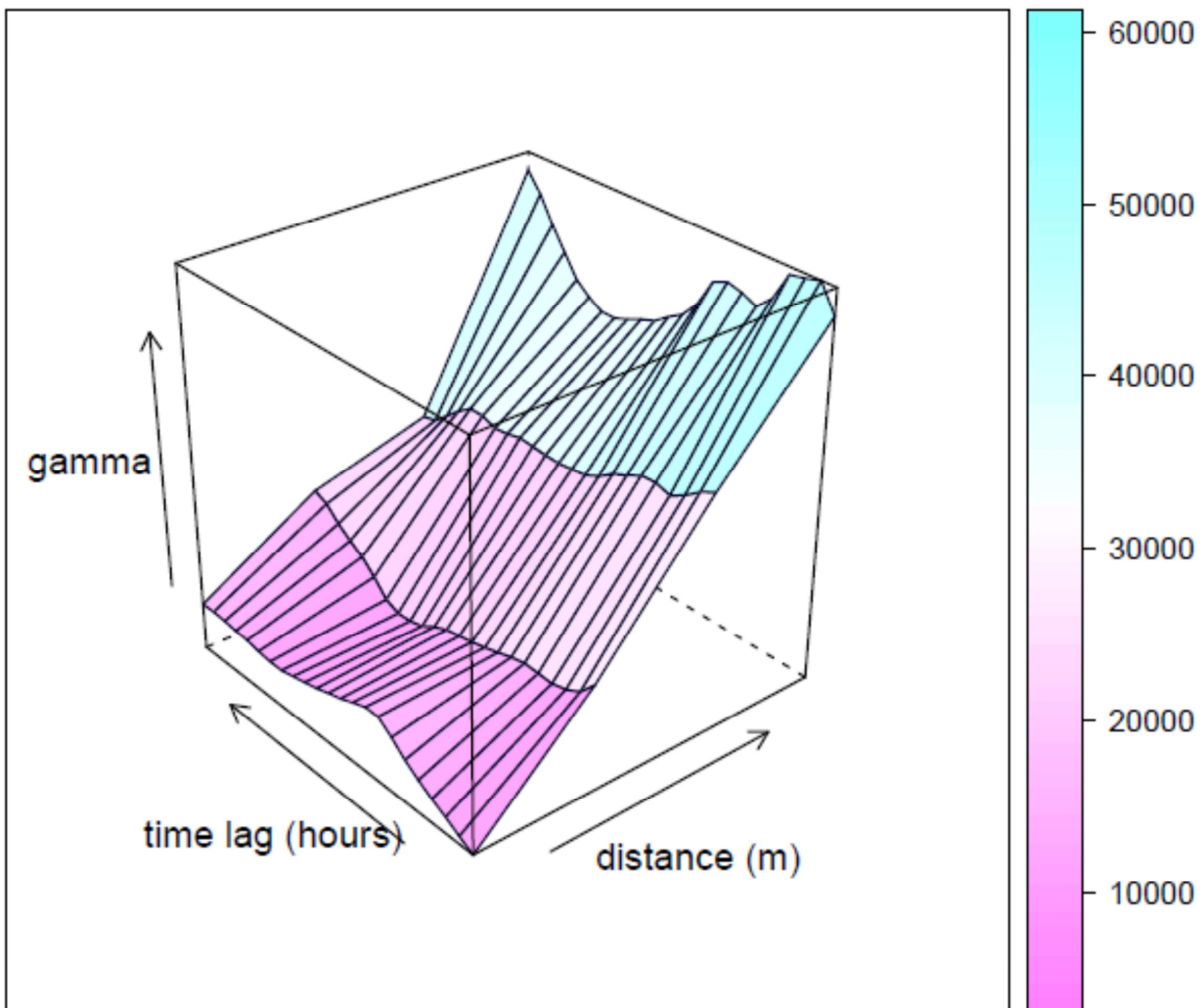
where  $\hat{\gamma}(d, t)$  represents the semivariance for distance  $d$  and time  $t$ , where  $d$  is a distance range and  $t$  is a time range. The term  $\frac{1}{n(d, t)}$  is the inverse of the number of observations of distance  $d$  and time  $t$  from one another. The term  $\sum_{d_{ij}=d}^n \sum_{t_{ij}=t}^n \{(z_{di} - z_{dj})^2 \times (z_{ti} - z_{tj})^2\}$  is the sum of squared differences between the attribute  $z$  at all pairs of points at distance  $d$  and time  $t$  from one another. The spatial and temporal lags are chosen so as to include a sufficient number of points within each lag (Schabenberger and Gotway 2005, p. 441).

As with the semivariance statistic introduced earlier, the spatiotemporal semivariance can be plotted as a three-dimensional spatiotemporal semivariogram, as illustrated in Figure 2.8. The time and distance lags are shown on the two bottom x and y

<sup>4</sup> The SaTScan software package was developed to implement Kulldorff's methodology.



axes, and the value of the semivariance statistic is shown on the vertical z-axis. The color ramp also indicates the value of the semivariance statistic.

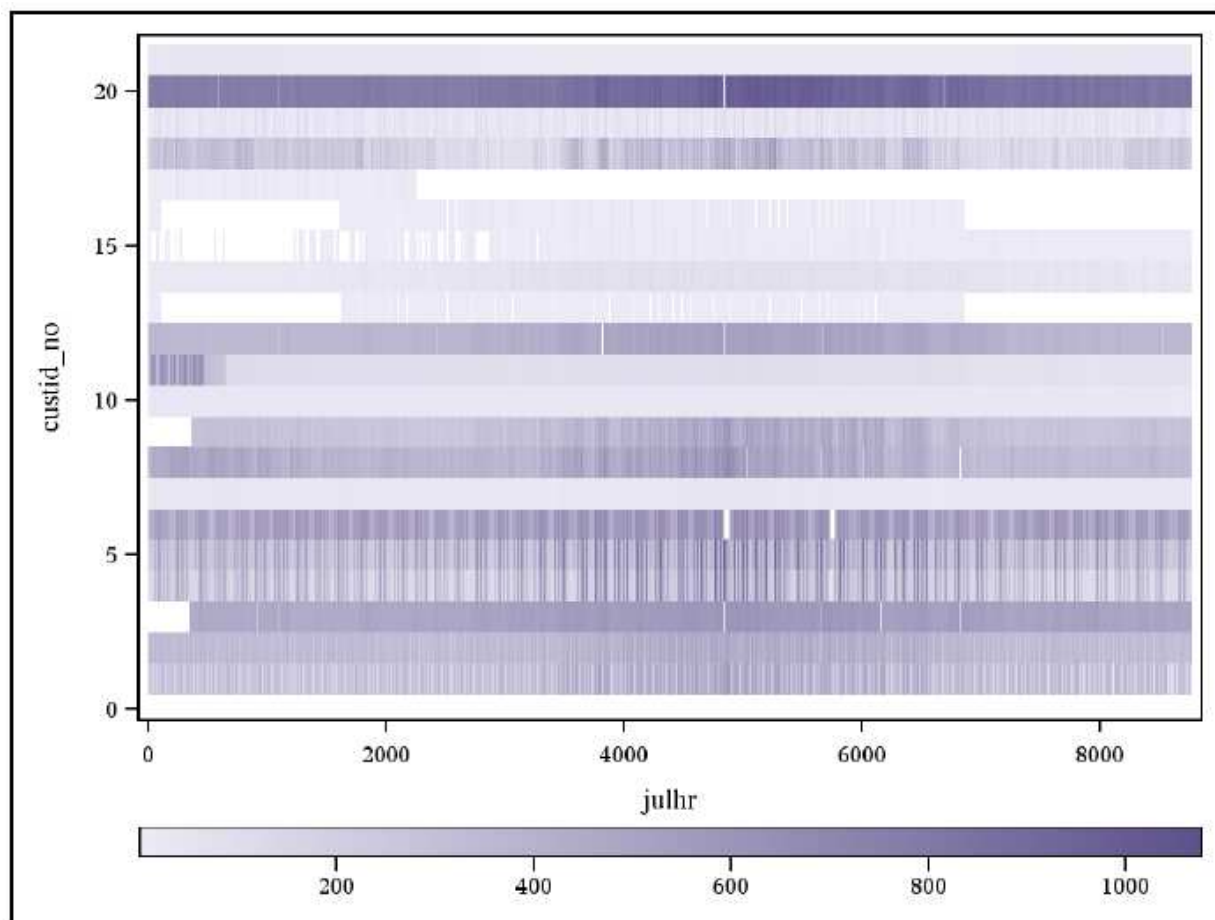


**Figure 2.8: Example of Spatiotemporal Semivariogram**

### 2.6.2 Spatiotemporal Graphics

As suggested by Andrienko et al (2010, p. 2) spatiotemporal data is often viewed in either (or both) of two ways: as a temporally-ordered sequence of spatial situations, or as set of spatially-arranged locations in which each location is characterized by the temporal variation of its attributes. Andrienko et al (2010, p. 2) propose that time-in-space self-

organizing maps (SOMs) can be used for exploratory visual analysis of local temporal variation. Specifically, SOMs can be used to identify the spatial locations of entities with similar temporal patterns, or the temporal patterns of entities that are close to each other in space. Figure 2.9 provides an illustration of a space-in-time map, also known as a heat map (Wicklin 2014). The vertical axis contains a customer and its nearest neighbors sorted by increasing distance as we move from the bottom to the top of the map, and the horizontal axis shows the hours of the year. A color ramp is used to indicate the level of usage during each hour of the year, with the darker colors representing times of higher usage and the white areas representing missing data.



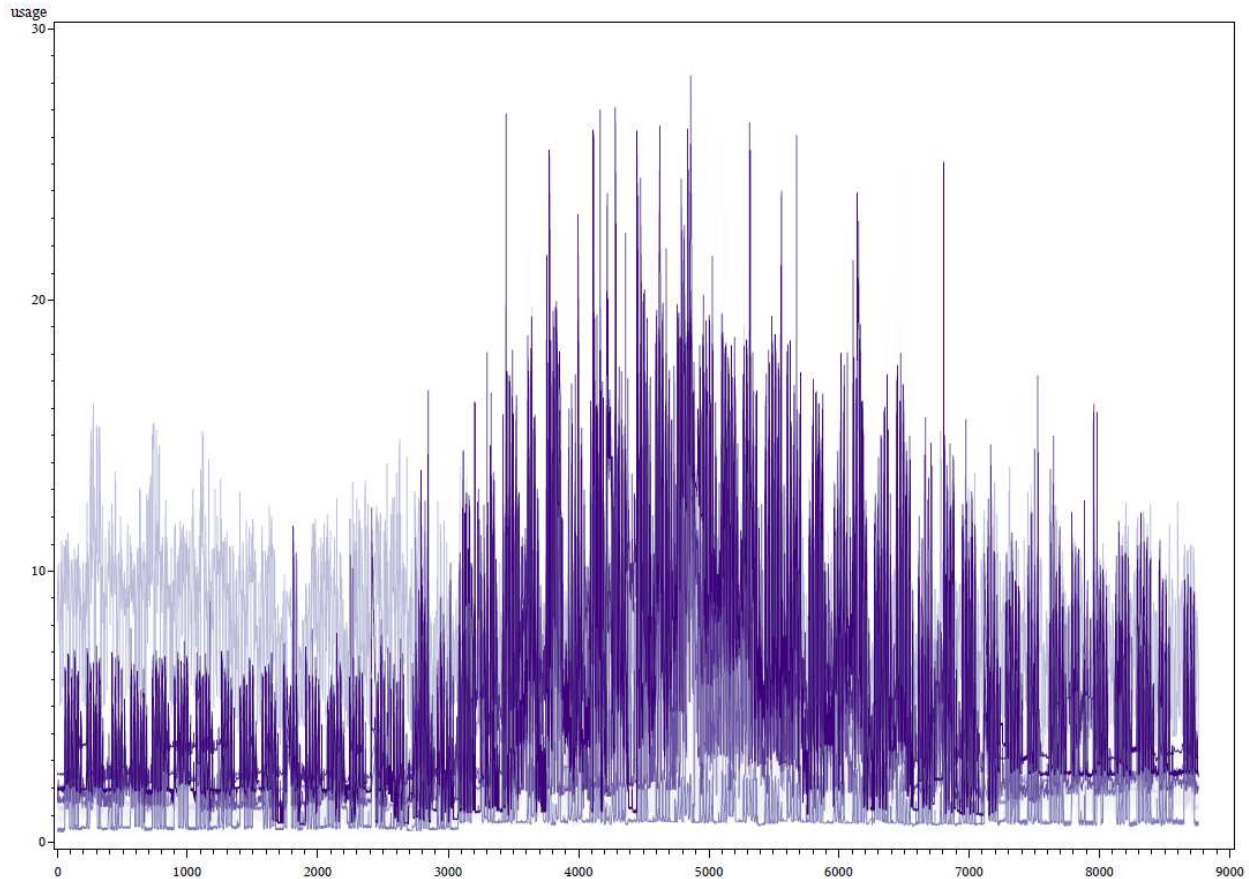
**Figure 2.9: Example of Space-in-Time or Heat Map**

Ideally, the heat map would help indicate a cut-off for the number of neighbors that are similar to the target customer (who is shown as the bottom row of the heat map). For example, in Figure 2.9, the bottom six rows (the target customer plus five neighbors) seem to have roughly the same level of usage, based on the similarity of the color ramp saturation. Additionally, it would be helpful if the heat map could ascertain whether or not the usage patterns are similar over time. In Figure 2.9, similar patterns of dark and light can be seen in the two bottom-most rows, as well as in the fourth, fifth, and sixth rows from the bottom; these are indicated by vertical bars that appear across multiple rows.

Although the heat map can help identify the existence of similarly-sized neighbors and similar usage patterns across neighbors, it does not provide much help in identifying appropriate temporal or distance lags to support further analytics.

Line graphs are another way of illustrating spatiotemporal relationships, and are illustrated in Figure 2.10. The horizontal x-axis shows the hours of the year, and the vertical y-axis shows the usage value during the hour. The darkest line shows the hourly interval energy usage pattern of a selected customer, with each of the other lines representing a neighbor. The lines get progressively lighter as the neighbors are farther away from the selected customer.

Because the lines in front (the nearer neighbors) hide the lines in back (the more distant neighbors), it is difficult to determine just how similar the usage patterns are. There is also no way to identify the distances that separate customers from one another.



**Figure 2.10: Example of a Line Graph**

For all spatiotemporal graphics, a decision must be made as to how to best group or sort the spatial locations. Additionally, although both of these graphics provide reasonably good information about the temporal patterns, neither provides good information about the spatial distances between the customer locations. Another problem is that each of these graphic styles focuses on a single customer and its neighbors, rather than being able to summarize data patterns about customers in general. If these problems could be overcome, graphical presentations can provide valuable visual information about the possible spatiotemporal autocorrelation of the underlying attribute data.

### 2.6.3 Spatiotemporal Modeling

Until recently, spatiotemporal literature in GIS was often focused either on theoretical frameworks through which spatiotemporal relationships can be viewed, or on exploratory data analysis and visualization methods through which the existence of spatiotemporal relationships can be explained and confirmed. More recently, these theoretical methods have been codified into statistical analysis packages that can implement these procedures and be used to interpolate attribute values that are missing at various spatial and/or temporal locations.

Christakos et al (2002, p. 17-18) propose the use of a Bayesian Maximum Entropy (BME) approach to truly integrate space and time. BME, often applied to the modeling of disease spreading, makes a distinction between two distinct knowledge bases that must be combined (Yu et al 2011, p. 488; Christakos et al 2002, p. 33). A core or general knowledge base (G-KB) includes scientific laws, epidemiologic theories, and theoretical or empirical relationships of space and time dependence. The G-KB affects all subjects in the same manner and requires that we have a practical grasp of the structural process through which the data affect the target of interest. A site-specific knowledge base (S-KB) includes both factual and fuzzy data related to the specific location. Christakos et al note that “good [knowledge bases] should provide the right knowledge and not merely a vast range of data” (2002, p. 33).

Ott and Swiaczny propose a data structure in which there are three dimensions to an object: a spatial dimension, a temporal dimension, and attribute values (2001, p. 29). Ott and Swiaczny suggest an Entity Relationship (ER) model that groups similar objects into classes, which can then be parts of other classes in a hierarchy of structured

relationships. Each object can be part of different classes, which allows a single object to be viewed from different perspectives. In a spatiotemporal environment, each attribute must be assigned not only a geometry but also a temporal location.

Ott and Swiaczny review a variety of methods that can be used to visually represent spatiotemporal data, depending upon the problem being studied. They specifically look at objects in which the geographic locations are static but other attributes vary over time (2001, p. 110). As an appropriate visual technique for this situation, they suggest a cartogram, in which each point is represented by a bar chart or graph that represents the changing value of the attribute over time.

Spatiotemporal data are often used to assist in spatial interpolations. For example, Pebesma (2012) provides an example of using spatiotemporal data to aid in spatial interpolation, but does not address temporal interpolation. Holdaway (1996) uses kriging to spatially interpolate monthly temperatures.

In a series of articles looking at air pollution measurements, Adam Szpiro and Johan Lindström developed a method to fill both temporal and spatial gaps in the dataset. Looking at air pollution in Los Angeles, for example, Szpiro et al (2010) were faced with the problem of having a sparse spatial representation of pollution monitoring stations, and a time series of observations that also had missing data. Their goal was to develop a model that could fill in the observations at missing times and locations. They developed a two-stage hierarchical model that attempts to decompose the complexity of space-time relationships into analyzable components. First, a set of mean temporal patterns is developed that, together, explain the overall seasonal and long-term trends. The residuals from this combination of temporal patterns are believed to incorporate the spatial

deviations from the mean temporal trends. A four-step process is used to estimate all the parameters of the model:

1. Using all available observations of air pollution  $Z_{st}$  (where  $s$  defines the spatial location and  $t$  defines the time of the observation), develop a set of  $i$  temporal trends (perhaps a seasonal and an annual trend),  $f_{it}$ , each of which is transformed to have a mean of zero.
2. For each location  $s$ , estimate a time-series regression equation using the temporal trends as explanatory variables:  $Z_{st} = \beta_{0s} + \beta_{1s}f_{1t} + \beta_{2s}f_{2t} + v_{st}$
3. For each location  $s$ , estimate a series of regressions using the newly-estimated parameters from the time-series equation as dependent variables that are explained by a set of locational variables  $L_{js}$ :
  - a.  $\beta_{0s} = a + a_1L_{1s} + a_2L_{2s} + e_{0s}$
  - b.  $\beta_{1s} = b + b_1L_{1s} + b_2L_{2s} + e_{1s}$
  - c.  $\beta_{2s} = c + c_1L_{1s} + c_2L_{2s} + e_{2s}$
4. The residual  $v_{st}$  (from the model in bullet 2, above) is assumed to have no temporal component. Its spatial structure is estimated by taking the model residual  $v_{st}$  from each location  $s$  and, assuming an exponential form, using a maximum likelihood estimator to obtain a semivariogram-style estimate of range, nugget, and sill parameters.

Once the model parameters have been estimated using the above steps, they can be combined to estimate the pollution value  $Z_{st}$  for any desired location  $s$  at any time  $t$ . Two additional articles, Lindström et al (2011) and Sampson et al (2009) expand the model so

that the locational variables, in step 3, above, may also vary temporally. A companion R package, SpatioTemporal, is available to implement this model.

The BME approach suggested by Christakos et al (2002) provides a useful structure for thinking about the data used in this thesis. The temperature data fall into the general category of G-KB, as it clearly affects all customers in the same way. The customer-specific data, whether it be the customer's energy usage or building characteristics, fall into the S-KB category.

The hierarchical model proposed by Szpiro et al (2010), Sampson et al (2009), and Lindström et al (2011) shows the most promise in developing a true spatiotemporal method for the data in this thesis. The sparse locational and temporal aspects of the air pollution data analyzed in these papers are similar to the hourly interval energy usage data being analyzed in this thesis. Additionally, the methodology is directly applicable to this thesis because its goal is to fill gaps in that sparse dataset.

#### **2.6.4 Spatiotemporal Literature Summary**

The focus of this thesis is the interpolation of missing attribute data in spatiotemporal locations. Therefore, methods that only look at point patterns but ignore attribute levels are not particularly relevant. For this reason, the only spatiotemporal statistical method that will be implemented is the spatiotemporal semivariance. The two graphical methods described above, the space-in-time self-organizing maps (also known as heat maps) and line graphs will be explored in an effort to help identify appropriate spatiotemporal relationships and lags. Additionally, as mentioned above, the Szpiro-Lindström hierarchical model will also be implemented.



## 2.7 Literature on Evaluating Success

With regard to the evaluation of gap-filling, there are two topics to be considered. The first is how to select the customers and data gaps to be filled. The second is how to conduct the evaluation once those customers have been selected.

### 2.7.1 Selecting Customers

For the analysis conducted by Mathis et al, customers were selected based on their having “a good mix of timing and length of data gaps” (2007, p. 17). For this reason, the customers selected by Mathis were not a random sample.

Cracknell (2009) notes that it is important to study customers from different service classes (e.g., residential and industrial) because different classes of customers can behave differently, essentially proposing a random sample stratified by service class.

Kirkeide, writing about evaluation of energy efficiency programs, discusses the use of control groups via a two-way stratification by neighborhood and energy usage level. Specifically, Kirkeide suggests the use of a matched pair in which the first customer is the treatment customer and the second is the customer in the same neighborhood who is closest in energy usage to the treatment customer (2010, p. 2).

Williamson also urges the use of matched pairs, proposing that customers be matched on their zip code and on-peak versus off-peak energy usage (2012, p. 14). The match for each targeted customer is found using a two-part method. First, customers are stratified by zip code. Within the zip code, a series of energy-usage differences is calculated between the target customer and each potential match and then weighted, as follows:

- Total summer on-peak consumption, weight =  $1/3$
- Total summer off-peak consumption, weight =  $1/3$

- Total winter on-peak consumption, weight = 1/6
- Total winter off-peak consumption, weight = 1/6

The customer having the lowest difference in weighted energy use from the target customer is selected as the match.

In this thesis, the use of different customer rate classes for analysis is important, because different rate classes do have significant differences in usage characteristics and usage levels. Therefore, Cracknell's (2009) suggestion of a stratified random sample will be used, allowing results to be provided for multiple classes of customers, while perhaps also offering some ability to extrapolate results to non-sampled customers in those classes. Additionally, Williamson's notion of matching customers by energy use will be implemented by making use of existing energy usage-based strata as an additional stratification level. The use of a wide range in gap lengths, in the timing of the gaps, and in usage levels will also be important in this research.

### 2.7.2 Conducting the Evaluation

Hennessey suggests that forecasts should be evaluated using measures for three different components: accuracy, bias, and variability (2011, p. 8). As an accuracy measure, Hennessey proposes the root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^I \sum_{d=1}^D \sum_{h=1}^H (actual\ z_{x,y,d,h} - calculated\ z_{x,y,d,h})^2}{n}}$$

where all variables are as previously defined. RMSE has a range from zero to infinity, with zero resulting from a perfect fit and larger numbers resulting from an imperfect fit. It differs with the magnitude of the values being calculated, so is not comparable across

locations with divergent usage levels. Because large differences are squared, their values are magnified in importance. RMSE is unaffected by the number of intervals that are filled.

To measure bias, Hennessey proposes using the average error:

$$\text{average error} = \frac{\sum_{i=1}^I \sum_{d=1}^D \sum_{h=1}^H (\text{actual } z_{x,y,d,h} - \text{calculated } z_{x,y,d,h})}{n}$$

where all variables are as defined previously. Average error has a range from negative infinity to infinity, with zero resulting from a perfect fit and larger numbers (in either direction) resulting from an imperfect fit. It differs with the magnitude of the values being calculated, so is not comparable across locations with divergent usage levels. With average error, errors in different directions can cancel each other out, which enables it to focus on measuring bias. Average error is unaffected by the number of intervals that are filled.

Hennessey's third measure, the error ratio, indicates how divergent (or variable) the errors are. The error ratio is the ratio between the sum of the residual standard deviations and the sum of the expected values (AEIC 2010):

$$ER = \frac{\sum_{i=1}^I \sum_{d=1}^D \sum_{h=1}^H \sigma_{x,y,d,h}}{\sum_{i=1}^I \sum_{d=1}^D \sum_{h=1}^H \mu_{x,y,d,h}}$$

where  $\sigma_{x,y,d,h}$  is the standard deviation and  $\mu_{x,y,d,h}$  is the mean of  $z_{x,y,d,h}$ . McMenamin

uses both the error ratio and the related measure of the coefficient of variation or CV (2011b, p. 24):

$$CV = \frac{\sigma_{x,y,d,h}}{\mu_{x,y,d,h}}$$

where all variables are as defined previously.

In a presentation on methods of modeling a system peak, Albrechtson uses two additional measures to determine how well the method is working (2009, p. 14). One of these is the mean absolute percentage error (MAPE), which is defined as follows:

$$MAPE = \frac{100}{n} * \sum_{i=1}^I \sum_{d=1}^D \sum_{h=1}^H \left[ \text{absolute value} \left( \frac{\text{actual } z_{x,y,d,h} - \text{calculated } z_{x,y,d,h}}{\text{calculated } z_{x,y,d,h}} \right) \right]$$

where all variables are as defined previously. Rupp (2008, p. 18) also uses MAPE as the primary means of evaluating alternative methods when analyzing interval data. Xu (2009, p. 15) includes MAPE as one of two evaluation methods.

MAPE has a range from zero to infinity, with zero resulting from a perfect fit and larger numbers resulting from an imperfect fit. Because the MAPE is measured in percentage terms, it is comparable across calculated values of different magnitudes. However, if the magnitudes are small, even small absolute differences can be large when measured as a percent. MAPE is unaffected by the number of intervals that are filled.

Albrechtson (2009, p. 14) and Xu (2009, p. 15) also suggest the mean absolute error (MAE), which is the simple average of the absolute values of the errors:

$$MAE = \frac{\sum_{i=1}^I \sum_{d=1}^D \sum_{h=1}^H [\text{absolute value} (\text{actual } z_{x,y,d,h} - \text{calculated } z_{x,y,d,h})]}{n}$$

where all variables are as defined previously. Similar to the MAPE, the MAE has a minimum value of zero and no upper limit. It differs with the magnitude of the values being calculated, so is not comparable across locations with divergent usage levels. MAE is unaffected by the number of intervals that are filled.

Richardson et al (2010) make use of a number of summary comparisons between measured customer usage and synthetic usage as developed by their model. Their measures consist of a series of tabular and graphical comparisons demonstrating the similarities and differences in the two usage datasets. In tabular form, they compare the sums of the maximum non-coincident demands, the maximum time-coincident demands, and the ratio of these two. Graphically, they compare load duration curves.

Because the focus of this research is to match the actual value of the hourly interval energy usage data values as closely as possible, the most appropriate methods for evaluation of success focus on accuracy and bias rather than on variability. Evaluation methods that will allow the comparison of the actual value during a gap to its estimated value are most appropriate.

Based on this criterion, the RMSE and MAPE methods will be used to assess the accuracy of the gap-filling methods, and the average error will be used for bias. The methods suggested by Richardson et al (2010), which compare sums of maximum values during some time period of interest, also suggest that the maximum and minimum values of filled hourly interval energy usage values should be compared to their actual counterparts.

## **2.8 Literature Review Summary**

Four bodies of literature have identified exploratory analytic and gap-filling methods that will be used in this thesis, including those that stem from Load Research, spatial analysis, temporal analysis, and spatiotemporal analysis. Specifically, the following methods will be used in the exploratory data analysis:

- General methods -- Simple correlations between the dependent variables of hourly interval energy usage data values and the potential independent variables.
- Spatial methods -- Statistics and graphics looking for spatial autocorrelation, including Moran's I, Geary's C, semivariograms, and IDW maps.
- Temporal methods -- Sample autocorrelation function to determine appropriate temporal lags.

- Spatiotemporal methods -- Spatiotemporal semivariograms, space-in-time self-organizing (heat) maps, and line graphs.

The following methods will be used to analyze and fill gaps in energy usage data:

- Load Research methods -- KEMA method of using a separate regression for each customer-hour, Smith and Hanna method of using dummy variables for each customer-hour, and McMenamin method of a neural network for each customer.
- Spatial methods -- Geographically-weighted regression and spatial regression.
- Spatiotemporal methods -- Szpiro's two-stage hierarchical model.

Additionally, the following methods will be used to evaluate the various gap-filling methods: root mean square error and mean absolute percentage error to evaluate accuracy, average error to evaluate bias, and comparison of maximum and minimum values of filled hourly interval energy usage data values to their actual counterparts for bias and accuracy.

The next chapter discusses the available data elements and the selection of the sample data for analysis, and provides an overview of the methodology.

### **3. Methods and Data**

#### **3.1 Research Design**

This research will compare several methods of filling gaps in hourly interval energy usage data. When used in a real-life situation, the desired outcome is for each filled interval to match the actual (but unknown) value of the actual energy usage during that interval. In order to replicate this situation as closely as possible, artificial gaps will be created in the hourly interval energy usage data and then filled; in this way, the filled value can be compared with the actual value. Therefore, in this research, the success or failure of each method is based on how close the filled value comes to the actual (and known) value of the hourly interval energy usage value.

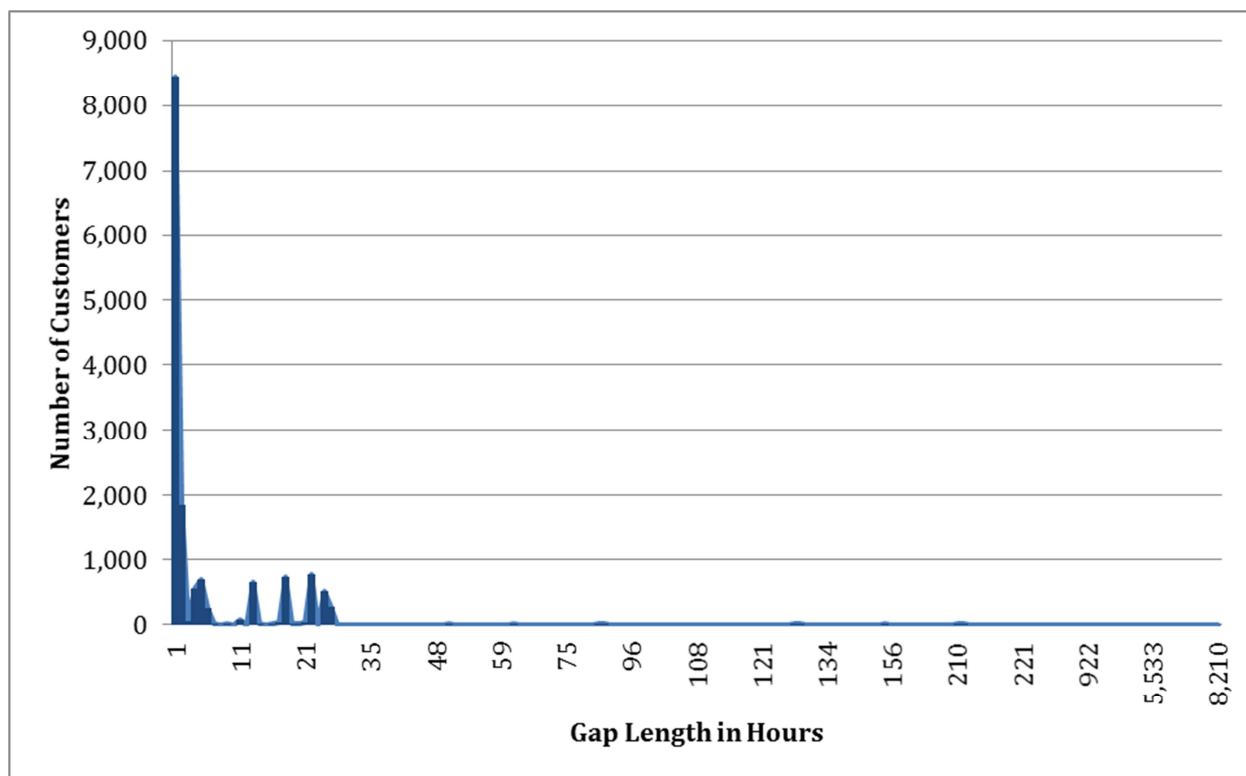
#### **3.2 Data and Variables**

The unit of analysis in this research is an hour-long energy usage interval data gap for a customer of an energy provider. Both residential and business customers are analyzed, so the range of customer size varies from a small residential apartment to a large commercial or industrial business premise. All data used in this study were collected by an electric energy provider in the northeastern United States. One year's worth of hourly

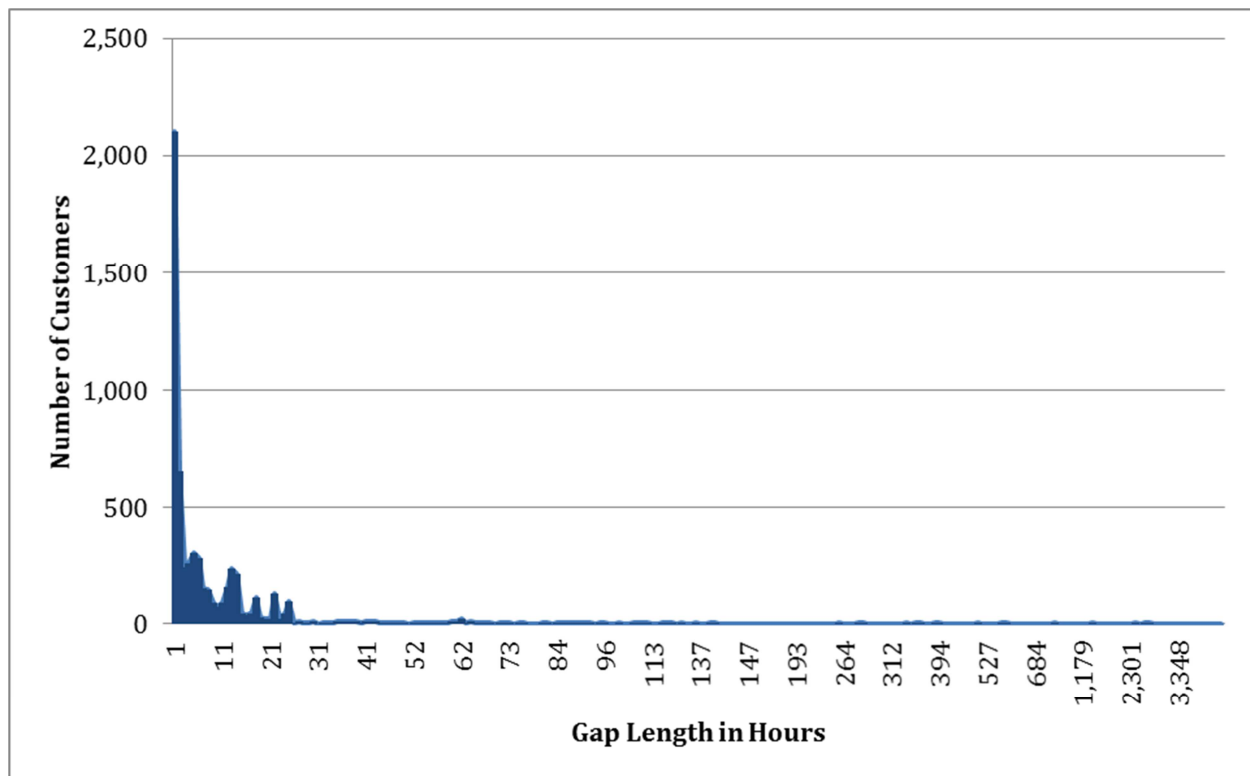
interval energy usage data were collected from 911 of the energy provider's residential customers and 364 business customers in a geographic area that is approximately 128 square miles.

For this research, the energy usage data are analyzed in hourly intervals, so that each energy customer has 8,760 possible intervals in a calendar year (24 intervals per day times 365 days per year). The hourly interval energy usage data were collected with electric metering equipment that meets or exceeds all state-mandated requirements for metering accuracy. Even so, a combination of equipment and communication failures results in a non-trivial set of missing hourly interval energy usage data. For residential customers, gap lengths are anywhere from one hour to 8,210 hours; for business customers, there are gaps of one hour to 8,160 hours in length. For both customer types, short gaps are much more common than longer gaps. Figures 3.1 and 3.2 show the number of interval data gaps for residential and business customers, respectively, by the length of the gap in intervals.



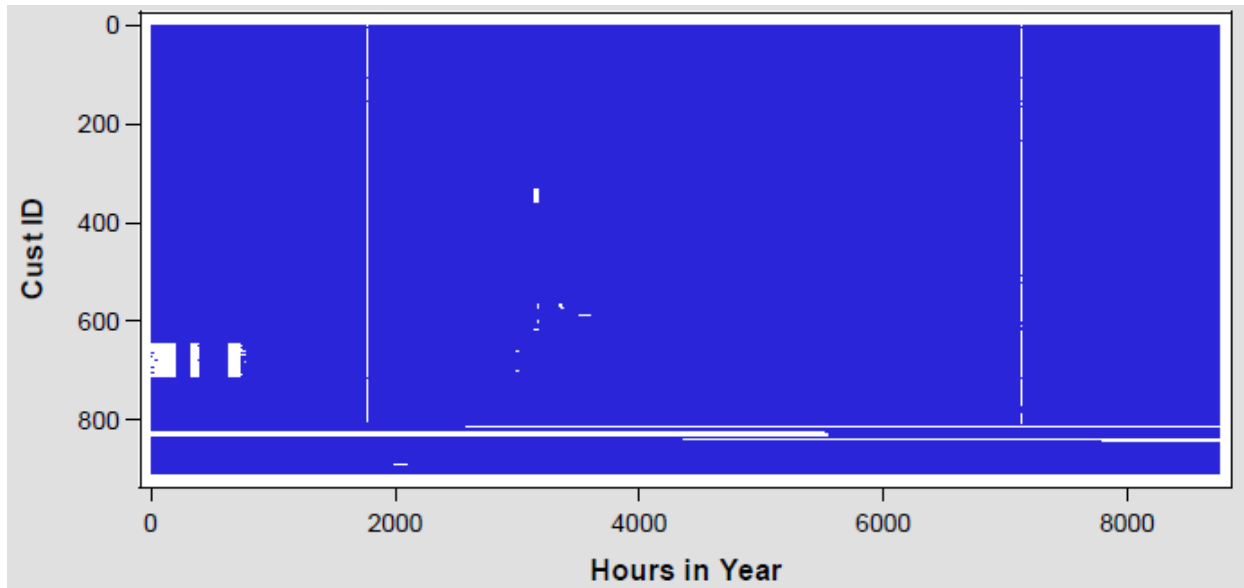


**Figure 3.1: Gap Lengths for Residential Customers**



**Figure 3.2: Gap Lengths for Business Customers**

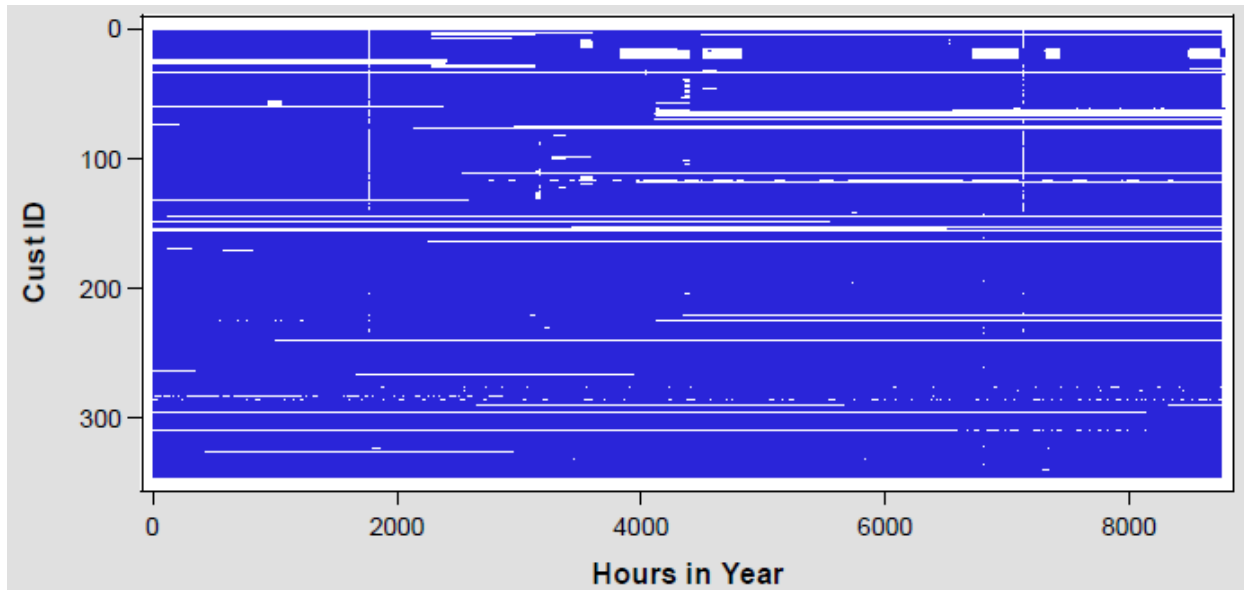
In addition to understanding the distribution of gap lengths, it is also important to understand the distribution of interval data gaps across time for each customer, as well as across customers. Figure 3.3 shows the distribution of hourly interval energy usage data gaps for residential customers. For residential customers, there are a series of data gaps (the white spaces) at hours 1,707 and 7,375, which are the hours during which the clock "springs forward" for daylight savings time and a day in early November (not the day of the daylight savings time shift). Additionally, there is a group of customers (with CustIDs of approximately 625 to 700) who are missing data in three blocks between hours 1 and approximately 1,000. A few customers (with CustIDs around 775) are missing data for six months or more. The remaining noticeable patterns appear between hours 3,000 and 4,000.



**Figure 3.3: Distribution of Interval Data Gaps for Residential Customers**

Figure 3.4 provides the distribution of hourly interval energy usage data gaps for business customers. Compared to residential customers, a larger percentage of business customers have long periods of missing data. About half the customers are missing data in hours 1,707 and 7,375. The missing data for business customers seems to be more randomly distributed in time, but also seems to be generally confined to a few customers who are missing a lot of data.

The energy provider is the source of the basic customer variables used in the analysis, as shown in Table 3.1. Customer latitude and longitude are taken from the energy provider's customer billing system, and are then projected to Universal Transverse Mercator coordinates using the North American Datum of 1983. Customer rate class and usage-based stratum are also taken from the energy provider's customer billing system.



**Figure 3.4: Distribution of Interval Data Gaps for Business Customers**

**Table 3.1. Overview of Basic Variables Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Customer Identification Number	custid	identification variable	incremental number	residential	118,682	145,463	5,175-769,294
				business	468,791	349,507	1,046-769,546
Longitude	x	selection of neighbors	UTM coordinates	residential	n/a	3,064	n/a
				business	n/a	5,472	n/a
Latitude	y	selection of neighbors	UTM coordinates	residential	n/a	958	n/a
				business	n/a	2,577	n/a
Rate Class	ratecd	stratification variable	residential or business	residential	n/a	n/a	n/a
				business	n/a	n/a	n/a
Stratum	stratum	stratification variable	stratum number	residential	2	1	1-6
				business	4	2	1-6
Date	date	defines temporal identity	day-month-year	residential	July 2	105 days	January 1-December 31
				business	June 30	106 days	January 1-December 31
Day-type	daytype	used to categorize observations	day of week (holidays treated as Sundays)	residential	4	2	1-7
				business	4	2	1-7

Hourly interval energy usage data are taken from the energy provider. Because hour-to-hour patterns in energy usage may be more predictable than absolute levels of energy usage, four transformations are made to the raw energy usage interval data and will be studied in this thesis: hourly usage as a percent of the daily maximum interval, hourly usage as a percent of monthly billed usage, hourly usage as a percent of the annual billed usage, and hourly usage as a percent of the prior hour's usage. An overview of these data elements is shown in Table 3.2.

Monthly kWh energy and kW demand billing values are taken from the energy provider's customer billing system. Annual kWh energy use and the maximum demand billed during the year are also available. A summary of these values is shown in Table 3.3.

Weather data, including hourly dry bulb temperatures, cooling degree hours, heating degree hours, cloud cover, humidity, minutes of sunshine, and wind speed are also obtained from the energy provider, as captured from a central location near the customers. Table 3.4 provides an overview of the weather data variables.

Customer addresses are used to identify specific premises in city and county databases that include basic data about the customer's building. A summary of these factors is provided in Table 3.5.

Customer addresses are also used to identify census tract locations. Demographic data about the census tract population is then captured. These data elements are summarized in Table 3.6.

Also captured from census tract data are building-related demographics, as shown in Table 3.7.

**Table 3.1. Overview of Basic Variables Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Customer Identification Number	custid	identification variable	incremental number	residential	118,682	145,463	5,175-769,294
				business	468,791	349,507	1,046-769,546
Longitude	x	selection of neighbors	UTM coordinates	residential	n/a	3,064	n/a
				business	n/a	5,472	n/a
Latitude	y	selection of neighbors	UTM coordinates	residential	n/a	958	n/a
				business	n/a	2,577	n/a
Rate Class	ratecd	stratification variable	residential or business	residential	n/a	n/a	n/a
				business	n/a	n/a	n/a
Stratum	stratum	stratification variable	stratum number	residential	2	1	1-6
				business	4	2	1-6
Date	date	defines temporal identity	day-month-year	residential	July 2	105 days	January 1-December 31
				business	June 30	106 days	January 1-December 31
Day-type	daytype	used to categorize observations	day of week (holidays treated as Sundays)	residential	4	2	1-7
				business	4	2	1-7

**Table 3.2. Overview of Hourly Interval Energy Usage Data Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Hourly Energy Usage, 24 Hours Ending 1AM to Midnight	k1 - k24	dependent variable, explanatory variable	kiloWatthours or kWh	residential	2	14	0-796
				business	165	210	0-1,406
Hourly Energy Usage as Percent of Daily Maximum Hourly Usage, 24 Hours Ending 1AM to Midnight	pctd1 - pctd24	dependent variable, explanatory variable	percent	residential	47	28	0-100
				business	70	30	0-100
Hourly Energy Usage as Percent of Monthly Billed Usage, 24 Hours Ending 1AM to Midnight	pctm1 - pctm24	dependent variable, explanatory variable	percent of monthly billed kWh for residential, percent of monthly billed kW for business	residential	0	0	0-60
				business	48	30	0-785
Hourly Energy Usage as Percent of Annual Billed Usage, 24 Hours Ending 1AM to Midnight	pcta1 - pcta24	dependent variable, explanatory variable	percent of annual billed kWh for residential, percent of maximum annual billed kW for business	residential	0	0	0-30
				business	38	26	0-110
Hourly Energy Usage as Percent of Prior Hour's Usage, 24 Hours Ending 1AM to Midnight	delt1 - delt24	dependent variable, explanatory variable	percent of prior hour's usage	residential	13	108	-100 to 154,000
				business	10	350	-100 to 188,867



**Table 3.3. Overview of Billing Determinants Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Annual Energy Consumption	annkwh	explanatory variable	kiloWatthours or kWh	residential	14,784	102,523	0-1,825,600
				business	1,432,207	1,714,547	0-9,780,000
Maximum Billed Demand During Year	maxdmd	explanatory variable	kiloWatts or kW	residential	321	239	2-803
				business	345	345	4-1,411
Monthly Billed Energy Consumption, 12 Months from January to December	c1 - c12	explanatory variable	kiloWatthours or kWh	residential	1,258	8,829	0-196,000
				business	120,311	146,233	0-955,200
Monthly Billed Demand, 12 Months from January to December	d1 - d12	explanatory variable	kiloWatts or kW	residential	236	193	1-803
				business	273	284	0-1,411

**Table 3.4. Overview of Weather Data Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Dry Bulb Temperature, 24 Hours Ending 1AM to Midnight	tmp1-tmp24	explanatory variable	degrees Fahrenheit	residential	56	17	6-103
				business	56	17	6-103
Cooling Degree Hour, 24 Hours Ending 1AM to Midnight	cdh1 - cdh24	explanatory variable	degrees Fahrenheit, differenced from 65°	residential	3	6	0-38
				business	3	6	0-38
Heating Degree Hour, 24 Hours Ending 1AM to Midnight	hdh1 - hdh24	explanatory variable	degrees Fahrenheit, differenced from 65°	residential	12	13	0-59
				business	13	13	0-59
Cloud Cover, 24 Hours Ending 1AM to Midnight	cc1-cc24	explanatory variable	percent of sky covered with clouds	residential	38	43	0-100
				business	38	43	0-100
Humidity, 24 Hours Ending 1AM to Midnight	hum1-hum24	explanatory variable	percent relative humidity	residential	65	20	15-100
				business	65	20	15-100
Sunshine Minutes, 24 Hours Ending 1AM to Midnight	ssm1-ssm24	explanatory variable	sunshine minutes per hour	residential	19	26	0-60
				business	19	26	0-60
Wind Speed, 24 Hours Ending 1AM to Midnight	wsp1-wsp24	explanatory variable	wind speed, miles per hour	residential	6	4	0-24
				business	6	4	0-24

**Table 3.5. Overview of Building-Specific Data Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Lot Footprint	footprint	explanatory variable	square footage	residential	100,007	182,609	1,450-718,100
				business	187,830	714,780	875-9,150,700
Number of Buildings	no_bldgs	explanatory variable	number of buildings	residential	2	2	1-6
				business	2	3	0-26
Year Built	year_built	explanatory variable	year	residential	1967	38	1890-2008
				business	1948	28	1890-2009
Number of Floors	no_floors	explanatory variable	number of floors	residential	17	15	1-42
				business	5	4	0-32
Building Area	bldg_area	explanatory variable	square footage	residential	274,985	257,948	952-724,475
				business	265,474	598,570	0-4,687,440
Number of Units or Residential Units	no_units, no_res_units	explanatory variable	number of units	residential	275	224	0-580
				business	22	93	0-1,201
Floor-Area Ratio	far	explanatory variable	ratio	residential	4	3	0-10
				business	3	2	0-14
Building Area Per Unit or Per Residential Unit	bapu or bapru	explanatory variable	square footage	residential	1,672	10,411	0-263,000
				business	45,117	136,792	0-1,675,000

**Table 3.6. Overview of Customer Demographics from Census Tract Data Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Census Tract Population	popn	transformation variable	population	residential	4,744	1,133	0-10,413
				business	3,284	1,998	0-12,786
Percent of Census Tract Population That are Less Than 5 Years Old, between 5 and 9 Years Old, between 10 and 14 Years Old, between 15 and 19 Years Old, 9 or Younger, 14 or Younger, 19 or Younger, or Over 65	kidlt5, kid59, kid1014, kid1519, kidle9, kidle14, kidle19, or srcit	explanatory variable	percent of population	residential	7	5	0-56
				business	9	6	0-39
Median Age of Census Tract Population	medage	explanatory variable	years	residential	34	3	2-48
				business	36	5	22-54
Median Household Income in the Census Tract	medinc	explanatory variable	dollars	residential	81,868	31,210	38,750-136,053
				business	65,380	23,319	16,063-123,929
Percent of Census Tract Population That is White, Black, Asian, or Hispanic/Latino	white, black, asian, or latino	explanatory variable	percent of population	residential	28	24	0-97
				business	29	25	0-97
Percent of Census Tract Population Reporting First Ancestry as American, Guyanese, Irish, Italian, Polish, West Indian, or Jamaican, or that is Foreign Born	america, guyana, ireland, italy, poland, windies, jamaica, or foreign	explanatory variable	percent of population	residential	8	14	0-70
				business	8	14	0-100
Household Size in the Census Tract	hhsz	explanatory variable	number of people	residential	2	0	2-5
				business	3	1	2-6
Percent of Census Tract	edlths,	explanatory	percent of	residential	21	21	0-100

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Householders With Less Than a High School Education, Who are High School Graduates, with some college, with a Bachelor's degree or higher, with a high school education or less, or with some college or less	edhs, edsmcoll, edcoll, edlehs, edltcoll	variable	householders	business	33	23	0-93

**Table 3.7. Overview of Building Demographics from Census Tract Data Used in the Analysis**

Variable	Variable Name	Expected Relevance	Measurement Scale	Customer Type	Mean Value	Standard Deviation	Value Range
Number of Housing Units in the Census Tract	houses	transformation variable	number of units	residential	2,443	327	0-4,067
				business	1,520	1,009	0-4,633
Number of Occupied Housing Units in the Census Tract	occhsng	transformation variable	number of units	residential	2,178	334	0-3,961
				business	1,355	914	0-1,355
Percent of Census Tract Occupied Housing Units That Have Electric Heat	elecheat	explanatory variable	percent of units	residential	25	24	0-58
				business	13	16	0-58

Appendix A provides the mean, standard deviation, minimum, and maximum values for each of several groups of data elements. Data are grouped when there are individual variables for different hours of the day or months of the year.

### **3.3 Selection of the Sample Data**

For the purpose of filling gaps in energy usage data, a sample of the previously-described customer base will be selected. Cracknell (2009) proposed that separate random samples of customers from different rate classes should be examined, so customers will be sampled from each of two customer groups: residential and business customers. Williamson (2012) suggested that customers be stratified by usage levels. The energy provider assigns each customer to one of six usage-based statistical strata that minimize within-stratum variance<sup>5</sup>. These stratum assignments allow customer groups to be analyzed separately, so that very small customers are separated from large ones. Customers will be separately sampled from each stratum.

Mathis et al (2007) recommended that a variety of interval data gaps be studied. A total of eleven gap lengths will be studied, varying from a gap of a single hour to a six-month-long gap.

In this thesis, the focus is on comparing the alternative gap-filling methods. No confidence intervals or power or precision estimates will be made surrounding the analytic results. Therefore, it is not imperative to select samples that will provide statistical estimates at any particular level of confidence. Instead, the sample size is arbitrarily set at

---

<sup>5</sup> It should be noted that, by state law, religiously-affiliated businesses, including churches, temples, schools, and colleges, and certain civic organizations are allowed to opt for residential rates for electric service. Although these religious and civic customers could be of any size, the sixth (and largest) residential stratum contains a mix of "McMansions" and large religious customers.

10 customers in each of two customer groups for each of six strata for each of 11 gap lengths for each of six gap-filling methods. This will provide an overall sample size of 1,320 (10 customers  $\times$  6 strata  $\times$  2 customer groups  $\times$  11 gap lengths) for each of the six gap-filling methods. Separate random samples will be drawn for each gap-filling method and gap length.

### **3.4 Overview of Data Analysis Procedures**

Although few if any customers have complete hourly interval energy usage data for all 8,760 hours of the year, the gap-filling techniques will be evaluated using only the hourly interval energy usage data gaps that are created especially for this thesis, because these are the hourly interval energy usage data values for which the actual values are known with certainty. In order to best compare the techniques used to fill each gap, each filled value will be compared to the actual value for that interval. Thus, each filled gap in hourly interval energy usage data provides its own matched pair for the analysis. Each combination of customer and gap length is analyzed separately, so that all available data can be used to aid in the evaluation of any particular hourly interval energy usage data gap for any particular customer.

Results will be evaluated using measurements related to difference. The literature review suggested that a total of five evaluation measures would be appropriate: RMSE and MAPE as measures of accuracy, average error as a measure of bias, the difference between the maximum actual value and the maximum filled value as a measure of accuracy at the top of the range, and the difference between the minimum actual value and the minimum filled value as a measure of accuracy at the bottom of the range.



Each of these measures will be computed for each hourly interval energy usage data gap length and for each method. The findings for each method will be presented in a table for each gap length. The results will be discussed and the best method(s) identified for each gap length. A separate table will summarize the performance of each method for each gap length, and the overall results will be discussed.

### **3.5 Expected Findings**

This research is not hypothesis-based, and as such there is no a priori assumption regarding which of the proposed methods will be evaluated as the best. The hope in conducting the research is that a single method can be found that provides accurate and unbiased filling of hourly interval energy usage data gaps. It is also possible, however, that one or more methods work well, perhaps for different gap lengths. Another less promising scenario is that no method provides good results. Whatever the results of the gap-filling, this research will have established a procedure by which alternative methods can be tested and evaluated – that in itself is a worthy goal.

## 4. Exploratory Data Analysis

This section explores the residential and business customer data in a variety of realms: correlational, spatial, temporal, and spatiotemporal. Data are viewed both overall (all residential or business customers as a single data set) and within each of six residential or business customer strata that have been pre-defined by the energy provider; for residential customers the strata are based on billed annual kWh, and for business customers the strata are based on the maximum summer billed demand kW. The stratum boundaries are statistically developed to minimize variance within a stratum while maximizing the variance between strata.

In the text below, summary results of the exploratory analyses are provided. Further details reside in the referenced appendices. The specific analyses discussed below are as follows:

- Correlation coefficients between the dependent and independent variables.
- Spatial statistics -- Geary's I, Moran's C, semivariograms, and IDW maps.
- Temporal statistics -- Sample autocorrelation function to determine appropriate temporal lags.

- Spatiotemporal statistics -- Spatiotemporal semivariogram, space-in-time self-organizing (heat) maps, and line graphs.

## 4.1 Correlation Coefficients

Correlation coefficients between dependent and explanatory variables often provide a useful guide as to which explanatory variables are most relevant to an analysis. In this thesis, the hourly interval energy usage value during any of the 24 hours of the day is the dependent variable (i.e., 24 dependent variables). Additionally, four transformations of the dependent variable are considered. The number of dependent variables is even larger if the same hour on different days is separately considered (i.e., 8,760 dependent variables). The number of correlations can thus grow exponentially depending on how one views the data and the problem of filling gaps in that data.

Appendix B provides correlation coefficient ranges for groups of data. Data are grouped when there are variables for different hours of the day or months of the year. Appendix B includes both overall and stratum-level results. Additionally, the raw hourly interval energy usage data and each of its transformations are included<sup>6</sup>. The tables provided in the text are limited to those independent variables with relatively high correlation coefficients of 0.67 or greater.

### 4.1.1 Correlation Coefficients for Residential Customers

Table 4.1 shows, for residential customers, the independent variables or groups of variables for which that variable or any member of its group has a correlation of 0.67 or

---

<sup>6</sup> Companion data files to this thesis provide a significant amount of additional details. The file `graves_correlations_STRno.csv` contains all overall correlations for both residential and business customers and for all transformations of the dependent variable. The file `graves_correlations_STRyes.csv` contains all correlations by stratum for both residential and business customers, and for all transformations of the dependent variable.

greater (in absolute value) with the raw hourly interval energy usage data. The results are provided overall and by stratum. If a cell is blank, that means that the correlation coefficients were all below the 0.67 cut-off for that variable. Considering raw hourly interval energy usage data values as the dependent variable, then, several measures of billing determinants are correlated, including hourly interval energy usage data, annual kWh consumption, maximum billed demand during the year, monthly energy consumption, and monthly billed demand. The sole other independent variable that is correlated with the hourly interval energy usage data values is the building area per unit.

**Table 4.1: Independent Variables Correlated with Raw Energy Usage Intervals (k1-k24) for Residential Customers**

Variable	Range	Overall: k1-k24	Stratum 1: k1-k24	Stratum 2: k1-k24	Stratum 3: k1-k24	Stratum 4: k1-k24	Stratum 5: k1-k24	Stratum 6: k1-k24
k1-k24	low	0.82	0.27	0.24	0.23	0.31	0.33	0.56
	high	1.00	0.91	0.92	0.93	0.93	0.96	0.99
annkwh	low	0.88						0.67
	high	0.95						0.83
maxdmd	low	0.66						0.49
	high	0.77						0.67
c1-c12	low	0.85						0.60
	high	0.95						0.83
d1-d12	low	0.57						0.40
	high	0.84						0.78
bapu	low	0.77						
	high	0.85						

When the dependent variable is transformed, the only correlated variables are the transformed variables themselves. Table 4.2 shows the correlation coefficients for hourly interval energy usage data as a percent of the daily maximum value. Table 4.3 provides the correlation coefficients for hourly interval energy usage data as a percent of billed monthly energy use in kWh. Table 4.4 provides the correlation coefficients for hourly interval energy usage data as a percent of billed annual energy use in kWh. For the fourth

transformation, hourly interval energy usage data as a percent of the prior hour's interval energy usage data value, no variables show a correlation coefficient higher than 0.67.

**Table 4.2: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of the Daily Maximum Value (pctd1-pctd24) for Residential Customers**

Variable	Range	Overall: pctd1- pctd24	Stratum 1: pctd1- pctd24	Stratum 2: pctd1- pctd24	Stratum 3: pctd1- pctd24	Stratum 4: pctd1- pctd24	Stratum 5: pctd1- pctd24	Stratum 6: pctd1- pctd24
pctd1- pctd24	low	0.11	0.11	0.08	0.09	0.11	0.11	-0.19
	high	0.88	0.86	0.89	0.91	0.92	0.94	0.99

**Table 4.3: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Monthly Energy Use in kWh (pctm1-pctm24) for Residential Customers**

Variable	Range	Overall: pctm1- pctm24	Stratum 1: pctm1- pctm24	Stratum 2: pctm1- pctm24	Stratum 3: pctm1- pctm24	Stratum 4: pctm1- pctm24	Stratum 5: pctm1- pctm24	Stratum 6: pctm1- pctm24
pctm1- pctm24	low	0.80	0.24	0.38	0.17	0.20	0.42	0.01
	high	0.94	0.92	0.97	0.91	0.90	0.94	0.97

**Table 4.4: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Annual Energy Use in kWh (pcta1-pcta24) for Residential Customers**

Variable	Range	Overall: pcta1- pcta24	Stratum 1: pcta1- pcta24	Stratum 2: pcta1- pcta24	Stratum 3: pcta1- pcta24	Stratum 4: pcta1- pcta24	Stratum 5: pcta1- pcta24	Stratum 6: pcta1- pcta24
pcta1- pcta24	low	0.04	0.03	0.23	0.22	0.29	0.24	0.06
	high	0.92	0.94	0.92	0.93	0.92	0.95	0.97

For residential customers, then, neither the use of strata nor the data transformations provide any incremental value to the correlation coefficients.

#### 4.1.2 Correlation Coefficients for Business Customers

For business customers, Table 4.5 shows the independent variables that are correlated with the raw hourly interval energy usage data. As with residential customers,

billing determinant values are correlated with the dependent variable, including hourly interval energy usage data, annual kWh consumption, maximum billed demand during the year, monthly energy consumption, and monthly billed demand. The sole other independent variable that is correlated with the hourly interval energy usage data is the building area per unit, just as it was correlated for residential customers.

**Table 4.5: Independent Variables Correlated with Raw Energy Usage Intervals (k1-k24) for Business Customers**

Variable	Range	Overall: k1-k24	Stratum 1: k1-k24	Stratum 2: k1-k24	Stratum 3: k1-k24	Stratum 4: k1-k24	Stratum 5: k1-k24	Stratum 6: k1-k24
k1-k24	low	0.88	0.90	0.35	0.69	0.64	0.74	0.72
	high	1.00	1.00	0.98	1.00	1.00	1.00	1.00
annkwh	low	0.93		0.46	0.67	0.66	0.72	0.79
	high	0.95		0.82	0.92	0.88	0.92	0.89
maxdmd	low	0.78						
	high	0.89						
c1-c12	low	0.88		0.30	0.62	0.59	0.61	0.67
	high	0.95		0.81	0.91	0.91	0.91	0.88
d1-d12	low	0.78	0.29		0.13	0.35		0.43
	high	0.91	0.68		0.74	0.68		0.72
bapu	low			0.28				
	high			0.70				

When the data transformation is hourly interval energy usage as a percent of the daily maximum value, the only correlated variables are the transformed variables themselves, as shown in Table 4.6. For hourly interval energy usage as a percent of billed monthly energy use in kWh, and for hourly interval energy usage as a percent of billed annual energy use in kWh, some of the strata show correlations for annual kWh consumption and monthly billed kWh energy. These results are shown in Tables 4.7 and 4.8, respectively. As with residential customers, there are no correlations higher than 0.67 when the fourth transformation, hourly interval energy usage as a percent of the prior hour's interval value, is used.

**Table 4.6: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of the Daily Maximum Value (pctd1-pctd24) for Business Customers**

Variable	Range	Overall: pctd1- pctd24	Stratum 1: pctd1- pctd24	Stratum 2: pctd1- pctd24	Stratum 3: pctd1- pctd24	Stratum 4: pctd1- pctd24	Stratum 5: pctd1- pctd24	Stratum 6: pctd1- pctd24
pctd1- pctd24	low	0.17	0.12	-0.12	-0.09	-0.04	0.02	0.10
	high	0.98	0.96	0.95	0.99	0.99	0.99	0.99

**Table 4.7: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Monthly Energy Use in kWh (pctm1-pctm24) for Business Customers**

Variable	Range	Overall: pctm1- pctm24	Stratum 1: pctm1- pctm24	Stratum 2: pctm1- pctm24	Stratum 3: pctm1- pctm24	Stratum 4: pctm1- pctm24	Stratum 5: pctm1- pctm24	Stratum 6: pctm1- pctm24
pctm1- pctm24	low	0.60	0.54	0.23	0.45	0.34	0.60	0.53
	high	0.99	0.99	0.97	0.99	0.99	1.00	0.99
annkwh	low		0.46	0.34	0.40		0.48	
	high		0.78	0.73	0.83		0.83	
c1-c12	low		0.38	0.18	0.35		0.36	
	high		0.79	0.70	0.83		0.83	

**Table 4.8: Independent Variables Correlated with Hourly Interval Energy Usage as a Percent of Billed Annual Energy Use in kWh (pcta1-pcta24) for Business Customers**

Variable	Range	Overall: pcta1- pcta24	Stratum 1: pcta1- pcta24	Stratum 2: pcta1- pcta24	Stratum 3: pcta1- pcta24	Stratum 4: pcta1- pcta24	Stratum 5: pcta1- pcta24	Stratum 6: pcta1- pcta24
pcta1- pcta24	low	0.66	0.55	0.28	0.56	0.35	0.67	0.63
	high	0.99	0.98	0.98	0.99	0.99	1.00	0.99
annkwh	low		0.52	0.33	0.44	0.35	0.52	
	high		0.84	0.72	0.80	0.70	0.83	
c1-c12	low		0.37	0.18	0.39	0.27	0.38	0.44
	high		0.84	0.68	0.82	0.69	0.83	0.67

For business customers, stratification does provide some increase in the number of correlated variables in two of the data transformations.

### 4.1.3 Correlation Summary

Only a relatively few independent variables meet or exceed the threshold value of having a correlation coefficient of 0.67. It should be noted, however, that these correlation coefficients make no distinction between temporal qualities and relationships such as different days of the week or months of the year. Additionally, the correlations only consider the relationship of any customer's data with itself; without any consideration of spatial relationships.

The consistent correlations of hourly interval energy usage values (both in raw form and transformed) with other values of hourly interval energy usage indicate that there is a temporal aspect or autocorrelation to the dependent variable. This aspect will be explored in a later Section 4.3, below, on temporal exploratory data analysis. The next section discusses spatial exploratory data analysis results.

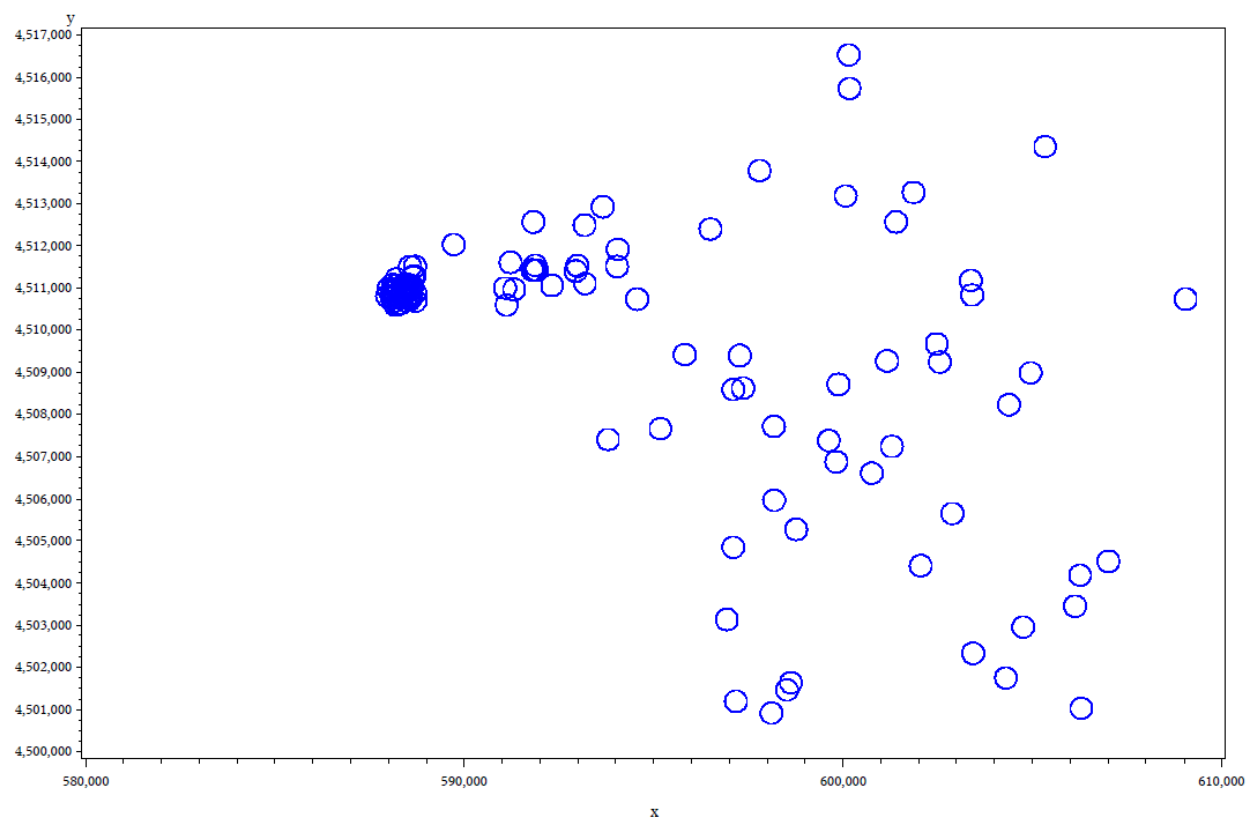
## 4.2 Spatial Exploratory Data Analysis

Figure 4.1 shows the spatial locations of residential customers<sup>7</sup>. The residential customers have a dense spatial cluster in the western-most portion of the data, with spatial dispersion elsewhere. Figure 4.2 illustrates the spatial locations of business customers, which show a similar spatial point pattern although the business customers seem to be a little more evenly dispersed.

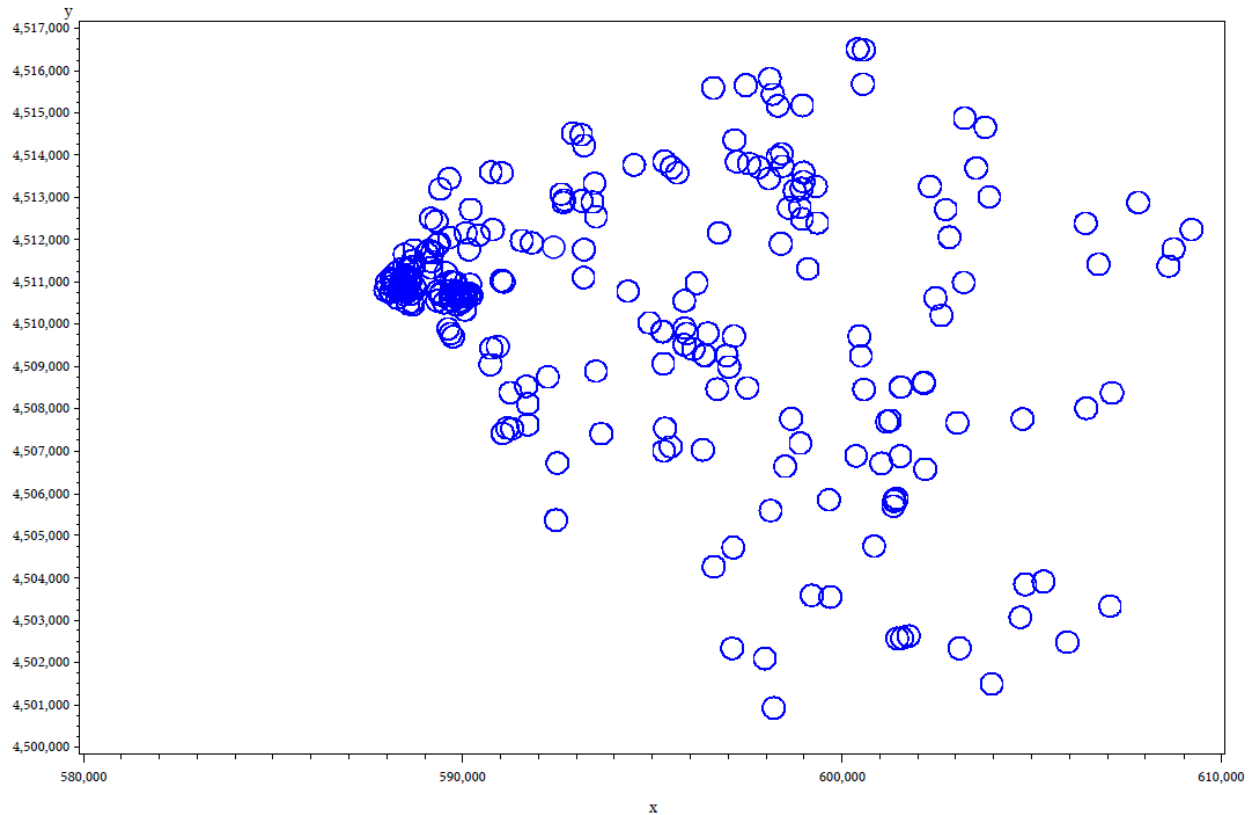
---

<sup>7</sup> Note that in all maps provided in this thesis, the longitude (x) and latitude (y) values have been altered in two ways. First, small random numbers have been added to individual longitude and latitude values to allow distinction between customers with overlapping longitude and latitude. Second, a large constant integer has been added to longitude and latitude values to alter their apparent global location to protect customer privacy.





**Figure 4.1: Residential Customer Locations**



**Figure 4.2: Business Customer Locations**

Spatial exploratory data analysis examines spatial differences with time held constant. Because there are 8,760 unique hours of data, only a sample of those hours are examined in this section of the thesis.

#### **4.2.1 Inverse Distance Weighted Maps**

Inverse distance weighted (IDW) mapping is a visual method that can provide information about the spatial distribution of hourly interval energy usage values. Although IDW maps do not have a statistical basis, the visual appearance of clusters of high and or low values is an indicator of spatial autocorrelation.

IDW maps were created for raw energy hourly interval energy usage data and for each of the four data transformations. Additionally, IDW maps were created overall and

separately for each stratum. The IDW maps were prepared for ten key hours, where the hours chosen for the exploratory analysis are the hours representing the five summer and winter peak hours of the energy provider's system load. These hours are of critical importance to the energy provider because delivery systems must be designed to meet peak loads. If the analytic measures don't show promise for peak times, then they are less attractive methods as a result. In the sections below, a sample of maps is provided<sup>8</sup>. All IDW maps were created using a 40 by 40 grid. A nearest neighbor selection with 6 neighbors was used because customers are unevenly distributed over the spatial range.

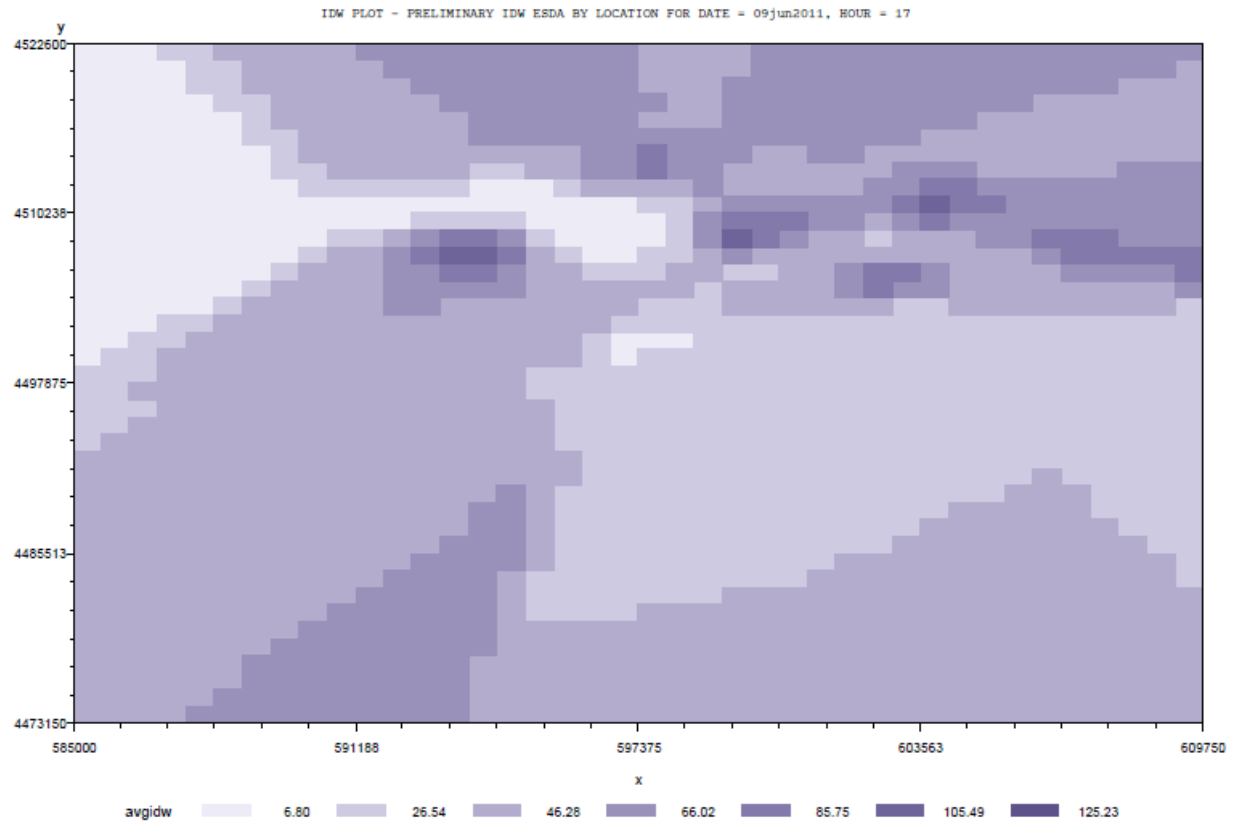
#### **4.2.1.1      *IDW Maps for Residential Customers***

The residential IDW maps indicate some spatial variability and the presence of spatial autocorrelation. For example, Figure 4.3 illustrates "hot spots" of high usage during the hour of interest at several locations across the territory.

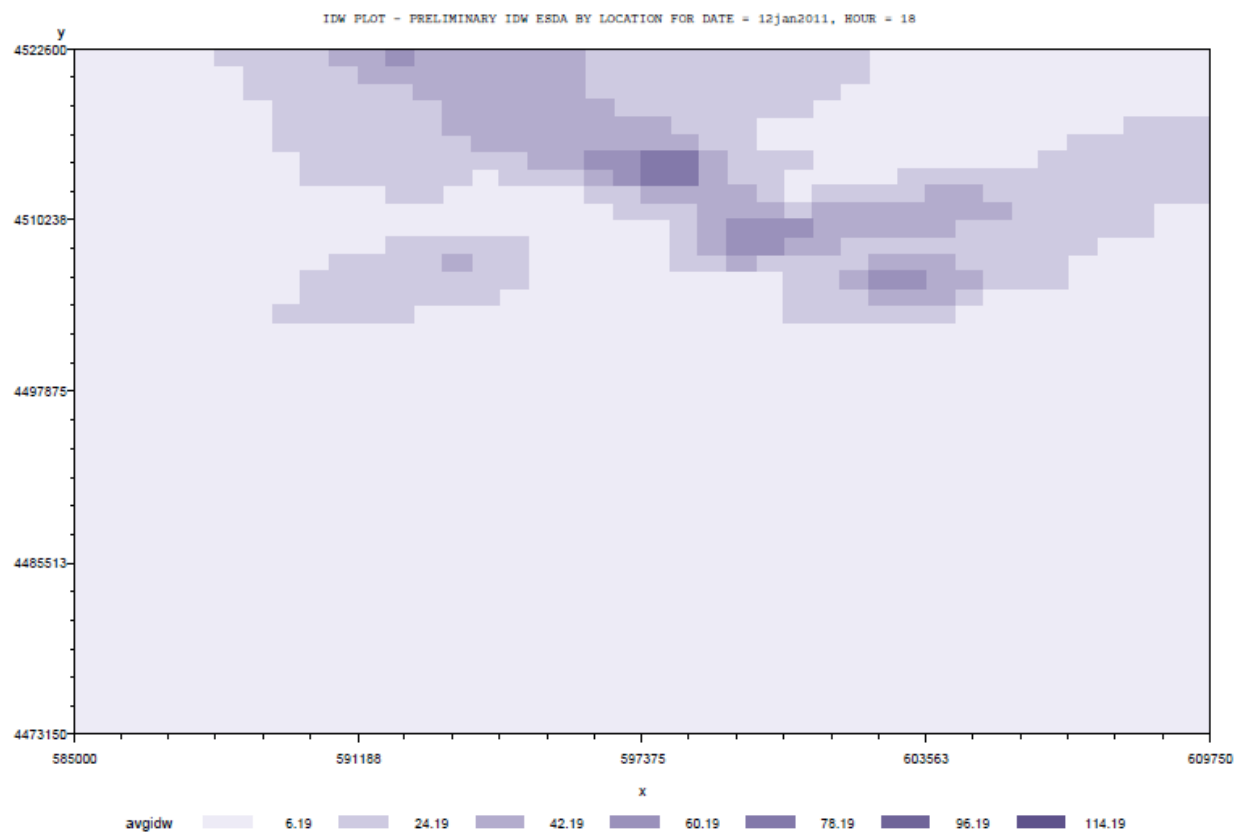
There is also some evidence that the same hours, on different days, have similar patterns of usage, indicating spatiotemporal autocorrelation. Figures 4.4 and 4.5, for example, provide IDW maps for two consecutive days at the same hour. The similarity in the spatial usage pattern can be readily seen, as compared to the IDW map shown in Figure 4.3, which is for a different date and time. These maps provide initial evidence not only of spatial autocorrelation, but of spatiotemporal patterns in the hourly interval energy usage data.

---

<sup>8</sup> A companion data file to this thesis provides a complete set of IDW maps. The file `graves_thesis_graphics.pdf` contains IDW maps for both residential and business customers, for all transformations of the dependent variable, both overall and by strata.



**Figure 4.3: Residential IDW Map for June 9 at Hour 17**



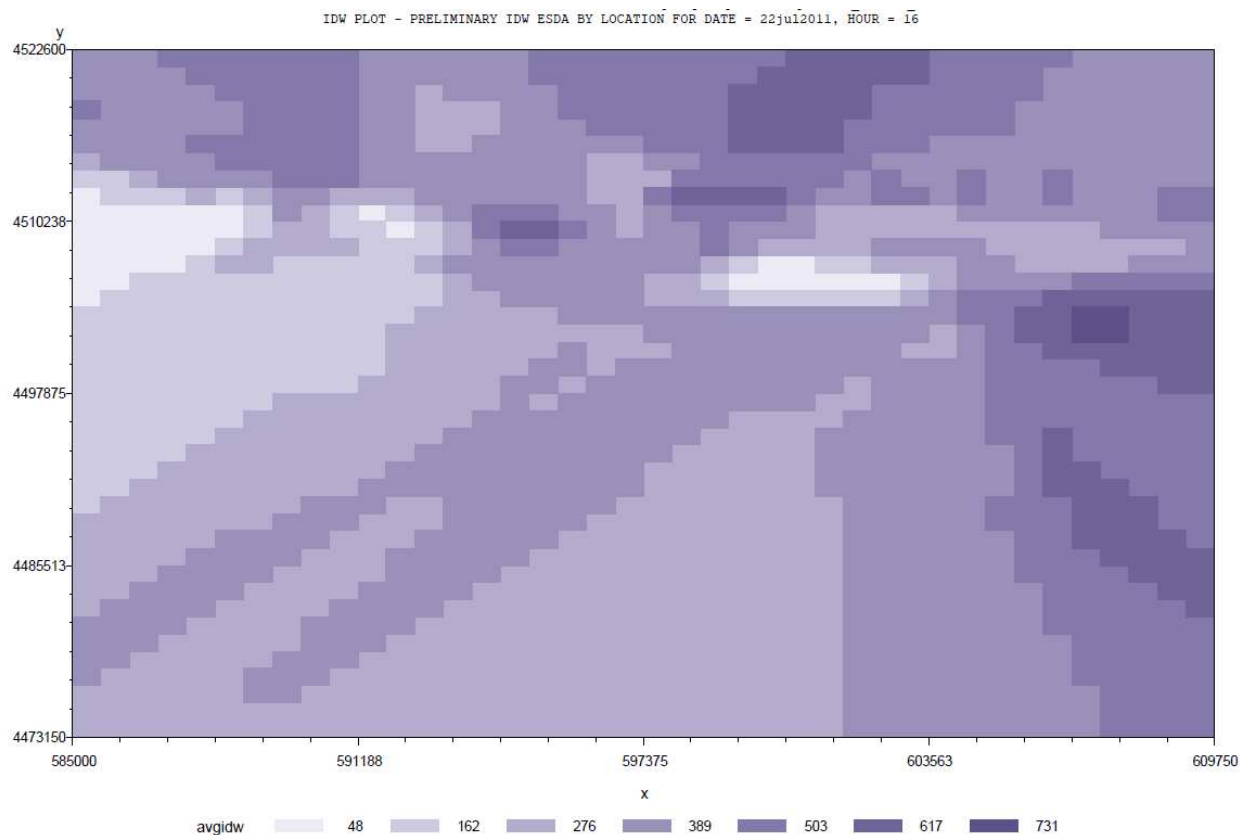
**Figure 4.4: Residential IDW Map for January 12 at Hour 16**



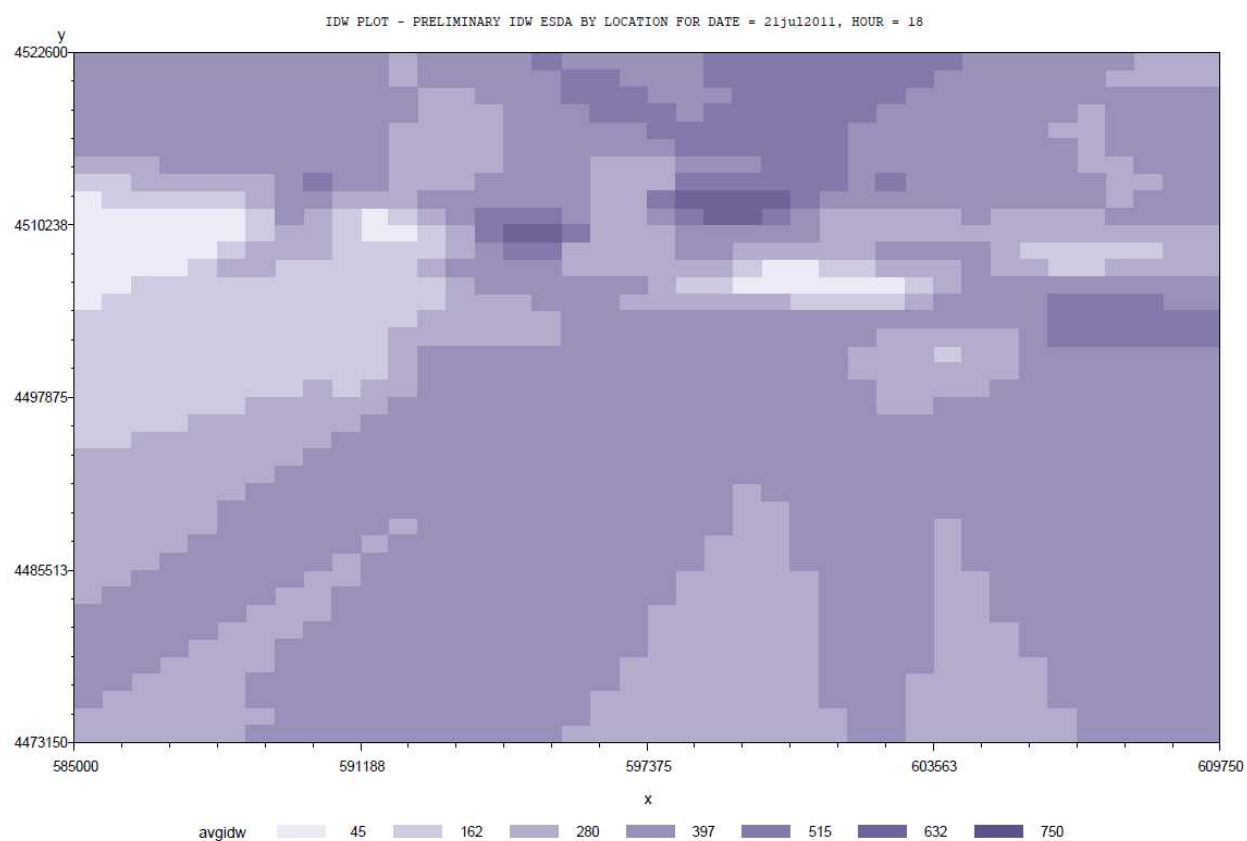
**Figure 4.5: Residential IDW Map for January 13 at Hour 16**

#### **4.2.1.2 IDW Maps for Business Customers**

Figures 4.6, 4.7, and 4.8 show IDW maps for three different day-hour combinations for business customers. Although the patterns are not identical, there is definite similarity even though the hours shown are different for each map. As for residential customers, the IDW maps for business customers indicate the presence of spatial autocorrelation in hourly interval energy usage data as well as indicating spatiotemporal patterns.

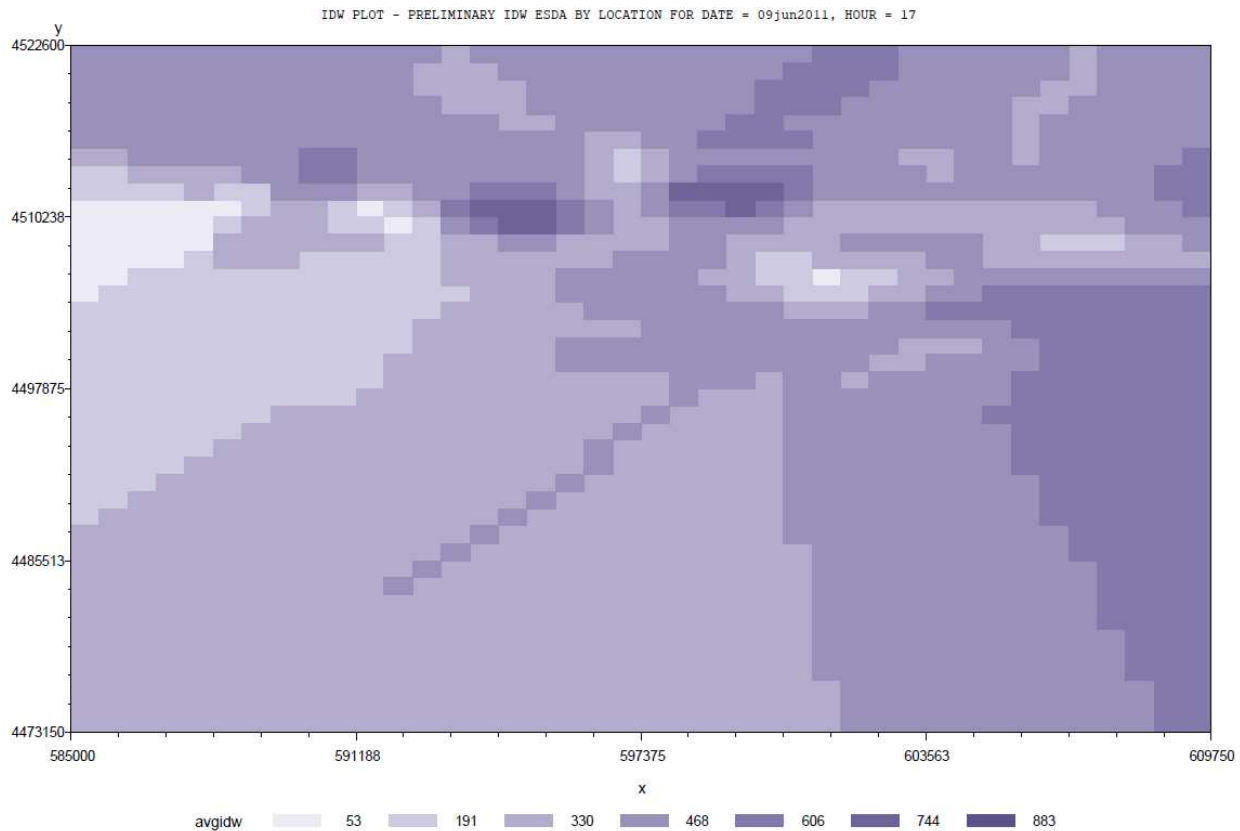


**Figure 4.6: Business IDW Map for July 22 at Hour 16**



**Figure 4.7: Business IDW Map for July 21 at Hour 18**





**Figure 4.8: Business IDW Map for June 9 at Hour 17**

#### **4.2.1.3 IDW Map Summary**

The IDW maps indicate the existence of spatial autocorrelation, a hypothesis that is subjected to statistical (rather than merely visual) tests in the next section. It is difficult to visually assess whether or not stratification and transformations of the dependent variable make a significant difference with regard to spatial autocorrelation. The companion graphics data file contains a complete set of maps available for browsing by the reader. A more concise indication is provided by the Geary's C and Moran's I statistics presented in the next section of this thesis.

#### 4.2.2 Moran's I and Geary's C

As discussed in Section 2.4.2 of this thesis, Moran's I and Geary's C are commonly used statistics to indicate the presence of spatial autocorrelation. The values of both statistics are heavily influenced by the choice of the spatial distance used to separate customers into distance categories. Several rules of thumb are available to help select an appropriate spatial distance (SAS Institute, 2012, p. 8565-8567):

- Each lag distance class should contain a sufficient number of data pairs to maintain computational accuracy. A minimum of 30 pairs per lag distance class is recommended.
- A sufficient number of lag distance classes is needed to capture the extent to which the data are spatially correlated and to define the shape of the semivariance function. Because the function typically has an "S" shape, a minimum of 5 lag distance classes would be needed.

For both residential and business customers, and both overall and within strata, several different values for lag distance classes were tested in an effort to find the right balance between the number of pairs within each lag distance class and the number of available lag distance classes. The final lag distances and number of lag classes used in the calculation of Geary's I and Moran's C statistics are as follows:

- Residential customers by stratum
  - Lag distance = 1,492.278
  - Number of lag distance classes considered= 6
- Residential customers overall, without stratification
  - Lag distance = 1,056.46

- Number of lag distance classes considered = 11
- Business customers by stratum
  - Lag distance = 2,060.758
  - Number of lag distance classes considered = 6
- Business customers overall, without stratification
  - Lag distance = 1,317.92
  - Number of lag distance classes considered = 11

Using these values, Moran's I and Geary's C statistics were calculated for residential and business customers at each of 20 different date-hour combinations: ten of the date-hour combinations are peak dates and times for energy provider operations, and ten are randomly selected from the universe of the 8,760 date-hour combinations in the year.

#### **4.2.2.1 *Moran's I and Geary's C for Residential Customers***

As described previously, Moran's I centers around zero, and is considered a significant indicator of spatial autocorrelation only if it falls outside the range of -0.3 to +0.3. For residential customers, Table 4.9 provides Moran's I ranges for groups of the dependent variables being examined in this thesis, including both raw energy usage data and its four transformations. Values falling outside the (-0.3, 0.3) range are highlighted in green. Appendix C provides additional details, including the Moran's I values for each of the potential independent variables with spatial variation, virtually all of which have at least one stratum outside the (-0.3, 0.3) range<sup>9</sup>.

---

<sup>9</sup> A companion data file to this thesis provides a significant amount of additional details. The file `graves_geary_moran.csv` contains Geary and Moran statistics and z-scores for both residential and business

**Table 4.9: Moran's I Ranges for Dependent Variables for Residential Customers**

Variable	Range	Overall	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6
kwh1-kwh24	low	0.010	-0.001	-0.017	-0.019	-0.154	-0.378	-0.093
	high	0.012	0.050	0.050	0.086	0.295	1.263	0.058
pctd1-pctd24	low	0.005	0.004	-0.015	0.006	-0.139	-0.581	-0.230
	high	0.078	0.113	0.139	0.159	0.450	0.801	-0.013
pctm1-pctm24	low	-0.005	-0.012	-0.039	-0.060	-0.148	-0.684	-0.252
	high	0.041	0.038	0.071	0.094	0.200	0.283	-0.008
pcta1-pcta24	low	0.011	0.002	-0.023	-0.021	-0.155	-0.427	-0.250
	high	0.065	0.100	0.057	0.073	0.249	0.959	-0.080
delt1-delt24	low	-0.009	-0.005	-0.033	-0.014	-0.198	-0.709	-0.192
	high	0.033	0.071	0.027	0.153	0.092	0.242	0.035

Geary's C values center around one, with larger positive values or values close to zero indicating significant spatial autocorrelation. Table 4.10 provides the range of Geary's C values for the dependent variables for residential customers. Cells are highlighted if the absolute z-value associated with the Geary's C exceeds 1.96, the 95% threshold. Appendix C includes Geary's C values for the independent variables.

For residential customers, Moran's I, the more global measure of spatial autocorrelation, shows little significance other than in stratum 5. Geary's C, however, indicates more local patterns of spatial autocorrelation. However, the use of stratification seems to provide little value. Three of the four transformations of hourly interval energy usage data also offer no advantage.

---

customers, for all transformations of the dependent variable, both overall and by stratum. Geary and Moran scores for the geodemographic variables are also included.

**Table 4.10: Geary's C Ranges for Dependent Variables for Residential Customers**

Variable	Range	Overall	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6
kwh1-kwh24	low	0.003	0.740	0.838	0.530	0.450	0.304	0.786
	high	0.007	1.059	1.293	1.181	0.946	1.190	1.013
pctd1-pctd24	low	0.906	0.867	0.820	0.781	0.683	0.087	0.936
	high	1.049	1.016	1.059	1.008	1.102	1.693	1.242
pctm1-pctm24	low	0.767	0.602	0.573	0.926	0.579	0.387	0.944
	high	1.239	1.200	1.364	1.095	1.159	1.998	1.294
pcta1-pcta24	low	0.833	0.807	0.908	0.592	0.479	0.271	0.841
	high	1.094	1.053	1.362	1.144	0.979	1.390	1.314
delt1-delt24	low	0.438	0.223	0.684	0.727	0.678	0.087	0.817
	high	1.117	1.264	1.718	1.689	1.741	2.429	1.197

#### 4.2.2.2 Moran's I and Geary's C for Business Customers

Table 4.11 provides a summary of Moran's I statistics for the independent variables for business customers. Values falling outside the (-0.3, 0.3) range are highlighted in green. Appendix C provides additional details, including the Moran's I values for each of the potential independent variables with spatial variation, virtually all of which have at least one stratum outside the (-0.3, 0.3) range.

**Table 4.11: Moran's I Ranges for Dependent Variables for Business Customers**

Variable	Range	Overall	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6
kwh1-kwh24	low	0.600	0.006	-0.094	-0.036	-0.341	0.072	0.130
	high	0.661	0.233	0.328	1.123	0.053	0.850	1.130
pctd1-pctd24	low	0.012	0.016	-0.148	-0.109	-0.172	0.009	-0.257
	high	0.425	0.379	0.469	1.232	0.208	0.328	0.435
pctm1-pctm24	low	0.012	0.092	-0.278	-0.074	-0.091	-0.113	-0.229
	high	0.547	0.425	0.129	1.081	0.170	0.705	0.232
pcta1-pcta24	low	0.026	0.151	-0.150	-0.040	-0.128	-0.107	-0.285
	high	0.523	0.380	0.537	1.118	0.103	0.726	0.238
delt1-delt24	low	-0.138	-0.155	-0.249	-0.300	-0.271	-0.222	-0.188
	high	0.241	0.239	0.608	1.166	0.044	0.169	0.112

Table 4.12 summarizes Geary's C statistics for business customers, showing the minimum and maximum values that were calculated for different day-hour combinations. Cells are highlighted if the absolute z-value associated with the Geary's C exceeds 1.96, the 95% threshold. Appendix C includes Geary's C values for the independent variables.

**Table 4.12: Geary's C Ranges for Dependent Variables for Business Customers**

Variable	Range	Overall	Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5	Stratum 6
kwh1-kwh24	low	0.082	0.096	0.430	0.086	0.726	0.126	0.176
	high	0.153	0.322	1.088	1.181	1.436	0.270	1.688
pctd1-pctd24	low	0.698	0.355	0.621	0.140	0.054	0.151	0.089
	high	1.075	1.001	1.233	1.501	1.200	0.248	2.149
pctm1-pctm24	low	0.559	0.297	0.413	0.309	0.374	0.182	0.196
	high	1.185	0.756	1.543	1.433	1.176	0.559	1.790
pcta1-pcta24	low	0.535	0.298	0.675	0.190	0.483	0.175	0.202
	high	1.049	0.891	1.143	1.272	1.254	0.623	2.121
delt1-delt24	low	0.025	0.301	0.149	0.170	0.004	0.090	0.091
	high	5.125	1.427	1.850	2.147	0.950	0.784	1.497

For business customers, both Moran's I and Geary's C results indicate a smattering of spatial autocorrelation for various transformations of the independent variable and at the stratum level. Neither Moran's I, the more global measure of spatial autocorrelation, nor Geary's C, the more local measure, seem to provide any evidence that stratification or transformations of the dependent variable will assist the analysis.

#### 4.2.2.3 *Moran's I and Geary's C Summary*

The results provided in the text and Appendix C both offer support for the presence of spatial autocorrelation in both the dependent and independent variables. It does not appear, however, that stratification or transformation of the dependent variable is necessary to detect spatial autocorrelation.

### 4.2.3 Semivariograms

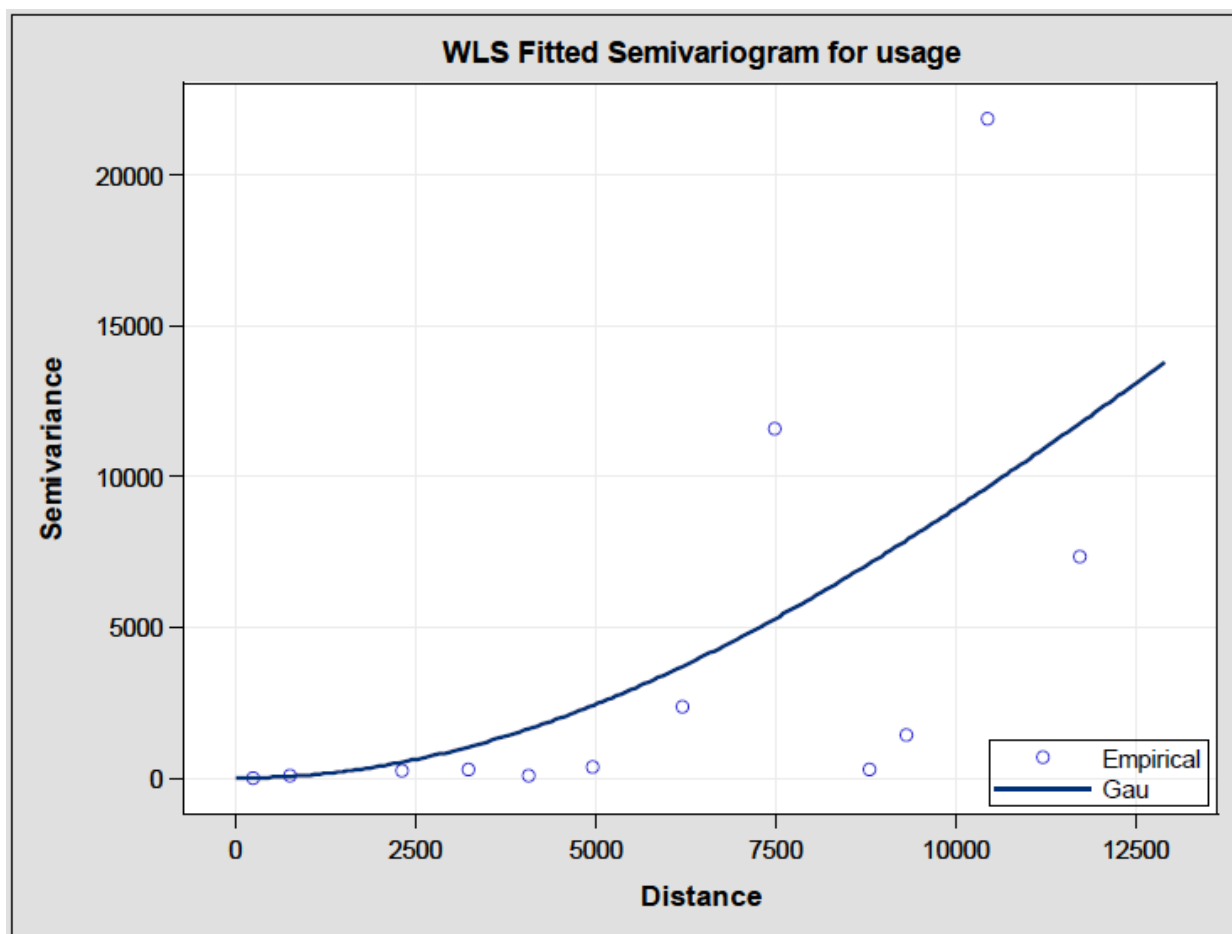
Using the lag distance information developed for the Moran's I and Geary's C statistics, as described previously, semivariograms were created to examine the spatial autocorrelation of hourly interval energy usage data. The Moran's I and Geary's C results indicate that, for spatial statistics, the raw hourly interval energy usage data without stratification provides the most robust results. Therefore, semivariograms were only created for customers overall and without using a transformation for the dependent variable. The resulting semivariograms are summarized in the sections below<sup>10</sup>.

#### 4.2.3.1 *Semivarograms for Residential Customers*

The ten semivariograms for residential customers are similar in shape for each of the ten day-hour combinations tested, although the maximum sill value of the semivariance statistic (shown on the y-axis) varies. In Figure 4.9, for example, the extrapolated semivariance statistic reaches a value of approximately 13,000 at a distance range of 12,500 meters. In Figure 4.10, however, the semivariance statistic reaches a maximum sill value of about 1,600 at the same distance range. The overall shapes are similar with a nugget of zero and no evident flattening.

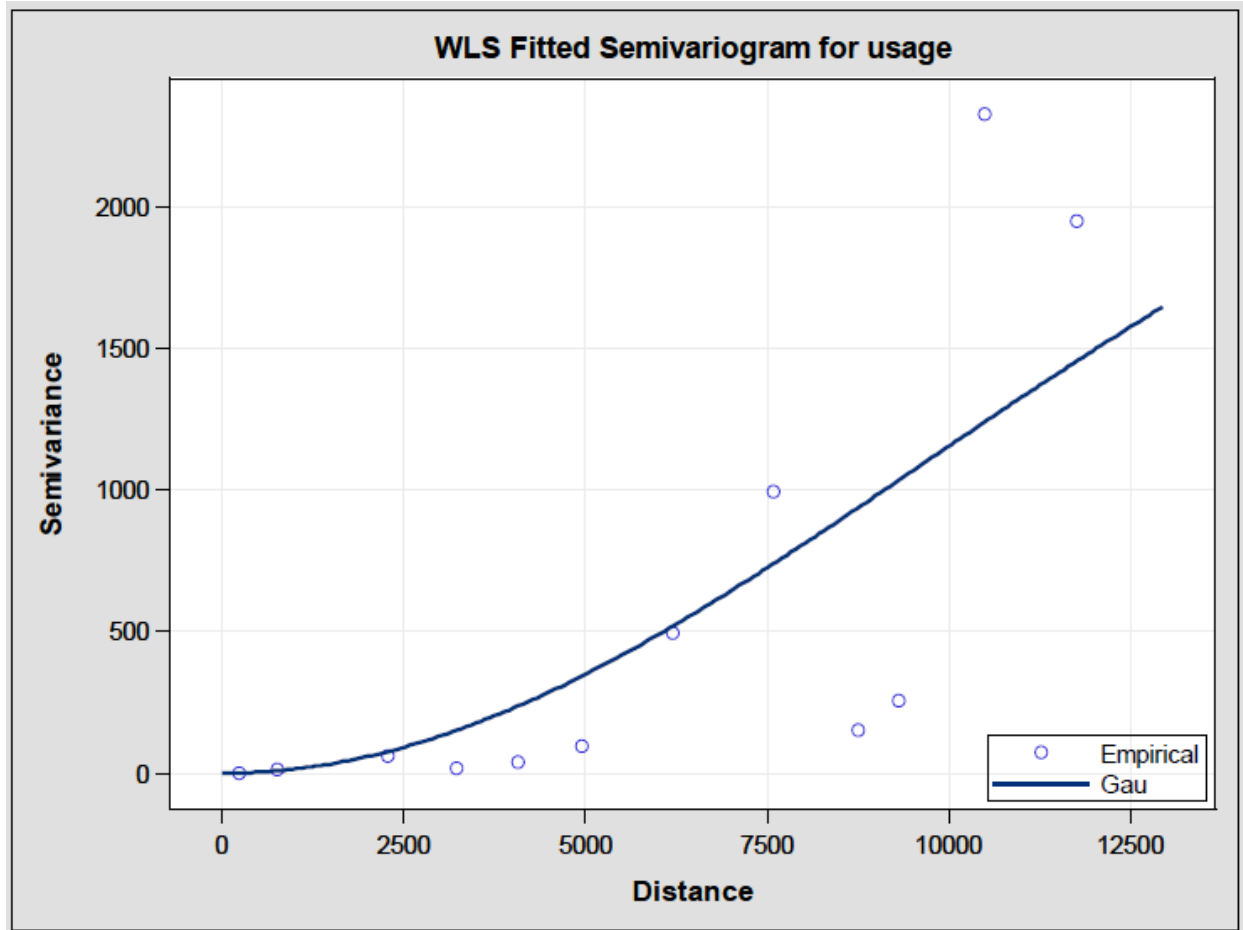
---

<sup>10</sup> A companion data file to this thesis provides a more complete set of semivariograms. The file *graves\_thesis\_graphics.pdf* contains semivariograms for residential and business customers for selected day-hour combinations.



**Figure 4.9: Residential Semivariogram for July 22, Hour 16**

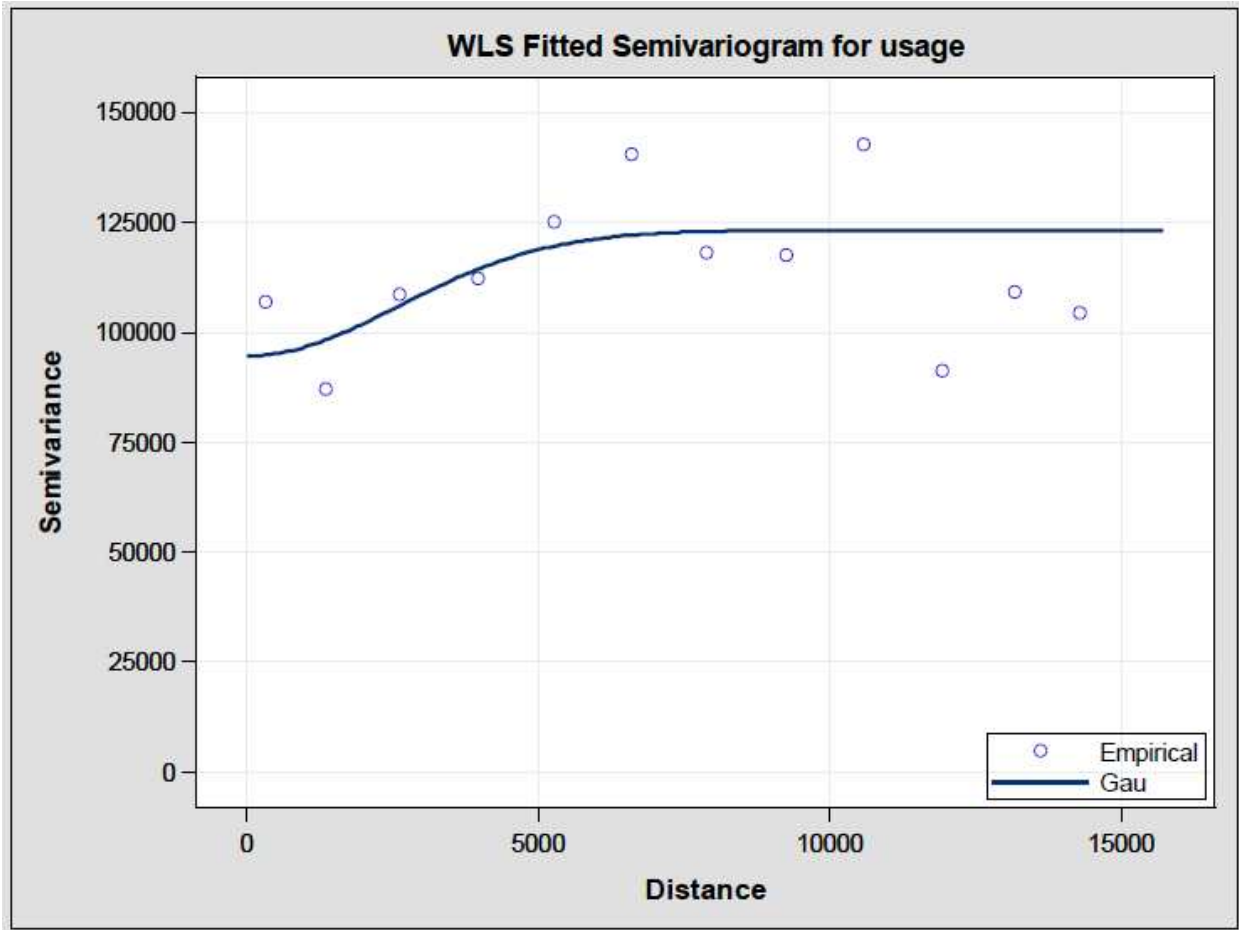




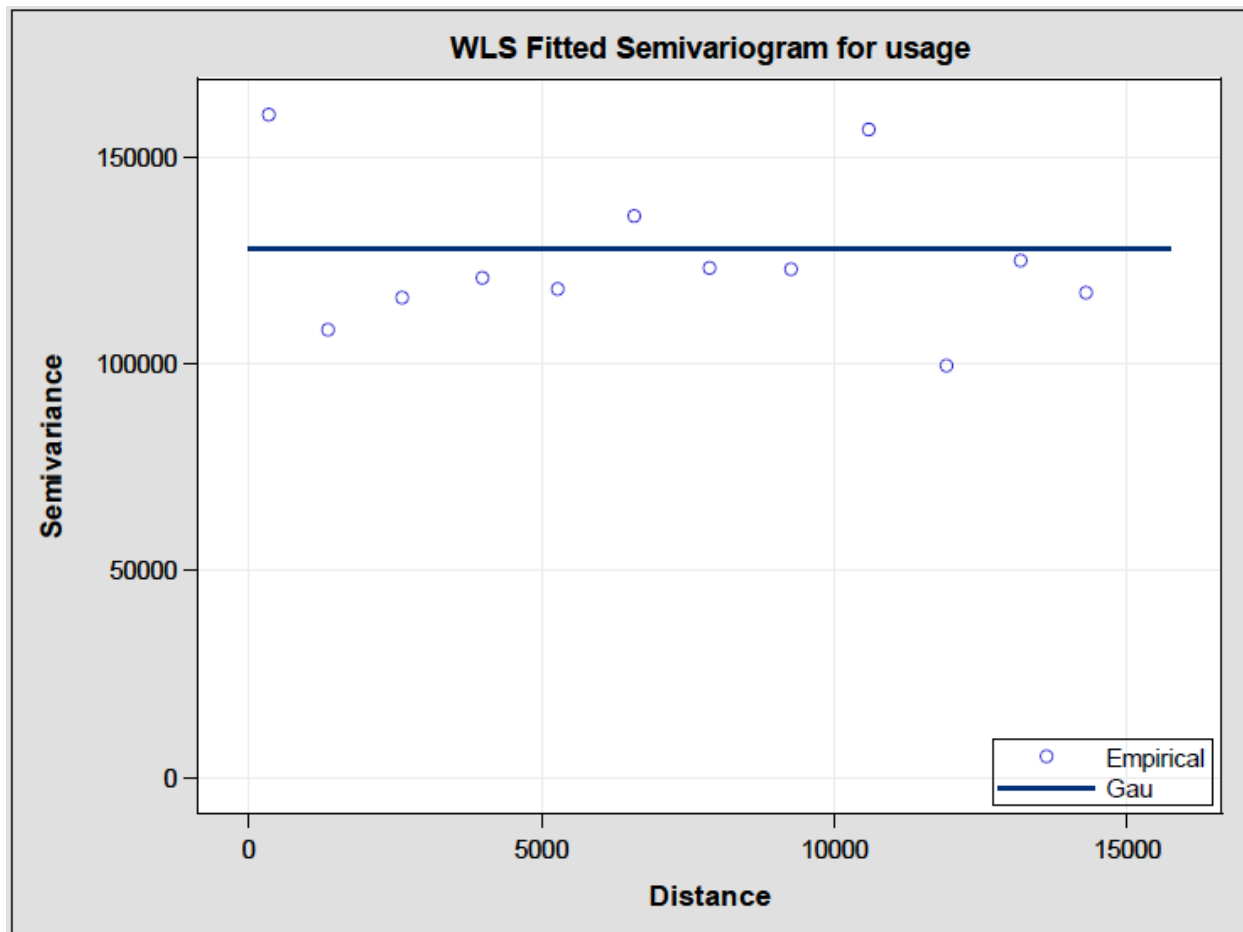
**Figure 4.10: Residential Semivariogram for January 12, Hour 18**

#### 4.2.3.2 *Semivariograms for Business Customers*

For business customers, two distinct semivariogram shapes are evident. Eight of the ten day-hour combinations are similar in shape to that shown in Figure 4.11, although the sill value varies between about 125,000 (as shown in Figure 4.11) and about 6,000, both over a distance of about 15,000 meters. In all cases, the nugget is fairly large, perhaps about 20 percent of the sill value. The other shape that appears for two of the ten day-hour combinations examined is illustrated in Figure 4.12, in which the semivariogram is perfectly horizontal. A flat semivariogram such as this demonstrates no spatial autocorrelation, because the estimated usage is the same at all spatial distances.



**Figure 4.11: Business Semivariogram for June 9, Hour 17**



**Figure 4.12: Business Semivariogram for July 12, Hour 17**

#### **4.2.3.3 *Semivariogram Summary***

The semivariogram results presented here and in companion files indicate that, at least for most of the day-hour combinations examined, spatial autocorrelation exists. The extent varies from one day-hour combination to another, indicating that the temporal aspect of the problem also must be addressed.

#### **4.2.4 Spatial Exploratory Data Analysis Summary**

For both residential and business customers, the spatial statistics presented in this section support the existence of spatial autocorrelation. The IDW maps illustrate "hot" and

"cold" areas for usage. The Moran's I and Geary's C statistics are significant, particularly for residential and business customers overall. There is little evidence that transformations of the dependent variable provide any value spatially.

The next section of this thesis examines temporal data patterns, with the goal of identifying appropriate temporal lags. After such lags are determined, spatiotemporal exploratory analysis is discussed in the following section.

### **4.3 Temporal Exploratory Data Analysis**

In this section, the sample autocorrelation function is used to develop correlograms, which are then examined to determine appropriate temporal lags. Transformations of the dependent variable are examined, as is the stratification of the residential and business customers. Temporal lags up to 672 hours (4 weeks) were examined. It is assumed that this period would be sufficiently long to identify any critically important temporal lags. Although longer temporal lags may also be important, such as a lag of one year, the dependent data is generally analyzed for a single calendar year, making a one-year lag impractical for gap-filling purposes.

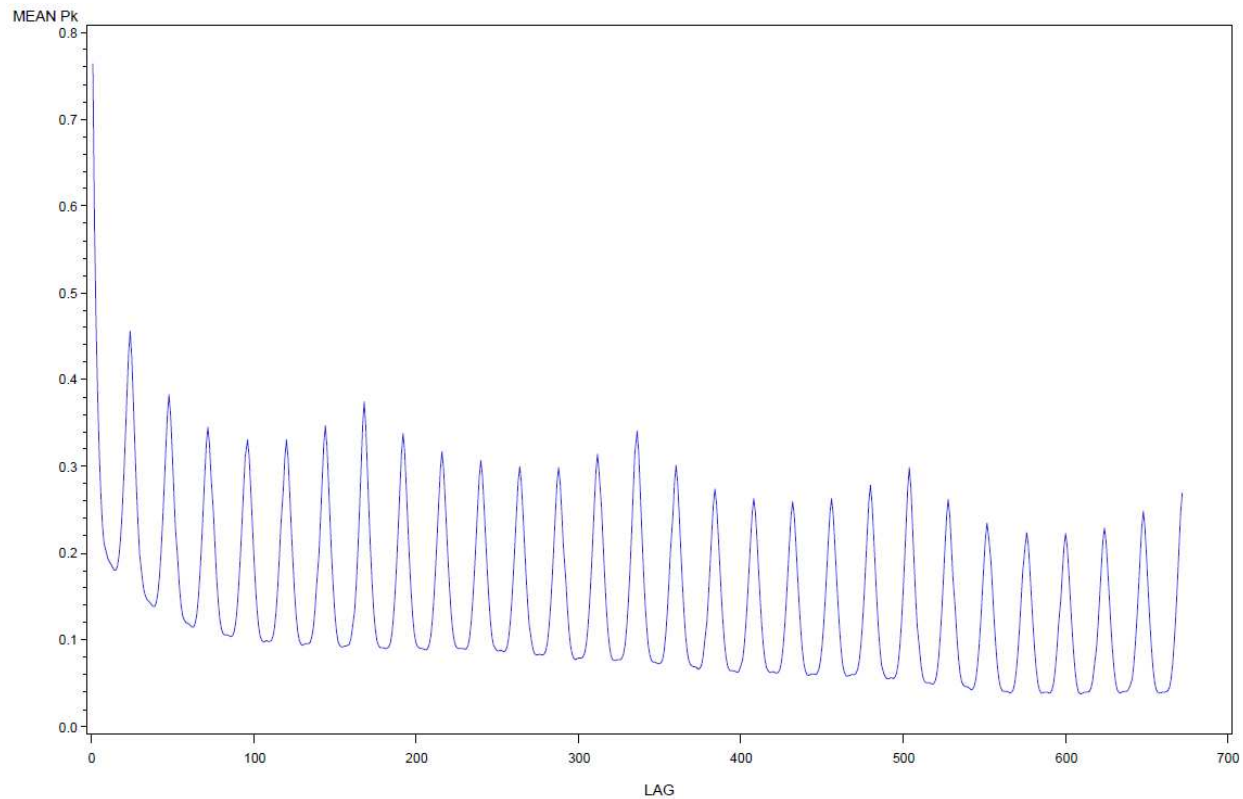
#### **4.3.1 Temporal Exploratory Data Analysis for Residential Customers**

For residential customers, a sample of correlograms is shown<sup>11</sup>. Figure 4.13 shows a typical correlogram, illustrated for all residential customers overall using the raw hourly interval energy usage data. The one-hour lag shows the strongest correlation, followed by a series of lags at 24-hour intervals. The weekly lags, at intervals of 168 hours, are also

---

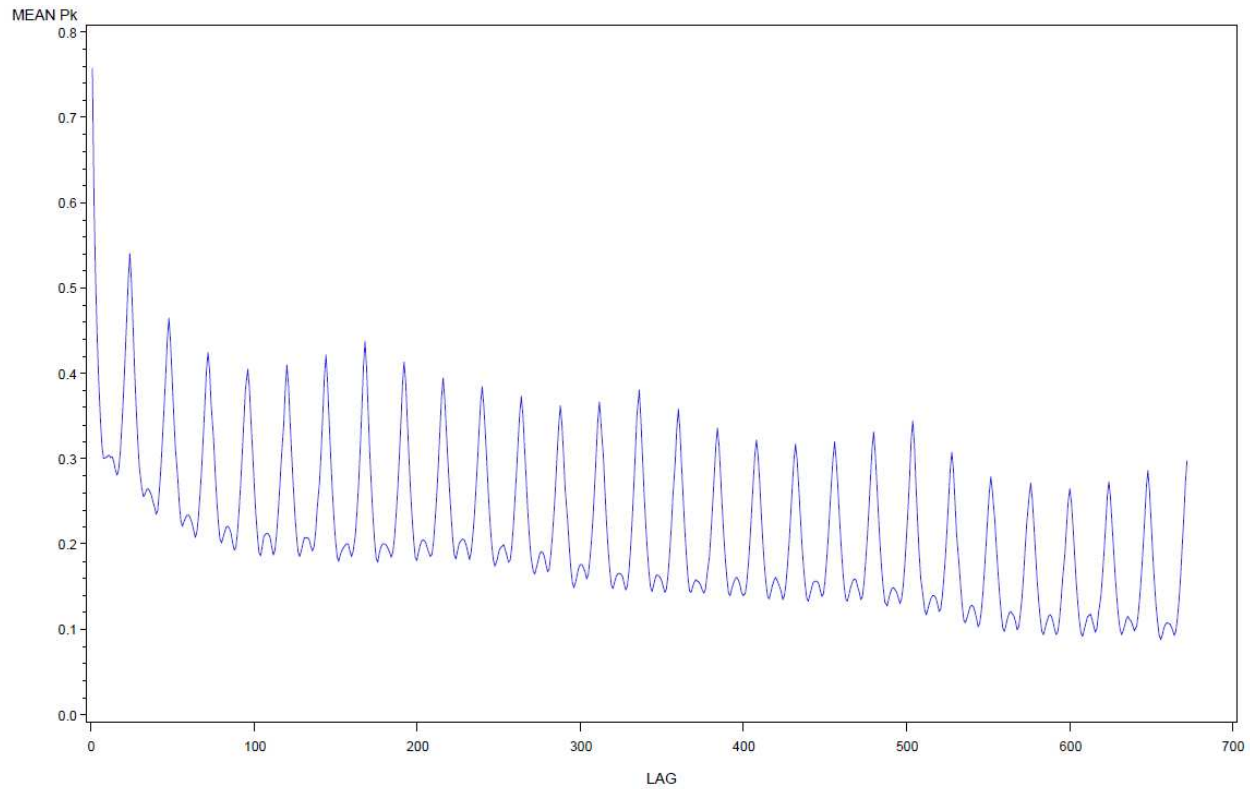
<sup>11</sup> A companion data file to this thesis provides a more complete set of correlograms. The file `graves_thesis_graphics.pdf` contains correlograms for both residential and business customers for all transformations of the dependent variable, both overall and by stratum.

strong but decline over time. The average correlation falls below 0.5, however, even for the initial 24-hour lag; only the one-hour lag has a high correlation.



**Figure 4.13: Residential Correlogram, Raw Energy Usage Interval Data, No Stratification**

Figure 4.14 illustrates a phenomenon that appears in some of the correlograms, in which an extra bump in correlation appears at the 12-hour mark in between the 24-hour lags. The same pattern of weekly lags appears, as does the tapering off over time. The mean correlation value of the temporal lags continues to be high only for the one-hour lag.



**Figure 4.14: Residential Correlogram, Raw Energy Usage Interval Data, Stratum 5**

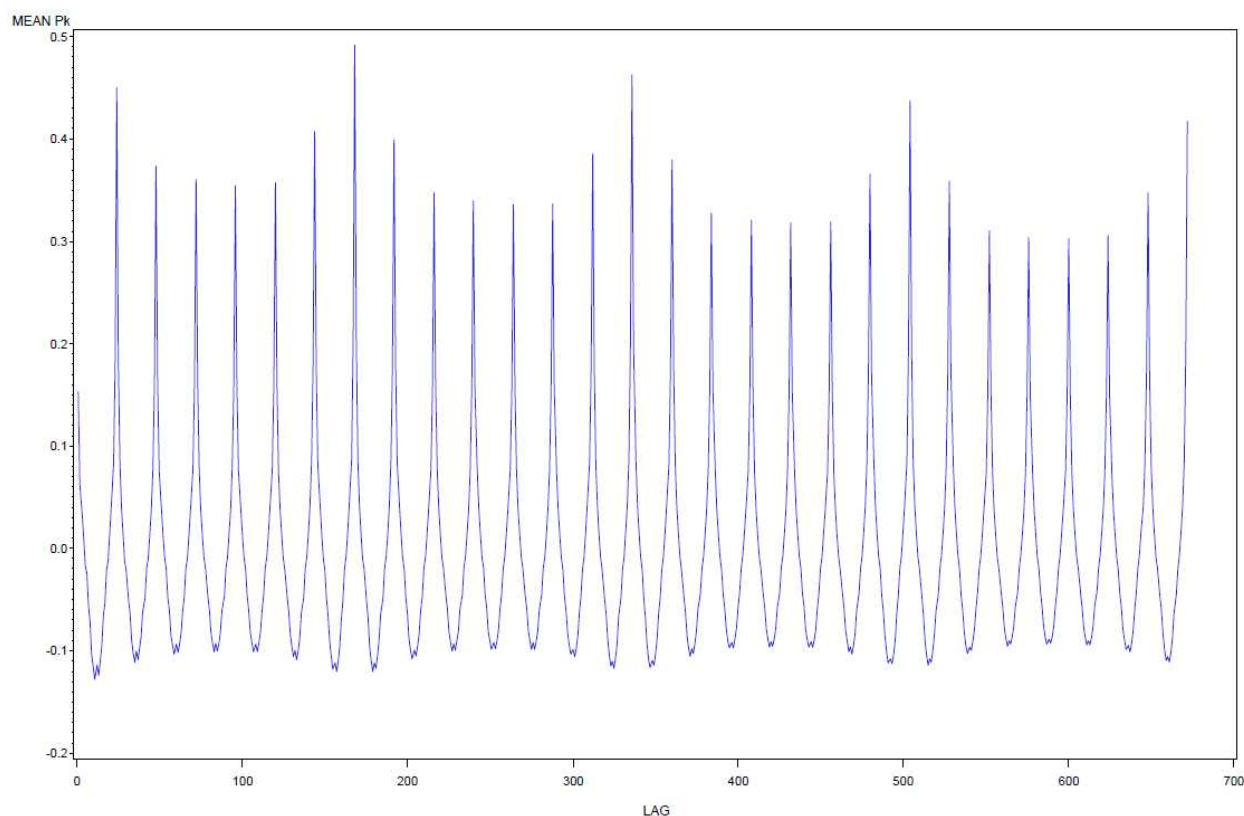
Table 4.13 summarizes the mean correlation values at the set of lags mentioned above: one-hour, 24-hour, 36-hour, and 168-hours. For each of the transformations of the dependent variable, the one-hour lag consistently has the highest average correlation coefficients. The raw hourly interval energy usage data has coefficients that are as high or higher than any other transformation of the dependent variable. Within the one-hour lag category for the raw hourly interval energy usage data (kwh1-kwh24), the larger strata typically have a higher correlation. The overall correlation, however, is still significant. All correlations with an average value of 0.67 or higher are highlighted in Table 4.13.

**Table 4.13: Average Correlation of Temporal Lags for Residential Customers**

Dependent Variable	Stratification	One-Hour Lag	24-Hour Lag	36-Hour Lag	168-Hour Lag
kwh1-kwh24	Overall	0.76	0.46	0.14	0.37
	Stratum 1	0.75	0.44	0.13	0.36
	Stratum 2	0.77	0.45	0.15	0.37
	Stratum 3	0.79	0.47	0.16	0.39
	Stratum 4	0.79	0.50	0.20	0.39
	Stratum 5	0.76	0.54	0.26	0.44
	Stratum 6	0.92	0.60	-0.11	0.69
pctd1-pctd24	Overall	0.70	0.35	-0.01	0.32
	Stratum 1	0.68	0.35	0.02	0.32
	Stratum 2	0.71	0.35	-0.02	0.32
	Stratum 3	0.72	0.34	-0.05	0.32
	Stratum 4	0.71	0.34	-0.02	0.30
	Stratum 5	0.70	0.35	-0.00	0.32
	Stratum 6	0.88	0.56	-0.33	0.68
pctm1-pctm24	Overall	0.74	0.41	0.08	0.32
	Stratum 1	0.72	0.40	0.08	0.30
	Stratum 2	0.74	0.41	0.09	0.32
	Stratum 3	0.75	0.42	0.08	0.32
	Stratum 4	0.77	0.45	0.13	0.31
	Stratum 5	0.74	0.49	0.22	0.35
	Stratum 6	0.92	0.58	-0.17	0.65
pcta1-pcta24	Overall	0.76	0.46	0.14	0.37
	Stratum 1	0.75	0.44	0.13	0.36
	Stratum 2	0.77	0.45	0.15	0.37
	Stratum 3	0.79	0.47	0.16	0.39
	Stratum 4	0.79	0.50	0.20	0.39
	Stratum 5	0.76	0.54	0.26	0.44
	Stratum 6	0.92	0.60	-0.11	0.69
delt1-delt24	Overall	-0.04	0.13	0.01	0.14
	Stratum 1	-0.07	0.12	0.01	0.13
	Stratum 2	-0.04	0.13	0.01	0.14
	Stratum 3	-0.03	0.13	0.00	0.13
	Stratum 4	-0.03	0.14	0.01	0.13
	Stratum 5	-0.02	0.15	0.02	0.15
	Stratum 6	0.27	0.42	-0.07	0.52

### 4.3.2 Temporal Exploratory Data Analysis for Business Customers

For business customers, the correlation with the 168-hour lag (one week) sometimes exceeds that of the one-hour lag. Figure 4.15 illustrates such an example, taken from the dependent variable transformation in which hourly interval energy usage data is shown as a percentage of its prior hourly value. In this transformation, the one-week lag has the strongest correlations.



**Figure 4.15: Business Correlogram, Energy Usage Interval Data as Percent of Prior Hour's Interval, Overall**

Table 4.14 presents the average correlations for the set of four temporal lags discussed above. For these customers, the one-hour and 168-hour lags have correlations that are consistently as high or higher than those of the 24-hour or 36-hour lags. The raw hourly interval energy usage data also provides correlations that are as high or higher than those of other dependent variable data transformations. Similarly to the finding with residential customers, the use of stratification with temporal lags does not seem to offer any significant improvement.



**Table 4.14: Average Correlation of Temporal Lags for Business Customers**

Dependent Variable	Stratification	One-Hour Lag	24-Hour Lag	36-Hour Lag	168-Hour Lag
kwh1-kwh24	Overall	0.91	0.72	0.01	0.75
	Stratum 1	0.84	0.62	-0.12	0.63
	Stratum 2	0.90	0.69	-0.06	0.72
	Stratum 3	0.93	0.70	-0.07	0.77
	Stratum 4	0.91	0.68	-0.13	0.77
	Stratum 5	0.93	0.78	0.04	0.78
	Stratum 6	0.94	0.79	0.11	0.79
pctd1-pctd24	Overall	0.84	0.58	-0.34	0.68
	Stratum 1	0.75	0.49	-0.24	0.56
	Stratum 2	0.82	0.57	-0.33	0.67
	Stratum 3	0.87	0.56	-0.33	0.71
	Stratum 4	0.86	0.55	-0.33	0.71
	Stratum 5	0.87	0.66	-0.40	0.74
	Stratum 6	0.86	0.63	-0.39	0.69
pctm1-pctm24	Overall	0.90	0.69	-0.10	0.69
	Stratum 1	0.83	0.62	-0.07	0.63
	Stratum 2	0.89	0.67	-0.14	0.67
	Stratum 3	0.92	0.67	-0.16	0.73
	Stratum 4	0.90	0.66	-0.16	0.72
	Stratum 5	0.91	0.73	-0.12	0.71
	Stratum 6	0.92	0.73	-0.05	0.69
pcta1-pcta24	Overall	0.91	0.72	0.01	0.75
	Stratum 1	0.84	0.63	-0.03	0.63
	Stratum 2	0.90	0.69	-0.06	0.72
	Stratum 3	0.93	0.69	-0.08	0.76
	Stratum 4	0.91	0.67	-0.11	0.76
	Stratum 5	0.92	0.76	0.02	0.77
	Stratum 6	0.94	0.79	0.12	0.79
delt1-delt24	Overall	0.15	0.45	-0.10	0.49
	Stratum 1	-0.00	0.32	-0.04	0.35
	Stratum 2	0.09	0.41	-0.09	0.45
	Stratum 3	0.16	0.45	-0.11	0.51
	Stratum 4	0.21	0.45	-0.11	0.52
	Stratum 5	0.17	0.54	-0.10	0.56
	Stratum 6	0.23	0.50	-0.13	0.54

### 4.3.3 Temporal Exploratory Data Analysis Summary

For both residential and business customers, significant temporal lags have been found. For residential customers only the one-hour lag is significant, while for business customers, significant lags appear at one-hour, 24-hours, and 168-hours. Neither stratification nor data transformations of the dependent variable seem to offer any real improvement in the results obtained.

## 4.4 Spatiotemporal Exploratory Data Analysis

The foregoing exploratory data analysis has consistently indicated that data transformations of the dependent variable and stratification of the customer groups do not offer improvements in the analytic results. This finding holds in the correlation coefficients, the spatial exploratory data analysis, and the temporal exploratory data analysis. For this reason, the spatiotemporal exploratory analysis will commence without the use of stratification or data transformation.

### 4.4.1 Line Graphs

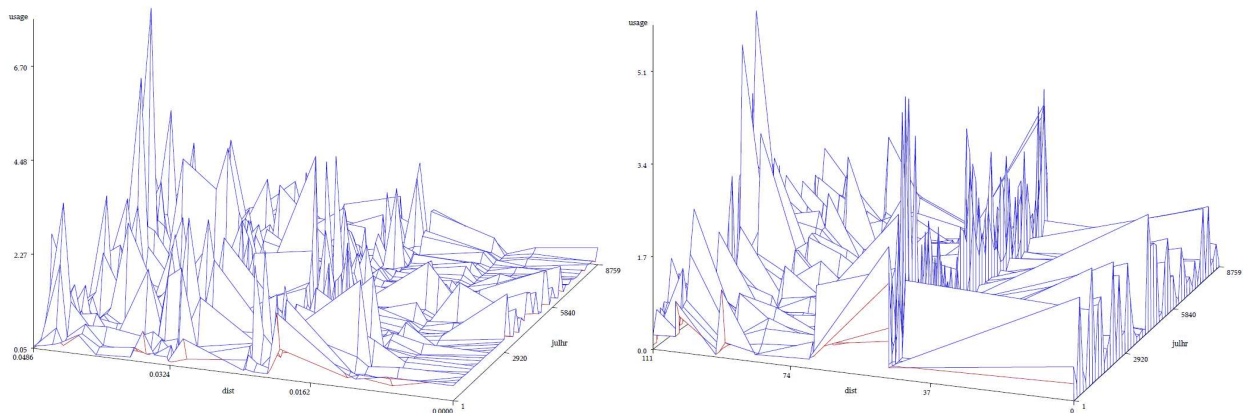
As presented previously, line graphs in a two-dimensional space make it impossible to distinguish the spatial lags between customer locations. For the purposes of exploratory spatiotemporal analysis, therefore, the line graphics were converted to a three-dimensional version in which the x-axis shows the hours of the year, the y-axis shows the spatial lag, and the z-axis shows the usage pattern.

Figure 4.16 provides three-dimensional line graphs for two randomly-chosen residential customers and their 20 nearest neighbors. The usage pattern over the year for each randomly-chosen customer is shown adjacent to the x-axis. The 20 nearest neighbors are shown with the y-axis indicating their spatial distances from the randomly chosen customer<sup>12</sup>. (Note that the red lines on the graph indicate that we are looking at the underside of the surface.) The interval data usage patterns of each the various neighbors can be seen here. On the left-hand graph, it appears that if a smoothed surface were fitted

---

<sup>12</sup> A companion file to this thesis provides a more complete set of line graphs. The file `graves_thesis_graphics.pdf` contains three-dimensional line graphs for both residential and business customers for all transformations of the dependent variable. Graphs are provided for a random sample of 10 customers in conjunction with the neighbors of those customers.

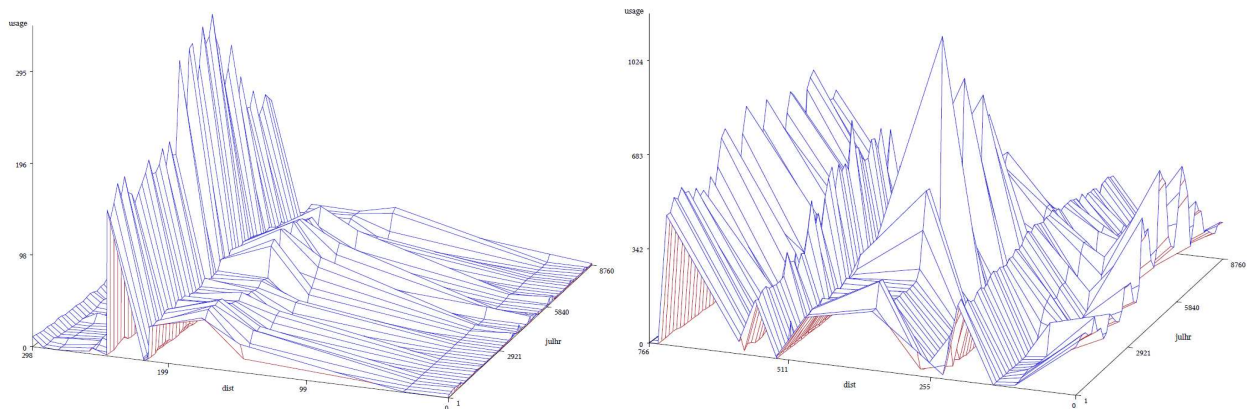
to the data, a steeper slope up and away from the randomly-chosen customer would begin at a distance of about 0.2 meters; the extremely short distance indicates that these neighbors are probably located within the same building as the randomly-chosen customer. With the exception of one neighbor located at about 50 meters, the right-hand graph shows an approximate U-shaped pattern over its much larger (but still small) range of approximately 110 meters. In both graphs it is difficult to compare the temporal patterns of the different customers, or to identify appropriate temporal or spatial lags for further analysis.



**Figure 4.16: Line Graphs for Two Random Residential Customers and Their Neighbors**

Figure 4.17 provides three-dimensional line graphs for two randomly-chosen business customers and the 20 nearest neighbors of each. The distance range on the left-hand graph is about 300 meters, and the right-hand graph has a range of about 750 meters. The general pattern is that values are generally upward sloping as the distance from the randomly-chosen customer increases. On the left-hand graph, there is a sharp increase at about 200 meters distance, while the right-hand graph has a more gradual slope (except for the jump in the neighbor located at a distance of about 300 meters). As with the residential

customer graphics, the identification of appropriate spatial and temporal lags is still problematic.



**Figure 4.17: Line Graphs for Two Random Business Customers and their Neighbors**

#### 4.4.2 Spatiotemporal Correlograms

As with the line graphs, the heat maps discussed previously in Section 2.6.2 gave little actionable information about appropriate spatial or temporal lags. Instead a spatiotemporal correlogram was developed that groups all data for a random set of 25 customers into a grid of spatial and temporal lags. The temporal lags that were examined range from a one-hour lag to a 672-hour (28-day or four-week) lag. The spatial lags are determined by calculating the distance of each of the 25 customers from each of the other 24 customers, identifying the centiles of all the distances, and then retrieving every 4th centile so that there are 25 grid cells in the spatial direction. Within each of the 672-by-25 grid cells, a correlation coefficient is calculated between for each customer in that grid cell between an hourly interval energy usage value and it's temporally-lagged counterpart for the same customer. The correlation values are then averaged for all lags in the grid cell. The results are graphed with the x-axis showing the temporal lag, the y-axis showing the distance lag, and the z-axis showing the correlation coefficient.

Figure 4.18 shows the results of the spatiotemporal correlogram for residential customers<sup>13</sup>. (Due to computer memory restrictions, each set of 168-hour temporal lags is shown on its own set of axes.) The temporal pattern replicates that found previously in Figure 4.14, which is logical because appropriate temporal lags are analogous to autocorrelation. Peak correlations appear at each 24-hour interval, with higher peaks at each 168-hour (one week) interval. Previously, the spatial semivariograms had shown only data for a single hour on a single day, as opposed to the data shown in Figure 4.18, which includes temporal lags up to 168 hours. The spatial lag pattern shown in Figure 4.18 is bi-modal, indicating that the fall-off in the influence of one's neighbors is not monotonically related to distance.

As a way of testing how pervasive the bi-modal spatial lag is, correlograms were created for 50 random customers that included the initial 25 customers, as well as a separate set of 25 random customers. These correlograms are shown in Figure 4.19. All residential results show an immediate drop in correlation on the spatial axis, followed by an increase in correlation, with a subsequent drop-off. The bi-modality is inconsistent, at least for the distances shown here. Significant drops in correlations are seen at distances of about 4,400 meters (Figure 4.18), 9,000 meters (Figure 4.19, left-hand side) and 3,000 meters (Figure 4.19, right-hand side).

Figure 4.20 provides the spatiotemporal correlogram for 25 randomly-chosen business customers. In this presentation, the distinct temporal lag pattern that was seen previously in Figure 4.15 is repeated. The spatial lag pattern differs among the results. The

---

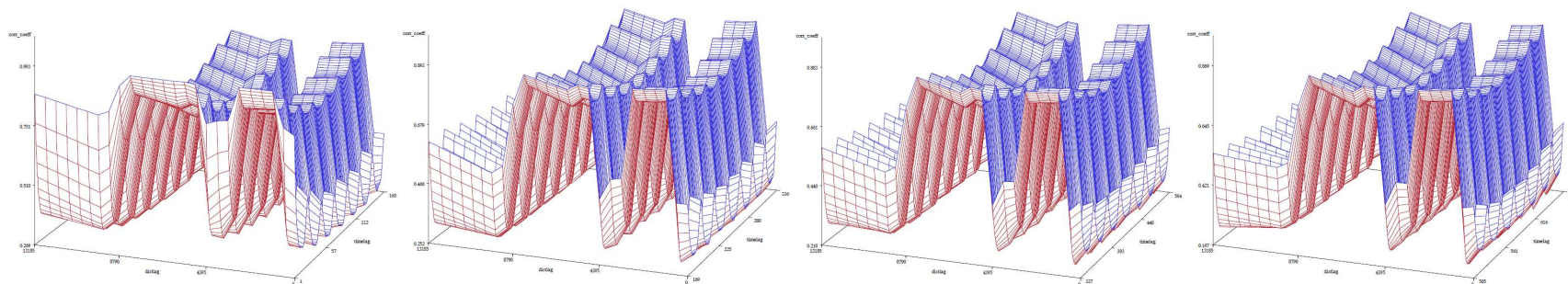
<sup>13</sup> A companion file to this thesis provides a more complete set of spatiotemporal correlograms. The file `graves_thesis_graphics.pdf` contains three-dimensional line graphs for both residential and business customers for random sets of customers.

left-hand side of Figure 4.21 is a correlogram for 50 randomly-chosen customers, including the 25 shown in Figure 4.20. The right-hand side of Figure 4.21 provides a correlogram for a different set of 50 random customers. The spatial lag in Figure 4.20 and 4.21 (left-hand side) go up to nearly 20,000 meters. For the right-hand side of Figure 4.21, the spatial lags go up to about 16,600 meters. As with residential customers, there is inconsistency in the correlogram results.

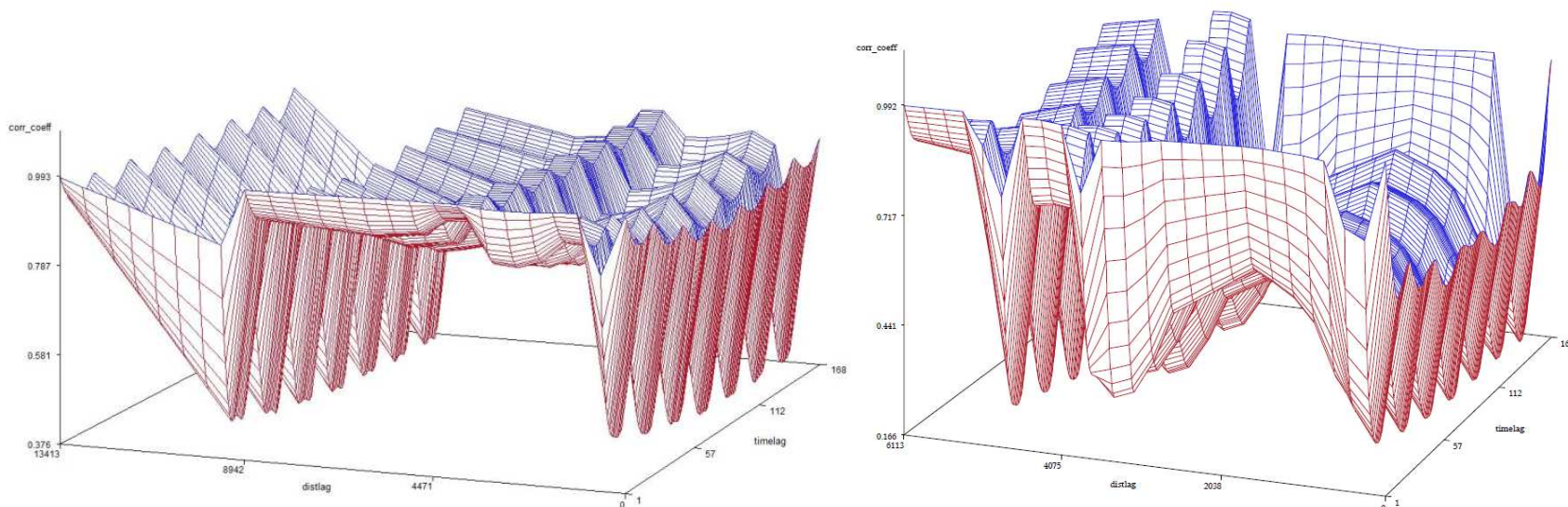
#### **4.4.3 Spatiotemporal Semivariograms**

Unlike the three-dimensional correlograms, which looked at spatial lags between customers and temporal lags within customers, spatiotemporal semivariograms examine both spatial and temporal lags between customers. Additionally, instead of focusing on correlations, semivariograms focus on semivariance of the hourly interval energy usage data within each particular cell combination of temporal and spatial lag.

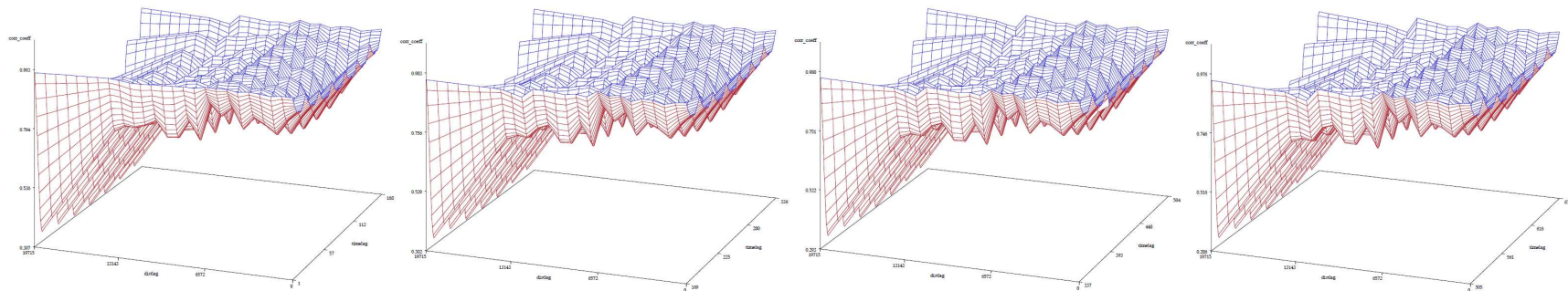




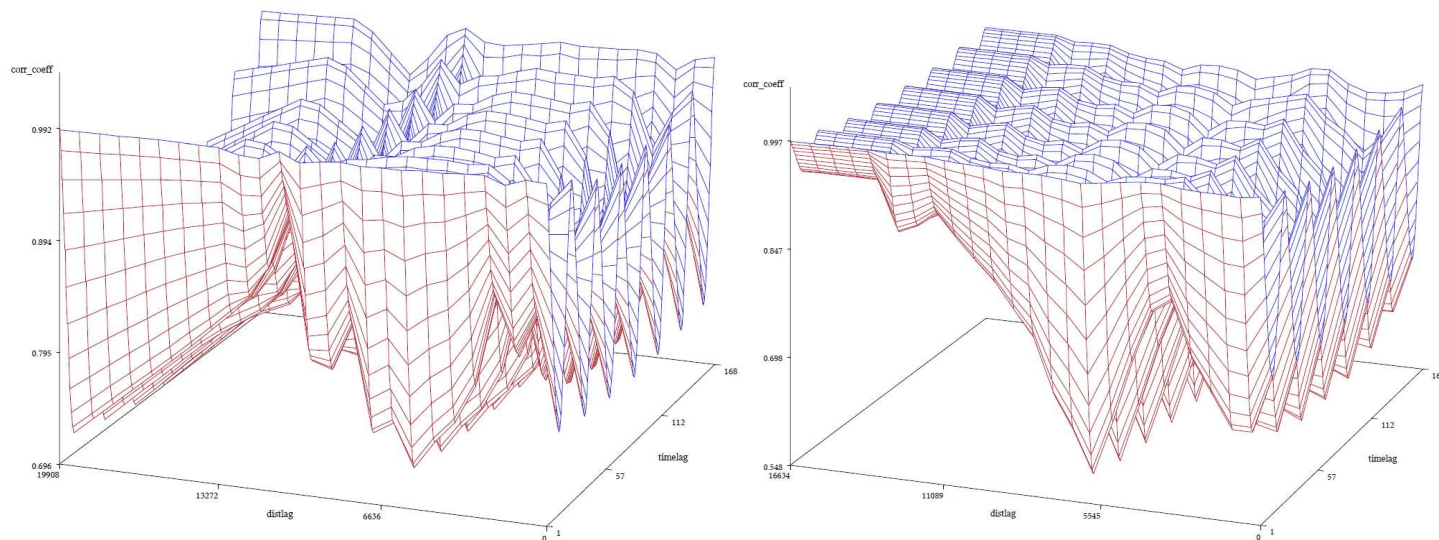
**Figure 4.18: Spatiotemporal Correlogram for Residential Customers**



**Figure 4.19: Spatiotemporal Correlograms for Additional Residential Customers**



**Figure 4.20: Spatiotemporal Correlogram for Business Customers**

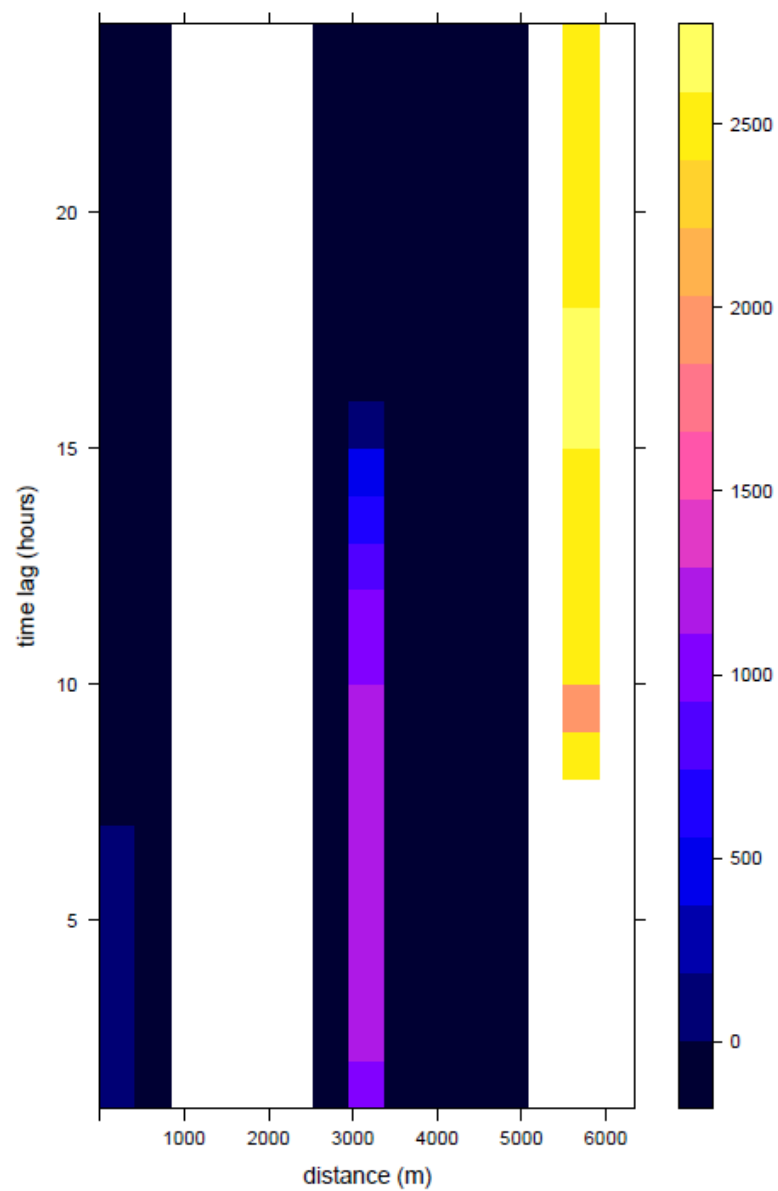


**Figure 4.21: Spatiotemporal Correlograms for Additional Business Customers**

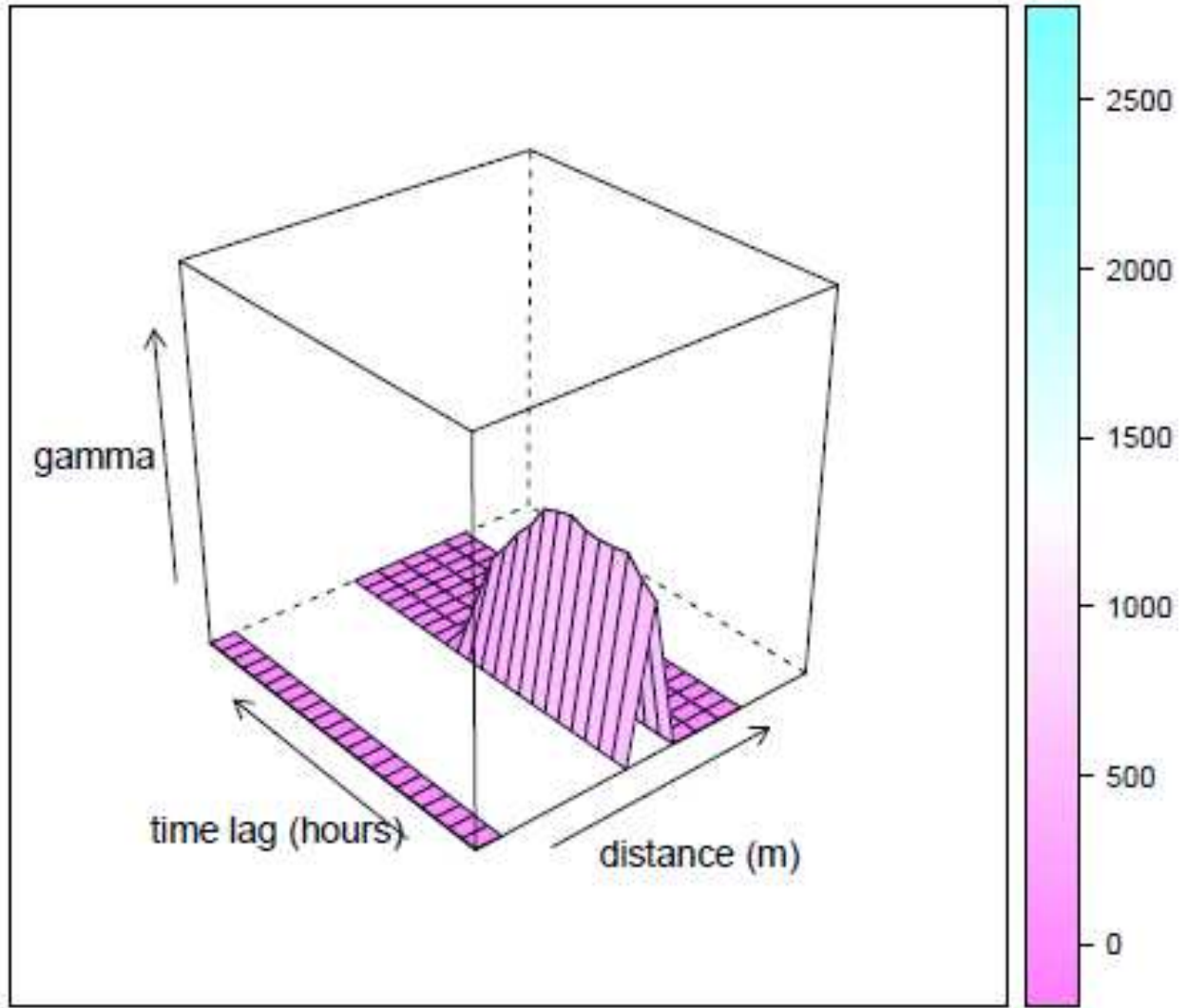


A two-dimensional spatiotemporal semivariogram for residential customers is shown in Figure 4.22. The results shown are for 100 randomly selected residential customers, using one calendar month of data, although the temporal lags examined are limited to between one and 24 hours. In Figure 4.22 the spatial lag is shown on the x-axis, the temporal lag is shown on the y-axis, and the semivariance is shown via a color ramp. In the Figure, the white space represents a grid cell with no available data. Figure 4.23 provides the three-dimensional version of the same information.

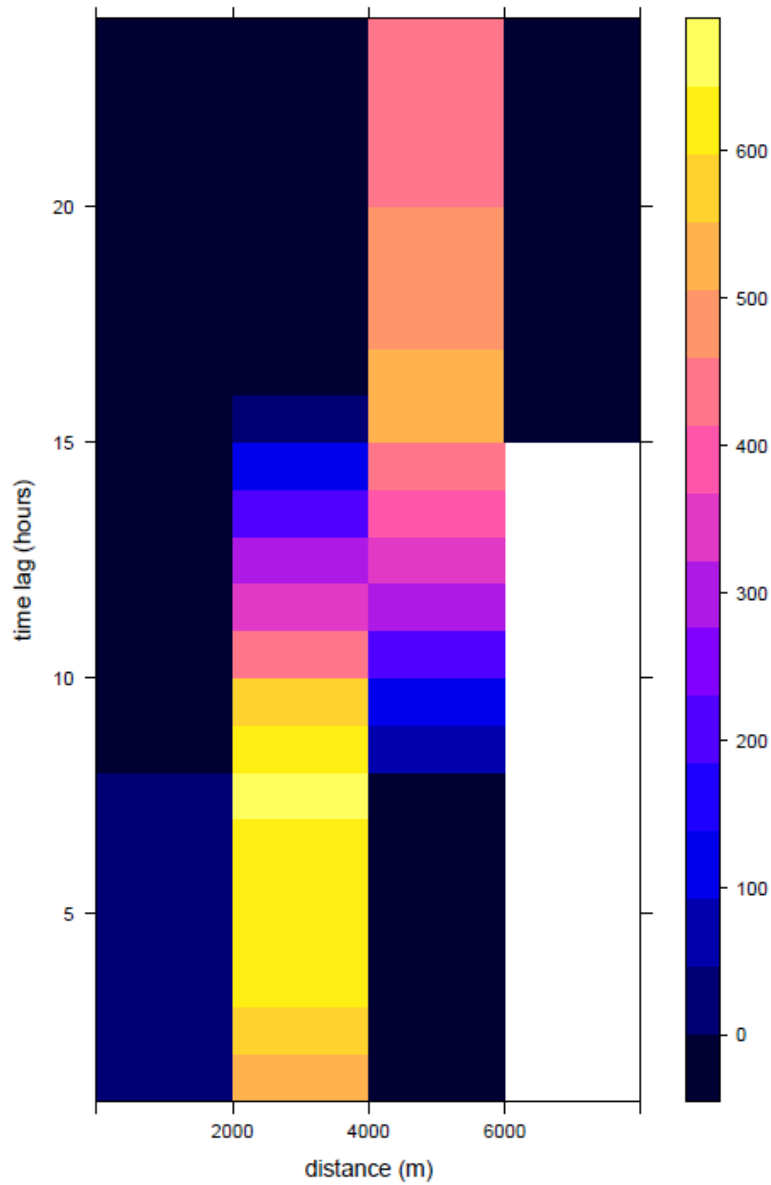
As illustrated in Figures 4.22 and 4.23, there are a significant number of grid cells without available data. By accumulating the existing grid cells on the distance (x-) axis into larger bins, the number of cells without missing data can be reduced. Figures 4.24 and 4.25, respectively, are the two-dimensional and three-dimensional spatially-binned versions of the data, using spatial bins of 2,000 meters. Until now, the auto-correlative temporal lags always showed a distinct U-shaped pattern with peaks at 24-hour intervals. In Figure 4.25, however, the now cross-correlative temporal lags no longer show this pattern, but are instead varying with the spatial lag value. The range is about 6,000 meters with a sill of about 400 to 500.



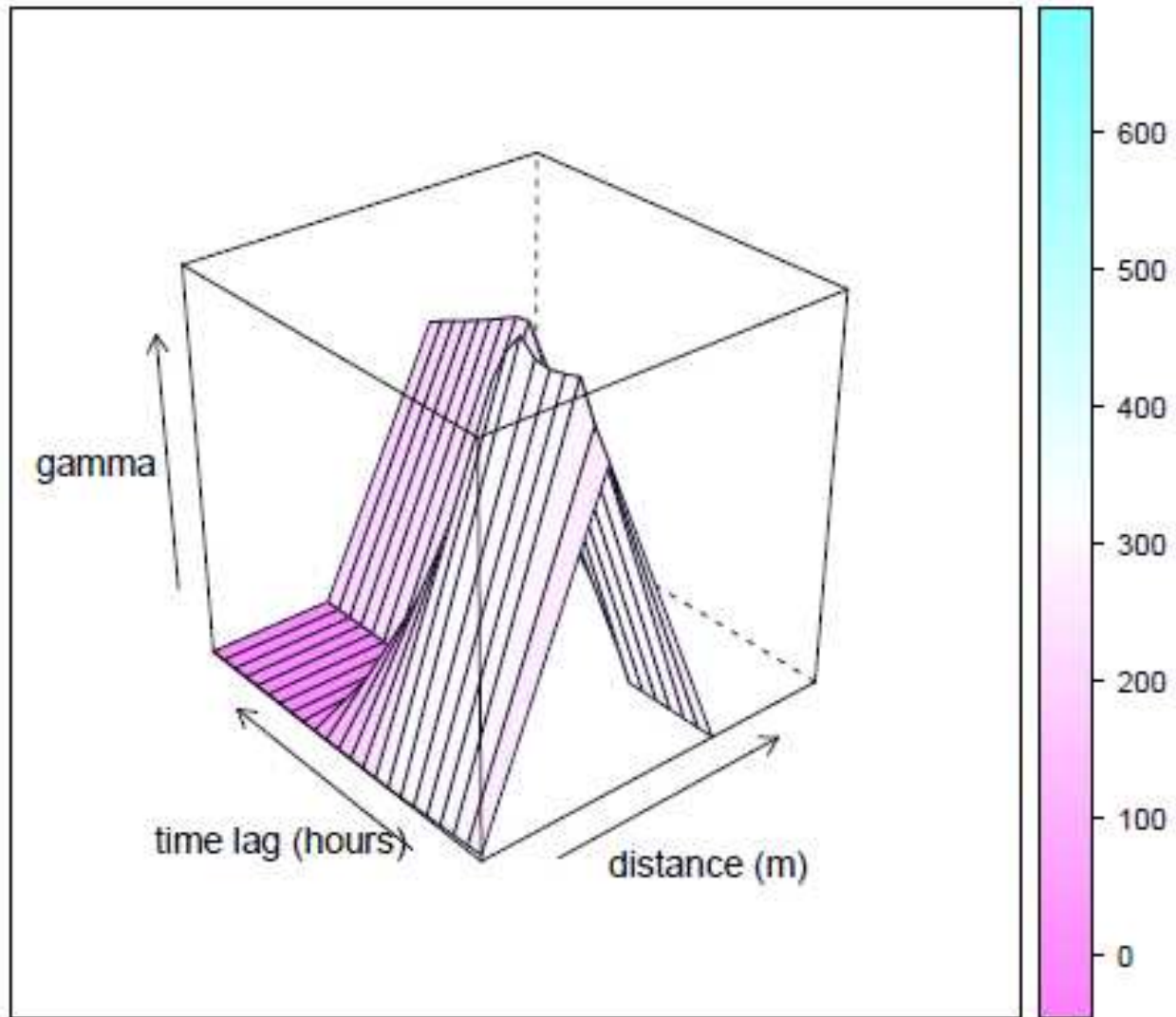
**Figure 4.22: Two-Dimensional Spatiotemporal Semivariogram for Residential Customers**



**Figure 4.23: Three-Dimensional Spatiotemporal Semivariogram for Residential Customers**



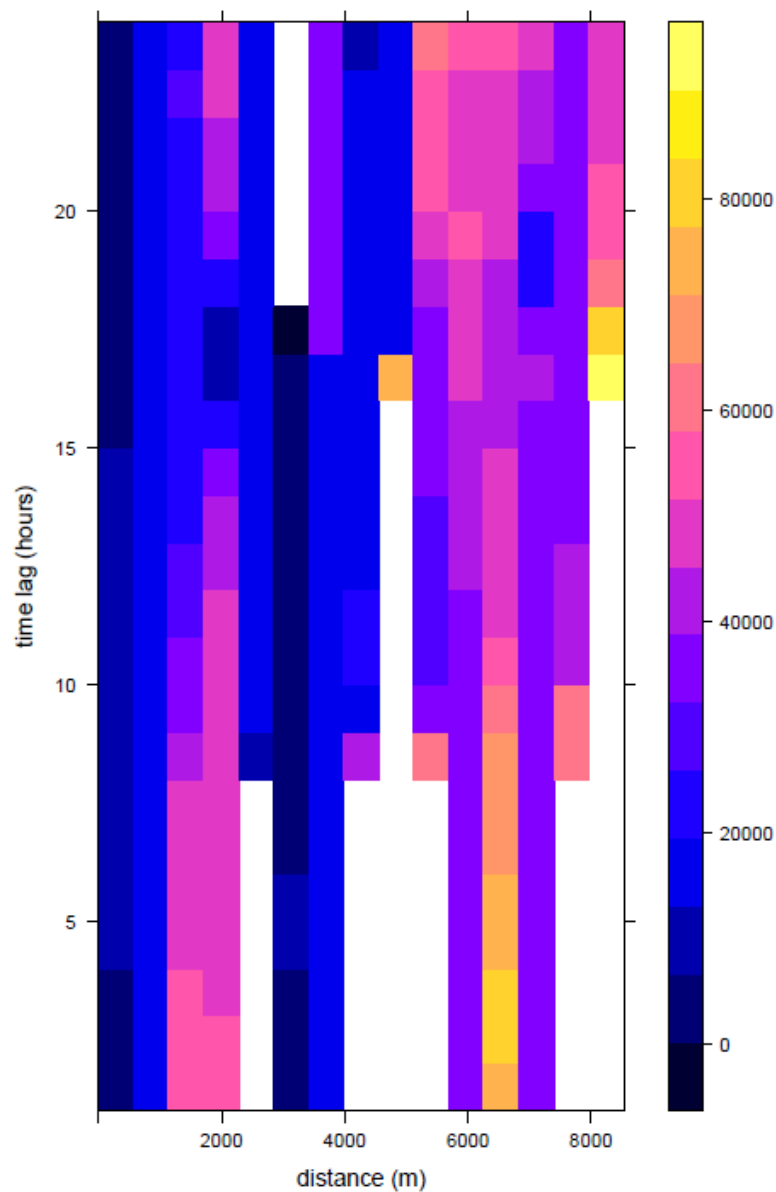
**Figure 4.24: Two-Dimensional Spatiotemporal Semivariogram for Residential Customers, With Boundaries**



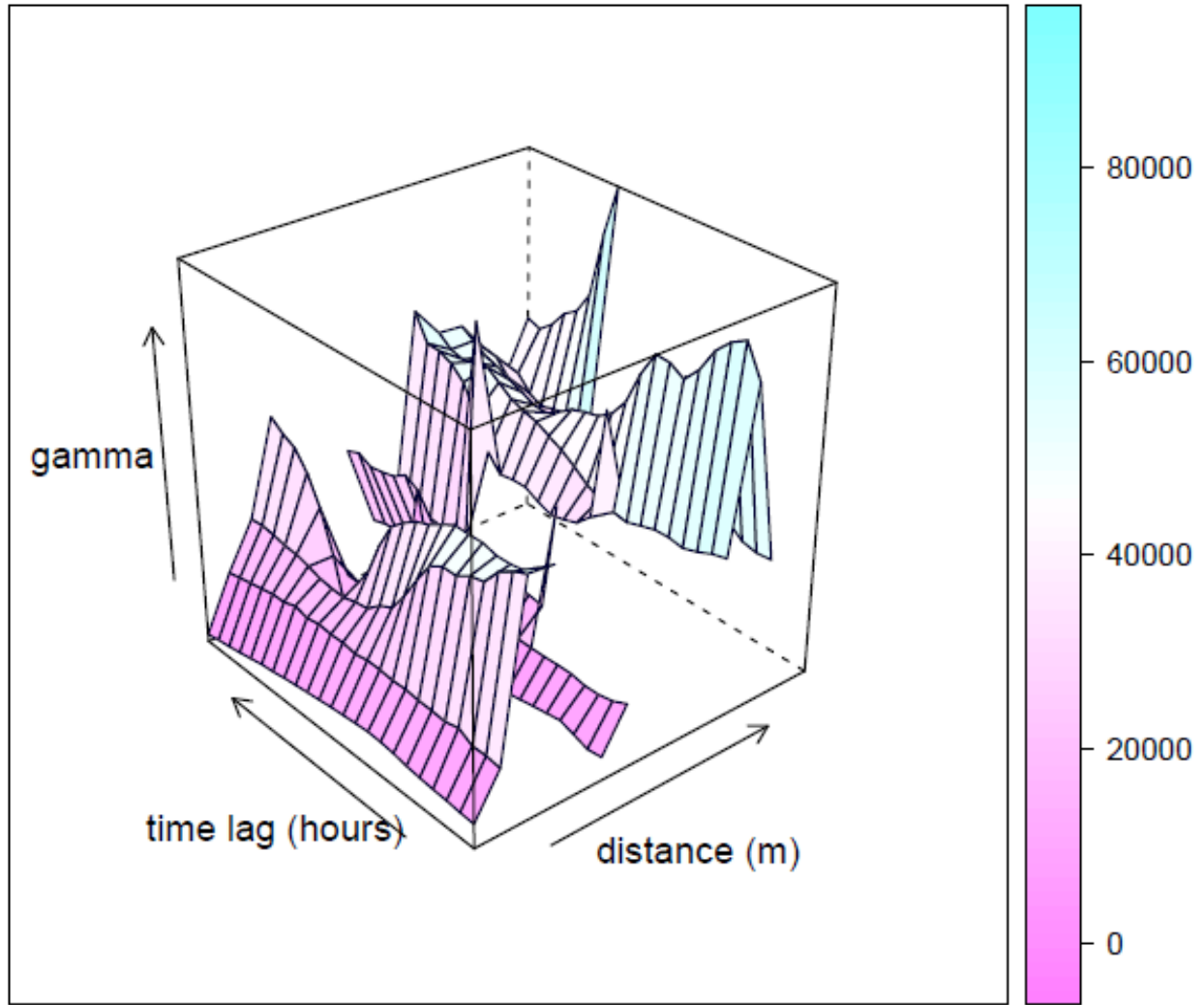
**Figure 4.25: Three-Dimensional Spatiotemporal Semivariogram for Residential Customers, With Boundaries**

For business customers, Figures 4.26 and 4.27, respectively, provide the two-dimensional and three-dimensional spatiotemporal semivariograms. As with the residential customers, the business semivariograms are based on 100 randomly selected customers for a single calendar month, and temporal lags up to 24 hours are included. Both the two-dimensional and three-dimensional spatiotemporal semivariograms for business customers contain cells with missing data. By binning the distance lags to 2,000

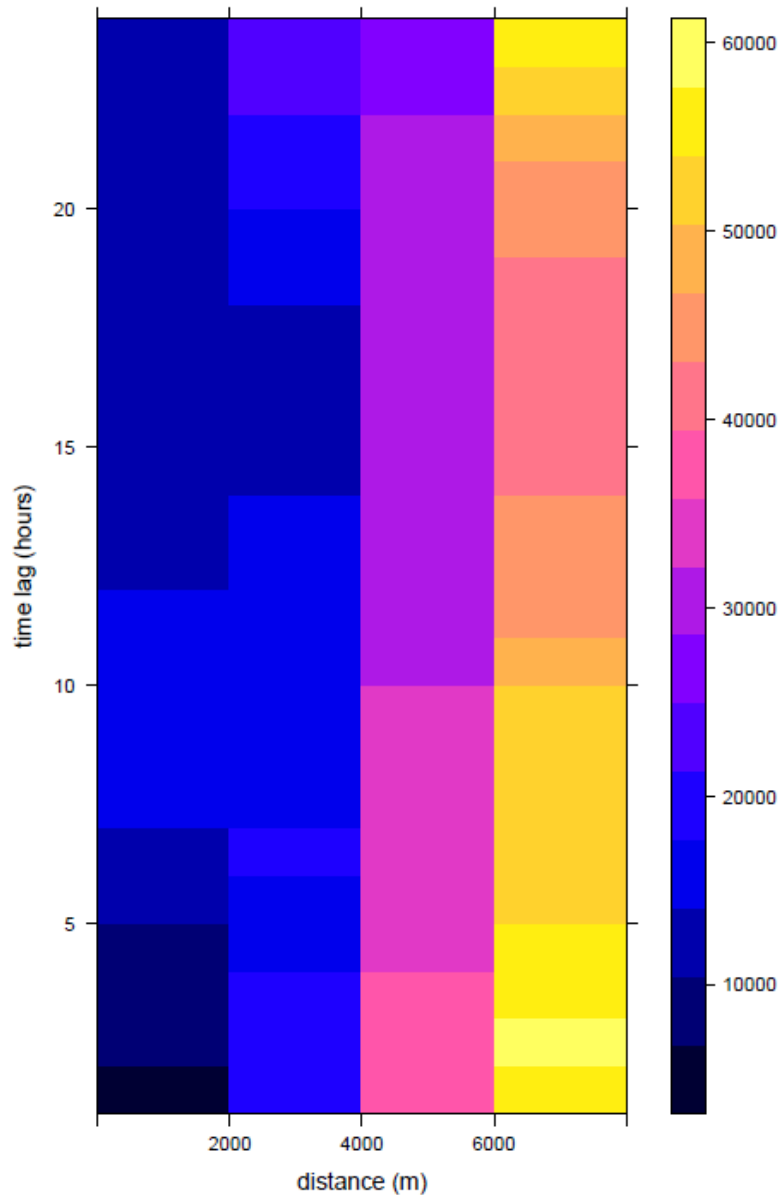
meter sections, all cells have values. These results are shown in Figures 4.28 and 4.29, respectively, for the two-dimensional and three-dimensional views. As with the residential customers, the previously-seen U-shaped temporal pattern no longer appears. Instead, the temporal pattern now varies with the values of the distance lags. The range for the business customers seems to exceed 8,000 meters, with a sill of perhaps 50,000 to 60,000.



**Figure 4.26: Two-Dimensional Spatiotemporal Semivariogram for Business Customers**

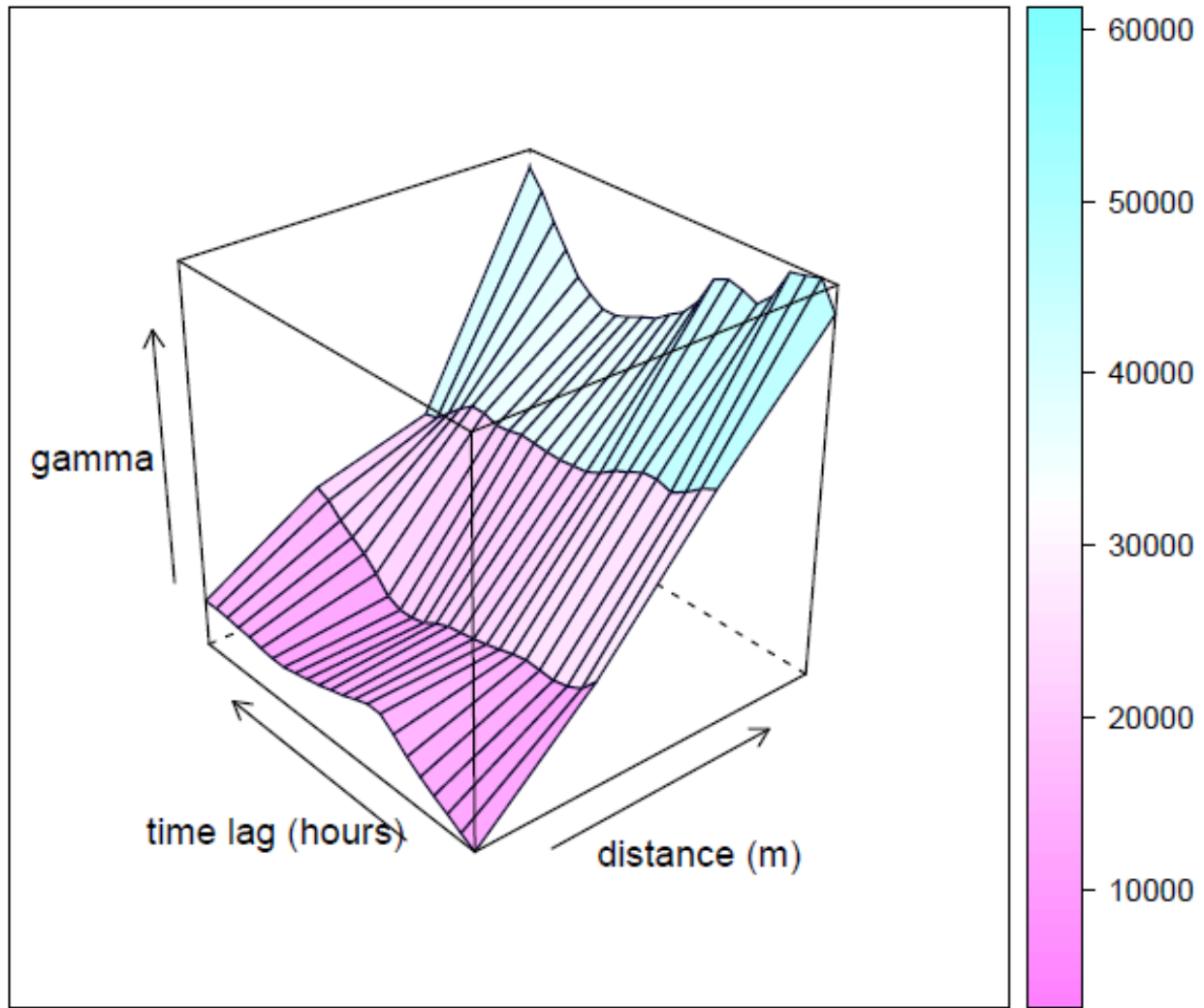


**Figure 4.27: Three-Dimensional Spatiotemporal Semivariogram for Business Customers**



**Figure 4.28: Two-Dimensional Spatiotemporal Semivariogram for Business Customers, With Boundaries**





**Figure 4.29: Three-Dimensional Spatiotemporal Semivariogram for Business Customers, With Boundaries**

#### 4.4.4 Spatiotemporal Exploratory Data Analysis Summary

The spatiotemporal data exploration looked at data using three techniques. Three-dimensional line graphs looked at two (for residential) and two (for business) randomly-chosen customers and the 20 nearest neighbors of each. One calendar year of temporal data was included. Although providing some general information about the presence of spatial patterns, they were not particularly useful in identifying spatial or temporal lags.

Three-dimensional correlograms looked at 25 or 50 randomly-chosen customers with temporal lags up to 672 hours (i.e., 28 days or four weeks). The correlograms made use of autocorrelation in the temporal lags but included temporal lags of different customers within the same spatially- and temporally-lagged cell. Results were inconsistent in terms of the spatial lag at which the correlations dropped off. Temporal lags continued the distinct U-shaped pattern first seen in Figures 4.13 and 4.14.

Finally, spatiotemporal semivariograms included cross-customer lags in both the spatial and temporal dimensions. The semivariograms used 100 randomly-chosen customers for a single calendar month, with temporal lags up to 24 hours. New temporal lag patterns resulted from the spatiotemporal semivariogram analysis.

#### **4.5 Summary of Exploratory Data Analysis**

A variety of exploratory techniques have been tested herein providing a variety of results, some of which are in conflict. Several tests indicated that none of the four transformations of the dependent variable of hourly interval energy usage data resulted in any improvement in the statistical results. Additionally, although the energy provider groups the data into statistical strata to reduce variation and thus minimize necessary sample sizes, the use of these strata provides no assistance in explaining associations between the independent and dependent variables. For these reasons, neither the data transformations nor stratification will be used in any of the gap-filling methods to be tested.

Although the results of the exploratory analysis are not always consistent, the analysis has offered some insight to the data structure and relationships. Temporal lags within a customer have a distinct U-shaped structure with peaks every 24 hours, and larger

peaks at one-week and four-week intervals. Temporal lags across customers have a different pattern that varies with the distance lag. Spatial lags for residential customers range from 0.2 meters (line graphs) to 12,500 meters (semivariograms), with the other methods somewhere in between. For business customers, spatial lags range from 300 meters (three-dimensional line graph) to 20,000 meters (spatiotemporal correlogram), with other methods falling in between. Despite the lack of evidence for a clearly-defined spatial lag, there is evidence that the spatial realm will be provide value in the gap-filling methods to be tested.

## 5. Drawing the Sample

Because each customer's hourly interval energy usage values are acting as their own control group, only customers having non-missing hourly interval energy usage data for the entire length of each gap have the potential to be useful in analyzing that gap length. So, for each gap length being analyzed, all customers were examined to identify all of the possible consecutive streaks of actual data of that length, thus providing a customer population for that gap length. From this "population" of customers a random number generator is used to select the specific customers and consecutive actual data streaks for which missing intervals are created. To ensure that alternative gap-filling methods are compared in an equitable manner, for each category of gap length all methods will be used to fill an identical set of gaps in the hourly interval energy usage data. This will be accomplished by using an identical random number seed for each gap length, while different gap lengths make use of a different random number seed.

Tables 5.1 and 5.2 show the lengths of the various gaps in hourly interval energy usage data that are analyzed in this study for the residential and business customer populations and samples, respectively. Appendix D lists the sampled customers and the specific hourly interval energy usage data gaps to which they are assigned.

**Table 5.1: Hourly Interval Energy Usage Data Gaps Used in the Research -- Residential**

Category	Category ID	Random Number Seed	Number of Hours to be Filled Per Customer	Population			Sample			
				Number of Customers	Mean of Intervals	Standard Deviation of Intervals	Number of Customers	Number of Hours	Mean of Intervals	Standard Deviation of Intervals
System Peak Hour	SPH	7095	1	859	3.16	22.71	10	10	2.09	1.66
Customer Peak Hour	CPH	27	1	910	8.25	46.65	10	10	3.72	2.09
1 Hour	1HR	9972	1	910	1.70	13.73	10	10	0.73	0.51
3 Hours	3HR	8934	3	910	1.73	14.10	10	30	0.38	0.21
12 Hours	12H	9562	12	910	1.66	14.25	10	120	1.07	1.33
Customer Peak Day	CPD	3638	24	910	4.13	27.96	10	240	2.07	1.05
24 Hours	24H	3292	24	910	1.76	14.04	10	240	0.50	0.50
7 Days	7DY	7751	168	910	1.70	13.73	10	1,680	0.46	0.41
1 Month	1MO	3735	720	903	1.70	13.74	10	7,200	14.84	31.65
3 Months	3MO	9687	2,160	127	9.91	36.23	10	21,600	6.60	20.98
6 Months	6MO	2319	4,344	75	14.77	43.95	10	43,392	11.96	46.22

**Table 5.2: Hourly Interval Energy Usage Data Gaps Used in the Research -- Business**

Category	Category ID	Random Number Seed	Number of Hours to be Filled Per Customer	Population			Sample			
				Number of Customers	Mean of Intervals	Standard Deviation of Intervals	Number of Customers	Number of Hours	Mean of Intervals	Standard Deviation of Intervals
System Peak Hour	SPH	7095	1	294	232.03	279.65	10	10	231.91	272.36
Customer Peak Hour	CPH	27	1	346	308.39	330.23	10	10	273.27	322.70
1 Hour	1HR	9972	1	346	165.23	210.37	10	10	176.59	234.31
3 Hours	3HR	8934	3	346	167.76	211.98	10	30	203.07	199.77
12 Hours	12H	9562	12	346	161.09	204.04	10	120	79.40	102.99
Customer Peak Day	CPD	3638	24	345	238.09	281.72	10	240	203.75	208.32
24 Hours	24H	3292	24	345	170.35	212.03	10	240	336.37	375.66
7 Days	7DY	7751	168	338	166.29	210.75	10	1,680	200.84	205.54
1 Month	1MO	3735	720	328	167.72	211.19	10	7,200	29.49	60.22
3 Months	3MO	9687	2,160	192	259.05	216.94	10	21,576	233.24	175.54
6 Months	6MO	2319	4,344	173	266.20	218.61	10	43,368	316.79	323.98

## 6. Results of the Analysis

### 6.1 Discussion of Gap Filling Methods

Six gap-filling methods are tested using each of eleven gap lengths. For each customer type (residential and business), eleven different data sets are used in the analysis, each being associated with one particular gap length. For the three methods that have been previously used for Load Research purposes (KEMA, dummy variable, and neural network), the independent variables were chosen to as closely as possible replicate the variables used in prior Load Research analyses. For the three methods that are new to Load Research (GWR, spatial regression, and spatiotemporal), variables were chosen that provided the greatest likelihood of success for gap-filling. Each of the six gap-filling methods is briefly reviewed below, and the specific independent variables used are listed.

The KEMA method, commonly used by load researchers, is a temporal method in which the dependent variable of hourly interval energy usage data is estimated via a linear regression with cooling and heating degree hourly values as the sole independent variables. Separate regressions are made for each combination of customer, day type, and hour with the input data being limited to the same customer, day-of-the-week, and hour-of-

the-day data values as that of the target interval being filled. This method is provided as a baseline against which other methods can be compared. The KEMA method was run using the SAS PROC REG function.

The dummy variable method, proposed by Smith and Hanna (2008) is also a temporal method. A separate regression is estimated for each customer, but all hours and days for that customer are handled in a single linear regression equation via a series of hourly, daily, and monthly dummy variables. Similar to the KEMA method, both heating and cooling degree days for all days and hours are included as independent variables. The dummy variable method was run using the SAS PROC REG function.

The neural network method, as used by McMenamin and Monforte (1997), is also a temporal method. The independent variables proposed by McMenamin and Monforte include weather variables, calendar variables, lagged loads, and interactions between all of those categories. Although lagged hourly interval energy usage values are included by McMenamin and Monforte, for the purposes of gap-filling there is no guarantee that the hourly interval for any particular lag length will be available. Therefore, lagged interval values are excluded as independent variables herein. As a substitute, billed monthly energy (and demand for business customers) variables were included in this analysis. The independent variables are the following:

- For the hour of the missing hourly interval energy usage value -- hour, dry bulb temperature, humidity, wind speed, cloud cover, and sunshine minutes.
- For the day of the missing hourly interval energy usage value -- day type, maximum and minimum hourly dry bulb temperature during the day, cumulative dry bulb temperature for the day up to and including the hour of the missing value, and the



temperature gradient as measured by the percent change from the 1am dry bulb temperature for the day to the dry bulb temperature at the hour of the missing value.

- For the month of the missing hourly interval energy usage value -- month, billed monthly energy, and billed monthly demand (business customers only).
- Interaction variables at the hour of the missing value -- dry bulb temperature and dry bulb temperature squared interacted with day type and month, and dry bulb temperature interacted with each of the following: humidity, wind speed, cloud cover, and sunshine minutes.

The neural network method was run using the R nnet package.

Both the GWR and spatial regression methods are spatial methods that examine all customers at once but only for a single hour at a time. Because they are purely spatial, the potential set of independent variables is limited to those that vary across space. Therefore, all weather variables and daytype are excluded by definition. Potential variables for inclusion are geodemographic and building-related variables, annual and monthly billed energy, and (for business customers only) billed demand. Using all available data, a backward elimination linear regression model was implemented for each of the 8,760 hours in the year. Each regression starts with all independent variables. At each step, the variable making the smallest contribution to the model is eliminated one by one until all variables that remain have an F statistic significant at the 0.10 level (SAS Institute 2009, p. 5524). The results of the regression equations were examined and the variables that remained in the most equations (i.e., that were significant for the most hours in the year)

were selected for inclusion in the GWR and spatial regression models. For residential and business customers, variables had to appear in at least 94 percent and 67 percent of the hourly equations, respectively. The variables chosen for inclusion in this way are the following:

- For residential customers -- annual kWh usage, monthly consumption value associated with the hour of interest, percent of the census tract population that is black, and percent of census tract householders who have less than a high school education<sup>14</sup>.
- For business customers -- annual kWh usage, monthly consumption value associated with the hour of interest, monthly demand value associated with the hour of interest, building area per unit, and median age for the census tract.

The GWR method was run using the R GWmodel package, and the spatial regression method was run using the R spdep package.

The spatiotemporal method allows both spatial and temporal lags to be modeled simultaneously, or at least in conjunction with one another. The independent variables selected for inclusion were those used in the spatial methods, plus additional temporal values. The variables included in the spatiotemporal analysis are the following:

- For residential customers -- annual kWh usage, monthly consumption value associated with the hour of interest, percent of the census tract population that is

---

<sup>14</sup> An additional variable, percent of the census tract householders who have at least some college, was originally included in the residential equations for GWR and spatial regression but was removed from the final equations to solve a recurring singular matrix problem.

black, percent of census tract householders who have less than a high school education, and hourly dry bulb temperature for the hour of interest.<sup>15</sup>

- For business customers -- annual kWh usage, monthly consumption value associated with the hour of interest, monthly demand value associated with the hour of interest, building area per unit, median age for the census tract, and hourly dry bulb temperature for the hour of interest.

The spatiotemporal method was run using the R spTimer package<sup>16</sup>.

## 6.2 Discussion of Evaluation Statistics

In order to evaluate the different gap-filling methods, eight different evaluation statistics are calculated, each of which is targeted at one of the twin evaluation measures of accuracy and bias. Overall accuracy is measured by the root mean square error (RMSE) and mean absolute percentage error (MAPE) statistics. Overall bias is measured by the average error statistic. Accuracy and bias for the largest and smallest hourly interval energy usage values are measured by taking the difference between the actual and estimated values for those hours, both in absolute terms and as a percent.

The goal of this research is to identify actual data, set it to missing, use the various methods to fill these newly-created gaps the data, and then compare the newly-filled value to the actual value. For the three load research methods, separate data bases were created

---

<sup>15</sup> For residential customers, no spatiotemporal results were generated. Repeated tinkering with the number of iterations either resulted in no predictions (perhaps because no stable results were achieved) or memory errors because the number of predictions was too large.

<sup>16</sup> The original plan was to use the R SpatioTemporal package. Although a small test dataset worked perfectly, numerous tests with larger datasets repeatedly failed with singular matrix errors. As a second choice, the spatiotemporal results were analyzed in spTimer, which uses Bayes logic in a spatiotemporal implementation.

for each of the 11 gap-lengths that correctly had gaps for the intervals that are to be filled. The spatial regression and GWR methods, however, each filling one hour at a time, took approximately 24 hours to fill one month's worth of data (approximately 720 hours of data), or approximately 12 days to fill a year's worth of data. Had separate datasets been created for each of the 11 gap-filling methods, this would have required approximately 130 days to run all of the necessary data calculations. Therefore, due to time constraints, a decision was made to use a single input dataset (the system peak hour dataset) for all spatial regression and GWR gap-filling. This decision provides the spatial regression and GWR methods with an advantage over the other methods, because the actual data values for customer-hours being filled are included as inputs to the algorithms that determine how to fill those hours.

For any particular hour analyzed, the following are the maximum number of customers that had actual data for spatial regression and GWR that "should" have been a data gap but was not (out of 911 residential customers and 364 business customers):

- System peak hour -- no residential or business customers.
- Customer peak hour -- no residential customers, 6 business customers.
- One hour -- no residential or business customers.
- Three hours -- 2 residential customers, no business customers.
- 12 hours -- no residential or business customers.
- Customer peak day -- 4 residential customers, 2 business customers.
- 24 hours -- no residential or business customers.
- 7 days -- 2 residential customers, no business customers.
- One month -- 5 residential customers, 6 business customers.

- 3 months -- 6 residential customers, 4 business customers.
- 6 months -- 7 residential customers, 9 business customers.

Additionally, as shall be seen, the spatial regression and GWR methods resulted in singular matrices for some of the hours analyzed. The result of this is that the number of gap-filled hours is smaller than for the three Load Research methods. These occurrences are noted in the sections below.

The spatiotemporal method is not shown to its full advantage because memory errors prevented a full year's worth of data from being analyzed in a single run. Therefore, monthly datasets were also used in the spatiotemporal analysis, thus limiting the availability of temporal input data with which to implement the spatiotemporal models. As with the spatial regression and GWR methods, not all of the hours were gap-filled. The reason for these omissions is not entirely clear, but based on a review of input and output data files, it appears that a combination of missing data and singular matrix problems may be the culprits. The R spTimer package requires a complete rectangular dataset, in the sense that all input independent variable values must be available for all timeframes and for all observations. The only values allowed to be missing are those of the dependent variable. Other packages used are less strict with regard to missing data. Therefore any data issues that may have been forgiven by other packages were rejected by spTimer.

### **6.3 Gap Filling for the System Peak Hour**

The hour of the energy provider's system peak is the single hour of the year during which the energy provider's customers use more energy than at any other time. It is a critical time for the energy provider due to the possible stress put on the energy supply and delivery systems. The length of the gap is one hour, and any customer having known actual

usage for that hour is a member of the population for this gap length. From that population, a sample of ten customers is selected for the evaluation of the gap filling methods.

Tables 6.1 and 6.2 provide the results for the residential and business customers, respectively, of gap-filling methods for the system peak hour. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for business customers, the spatiotemporal method results in a smaller sample size than other methods.

**Table 6.1: Analysis Results for System Peak Hour -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	10	0.88	124%	0.40	5.00	4.70	0.30	6%	0.06	0.11	-0.04	-42%
Dummy Variable	10	0.92	58%	0.50	4.21	4.70	-0.49	-10%	0.57	0.11	0.47	445%
Neural Network	10	1.54	152%	1.05	2.34	4.70	-2.36	-50%	0.29	0.11	0.18	172%
GWR	10	1.08	81%	0.48	3.72	4.70	-0.97	-21%	0.48	0.11	0.36	356%
Spatial Regression	10	1.69	239%	0.68	3.48	4.70	-1.22	-27%	-0.13	0.11	-0.24	-227%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.2: Analysis Results for System Peak Hour -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	10	355.15	34%	-171.44	1,132.50	776.16	356.34	46%	6.73	4.15	2.58	62%
Dummy Variable	10	297.13	25%	-116.49	984.85	776.16	208.69	27%	5.22	4.15	1.07	26%
Neural Network	10	301.45	66%	5.38	853.63	776.16	77.47	10%	4.50	4.15	0.35	8%
GWR	10	228.78	47%	-69.51	697.38	776.16	-78.78	-10%	16.70	4.15	12.55	302%
Spatial Regression	10	281.32	33%	-83.67	895.08	776.16	118.92	15%	4.24	4.15	0.09	2%
Spatio-Temporal	9	1,295.87	138%	-307.41	2,801.85	776.16	2,025.69	261%	-1,103.85	4.15	-1,108.00	-26,699%

## 6.4 Gap Filling for Customer Peak Hour

Although not critical to the energy provider, the maximum value of each individual customer's hourly interval energy usage value can be a critical value for the customer. For some customers, their peak value could determine a significant portion of their billing charges for the next 18 months. The length of the gap is one hour, and all customers have a maximum value. From the population of customers, a sample of ten customers is selected for the evaluation of the gap filling methods.

For the customer peak hour, Tables 6.3 and 6.4 provide the results for the residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for business customers, the spatial regression and spatiotemporal methods result in a smaller sample size than other methods.



**Table 6.3: Analysis Results for Customer Peak Hour -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	10	3.32	421%	2.87	1.86	7.65	-5.79	-76%	0.23	0.69	-0.45	-66%
Dummy Variable	10	3.58	524%	3.11	1.36	7.65	-6.29	-82%	0.18	0.69	-0.51	-74%
Neural Network	10	3.71	684%	3.23	0.99	7.65	-6.66	-87%	0.17	0.69	-0.52	-76%
GWR	10	3.48	537%	3.05	1.33	7.65	-6.32	-83%	-0.07	0.69	-0.61	-90%
Spatial Regression	10	3.20	291%	2.42	4.50	7.65	-3.15	-41%	-0.65	0.69	-1.33	-194%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.4: Analysis Results for Customer Peak Hour -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	10	25.11	32%	2.90	723.10	720.12	2.98	0%	2.08	5.50	-3.41	-62%
Dummy Variable	10	73.19	60%	40.87	599.48	720.12	-120.64	-17%	1.53	5.50	-3.97	-72%
Neural Network	10	191.73	156%	110.43	518.40	720.12	-201.72	-28%	0.98	5.50	-4.51	-82%
GWR	10	52.73	22%	32.04	639.19	720.12	-80.93	-11%	6.43	5.50	0.94	17%
Spatial Regression	9	60.83	40%	35.94	579.51	720.12	-140.61	-20%	7.98	5.50	2.49	45%
Spatio-Temporal	8	2,672.64	104%	1,004.51	1,558.68	720.12	838.56	116%	-5,739.10	5.50	-5,744.59	-104,523%

## 6.5 Gap Filling for 1 Hour

Unlike the first two gap lengths, the one hour length simply chooses a random hour during the year. All customers have hourly interval energy usage values that are part of the population. A sample of ten customer-hour combinations is selected for the evaluation of the gap filling methods.

For the one hour gap, Tables 6.5 and 6.6 provide the results for the residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for business customers, the spatiotemporal method results in a smaller sample size than other methods.

**Table 6.5: Analysis Results for One Hour -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	10	0.51	92%	0.23	1.14	1.51	-0.37	-25%	0.14	0.03	0.11	380%
Dummy Variable	10	0.40	59%	0.13	1.51	1.51	-0.01	-1%	-0.16	0.03	-0.19	-626%
Neural Network	10	0.44	65%	0.15	1.31	1.51	-0.20	-13%	0.05	0.03	0.02	65%
GWR	10	0.45	67%	0.03	1.71	1.51	0.20	14%	0.15	0.03	0.12	405%
Spatial Regression	10	0.73	17,721%	0.18	1.77	1.51	0.26	17%	-0.50	0.03	-0.53	-1,776%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.6: Analysis Results for One Hour -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	10	20.07	16%	6.23	532.81	552.96	-20.15	-4%	1.80	1.11	0.69	62%
Dummy Variable	10	37.47	28%	12.35	548.38	552.96	-4.58	-1%	-7.39	1.11	-8.49	-767%
Neural Network	10	141.62	75%	42.93	386.30	552.96	-166.67	-30%	0.98	1.11	-0.12	-11%
GWR	10	79.19	117%	21.79	518.36	552.96	-34.60	-6%	-0.37	1.11	-1.47	-133%
Spatial Regression	10	78.82	65%	22.75	514.29	552.96	-38.67	-7%	-2.73	1.11	-3.84	-347%
Spatio-Temporal	9	622.00	128%	-354.64	1,640.32	552.96	1,087.36	197%	-141.16	1.11	-142.27	-12,852%

## 6.6 Gap Filling for 3 Hours

The three-hour gap length uses as its population any set of three consecutive hours. All customers having three consecutive hours of actual data are identified, along with all of the customers' three-hour-long stretches of actual data. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. Each gap is three hours in length, so that three hours are sampled for each of ten customers for a total of 30 sampled hours.

Tables 6.7 provides results for the residential customers and Table 6.8 provides the results for the business customers. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for residential customers, the GWR and spatial regression methods result in a smaller sample size than other methods. For business customers, the spatiotemporal method results in a smaller sample size than other methods.

**Table 6.7: Analysis Results for Three Hours -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	30	0.18	47%	-0.01	0.83	1.06	-0.23	-22%	0.07	0.04	0.03	93%
Dummy Variable	30	0.23	50%	-0.01	0.86	1.06	-0.20	-19%	0.17	0.04	0.13	364%
Neural Network	30	0.22	41%	-0.06	0.69	1.06	-0.36	-34%	0.27	0.04	0.23	651%
GWR	27	0.17	36%	-0.03	0.76	1.06	-0.29	-28%	0.10	0.04	0.06	180%
Spatial Regression	27	0.75	195%	-0.28	3.00	1.06	1.95	184%	-0.06	0.04	-0.09	-255%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.8: Analysis Results for Three Hours -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	30	34.28	19%	21.99	558.80	632.16	-73.36	-12%	5.72	7.41	-1.68	-23%
Dummy Variable	30	25.05	14%	9.78	577.07	632.16	-55.09	-9%	9.71	7.41	2.31	31%
Neural Network	30	42.37	32%	15.97	548.29	632.16	-83.87	-13%	5.73	7.41	-1.68	-23%
GWR	30	57.02	30%	4.64	552.47	632.16	-79.70	-13%	7.02	7.41	-0.39	-5%
Spatial Regression	30	62.81	2,823%	8.81	516.65	632.16	-115.51	-18%	-0.26	7.41	-7.66	-103%
Spatio-Temporal	27	1,322.06	127%	470.41	1,804.45	632.16	1,172.29	185%	-3,701.40	13.79	-3,715.19	-26,945%

## 6.7 Gap Filling for 12 Hours

Similarly to the three hour gap, the twelve hour gap length is any set of 12 consecutive hours occurring during the year. All customers having 12 consecutive hours of actual data are identified, along with all of the customers' 12-hour-long stretches of actual data. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. Each gap is 12 hours in length, so the 10 sampled customers have a total of 36 sampled hours.

Tables 6.9 and 6.10 provide results for the residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for business customers, the GWR, spatial regression, and spatiotemporal methods result in a smaller sample size than other methods.

**Table 6.9: Analysis Results for Twelve Hours -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	120	0.80	75%	0.05	5.16	7.39	-2.23	-30%	0.09	0.05	0.04	93%
Dummy Variable	120	0.85	44%	-0.18	5.98	7.39	-1.41	-19%	0.07	0.05	0.02	51%
Neural Network	120	0.88	61%	-0.04	5.79	7.39	-1.60	-22%	0.15	0.05	0.10	214%
GWR	120	0.65	69%	0.18	5.05	7.39	-2.34	-32%	-0.38	0.05	-0.43	-887%
Spatial Regression	120	0.96	154%	0.12	4.30	7.39	-3.09	-42%	-1.20	0.05	-1.24	-2,591%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.10: Analysis Results for Twelve Hours -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	120	16.96	35%	6.13	301.40	321.12	-19.72	-6%	0.14	0.08	0.06	81%
Dummy Variable	120	23.67	38%	8.00	297.17	321.12	-23.95	-7%	0.03	0.08	-0.05	-63%
Neural Network	120	27.82	49%	10.10	267.82	321.12	-53.30	-17%	0.49	0.08	0.41	535%
GWR	108	19.70	117%	8.31	284.07	321.12	-37.05	-12%	-10.97	0.08	-11.04	-14,434%
Spatial Regression	108	19.24	87%	10.37	284.09	321.12	-37.03	-12%	-8.11	0.08	-8.19	-10,703%
Spatio-Temporal	96	1,075.71	128%	61.17	2,222.84	321.12	1,901.72	592%	-4,565.86	0.08	-4,565.93	-5,968,540%

## 6.8 Gap Filling for Customer Peak Day

Similar to the customer peak hour, the customer peak day examines the calendar day for each customer when their total usage during that day (i.e., the sum of the hourly interval energy usage values for the day) is maximized. Each customer has a maximum value, and ten customers are sampled. Because each sampled customer has 24 hours, a total of 240 hours are sampled.

For the customer peak hour, Tables 6.11 and 6.12 provide the results for the residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for business customers, the GWR, spatial regression, and spatiotemporal methods results in a smaller sample size than other methods.



**Table 6.11: Analysis Results for Customer Peak Day -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	240	1.45	213%	1.11	2.85	5.19	-2.34	-45%	0.12	0.15	-0.04	-23%
Dummy Variable	240	1.38	123%	1.04	2.53	5.19	-2.66	-51%	0.32	0.15	0.17	111%
Neural Network	240	1.64	254%	1.30	3.35	5.19	-1.84	-36%	0.10	0.15	-0.06	-37%
GWR	240	1.46	149%	1.10	3.32	5.19	-1.87	-36%	0.29	0.15	0.14	90%
Spatial Regression	240	1.80	160%	0.68	5.89	5.19	0.70	13%	-3.30	0.15	-3.45	-2,255%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.12: Analysis Results for Customer Peak Day -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	240	51.33	49%	23.11	565.34	674.28	-108.94	-16%	0.75	0.55	0.20	36%
Dummy Variable	240	40.61	27%	21.23	550.77	674.28	-123.51	-18%	1.67	0.55	1.12	203%
Neural Network	240	108.38	163%	64.41	385.71	674.28	-288.57	-43%	-0.20	0.55	-0.75	-137%
GWR	216	42.85	82%	21.60	602.45	674.28	-71.83	-11%	0.04	0.55	-0.51	-93%
Spatial Regression	216	48.50	184%	25.43	577.12	674.28	-97.16	-14%	-4.52	0.55	-5.07	-919%
Spatio-Temporal	144	2,053.81	122%	-5.15	4,337.38	674.28	3,663.10	543%	-5,787.61	0.55	-5,788.16	-1,048,580%

## 6.9 Gap Filling for 24 Hours

The population for the 24 hour gap is any calendar day with actual values for all 24 consecutive hourly interval energy usage values. All customers having a complete day of actual data are identified, along with all of the customers' actual-data-days. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. A total of 240 hours are sampled in this way.

Table 6.13 contains the results for the residential customers, and Table 6.14 shows the business customer results. For each of the evaluation methods, the "best" result is highlighted in green.

**Table 6.13: Analysis Results for Twenty-Four Hours -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	240	0.36	45%	0.04	1.77	3.49	-1.72	-49%	0.06	0.06	-0.00	-1%
Dummy Variable	240	0.32	41%	0.03	1.28	3.49	-2.20	-63%	0.09	0.06	0.03	42%
Neural Network	240	0.42	54%	-0.02	3.49	3.49	-2.57	-74%	0.09	0.06	0.03	50%
GWR	240	0.42	88%	-0.06	1.74	3.49	-1.75	-50%	-0.06	0.06	-0.12	-205%
Spatial Regression	240	0.75	86%	-0.23	3.16	3.49	-0.32	-9%	-1.14	0.06	-1.20	-2,008%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.14: Analysis Results for Twenty-Four Hours -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	240	33.78	12%	3.31	952.89	1,014.00	-61.11	-6%	1.13	1.32	-0.18	-14%
Dummy Variable	240	34.05	59%	3.02	1,034.58	1,014.00	20.58	2%	-8.50	1.32	-9.82	-744%
Neural Network	240	86.12	41%	34.89	853.33	1,014.00	-160.67	-16%	0.93	1.32	-0.39	-29%
GWR	240	47.36	44%	-2.01	1,000.64	1,014.00	-13.37	-1%	-7.60	1.32	-8.92	-676%
Spatial Regression	240	54.43	46%	-3.92	1,044.80	1,014.00	30.80	3%	-5.85	1.32	-7.17	-544%
Spatio-Temporal	240	2,541.80	122%	345.32	6,672.83	1,014.00	5,658.83	558%	-7,877.36	1.32	-7,878.68	-597,549%

## 6.10 Gap Filling for 7 Days

The seven day gap is any period of seven consecutive calendar days (168 hours) with actual hourly interval energy usage data in all 168 hours. All customers having seven complete days of actual data are identified, along with all of the customers' actual-seven-data-day-stretches. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. A total of 1,680 hours are sampled.

Tables 6.15 and 6.16 include results for the residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for business customers, the spatiotemporal method results in a smaller sample size than other methods.

**Table 6.15: Analysis Results for Seven Days -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	1,680	0.33	49%	-0.01	1.88	3.19	-1.30	-41%	0.07	0.00	0.07	224%
Dummy Variable	1,680	0.35	47%	-0.02	1.17	3.19	-2.02	-63%	0.07	0.00	0.07	217%
Neural Network	1,680	0.42	63%	0.00	1.39	3.19	-1.80	-57%	0.06	0.00	0.06	199%
GWR	1,680	0.51	89%	-0.08	3.42	3.19	0.23	7%	-3.01	0.00	-3.01	-100,465%
Spatial Regression	1,680	1.23	140%	-0.16	8.63	3.19	5.44	171%	-5.25	0.00	-5.26	-175,209%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.16: Analysis Results for Seven Days -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	1,680	83.04	20%	11.60	652.02	775.92	-123.90	-16%	0.81	0.71	0.11	15%
Dummy Variable	1,680	86.24	54%	11.23	615.98	775.92	-159.95	-21%	-8.49	0.71	-9.20	-130%
Neural Network	1,680	119.49	49%	31.03	415.91	775.92	-360.01	-46%	-0.70	0.71	-1.41	-200%
GWR	1,680	70.94	122%	17.95	641.07	775.92	-134.85	-17%	-14.05	0.71	-14.75	-2,093%
Spatial Regression	1,680	78.53	136%	17.27	613.06	775.92	-162.86	-21%	-10.15	0.71	-10.86	-1,540%
Spatio-Temporal	1,272	1,135.07	221%	48.87	4,922.88	775.92	4,146.96	534%	-4,745.75	0.71	-4,746.46	-673,256%

### 6.11 Gap Filling for 1 Month

The one month gap is any period of 30 consecutive calendar days (720 hours) with actual values for all 720 hourly interval energy usage periods. All customers having 30 complete days of actual data are identified, along with all of the customers' actual-streaks of 30-data-days. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. A total of 7,200 hours is sampled in this way.

Tables 6.17 and 6.18 contain the gap-filling results for residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for residential customers, the GWR and spatial regression methods result in a smaller sample size than other methods. For business customers, the spatial regression and spatiotemporal methods result in a smaller sample size.

**Table 6.17: Analysis Results for One Month -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	7,200	8.52	43%	-0.21	177.87	174.60	3.27	2%	-0.29	0.00	-0.29	~-29%
Dummy Variable	7,200	8.90	41%	-0.62	128.39	174.60	-46.21	-26%	0.00	0.00	0.00	~0%
Neural Network	7,200	15.07	52%	2.70	99.41	174.60	-75.19	-43%	-0.21	0.00	-0.21	~-21%
GWR	7,198	7.47	129%	1.12	171.04	174.60	-3.56	-2%	-0.77	0.06	-0.84	-1,326%
Spatial Regression	7,198	12.73	304%	1.66	239.90	174.60	65.30	37%	-3.42	0.06	-3.48	-5,524%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.18: Analysis Results for One Month -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	7,200	4.42	26%	0.66	224.36	228.48	-4.12	2%	-4.60	0.00	-4.60	~460%
Dummy Variable	7,200	4.70	69%	-0.58	223.66	228.48	-4.82	-2%	-7.60	0.00	-7.60	~-760%
Neural Network	7,200	6.93	64%	-0.68	213.99	228.48	-14.49	-6%	-4.73	0.00	-4.73	~-473%
GWR	7,200	9.06	134%	1.01	231.99	228.48	3.51	2%	-40.97	0.00	-40.97	~-4,097%
Spatial Regression	6,479	12.10	254%	1.30	45.06	44.54	0.51	1%	-43.18	0.00	-43.18	-959,643%
Spatio-Temporal	6,216	777.85	114%	-68.09	3,467.68	228.48	3,239.20	1,418%	-3,326.52	0.00	-3,326.52	-73,922,656%

## 6.12 Gap Filling for 3 Months

For the three month gap, 90 calendar days (2,160 hours) of actual values of consecutive hourly interval energy usage data must be available. For the three month gap, however, the consecutive period can "wrap" around the end of the year (December 31) to the first of the same calendar year (January 1). All customers having 90 days of actual data are identified, along with all of the customers' actual-90-data-day-streaks. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. A total of 21,600 hours are sampled in this way.

Tables 6.19 and 6.20, respectively, contain the results for the residential and business customers. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for residential and business customers, the GWR and spatial regression methods result in a smaller sample size than other methods, as does the spatiotemporal method for business customers.



**Table 6.19: Analysis Results for Three Months -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	24,792	10.16	38%	-0.96	399.70	402.96	-3.26	-1%	-8.43	0.00	-8.43	~-843%
Dummy Variable	24,792	14.06	243%	-1.64	173.93	402.96	-229.03	-57%	-11.86	0.00	-11.86	~-1,186%
Neural Network	24,792	17.26	828%	-1.71	71.40	402.96	-331.56	-82%	-0.00	0.00	-0.00	~-0%
GWR	19,438	3.62	216%	0.08	61.70	108.72	-47.02	-43%	-3.81	0.00	-3.81	~-381%
Spatial Regression	19,438	4.55	191%	0.18	54.82	108.72	-53.90	-50%	-15.09	0.00	-15.09	~-1,509%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.20: Analysis Results for Three Months -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	24,792	62.75	22%	2.70	1,078.78	1,102.80	-24.02	-2%	-4.08	0.00	-4.08	~-408%
Dummy Variable	24,792	77.87	37%	6.72	589.80	1,102.80	-513.00	-47%	-54.59	0.00	-54.59	~-5,459%
Neural Network	24,792	113.69	40%	3.18	446.76	1,102.80	-656.04	-59%	51.70	0.00	50.71	~5,071%
GWR	21,576	70.65	801%	2.94	941.48	1,102.80	-161.32	-15%	-42.56	0.00	-42.56	~-4,256%
Spatial Regression	19,413	76.11	52%	3.55	854.40	1,102.80	-248.40	-23%	-14.56	18.36	-32.92	-179%
Spatio-Temporal	17,664	2,293.00	158%	49.84	10,574.69	1,102.80	9,471.89	859%	-11,203.50	0.00	-11,203.50	~-1,120,350%

### 6.13 Gap Filling for 6 Months

The longest gap studied is six months (4,344 consecutive hours). As with the three month gap, the consecutive period can wrap around from the end of the year to the first of the year. All customers having 181 days of actual data are identified, along with all of the customers' actual-181-data-day-streaks. Within each of those customer-data combinations, ten sets of customer-data combinations are randomly sampled. A total of 43,440 hours are chosen in this way.

Tables 6.21 and 6.22 provide the results for the filling of the six month gap for residential and business customers, respectively. For each of the evaluation methods, the "best" result is highlighted in green. Note that, for residential customers, the GWR and spatial regression methods result in a smaller sample size than other methods. For business customers, the GWR, spatial regression, and spatiotemporal methods result in a smaller sample size than other methods.

**Table 6.21: Analysis Results for Six Months -- Residential**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	61,944	22.87	52%	0.43	511.44	547.92	-36.48	-7%	-37.72	0.00	-37.72	~-3,772%
Dummy Variable	61,944	26.06	129%	0.80	243.54	547.92	-304.38	-56%	-0.51	0.00	-0.51	~-51%
Neural Network	61,944	25.97	60%	0.57	173.47	547.92	-374.45	-68%	-0.03	0.00	-0.03	~-3%
GWR	43,389	14.40	620%	-0.11	464.61	547.92	-83.31	-15%	-25.75	0.00	-25.75	~-2,575%
Spatial Regression	43,389	16.42	309%	-0.06	464.23	547.92	-83.69	-15%	-25.23	0.00	-25.23	~-2,523%
Spatio-Temporal	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

**Table 6.22: Analysis Results for Six Months -- Business**

Gap-Filling Method	No. of Gaps	RMSE	MAPE	Average Error	Max. Filled Value	Max. Actual Value	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Min. Filled Value	Min. Actual Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
KEMA	61,944	50.07	20%	-3.50	1,387.90	1,337.28	50.62	4%	-8.89	0.00	-8.89	~-889%
Dummy Variable	61,944	64.06	105%	-6.51	1,272.32	1,337.28	-64.96	-5%	-88.73	0.00	-88.73	~-8,873%
Neural Network	61,944	89.93	35%	-24.26	1,116.39	1,337.28	-220.89	-17%	-35.07	0.00	-35.07	~-3,507%
GWR	43,368	45.82	82%	-1.07	1,406.62	1,337.28	69.34	5%	-29.40	0.00	-29.40	~-2,940%
Spatial Regression	30,381	46.00	125%	-3.42	854.40	1,016.88	-162.48	-16%	-28.21	0.12	-28.33	-23,605%
Spatio-Temporal	39,388	1,144.94	1,076%	151.87	6,608.71	1,337.28	5,271.43	394%	-13,881.14	0.00	-1,388,114.00	~-138,811,400%

## 6.14 Summary

Six alternative gap-filling methods have been used to fill gaps ranging in length from one hour to six months. Tables 6.23 and 6.24 show, for each gap length, the number of "best" evaluation statistics received by each gap-filling method for residential and business customers, respectively. Therefore, the higher the number of "best" ratings, the better the method compared to the other methods. The KEMA method has the best performance for both residential and business customers. For residential customers, the dummy variable method is second best, and three methods are tied for third: neural network, GWR, and spatial regression. For business customers GWR is second best, followed by the dummy variable method, while the neural network and spatial regression methods lag behind. The spatiotemporal method either was unable to provide results, or came in a poor sixth place.

No apparent pattern is demonstrated regarding whether certain gap-filling methods perform better for shorter or longer gaps. For example, the first three gap lengths are all one-hour in length but, for residential customers, three different methods perform best in filling them (including both temporal and spatial methods). For business customers there are two different methods that perform best for these gap lengths, both temporal methods. Similarly, for the two gap lengths that cover a 24-hour period the results are split among several gap-filling methods for both residential and business customers. Spatial methods are among the best performers for one-day gaps. For the longest gaps of one month or more, the KEMA method provides the best results.

**Table 6.23: Rating of Gap-Filling Methods by Gap Length -- Residential**

Gap Length	KEMA	Dummy Variable	Neural Network	GWR	Spatial Regression	Spatio-Temporal
System Peak Hour	6	1	0	0	0	n/a
Customer Peak Hour	2	0	0	0	5	n/a
1 Hour	0	4	2	1	0	n/a
3 Hours	3	3	0	2	0	n/a
12 Hours	0	5	1	1	0	n/a
Customer Peak Day	2	2	0	0	3	n/a
24 Hours	2	2	1	0	2	n/a
7 Days	1	1	3	2	0	n/a
1 Month	3	3	0	2	0	n/a
3 Months	3	0	2	2	0	n/a
6 Months	3	0	2	1	1	n/a
Total for all Gap Lengths	25	21	11	11	11	n/a

**Table 6.24: Rating of Gap-Filling Methods by Gap Length -- Business**

Gap Length	KEMA	Dummy Variable	Neural Network	GWR	Spatial Regression	Spatio-Temporal
System Peak Hour	0	1	3	2	2	0
Customer Peak Hour	4	0	0	3	0	0
1 Hour	3	2	2	0	0	0
3 Hours	0	4	0	3	0	0
12 Hours	5	2	0	0	0	0
Customer Peak Day	2	2	0	2	0	1
24 Hours	3	0	1	3	0	0
7 Days	5	1	0	1	0	0
1 Month	4	1	0	0	2	0
3 Months	6	0	0	0	1	0
6 Months	5	0	0	2	0	0
Total for all Gap Lengths	37	13	6	16	5	1

Tables 6.25 and 6.26 show, for residential and business customers respectively, how each of the gap-filling methods performed according to each of the evaluation statistics.

For residential customers using RMSE as the evaluation statistic, GWR is the best

performer. The dummy variable method has the best results using MAPE, and three methods tie using the average error: KEMA, neural network, and spatial regression. The KEMA method performs the best for getting close to the minimum and maximum values. For business customers, the KEMA method performs the best using any of the evaluation statistics.

**Table 6.25: Rating of Gap-Filling Methods by Evaluation Statistic -- Residential**

Gap Length	KEMA	Dummy Variable	Neural Network	GWR	Spatial Regression	Spatio-Temporal
RMSE	2	3	0	5	1	n/a
MAPE	2	7	0	1	1	n/a
Average Error	3	1	3	2	3	n/a
Absolute Difference in Max Value	4	3	0	1	3	n/a
Percent Difference in Max Value	4	3	0	2	3	n/a
Absolute Difference in Min Value	5	2	4	0	0	n/a
Percent Difference in Min Value	5	2	4	0	0	n/a

**Table 6.26: Rating of Gap-Filling Methods by Evaluation Statistic -- Business**

Gap Length	KEMA	Dummy Variable	Neural Network	GWR	Spatial Regression	Spatio-Temporal
RMSE	6	2	0	3	0	0
MAPE	6	3	1	1	0	0
Average Error	4	2	1	3	0	1
Absolute Difference in Max Value	5	2	1	2	1	0
Percent Difference in Max Value	5	2	1	3	1	0
Absolute Difference in Min Value	6	1	1	2	1	0
Percent Difference in Min Value	5	1	1	2	2	0

This research utilized a wide variety of energy usage interval gap lengths, meant to mimic real-life experience of data gaps. Another means of analyzing the results is to determine if, in general, gap-filling methods do a better job of filling short gaps than they do of filling long gaps. Tables 6.27 and 6.28 show, for residential and business customers respectively, the best available score for each of the evaluation statistics for each gap length, regardless of the gap-filling method used. For residential customers, the longer gap lengths of one month or more have poorer results when using RMSE and absolute difference in maximum value. Other statistics, including MAPE, average error, percent difference in maximum value, absolute difference in minimum value, and percent difference in minimum value do not show this trend. For business customers, the average error evaluation statistic indicates that better results are obtained for longer gaps. The absolute and percentage difference in minimum value statistics, however, show better results at shorter gap lengths. Other statistics show no apparent pattern.

Although the KEMA method performs the best as compared to the other methods tested, as was demonstrated in Tables 6.23 and 6.24, it is clear from the results in Tables 6.1 through 6.22 that none of the results are good on a consistent basis. The RMSE, for example, often exceeds the mean values of 2 (for residential) and 165 (for business), and the MAPE is frequently in excess of 100%. The percent differences in the calculation of the maximum value often exceed 50%. If nothing else, these results demonstrate that an improved gap-filling method is clearly needed.

**Table 6.27: Best Evaluation Results for Each Gap Length -- Residential**

Gap Length	RMSE	MAPE	Average Error	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
System Peak Hour	0.88	58%	0.40	0.30	6%	-0.04	-42%
Customer Peak Hour	3.20	291%	2.42	-3.15	-41%	-1.45	-66%
1 Hour	0.40	59%	0.03	-0.01	-1%	0.02	65%
3 Hours	0.17	36%	-0.01	-0.20	-19%	0.03	93%
12 Hours	0.65	44%	-0.04	-1.41	-19%	0.02	51%
Customer Peak Day	1.38	123%	0.68	0.70	13%	-0.04	-23%
24 Hours	0.32	41%	-0.02	-0.32	-9%	-0.00	-1%
7 Days	0.33	47%	0.00	0.23	7%	0.06	199%
1 Month	7.47	41%	-0.21	3.27	+/-2%	0.00	~0%
3 Months	3.62	38%	0.08	-3.26	-1%	-0.00	~-0%
6 Months	14.40	52%	-0.06	-36.48	-7%	-0.03	~-3%

**Table 6.28: Best Evaluation Results for Each Gap Length -- Business**

Gap Length	RMSE	MAPE	Average Error	Absolute Diff. in Max. Value	Percent Diff. in Max. Value	Absolute Diff. in Min. Value	Percent Diff. in Min. Value
System Peak Hour	228.78	25%	5.38	77.47	+/-10%	0.00	2%
Customer Peak Hour	25.11	22%	2.90	2.98	0%	0.94	17%
1 Hour	20.07	16%	6.23	-4.58	-1%	-0.12	-11%
3 Hours	25.05	14%	4.64	-55.09	-9%	-0.39	-5%
12 Hours	16.96	35%	6.13	-19.72	-6%	-0.05	-63%
Customer Peak Day	40.61	27%	-5.15	-71.83	-11%	0.20	36%
24 Hours	33.78	41%	-2.01	-13.37	-1%	-0.18	-14%
7 Days	70.94	20%	11.23	-123.90	-16%	0.11	15%
1 Month	4.42	26%	-0.58	0.51	1%	-4.60	~460%
3 Months	62.75	22%	2.70	-24.02	-2%	-4.08	-179%
6 Months	45.82	20%	-1.07	50.62	4%	-8.89	~-889%

The spatial methods generally performed well or nearly as well as the existing Load Research methods, which are largely temporal in nature. This clearly supports the notion that a good implementation of a spatiotemporal method should be able to outperform both



the spatial and the temporal methods. Unfortunately, the difficulty in implementing the spatiotemporal methods precluded a successful proof of that hypothesis. Further investigation is definitely needed in this realm.

## **7. Summary of Findings and Call for Further Research**

### **7.1 Summary of Findings**

The research conducted in this thesis clearly supports two major findings. First, improved gap-filling methods are clearly needed. Second, based on the mixed results between temporal and spatial methods, there is every reason to believe that a spatiotemporal method should be able to offer this significant improvement. Unfortunately, the two spatiotemporal methods that I presented in this research were unable to provide the desired results. I was unable to get one spatiotemporal method working with large data sets, and the other provided poor results.

### **7.2 Suggestions for Future Research**

During the course of this research, numerous decisions were made regarding the data exploration and analysis that limited the results in some way. These included limits on sample sizes, data transformations, analytic methods, and even the amount of time to be devoted to problem-solving in running certain computer programs. In the sections below, several of these limits are briefly discussed and expansions or improvements are proposed. Anyone conducting future research in this area may wish to consider some or all of these

improvements. The sections below are presented in roughly increasing order of importance.

### **7.2.1 Additional Data Transformations**

In this research, several transformations of the dependent variable were tested in the exploratory data analysis section, but none provided results that were superior to those provided by the un-transformed version of the dependent variable. In future research, additional transformations could be tested, including taking the log or natural log of the dependent and/or independent variables, and standardizing all variables to their mean.

### **7.2.2 Exploratory Data Analysis**

In this research, the exploratory data analysis was largely based on small samples of data. Future research may wish to expand the exploratory data analysis to include the entire data set, to avoid inconsistency of results from different samples. Alternatively, a common data sample could be chosen and all exploratory analysis run against that sample.

### **7.2.3 Statistical Sampling**

The gap-filling methods tested in this research were implemented on a small sample of 10 customers for each method, as a way of providing a baseline of comparative results. Further research could include a larger sample that would provide results with a desired power or level of confidence in the results. An appropriate sample could be calculated using Cochran's (1977, p. 77) formula for a normal distribution with Lehmann's upward adjustment of 15 percent for a non-normal distribution (Lehmann 1975, p. 76-81), or by assuming a Poisson (or other non-normal) distribution.

#### **7.2.4 Additional Evaluation Methods for Gap Filling**

Several statistical tests were included to evaluate and compare the various gap-filling methods. An additional method not included here would be to identify the worst hour filled, i.e., the largest difference between the actual and predicted values for any single hour. Other evaluation methods from statistical literature may also be available.

#### **7.2.5 Optimize Each Gap-Filling Method**

In this research, the three Load Research methods made use of the independent variables suggested by prior researchers. For the spatial and spatiotemporal methods, a limited set of independent variables suggested by the exploratory data analysis were included. An alternative approach would have been to select a single "master" set of independent variables and to make use of those same variables in each of the gap-filling methods. A second alternative would be to optimize each of the gap-filling methods by including different independent variables or experimenting with alternative functions to produce the best set of results for each method. Further experimentation along these lines could well result in improved gap-filling.

#### **7.2.6 Alternative Gap-Filling Methods**

Although the spatial locations of customers of energy providers are not amenable to raster specification, the temporal nature of their interval usage data can readily be placed into a raster framework of 365 days per year by 24 hours per day. Because these raster assignments are not based on geographic location, the question becomes whether or not such an assignment can still be geographic in nature. I believe that the answer is yes, and that spatial analytic methods can still be applied to this situation. One possible extension of

the research conducted for this thesis would be to apply co-kriging methodology to a set of temporal rasters, one for each customer. To fill any particular data gap, the spatial distance between customers would be used to identify and select appropriate neighbors, whose time-based raster sets would then be utilized to help predict the gaps for the target customer.

### **7.2.7 Solve Computer-Related Issues**

During the analytic phase of my research, I ran into several computer problems, particularly related to the use of various R packages. Each of these problems had repercussions for the results presented in this thesis. The problems and their repercussions are discussed in the sections below.

#### **7.2.7.1 *Memory Issues***

Memory issues limiting the amount of data that could be analyzed within a single run -- The memory issues necessitated running the spatiotemporal analysis one month at a time rather than the entire year in a single run, thus severely limiting the temporal aspects of the spatiotemporal analysis. Memory issues also limited the number of iterations that could be used for the spatiotemporal analysis.

#### **7.2.7.2 *Excessive Run Times***

Long run times, of up to 36 hours or more, resulted in the use of a single input dataset for the GWR and spatial regression runs, rather than separate datasets for each gap length. Use of the single input dataset gives the GWR and spatial regression results a slight advantage over the other gap filling methods. Long run times were also associated with the ultimately unsuccessful spatiotemporal gap-filling methods for residential customers.

### **7.2.7.3      *Singular Matrices***

Singular matrices that could not be inverted prevented certain data gaps from being filled by the GWR and spatial regression methods. A singular matrix problem also prevented me from successfully using the SpatioTemporal R package.

### **7.2.7.4      *Not All Data Gaps Are Filled***

Whether resulting from matrices that could not be inverted, memory issues, missing data issues, or other unknown factors, not all of the methods filled all of the data gaps. The ideal gap-filling method should be able to make a prediction for all missing values of the dependent variable.

### **7.2.7.5      *Poor Results from Spatiotemporal Gap Filling***

Tests with the SpatioTemporal and spTimer R packages produced reasonably good filled values when using a small test dataset. When working with the real datasets, however, I was unable to produce any results with the SpatioTemporal package and only very poor results with the spTimer package. Repeated tinkering with the parameters offered no significant improvements. It seems counterintuitive to receive poorer results from a spatiotemporal analysis than from either of the separate spatial or temporal analyses. Therefore, further investigation of spatiotemporal methods for gap-filling and prediction is clearly needed.

## **7.3      Overall Summary**

The work summarized in this thesis provides several contributions. First, a methodology for evaluating gap-filling methods was established, incorporating choices of gap-filling methods, gap lengths, and evaluation statistics. Second, the importance of using

multiple evaluation statistics in a study of this nature was demonstrated. Third, a head-to-head comparison of temporal, spatial, and spatiotemporal methods was conducted. As a result, limits were identified in the ability of two R spatiotemporal packages to predict values. Fourth, based on the exploratory spatial data analysis, there is value in incorporating geodemographic data variables into the gap-filling process. Finally, several improvements to the process have been identified, which may aid in the development of improved methods in the future.

## Appendix A: Summary Statistics for Each Variable Used In the Analysis

In this appendix, summary statistics are provided for each analytic variable included in the analysis, including the mean, standard deviation, and range. Table A.1 provides the results for the residential analysis variables.

**Table A.1: Summary Statistics for Residential Analysis Variables**

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Customer Identification Number	custid	118,682	145,463	5,175	769,294
Date	date	July 2	105 days	January 1	December 31
Latitude	y	confidential	958	confidential	confidential
Longitude	x	confidential	3,064	confidential	confidential
Annual kWh Energy Usage	annkwh	14,784	102,523	0	1,825,600
Maximum kW Demand During Year	maxdmd	321	239	2	803
Energy Usage 1 am	k1	1.28	8	0	237
Energy Usage 2am	k2	1.18	8	0	259
Energy Usage 3am	k3	1.13	7	0	254
Energy Usage 4am	k4	1.11	7	0	259
Energy Usage 5am	k5	1.12	8	0	364
Energy Usage 6am	k6	1.26	9	0	531
Energy Usage 7am	k7	1.47	12	0	547
Energy Usage 8am	k8	1.74	15	0	642
Energy Usage 9am	k9	2.01	18	0	752
Energy Usage 10am	k10	2.15	19	0	756
Energy Usage 11am	k11	2.23	20	0	796
Energy Usage noon	k12	2.25	20	0	777
Energy Usage 1pm	k13	2.21	19	0	766
Energy Usage 2pm	k14	2.15	19	0	784
Energy Usage 3pm	k15	2.05	18	0	746
Energy Usage 4pm	k16	1.90	15	0	535
Energy Usage 5pm	k17	1.78	13	0	449
Energy Usage 6pm	k18	1.77	13	0	418
Energy Usage 7pm	k19	1.77	12	0	420
Energy Usage 8pm	k20	1.81	12	0	485
Energy Usage 9pm	k21	1.79	12	0	482
Energy Usage 10pm	k22	1.70	11	0	387
Energy Usage 11pm	k23	1.54	9	0	356
Energy Usage midnight	k24	1.41	9	0	261
Energy Usage As Percent of Daily Maximum 1 am	pctd1	48	28	0	100
Energy Usage As Percent of Daily Maximum 2am	pctd2	43	26	0	100
Energy Usage As Percent of	pctd3	40	25	0	100



Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Daily Maximum 3am					
Energy Usage As Percent of Daily Maximum 4am	pctd4	38	24	0	100
Energy Usage As Percent of Daily Maximum 5am	pctd5	37	24	0	100
Energy Usage As Percent of Daily Maximum 6am	pctd6	37	24	0	100
Energy Usage As Percent of Daily Maximum 7am	pctd7	39	25	0	100
Energy Usage As Percent of Daily Maximum 8am	pctd8	43	26	0	100
Energy Usage As Percent of Daily Maximum 9am	pctd9	44	26	0	100
Energy Usage As Percent of Daily Maximum 10am	pctd10	43	26	0	100
Energy Usage As Percent of Daily Maximum 11am	pctd11	43	26	0	100
Energy Usage As Percent of Daily Maximum noon	pctd12	43	26	0	100
Energy Usage As Percent of Daily Maximum 1pm	pctd13	44	27	0	100
Energy Usage As Percent of Daily Maximum 2pm	pctd14	44	27	0	100
Energy Usage As Percent of Daily Maximum 3pm	pctd15	44	27	0	100
Energy Usage As Percent of Daily Maximum 4pm	pctd16	44	27	0	100
Energy Usage As Percent of Daily Maximum 5pm	pctd17	45	27	0	100
Energy Usage As Percent of Daily Maximum 6pm	pctd18	49	28	0	100
Energy Usage As Percent of Daily Maximum 7pm	pctd19	53	28	0	100
Energy Usage As Percent of Daily Maximum 8pm	pctd20	58	28	0	100
Energy Usage As Percent of Daily Maximum 9pm	pctd21	61	28	0	100
Energy Usage As Percent of Daily Maximum 10pm	pctd22	62	28	0	100
Energy Usage As Percent of Daily Maximum 11pm	pctd23	60	27	0	100
Energy Usage As Percent of Daily Maximum midnight	pctd24	55	28	0	100
Energy Usage As Percent of Monthly Billed kWh 1 am	pctm1	0.15	0.15	0	17
Energy Usage As Percent of Monthly Billed kWh 2am	pctm2	0.13	0.14	0	17
Energy Usage As Percent of Monthly Billed kWh 3am	pctm3	0.12	0.13	0	17
Energy Usage As Percent of Monthly Billed kWh 4am	pctm4	0.11	0.13	0	17
Energy Usage As Percent of	pctm5	0.11	0.12	0	17

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Monthly Billed kWh 5am					
Energy Usage As Percent of Monthly Billed kWh 6am	pctm6	0.11	0.11	0	17
Energy Usage As Percent of Monthly Billed kWh 7am	pctm7	0.11	0.12	0	17
Energy Usage As Percent of Monthly Billed kWh 8am	pctm8	0.13	0.13	0	17
Energy Usage As Percent of Monthly Billed kWh 9am	pctm9	0.13	0.14	0	17
Energy Usage As Percent of Monthly Billed kWh 10am	pctm10	0.13	0.18	0	60
Energy Usage As Percent of Monthly Billed kWh 11am	pctm11	0.13	0.15	0	17
Energy Usage As Percent of Monthly Billed kWh noon	pctm12	0.13	0.16	0	17
Energy Usage As Percent of Monthly Billed kWh 1pm	pctm13	0.13	0.16	0	17
Energy Usage As Percent of Monthly Billed kWh 2pm	pctm14	0.13	0.17	0	17
Energy Usage As Percent of Monthly Billed kWh 3pm	pctm15	0.13	0.17	0	17
Energy Usage As Percent of Monthly Billed kWh 4pm	pctm16	0.14	0.17	0	17
Energy Usage As Percent of Monthly Billed kWh 5pm	pctm17	0.14	0.17	0	17
Energy Usage As Percent of Monthly Billed kWh 6pm	pctm18	0.15	0.19	0	47
Energy Usage As Percent of Monthly Billed kWh 7pm	pctm19	0.17	0.20	0	47
Energy Usage As Percent of Monthly Billed kWh 8pm	pctm20	0.19	0.20	0	30
Energy Usage As Percent of Monthly Billed kWh 9pm	pctm21	0.20	0.21	0	47
Energy Usage As Percent of Monthly Billed kWh 10pm	pctm22	0.20	0.19	0	17
Energy Usage As Percent of Monthly Billed kWh 11pm	pctm23	0.20	0.18	0	17
Energy Usage As Percent of Monthly Billed kWh midnight	pctm24	0.17	0.17	0	17
Energy Usage As Percent of Annual Billed kWh 1 am	pcta1	0.01	0.01	0	0.26
Energy Usage As Percent of Annual Billed kWh 2am	pcta2	0.01	0.01	0	0.22
Energy Usage As Percent of Annual Billed kWh 3am	pcta3	0.01	0.01	0	0.22
Energy Usage As Percent of Annual Billed kWh 4am	pcta4	0.01	0.01	0	0.22
Energy Usage As Percent of Annual Billed kWh 5am	pcta5	0.01	0.01	0	0.22
Energy Usage As Percent of Annual Billed kWh 6am	pcta6	0.01	0.01	0	0.22
Energy Usage As Percent of	pcta7	0.01	0.01	0	0.22

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Annual Billed kWh 7am					
Energy Usage As Percent of Annual Billed kWh 8am	pcta8	0.01	0.01	0	0.22
Energy Usage As Percent of Annual Billed kWh 9am	pcta9	0.01	0.01	0	0.56
Energy Usage As Percent of Annual Billed kWh 10am	pcta10	0.01	0.02	0	10.00
Energy Usage As Percent of Annual Billed kWh 11am	pcta11	0.01	0.07	0	30.00
Energy Usage As Percent of Annual Billed kWh noon	pcta12	0.01	0.06	0	30.00
Energy Usage As Percent of Annual Billed kWh 1pm	pcta13	0.01	0.02	0	11.00
Energy Usage As Percent of Annual Billed kWh 2pm	pcta14	0.01	0.01	0	0.74
Energy Usage As Percent of Annual Billed kWh 3pm	pcta15	0.01	0.01	0	0.66
Energy Usage As Percent of Annual Billed kWh 4pm	pcta16	0.01	0.01	0	0.65
Energy Usage As Percent of Annual Billed kWh 5pm	pcta17	0.01	0.01	0	0.70
Energy Usage As Percent of Annual Billed kWh 6pm	pcta18	0.01	0.01	0	0.68
Energy Usage As Percent of Annual Billed kWh 7pm	pcta19	0.01	0.01	0	0.59
Energy Usage As Percent of Annual Billed kWh 8pm	pcta20	0.01	0.01	0	0.72
Energy Usage As Percent of Annual Billed kWh 9pm	pcta21	0.02	0.01	0	0.71
Energy Usage As Percent of Annual Billed kWh 10pm	pcta22	0.02	0.01	0	0.73
Energy Usage As Percent of Annual Billed kWh 11pm	pcta23	0.02	0.01	0	0.56
Energy Usage As Percent of Annual Billed kWh midnight	pcta24	0.01	0.01	0	0.40
Energy Usage As Percent of Prior Interval 1 am	delt1	-5	66	-100	6,500
Energy Usage As Percent of Prior Interval 2am	delt2	-4	61	-100	8,467
Energy Usage As Percent of Prior Interval 3am	delt3	-2	57	-100	10,700
Energy Usage As Percent of Prior Interval 4am	delt4	1	54	-100	13,700
Energy Usage As Percent of Prior Interval 5am	delt5	3	61	-100	12,400
Energy Usage As Percent of Prior Interval 6am	delt6	9	74	-100	7,700
Energy Usage As Percent of Prior Interval 7am	delt7	18	81	-100	7,500
Energy Usage As Percent of Prior Interval 8am	delt8	30	113	-100	10,150
Energy Usage As Percent of	delt9	21	126	-100	36,400

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Prior Interval 9am					
Energy Usage As Percent of Prior Interval 10am	delt10	12	90	-100	7,537
Energy Usage As Percent of Prior Interval 11am	delt11	13	83	-100	8,400
Energy Usage As Percent of Prior Interval noon	delt12	14	292	-100	154,000
Energy Usage As Percent of Prior Interval 1pm	delt13	13	89	-100	9,200
Energy Usage As Percent of Prior Interval 2pm	delt14	11	83	-100	10,700
Energy Usage As Percent of Prior Interval 3pm	delt15	12	94	-100	21,250
Energy Usage As Percent of Prior Interval 4pm	delt16	14	87	-100	8,800
Energy Usage As Percent of Prior Interval 5pm	delt17	17	110	-100	33,800
Energy Usage As Percent of Prior Interval 6pm	delt18	23	101	-100	11,150
Energy Usage As Percent of Prior Interval 7pm	delt19	29	115	-100	13,250
Energy Usage As Percent of Prior Interval 8pm	delt20	29	130	-100	24,200
Energy Usage As Percent of Prior Interval 9pm	delt21	25	112	-100	9,100
Energy Usage As Percent of Prior Interval 10pm	delt22	18	99	-100	7,700
Energy Usage As Percent of Prior Interval 11pm	delt23	10	89	-100	12,000
Energy Usage As Percent of Prior Interval midnight	delt24	1	81	-100	11,000
Dry Bulb Temperature 1am	tmp1	54	17	10	89
Dry Bulb Temperature 2am	tmp2	53	16	9	89
Dry Bulb Temperature 3am	tmp3	53	16	8	86
Dry Bulb Temperature 4am	tmp4	53	16	7	85
Dry Bulb Temperature 5am	tmp5	52	16	7	85
Dry Bulb Temperature 6am	tmp6	52	16	6	85
Dry Bulb Temperature 7am	tmp7	52	16	6	85
Dry Bulb Temperature 8am	tmp8	53	17	6	87
Dry Bulb Temperature 9am	tmp9	54	17	8	90
Dry Bulb Temperature 10am	tmp10	55	17	10	93
Dry Bulb Temperature 11am	tmp11	57	18	10	97
Dry Bulb Temperature noon	tmp12	59	18	13	100
Dry Bulb Temperature 1pm	tmp13	60	18	15	102
Dry Bulb Temperature 2pm	tmp14	61	18	17	103
Dry Bulb Temperature 3pm	tmp15	61	18	19	102
Dry Bulb Temperature 4pm	tmp16	61	18	19	102
Dry Bulb Temperature 5pm	tmp17	60	18	19	101
Dry Bulb Temperature 6pm	tmp18	59	18	19	101
Dry Bulb Temperature 7pm	tmp19	58	17	19	99
Dry Bulb Temperature 8pm	tmp20	57	17	18	97

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Dry Bulb Temperature 9pm	tmp21	56	17	17	94
Dry Bulb Temperature 10pm	tmp22	56	17	15	92
Dry Bulb Temperature 11pm	tmp23	55	17	13	90
Dry Bulb Temperature midnight	tmp24	55	17	11	89
Cooling Degree Hours 1am	cdh1	2.44	4.38	0	24
Cooling Degree Hours 2am	cdh2	2.28	4.16	0	24
Cooling Degree Hours 3am	cdh3	2.09	3.89	0	21
Cooling Degree Hours 4am	cdh4	1.93	3.68	0	20
Cooling Degree Hours 5am	cdh5	1.75	3.44	0	20
Cooling Degree Hours 6am	cdh6	1.66	3.36	0	20
Cooling Degree Hours 7am	cdh7	1.72	3.41	0	20
Cooling Degree Hours 8am	cdh8	2.10	3.92	0	2
Cooling Degree Hours 9am	cdh9	2.61	4.59	0	25
Cooling Degree Hours 10am	cdh10	3.22	5.38	0	28
Cooling Degree Hours 11am	cdh11	4.06	6.39	0	32
Cooling Degree Hours noon	cdh12	5.00	7.47	0	35
Cooling Degree Hours 1pm	cdh13	5.56	8.10	0	37
Cooling Degree Hours 2pm	cdh14	5.87	8.42	0	38
Cooling Degree Hours 3pm	cdh15	5.92	8.42	0	37
Cooling Degree Hours 4pm	cdh16	5.57	8.09	0	37
Cooling Degree Hours 5pm	cdh17	5.20	7.74	0	36
Cooling Degree Hours 6pm	cdh18	4.75	7.37	0	36
Cooling Degree Hours 7pm	cdh19	4.20	6.85	0	34
Cooling Degree Hours 8pm	cdh20	3.73	6.25	0	32
Cooling Degree Hours 9pm	cdh21	3.34	5.70	0	29
Cooling Degree Hours 10pm	cdh22	3.07	5.31	0	27
Cooling Degree Hours 11pm	cdh23	2.89	5.01	0	25
Cooling Degree Hours midnight	cdh24	2.67	4.68	0	24
Heating Degree Hours 1am	hdh1	13.51	13.74	0	55
Heating Degree Hours 2am	hdh2	13.80	13.83	0	56
Heating Degree Hours 3am	hdh3	14.12	13.97	0	57
Heating Degree Hours 4am	hdh4	14.39	14.12	0	58
Heating Degree Hours 5am	hdh5	14.56	14.14	0	58
Heating Degree Hours 6am	hdh6	14.79	14.24	0	59
Heating Degree Hours 7am	hdh7	14.82	14.31	0	59
Heating Degree Hours 8am	hdh8	14.48	14.31	0	59
Heating Degree Hours 9am	hdh9	13.72	14.06	0	57
Heating Degree Hours 10am	hdh10	12.76	13.56	0	55
Heating Degree Hours 11am	hdh11	11.79	13.15	0	55
Heating Degree Hours noon	hdh12	10.94	12.66	0	52
Heating Degree Hours 1pm	hdh13	10.28	12.25	0	50
Heating Degree Hours 2pm	hdh14	9.88	12.09	0	48
Heating Degree Hours 3pm	hdh15	9.76	11.93	0	46
Heating Degree Hours 4pm	hdh16	9.90	11.96	0	46
Heating Degree Hours 5pm	hdh17	10.25	12.16	0	46
Heating Degree Hours 6pm	hdh18	10.70	12.35	0	46
Heating Degree Hours 7pm	hdh19	11.20	12.55	0	46
Heating Degree Hours 8pm	hdh20	11.66	12.78	0	47
Heating Degree Hours 9pm	hdh21	12.05	12.99	0	48
Heating Degree Hours 10pm	hdh22	12.42	13.14	0	50

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Heating Degree Hours 11pm	hdh23	12.80	13.34	0	52
Heating Degree Hours midnight	hdh24	13.16	13.52	0	54
Humidity 1am	hum1	70	18	19	100
Humidity 2am	hum2	71	17	22	100
Humidity 3am	hum3	72	17	23	100
Humidity 4am	hum4	73	17	25	100
Humidity 5am	hum5	74	17	23	100
Humidity 6am	hum6	74	17	27	100
Humidity 7am	hum7	74	16	29	100
Humidity 8am	hum8	73	16	23	100
Humidity 9am	hum9	70	17	33	100
Humidity 10am	hum10	66	17	17	100
Humidity 11am	hum11	62	18	20	100
Humidity noon	hum12	59	19	23	100
Humidity 1pm	hum13	57	20	23	100
Humidity 2pm	hum14	56	20	21	100
Humidity 3pm	hum15	55	20	19	100
Humidity 4pm	hum16	56	21	21	100
Humidity 5pm	hum17	58	20	19	100
Humidity 6pm	hum18	59	20	19	100
Humidity 7pm	hum19	61	20	21	100
Humidity 8pm	hum20	63	20	24	100
Humidity 9pm	hum21	65	20	24	100
Humidity 10pm	hum22	66	19	25	100
Humidity 11pm	hum23	67	19	17	100
Humidity midnight	hum24	69	18	15	100
Wind Speed 1am	wsp1	5	4	0	18
Wind Speed 2am	wsp2	5	4	0	23
Wind Speed 3am	wsp3	5	4	0	20
Wind Speed 4am	wsp4	5	4	0	20
Wind Speed 5am	wsp5	5	4	0	20
Wind Speed 6am	wsp6	5	4	0	23
Wind Speed 7am	wsp7	5	4	0	18
Wind Speed 8am	wsp8	5	4	0	20
Wind Speed 9am	wsp9	6	4	0	22
Wind Speed 10am	wsp10	6	4	0	16
Wind Speed 11am	wsp11	6	4	0	18
Wind Speed noon	wsp12	7	4	0	17
Wind Speed 1pm	wsp13	7	4	0	18
Wind Speed 2pm	wsp14	7	4	0	18
Wind Speed 3pm	wsp15	7	4	0	20
Wind Speed 4pm	wsp16	7	4	0	20
Wind Speed 5pm	wsp17	6	4	0	21
Wind Speed 6pm	wsp18	7	4	0	20
Wind Speed 7pm	wsp19	6	4	0	20
Wind Speed 8pm	wsp20	6	4	0	21
Wind Speed 9pm	wsp21	6	4	0	24
Wind Speed 10pm	wsp22	6	4	0	23
Wind Speed 11pm	wsp23	5	4	0	18
Wind Speed midnight	wsp24	5	4	0	17
Cloud Cover 1am	cc1	38	45	0	100

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Cloud Cover 2am	cc2	36	43	0	100
Cloud Cover 3am	cc3	38	45	0	100
Cloud Cover 4am	cc4	38	45	0	100
Cloud Cover 5am	cc5	41	45	0	100
Cloud Cover 6am	cc6	40	45	0	100
Cloud Cover 7am	cc7	39	45	0	100
Cloud Cover 8am	cc8	39	45	0	100
Cloud Cover 9am	cc9	36	44	0	100
Cloud Cover 10am	cc10	36	44	0	100
Cloud Cover 11am	cc11	35	43	0	100
Cloud Cover noon	cc12	34	41	0	100
Cloud Cover 1pm	cc13	38	41	0	100
Cloud Cover 2pm	cc14	38	40	0	100
Cloud Cover 3pm	cc15	36	41	0	100
Cloud Cover 4pm	cc16	37	41	0	100
Cloud Cover 5pm	cc17	36	40	0	100
Cloud Cover 6pm	cc18	38	43	0	100
Cloud Cover 7pm	cc19	38	43	0	100
Cloud Cover 8pm	cc20	39	43	0	100
Cloud Cover 9pm	cc21	38	43	0	100
Cloud Cover 10pm	cc22	37	44	0	100
Cloud Cover 11pm	cc23	38	44	0	100
Cloud Cover midnight	cc24	39	44	0	100
Sunshine Minutes 1am	ssm1	0	0	0	0
Sunshine Minutes 2am	ssm2	0	0	0	0
Sunshine Minutes 3am	ssm3	0	0	0	0
Sunshine Minutes 4am	ssm4	0	0	0	0
Sunshine Minutes 5am	ssm5	4	10	0	37
Sunshine Minutes 6am	ssm6	17	22	0	60
Sunshine Minutes 7am	ssm7	35	26	0	60
Sunshine Minutes 8am	ssm8	37	27	0	60
Sunshine Minutes 9am	ssm9	38	27	0	60
Sunshine Minutes 10am	ssm10	38	26	0	60
Sunshine Minutes 11am	ssm11	39	26	0	60
Sunshine Minutes noon	ssm12	39	25	0	60
Sunshine Minutes 1pm	ssm13	37	25	0	60
Sunshine Minutes 2pm	ssm14	37	24	0	60
Sunshine Minutes 3pm	ssm15	38	24	0	60
Sunshine Minutes 4pm	ssm16	35	24	0	60
Sunshine Minutes 5pm	ssm17	28	25	0	60
Sunshine Minutes 6pm	ssm18	22	26	0	60
Sunshine Minutes 7pm	ssm19	15	22	0	60
Sunshine Minutes 8pm	ssm20	4	8	0	31
Sunshine Minutes 9pm	ssm21	0	0	0	0
Sunshine Minutes 10pm	ssm22	0	0	0	0
Sunshine Minutes 11pm	ssm23	0	0	0	0
Sunshine Minutes midnight	ssm24	0	0	0	0
Billed Energy January	c1	1,307	9,058	0	134,400
Billed Energy Febuary	c2	1,290	9,340	0	157,600
Billed Energy March	c3	1,209	8,821	0	144,800
Billed Energy April	c4	1,091	7,694	0	115,200



Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Billed Energy May	c5	1,045	7,435	0	128,000
Billed Energy June	c6	1,333	9,983	0	196,000
Billed Energy July	c7	1,474	9,946	0	174,400
Billed Energy August	c8	1,538	9,419	0	144,800
Billed Energy September	c9	1,452	9,737	0	192,800
Billed Energy October	c10	1,227	8,755	0	167,200
Billed Energy November	c11	1,084	7,847	0	149,600
Billed Energy December	c12	1,055	7,342	0	130,400
Billed Demand January	d1	184	155	1.32	472
Billed Demand Febuary	d2	189	168	1.20	544
Billed Demand March	d3	191	176	0.96	554
Billed Demand April	d4	205	170	2.52	569
Billed Demand May	d5	238	174	2.52	579
Billed Demand June	d6	295	202	3.42	736
Billed Demand July	d7	317	203	4.50	721
Billed Demand August	d8	307	189	3.96	622
Billed Demand September	d9	305	240	0.84	803
Billed Demand October	d10	271	217	2.04	778
Billed Demand November	d11	190	165	0.72	555
Billed Demand December	d12	151	131	1.20	446
Lot Footprint	footprint	100,007	182,609	1,450	718,100
Number of Buildings	no_bldgs	1.92	2	1	6
Year Built	year_built	1967	38	1890	2008
Number of Floors	no_floors	17	15	1	42
Building Area	bldg_area	274,985	257,948	952	724,475
Number of Units	no_units	281	226	1	580
Number of Residential Units	no_res_units	269	221	0	578
Floor-Area Ratio	far	4.32	2.94	0.26	9.54
Building Area Per Unit	bapu	2,339	14,685	278	263,000
Building Area Per Residential Unit	bapru	1,005	486	0	7,536
Children Under 5 As Percent of Census Tract Population	kidlt5	7.13	2.32	1.28	55.74
Children Between 5 and 9 As Percent of Census Tract Population	kid59	2.74	1.75	0	12.27
Children Between 10 and 14 As Percent of Census Tract Population	kid1014	1.81	1.55	0	12.41
Children Between 15 and 19 As Percent of Census Tract Population	kid1519	3.03	1.75	0	16.44
Children Under 10 As Percent of Census Tract Population	kidle9	9.86	2.56	4.69	55.74
Children Under 15 As Percent of Census Tract Population	kidle14	11.68	3.29	7.98	55.74
Children Under 20 As Percent of Census Tract Population	kidle19	14.71	4.28	9.92	55.74
Median Age of Census Tract Population	medage	34.44	3.16	1.90	48.40
People 65 And Over As Percent	srcit	8.04	3.95	0	26.68



Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
of Census Tract Population					
Median Household Income of Census Tract	medinc	81,868	31,210	38,750	136,053
People Who Are One Race - White As Percent of Census Tract Population	white	63.50	12.09	0	97.46
People Who Are One Race - Black As Percent of Census Tract Population	black	4.83	6.18	0	96.66
People Who Are One Race - Asian As Percent of Census Tract Population	asian	24.16	10.50	0	64.53
People Who Are One Race - Latino As Percent of Census Tract Population	latino	18.89	8.06	0	78.28
People Of American Ancestry As Percent of Census Tract Population	america	2.59	1.80	0	10.39
People Of Guyanese Ancestry As Percent of Census Tract Population	guyana	0.22	0.70	0	12.41
People Of Irish Ancestry As Percent of Census Tract Population	ireland	8.14	3.59	0	55.74
People Of Italian Ancestry As Percent of Census Tract Population	italy	6.95	3.59	0	57.56
People Of Polish Ancestry As Percent of Census Tract Population	poland	2.71	1.64	0	6.65
People Of West Indian Ancestry As Percent of Census Tract Population	windies	0.77	2.35	0	44.77
People Of Jamaican Ancestry As Percent of Census Tract Population	jamaica	0.37	1.44	0	27.36
People Who Are Foreign Born As Percent of Census Tract Population	foreign	42.32	11.53	0	70.15
Percent of Occupied Housing Units With Electric Heat	elecheat	25.27	23.90	0	58.10
Average Household Size	hhsiz	2.18	0.35	1.88	4.80
People With Less Than a High School Education As Percent of Census Tract Population	edlths	9.59	9.06	0	33.00
People With a High School Diploma As Percent of Census Tract Population	edhs	14.54	6.73	0	49.50
People With Some College Education As Percent of Census Tract Population	edsmcoll	16.18	4.80	0	42.10
People With a College Diploma	edcoll	59.72	17.25	10.80	100.00

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
or More As Percent of Census Tract Population					
People With a High School Diploma or Less As Percent of Census Tract Population	edlehs	24.13	13.53	0	64.20
People With Some College or Less As Percent of Census Tract Population	edltcoll	40.31	17.28	0	89.30

Table A.2 provides the results for the business analysis variables.

**Table A.2: Summary Statistics for Business Analysis Variables**

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Customer Identification Number	custid	468,791	349,507	1,046	769,546
Date	date	June 30	106 days	January 1	December 31
Latitude	y	confidential	2,577	confidential	confidential
Longitude	x	confidential	5,472	confidential	confidential
Annual kWh Energy Usage	annkwh	1,432,207	1,714,547	0	9,780,000
Maximum kW Demand During Year	maxdmd	344.96	344.68	4	1,411
Energy Usage 1 am	k1	135.22	185.33	0	1,324.56
Energy Usage 2am	k2	132.33	182.58	0	1,326.24
Energy Usage 3am	k3	130.45	180.89	0	1,314.48
Energy Usage 4am	k4	129.87	179.89	0	1,347.12
Energy Usage 5am	k5	131.05	179.98	0	1,354.08
Energy Usage 6am	k6	137.64	184.95	0	1,356.72
Energy Usage 7am	k7	151.42	193.82	0	1,355.52
Energy Usage 8am	k8	164.40	202.52	0	1,367.28
Energy Usage 9am	k9	176.94	211.40	0	1,382.64
Energy Usage 10am	k10	185.70	218.47	0	1,389.36
Energy Usage 11am	k11	192.27	226.20	0	1,397.52
Energy Usage noon	k12	194.95	229.58	0	1,386.00
Energy Usage 1pm	k13	194.96	230.09	0	1,402.80
Energy Usage 2pm	k14	196.41	231.26	0	1,405.20
Energy Usage 3pm	k15	194.95	231.05	0	1,405.44
Energy Usage 4pm	k16	192.48	230.06	0	1,405.48
Energy Usage 5pm	k17	187.84	228.17	0	1,376.16
Energy Usage 6pm	k18	182.08	225.63	0	1,370.16
Energy Usage 7pm	k19	175.13	220.80	0	1,372.08
Energy Usage 8pm	k20	169.05	216.99	0	1,360.32
Energy Usage 9pm	k21	163.69	212.40	0	1,349.28
Energy Usage 10pm	k22	158.51	206.90	0	1,341.12
Energy Usage 11pm	k23	148.41	196.74	0	1,334.64
Energy Usage midnight	k24	139.80	188.68	0	1,326.00
Energy Usage As Percent of Daily Maximum 1 am	pctd1	55	31	0.02	100
Energy Usage As Percent of Daily Maximum 2am	pctd2	54	31	0.02	100
Energy Usage As Percent of Daily Maximum 3am	pctd3	53	31	0.02	100
Energy Usage As Percent of Daily Maximum 4am	pctd4	53	31	0.02	100
Energy Usage As Percent of Daily Maximum 5am	pctd5	53	30	0.02	100
Energy Usage As Percent of Daily Maximum 6am	pctd6	56	30	0.02	100
Energy Usage As Percent of Daily Maximum 7am	pctd7	62	29	0.02	100

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Energy Usage As Percent of Daily Maximum 8am	pctd8	68	27	0.11	100
Energy Usage As Percent of Daily Maximum 9am	pctd9	76	24	0.10	100
Energy Usage As Percent of Daily Maximum 10am	pctd10	82	21	0.10	100
Energy Usage As Percent of Daily Maximum 11am	pctd11	85	19	0.03	100
Energy Usage As Percent of Daily Maximum noon	pctd12	87	18	0.18	100
Energy Usage As Percent of Daily Maximum 1pm	pctd13	87	18	0.14	100
Energy Usage As Percent of Daily Maximum 2pm	pctd14	87	18	0.05	100
Energy Usage As Percent of Daily Maximum 3pm	pctd15	86	18	0.03	100
Energy Usage As Percent of Daily Maximum 4pm	pctd16	85	19	0.04	100
Energy Usage As Percent of Daily Maximum 5pm	pctd17	81	22	0.02	100
Energy Usage As Percent of Daily Maximum 6pm	pctd18	76	26	0.02	100
Energy Usage As Percent of Daily Maximum 7pm	pctd19	71	29	0.02	100
Energy Usage As Percent of Daily Maximum 8pm	pctd20	67	31	0.02	100
Energy Usage As Percent of Daily Maximum 9pm	pctd21	65	32	0.02	100
Energy Usage As Percent of Daily Maximum 10pm	pctd22	62	32	0.02	100
Energy Usage As Percent of Daily Maximum 11pm	pctd23	59	31	0.02	100
Energy Usage As Percent of Daily Maximum midnight	pctd24	56	31	0.02	100
Energy Usage As Percent of Monthly Billed kW 1 am	pctm1	37	28	0.02	317
Energy Usage As Percent of Monthly Billed kW 2am	pctm2	36	28	0.02	319
Energy Usage As Percent of Monthly Billed kW 3am	pctm3	36	27	0.02	313
Energy Usage As Percent of Monthly Billed kW 4am	pctm4	36	27	0.02	313
Energy Usage As Percent of Monthly Billed kW 5am	pctm5	36	27	0.02	313
Energy Usage As Percent of Monthly Billed kW 6am	pctm6	38	27	0.02	319
Energy Usage As Percent of Monthly Billed kW 7am	pctm7	42	28	0.02	316
Energy Usage As Percent of Monthly Billed kW 8am	pctm8	47	28	0.02	328
Energy Usage As Percent of Monthly Billed kW 9am	pctm9	53	28	0.02	350

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Energy Usage As Percent of Monthly Billed kW 10am	pctm10	57	28	0.02	400
Energy Usage As Percent of Monthly Billed kW 11am	pctm11	60	28	0.02	618
Energy Usage As Percent of Monthly Billed kW noon	pctm12	61	28	0.02	760
Energy Usage As Percent of Monthly Billed kW 1pm	pctm13	61	29	0.02	764
Energy Usage As Percent of Monthly Billed kW 2pm	pctm14	61	29	0.02	774
Energy Usage As Percent of Monthly Billed kW 3pm	pctm15	61	29	0.02	785
Energy Usage As Percent of Monthly Billed kW 4pm	pctm16	60	29	0.02	569
Energy Usage As Percent of Monthly Billed kW 5pm	pctm17	57	29	0.02	372
Energy Usage As Percent of Monthly Billed kW 6pm	pctm18	53	30	0.02	410
Energy Usage As Percent of Monthly Billed kW 7pm	pctm19	50	31	0.02	362
Energy Usage As Percent of Monthly Billed kW 8pm	pctm20	47	31	0.02	347
Energy Usage As Percent of Monthly Billed kW 9pm	pctm21	45	31	0.02	347
Energy Usage As Percent of Monthly Billed kW 10pm	pctm22	43	31	0.02	338
Energy Usage As Percent of Monthly Billed kW 11pm	pctm23	41	29	0.02	327
Energy Usage As Percent of Monthly Billed kW midnight	pctm24	38	28	0.02	326
Energy Usage As Percent of Annual Max Billed kW 1 am	pcta1	29	24	0.02	98
Energy Usage As Percent of Annual Max Billed kW 2am	pcta2	28	23	0.02	98
Energy Usage As Percent of Annual Max Billed kW 3am	pcta3	28	23	0.02	98
Energy Usage As Percent of Annual Max Billed kW 4am	pcta4	28	23	0.02	98
Energy Usage As Percent of Annual Max Billed kW 5am	pcta5	28	23	0.02	98
Energy Usage As Percent of Annual Max Billed kW 6am	pcta6	30	23	0.02	98
Energy Usage As Percent of Annual Max Billed kW 7am	pcta7	33	24	0.02	99
Energy Usage As Percent of Annual Max Billed kW 8am	pcta8	37	24	0.02	100
Energy Usage As Percent of Annual Max Billed kW 9am	pcta9	42	25	0.02	99
Energy Usage As Percent of Annual Max Billed kW 10am	pcta10	45	25	0.02	99
Energy Usage As Percent of Annual Max Billed kW 11am	pcta11	47	25	0.02	104

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Energy Usage As Percent of Annual Max Billed kW noon	pcta12	48	25	0.02	108
Energy Usage As Percent of Annual Max Billed kW 1pm	pcta13	47	25	0.02	110
Energy Usage As Percent of Annual Max Billed kW 2pm	pcta14	48	26	0.02	104
Energy Usage As Percent of Annual Max Billed kW 3pm	pcta15	47	26	0.02	106
Energy Usage As Percent of Annual Max Billed kW 4pm	pcta16	46	26	0.02	106
Energy Usage As Percent of Annual Max Billed kW 5pm	pcta17	44	25	0.02	101
Energy Usage As Percent of Annual Max Billed kW 6pm	pcta18	41	26	0.02	101
Energy Usage As Percent of Annual Max Billed kW 7pm	pcta19	39	26	0.02	104
Energy Usage As Percent of Annual Max Billed kW 8pm	pcta20	37	27	0.02	102
Energy Usage As Percent of Annual Max Billed kW 9pm	pcta21	35	27	0.02	100
Energy Usage As Percent of Annual Max Billed kW 10pm	pcta22	34	26	0.02	101
Energy Usage As Percent of Annual Max Billed kW 11pm	pcta23	32	25	0.02	99
Energy Usage As Percent of Annual Max Billed kW midnight	pcta24	30	24	0.02	98
Energy Usage As Percent of Prior Interval 1 am	delt1	-1	48	-100	3,150
Energy Usage As Percent of Prior Interval 2am	delt2	1	40	-100	2,850
Energy Usage As Percent of Prior Interval 3am	delt3	1	41	-100	2,650
Energy Usage As Percent of Prior Interval 4am	delt4	2	43	-99	2,850
Energy Usage As Percent of Prior Interval 5am	delt5	5	68	-98	14,437
Energy Usage As Percent of Prior Interval 6am	delt6	15	84	-97	5,500
Energy Usage As Percent of Prior Interval 7am	delt7	31	138	-99	12,300
Energy Usage As Percent of Prior Interval 8am	delt8	91	1,309	-100	154,600
Energy Usage As Percent of Prior Interval 9am	delt9	47	353	-99	27,600
Energy Usage As Percent of Prior Interval 10am	delt10	24	141	-99	12,600
Energy Usage As Percent of Prior Interval 11am	delt11	14	140	-100	15,640
Energy Usage As Percent of Prior Interval noon	delt12	11	736	-99	188,867
Energy Usage As Percent of Prior Interval 1pm	delt13	4	83	-100	9,500

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Energy Usage As Percent of Prior Interval 2pm	delt14	4	112	-100	21,700
Energy Usage As Percent of Prior Interval 3pm	delt15	4	222	-100	53,300
Energy Usage As Percent of Prior Interval 4pm	delt16	2	158	-100	26,400
Energy Usage As Percent of Prior Interval 5pm	delt17	-1	155	-100	41,740
Energy Usage As Percent of Prior Interval 6pm	delt18	-3	483	-100	136,700
Energy Usage As Percent of Prior Interval 7pm	delt19	-4	226	-100	67,000
Energy Usage As Percent of Prior Interval 8pm	delt20	-5	167	-100	52,300
Energy Usage As Percent of Prior Interval 9pm	delt21	-2	175	-100	53,100
Energy Usage As Percent of Prior Interval 10pm	delt22	-2	40	-100	4,188
Energy Usage As Percent of Prior Interval 11pm	delt23	-3	57	-99	13,367
Energy Usage As Percent of Prior Interval midnight	delt24	-3	42	-99	2,900
Dry Bulb Temperature 1am	tmp1	54	17	10	89
Dry Bulb Temperature 2am	tmp2	53	17	9	89
Dry Bulb Temperature 3am	tmp3	53	16	8	86
Dry Bulb Temperature 4am	tmp4	52	16	7	85
Dry Bulb Temperature 5am	tmp5	52	16	7	85
Dry Bulb Temperature 6am	tmp6	52	16	6	85
Dry Bulb Temperature 7am	tmp7	52	16	6	85
Dry Bulb Temperature 8am	tmp8	52	17	6	87
Dry Bulb Temperature 9am	tmp9	54	17	8	90
Dry Bulb Temperature 10am	tmp10	55	17	10	93
Dry Bulb Temperature 11am	tmp11	57	18	10	97
Dry Bulb Temperature noon	tmp12	59	18	13	100
Dry Bulb Temperature 1pm	tmp13	60	18	15	102
Dry Bulb Temperature 2pm	tmp14	61	18	17	103
Dry Bulb Temperature 3pm	tmp15	61	18	19	102
Dry Bulb Temperature 4pm	tmp16	60	18	19	102
Dry Bulb Temperature 5pm	tmp17	60	18	19	101
Dry Bulb Temperature 6pm	tmp18	59	18	19	101
Dry Bulb Temperature 7pm	tmp19	58	17	19	99
Dry Bulb Temperature 8pm	tmp20	57	17	18	97
Dry Bulb Temperature 9pm	tmp21	56	17	17	94
Dry Bulb Temperature 10pm	tmp22	55	17	15	92
Dry Bulb Temperature 11pm	tmp23	55	17	13	90
Dry Bulb Temperature midnight	tmp24	54	17	11	89
Cooling Degree Hours 1am	cdh1	2.42	4.37	0	24
Cooling Degree Hours 2am	cdh2	2.25	4.14	0	24
Cooling Degree Hours 3am	cdh3	2.06	3.88	0	21
Cooling Degree Hours 4am	cdh4	1.91	3.67	0	20

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Cooling Degree Hours 5am	cdh5	1.73	3.42	0	20
Cooling Degree Hours 6am	cdh6	1.64	3.35	0	20
Cooling Degree Hours 7am	cdh7	1.70	3.40	0	20
Cooling Degree Hours 8am	cdh8	2.07	3.91	0	22
Cooling Degree Hours 9am	cdh9	2.58	4.58	0	25
Cooling Degree Hours 10am	cdh10	3.19	5.37	0	28
Cooling Degree Hours 11am	cdh11	4.02	6.37	0	32
Cooling Degree Hours noon	cdh12	4.96	7.45	0	35
Cooling Degree Hours 1pm	cdh13	5.50	8.08	0	37
Cooling Degree Hours 2pm	cdh14	5.82	8.40	0	38
Cooling Degree Hours 3pm	cdh15	5.86	8.40	0	37
Cooling Degree Hours 4pm	cdh16	5.52	8.07	0	37
Cooling Degree Hours 5pm	cdh17	5.16	7.72	0	36
Cooling Degree Hours 6pm	cdh18	4.71	7.35	0	36
Cooling Degree Hours 7pm	cdh19	4.16	6.83	0	34
Cooling Degree Hours 8pm	cdh20	3.69	6.23	0	32
Cooling Degree Hours 9pm	cdh21	3.31	5.68	0	29
Cooling Degree Hours 10pm	cdh22	3.04	5.29	0	27
Cooling Degree Hours 11pm	cdh23	2.85	4.9	0	25
Cooling Degree Hours midnight	cdh24	2.64	4.66	0	24
Heating Degree Hours 1am	hdh1	13.69	13.81	0	55
Heating Degree Hours 2am	hdh2	13.97	13.91	0	56
Heating Degree Hours 3am	hdh3	14.29	14.04	0	57
Heating Degree Hours 4am	hdh4	14.58	14.19	0	58
Heating Degree Hours 5am	hdh5	14.74	14.21	0	58
Heating Degree Hours 6am	hdh6	14.97	14.31	0	59
Heating Degree Hours 7am	hdh7	15.00	14.38	0	59
Heating Degree Hours 8am	hdh8	14.66	14.38	0	59
Heating Degree Hours 9am	hdh9	13.90	14.14	0	57
Heating Degree Hours 10am	hdh10	12.94	13.64	0	55
Heating Degree Hours 11am	hdh11	11.96	13.24	0	55
Heating Degree Hours noon	hdh12	11.11	12.75	0	52
Heating Degree Hours 1pm	hdh13	10.44	12.33	0	50
Heating Degree Hours 2pm	hdh14	10.04	12.17	0	48
Heating Degree Hours 3pm	hdh15	9.91	12.01	0	46
Heating Degree Hours 4pm	hdh16	10.05	12.05	0	46
Heating Degree Hours 5pm	hdh17	10.40	12.25	0	46
Heating Degree Hours 6pm	hdh18	10.86	12.44	0	46
Heating Degree Hours 7pm	hdh19	11.36	12.34	0	46
Heating Degree Hours 8pm	hdh20	11.83	12.87	0	47
Heating Degree Hours 9pm	hdh21	12.22	13.07	0	48
Heating Degree Hours 10pm	hdh22	12.59	13.22	0	50
Heating Degree Hours 11pm	hdh23	12.98	13.42	0	52
Heating Degree Hours midnight	hdh24	13.34	13.60	0	54
Humidity 1am	hum1	70	18	19	100
Humidity 2am	hum2	71	17	22	100
Humidity 3am	hum3	72	17	23	100
Humidity 4am	hum4	73	17	25	100
Humidity 5am	hum5	74	17	23	100
Humidity 6am	hum6	74	17	27	100
Humidity 7am	hum7	16	29	29	100



Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Humidity 8am	hum8	73	16	23	100
Humidity 9am	hum9	70	17	33	100
Humidity 10am	hum10	66	17	17	100
Humidity 11am	hum11	62	18	20	100
Humidity noon	hum12	59	19	23	100
Humidity 1pm	hum13	57	20	23	100
Humidity 2pm	hum14	56	20	21	100
Humidity 3pm	hum15	55	20	19	100
Humidity 4pm	hum16	56	21	21	100
Humidity 5pm	hum17	58	20	19	100
Humidity 6pm	hum18	59	20	19	100
Humidity 7pm	hum19	61	20	21	100
Humidity 8pm	hum20	63	20	24	100
Humidity 9pm	hum21	64	20	24	100
Humidity 10pm	hum22	66	19	25	100
Humidity 11pm	hum23	67	19	17	100
Humidity midnight	hum24	69	18	15	100
Wind Speed 1am	wsp1	5	4	0	18
Wind Speed 2am	wsp2	5	4	0	23
Wind Speed 3am	wsp3	5	4	0	20
Wind Speed 4am	wsp4	5	4	0	20
Wind Speed 5am	wsp5	5	4	0	20
Wind Speed 6am	wsp6	5	4	0	23
Wind Speed 7am	wsp7	5	4	0	18
Wind Speed 8am	wsp8	5	4	0	20
Wind Speed 9am	wsp9	6	4	0	22
Wind Speed 10am	wsp10	6	4	0	16
Wind Speed 11am	wsp11	6	4	0	18
Wind Speed noon	wsp12	7	4	0	17
Wind Speed 1pm	wsp13	7	4	0	18
Wind Speed 2pm	wsp14	7	4	0	18
Wind Speed 3pm	wsp15	7	4	0	20
Wind Speed 4pm	wsp16	7	4	0	20
Wind Speed 5pm	wsp17	6	4	0	21
Wind Speed 6pm	wsp18	7	4	0	20
Wind Speed 7pm	wsp19	3	4	0	20
Wind Speed 8pm	wsp20	6	4	0	21
Wind Speed 9pm	wsp21	6	4	0	24
Wind Speed 10pm	wsp22	6	4	0	23
Wind Speed 11pm	wsp23	5	4	0	18
Wind Speed midnight	wsp24	5	4	0	17
Cloud Cover 1am	cc1	38	45	0	100
Cloud Cover 2am	cc2	37	43	0	100
Cloud Cover 3am	cc3	39	45	0	100
Cloud Cover 4am	cc4	38	45	0	100
Cloud Cover 5am	cc5	41	46	0	100
Cloud Cover 6am	cc6	40	45	0	100
Cloud Cover 7am	cc7	39	45	0	100
Cloud Cover 8am	cc8	39	45	0	100
Cloud Cover 9am	cc9	36	45	0	100
Cloud Cover 10am	cc10	36	44	0	100

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Cloud Cover 11am	cc11	35	43	0	100
Cloud Cover noon	cc12	34	41	0	100
Cloud Cover 1pm	cc13	38	41	0	100
Cloud Cover 2pm	cc14	38	40	0	100
Cloud Cover 3pm	cc15	37	41	0	100
Cloud Cover 4pm	cc16	37	41	0	100
Cloud Cover 5pm	cc17	37	41	0	100
Cloud Cover 6pm	cc18	38	43	0	100
Cloud Cover 7pm	cc19	38	43	0	100
Cloud Cover 8pm	cc20	39	43	0	100
Cloud Cover 9pm	cc21	38	43	0	100
Cloud Cover 10pm	cc22	37	44	0	100
Cloud Cover 11pm	cc23	39	44	0	100
Cloud Cover midnight	cc24	39	44	0	100
Sunshine Minutes 1am	ssm1	0	0	0	0
Sunshine Minutes 2am	ssm2	0	0	0	0
Sunshine Minutes 3am	ssm3	0	0	0	0
Sunshine Minutes 4am	ssm4	0	0	0	0
Sunshine Minutes 5am	ssm5	4	10	0	37
Sunshine Minutes 6am	ssm6	17	22	0	60
Sunshine Minutes 7am	ssm7	34	26	0	60
Sunshine Minutes 8am	ssm8	37	27	0	60
Sunshine Minutes 9am	ssm9	38	27	0	60
Sunshine Minutes 10am	ssm10	38	26	0	60
Sunshine Minutes 11am	ssm11	39	26	0	60
Sunshine Minutes noon	ssm12	39	25	0	60
Sunshine Minutes 1pm	ssm13	37	25	0	60
Sunshine Minutes 2pm	ssm14	37	24	0	60
Sunshine Minutes 3pm	ssm15	38	24	0	60
Sunshine Minutes 4pm	ssm16	35	24	0	60
Sunshine Minutes 5pm	ssm17	28	25	0	60
Sunshine Minutes 6pm	ssm18	22	26	0	60
Sunshine Minutes 7pm	ssm19	15	22	0	60
Sunshine Minutes 8pm	ssm20	3	8	0	31
Sunshine Minutes 9pm	ssm21	0	0	0	0
Sunshine Minutes 10pm	ssm22	0	0	0	0
Sunshine Minutes 11pm	ssm23	0	0	0	0
Sunshine Minutes midnight	ssm24	0	0	0	0
Billed Energy January	c1	119,769	148,848	0	772,800
Billed Energy Febuary	c2	112,574	135,536	0	769,600
Billed Energy March	c3	108,366	131,059	0	722,400
Billed Energy April	c4	103,557	126,657	0	704,800
Billed Energy May	c5	107,617	132,545	0	751,200
Billed Energy June	c6	126,516	152,503	0	905,600
Billed Energy July	c7	140,941	167,306	0	891,200
Billed Energy August	c8	141,726	162,790	0	864,800
Billed Energy September	c9	135,134	159,739	0	955,200
Billed Energy October	c10	121,990	146,192	0	824,800
Billed Energy November	c11	111,787	137,510	0	751,200
Billed Energy December	c12	113,720	140,880	0	866,400
Billed Demand January	d1	241.13	247.86	0	1,084.32

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
Billed Demand Febuary	d2	237.58	243.86	0	1,090.56
Billed Demand March	d3	240.23	244.92	0	1,101.60
Billed Demand April	d4	242.09	248.88	0	1,127.52
Billed Demand May	d5	260.39	269.81	0.60	1,197.60
Billed Demand June	d6	307.65	316.00	2.16	1,364.64
Billed Demand July	d7	321.66	327.81	2.34	1,411.20
Billed Demand August	d8	326.48	332.52	0	1,409.28
Billed Demand September	d9	303.87	309.01	0	1,341.12
Billed Demand October	d10	291.80	294.77	0	1,332.48
Billed Demand November	d11	259.98	265.32	0	1,223.52
Billed Demand December	d12	245.31	255.12	0	1,200.48
Lot Footprint	footprint	187,830	714,780	875	9,150,700
Number of Buildings	no_bldgs	1.66	2.65	0	26
Year Built	year_built	1948	28	1890	2009
Number of Floors	no_floors	5	4	0	32
Building Area	bldg_area	265,474	598,570	0	4,687,440
Number of Units	no_units	29	97	0	1,201
Number of Residential Units	no_res_units	15	88	0	1,199
Floor-Area Ratio	far	2.67	2.25	0	13.65
Building Area Per Unit	bapu	79,487	164,309	0	1,675,000
Building Area Per Residential Unit	bapru	10,747	89,802	0	781,240
Children Under 5 As Percent of Census Tract Population	kidlt5	5.79	2.01	0	11.64
Children Between 5 and 9 As Percent of Census Tract Population	kid59	3.65	2.08	0	14.38
Children Between 10 and 14 As Percent of Census Tract Population	kid1014	4.36	3.06	0	12.41
Children Between 15 and 19 As Percent of Census Tract Population	kid1519	5.07	3.81	0	27.93
Children Under 10 As Percent of Census Tract Population	kidle9	9.43	2.76	0	22.97
Children Under 15 As Percent of Census Tract Population	kidle14	13.80	4.38	0	29.13
Children Under 20 As Percent of Census Tract Population	kidle19	18.86	6.38	0	39.35
Median Age of Census Tract Population	medage	35.51	4.99	21.50	54.40
People 65 And Over As Percent of Census Tract Population	srcit	9.52	6.02	0	35.52
Median Household Income of Census Tract	medinc	65,380	23,319	16,063	123,929
People Who Are One Race - White As Percent of Census Tract Population	white	52.22	23.73	0.28	96.66
People Who Are One Race - Black As Percent of Census Tract Population	black	10.78	17.18	0	96.43

Variable	Variable Name	Mean Value	Standard Deviation	Minimum Value	Maximum Value
People Who Are One Race - Asian As Percent of Census Tract Population	asian	21.02	14.56	0	7.11
People Who Are One Race - Latino As Percent of Census Tract Population	latino	30.08	20.52	0	82.62
People Of American Ancestry As Percent of Census Tract Population	america	2.98	2.98	0	16.89
People Of Guyanese Ancestry As Percent of Census Tract Population	guyana	0.94	2.88	0	23.57
People Of Irish Ancestry As Percent of Census Tract Population	ireland	4.34	4.13	0	16.29
People Of Italian Ancestry As Percent of Census Tract Population	italy	6.95	6.81	0	57.56
People Of Polish Ancestry As Percent of Census Tract Population	poland	2.05	3.10	0	33.44
People Of West Indian Ancestry As Percent of Census Tract Population	windies	2.38	6.72	0	45.93
People Of Jamaican Ancestry As Percent of Census Tract Population	jamaica	0.78	2.89	0	28.70
People Who Are Foreign Born As Percent of Census Tract Population	foreign	41.51	13.71	15.35	100
Percent of Occupied Housing Units With Electric Heat	elecheat	12.96	16.08	0	58.10
Average Household Size	hhsz	2.89	1.09	1.82	5.62
People With Less Than a High School Education As Percent of Census Tract Population	edlths	16.30	13.07	0	46.90
People With a High School Diploma As Percent of Census Tract Population	edhs	23.46	11.09	0	49.50
People With Some College Education As Percent of Census Tract Population	edsmcoll	18.59	7.20	7.30	41.50
People With a College Diploma or More As Percent of Census Tract Population	edcoll	41.65	23.21	7.10	80.20
People With a High School Diploma or Less As Percent of Census Tract Population	edlehs	39.76	20.71	8.70	75.00
People With Some College or Less As Percent of Census Tract Population	edltcoll	58.35	23.22	19.70	93.00

## Appendix B: Correlation Coefficients for Data Groups

This appendix provides additional detail beyond the summary of correlation coefficients provided in the text. In this appendix, correlation coefficient ranges are provided for groups of data elements. The complete set of correlation coefficients is provided in two companion comma-separated data files to this thesis, called "graves\_correlations\_STRno.csv" and "graves\_correlations\_STRyes.csv".

Each table below provides results for a different set of potential dependent variables, including the raw energy usage intervals and their transformations. The data elements, both dependent and independent, are grouped together if there are individual data elements for different hours of the day or months of the year.

**Table B.1: Correlation Coefficient Ranges for Residential Customer Data Groups - Raw Energy Usage Intervals (k1-k24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.02	-0.01	-0.04	0.02	-0.06	0.05	-0.06	0.05	-0.14	-0.02	-0.07	0.21	-0.21	-0.06
annkwh	0.88	0.95	0.25	0.31	0.11	0.15	0.07	0.15	0.11	0.23	0.19	0.31	0.67	0.83
asian	-0.03	0.00	-0.10	-0.02	-0.07	0.02	-0.09	0.02	-0.04	0.13	-0.16	0.02	-0.20	0.13
bapru	-0.14	-0.11	-0.01	0.03	-0.06	0.05	-0.02	0.05	0.00	0.11	-0.05	0.27	0.03	0.35
bapu	0.77	0.85	-0.05	-0.01	-0.07	0.05	-0.03	0.08	-0.04	0.10	-0.05	0.26	0.39	0.60
black	0.13	0.20	0.05	0.12	0.00	0.07	-0.04	0.02	-0.03	0.07	-0.09	0.07	0.12	0.37
bldg_area	-0.06	-0.05	-0.02	0.06	-0.03	0.09	-0.04	0.01	-0.13	-0.01	-0.04	0.12	0.42	0.61
c1-c12	0.85	0.95	0.12	0.25	-0.02	0.09	-0.05	0.11	-0.07	0.21	-0.25	0.30	0.60	0.83
cc1-cc24	-0.01	0.00	-0.10	0.00	-0.12	-0.01	-0.14	-0.01	-0.12	-0.01	-0.11	0.02	-0.07	0.00
cdh1-cdh24	0.01	0.03	0.16	0.39	0.24	0.48	0.29	0.51	0.28	0.46	0.01	0.20	0.05	0.19
d1-d12	0.57	0.84											0.40	0.78
daytype	-0.01	0.00	-0.06	0.00	-0.07	0.01	-0.09	0.01	-0.06	0.01	-0.06	0.03	-0.07	0.09
edcoll	-0.09	-0.07	0.05	0.13	-0.03	0.09	-0.03	0.09	-0.16	0.00	-0.05	0.26	-0.14	0.03
edhs	0.11	0.16	-0.09	-0.02	-0.09	0.03	-0.04	0.05	0.00	0.12	-0.15	0.11	-0.01	0.27
edlehs	0.04	0.06	-0.12	-0.04	-0.09	0.03	-0.08	0.04	0.02	0.14	-0.27	0.07	-0.15	0.06
editcoll	0.07	0.09	-0.13	-0.05	-0.09	0.03	-0.09	0.03	0.00	0.16	-0.26	0.05	-0.02	0.14
edlths	-0.02	-0.02	-0.11	-0.03	-0.08	0.04	-0.09	0.02	-0.02	0.13	-0.25	0.04	-0.35	-0.28
edsmcoll	0.12	0.13	-0.12	-0.05	-0.08	0.01	-0.10	0.02	-0.06	0.16	-0.21	-0.02	0.10	0.23
elecheat	-0.08	-0.07	0.05	0.13	-0.01	0.09	-0.04	0.06	-0.12	-0.02	-0.11	0.13	-0.19	0.03
far	-0.10	-0.09	0.02	0.08	-0.03	0.07	-0.02	0.02	-0.10	0.00	-0.05	0.09	0.10	0.24
footprint	0.00	0.01	0.00	0.05	-0.01	0.04	-0.03	0.02	-0.09	0.02	0.00	0.17	0.27	0.45
foreign	-0.03	-0.02	-0.11	-0.03	-0.08	0.03	-0.10	0.04	-0.03	0.14	-0.25	0.04	-0.25	-0.09
guyana	0.05	0.06	0.00	0.06	-0.03	0.02	-0.04	0.02	-0.04	0.09	-0.09	0.06	-0.23	-0.08
hdh1-hdh24	-0.01	0.00	-0.22	-0.04	-0.30	-0.08	-0.33	-0.10	-0.28	-0.07	0.06	0.22	-0.11	0.07
hhsz	0.13	0.17	-0.05	0.01	-0.04	0.04	-0.09	-0.01	-0.08	0.14	-0.13	0.09	-0.17	0.03
houses	-0.25	-0.13	-0.07	-0.03	-0.08	0.06	-0.05	0.04	-0.07	0.05	-0.22	0.12	0.11	0.39
hum9	0.00	0.01	-0.01	0.08	0.00	0.12	-0.02	0.12	-0.01	0.12	-0.11	0.01	-0.03	0.05
ireland	-0.02	0.00	-0.01	0.05	-0.03	0.06	-0.03	0.04	-0.12	-0.02	-0.12	0.18	-0.04	0.05
italy	-0.06	-0.03	0.01	0.08	-0.01	0.04	-0.01	0.07	-0.06	0.02	-0.07	0.19	-0.04	0.20
jamaica	0.16	0.26	0.02	0.08	-0.01	0.05	-0.03	0.01	-0.03	0.08	-0.09	0.06	0.03	0.28
k1-k24	0.82	1.00	0.27	0.91	0.24	0.92	0.23	0.93	0.31	0.93	0.33	0.96	0.56	0.99
kid1014	0.19	0.23	-0.12	-0.04	-0.08	0.02	-0.08	0.03	-0.03	0.18	-0.09	0.20	0.12	0.35
kid1519	0.11	0.14	-0.09	-0.03	-0.07	0.02	-0.03	0.02	0.01	0.12	-0.12	0.10	-0.22	-0.09

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	0.13	0.17	-0.10	-0.02	-0.08	0.04	-0.07	0.02	-0.02	0.20	-0.12	0.06	0.19	0.40
kidle14	0.29	0.33	-0.08	0.00	-0.06	0.04	-0.09	0.00	-0.05	0.20	-0.13	0.08	0.20	0.31
kidle19	0.27	0.31	-0.09	-0.02	-0.07	0.04	-0.07	0.01	-0.02	0.20	-0.10	0.08	0.16	0.28
kidle9	0.24	0.30	-0.03	0.03	-0.03	0.05	-0.07	0.00	-0.07	0.17	-0.15	-0.05	0.12	0.26
kidlt5	0.16	0.21	0.02	0.08	0.00	0.07	-0.05	0.01	-0.10	0.02	-0.14	-0.03	0.06	0.16
latino	-0.01	0.00	-0.10	-0.02	-0.08	0.04	-0.07	0.04	0.03	0.14	-0.26	0.11	-0.24	-0.09
maxdmd	0.66	0.77											0.49	0.67
medage	-0.07	-0.03	-0.13	-0.08	-0.07	-0.02	-0.07	0.03	-0.02	0.05	-0.08	0.16	-0.19	-0.04
medinc	-0.07	-0.06	0.04	0.12	-0.03	0.10	-0.05	0.05	-0.13	-0.02	-0.15	0.23	-0.40	-0.23
no_bldgs	-0.03	-0.02	-0.10	-0.04	-0.05	0.03	-0.08	0.02	-0.02	0.10	-0.05	0.21	0.08	0.38
no_floors	-0.09	-0.08	0.02	0.11	-0.03	0.09	-0.04	0.04	-0.13	-0.03	-0.05	0.13	0.14	0.39
no_res_units	-0.13	-0.11	-0.02	0.04	-0.03	0.08	-0.05	0.02	-0.12	-0.01	-0.07	0.13	0.03	0.35
no_units	-0.13	-0.11	-0.02	0.04	-0.03	0.08	-0.05	0.03	-0.13	-0.01	-0.08	0.13	-0.10	0.04
occhsng	-0.19	-0.08	-0.08	-0.04	-0.08	0.05	-0.07	0.04	-0.04	0.04	-0.22	0.12	0.12	0.41
poland	-0.07	-0.05	-0.01	0.06	-0.04	0.07	-0.05	0.04	-0.18	-0.01	-0.09	0.17	-0.13	0.02
popn	-0.06	0.00	-0.08	-0.02	-0.06	0.03	-0.10	0.01	-0.04	0.11	-0.25	0.05	0.10	0.38
srcit	0.08	0.10	-0.12	-0.08	-0.07	-0.02	-0.08	0.03	-0.06	0.06	-0.08	0.15	-0.22	-0.07
ssm1-ssm24	0.00	0.02	0.02	0.24	0.03	0.30	0.03	0.33	0.03	0.29	-0.10	0.11	-0.02	0.13
tmp1-tmp24	0.01	0.02	0.10	0.28	0.16	0.37	0.18	0.39	0.18	0.34	-0.16	-0.01	-0.02	0.13
white	-0.15	-0.13	-0.01	0.06	-0.03	0.04	0.00	0.08	-0.15	0.07	0.00	0.17	-0.32	-0.11
windies	0.15	0.22	0.02	0.09	-0.02	0.05	-0.03	0.01	-0.03	0.09	-0.08	0.07	0.08	0.32
wsp1-wsp24	-0.01	0.00	-0.11	-0.01	-0.14	-0.01	-0.16	-0.02	-0.14	-0.02	-0.06	0.06	-0.08	0.02
year_built	-0.01	0.00	0.03	0.11	-0.03	0.09	-0.09	0.08	-0.13	0.00	0.05	0.21	0.32	0.54

**Table B.2: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Daily Maximum Value (pctd1-pctd24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.05	0.03	-0.05	0.03	-0.11	0.05	-0.07	0.05	-0.06	0.07	-0.37	-0.04	0.00	0.23
annkwh	-0.03	0.18	-0.04	0.00	-0.03	0.03	-0.05	0.04	-0.01	0.12	-0.08	0.16	-0.01	0.27
asian	-0.06	0.11	-0.09	0.08	-0.05	0.16	-0.02	0.16	-0.04	0.15	0.08	0.38	-0.27	0.10
bapru	-0.04	0.01	-0.05	0.02	-0.11	0.04	-0.03	0.05	0.00	0.16	-0.14	0.23	0.08	0.34
bapu	-0.03	0.16	-0.04	0.02	-0.09	0.06	-0.05	0.05	-0.01	0.14	-0.07	0.28	-0.06	0.26
black	-0.05	0.02	-0.07	0.00	-0.07	0.03	-0.08	0.01	-0.01	0.13	-0.01	0.18	-0.19	0.17
bldg_area	-0.07	0.00	-0.07	0.00	-0.12	0.06	-0.05	0.03	-0.07	0.07	-0.19	0.03	-0.08	0.24
c1-c12	-0.03	0.18	-0.08	0.07	-0.07	0.06	-0.10	0.08	-0.22	0.27	-0.28	0.29	-0.06	0.34
cc1-cc24	-0.05	0.03	-0.04	0.04	-0.05	0.04	-0.06	0.03	-0.05	0.02	-0.08	0.05	-0.03	0.05
cdh1-cdh24	-0.09	0.16	-0.11	0.14	-0.10	0.18	-0.06	0.20	-0.01	0.19	-0.06	0.21	-0.17	-0.01
d1-d12	-0.20	0.37											-0.32	0.10
daytype	-0.06	0.03	-0.06	0.03	-0.06	0.04	-0.07	0.05	-0.06	0.03	-0.09	0.04	-0.15	0.11
edcoll	-0.13	0.03	-0.12	0.06	-0.17	0.05	-0.17	0.01	-0.13	0.06	-0.42	-0.03	-0.03	0.24
edhs	0.02	0.10	0.03	0.10	-0.04	0.13	0.00	0.12	-0.07	0.06	-0.06	0.25	-0.21	0.16
edlehs	-0.03	0.13	-0.05	0.11	-0.06	0.18	-0.01	0.16	-0.04	0.09	0.00	0.44	-0.20	0.08
editcoll	-0.03	0.13	-0.06	0.12	-0.05	0.17	-0.01	0.17	-0.06	0.13	0.03	0.42	-0.24	0.03
edlths	-0.06	0.12	-0.09	0.10	-0.05	0.17	-0.02	0.17	-0.06	0.13	0.02	0.41	-0.20	0.05
edsmcoll	-0.03	0.12	-0.07	0.11	-0.03	0.16	0.00	0.19	-0.07	0.20	0.12	0.36	-0.21	-0.04
elecheat	-0.11	0.00	-0.11	0.01	-0.14	0.05	-0.14	0.00	-0.04	0.08	-0.27	0.00	-0.27	-0.14
far	-0.04	0.03	-0.03	0.05	-0.07	0.09	-0.06	0.03	-0.01	0.12	-0.13	0.03	0.07	0.24
footprint	-0.09	-0.04	-0.08	-0.03	-0.12	-0.03	-0.10	-0.01	-0.10	0.02	-0.24	0.00	-0.10	0.29
foreign	-0.07	0.11	-0.10	0.09	-0.05	0.17	-0.02	0.17	-0.06	0.14	0.01	0.43	-0.16	0.00
guyana	-0.02	0.04	-0.02	0.05	-0.03	0.05	-0.02	0.06	-0.03	0.12	-0.06	0.16	-0.29	0.02
hdh1-hdh24	-0.11	0.11	-0.09	0.11	-0.12	0.12	-0.15	0.10	-0.13	0.06	-0.08	0.15	0.05	0.20
hhsz	-0.03	0.11	-0.10	0.06	-0.03	0.13	-0.02	0.14	-0.05	0.20	0.11	0.38	-0.27	0.10
houses	-0.06	0.07	-0.01	0.06	-0.07	0.13	-0.02	0.11	-0.11	0.05	-0.19	0.26	-0.11	0.37
hum9	-0.04	0.03	-0.05	0.02	-0.05	0.04	-0.04	0.05	-0.02	0.06	-0.07	0.03	-0.10	-0.01
ireland	-0.09	-0.01	-0.10	-0.02	-0.13	0.03	-0.12	0.01	-0.07	0.03	-0.34	0.05	-0.04	0.11
italy	-0.07	0.04	-0.06	0.10	-0.11	0.02	-0.13	-0.01	-0.03	0.07	-0.32	0.07	-0.24	0.07
jamaica	-0.03	0.03	-0.05	0.01	-0.06	0.02	-0.06	0.02	-0.02	0.13	-0.05	0.17	-0.20	0.17
pctd1-pctd24	0.11	0.88	0.11	0.86	0.08	0.89	0.09	0.91	0.11	0.92	0.11	0.94	-0.19	0.99
kid1014	0.02	0.13	-0.03	0.11	-0.03	0.15	-0.01	0.16	-0.05	0.21	0.03	0.37	-0.31	-0.02
kid1519	0.03	0.11	0.03	0.11	-0.02	0.11	0.02	0.12	-0.07	0.05	-0.14	0.14	-0.30	-0.01



Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	-0.01	0.13	-0.04	0.12	-0.05	0.18	-0.02	0.15	-0.08	0.19	0.14	0.36	-0.04	0.23
kidle14	0.01	0.12	-0.07	0.09	-0.04	0.16	-0.03	0.15	-0.06	0.23	0.14	0.38	0.02	0.16
kidle19	0.02	0.14	-0.04	0.11	-0.04	0.16	-0.01	0.15	-0.06	0.20	0.14	0.36	-0.04	0.14
kidle9	0.00	0.08	-0.08	0.06	-0.04	0.12	-0.03	0.10	-0.08	0.20	0.10	0.27	0.07	0.17
kidlt5	-0.03	0.01	-0.11	-0.04	-0.06	0.03	-0.07	-0.02	-0.03	0.07	-0.04	0.13	0.03	0.16
latino	-0.04	0.11	-0.05	0.10	-0.05	0.16	-0.02	0.14	-0.05	0.06	-0.08	0.42	-0.24	-0.11
maxdmd	-0.12	0.31											-0.31	0.05
medage	-0.02	0.08	-0.03	0.09	-0.02	0.11	0.01	0.17	-0.01	0.10	-0.01	0.27	-0.22	-0.08
medinc	-0.11	0.01	-0.11	0.01	-0.15	0.05	-0.14	0.00	-0.03	0.08	-0.38	0.14	-0.29	0.00
no_bldgs	-0.07	0.08	-0.09	0.08	-0.04	0.13	-0.02	0.15	0.01	0.12	0.01	0.33	-0.23	0.15
no_floors	-0.10	0.00	-0.09	0.02	-0.15	0.05	-0.12	0.00	-0.06	0.06	-0.22	0.02	0.03	0.25
no_res_units	-0.08	-0.03	-0.08	-0.02	-0.13	0.05	-0.06	0.03	-0.05	0.10	-0.18	0.06	0.08	0.34
no_units	-0.09	-0.03	-0.08	-0.03	-0.14	0.04	-0.08	0.02	-0.05	0.11	-0.20	0.06	-0.13	0.11
occhsng	-0.05	0.09	-0.03	0.07	-0.05	0.15	-0.02	0.13	-0.07	0.07	-0.16	0.34	-0.11	0.37
poland	-0.11	0.00	-0.09	-0.01	-0.15	0.05	-0.12	0.01	-0.13	0.07	-0.37	0.00	0.07	0.25
popn	-0.05	0.10	-0.08	0.09	-0.04	0.16	-0.03	0.15	-0.01	0.15	-0.02	0.42	-0.11	0.34
srcit	0.00	0.09	-0.03	0.08	-0.01	0.11	0.01	0.19	-0.04	0.13	-0.01	0.28	-0.14	0.07
ssm1-ssm24	-0.10	0.11	-0.10	0.11	-0.11	0.12	-0.08	0.13	-0.06	0.12	-0.12	0.11	-0.15	0.03
tmp1-tmp24	-0.11	0.13	-0.12	0.11	-0.12	0.14	-0.09	0.17	-0.04	0.15	-0.13	0.11	-0.20	-0.04
white	-0.09	0.03	-0.06	0.10	-0.12	0.02	-0.14	0.01	-0.19	0.05	-0.31	-0.07	-0.05	0.21
windies	-0.03	0.03	-0.05	0.01	-0.05	0.02	-0.06	0.02	-0.02	0.14	-0.01	0.17	-0.16	0.19
wsp1-wsp24	-0.04	0.05	-0.04	0.05	-0.05	0.05	-0.07	0.04	-0.07	0.02	-0.06	0.08	-0.03	0.10
year_built	-0.11	0.04	-0.06	0.07	-0.19	0.04	-0.17	0.00	-0.15	-0.01	-0.31	-0.03	0.05	0.23

**Table B.3: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Monthly Billed kWh (pctm1-pctm24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.03	0.02	-0.03	0.02	-0.04	0.03	-0.06	0.05	-0.09	0.03	-0.17	0.10	-0.05	0.12
annkwh	-0.05	0.05	-0.01	0.01	-0.03	0.00	-0.04	0.05	-0.07	0.08	-0.07	0.05	-0.15	0.21
asian	-0.06	0.01	-0.06	0.00	-0.06	0.01	-0.08	0.03	-0.09	0.09	-0.02	0.15	-0.26	0.12
bapru	-0.02	0.02	-0.02	0.02	-0.05	0.05	-0.02	0.04	-0.01	0.14	-0.10	0.10	-0.08	0.29
bapu	-0.05	0.05	-0.03	0.02	-0.05	0.05	-0.04	0.07	-0.06	0.11	-0.09	0.11	-0.13	0.13
black	-0.02	0.02	-0.02	0.04	-0.01	0.03	-0.03	0.05	-0.04	0.09	-0.04	0.11	-0.07	0.17
bldg_area	-0.02	0.02	-0.03	0.03	-0.03	0.05	-0.03	0.02	-0.10	0.03	-0.07	0.07	-0.14	0.17
c1-c12	-0.06	0.05	-0.06	0.04	-0.07	0.09	-0.09	0.09	-0.12	0.13	-0.23	0.18	-0.16	0.26
cc1-cc24	-0.05	0.01	-0.05	0.01	-0.04	0.01	-0.08	0.00	-0.07	0.00	-0.09	0.03	-0.06	0.02
cdh1-cdh24	0.04	0.22	0.03	0.24	0.04	0.18	0.06	0.33	0.09	0.29	0.02	0.19	-0.04	0.19
d1-d12	-0.15	0.29											-0.31	0.19
daytype	-0.06	0.00	-0.05	0.00	-0.06	0.01	-0.08	0.01	-0.07	0.01	-0.05	0.03	-0.18	0.14
edcoll	-0.01	0.06	-0.01	0.06	-0.02	0.07	-0.04	0.08	-0.13	0.06	-0.20	0.09	-0.09	0.11
edhs	-0.03	0.03	-0.03	0.04	-0.05	0.03	-0.05	0.04	-0.03	0.10	-0.15	0.08	-0.15	0.21
edlehs	-0.06	0.02	-0.06	0.01	-0.07	0.03	-0.08	0.04	-0.03	0.10	-0.09	0.22	-0.14	0.12
edltcoll	-0.06	0.01	-0.06	0.01	-0.07	0.02	-0.08	0.04	-0.06	0.13	-0.09	0.20	-0.11	0.09
edlths	-0.07	0.01	-0.07	0.01	-0.07	0.02	-0.08	0.03	-0.08	0.08	-0.05	0.22	-0.09	0.06
edsmcoll	-0.05	0.01	-0.06	0.01	-0.05	0.02	-0.09	0.03	-0.10	0.16	-0.06	0.12	-0.16	0.07
elecheat	-0.02	0.04	-0.02	0.04	-0.03	0.05	-0.05	0.06	-0.08	0.04	-0.11	0.12	-0.22	0.11
far	-0.01	0.02	-0.01	0.03	-0.03	0.05	-0.03	0.01	-0.09	0.03	-0.06	0.07	-0.05	0.14
footprint	-0.02	0.02	-0.03	0.02	-0.02	0.02	-0.03	0.04	-0.07	0.08	-0.10	0.04	-0.14	0.24
foreign	-0.07	0.01	-0.07	0.01	-0.07	0.02	-0.09	0.04	-0.09	0.08	-0.06	0.21	-0.17	0.03
guyana	-0.02	0.02	-0.02	0.03	-0.02	0.01	-0.03	0.04	-0.04	0.10	-0.08	0.07	-0.24	0.15
hdh1-hdh24	-0.11	0.02	-0.12	0.02	-0.10	0.02	-0.19	0.04	-0.15	0.03	-0.10	0.03	-0.13	0.11
hhsz	-0.04	0.01	-0.05	0.01	-0.04	0.02	-0.07	0.02	-0.11	0.14	0.00	0.15	-0.14	0.14
houses	-0.04	0.04	-0.02	0.02	-0.08	0.04	-0.06	0.03	-0.09	0.05	-0.07	0.16	-0.18	0.15
hum9	-0.02	0.04	-0.02	0.04	-0.01	0.04	-0.04	0.06	-0.03	0.06	-0.04	0.05	-0.06	0.07
ireland	-0.01	0.02	-0.02	0.03	-0.03	0.04	-0.02	0.04	-0.08	0.04	-0.23	0.06	-0.06	0.09
italy	0.00	0.04	0.00	0.05	0.00	0.04	-0.02	0.05	-0.03	0.05	-0.19	0.03	-0.20	0.14
jamaica	-0.02	0.01	-0.02	0.03	-0.01	0.02	-0.02	0.04	-0.04	0.10	-0.06	0.08	-0.07	0.17
pctm1-pctm24	0.30	0.94	0.24	0.92	0.38	0.97	0.17	0.91	0.20	0.90	0.42	0.94	0.01	0.97
kid1014	-0.04	0.02	-0.05	0.01	-0.05	0.03	-0.07	0.03	-0.07	0.18	-0.07	0.13	-0.23	0.10
kid1519	-0.03	0.02	-0.02	0.02	-0.03	0.03	-0.04	0.02	-0.04	0.10	-0.12	0.10	-0.16	0.12

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	-0.04	0.02	-0.05	0.01	-0.06	0.03	-0.07	0.03	-0.08	0.18	-0.04	0.12	-0.07	0.18
kidle14	-0.03	0.01	-0.06	0.01	-0.05	0.02	-0.07	0.02	-0.09	0.19	0.00	0.14	-0.06	0.16
kidle19	-0.03	0.02	-0.05	0.01	-0.05	0.03	-0.07	0.02	-0.07	0.18	-0.01	0.14	-0.06	0.13
kidle9	-0.02	0.01	-0.05	0.01	-0.04	0.02	-0.05	0.02	-0.11	0.16	0.01	0.12	-0.05	0.15
kidlt5	-0.02	0.01	-0.05	0.01	-0.03	0.03	-0.04	0.02	-0.07	0.04	0.02	0.10	-0.04	0.11
latino	-0.05	0.02	-0.06	0.01	-0.06	0.03	-0.07	0.04	-0.03	0.09	-0.07	0.24	-0.19	0.10
maxdmd	-0.15	0.28											-0.27	0.13
medage	-0.04	0.00	-0.04	0.01	-0.03	0.01	-0.08	0.03	-0.04	0.05	-0.10	0.06	-0.19	0.04
medinc	-0.03	0.04	-0.03	0.04	-0.03	0.06	-0.05	0.06	-0.07	0.04	-0.20	0.11	-0.19	0.11
no_bldgs	-0.06	0.00	-0.06	0.01	-0.05	0.01	-0.08	0.02	-0.04	0.08	-0.09	0.08	-0.20	0.14
no_floors	-0.03	0.03	-0.03	0.04	-0.03	0.05	-0.05	0.05	-0.10	0.02	-0.07	0.09	-0.07	0.25
no_res_units	-0.03	0.02	-0.04	0.02	-0.04	0.04	-0.04	0.03	-0.10	0.04	-0.08	0.12	-0.08	0.29
no_units	-0.03	0.02	-0.04	0.02	-0.04	0.04	-0.05	0.04	-0.10	0.04	-0.09	0.12	-0.09	0.12
occhsng	-0.04	0.03	-0.02	0.01	-0.07	0.03	-0.08	0.03	-0.07	0.04	-0.08	0.19	-0.18	0.15
poland	-0.02	0.03	-0.03	0.03	-0.03	0.05	-0.05	0.05	-0.14	0.06	-0.16	0.09	-0.04	0.23
popn	-0.05	0.01	-0.05	0.01	-0.06	0.01	-0.08	0.02	-0.06	0.09	-0.05	0.22	-0.18	0.14
srcit	-0.04	0.00	-0.04	0.01	-0.03	0.00	-0.08	0.03	-0.07	0.07	-0.10	0.06	-0.17	0.07
ssm1-ssm24	-0.02	0.17	-0.02	0.17	-0.02	0.16	-0.02	0.27	0.00	0.22	-0.03	0.15	-0.06	0.15
tmp1-tmp24	0.00	0.15	-0.01	0.16	0.00	0.13	-0.01	0.23	0.02	0.20	0.00	0.12	-0.09	0.16
white	0.00	0.05	0.00	0.06	0.00	0.05	-0.01	0.06	-0.14	0.10	-0.15	0.03	-0.09	0.16
windies	-0.02	0.01	-0.02	0.03	-0.01	0.02	-0.03	0.04	-0.04	0.10	-0.05	0.10	-0.07	0.18
wsp1-wsp24	-0.06	0.02	-0.06	0.02	-0.05	0.02	-0.10	0.03	-0.08	0.03	-0.08	0.03	-0.09	0.06
year_built	-0.03	0.05	-0.01	0.05	-0.03	0.06	-0.10	0.07	-0.15	0.01	-0.10	0.05	-0.06	0.17

**Table B.4: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Annual Billed kWh (pcta1-pcta24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.04	0.03	-0.04	0.02	-0.04	0.06	-0.05	0.06	-0.07	0.03	-0.14	0.20	-0.05	0.11
annkwh	-0.06	0.05	-0.02	0.01	-0.02	0.01	-0.05	0.04	-0.06	0.07	-0.07	0.10	-0.14	0.20
asian	-0.07	0.02	-0.06	0.02	-0.08	0.02	-0.08	0.04	-0.08	0.08	-0.14	0.08	-0.28	0.10
bapru	-0.03	0.04	-0.02	0.05	-0.06	0.05	-0.04	0.04	-0.01	0.11	-0.11	0.18	-0.07	0.28
bapu	-0.05	0.04	-0.03	0.05	-0.07	0.05	-0.04	0.06	-0.06	0.08	-0.11	0.18	-0.13	0.12
black	-0.02	0.03	-0.02	0.04	-0.01	0.06	-0.03	0.04	-0.04	0.07	-0.03	0.15	-0.03	0.23
bldg_area	-0.02	0.03	-0.04	0.03	-0.02	0.10	-0.03	0.02	-0.07	0.04	-0.04	0.13	-0.15	0.13
c1-c12	-0.06	0.05	-0.03	0.05	-0.05	0.07	-0.08	0.07	-0.11	0.13	-0.21	0.18	-0.15	0.25
cc1-cc24	-0.11	0.00	-0.10	0.00	-0.12	-0.01	-0.14	-0.01	-0.13	-0.02	-0.12	0.02	-0.10	0.01
cdh1-cdh24	0.04	0.43	0.03	0.40	0.24	0.47	0.29	0.51	0.28	0.48	0.03	0.26	0.07	0.29
d1-d12	-0.22	0.24											-0.28	0.14
daytype	-0.06	0.01	-0.06	0.00	-0.07	0.01	-0.09	0.01	-0.06	0.01	-0.07	0.03	-0.17	0.14
edcoll	-0.02	0.07	-0.01	0.06	-0.03	0.09	-0.04	0.09	-0.09	0.06	-0.15	0.22	-0.08	0.10
edhs	-0.03	0.03	-0.03	0.04	-0.08	0.03	-0.06	0.03	-0.04	0.07	-0.18	0.10	-0.14	0.19
edlehs	-0.07	0.02	-0.06	0.02	-0.10	0.03	-0.08	0.04	-0.04	0.07	-0.23	0.16	-0.14	0.10
edltcoll	-0.07	0.02	-0.06	0.01	-0.09	0.03	-0.09	0.04	-0.06	0.09	-0.22	0.15	-0.10	0.08
edlths	-0.07	0.02	-0.06	0.02	-0.08	0.03	-0.09	0.03	-0.08	0.05	-0.19	0.15	-0.13	0.03
edsmcoll	-0.06	0.01	-0.06	0.01	-0.06	0.03	-0.09	0.03	-0.10	0.13	-0.15	0.08	-0.13	0.08
elecheat	-0.02	0.05	-0.02	0.05	-0.02	0.09	-0.04	0.07	-0.06	0.04	-0.11	0.18	-0.25	0.06
far	-0.02	0.02	-0.02	0.02	-0.04	0.06	-0.02	0.02	-0.06	0.03	-0.03	0.12	-0.03	0.14
footprint	-0.01	0.03	-0.02	0.03	0.00	0.05	-0.02	0.03	-0.05	0.08	-0.06	0.09	-0.16	0.19
foreign	-0.07	0.02	-0.06	0.02	-0.08	0.03	-0.09	0.05	-0.09	0.07	-0.19	0.16	-0.17	0.01
guyana	-0.02	0.02	-0.02	0.03	-0.02	0.03	-0.04	0.02	-0.04	0.09	-0.06	0.12	-0.25	0.12
hdh1-hdh24	-0.26	-0.02	-0.23	-0.01	-0.30	-0.08	-0.33	-0.10	-0.29	-0.08	0.03	0.22	-0.17	0.09
hhsz	-0.05	0.02	-0.05	0.02	-0.05	0.03	-0.07	0.01	-0.11	0.11	-0.08	0.11	-0.12	0.17
houses	-0.04	0.05	-0.04	0.02	-0.08	0.07	-0.06	0.04	-0.08	0.04	-0.16	0.17	-0.18	0.15
hum9	-0.01	0.09	-0.01	0.08	0.00	0.12	-0.02	0.12	-0.01	0.12	-0.09	0.02	-0.05	0.07
ireland	-0.01	0.03	-0.02	0.03	-0.02	0.06	-0.02	0.05	-0.06	0.03	-0.20	0.15	-0.06	0.08
italy	-0.01	0.04	-0.02	0.05	-0.02	0.04	-0.02	0.06	-0.04	0.04	-0.17	0.12	-0.20	0.09
jamaica	-0.02	0.02	-0.02	0.03	-0.01	0.05	-0.02	0.02	-0.03	0.08	-0.02	0.15	-0.03	0.24
pcta1-pcta24	0.04	0.92	0.03	0.94	0.23	0.92	0.22	0.93	0.29	0.92	0.24	0.95	0.06	0.97
kid1014	-0.04	0.02	-0.05	0.04	-0.08	0.02	-0.08	0.02	-0.07	0.14	-0.10	0.18	-0.20	0.10
kid1519	-0.03	0.02	-0.03	0.02	-0.06	0.02	-0.05	0.01	-0.04	0.07	-0.20	0.05	-0.15	0.12

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	-0.05	0.02	-0.06	0.01	-0.09	0.03	-0.08	0.02	-0.07	0.14	-0.12	0.08	-0.03	0.23
kidle14	-0.04	0.02	-0.06	0.02	-0.07	0.03	-0.08	0.01	-0.09	0.15	-0.08	0.10	-0.05	0.16
kidle19	-0.04	0.02	-0.05	0.02	-0.07	0.03	-0.08	0.01	-0.07	0.14	-0.11	0.08	-0.05	0.14
kidle9	-0.02	0.01	-0.04	0.01	-0.04	0.04	-0.06	0.02	-0.11	0.13	-0.11	0.03	-0.03	0.14
kidlt5	-0.02	0.01	-0.04	0.01	-0.01	0.06	-0.02	0.02	-0.07	0.04	-0.09	0.03	-0.04	0.11
latino	-0.06	0.02	-0.06	0.02	-0.09	0.03	-0.08	0.03	-0.02	0.07	-0.21	0.19	-0.21	0.05
maxdmd	-0.23	0.24											-0.25	0.12
medage	-0.04	0.00	-0.05	0.01	-0.05	0.01	-0.07	0.04	-0.04	0.03	-0.08	0.15	-0.17	0.04
medinc	-0.03	0.05	-0.03	0.05	-0.03	0.09	-0.04	0.07	-0.05	0.04	-0.20	0.21	-0.13	0.16
no_bldgs	-0.06	0.02	-0.06	0.00	-0.05	0.03	-0.07	0.04	-0.03	0.07	-0.06	0.18	-0.21	0.12
no_floors	-0.03	0.04	-0.03	0.04	-0.03	0.09	-0.04	0.05	-0.07	0.02	-0.05	0.15	-0.08	0.20
no_res_units	-0.02	0.03	-0.04	0.01	-0.02	0.09	-0.03	0.04	-0.07	0.04	-0.06	0.18	-0.07	0.28
no_units	-0.03	0.03	-0.04	0.02	-0.03	0.09	-0.02	0.05	-0.07	0.04	-0.07	0.18	-0.10	0.09
occhsng	-0.04	0.04	-0.04	0.01	-0.08	0.06	-0.08	0.04	-0.06	0.03	-0.15	0.19	-0.18	0.16
poland	-0.03	0.04	-0.03	0.03	-0.03	0.08	-0.03	0.06	-0.11	0.05	-0.12	0.21	-0.03	0.21
popn	-0.06	0.02	-0.05	0.01	-0.07	0.03	-0.09	0.02	-0.07	0.06	-0.18	0.16	-0.18	0.14
srcit	-0.04	0.00	-0.04	0.01	-0.04	0.01	-0.07	0.03	-0.07	0.05	-0.07	0.15	-0.16	0.06
ssm1-ssm24	0.00	0.27	0.00	0.24	0.03	0.30	0.03	0.33	0.04	0.31	-0.10	0.13	-0.02	0.20
tmp1-tmp24	0.03	0.32	0.02	0.29	0.16	0.36	0.19	0.40	0.18	0.36	-0.12	0.03	-0.03	0.21
white	-0.02	0.05	-0.02	0.05	-0.03	0.04	-0.02	0.07	-0.11	0.10	-0.12	0.10	-0.12	0.11
windies	-0.02	0.02	-0.02	0.04	-0.01	0.05	-0.03	0.02	-0.04	0.08	-0.03	0.15	-0.02	0.24
wsp1-wsp24	-0.12	0.00	-0.11	0.00	-0.14	-0.01	-0.16	-0.02	-0.14	-0.02	-0.07	0.06	-0.11	0.03
year_built	-0.04	0.05	-0.03	0.05	-0.03	0.08	-0.09	0.08	-0.10	0.04	-0.09	0.13	-0.02	0.20

**Table B.5: Correlation Coefficient Ranges for Residential Customer Data Groups - Energy Usage Intervals as a Percent of the Prior Interval (delt1-delt24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.02	0.02	-0.03	0.03	-0.04	0.06	-0.03	0.05	-0.05	0.02	-0.01	0.11	-0.13	0.09
annkwh	-0.03	0.02	-0.03	0.00	-0.02	0.01	-0.02	0.03	-0.09	0.01	-0.05	0.08	-0.15	0.16
asian	-0.03	0.03	-0.04	0.04	-0.06	0.03	-0.05	0.03	-0.08	0.01	-0.12	0.00	-0.17	0.21
bapru	-0.01	0.02	0.00	0.06	-0.02	0.11	-0.07	0.01	-0.06	0.01	-0.07	0.09	-0.40	0.15
bapu	-0.03	0.02	-0.01	0.06	-0.02	0.11	-0.03	0.19	-0.10	0.01	-0.10	0.07	-0.12	0.09
black	-0.01	0.00	-0.03	0.02	-0.01	0.04	-0.02	0.02	-0.06	0.10	-0.07	0.07	-0.18	0.10
bldg_area	-0.03	0.01	-0.02	0.03	-0.06	0.04	-0.06	0.03	-0.07	0.03	-0.09	0.03	-0.15	0.12
c1-c12	-0.03	0.03	-0.05	0.02	-0.04	0.09	-0.06	0.09	-0.14	0.17	-0.13	0.10	-0.17	0.31
cc1-cc24	-0.02	0.03	-0.02	0.02	-0.03	0.03	-0.03	0.04	-0.02	0.03	-0.05	0.05	-0.05	0.05
cdh1-cdh24	-0.05	0.07	-0.05	0.08	-0.06	0.07	-0.08	0.07	-0.05	0.07	-0.08	0.04	-0.07	0.12
d1-d12	-0.17	0.14											-0.18	0.51
daytype	-0.02	0.03	-0.02	0.02	-0.02	0.04	-0.03	0.03	-0.02	0.05	-0.04	0.03	-0.15	0.10
edcoll	-0.03	0.04	-0.04	0.04	-0.03	0.05	-0.03	0.05	-0.04	0.06	0.00	0.12	-0.20	0.16
edhs	-0.02	0.02	-0.04	0.02	-0.05	0.02	-0.05	0.02	-0.04	0.05	-0.02	0.07	-0.21	0.31
edlehs	-0.03	0.03	-0.03	0.04	-0.06	0.03	-0.06	0.02	-0.04	0.02	-0.11	0.00	-0.17	0.26
editcoll	-0.04	0.03	-0.04	0.04	-0.05	0.03	-0.05	0.03	-0.06	0.04	-0.12	0.00	-0.16	0.20
edlths	-0.03	0.03	-0.04	0.04	-0.05	0.03	-0.05	0.03	-0.07	0.02	-0.11	-0.01	-0.22	0.13
edsmcoll	-0.04	0.02	-0.04	0.04	-0.04	0.03	-0.05	0.05	-0.09	0.07	-0.13	-0.01	-0.15	0.09
elecheat	-0.03	0.02	-0.03	0.02	-0.03	0.06	-0.03	0.04	-0.05	0.03	-0.03	0.06	-0.14	0.23
far	-0.03	0.01	-0.04	0.01	-0.04	0.03	-0.04	0.01	-0.06	0.02	-0.09	0.03	-0.16	0.10
footprint	-0.01	0.02	-0.01	0.03	-0.03	0.04	-0.02	0.06	-0.06	0.07	-0.06	0.06	-0.13	0.22
foreign	-0.03	0.03	-0.03	0.05	-0.05	0.03	-0.05	0.03	-0.08	0.03	-0.12	-0.01	-0.24	0.14
guyana	-0.02	0.02	-0.02	0.03	-0.03	0.03	-0.02	0.01	-0.05	0.03	-0.07	0.08	-0.18	0.29
hdh1-hdh24	-0.05	0.04	-0.05	0.05	-0.06	0.05	-0.06	0.06	-0.05	0.06	-0.03	0.11	-0.09	0.08
hhsz	-0.04	0.02	-0.04	0.03	-0.04	0.03	-0.05	0.02	-0.11	0.05	-0.13	0.02	-0.10	0.13
houses	-0.03	0.02	-0.03	0.01	-0.07	0.03	-0.04	0.02	-0.09	0.06	-0.09	0.08	-0.14	0.16
hum9	-0.01	0.02	-0.02	0.02	-0.01	0.02	-0.02	0.03	-0.03	0.02	-0.08	0.03	-0.05	0.06
ireland	0.00	0.03	0.00	0.04	-0.03	0.04	-0.03	0.05	-0.03	0.04	-0.03	0.08	-0.06	0.16
italy	-0.02	0.02	-0.03	0.04	-0.04	0.04	-0.04	0.05	-0.03	0.04	-0.03	0.09	-0.23	0.50
jamaica	-0.01	0.01	-0.01	0.02	-0.01	0.04	-0.01	0.02	-0.06	0.11	-0.07	0.06	-0.20	0.10
delt1-delt24	-0.09	0.06	-0.11	0.06	-0.11	0.05	-0.08	0.09	-0.12	0.24	-0.07	0.28	-0.22	0.21
kid1014	-0.04	0.03	-0.03	0.04	-0.05	0.05	-0.06	0.02	-0.07	0.09	-0.11	0.06	-0.22	0.44
kid1519	-0.03	0.02	-0.03	0.02	-0.04	0.05	-0.04	0.02	-0.04	0.05	-0.06	0.07	-0.12	0.14

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	-0.04	0.02	-0.04	0.02	-0.05	0.04	-0.05	0.02	-0.07	0.09	-0.11	0.01	-0.17	0.07
kidle14	-0.04	0.01	-0.03	0.02	-0.05	0.03	-0.05	0.02	-0.10	0.08	-0.12	0.02	-0.13	0.19
kidle19	-0.04	0.02	-0.04	0.02	-0.05	0.04	-0.05	0.02	-0.07	0.08	-0.10	0.02	-0.15	0.17
kidle9	-0.03	0.00	-0.03	0.02	-0.04	0.02	-0.04	0.02	-0.10	0.05	-0.08	-0.01	-0.10	0.18
kidlt5	-0.02	0.00	-0.03	0.02	-0.06	0.04	-0.03	0.03	-0.06	0.04	-0.08	0.00	-0.09	0.17
latino	-0.03	0.03	-0.03	0.03	-0.05	0.04	-0.06	0.02	-0.02	0.02	-0.09	0.03	-0.23	0.16
maxdmd	-0.14	-0.01											-0.17	0.31
medage	-0.02	0.03	-0.03	0.04	-0.03	0.04	-0.04	0.06	-0.03	0.02	-0.09	0.06	-0.13	0.15
medinc	-0.02	0.02	-0.02	0.03	-0.03	0.06	-0.02	0.05	-0.04	0.03	-0.07	0.09	-0.24	0.15
no_bldgs	-0.02	0.02	-0.03	0.03	-0.05	0.03	-0.04	0.05	-0.06	0.03	-0.09	0.02	-0.16	0.20
no_floors	-0.03	0.02	-0.03	0.03	-0.04	0.06	-0.03	0.03	-0.06	0.02	-0.07	0.03	-0.23	0.31
no_res_units	-0.02	0.02	-0.02	0.03	-0.06	0.05	-0.06	0.04	-0.08	0.03	-0.06	0.04	-0.40	0.15
no_units	-0.02	0.02	-0.02	0.03	-0.05	0.05	-0.06	0.05	-0.08	0.03	-0.06	0.04	-0.13	0.23
occhsng	-0.02	0.02	-0.02	0.01	-0.06	0.04	-0.05	0.03	-0.09	0.04	-0.09	0.05	-0.14	0.15
poland	-0.01	0.03	-0.01	0.03	-0.04	0.04	-0.03	0.06	-0.06	0.05	-0.01	0.09	-0.16	0.18
popn	-0.02	0.02	-0.04	0.02	-0.07	0.03	-0.05	0.03	-0.06	0.02	-0.11	0.00	-0.14	0.16
srcit	-0.02	0.03	-0.03	0.03	-0.03	0.05	-0.05	0.06	-0.06	0.03	-0.10	0.05	-0.07	0.19
ssm1-ssm24	-0.04	0.06	-0.04	0.06	-0.05	0.06	-0.05	0.05	-0.06	0.05	-0.07	0.05	-0.07	0.10
tmp1-tmp24	-0.05	0.06	-0.05	0.06	-0.06	0.07	-0.07	0.07	-0.06	0.06	-0.11	0.02	-0.08	0.11
white	-0.02	0.03	-0.03	0.04	-0.03	0.04	-0.03	0.05	-0.06	0.10	0.00	0.12	-0.22	0.18
windies	-0.01	0.01	-0.02	0.02	-0.01	0.04	-0.01	0.02	-0.06	0.11	-0.07	0.07	-0.16	0.09
wsp1-wsp24	-0.03	0.03	-0.03	0.02	-0.03	0.03	-0.03	0.04	-0.03	0.03	-0.04	0.06	-0.06	0.04
year_built	-0.04	0.03	-0.05	0.02	-0.04	0.05	-0.04	0.05	-0.05	0.04	-0.02	0.07	-0.16	0.10

**Table B.6: Correlation Coefficient Ranges for Business Customer Data Groups - Raw Energy Usage Intervals (k1-k24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.19	-0.15	0.40	0.49	-0.15	-0.07	-0.17	0.04	-0.13	0.04	-0.12	-0.02	-0.07	0.02
annkwh	0.93	0.95	0.16	0.32	0.46	0.82	0.67	0.92	0.66	0.88	0.72	0.92	0.79	0.89
asian	0.21	0.27	0.13	0.20	0.24	0.40	-0.18	0.00	-0.03	0.17	0.14	0.21	0.04	0.14
bapru	0.07	0.09	-0.05	0.02	-0.07	0.22	0.23	0.48	0.07	0.29	-0.07	0.03	0.03	0.06
bapu	0.29	0.38	-0.04	0.02	0.28	0.70	-0.26	-0.09	-0.08	0.13	-0.10	-0.01	0.02	0.14
black	0.09	0.14	0.48	0.58	-0.12	0.07	0.15	0.28	-0.09	0.01	-0.29	-0.18	0.00	0.05
bldg_area	0.16	0.24	-0.05	-0.01	-0.05	0.21	0.22	0.31	-0.16	0.04	-0.07	0.07	0.04	0.21
c1-c12	0.88	0.95	0.08	0.33	0.30	0.81	0.62	0.91	0.59	0.91	0.61	0.91	0.67	0.88
cc1-cc24	-0.03	0.00	-0.03	0.00	-0.10	-0.01	-0.07	0.00	-0.03	0.01	-0.05	0.00	-0.07	0.00
cdh1-cdh24	0.06	0.12	0.04	0.12	0.10	0.34	0.03	0.26	0.07	0.17	0.10	0.25	0.12	0.28
d1-d12	0.78	0.91	0.29	0.68	-0.11	0.44	0.13	0.74	0.35	0.68	0.39	0.66	0.43	0.72
daytype	0.00	0.02	-0.01	0.00	0.00	0.05	0.01	0.05	-0.02	0.03	0.00	0.04	0.00	0.04
edcoll	-0.25	-0.23	-0.32	-0.23	-0.14	0.11	-0.12	-0.02	-0.09	0.06	0.08	0.19	-0.07	0.06
edhs	0.17	0.20	0.08	0.14	-0.15	0.09	0.00	0.13	-0.06	0.12	-0.11	-0.01	-0.20	-0.09
edlehs	0.13	0.15	0.20	0.29	-0.09	0.09	-0.03	0.08	-0.05	0.09	-0.24	-0.10	-0.09	0.02
edltcoll	0.24	0.25	0.23	0.32	-0.11	0.14	0.02	0.12	-0.06	0.09	-0.19	-0.08	-0.06	0.07
edlths	0.06	0.08	0.24	0.33	-0.05	0.12	-0.06	0.05	-0.06	0.08	-0.29	-0.14	0.07	0.14
edsmcoll	0.37	0.38	0.23	0.34	-0.13	0.25	0.16	0.26	-0.09	0.00	0.01	0.15	0.04	0.14
elecheat	-0.20	-0.17	-0.12	-0.07	-0.09	0.06	-0.16	0.00	-0.10	0.13	-0.01	0.07	-0.20	-0.08
far	-0.11	-0.02	-0.14	-0.06	-0.09	0.07	0.13	0.29	-0.24	-0.03	-0.25	-0.11	-0.07	0.14
footprint	0.09	0.12	0.00	0.07	-0.02	0.19	0.16	0.25	0.03	0.11	-0.02	0.09	-0.03	0.04
foreign	0.30	0.39	0.13	0.20	0.33	0.53	-0.05	0.15	-0.11	0.12	0.07	0.11	-0.02	0.15
guyana	0.12	0.16	0.46	0.55	0.08	0.22	0.11	0.28	-0.09	0.00	-0.10	0.00	-0.14	0.00
hdh1-hdh24	-0.10	-0.04	-0.12	-0.03	-0.24	-0.05	-0.14	0.05	-0.13	-0.02	-0.19	-0.04	-0.23	-0.06
hhsz	0.04	0.05	0.09	0.17	-0.09	0.07	0.03	0.13	-0.03	0.12	-0.28	-0.20	-0.05	0.05
houses	-0.08	-0.05	-0.22	-0.13	-0.17	0.01	-0.11	-0.02	-0.10	0.02	0.19	0.31	0.04	0.13
hum9	0.01	0.05	0.01	0.06	-0.01	0.10	-0.03	0.07	0.00	0.07	0.00	0.09	0.01	0.11
ireland	-0.23	-0.18	-0.25	-0.19	-0.17	0.09	-0.18	-0.10	-0.08	0.09	0.24	0.32	-0.14	-0.04
italy	-0.03	0.01	-0.40	-0.32	-0.21	-0.03	-0.17	-0.10	0.03	0.13	0.21	0.29	-0.21	-0.11
jamaica	0.14	0.20	0.46	0.55	-0.07	0.17	0.13	0.24	-0.12	0.07	-0.03	0.07	-0.01	0.09
k1-k24	0.88	1.00	0.90	1.00	0.35	0.98	0.69	1.00	0.64	1.00	0.74	1.00	0.72	1.00
kid1014	0.25	0.28	0.18	0.26	-0.01	0.13	0.17	0.24	-0.06	0.09	-0.03	0.08	-0.18	-0.10
kid1519	0.20	0.23	0.07	0.11	0.10	0.35	0.00	0.22	-0.17	-0.03	0.04	0.15	-0.03	0.01
kid59	0.31	0.33	0.30	0.40	-0.05	0.06	0.12	0.24	-0.10	0.11	0.11	0.18	0.03	0.11
kid14	0.29	0.33	0.27	0.38	-0.08	0.09	0.13	0.24	-0.10	0.14	0.05	0.15	-0.06	0.09



Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kidle19	0.33	0.36	0.24	0.32	0.03	0.19	0.15	0.26	-0.15	0.07	0.07	0.20	-0.05	0.05
kidle9	0.15	0.24	0.23	0.34	-0.11	0.08	0.06	0.23	-0.08	0.12	0.08	0.14	0.04	0.23
kidlt5	-0.12	-0.01	-0.02	0.03	-0.13	0.06	-0.08	0.17	-0.08	0.07	-0.01	0.05	0.01	0.22
latino	-0.06	-0.04	-0.12	-0.06	-0.12	0.03	-0.11	0.03	-0.08	0.04	-0.39	-0.28	0.04	0.10
maxdmd	0.78	0.89	0.43	0.57	0.16	0.30	0.45	0.51	0.50	0.66	0.37	0.50	0.44	0.61
medage	0.21	0.33	-0.07	0.01	0.23	0.34	0.01	0.22	0.04	0.19	0.14	0.22	-0.22	-0.02
medinc	-0.18	-0.15	-0.20	-0.14	-0.13	0.00	-0.11	0.07	-0.03	0.18	0.14	0.20	-0.14	-0.05
no_bldgs	0.09	0.10	-0.28	-0.21	-0.02	0.14	-0.18	0.06	-0.09	0.05	-0.02	0.05	-0.02	0.02
no_floors	0.07	0.10	-0.10	-0.02	-0.04	0.16	0.42	0.61	0.04	0.24	0.01	0.13	0.02	0.13
no_res_units	-0.08	-0.06	0.01	0.11	-0.05	0.13	0.39	0.65	0.01	0.26	0.07	0.18	-0.11	-0.09
no_units	-0.04	0.01	0.00	0.09	-0.07	0.12	0.40	0.65	-0.06	0.19	0.04	0.18	0.03	0.16
occhsng	-0.05	-0.01	-0.19	-0.10	-0.15	0.02	-0.11	-0.01	-0.09	0.02	0.20	0.32	0.05	0.14
poland	-0.03	-0.02	-0.15	-0.07	-0.18	-0.02	-0.09	0.09	-0.13	0.09	0.05	0.20	-0.02	0.01
popn	0.09	0.14	-0.14	-0.06	-0.09	0.07	-0.04	0.06	-0.08	0.06	0.18	0.32	0.03	0.12
srcit	0.29	0.38	0.21	0.28	0.22	0.36	0.07	0.23	0.04	0.17	0.18	0.28	-0.10	0.07
ssm1-ssm24	0.01	0.09	0.00	0.09	0.01	0.23	-0.01	0.18	0.00	0.13	0.00	0.19	0.01	0.21
tmp1-tmp24	0.05	0.11	0.04	0.13	0.08	0.28	-0.02	0.17	0.05	0.14	0.07	0.21	0.10	0.26
white	-0.22	-0.19	-0.45	-0.35	-0.26	-0.04	-0.19	-0.12	-0.07	0.04	0.16	0.24	-0.09	-0.05
windies	0.12	0.18	0.53	0.66	0.04	0.20	0.14	0.30	-0.12	0.05	-0.17	-0.01	-0.01	0.07
wsp1-wsp24	-0.05	0.00	-0.05	-0.01	-0.14	0.01	-0.09	0.02	-0.08	0.02	-0.10	0.01	-0.11	0.00
year_built	0.35	0.41	0.03	0.19	-0.09	0.02	0.37	0.59	0.06	0.26	0.31	0.45	-0.08	0.06

**Table B.7: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Daily Maximum Value (pctd1-pctd24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.11	0.01	-0.14	0.21	-0.13	0.07	-0.27	0.20	-0.05	0.18	-0.22	0.07	-0.08	0.07
annkwh	0.29	0.47	0.27	0.41	0.15	0.45	0.14	0.56	0.21	0.47	0.10	0.64	0.24	0.44
asian	0.07	0.17	0.00	0.15	0.02	0.21	-0.31	0.06	-0.18	0.18	0.09	0.24	-0.12	0.09
bapru	0.04	0.08	-0.13	0.12	-0.18	0.19	-0.12	0.36	-0.03	0.31	-0.06	0.12	0.01	0.06
bapu	0.09	0.20	-0.02	0.11	0.06	0.31	-0.30	0.16	-0.11	0.09	-0.09	0.05	-0.06	0.13
black	0.04	0.14	-0.15	0.27	-0.19	0.05	0.11	0.29	-0.10	0.05	-0.25	0.07	0.02	0.11
bldg_area	0.04	0.15	0.00	0.17	-0.12	0.20	0.03	0.23	-0.10	0.18	-0.11	0.12	-0.08	0.15
c1-c12	0.27	0.47	0.19	0.43	0.04	0.46	0.11	0.58	0.15	0.50	-0.04	0.65	0.19	0.45
cc1-cc24	-0.02	0.02	-0.04	0.02	-0.03	0.03	-0.05	0.04	-0.04	0.03	-0.03	0.03	-0.03	0.05
cdh1-cdh24	-0.06	0.06	-0.05	0.07	-0.08	0.03	-0.09	0.08	-0.05	0.04	-0.07	0.09	-0.10	0.08
d1-d12	0.24	0.39	-0.08	0.32	-0.25	0.20	-0.16	0.40	-0.16	0.26	-0.24	0.31	-0.04	0.19
daytype	-0.07	0.02	-0.06	0.02	-0.08	0.03	-0.13	0.00	-0.09	0.04	-0.09	0.02	-0.08	0.03
edcoll	-0.20	-0.08	-0.15	0.13	-0.18	0.13	-0.16	0.01	-0.20	0.05	-0.10	0.05	-0.19	0.10
edhs	0.05	0.18	-0.11	0.08	-0.16	0.14	-0.03	0.22	-0.07	0.25	-0.11	0.13	-0.15	0.14
edlehs	0.06	0.16	-0.15	0.14	-0.11	0.15	-0.04	0.11	-0.03	0.21	-0.07	0.08	-0.10	0.16
editcoll	0.08	0.20	-0.13	0.15	-0.13	0.18	-0.01	0.16	-0.05	0.20	-0.05	0.10	-0.10	0.19
edlths	0.04	0.13	-0.15	0.16	-0.07	0.13	-0.04	0.10	-0.01	0.16	-0.10	0.10	-0.01	0.15
edsmcoll	0.11	0.22	-0.10	0.23	-0.19	0.20	0.00	0.27	-0.16	0.03	-0.08	0.21	-0.03	0.12
elecheat	-0.19	-0.08	-0.10	0.14	-0.26	0.05	-0.27	0.07	-0.25	0.13	-0.10	0.00	-0.27	0.00
far	-0.11	0.06	-0.07	0.18	-0.09	0.10	-0.12	0.22	-0.21	0.10	-0.31	0.05	-0.16	0.14
footprint	0.01	0.09	-0.14	0.24	-0.07	0.17	0.02	0.25	0.07	0.21	-0.16	0.06	-0.05	0.08
foreign	0.10	0.23	-0.09	0.20	0.08	0.21	-0.30	0.17	-0.34	0.00	-0.09	0.08	-0.13	0.15
guyana	0.04	0.19	-0.13	0.22	-0.11	0.06	0.05	0.28	-0.24	-0.05	-0.11	0.12	-0.01	0.23
hdh1-hdh24	-0.06	0.07	-0.08	0.05	-0.01	0.08	-0.04	0.12	-0.06	0.06	-0.09	0.08	-0.07	0.13
hhsz	0.04	0.17	-0.13	0.06	-0.12	0.11	0.00	0.20	0.02	0.16	-0.08	0.08	-0.03	0.13
houses	-0.15	-0.01	-0.07	0.12	-0.18	0.02	-0.17	-0.02	-0.15	0.03	-0.01	0.25	-0.10	0.10
hum9	-0.05	0.03	-0.04	0.05	-0.06	0.02	-0.08	0.03	-0.05	0.04	-0.05	0.05	-0.06	0.04
ireland	-0.16	-0.08	-0.12	0.13	-0.19	0.00	-0.27	-0.05	-0.18	0.09	0.00	0.13	-0.17	0.02
italy	-0.10	0.01	-0.22	0.04	-0.13	0.08	-0.16	-0.05	-0.11	0.07	-0.08	0.17	-0.29	0.01
jamaica	0.06	0.10	-0.10	0.23	-0.19	-0.01	0.08	0.23	-0.18	0.03	-0.08	0.10	-0.08	0.10
pctd1-pctd24	0.17	0.98	0.12	0.96	-0.12	0.95	-0.09	0.99	-0.04	0.99	0.02	0.99	0.10	0.99
kid1014	0.06	0.20	-0.11	0.07	0.04	0.18	0.08	0.27	-0.18	0.15	-0.21	0.06	-0.22	0.04
kid1519	0.08	0.17	-0.09	0.07	0.06	0.24	-0.06	0.39	-0.29	0.08	-0.02	0.19	-0.03	0.04

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	0.10	0.22	-0.16	0.10	-0.04	0.19	0.07	0.30	-0.26	0.18	-0.05	0.14	0.07	0.17
kidle14	0.08	0.24	-0.16	0.10	-0.07	0.15	0.02	0.26	-0.32	0.17	-0.15	0.13	-0.04	0.20
kidle19	0.10	0.25	-0.13	0.08	0.02	0.16	0.08	0.31	-0.35	0.13	-0.08	0.22	-0.04	0.17
kidle9	0.04	0.18	-0.19	0.10	-0.16	0.11	-0.08	0.28	-0.24	0.07	0.00	0.13	0.02	0.31
kidlt5	-0.13	0.05	-0.10	0.11	-0.23	0.05	-0.24	0.25	-0.27	0.04	-0.02	0.09	-0.07	0.28
latino	-0.06	0.06	-0.15	-0.01	-0.09	0.09	-0.09	0.09	-0.05	0.17	-0.23	-0.05	-0.01	0.11
maxdmd	0.27	0.39	-0.02	0.29	-0.09	0.11	0.05	0.19	0.00	0.20	-0.12	0.10	-0.02	0.15
medage	0.06	0.25	0.05	0.16	0.07	0.20	-0.26	0.24	-0.02	0.26	-0.03	0.20	-0.32	0.08
medinc	-0.15	-0.03	-0.12	0.13	-0.22	0.03	-0.22	0.12	-0.20	0.16	-0.07	0.10	-0.18	0.01
no_bldgs	0.03	0.09	-0.16	-0.01	-0.05	0.11	-0.39	0.18	-0.14	0.11	-0.03	0.09	0.01	0.09
no_floors	0.07	0.13	-0.12	0.19	-0.08	0.16	0.09	0.34	-0.03	0.17	-0.04	0.17	-0.06	0.19
no_res_units	-0.03	0.10	-0.19	0.29	-0.06	0.13	0.11	0.40	-0.26	0.30	-0.01	0.20	-0.07	0.03
no_units	0.01	0.13	-0.18	0.28	-0.07	0.14	0.11	0.40	-0.29	0.30	0.00	0.23	0.00	0.18
occhsng	-0.13	0.00	-0.05	0.13	-0.17	0.02	-0.18	-0.01	-0.14	0.02	-0.01	0.26	-0.10	0.10
poland	-0.06	0.01	-0.10	0.17	-0.24	0.00	-0.27	0.04	-0.22	0.00	-0.25	0.03	0.02	0.11
popn	-0.03	0.09	-0.04	0.14	-0.14	0.03	-0.11	0.03	-0.16	0.01	-0.05	0.29	-0.09	0.11
srcit	0.10	0.29	-0.01	0.23	0.09	0.19	-0.16	0.24	-0.10	0.14	-0.04	0.24	-0.22	0.12
ssm1-ssm24	-0.06	0.04	-0.06	0.05	-0.07	0.03	-0.09	0.05	-0.05	0.03	-0.08	0.08	-0.10	0.06
tmp1-tmp24	-0.06	0.06	-0.05	0.08	-0.07	0.02	-0.12	0.06	-0.06	0.05	-0.09	0.09	-0.13	0.08
white	-0.20	-0.10	-0.19	0.14	-0.21	0.11	-0.23	-0.08	-0.11	0.02	-0.12	0.06	-0.12	-0.03
windies	0.02	0.13	-0.13	0.30	-0.11	0.04	0.11	0.30	-0.22	0.07	-0.23	0.05	-0.03	0.12
wsp1-wsp24	-0.03	0.04	-0.04	0.04	-0.04	0.06	-0.04	0.07	-0.04	0.04	-0.05	0.04	-0.05	0.06
year_built	0.07	0.32	-0.13	0.20	-0.13	-0.02	0.14	0.45	-0.16	0.31	-0.12	0.28	-0.10	0.20

**Table B.8: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Monthly Billed kW (pctm1-pctm24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.16	-0.11	-0.09	0.01	-0.17	-0.01	-0.30	0.04	-0.06	0.12	-0.22	-0.07	-0.11	0.01
annkwh	0.46	0.63	0.46	0.78	0.34	0.73	0.40	0.83	0.35	0.66	0.48	0.83	0.43	0.60
asian	0.14	0.21	0.06	0.16	0.12	0.29	-0.25	-0.02	-0.04	0.22	0.22	0.32	-0.10	0.04
bapru	0.04	0.08	-0.02	0.17	-0.09	0.21	0.10	0.46	0.11	0.40	-0.06	0.06	0.01	0.04
bapu	0.15	0.23	-0.01	0.13	0.13	0.52	-0.30	0.01	-0.13	0.02	-0.12	-0.01	-0.04	0.09
black	0.09	0.16	-0.05	0.04	-0.21	-0.04	0.21	0.42	-0.12	0.00	-0.26	-0.09	0.05	0.10
bldg_area	0.08	0.16	-0.01	0.15	-0.12	0.14	0.09	0.20	-0.18	0.05	-0.08	0.07	-0.04	0.13
c1-c12	0.43	0.62	0.38	0.79	0.18	0.70	0.35	0.83	0.29	0.66	0.36	0.83	0.37	0.60
cc1-cc24	-0.03	0.02	-0.05	0.01	-0.05	0.02	-0.04	0.04	-0.03	0.03	-0.02	0.03	-0.03	0.03
cdh1-cdh24	0.01	0.14	0.04	0.19	0.04	0.22	0.00	0.20	0.03	0.12	0.01	0.12	-0.01	0.13
d1-d12	0.37	0.51	0.10	0.54	-0.22	0.25	-0.10	0.57	-0.11	0.20	0.01	0.44	0.00	0.23
daytype	0.00	0.05	0.00	0.04	0.00	0.06	0.01	0.07	-0.03	0.04	-0.01	0.05	0.01	0.06
edcoll	-0.25	-0.18	-0.07	0.07	-0.08	0.17	-0.16	-0.04	-0.12	0.08	0.02	0.10	-0.25	-0.08
edhs	0.14	0.22	-0.08	0.02	-0.19	0.07	0.05	0.23	-0.08	0.16	-0.06	0.02	0.00	0.15
edlehs	0.11	0.17	-0.11	0.00	-0.14	0.05	-0.02	0.08	-0.06	0.12	-0.14	-0.04	0.04	0.19
editcoll	0.18	0.25	-0.07	0.07	-0.17	0.08	0.04	0.16	-0.07	0.12	-0.10	-0.02	0.08	0.25
edlths	0.05	0.11	-0.12	-0.01	-0.10	0.09	-0.08	0.07	-0.04	0.10	-0.22	-0.08	0.06	0.19
edsmcoll	0.25	0.32	0.02	0.27	-0.21	0.16	0.18	0.39	-0.11	0.02	0.01	0.18	0.09	0.19
elecheat	-0.21	-0.12	-0.02	0.13	-0.14	0.05	-0.26	-0.03	-0.18	0.10	-0.04	0.02	-0.32	-0.16
far	-0.13	-0.01	0.07	0.24	-0.09	0.08	-0.06	0.17	-0.25	0.03	-0.26	-0.05	-0.18	0.03
footprint	0.05	0.10	-0.02	0.31	-0.07	0.15	0.09	0.23	0.06	0.20	-0.07	0.05	-0.04	0.06
foreign	0.20	0.29	0.04	0.25	0.23	0.42	-0.15	0.17	-0.27	-0.03	0.07	0.15	-0.07	0.12
guyana	0.12	0.24	-0.05	0.13	-0.04	0.10	0.15	0.40	-0.18	-0.05	-0.10	0.03	0.04	0.24
hdh1-hdh24	-0.08	0.03	-0.16	-0.03	-0.11	0.02	-0.07	0.08	-0.06	0.03	-0.05	0.05	-0.05	0.09
hhsz	0.05	0.13	-0.17	-0.02	-0.16	0.01	0.03	0.17	0.00	0.13	-0.22	-0.12	0.02	0.14
houses	-0.13	-0.04	0.08	0.22	-0.16	0.06	-0.16	-0.04	-0.06	0.06	0.21	0.36	-0.08	0.04
hum9	-0.01	0.04	0.00	0.08	-0.02	0.05	-0.03	0.04	-0.01	0.05	-0.02	0.04	-0.03	0.04
ireland	-0.17	-0.11	-0.02	0.09	-0.17	0.08	-0.23	-0.10	-0.11	0.10	0.17	0.23	-0.15	-0.03
italy	-0.06	0.03	-0.24	-0.06	-0.14	0.06	-0.19	-0.12	-0.02	0.11	0.13	0.25	-0.18	-0.03
jamaica	0.09	0.13	0.03	0.20	-0.13	0.02	0.18	0.35	-0.19	-0.01	0.02	0.14	-0.05	0.06
pctm1-pctm24	0.60	0.99	0.54	0.99	0.23	0.97	0.45	0.99	0.34	0.99	0.60	1.00	0.53	0.99
kid1014	0.14	0.24	-0.18	-0.06	0.01	0.19	0.20	0.36	-0.17	0.03	-0.10	0.05	-0.11	0.05
kid1519	0.12	0.20	-0.10	-0.02	0.12	0.37	0.08	0.41	-0.22	0.02	0.03	0.18	-0.02	0.02

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	0.21	0.30	-0.14	0.00	-0.04	0.12	0.17	0.39	-0.19	0.08	0.09	0.18	0.14	0.22
kidle14	0.19	0.31	-0.21	0.01	-0.09	0.09	0.18	0.38	-0.24	0.06	0.02	0.16	0.04	0.22
kidle19	0.20	0.33	-0.17	-0.01	0.03	0.21	0.21	0.43	-0.27	0.04	0.05	0.23	0.04	0.17
kidle9	0.14	0.24	-0.11	0.11	-0.16	0.06	0.09	0.30	-0.15	0.06	0.11	0.16	0.10	0.33
kidlt5	-0.09	0.03	0.00	0.16	-0.19	0.00	-0.13	0.18	-0.15	0.06	0.02	0.11	-0.03	0.24
latino	-0.09	-0.02	-0.22	-0.10	-0.13	0.04	-0.16	0.03	-0.09	0.06	-0.40	-0.29	0.01	0.10
maxdmd	0.37	0.49	0.13	0.42	-0.05	0.13	0.13	0.25	0.08	0.20	0.07	0.15	0.01	0.11
medage	0.17	0.31	0.14	0.34	0.16	0.24	-0.10	0.24	0.04	0.28	0.17	0.29	-0.22	0.07
medinc	-0.15	-0.07	-0.01	0.11	-0.12	0.04	-0.19	0.06	-0.11	0.14	0.06	0.15	-0.22	-0.10
no_bldgs	0.04	0.09	-0.08	0.03	-0.06	0.11	-0.21	0.15	-0.10	0.05	-0.03	0.06	0.00	0.05
no_floors	0.09	0.14	-0.01	0.30	-0.09	0.13	0.21	0.47	0.01	0.20	0.05	0.17	-0.07	0.07
no_res_units	0.02	0.11	-0.02	0.40	-0.07	0.11	0.25	0.60	-0.03	0.38	0.07	0.22	-0.08	-0.04
no_units	0.06	0.14	-0.02	0.38	-0.09	0.11	0.25	0.59	-0.06	0.35	0.08	0.25	0.04	0.19
occhsng	-0.10	0.00	0.12	0.26	-0.15	0.07	-0.16	-0.03	-0.05	0.06	0.21	0.37	-0.08	0.05
poland	-0.04	0.02	0.09	0.21	-0.17	0.00	-0.15	0.02	-0.19	-0.04	-0.13	0.04	0.03	0.10
popn	0.03	0.12	0.15	0.26	-0.11	0.10	-0.08	0.06	-0.07	0.07	0.19	0.38	-0.06	0.07
srcit	0.22	0.36	0.12	0.31	0.13	0.31	0.00	0.30	-0.03	0.16	0.17	0.33	-0.12	0.12
ssm1-ssm24	-0.02	0.11	0.00	0.12	-0.02	0.16	-0.04	0.16	-0.02	0.10	-0.02	0.10	-0.04	0.11
tmp1-tmp24	-0.02	0.10	0.03	0.17	0.00	0.14	-0.05	0.10	-0.01	0.08	-0.03	0.07	-0.06	0.08
white	-0.22	-0.16	-0.02	0.14	-0.15	0.10	-0.27	-0.16	-0.08	0.04	0.03	0.10	-0.13	-0.06
windies	0.08	0.15	0.20	0.39	-0.08	0.04	0.21	0.45	-0.17	0.01	-0.19	0.00	-0.02	0.08
wsp1-wsp24	-0.06	0.03	-0.10	0.01	-0.09	0.03	-0.08	0.06	-0.06	0.05	-0.05	0.04	-0.05	0.06
year_built	0.23	0.41	0.00	0.26	-0.15	-0.02	0.30	0.64	0.06	0.41	0.16	0.36	0.01	0.22

**Table B.9: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Annual Maximum Billed kW (pcta1-pcta24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.13	-0.08	-0.07	0.04	-0.15	0.02	-0.33	-0.09	-0.09	0.11	-0.20	-0.06	-0.05	0.05
annkwh	0.48	0.64	0.52	0.84	0.33	0.72	0.44	0.80	0.35	0.70	0.52	0.83	0.51	0.66
asian	0.12	0.19	0.04	0.13	0.13	0.29	-0.29	-0.05	-0.09	0.18	0.20	0.28	-0.09	0.01
bapru	0.07	0.10	-0.02	0.16	-0.06	0.23	0.11	0.43	0.10	0.40	-0.04	0.07	0.06	0.10
bapu	0.13	0.21	-0.08	0.04	0.08	0.46	-0.25	0.01	-0.13	0.01	-0.11	0.00	-0.05	0.07
black	0.10	0.15	-0.05	0.03	-0.21	-0.05	0.26	0.48	-0.09	0.02	-0.26	-0.10	0.03	0.07
bldg_area	0.10	0.18	-0.03	0.12	-0.12	0.14	0.09	0.19	-0.15	0.06	-0.07	0.07	0.02	0.18
c1-c12	0.45	0.65	0.37	0.84	0.18	0.68	0.39	0.82	0.27	0.69	0.38	0.83	0.44	0.67
cc1-cc24	-0.07	-0.01	-0.07	0.00	-0.11	-0.01	-0.10	0.00	-0.05	0.00	-0.06	-0.01	-0.09	0.00
cdh1-cdh24	0.11	0.29	0.05	0.24	0.15	0.39	0.07	0.36	0.12	0.27	0.12	0.28	0.15	0.36
d1-d12	0.36	0.53	0.18	0.54	-0.23	0.27	-0.10	0.57	-0.02	0.28	0.06	0.44	0.05	0.35
daytype	0.00	0.04	0.00	0.04	0.00	0.06	0.01	0.06	-0.03	0.04	0.00	0.04	0.00	0.05
edcoll	-0.24	-0.16	-0.10	0.05	-0.06	0.19	-0.16	-0.07	-0.12	0.07	0.06	0.12	-0.20	-0.04
edhs	0.13	0.21	-0.06	0.04	-0.21	0.05	0.08	0.24	-0.09	0.14	-0.09	-0.01	-0.02	0.12
edlehs	0.09	0.15	-0.09	0.02	-0.16	0.02	-0.02	0.07	-0.06	0.13	-0.17	-0.07	-0.01	0.12
editcoll	0.16	0.24	-0.05	0.10	-0.19	0.06	0.07	0.16	-0.07	0.12	-0.12	-0.06	0.04	0.20
edlths	0.03	0.08	-0.11	0.00	-0.12	0.05	-0.10	-0.01	-0.03	0.11	-0.24	-0.09	0.00	0.12
edsmcoll	0.27	0.33	0.05	0.33	-0.19	0.21	0.28	0.44	-0.10	0.03	0.00	0.18	0.11	0.20
elecheat	-0.20	-0.12	-0.08	0.08	-0.11	0.06	-0.27	-0.09	-0.14	0.15	-0.04	0.01	-0.25	-0.09
far	-0.14	-0.02	0.09	0.25	-0.14	0.05	-0.07	0.11	-0.23	0.04	-0.28	-0.10	-0.17	0.00
footprint	0.07	0.11	-0.06	0.25	-0.04	0.19	0.10	0.21	0.05	0.19	-0.05	0.05	-0.01	0.09
foreign	0.18	0.26	0.06	0.26	0.21	0.40	-0.07	0.18	-0.25	0.01	0.06	0.14	-0.09	0.06
guyana	0.07	0.17	-0.05	0.15	-0.04	0.10	0.20	0.45	-0.17	-0.04	-0.12	0.00	-0.06	0.11
hdh1-hdh24	-0.23	-0.05	-0.18	-0.01	-0.29	-0.10	-0.23	0.02	-0.22	-0.07	-0.23	-0.06	-0.30	-0.07
hhsz	0.05	0.11	-0.18	-0.03	-0.18	-0.03	0.03	0.11	0.01	0.14	-0.26	-0.18	0.04	0.15
houses	-0.12	-0.03	0.07	0.22	-0.14	0.07	-0.11	-0.04	-0.07	0.07	0.25	0.38	-0.08	0.03
hum9	0.00	0.11	0.00	0.09	0.00	0.13	-0.01	0.11	0.00	0.12	0.00	0.11	0.01	0.14
ireland	-0.15	-0.10	-0.05	0.07	-0.14	0.14	-0.22	-0.12	-0.08	0.12	0.16	0.22	-0.11	0.00
italy	-0.04	0.02	-0.26	-0.07	-0.10	0.07	-0.18	-0.05	-0.02	0.08	0.15	0.26	-0.15	-0.03
jamaica	0.10	0.13	0.03	0.20	-0.11	0.07	0.25	0.41	-0.16	0.03	0.04	0.15	-0.04	0.05
pcta1-pcta24	0.66	0.99	0.55	0.98	0.28	0.98	0.56	0.99	0.35	0.99	0.67	1.00	0.63	0.99
kid1014	0.15	0.24	-0.16	-0.03	0.04	0.22	0.26	0.41	-0.18	0.02	-0.03	0.09	-0.13	0.01
kid1519	0.14	0.21	-0.06	0.02	0.11	0.37	0.16	0.45	-0.23	0.00	0.06	0.18	-0.01	0.04

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	0.22	0.30	-0.11	0.05	-0.04	0.09	0.26	0.45	-0.22	0.05	0.15	0.20	0.12	0.20
kidle14	0.21	0.31	-0.19	0.04	-0.08	0.09	0.25	0.43	-0.22	0.10	0.10	0.19	0.04	0.19
kidle19	0.23	0.33	-0.14	0.03	0.04	0.22	0.29	0.49	-0.26	0.06	0.12	0.25	0.04	0.15
kidle9	0.16	0.24	-0.11	0.12	-0.17	0.04	0.16	0.39	-0.13	0.12	0.13	0.19	0.12	0.31
kidlt5	-0.07	0.06	-0.02	0.12	-0.21	-0.01	-0.09	0.21	-0.07	0.14	0.02	0.10	0.02	0.24
latino	-0.09	-0.03	-0.21	-0.08	-0.15	0.00	-0.18	-0.05	-0.08	0.07	-0.39	-0.29	0.03	0.11
maxdmd	0.35	0.47	0.08	0.38	-0.11	0.07	0.05	0.19	0.10	0.21	0.01	0.12	0.01	0.08
medage	0.13	0.26	0.18	0.40	0.16	0.26	-0.01	0.25	-0.01	0.22	0.14	0.24	-0.25	-0.02
medinc	-0.13	-0.06	-0.07	0.06	-0.11	0.04	-0.18	0.03	-0.09	0.17	0.09	0.17	-0.16	-0.03
no_bldgs	0.05	0.09	-0.05	0.09	-0.02	0.17	-0.19	0.13	-0.11	0.03	0.00	0.07	-0.01	0.03
no_floors	0.07	0.13	0.00	0.30	-0.09	0.13	0.18	0.42	0.04	0.26	0.00	0.12	-0.08	0.02
no_res_units	0.04	0.11	0.00	0.41	-0.04	0.17	0.24	0.56	0.00	0.39	0.07	0.19	-0.10	-0.07
no_units	0.07	0.14	0.00	0.40	-0.06	0.16	0.24	0.56	-0.03	0.35	0.07	0.21	0.04	0.18
occhsng	-0.09	-0.01	0.12	0.26	-0.12	0.09	-0.11	-0.03	-0.07	0.07	0.25	0.39	-0.08	0.03
poland	-0.02	0.06	0.08	0.19	-0.13	0.03	-0.12	0.11	-0.15	0.02	-0.09	0.07	0.07	0.14
popn	0.03	0.11	0.13	0.25	-0.10	0.11	-0.01	0.07	-0.07	0.09	0.24	0.39	-0.06	0.05
srcit	0.20	0.33	0.16	0.38	0.19	0.35	0.13	0.33	-0.07	0.10	0.15	0.28	-0.17	0.03
ssm1-ssm24	0.01	0.21	0.00	0.15	0.02	0.27	0.01	0.25	0.01	0.20	0.01	0.22	0.01	0.27
tmp1-tmp24	0.08	0.26	0.03	0.20	0.14	0.33	0.02	0.27	0.10	0.24	0.09	0.25	0.12	0.33
white	-0.20	-0.14	-0.02	0.15	-0.13	0.11	-0.28	-0.14	-0.09	0.02	0.08	0.13	-0.10	-0.03
windies	0.09	0.15	0.21	0.42	-0.06	0.06	0.29	0.52	-0.15	0.03	-0.17	0.01	-0.02	0.05
wsp1-wsp24	-0.12	0.00	-0.10	0.00	-0.16	0.00	-0.14	0.02	-0.12	0.03	-0.12	0.01	-0.14	0.00
year_built	0.22	0.38	0.02	0.26	-0.16	-0.02	0.28	0.57	0.06	0.38	0.17	0.34	-0.01	0.16

**Table B.10: Correlation Coefficient Ranges for Business Customer Data Groups - Energy Usage Intervals as a Percent of the Prior Interval (delt1-delt24)**

Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
america	-0.03	0.04	-0.07	0.14	-0.09	0.08	-0.20	0.17	-0.15	0.11	-0.08	0.13	-0.06	0.11
annkwh	-0.11	0.02	-0.12	-0.01	-0.13	0.00	-0.19	0.21	-0.26	0.10	-0.39	0.31	-0.12	0.06
asian	-0.05	0.02	-0.09	0.19	-0.06	0.12	-0.11	0.13	-0.19	0.21	-0.18	0.13	-0.10	0.08
bapru	-0.03	0.00	-0.08	0.06	-0.11	0.05	-0.16	0.19	-0.18	0.05	-0.24	0.16	-0.06	0.03
bapu	-0.07	0.01	-0.05	0.03	-0.04	0.03	-0.20	0.30	-0.07	0.10	-0.25	0.11	-0.17	0.07
black	-0.04	0.02	-0.07	0.17	-0.05	0.09	-0.10	0.09	-0.10	0.10	-0.11	0.12	-0.04	0.08
bldg_area	-0.06	0.00	-0.07	0.05	-0.04	0.10	-0.12	0.09	-0.13	0.08	-0.22	0.11	-0.21	0.07
c1-c12	-0.11	0.02	-0.13	0.01	-0.17	0.04	-0.20	0.25	-0.29	0.13	-0.40	0.35	-0.13	0.08
cc1-cc24	-0.01	0.01	-0.02	0.02	-0.02	0.03	-0.04	0.04	-0.03	0.03	-0.03	0.02	-0.02	0.01
cdh1-cdh24	-0.01	0.03	-0.02	0.03	-0.04	0.05	-0.07	0.14	-0.03	0.04	-0.05	0.06	-0.04	0.06
d1-d12	-0.11	0.06	-0.08	0.10	-0.16	0.13	-0.22	0.20	-0.22	0.11	-0.22	0.24	-0.09	0.09
daytype	-0.01	0.02	-0.02	0.03	-0.02	0.02	-0.06	0.03	-0.03	0.05	-0.05	0.02	-0.06	0.03
edcoll	-0.01	0.06	-0.09	0.06	-0.07	0.08	-0.10	0.20	-0.09	0.08	-0.15	0.10	-0.12	0.08
edhs	-0.05	0.01	-0.10	0.06	-0.12	0.10	-0.25	0.17	-0.15	0.10	-0.13	0.13	-0.07	0.12
edlehs	-0.04	0.01	-0.06	0.09	-0.07	0.07	-0.19	0.07	-0.10	0.10	-0.13	0.15	-0.08	0.11
editcoll	-0.06	0.01	-0.06	0.09	-0.08	0.07	-0.20	0.10	-0.08	0.09	-0.10	0.15	-0.08	0.12
edlths	-0.04	0.01	-0.06	0.10	-0.04	0.07	-0.13	0.09	-0.06	0.11	-0.18	0.17	-0.05	0.06
edsmcoll	-0.07	0.02	-0.10	0.14	-0.09	0.20	-0.12	0.20	-0.05	0.12	-0.17	0.18	-0.06	0.08
elecheat	-0.01	0.09	-0.07	0.05	-0.03	0.19	-0.20	0.29	-0.12	0.14	-0.07	0.13	-0.07	0.07
far	-0.03	0.04	-0.08	0.03	-0.10	0.17	-0.12	0.11	-0.14	0.04	-0.29	0.30	-0.11	0.10
footprint	-0.03	0.01	-0.09	0.05	-0.04	0.02	-0.14	0.13	-0.11	0.06	-0.09	0.14	-0.12	0.04
foreign	-0.07	0.02	-0.05	0.10	-0.11	0.11	-0.12	0.15	-0.12	0.21	-0.15	0.10	-0.17	0.10
guyana	-0.04	0.02	-0.07	0.16	-0.05	0.04	-0.08	0.10	-0.17	0.30	-0.12	0.17	-0.06	0.12
hdh1-hdh24	-0.02	0.02	-0.02	0.03	-0.02	0.06	-0.08	0.07	-0.05	0.04	-0.06	0.08	-0.08	0.05
hhsz	-0.05	0.01	-0.06	0.06	-0.04	0.09	-0.13	0.10	-0.10	0.11	-0.15	0.14	-0.09	0.05
houses	-0.02	0.05	-0.13	0.07	-0.07	0.02	-0.10	0.11	-0.10	0.11	-0.13	0.08	-0.09	0.08
hum9	-0.01	0.01	-0.02	0.02	-0.02	0.02	-0.05	0.05	-0.03	0.03	-0.05	0.04	-0.03	0.04
ireland	-0.01	0.09	-0.10	0.05	-0.03	0.22	-0.09	0.16	-0.08	0.10	-0.12	0.06	-0.04	0.12
italy	-0.03	0.03	-0.10	0.09	-0.10	0.01	-0.14	0.11	-0.11	0.07	-0.11	0.10	-0.11	0.09
jamaica	-0.03	0.01	-0.06	0.15	-0.03	0.13	-0.08	0.09	-0.10	0.03	-0.08	0.09	-0.07	0.07
delt1-delt24	-0.13	0.19	-0.16	0.22	-0.16	0.28	-0.31	0.21	-0.31	0.21	-0.33	0.52	-0.14	0.23
kid1014	-0.06	0.02	-0.07	0.17	-0.06	0.04	-0.19	0.16	-0.09	0.10	-0.09	0.13	-0.06	0.11
kid1519	-0.05	0.01	-0.05	0.16	-0.11	0.03	-0.26	0.22	-0.12	0.08	-0.22	0.24	-0.05	0.07



Variable or Data Group	Overall Min.	Overall Max.	Str. 1 Min.	Str. 1 Max.	Str. 2 Min.	Str. 2 Max.	Str. 3 Min.	Str. 3 Max.	Str. 4 Min.	Str. 4 Max.	Str. 5 Min.	Str. 5 Max.	Str. 6 Min.	Str. 6 Max.
kid59	-0.07	0.01	-0.08	0.17	-0.12	-0.01	-0.19	0.20	-0.09	0.15	-0.09	0.08	-0.07	0.11
kidle14	-0.06	0.02	-0.07	0.14	-0.07	0.01	-0.12	0.12	-0.11	0.19	-0.08	0.13	-0.10	0.15
kidle19	-0.07	0.02	-0.08	0.13	-0.09	0.00	-0.15	0.17	-0.13	0.13	-0.18	0.27	-0.10	0.08
kidle9	-0.05	0.02	-0.12	0.16	-0.08	0.06	-0.11	0.09	-0.11	0.22	-0.08	0.10	-0.12	0.24
kidlt5	-0.01	0.06	-0.16	0.04	-0.07	0.14	-0.12	0.25	-0.08	0.22	-0.08	0.10	-0.13	0.24
latino	-0.02	0.05	-0.03	0.06	-0.07	0.06	-0.12	0.09	-0.06	0.09	-0.20	0.14	-0.05	0.07
maxdmd	-0.11	0.05	-0.08	0.11	-0.01	0.11	-0.11	0.11	-0.21	0.09	-0.18	0.08	-0.08	0.08
medage	-0.07	0.01	-0.06	0.04	-0.12	0.18	-0.13	0.21	-0.16	0.06	-0.12	0.07	-0.23	0.17
medinc	-0.01	0.07	-0.08	0.06	-0.05	0.17	-0.18	0.28	-0.13	0.11	-0.08	0.12	-0.07	0.06
no_bldgs	-0.03	0.00	-0.09	0.17	-0.04	0.01	-0.10	0.21	-0.07	0.07	-0.19	0.16	-0.04	0.05
no_floors	-0.04	0.00	-0.10	0.05	-0.10	0.14	-0.11	0.10	-0.27	0.08	-0.09	0.08	-0.11	0.10
no_res_units	-0.03	0.02	-0.10	0.06	-0.03	0.00	-0.12	0.14	-0.20	0.05	-0.09	0.09	-0.12	0.06
no_units	-0.04	0.01	-0.11	0.06	-0.03	0.01	-0.12	0.14	-0.23	0.05	-0.10	0.09	-0.06	0.07
occhsng	-0.02	0.04	-0.12	0.06	-0.07	0.03	-0.10	0.11	-0.11	0.11	-0.13	0.09	-0.09	0.09
poland	-0.01	0.06	-0.08	0.05	-0.01	0.14	-0.10	0.14	-0.09	0.19	-0.22	0.32	-0.04	0.02
popn	-0.03	0.01	-0.14	0.05	-0.05	0.02	-0.09	0.13	-0.13	0.15	-0.11	0.11	-0.08	0.06
srcit	-0.07	0.01	-0.06	0.07	-0.10	0.13	-0.13	0.20	-0.12	0.07	-0.14	0.12	-0.15	0.15
ssm1-ssm24	-0.01	0.02	-0.03	0.03	-0.03	0.04	-0.08	0.08	-0.03	0.04	-0.04	0.05	-0.04	0.05
tmp1-tmp24	-0.02	0.02	-0.03	0.02	-0.06	0.03	-0.07	0.10	-0.03	0.05	-0.07	0.06	-0.05	0.08
white	-0.02	0.05	-0.13	0.07	-0.09	0.03	-0.13	0.13	-0.12	0.12	-0.09	0.10	-0.07	0.05
windies	-0.04	0.02	-0.07	0.18	-0.03	0.05	-0.09	0.10	-0.12	0.04	-0.16	0.15	-0.07	0.09
wsp1-wsp24	-0.01	0.01	-0.02	0.02	-0.03	0.03	-0.05	0.04	-0.03	0.02	-0.04	0.04	-0.04	0.03
year_built	-0.06	0.04	-0.05	0.09	-0.06	0.06	-0.17	0.15	-0.17	0.11	-0.22	0.19	-0.18	0.04

## Appendix C: Summary Ranges for Moran's I and Geary's C Statistics

Tables C.1 and C.2 provide detailed results of the Geary's C and Moran's I statistics for residential and business customers, respectively. For variables with a temporal aspect, such as hourly and monthly data, there are minimum and maximum values of the Geary's C and Moran's I statistics, because these have been calculated for different hours of the year. A total of 20 day-hour combinations are used to calculate these values: 5 summer and 5 winter system peak day-hour combinations for the energy provider, and 10 randomly chosen day-hour combinations. Both raw energy usage interval data and four transformations are also examined, but the results are provided for the grouped values overall. Results are provided for each residential stratum as well as for all residential data as a whole.

Because the geodemographic variables related to the customer buildings and census tract data do not vary day-by-day or hour-by-hour, the minimum and maximum values for those variables are identical.

The complete set of Geary's C and Moran's I values is provided in a companion comma-separated data files to this thesis, called "graves\_geary\_moran.csv".

**Table C.1: Moran's I and Geary's C Ranges for Residential Customer Data Groups**

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
bapru	1	0.310	0.310	0.030	0.030
	2	0.246	0.246	0.034	0.034
	3	0.774	0.774	0.065	0.065
	4	0.113	0.113	0.080	0.080
	5	0.055	0.055	0.003	0.003
	6	-0.104	-0.104	1.231	1.231
	overall	0.314	0.314	0.040	0.040
bapu	1	0.253	0.253	0.030	0.030
	2	0.287	0.287	0.030	0.030
	3	0.417	0.417	0.051	0.051
	4	0.249	0.249	0.080	0.080
	5	0.164	0.164	0.002	0.002
	6	-0.111	-0.111	0.943	0.943
	overall	0.009	0.009	0.004	0.004
bldg_area	1	1.014	1.014	0.032	0.032
	2	1.367	1.367	0.040	0.040
	3	1.121	1.121	0.041	0.041
	4	1.403	1.403	0.067	0.067
	5	0.449	0.449	0.023	0.023
	6	-0.112	-0.112	0.946	0.946
	overall	1.242	1.242	0.042	0.042
delt1-delt24	1	-0.005	0.071	0.223	1.264
	2	-0.033	0.027	0.684	1.718
	3	-0.014	0.153	0.727	1.689
	4	-0.198	0.092	0.678	1.741
	5	-0.709	0.242	0.087	2.429
	6	-0.192	0.035	0.817	1.197
	overall	-0.009	0.033	0.438	1.117
far	1	0.807	0.807	0.175	0.175
	2	0.965	0.965	0.359	0.359
	3	0.894	0.894	0.260	0.260
	4	1.295	1.295	0.156	0.156
	5	0.319	0.319	0.022	0.022
	6	-0.116	-0.116	1.031	1.031
	overall	0.945	0.945	0.252	0.252
footprint	1	0.972	0.972	0.286	0.286
	2	1.376	1.376	0.507	0.507
	3	1.281	1.281	0.390	0.390
	4	0.485	0.485	0.502	0.502
	5	1.039	1.039	0.010	0.010
	6	-0.146	-0.146	1.046	1.046
	overall	1.279	1.279	0.415	0.415
no_bldgs	1	1.068	1.068	0.021	0.021
	2	0.805	0.805	0.032	0.032
	3	0.678	0.678	0.031	0.031
	4	0.054	0.054	0.118	0.118
	5	0.180	0.180	0.007	0.007
	6	-0.134	-0.134	0.989	0.989
	overall	0.962	0.962	0.040	0.040

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
no_floors	1	1.067	1.067	0.026	0.026
	2	1.347	1.347	0.032	0.032
	3	1.057	1.057	0.032	0.032
	4	1.360	1.360	0.065	0.065
	5	0.305	0.305	0.024	0.024
	6	-0.055	-0.055	0.880	0.880
	overall	1.283	1.283	0.034	0.034
no_res_units	1	0.954	0.954	0.032	0.032
	2	1.050	1.050	0.038	0.038
	3	0.799	0.799	0.036	0.036
	4	1.191	1.191	0.061	0.061
	5	0.083	0.083	0.028	0.028
	6	-0.104	-0.104	1.231	1.231
	overall	1.004	1.004	0.040	0.040
no_units	1	0.924	0.924	0.031	0.031
	2	1.015	1.015	0.035	0.035
	3	0.777	0.777	0.032	0.032
	4	1.133	1.133	0.058	0.058
	5	0.064	0.064	0.028	0.028
	6	-0.016	-0.016	0.894	0.894
	overall	0.975	0.975	0.038	0.038
kwh1-kwh24	1	-0.001	0.050	0.740	1.059
	2	-0.017	0.050	0.838	1.293
	3	-0.019	0.086	0.530	1.181
	4	-0.154	0.295	0.450	0.946
	5	-0.378	1.263	0.304	1.190
	6	-0.093	0.058	0.786	1.013
	overall	0.010	0.012	0.003	0.007
pcta1-pcta24	1	0.002	0.100	0.807	1.053
	2	-0.023	0.057	0.908	1.362
	3	-0.021	0.073	0.592	1.144
	4	-0.155	0.249	0.479	0.979
	5	-0.427	0.959	0.271	1.390
	6	-0.250	-0.080	0.841	1.314
	overall	0.011	0.065	0.833	1.094
pctd1-pctd24	1	0.004	0.113	0.867	1.016
	2	-0.015	0.139	0.820	1.059
	3	0.006	0.159	0.781	1.008
	4	-0.139	0.450	0.683	1.102
	5	-0.581	0.801	0.087	1.693
	6	-0.230	-0.013	0.936	1.242
	overall	0.005	0.078	0.906	1.049
pctm1-pctm24	1	-0.012	0.038	0.602	1.200
	2	-0.039	0.071	0.573	1.364
	3	-0.060	0.094	0.926	1.095
	4	-0.148	0.200	0.579	1.159
	5	-0.684	0.283	0.387	1.998
	6	-0.252	-0.008	0.944	1.294
	overall	-0.005	0.041	0.767	1.239
year_built	1	0.959	0.959	0.026	0.026
	2	0.885	0.885	0.028	0.028

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	3	0.664	0.664	0.023	0.023
	4	0.735	0.735	0.059	0.059
	5	0.659	0.659	0.005	0.005
	6	0.066	0.066	0.875	0.875
	overall	0.935	0.935	0.031	0.031
kidlt5	1	0.796	0.796	0.022	0.022
	2	0.826	0.826	0.024	0.024
	3	0.691	0.691	0.028	0.028
	4	0.985	0.985	0.044	0.044
	5	0.062	0.062	0.008	0.008
	6	-0.049	-0.049	0.790	0.790
	overall	0.420	0.420	0.016	0.016
kid59	1	0.725	0.725	0.012	0.012
	2	0.998	0.998	0.012	0.012
	3	0.749	0.749	0.011	0.011
	4	0.733	0.733	0.015	0.015
	5	0.378	0.378	0.004	0.004
	6	0.038	0.038	0.823	0.823
	overall	0.795	0.795	0.013	0.013
kid1014	1	0.654	0.654	0.011	0.011
	2	0.673	0.673	0.010	0.010
	3	0.585	0.585	0.008	0.008
	4	0.519	0.519	0.008	0.008
	5	0.372	0.372	0.002	0.002
	6	-0.019	-0.019	0.766	0.766
	overall	0.596	0.596	0.010	0.010
kid1519	1	0.857	0.857	0.017	0.017
	2	0.881	0.881	0.019	0.019
	3	0.597	0.597	0.016	0.016
	4	1.049	1.049	0.034	0.034
	5	0.209	0.209	0.024	0.024
	6	-0.087	-0.087	0.936	0.936
	overall	0.832	0.832	0.019	0.019
kidle9	1	0.558	0.558	0.008	0.008
	2	0.610	0.610	0.008	0.008
	3	0.357	0.357	0.007	0.007
	4	0.206	0.206	0.008	0.008
	5	0.380	0.380	0.003	0.003
	6	-0.066	-0.066	0.835	0.835
	overall	0.313	0.313	0.007	0.007
kidle14	1	0.592	0.592	0.006	0.006
	2	0.704	0.704	0.005	0.005
	3	0.481	0.481	0.004	0.004
	4	0.326	0.326	0.004	0.004
	5	0.559	0.559	0.002	0.002
	6	-0.068	-0.068	0.859	0.859
	overall	0.437	0.437	0.005	0.005
kidle19	1	0.683	0.683	0.009	0.009
	2	0.814	0.814	0.008	0.008
	3	0.586	0.586	0.007	0.007
	4	0.572	0.572	0.009	0.009

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	5	0.353	0.353	0.003	0.003
	6	-0.059	-0.059	0.882	0.882
	overall	0.614	0.614	0.008	0.008
medage	1	1.211	1.211	0.007	0.007
	2	0.751	0.751	0.008	0.008
	3	0.784	0.784	0.009	0.009
	4	0.430	0.430	0.010	0.010
	5	0.259	0.259	0.002	0.002
	6	-0.044	-0.044	0.834	0.834
	overall	0.844	0.844	0.008	0.008
srcit	1	1.163	1.163	0.006	0.006
	2	0.750	0.750	0.006	0.006
	3	0.692	0.692	0.006	0.006
	4	0.350	0.350	0.007	0.007
	5	0.312	0.312	0.001	0.001
	6	-0.060	-0.060	0.873	0.873
	overall	0.940	0.940	0.006	0.006
medinc	1	0.968	0.968	0.021	0.021
	2	1.245	1.245	0.024	0.024
	3	0.986	0.986	0.024	0.024
	4	1.576	1.576	0.061	0.061
	5	0.085	0.085	0.011	0.011
	6	-0.033	-0.033	0.962	0.962
	overall	1.185	1.185	0.027	0.027
white	1	0.524	0.524	0.007	0.007
	2	0.407	0.407	0.006	0.006
	3	0.265	0.265	0.006	0.006
	4	0.175	0.175	0.006	0.006
	5	0.494	0.494	0.003	0.003
	6	-0.086	-0.086	0.905	0.905
	overall	0.377	0.377	0.007	0.007
black	1	0.367	0.367	0.007	0.007
	2	0.104	0.104	0.004	0.004
	3	0.078	0.078	0.004	0.004
	4	0.036	0.036	0.004	0.004
	5	0.128	0.128	0.002	0.002
	6	-0.025	-0.025	0.754	0.754
	overall	0.129	0.129	0.004	0.004
asian	1	0.677	0.677	0.005	0.005
	2	0.660	0.660	0.005	0.005
	3	0.464	0.464	0.004	0.004
	4	0.197	0.197	0.005	0.005
	5	0.436	0.436	0.003	0.003
	6	-0.034	-0.034	0.888	0.888
	overall	0.612	0.612	0.005	0.005
latino	1	0.685	0.685	0.011	0.011
	2	0.665	0.665	0.009	0.009
	3	0.557	0.557	0.009	0.009
	4	0.576	0.576	0.021	0.021
	5	0.006	0.006	0.004	0.004
	6	0.005	0.005	0.880	0.880

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	overall	0.665	0.665	0.011	0.011
america	1	0.858	0.858	0.007	0.007
	2	0.691	0.691	0.010	0.010
	3	0.530	0.530	0.012	0.012
	4	0.585	0.585	0.024	0.024
	5	0.159	0.159	0.010	0.010
	6	-0.056	-0.056	1.036	1.036
	overall	0.759	0.759	0.010	0.010
guyana	1	0.177	0.177	0.002	0.002
	2	0.144	0.144	0.003	0.003
	3	0.031	0.031	0.002	0.002
	4	0.052	0.052	0.003	0.003
	5	0.151	0.151	0.001	0.001
	6	-0.059	-0.059	0.928	0.928
	overall	0.104	0.104	0.003	0.003
ireland	1	0.910	0.910	0.011	0.011
	2	0.809	0.809	0.013	0.013
	3	0.695	0.695	0.013	0.013
	4	0.677	0.677	0.021	0.021
	5	0.002	0.002	0.004	0.004
	6	-0.039	-0.039	0.776	0.776
	overall	0.668	0.668	0.012	0.012
italy	1	0.582	0.582	0.006	0.006
	2	0.266	0.266	0.005	0.005
	3	0.115	0.115	0.003	0.003
	4	0.045	0.045	0.004	0.004
	5	0.022	0.022	0.002	0.002
	6	-0.039	-0.039	0.741	0.741
	overall	0.274	0.274	0.005	0.005
poland	1	0.858	0.858	0.016	0.016
	2	1.000	1.000	0.018	0.018
	3	0.797	0.797	0.018	0.018
	4	1.175	1.175	0.038	0.038
	5	0.028	0.028	0.012	0.012
	6	-0.064	-0.064	1.029	1.029
	overall	0.942	0.942	0.020	0.020
windies	1	0.124	0.124	0.004	0.004
	2	0.045	0.045	0.003	0.003
	3	0.024	0.024	0.002	0.002
	4	0.039	0.039	0.003	0.003
	5	0.112	0.112	0.001	0.001
	6	-0.006	-0.006	0.728	0.728
	overall	0.039	0.039	0.002	0.002
jamaica	1	0.081	0.081	0.003	0.003
	2	0.034	0.034	0.002	0.002
	3	0.018	0.018	0.002	0.002
	4	0.033	0.033	0.003	0.003
	5	0.209	0.209	0.009	0.009
	6	-0.030	-0.030	0.739	0.739
	overall	0.032	0.032	0.002	0.002
foreign	1	0.792	0.792	0.005	0.005

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	2	0.786	0.786	0.005	0.005
	3	0.590	0.590	0.004	0.004
	4	0.343	0.343	0.006	0.006
	5	0.314	0.314	0.003	0.003
	6	0.009	0.009	0.778	0.778
	overall	0.750	0.750	0.005	0.005
elecheat	1	1.069	1.069	0.020	0.020
	2	1.269	1.269	0.024	0.024
	3	1.028	1.028	0.024	0.024
	4	1.574	1.574	0.055	0.055
	5	0.001	0.001	0.024	0.024
	6	0.145	0.145	0.754	0.754
	overall	1.258	1.258	0.026	0.026
hhsz	1	0.590	0.590	0.007	0.007
	2	0.488	0.488	0.007	0.007
	3	0.311	0.311	0.006	0.006
	4	0.200	0.200	0.005	0.005
	5	0.698	0.698	0.003	0.003
	6	-0.137	-0.137	1.014	1.014
	overall	0.416	0.416	0.007	0.007
edlths	1	0.805	0.805	0.005	0.005
	2	0.850	0.850	0.006	0.006
	3	0.665	0.665	0.005	0.005
	4	0.563	0.563	0.009	0.009
	5	0.259	0.259	0.003	0.003
	6	0.129	0.129	0.810	0.810
	overall	0.814	0.814	0.006	0.006
edhs	1	0.869	0.869	0.023	0.023
	2	0.960	0.960	0.022	0.022
	3	0.882	0.882	0.026	0.026
	4	1.267	1.267	0.044	0.044
	5	0.023	0.023	0.014	0.014
	6	-0.098	-0.098	0.905	0.905
	overall	0.949	0.949	0.026	0.026
edsncoll	1	1.050	1.050	0.005	0.005
	2	0.663	0.663	0.004	0.004
	3	0.579	0.579	0.003	0.003
	4	0.326	0.326	0.003	0.003
	5	0.534	0.534	0.002	0.002
	6	-0.099	-0.099	0.855	0.855
	overall	0.737	0.737	0.005	0.005
edcoll	1	0.921	0.921	0.009	0.009
	2	0.956	0.956	0.009	0.009
	3	0.829	0.829	0.009	0.009
	4	0.914	0.914	0.017	0.017
	5	0.243	0.243	0.005	0.005
	6	-0.066	-0.066	0.845	0.845
	overall	0.953	0.953	0.011	0.011
edlehs	1	0.854	0.854	0.011	0.011
	2	1.001	1.001	0.012	0.012
	3	0.849	0.849	0.012	0.012



Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	4	1.058	1.058	0.026	0.026
	5	0.123	0.123	0.007	0.007
	6	-0.049	-0.049	0.865	0.865
	overall	0.962	0.962	0.014	0.014
edlecoll	1	0.922	0.922	0.009	0.009
	2	0.957	0.957	0.009	0.009
	3	0.829	0.829	0.009	0.009
	4	0.913	0.913	0.017	0.017
	5	0.244	0.244	0.005	0.005
	6	-0.066	-0.066	0.845	0.845
	overall	0.954	0.954	0.011	0.011

**Table C.2: Moran's I and Geary's C Ranges for Business Customer Data Groups**

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
bapru	1	0.452	0.452	0.062	0.062
	2	0.187	0.187	0.104	0.104
	3	0.129	0.129	0.035	0.035
	4	0.073	0.073	0.028	0.028
	5	0.935	0.935	0.072	0.072
	6	0.127	0.127	0.039	0.039
	overall	0.049	0.049	0.019	0.019
bapu	1	0.111	0.111	0.038	0.038
	2	0.035	0.035	0.006	0.006
	3	1.097	1.097	0.832	0.832
	4	0.240	0.240	0.044	0.044
	5	0.992	0.992	0.077	0.077
	6	0.728	0.728	0.061	0.061
	overall	0.283	0.283	0.036	0.036
bldg_area	1	0.113	0.113	0.072	0.072
	2	0.913	0.913	0.031	0.031
	3	0.553	0.553	0.073	0.073
	4	1.688	1.688	0.067	0.067
	5	0.845	0.845	0.084	0.084
	6	1.348	1.348	0.042	0.042
	overall	0.347	0.347	0.036	0.036
delt1-delt24	1	-0.155	0.239	0.301	1.427
	2	-0.249	0.608	0.149	1.850
	3	-0.300	1.166	0.170	2.147
	4	-0.271	0.044	0.004	0.950
	5	-0.222	0.169	0.090	0.784
	6	-0.188	0.112	0.091	1.497
	overall	-0.138	0.241	0.025	5.125
far	1	0.727	0.727	0.077	0.077
	2	1.341	1.341	0.062	0.062
	3	0.750	0.750	0.049	0.049
	4	0.510	0.510	0.043	0.043
	5	0.894	0.894	0.130	0.130
	6	1.364	1.364	0.065	0.065
	overall	0.945	0.945	0.094	0.094
footprint	1	0.073	0.073	0.070	0.070
	2	0.051	0.051	0.005	0.005
	3	0.578	0.578	0.080	0.080
	4	0.552	0.552	0.029	0.029
	5	0.287	0.287	0.082	0.082
	6	0.180	0.180	0.049	0.049
	overall	0.061	0.061	0.026	0.026
no_bldgs	1	0.122	0.122	0.108	0.108
	2	0.048	0.048	0.010	0.010
	3	0.495	0.495	0.014	0.014
	4	0.087	0.087	0.008	0.008
	5	0.816	0.816	0.077	0.077
	6	0.124	0.124	0.047	0.047
	overall	0.071	0.071	0.027	0.027

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
no_floors	1	0.232	0.232	0.074	0.074
	2	0.781	0.781	0.043	0.043
	3	0.126	0.126	0.048	0.048
	4	0.134	0.134	0.038	0.038
	5	0.099	0.099	0.144	0.144
	6	0.065	0.065	0.065	0.065
	overall	0.508	0.508	0.107	0.107
no_res_units	1	0.074	0.074	0.064	0.064
	2	0.023	0.023	0.004	0.004
	3	0.088	0.088	0.018	0.018
	4	0.071	0.071	0.027	0.027
	5	0.073	0.073	0.145	0.145
	6	0.011	0.011	0.048	0.048
	overall	0.225	0.225	0.060	0.060
no_units	1	0.081	0.081	0.065	0.065
	2	0.022	0.022	0.004	0.004
	3	0.118	0.118	0.020	0.020
	4	0.044	0.044	0.028	0.028
	5	0.077	0.077	0.139	0.139
	6	0.060	0.060	0.051	0.051
	overall	0.204	0.204	0.060	0.060
kwh1-kwh24	1	0.006	0.233	0.096	0.322
	2	-0.094	0.328	0.430	1.088
	3	-0.036	1.123	0.086	1.181
	4	-0.341	0.053	0.726	1.436
	5	0.072	0.850	0.126	0.270
	6	0.130	1.130	0.176	1.688
	overall	0.600	0.661	0.082	0.153
pcta1-pcta24	1	0.151	0.380	0.298	0.891
	2	-0.150	0.537	0.675	1.143
	3	-0.040	1.118	0.190	1.272
	4	-0.128	0.103	0.483	1.254
	5	-0.107	0.726	0.175	0.623
	6	-0.285	0.238	0.202	2.121
	overall	0.026	0.523	0.535	1.049
pctd1-pctd24	1	0.016	0.379	0.355	1.001
	2	-0.148	0.469	0.621	1.233
	3	-0.109	1.232	0.140	1.501
	4	-0.172	0.208	0.054	1.200
	5	0.009	0.328	0.151	0.248
	6	-0.257	0.435	0.089	2.149
	overall	0.012	0.425	0.698	1.075
pctm1-pctm24	1	0.092	0.425	0.297	0.756
	2	-0.278	0.129	0.413	1.543
	3	-0.074	1.081	0.309	1.433
	4	-0.091	0.170	0.374	1.176
	5	-0.113	0.705	0.182	0.559
	6	-0.229	0.232	0.196	1.790
	overall	0.012	0.547	0.559	1.185
year_built	1	1.165	1.165	0.179	0.179
	2	0.226	0.226	0.094	0.094

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	3	0.256	0.256	0.242	0.242
	4	0.658	0.658	0.032	0.032
	5	1.576	1.576	0.124	0.124
	6	0.181	0.181	0.052	0.052
	overall	0.948	0.948	0.149	0.149
kidlt5	1	1.039	1.039	0.056	0.056
	2	0.761	0.761	0.036	0.036
	3	1.381	1.381	0.030	0.030
	4	0.028	0.028	0.029	0.029
	5	0.064	0.064	0.110	0.110
	6	0.994	0.994	0.067	0.067
	overall	0.927	0.927	0.070	0.070
kid59	1	0.513	0.513	0.023	0.023
	2	0.539	0.539	0.019	0.019
	3	1.020	1.020	0.015	0.015
	4	0.039	0.039	0.019	0.019
	5	0.344	0.344	0.069	0.069
	6	0.124	0.124	0.052	0.052
	overall	0.646	0.646	0.038	0.038
kid1014	1	0.621	0.621	0.032	0.032
	2	0.907	0.907	0.034	0.034
	3	1.112	1.112	0.011	0.011
	4	0.032	0.032	0.032	0.032
	5	0.216	0.216	0.119	0.119
	6	0.208	0.208	0.067	0.067
	overall	0.924	0.924	0.057	0.057
kid1519	1	0.860	0.860	0.046	0.046
	2	0.924	0.924	0.035	0.035
	3	1.408	1.408	0.030	0.030
	4	0.072	0.072	0.020	0.020
	5	0.941	0.941	0.070	0.070
	6	0.094	0.094	0.056	0.056
	overall	0.576	0.576	0.039	0.039
kidle9	1	0.023	0.023	0.195	0.195
	2	0.010	0.010	0.181	0.181
	3	0.007	0.007	0.087	0.087
	4	0.013	0.013	0.083	0.083
	5	0.072	0.072	0.098	0.098
	6	0.057	0.057	0.841	0.841
	overall	0.037	0.037	0.239	0.239
kidle14	1	0.368	0.368	0.017	0.017
	2	0.252	0.252	0.009	0.009
	3	0.306	0.306	0.005	0.005
	4	0.011	0.011	0.016	0.016
	5	0.307	0.307	0.069	0.069
	6	0.766	0.766	0.058	0.058
	overall	0.462	0.462	0.034	0.034
kidle19	1	0.606	0.606	0.026	0.026
	2	0.466	0.466	0.016	0.016
	3	0.666	0.666	0.008	0.008
	4	0.037	0.037	0.019	0.019

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	5	1.000	1.000	0.080	0.080
	6	0.523	0.523	0.056	0.056
	overall	0.744	0.744	0.039	0.039
medage	1	0.311	0.311	0.018	0.018
	2	0.535	0.535	0.020	0.020
	3	0.351	0.351	0.010	0.010
	4	0.117	0.117	0.005	0.005
	5	0.649	0.649	0.060	0.060
	6	1.395	1.395	0.054	0.054
	overall	0.544	0.544	0.028	0.028
srcit	1	0.368	0.368	0.015	0.015
	2	0.658	0.658	0.017	0.017
	3	0.229	0.229	0.009	0.009
	4	0.522	0.522	0.009	0.009
	5	1.430	1.430	0.058	0.058
	6	0.479	0.479	0.054	0.054
	overall	0.750	0.750	0.028	0.028
medinc	1	1.630	1.630	0.069	0.069
	2	1.789	1.789	0.072	0.072
	3	2.295	2.295	0.034	0.034
	4	0.327	0.327	0.050	0.050
	5	0.374	0.374	0.109	0.109
	6	0.112	0.112	0.074	0.074
	overall	2.019	2.019	0.104	0.104
white	1	0.380	0.380	0.016	0.016
	2	0.719	0.719	0.018	0.018
	3	0.416	0.416	0.012	0.012
	4	0.719	0.719	0.029	0.029
	5	0.628	0.628	0.063	0.063
	6	0.194	0.194	0.050	0.050
	overall	0.586	0.586	0.039	0.039
black	1	0.164	0.164	0.010	0.010
	2	0.274	0.274	0.010	0.010
	3	0.077	0.077	0.004	0.004
	4	0.203	0.203	0.010	0.010
	5	0.469	0.469	0.048	0.048
	6	0.083	0.083	0.035	0.035
	overall	0.124	0.124	0.020	0.020
asian	1	0.065	0.065	0.013	0.013
	2	0.219	0.219	0.008	0.008
	3	0.407	0.407	0.011	0.011
	4	0.227	0.227	0.019	0.019
	5	0.462	0.462	0.055	0.055
	6	0.304	0.304	0.062	0.062
	overall	0.218	0.218	0.030	0.030
latino	1	0.409	0.409	0.033	0.033
	2	1.262	1.262	0.028	0.028
	3	0.926	0.926	0.016	0.016
	4	0.814	0.814	0.035	0.035
	5	2.851	2.851	0.080	0.080
	6	0.787	0.787	0.064	0.064

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	overall	1.395	1.395	0.065	0.065
america	1	0.270	0.270	0.018	0.018
	2	0.467	0.467	0.022	0.022
	3	0.359	0.359	0.017	0.017
	4	0.376	0.376	0.022	0.022
	5	0.487	0.487	0.087	0.087
	6	0.617	0.617	0.060	0.060
	overall	0.510	0.510	0.050	0.050
guyana	1	0.144	0.144	0.010	0.010
	2	0.060	0.060	0.004	0.004
	3	0.031	0.031	0.002	0.002
	4	0.119	0.119	0.025	0.025
	5	0.174	0.174	0.070	0.070
	6	0.156	0.156	0.042	0.042
	overall	0.097	0.097	0.020	0.020
ireland	1	0.732	0.732	0.030	0.030
	2	0.913	0.913	0.029	0.029
	3	1.160	1.160	0.018	0.018
	4	0.538	0.538	0.046	0.046
	5	0.679	0.679	0.081	0.081
	6	0.443	0.443	0.070	0.070
	overall	1.255	1.255	0.066	0.066
italy	1	0.601	0.601	0.021	0.021
	2	0.917	0.917	0.024	0.024
	3	0.373	0.373	0.014	0.014
	4	0.610	0.610	0.020	0.020
	5	0.689	0.689	0.053	0.053
	6	0.470	0.470	0.050	0.050
	overall	0.326	0.326	0.030	0.030
poland	1	1.231	1.231	0.054	0.054
	2	1.520	1.520	0.052	0.052
	3	0.092	0.092	0.009	0.009
	4	0.247	0.247	0.021	0.021
	5	0.180	0.180	0.059	0.059
	6	1.208	1.208	0.051	0.051
	overall	0.433	0.433	0.040	0.040
windies	1	0.137	0.137	0.007	0.007
	2	0.045	0.045	0.005	0.005
	3	0.019	0.019	0.002	0.002
	4	0.028	0.028	0.003	0.003
	5	0.056	0.056	0.033	0.033
	6	0.073	0.073	0.030	0.030
	overall	0.068	0.068	0.013	0.013
jamaica	1	0.284	0.284	0.018	0.018
	2	0.101	0.101	0.008	0.008
	3	0.018	0.018	0.002	0.002
	4	0.084	0.084	0.008	0.008
	5	0.155	0.155	0.048	0.048
	6	0.031	0.031	0.030	0.030
	overall	0.045	0.045	0.013	0.013
foreign	1	0.184	0.184	0.012	0.012

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	2	0.379	0.379	0.009	0.009
	3	0.288	0.288	0.011	0.011
	4	0.636	0.636	0.021	0.021
	5	0.935	0.935	0.066	0.066
	6	2.019	2.019	0.061	0.061
	overall	0.714	0.714	0.037	0.037
elecheat	1	1.926	1.926	0.080	0.080
	2	1.951	1.951	0.075	0.075
	3	2.326	2.326	0.033	0.033
	4	0.223	0.223	0.051	0.051
	5	0.149	0.149	0.127	0.127
	6	0.044	0.044	0.089	0.089
	overall	2.437	2.437	0.115	0.115
hhsz	1	0.327	0.327	0.028	0.028
	2	1.345	1.345	0.028	0.028
	3	0.734	0.734	0.016	0.016
	4	1.073	1.073	0.037	0.037
	5	2.143	2.143	0.089	0.089
	6	0.203	0.203	0.073	0.073
	overall	1.490	1.490	0.070	0.070
edlths	1	0.454	0.454	0.022	0.022
	2	1.361	1.361	0.028	0.028
	3	0.974	0.974	0.017	0.017
	4	0.957	0.957	0.037	0.037
	5	3.246	3.246	0.096	0.096
	6	0.235	0.235	0.074	0.074
	overall	1.525	1.525	0.069	0.069
edhs	1	0.817	0.817	0.050	0.050
	2	1.078	1.078	0.037	0.037
	3	1.830	1.830	0.021	0.021
	4	0.470	0.470	0.041	0.041
	5	0.193	0.193	0.079	0.079
	6	0.026	0.026	0.071	0.071
	overall	1.122	1.122	0.064	0.064
edsncoll	1	0.294	0.294	0.017	0.017
	2	0.308	0.308	0.010	0.010
	3	0.429	0.429	0.006	0.006
	4	0.004	0.004	0.018	0.018
	5	0.239	0.239	0.074	0.074
	6	0.281	0.281	0.047	0.047
	overall	0.490	0.490	0.029	0.029
edcoll	1	0.581	0.581	0.028	0.028
	2	1.215	1.215	0.028	0.028
	3	1.478	1.478	0.018	0.018
	4	0.778	0.778	0.043	0.043
	5	1.092	1.092	0.101	0.101
	6	0.025	0.025	0.074	0.074
	overall	1.544	1.544	0.068	0.068
edlehs	1	0.594	0.594	0.030	0.030
	2	1.369	1.369	0.031	0.031
	3	1.430	1.430	0.019	0.019

Variable	Stratum	Moran Minimum Value	Moran Maximum Value	Geary Minimum Value	Geary Maximum Value
	4	0.855	0.855	0.042	0.042
	5	1.675	1.675	0.099	0.099
	6	0.102	0.102	0.078	0.078
	overall	1.568	1.568	0.073	0.073
edlecoll	1	0.580	0.580	0.028	0.028
	2	1.215	1.215	0.028	0.028
	3	1.479	1.479	0.018	0.018
	4	0.778	0.778	0.043	0.043
	5	1.092	1.092	0.101	0.101
	6	0.025	0.025	0.074	0.074
	overall	1.544	1.544	0.068	0.068



## Appendix D: Customer Sampling for Gap-Filling

The tables below show the residential and business customers sampled for each gap-filling length, along with the the number of hourly energy usage intervals, and the start and end dates and hours for the gap.

**Table D.1: Residential Customer Sample for Gap-Filling of System Peak Hour**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54308	SPH	1	July 22	16	July 22	16
54548	SPH	1	July 22	16	July 22	16
67965	SPH	1	July 22	16	July 22	16
67992	SPH	1	July 22	16	July 22	16
68863	SPH	1	July 22	16	July 22	16
69038	SPH	1	July 22	16	July 22	16
69117	SPH	1	July 22	16	July 22	16
70508	SPH	1	July 22	16	July 22	16
70802	SPH	1	July 22	16	July 22	16
72575	SPH	1	July 22	16	July 22	16

**Table D.2: Business Customer Sample for Gap-Filling of System Peak Hour**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54308	SPH	1	July 22	16	July 22	16
54548	SPH	1	July 22	16	July 22	16
67965	SPH	1	July 22	16	July 22	16
67992	SPH	1	July 22	16	July 22	16
68863	SPH	1	July 22	16	July 22	16
69038	SPH	1	July 22	16	July 22	16
69117	SPH	1	July 22	16	July 22	16
70508	SPH	1	July 22	16	July 22	16
70802	SPH	1	July 22	16	July 22	16
72575	SPH	1	July 22	16	July 22	16

**Table D.3: Residential Customer Sample for Gap-Filling of Customer Peak Hour**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5190	CPH	1	July 29	21	July 29	21
61717	CPH	1	August 10	9	August 10	9
67735	CPH	1	October 9	13	October 9	13
68516	CPH	1	July 28	20	July 28	20
69863	CPH	1	May 30	16	May 30	16
70797	CPH	1	February 15	21	February 15	21
70813	CPH	1	January 8	12	January 8	12
72732	CPH	1	July 23	21	July 23	21
73847	CPH	1	July 31	15	July 31	15
596253	CPH	1	December 19	22	December 19	22

**Table D.4: Business Customer Sample for Gap-Filling of Customer Peak Hour**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5316	CPH	1	June 9	16	June 9	16
5346	CPH	1	July 22	12	July 22	12
5571	CPH	1	July 22	15	July 22	15
69702	CPH	1	July 26	14	July 26	14
72359	CPH	1	July 22	15	July 22	15
764415	CPH	1	June 9	14	June 9	14
768095	CPH	1	January 14	14	January 14	14
768397	CPH	1	July 22	13	July 22	13
768602	CPH	1	July 22	21	July 22	21
768644	CPH	1	July 22	12	July 22	12

**Table D.5: Residential Customer Sample for Gap-Filling of One Hour**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5190	1HR	1	June 17	15	June 17	15
5224	1HR	1	January 27	13	January 27	13
54908	1HR	1	December 19	11	December 19	11
61709	1HR	1	January 21	21	January 21	21
67757	1HR	1	July 22	23	July 22	23
67962	1HR	1	November 17	10	November 17	10
68368	1HR	1	December 12	5	December 12	5
68866	1HR	1	July 25	6	July 25	6
73359	1HR	1	January 16	8	January 16	8
354137	1HR	1	February 5	24	February 5	24

**Table D.6: Business Customer Sample for Gap-Filling of One Hour**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5319	1HR	1	January 17	15	January 17	15
5319	1HR	1	October 23	13	October 23	13
5601	1HR	1	March 19	6	March 19	6
68859	1HR	1	March 29	11	March 29	11
72707	1HR	1	December 9	21	December 9	21
73885	1HR	1	October 17	9	October 17	9
768043	1HR	1	November 6	23	November 6	23
768136	1HR	1	November 3	10	November 3	10
768147	1HR	1	February 22	15	February 22	15
768632	1HR	1	August 14	5	August 14	5

**Table D.7: Residential Customer Sample for Gap-Filling of Three Hours**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54852	3HR	3	October 14	11	October 14	13
61707	3HR	3	November 2	19	November 2	21
68017	3HR	3	October 26	5	October 26	7
68180	3HR	3	May 31	21	May 31	23
68453	3HR	3	August 24	2	August 24	4
68607	3HR	3	April 10	1	April 10	3
72474	3HR	3	August 24	13	August 24	15
73855	3HR	3	October 22	1	October 22	3
154083	3HR	3	May 5	5	May 5	7
382740	3HR	3	August 29	21	August 29	23

**Table D.8: Business Customer Sample for Gap-Filling of Three Hours**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5577	3HR	3	October 12	12	October 12	14
65096	3HR	3	January 22	11	January 22	13
72697	3HR	3	September 10	19	September 10	21
73884	3HR	3	August 18	10	August 18	12
768168	3HR	3	September 16	5	September 16	7
768408	3HR	3	September 27	21	September 27	23
768515	3HR	3	June 9	21	June 9	23
768656	3HR	3	November 8	8	November 8	10
768712	3HR	3	July 26	2	July 26	4
768979	3HR	3	February 3	1	February 3	3

**Table D.9: Residential Customer Sample for Gap-Filling of 12 Hours**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54836	12H	12	July 25	12	July 25	23
65485	12H	12	January 3	3	January 3	14
69687	12H	12	April 5	13	April 5	24
69978	12H	12	July 19	8	July 19	19
72363	12H	12	July 20	7	July 20	18
72574	12H	12	September 2	12	September 2	23
73832	12H	12	August 15	11	August 15	22
154057	12H	12	October 11	11	October 11	22
154103	12H	12	April 15	3	April 15	14
382739	12H	12	September 1	11	September 1	22

**Table D.10: Business Customer Sample for Gap-Filling of 12 Hours**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5598	12H	12	September 7	12	September 7	23
71690	12H	12	July 5	8	July 5	19
72738	12H	12	March 5	6	March 5	17
72739	12H	12	February 19	8	February 19	19
73883	12H	12	August 19	13	August 19	24
565680	12H	12	December 13	6	December 13	17
765812	12H	12	February 10	3	February 10	14
768052	12H	12	September 8	3	September 8	14
768253	12H	12	October 18	4	October 18	15
768684	12H	12	November 10	1	November 10	12

**Table D.11: Residential Customer Sample for Gap-Filling of Customer Peak Day**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
67962	CPD	24	July 23	1	July 23	24
68147	CPD	24	January 16	1	January 16	24
69347	CPD	24	January 23	1	January 23	24
69455	CPD	24	August 27	1	August 27	24
69665	CPD	24	July 23	1	July 23	24
69837	CPD	24	July 31	1	July 31	24
72461	CPD	24	July 23	1	July 23	24
73777	CPD	24	July 21	1	July 21	24
73795	CPD	24	July 31	1	July 31	24
159612	CPD	24	July 23	1	July 23	24

**Table D.12: Business Customer Sample for Gap-Filling of Customer Peak Day**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5352	CPD	24	August 9	1	August 9	24
5357	CPD	24	July 22	1	July 22	24
68859	CPD	24	January 27	1	January 27	24
71760	CPD	24	July 20	1	July 20	24
765819	CPD	24	July 22	1	July 22	24
768030	CPD	24	December 6	1	December 6	24
768043	CPD	24	September 14	1	September 14	24
768684	CPD	24	June 19	1	June 19	24
768715	CPD	24	July 6	1	July 6	24
768768	CPD	24	December 8	1	December 8	24

**Table D.13: Residential Customer Sample for Gap-Filling of 24 Hours**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54309	24H	24	September 5	1	September 5	24
54985	24H	24	July 4	1	July 4	24
67809	24H	24	April 18	1	April 18	24
67884	24H	24	September 29	1	September 29	24
68208	24H	24	June 1	1	June 1	24
70258	24H	24	November 27	1	November 27	24
70802	24H	24	November 6	1	November 6	24
154081	24H	24	September 25	1	September 25	24
154158	24H	24	February 1	1	February 1	24
596343	24H	24	February 23	1	February 23	24

**Table D.14: Business Customer Sample for Gap-Filling of 24 Hours**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5571	24H	24	September 18	1	September 18	24
5573	24H	24	October 9	1	October 9	24
71266	24H	24	September 3	1	September 3	24
570792	24H	24	September 27	1	September 27	24
768060	24H	24	August 2	1	August 2	24
768088	24H	24	April 12	1	April 12	24
768168	24H	24	June 29	1	June 29	24
768219	24H	24	August 28	1	August 28	24
768434	24H	24	September 28	1	September 28	24
768775	24H	24	June 28	1	June 28	24

**Table D.15: Residential Customer Sample for Gap-Filling of 7 Days**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54838	7DY	168	December 21	1	December 27	24
54851	7DY	168	January 3	1	January 9	24
54910	7DY	168	July 10	1	July 16	24
67965	7DY	168	October 5	1	October 11	24
69163	7DY	168	July 13	1	July 19	24
69347	7DY	168	August 24	1	August 30	24
70213	7DY	168	March 23	1	March 29	24
70454	7DY	168	October 5	1	October 11	24
70970	7DY	168	July 26	1	August 1	24
72672	7DY	168	August 31	1	September 6	24

**Table D.16: Business Customer Sample for Gap-Filling of 7 Days**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5352	7DY	168	April 24	1	April 30	24
5573	7DY	168	January 6	1	January 12	24
5579	7DY	168	May 10	1	May 16	24
73885	7DY	168	January 29	1	February 4	24
332538	7DY	168	May 22	1	May 28	24
768151	7DY	168	December 17	1	December 23	24
768716	7DY	168	November 24	1	November 30	24
768797	7DY	168	July 20	1	July 26	24
768859	7DY	168	June 20	1	June 26	24
769254	7DY	168	September 24	1	September 30	24



**Table D.17: Residential Customer Sample for Gap-Filling of One Month**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
54287	1MO	720	January 28	1	February 26	24
55019	1MO	720	February 12	1	March 13	24
69467	1MO	720	January 27	1	February 25	24
69511	1MO	720	February 8	1	March 9	24
72362	1MO	720	February 13	1	March 14	24
73867	1MO	720	August 13	1	September 11	24
111748	1MO	720	July 17	1	August 15	24
154054	1MO	720	August 10	1	September 8	24
583327	1MO	720	April 23	1	May 22	24
596261	1MO	720	March 19	1	April 17	24

**Table D.18: Business Customer Sample for Gap-Filling of One Month**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5352	1MO	720	February 10	1	March 11	24
5518	1MO	720	January 31	1	March 1	24
5598	1MO	720	January 22	1	February 20	24
55318	1MO	720	July 29	1	August 27	24
68859	1MO	720	August 15	1	September 13	24
68879	1MO	720	February 9	1	March 10	24
70442	1MO	720	February 8	1	March 9	24
72739	1MO	720	February 7	1	March 8	24
481039	1MO	720	March 6	1	April 4	24
768511	1MO	720	April 10	1	May 9	24

**Table D.19: Residential Customer Sample for Gap-Filling of Three Months**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
74092	3MO	2,160	February 14	1	May 14	24
154011	3MO	2,160	June 21	1	September 18	24
154034	3MO	2,160	July 19	1	October 16	24
313070	3MO	2,160	July 18	1	October 15	24
382735	3MO	2,160	September 16	1	December 14	24
489726	3MO	2,160	January 7	1	April 6	24
596260	3MO	2,160	May 28	1	August 25	24
596267	3MO	2,160	May 31	1	August 28	24
596343	3MO	2,160	June 20	1	September 17	24
768139	3MO	2,160	March 29	1	June 26	24

**Table D.20: Business Customer Sample for Gap-Filling of Three Months**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
429118	3MO	2,160	October 28	1	January 25	24
764384	3MO	2,160	January 15	1	April 14	24
764403	3MO	2,160	July 25	1	October 22	24
768057	3MO	2,160	December 16	1	March 14	24
768091	3MO	2,160	May 26	1	August 23	24
768225	3MO	2,160	February 4	1	May 4	24
768712	3MO	2,160	September 10	1	December 8	24
768719	3MO	2,160	September 9	1	December 7	24
768786	3MO	2,160	August 17	1	November 14	24
769143	3MO	2,160	May 2	1	July 30	24

**Table D.21: Residential Customer Sample for Gap-Filling of Six Months**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
53054	6MO	4,344	April 29	1	October 26	24
154146	6MO	4,344	April 24	1	October 21	24
154153	6MO	4,344	April 24	1	October 21	24
596259	6MO	4,344	February 18	1	August 17	24
596266	6MO	4,344	June 23	1	December 20	24
596267	6MO	4,344	August 2	1	January 29	24
596330	6MO	4,344	November 29	1	May 27	24
596336	6MO	4,344	January 11	1	July 10	24
669493	6MO	4,344	September 10	1	March 8	24
764409	6MO	4,344	June 12	1	December 9	24

**Table D.22: Business Customer Sample for Gap-Filling of Six Months**

Customer ID	Usage Interval Gap Category	Number of Intervals in Gap Category	Start Date of Missing Intervals	Start Hour of Missing Intervals	End Date of Missing Intervals	End Hour of Missing Intervals
5535	6MO	4,344	March 28	1	September 24	24
597485	6MO	4,344	February 17	1	August 16	24
768054	6MO	4,344	December 23	1	June 20	24
768091	6MO	4,344	February 3	1	August 2	24
768147	6MO	4,344	April 13	1	October 10	24
768509	6MO	4,344	January 3	1	July 2	24
768632	6MO	4,344	November 30	1	May 28	24
768656	6MO	4,344	March 19	1	September 15	24
768959	6MO	4,344	July 27	1	January 23	24
769092	6MO	4,344	November 26	1	May 24	24

## References

- Albrechtson, Scott. 2009. *Probabilistic System Peak Modeling & Hourly Weather Model*. Presentation slides. Presented at Western Load Research Association Fall Conference.
- Andrienko, G., N. Andrienko, S. Bremm, T. Schreck, T. von Landesberger, P. Bak, and D. Keim. 2010. *Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns*. Eurographics/IEEE-VGTC Symposium on Visualization 2010, guest editors G. Melançon, T. Munzner, and D. Weiskopf, Volume 29, Number 3, p. 1-10.
- Anselin, Luc. 2006. "Spatial Regression," Chapter 14, *SAGE Handbook of Spatial Analysis*. London: SAGE Publications Inc.
- AEIC. 2001. Association of Edison Illuminating Companies, Load Research Committee, *Load Research Manual*, Second Edition, Birmingham, AL: Association of Edison Illuminating Companies.
- AEIC. 2010. *Advanced Applications in Load Research Seminar*. Class notes. Birmingham, AL: Association of Edison Illuminating Companies, October 4-7, 2010.
- AEIC. 2012. *Modeling Changes in Interval Period Demands in Response to Dynamic Prices: Description of the Almon Lag Estimator*. 2012 AEIC Advanced Applications in Load Research Seminar. Columbus, OH. October 2012.
- Auchincloss, Amy H., Ana V. Diez Roux, Daniel G. Brown, Trivellore E. Raghunathan, and Christine A. Erdmann. 2007. *Filling the Gaps: Spatial Interpolation of Residential Survey Data in the Estimation of Neighborhood Characteristics*. Epidemiology. Volume 18, Number 4, July 2007, p. 469-478.
- Charlton, Martin, and A. Stewart Fotheringham. 2009. *Geographically Weighted Regression White Paper*. Maynooth, Ireland: National Centre for Geocomputation, National University of Ireland, March 3, 2009.
- Chetty, Marshini, David Tran, and Rebecca E. Grinter. 2008. *Getting to Green: Understanding Resource Consumption in the Home*. International Conference on Ubiquitous Computing, September.
- Christakos, George, Patrick Bogaert, and Marc L. Serre. 2002. *Temporal GIS: Advanced Functions for Field-Based Applications*. New York: Springer.
- Cochran, William G. 1976. *Sampling Techniques*, third edition. New York: John Wiley & Sons.
- Cracknell, Kevin. 2009. *VEE Process Development*. Presentation slides. Presented at Western Load Research Association Spring Conference.

Dennis, Michael L., E. Jonathan Soderstrom, Walter S. Koncinski, Jr., and Betty Cavanaugh. 1990. *Effective Dissemination of Energy-Related Information*. American Psychologist, Volume 45, Number 10, October, p. 1109-1117.

Dirks, K. N., J. E. Hay, C. D. Snow, and D. Harris. 1998. *High-resolution Studies of Rainfall on Norfolk Island Part II: Interpolation of Rainfall Data*. Journal of Hydrology. Volume 208, p. 187-193.

Fels, Margaret F. 1986. "PRISM: An Introduction." *Energy and Buildings*, Volume 9, p. 5-18.

Foster, Derek, Shaun Lawson, Mark Blythe, and Paul Cairns. 2010. *Wattsup?: Motivating Reductions in Domestic Energy Consumption Using Social Networks*. Nordic Conference on Human-Computer Interaction, October.

Francis, Louise. 2001. *Neural Networks Demystified*. Presented at Casualty Actuarial Society, Winter Forum.

Goodchild, M. F. 2009. "Challenges in Spatial Analysis." In *The Sage Handbook of Spatial Analysis*, edited by S. A. Fotheringham and P. A. Rogerson, London: Sage.

Hartshorn, Truman A. 1992. *Interpreting the City: An Urban Geography*, second edition, New York: John Wiley & Sons, Inc.

Hayes, Steven C. and John D. Cone. 1977. *Reducing Residential Electrical Energy Use: Payments, Information, and Feedback*. Journal of Applied Behavior Analysis, Volume 10, Number 3, Fall, p. 425-435.

Hennessey, Tim. 2011. *PJM Empirical Analysis of Demand Response Baseline Methods*. Presentation slides. Presented at Western Load Research Association Fall Conference.

Holdaway, Margaret R. 1996. *Spatial Modeling and Interpolation of Monthly Temperature Using Kriging*. Climate Research. Volume 6, p. 215-225.

Jacquez, Geoffrey M. 1996. *A k Nearest Neighbor Test for Space-Time Interaction*. Statistics in Medicine. Volume 15, p. 1935-1996.

Kaplan, David, James Wheeler, and Steven Holloway. 2009. *Urban Geography*, second edition. Hoboken: John Wiley & Sons, Inc.

KEMA, Inc. 2011. *The RLW Load Research System: KEMA's Software System for Load Research and Program Evaluation*. Version 2.0, October 14, 2011.

Kirkeide, Loren. 2010. *Neighborhood Stratification in Studies with Volunteers*. Presented at Western Load Research Association Fall Conference.

Knox, E. G., and M. S. Bartlett. 1964. *The Detection of Space-Time Interactions*. Journal of the Royal Statistical Society, Series C (Applied Statistics). Volume 13, Number 1, p. 25-30.

Kulldorff, Martin. 1997. *A Spatial Scan Statistic*. *Communications in Statistics: Theory and Methods*. Volume 26, Number 6, p. 1481-1496.

Kulldorff, Martin. 2014. *SaTScan User Guide for Version 9.3*. <http://www.satscan.org/>. March 2014.

Lehmann, E. L., and H. J. M. D'Abrera. 1975. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day Series in Probability and Statistics. San Francisco: Holden-Day, Inc.

Lim, David, and Runming Yao. 2012. *A Combined Engineering and Statistical Model of UK Domestic Appliance Electrical Load Profiles*. Presentation slides. Presented at Western Load Research Association Spring Conference.

Lindström, Johan, Adam A. Szpiro, Paul D. Sampson, Lianne Sheppard, and Assaf Oran. 2011. *A Flexible Spatio-Temporal Model for Air Pollution: Allowing for Spatio-Temporal Covariates*. UW Biostatistics Working Paper Series. Working Paper 370.

Malizia, Nicholas and Elizabeth A. Mack. 2012. *Enhancing the Jacquez  $k$  Nearest Neighbor Test for Space-Time Interaction*. GeoDA Center for Geospatial Analysis and Computation, Working Paper Number 04. Tempe: Arizona State University.

Mankoff, Jennifer, Deanna Matthews, Susan R. Fussell, and Michael Johnson. 2007. *Leveraging Social Networks to Motivate Individuals to Reduce Their Ecological Footprints*. Carnegie Mellon University Research Showcase, Human-Computer Interaction Institute, Paper 47.

Manski, Charles F. 2000. *Economic Analysis of Social Interactions*. NBER Working Paper Series, Working Paper 7580. Cambridge, MA: National Bureau of Economic Research, March.

Mantel, Nathan. 1967. *The Detection of Disease Clustering and a Generalized Regression Approach*. *Cancer Research*. Volume 27, Number 2, February 1967, p. 209-220.

Mathis, Kent, Craig Williamson, and Bridget Kester. 2007. *VEE Testing at JEA*. Presentation slides. Presented at Western Load Research Association Fall Conference.

McCaffrey, James. 2014. "How to Standardize Data for Neural Networks." *Visual Studio Magazine*. <http://visualstudiomagazine.com/articles/2014/01/01/how-to-standardize-data-for-neural-networks.aspx>.

McMenamin, Stuart, and Frank Monforte. 1997. "Hourly Load Analysis Using Neural Networks," in *AEIC Load Research Committee, Report of the Load Research Committee 1997*. Birmingham, AL: Association of Edison Illuminating Companies.

- McMenamin, Stuart. 2011a. *Forecasting, Load Research, and Energy Efficiency*. Presentation slides. Presented at Western Load Research Association Fall Conference, September 14, 2011.
- McMenamin, Stuart. 2011b. *The Pros and Cons of Ratio Estimation*. Presentation slides. Presented at Western Load Research Association Fall Conference.
- Mondragon, Josh. 2009. *Estimating Missing Interval Data for Smart Meters*. Presentation slides. Presented at Western Load Research Association Spring Conference.
- Ned Levine & Associates and The National Institute of Justice. 2004. *CrimeStat III Version 3.0: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. November 2004.
- O'Sullivan, David, and David J. Unwin. 2010. *Geographic Information Analysis, Second Edition*. Hoboken, NJ: John Wiley & Sons, Inc.
- Ott, Thomas, and Frank Swiaczny. 2001. *Time-Integrative Geographic Information Systems: Management and Analysis of Spatio-Temporal Data*. New York: Springer-Verlag.
- Pebesma, Edzer. 2012. spactime: Spatio-Temporal Data in R. *Journal of Statistical Software*. Volume 51, Issue 7, November.
- Peschiera, Gabriel, John E. Taylor, and Jeffrey A. Siegel. 2010. *Response-Relapse Patterns of Building Occupant Electricity Consumption Following Exposure to Personal, Contextualized and Occupant Peer Network Utilization Data*. *Energy and Buildings*, Volume 42, p. 1329-1336.
- Pindyck, Robert S., and Daniel L. Rubinfeld. 1976. *Econometric Models and Economic Forecasts*. New York: McGraw-Hill Book Company.
- Putnam, Robert D. 1995. *Bowling Alone: America's Declining Social Capital*. *Journal of Democracy*, volume 6, number 1, January, p. 65-78.
- Raish, Carl. 2007. *Evaluation and Implementation of an Oil & Gas Profile in the ERCOT Market*. Presentation slides. Presented at Western Load Research Association Spring Conference, March 14, 2007.
- Richardson, Ian, Murray Thomson, David Infield, and Conor Clifford. 2010. "Domestic Electricity Use: A High-Resolution Energy Demand Model." *Energy and Buildings*, Volume 42, p. 1878-1887.
- Ritchie, J. R. Brent, Gordon H. G. McDougall, and John D. Claxton. 1981. *Complexities of Household Energy Consumption and Conservation*. *Journal of Consumer Research*, Volume 8, December, p. 233-242.



Rosenshein, Lauren, Lauren Scott, and Monica Pratt. 2011. "Finding a Meaningful Model: This Checklist Will Help You Evaluate Regression Models." *ArcUser*, Winter 2011, p. 40-45.

Rupp, Cathy. 2008. *Earth Hour Community Challenge -- How Well Did We Do?* Presentation slides. Presented at Western Load Research Association Fall Conference, September 26, 2008.

Sampson, Paul D., Adam A. Szpiro, Lianne Sheppard, Johan Lindström, and Joel D. Kaufman. 2009. *Pragmatic Estimation of a Spatio-Temporal Air Quality Model With Irregular Monitoring Data*. UW Biostatistics Working Paper Series. Working Paper 353. November.

SAS Institute Inc. 2009. *SAS/STAT® 9.1 User's Guide, Second Edition*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2012. *SAS/STAT® 12.1 User's Guide*. Cary, NC: SAS Institute Inc.

Schabenberger, Oliver, and Carol A. Gotway. 2005. *Statistical Methods for Spatial Data Analysis*. Texts in Statistical Science Series. New York: Chapman & Hall/CRC, Taylor & Francis Group.

Schiermeyer, Ken. 2006. *Regression Modeling for Dynamic Load Profiles*. Presented at Western Load Research Association Spring Conference.

Shai, Donna. 2006. *Income, Housing, and Fire Injuries: Census Tract Analysis*. Public Health Reports. Volume 121, March-April 2006, p. 149-154.

Shepard, Donald. 1968. *A Two-Dimensional Interpolation Function for Irregularly Spaced Data*. Proceedings: 1968 ACM National Conference.

Smith, Bob, and Dave Hanna. 2008. *Hourly Weather Normalization*. Presentation slides. Presented at Western Load Research Association Fall Conference, September 25, 2008.

Stern, Paul C. 1992. *What Psychology Knows About Energy Conservation*. American Psychologist, Volume 47, Number 10, October, p. 1224-1232.

Szpiro, Adam A., Paul D. Sampson, Lianne Sheppard, Thomas Lumley, Sara D. Adar, and Joel D. Kaufman. 2010. "Predicting Intra-Urban Variation in Air Pollution Concentrations With Complex Spatio-Temporal Dependencies." *Environmetrics*, Volume 21, p. 606-631.

Tobler, Walter. 1970. *A Computer Movie Simulating Urban Growth in the Detroit Region*. *Economic Geography*, Volume 46, Supplement: Proceedings, International Geographical Union Commission on Quantitative Methods, June 1970, p. 234-240.

Van Raaij, W. Fred, and Theo M. M. Verhallen. 1983. "A Behavioral Model of Residential Energy Use," *Journal of Economic Psychology*, Volume 3, p. 39-63.



Waldfoegel, Joel. 2010. "Who Benefits Whom in the Neighborhood? Demographics and Retail Product Geography," Chapter 6 in *Agglomeration Economics*. Chicago: University of Chicago Press.

Weber, Alfred. 1929. *Theory of the Location of Industry*. Chicago: University of Chicago Press, 1929, originally published in German in 1909.

Wicklin, Rick. 2014. "Creating a Basic Heat Map in SAS." *The DO Loop: Statistical Programming in SAS with an Emphasis on SAS/IML Programs*, <http://blogs.sas.com/content/iml/2014/08/18/heat-map-in-sas/>, August 18, 2014

Williamson, Craig. 2012. *Matchmaker, Matchmaker, Make Me a Match: When and Why Matched Control Groups Work Better*. Presentation slides. Presented at Western Load Research Association Spring Conference, March 8, 2012.

Wood, Stacey. 2010. *PG&E Advanced Metering Assessment Report*. Commissioned by the California Public Utilities Commission, dated September 2, 2010. Houston, TX: Structure Consulting Group, LLC.

Xu, Tracy. 2009. *Critical Peak Pricing Model*. Presentation slides. Presented at Western Load Research Association Spring Conference, March 13, 2009.

Yates, Suzanne M., and Elliot Aronson. 1983. "A Social Psychological Perspective on Energy Conservation in Residential Buildings," *American Psychologist*, Volume 38, Number 4, April, p. 435-444.

Yu, Hwa-Lung, Shang-Jen Yang, Hsin-Ju Yen, and George Christakos. 2011. "A Spatio-Temporal Climate-Based Model of Early Dengue Fever Warning in Southern Taiwan." *Stochastic Environmental Research and Risk Assessment*, Volume 25, p. 485-494.