**Advanced Regression Analysis to predict the Future Energy Consumption of Buildings in Canada with Categorical Variables**

**By Prajna Paramita Dash, PhD, Energy Modeling Engineer,** *Energent Incorporated*

1. **Abstract:**

Remarkable change in climate has led to an interest in how this will affect energy consumption in buildings. Many a time, it is difficult task to come up with a regression model that can accurately predict future energy consumption pattern. Automatizing derivation of set points for Cooling Degree Days (CDDs) and Heating Degree Days (HDDs) where weather is not a primary driving factor, had always been a challenge. In addition, recognizing a data point as an outlier in a multivariable consumption pattern requires an evaluation of dependency of consumption on many dependent factors such as weather, day-of-week, and related production variables for industrial, commercial, and institutional buildings. Presence of irrelevant data point in a dataset may result in poor goodness-of-fit values. To overcome these modeling drawbacks, we propose an optimized regression method with automated outlier detection for buildings in Canada. Different methods of outlier detections are tested and verified on consumption data ranging over a year. Results from proposed regression model with and without exclusions are also provided. Findings from automated outlier detection methods verify that goodness-of-fit values, R2 (Coefficient of Determination) and CV, RMSE (Coefficient of Variation, Root Mean Square Error) considerably increase for datasets with excluded outliers.

*Keywords: CDD, HDD, Regression, Exclusion, Outliers, R2, CV, modeling, consumption, prediction, baseline, inspection, energy, R, optimization*

1. **Literature Review:**

Prediction of building energy performance has become necessary to project building energy consumption. Many tools and techniques are available to predict energy performance of the building. Prediction of energy consumption helps building owners or energy managers to project their annual energy cost, assist designers to decide on the energy conservation measures and design based on the goal of projects for energy saving, and also enable decision makers to decide on how much energy they can save during life cycle of a construction project or in any change of operation. Reference [1] discusses applicability of regression analysis for predicting building energy consumption. Regression techniques identify relationships, in the form of mathematical models, between one or more independent variables and dependent variables using sets of raw data. Method developed in [1] is called Lean Energy Analysis (LEAN). LEAN develops inverse models for building energy performance and describes method to benchmark buildings. Though authors in [1] develop models during baseline period, the article doesn’t discuss how model performs beyond baseline period, sometimes referred as “Inspection Phase”. Reference [2] describes development of a multi-linear regression model to predict impact of building shape on total energy consumption in two different climate regions (col-dry and warm-marine). Seven building shapes were used for this purpose. Study in [2] concluded there is strong correlation between shape of the building and their location to the total energy consumption. However, researchers have not provided any strong evidence that this relationship can be generalized for other locations and climatic condition. A detailed analysis on drawbacks on recent energy consumption models is performed by researchers in reference [3]. Work in [3] proposes and develops a residential building energy consumption demand model based on Neural Network (NN). With this methodology, predicted results from the model show that energy consumption demand model is precise and reliable. However, consumption model developed in [3] considers 16 dependent variables; which is not very realistic. Implementation of such models with large set of data will not be very feasible. An elaborated analysis on regression models for energy use in air-conditioned office buildings in different climates in [4]. Reference [4] develops multiple regression models for 5 different climates: sever cold, cold, hot summer, and cold winter. A total of 12 key building design variables were identified through parametric and sensitivity analysis. Through analysis, researchers in [4] conclude that regression models developed for warm weather tends to have strong correlation between the annual building energy use and the design parameters. A more advanced predictive model based on decision tree is proposed in [5]. Decision tree method discussed in [5] can provide combination of as well as the threshold values that will lead to high building energy performance. Generation of decision tree is based on training data and the evaluation of decision tree based on training test data. Energy prediction method proposed in [5] is flexible enough for users who don’t require much computational knowledge. Results of using decision tree demonstrate that use of decision tree method can classify and predict energy demand levels accurately, identify and rank significant factors of building energy use levels automatically. However, energy consumption model developed in [5] doesn’t take into account consumption pattern during weekdays and weekends. A new type of regression method named as “Kernel Regression” for real-time building energy analysis is proposed in [6]. Regression method used in [6] uses dataset of power and weather measurements for 4 buildings over 1.5 years, and performance of the method is compared to a standard neural network algorithm. Findings in [6] claim that kernel smoothers can outperform results produced by neural networks. Moreover, kernel regression models have been found to provide an additional level of information about buildings and are capable of showing sensitivity of building’s power consumption to time and weather parameters. Main drawback of the method proposed in [6] is that kernel regression methods works best for small set of data. Researchers have not validated this modeling process with large set and they have not taken seasonal variation into account. To address these shortcomings, and develop a predictive model that can perform and predict consumption pattern with permissible deviance percentage from actual consumption, we propose an optimized regression model with advanced outlier detection method for predicting consumption pattern for buildings located at Canada.

Rest of the paper is organized as follows. Section 3 discusses mathematical modeling and formulation for energy consumption. Section 4 discusses methodologies, and implementation of the model developed in Section 4. Test Results from this development are presented in Section 5. Finally, Section 6 concludes the paper.

1. **Description of Data and Software Tools**

In our study, we have used two types of data for linear regression modeling, data with daily variation, and data with billing period variation. Since data points with billing period variation for a year are significantly lower compared to data with daily variation, exclusion operations are not performed on them. Figures Figu*re 1* and Fi*gure 2*, respectively show consumption data with billing period and daily variation. Outlier Exclusions are performed only on data with daily variation. As ASHARE (American Society of Heating, Refrigeration, and Air-Conditioning Engineers) and IPMVP (International Performance Measurement and Verification Protocol) mandates no more than 25% of data can be excluded, we made sure combined missing and excluded data don’t go beyond 25% as per standards. In case, more than 25% of required data are missing, modeling software (Programming environment, we have used is R) [12] will deliver an error message showing that there are no enough data for modeling. At present, we have set a cut-off of 85% of available data for modeling. If less than 85% data are present for the required baseline period, regression modeling will not be performed on that data set. Different data outlier detection methods have been explored. Multivariable Adaptive Regression Spline (MARS) package available R is used for generating initial fit. Function “Earth” in MARS package is used for generating set points for (Cooling Degree Day) CDDs and (Heating Degree Days) HDDs. In some cases, when consumption data don’t follow regular consumption pattern, more than one CDD set points, in Case of electricity consumption and more than one HDD set points, in case of natural gas consumption. There are some cases function earth results more than two Cooling Degree Days that is not a very realistic case, therefore, when earth results in such drivers, the modeling has to be redone with different baseline period.

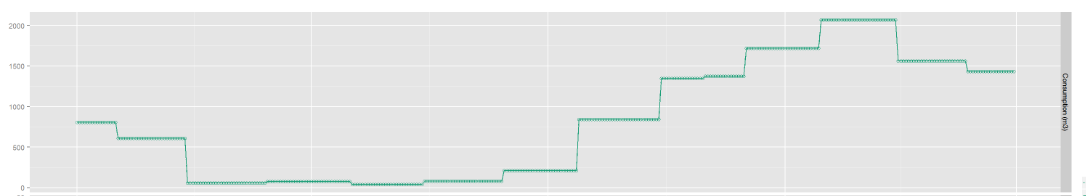


Figure 1 Energy Consumption Data with Billing Period Variation

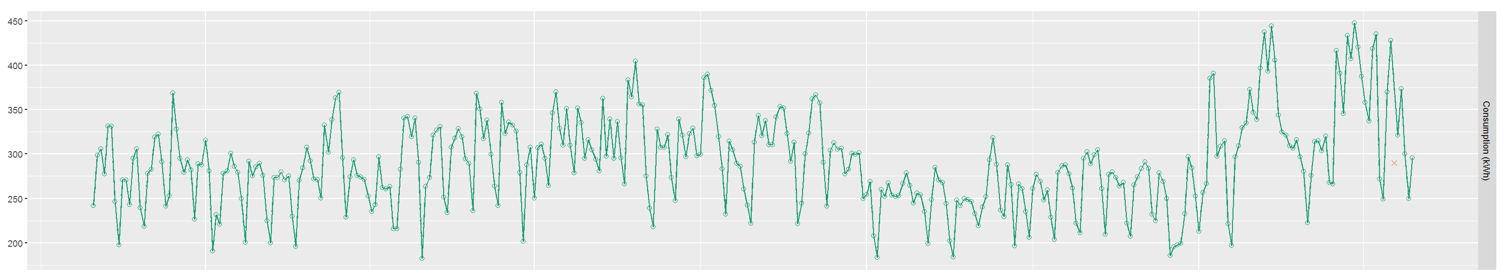


Figure 2 Energy Consumption Data with Daily Variation

1. **Mathematical Formulation**

Energy consumption of a building is dependent on many factors. These factors can be baseload, outside temperature, seasonal variation, wind speed, daily activity factor, amount of materials produced in case of a production plant, duration of terms and holidays in case of educational institution. Taking into account these factors, regression formula reflecting total energy consumption as a function of these factors can be formulated as:

(1)

Impact of temperature on energy consumption is measured by cooling degree days or heating degree days [1]. Degree Days are essentially simplified representation of outside temperature data. Heating Degree Days (HDD) are a measure of how much (in degrees), and for how long (in days), outside air temperature was lower than specific set points**.** These measurements are useful for calculations of energy consumption required to heat building. Similarly, Cooling Degree Days (CDD) are a measure of how much (in degrees), and for how long (in days), outside air temperature was higher than a specific base temperature. These measurements are used for calculations relating to energy consumption required to cool buildings. Humidity is measured as relative humidity and wind speed is measured as WindHDD. For

analysis purpose, we name dependent variables as “*Energy* *Drivers*”. Energy Drivers are driving factors that contribute towards total energy consumption. Figure 3 shows breakdown of these energy drivers for total energy consumption of electricity for a year for a building and Figure 4 shows output from linear regression model and actual consumption for the same site.

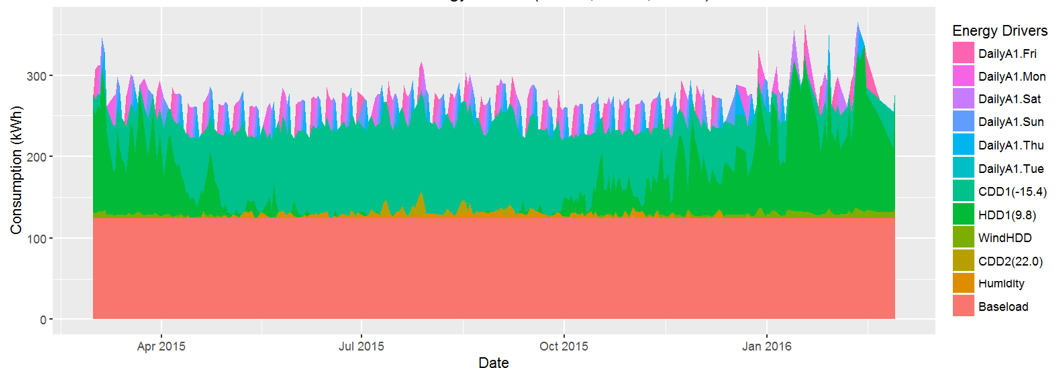


Figure 3 Energy Drivers Contribution Towards Total Energy Consumption

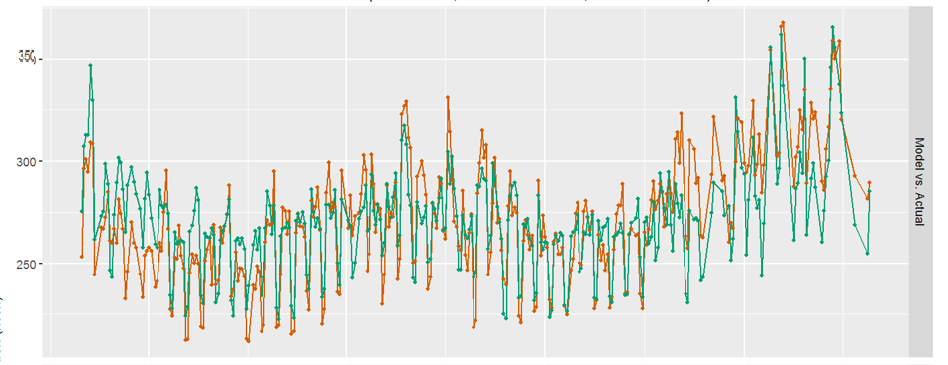
****

Figure 4 Actual Consumption Vs Regression Model Output

**2. Methodologies**

There are two types of exclusion methods: manual and auto. Manual method of exclusion offers user to exclude data that user visualizes to be an outlier, however if the exclusion method is set to “Auto Exclusion”, program calls different techniques and methods that are pre-defined in R such as tsoutliers, boxplot outlier detection methods. Iteration steps are set by the user, if model fitting values are not good enough (acceptable by ASHRAE standards) data will undergo optimization again, and go through exclusion process again. Exclusion process performs the checking for that the percentage of exclusion is not above acceptable 25%. There is another exclusion process which facilitates user to select the data point to be excluded that the user finds visually as an outlier. Therefore, there are three exclusion processes in the modeling process developed in R programming environment:

1. Manual Exclusion: In this exclusion option, user can select the point to be excluded from time series plot of consumption.
2. Auto Exclusion: In this exclusion option, in-built functions like boxplot, mvoutlier, k-means clustering function, residuals detect outliers and remove them.
3. No exclusion: Both auto and manual exclusion options are disabled, all data points present during modeling period are considered for modeling.

When no exclusion option is enabled, all data present during modeling phase are considered for modeling purpose. No data point is excluded for better modeling evaluation. In addition to those exclusion options, while observing the data pattern it is observed that, if consumption data are missing from for 3 to 4 days, the data that immediately appears, doesn’t capture the consumption for the whole day, in that case, that days is automatically excluded from modeling calculation. Packages used for exclusion, optimization, and estimation of coefficients for non negative least squares function are provided in reference lists.

1. Modeling Results with different exclusion options.
2. We implemented clustering by k-means on consumption data. We divided the data set (baseline period data) into 10 clusters by k-means function. K-means is an unsupervised method of clustering. We applied box plot in each clusters for detecting outliers. The improvement was not significant compared to single box plot method. In the next step, we implemented bivariate clustering method. Instead of applying consumption method only on consumption data, we applied clustering method on data matrix that consists of consumption and temperature. Improvements in outlier detection were not significant as compared to single box plot.
3. Tsoutliers: Package “tsoutliers” is available in R. The package implements procedure to detect any sudden change in pattern in time series data. We used tsoutliers function in conjunction with box plot and clustering method. Results obtained with the package are quite similar to that of clustering method.
4. Residuals: Outliers are detected from residual data after the model fit. The approach looked promising. We were getting maximum successful models so far with this approach; however, number of data points detected as “outliers” was very high.
5. Mvoutlier: Package “mvoutlier” in R detects outliers in a multivariable dataset. The dataset, we are working are mainly multivariable; consumption is dependent on temperature, humidity, wind, and schedules. We could get more success with this approach compared to clustering methods and tsoutliers. However,, we noticed, number of data points identified as outliers are significantly higher compared to previously mentioned methods.
6. Before-After missing: When there is a series of missing data, data collected on the day after such period doesn’t reflect the actual consumption. Therefore, we excluded the data reported on such days.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Exclusion Method | Avg. R2 | Avg. CVRMSE | Total Intervals Excluded | % Test Models for which method was successful \* |
| tsoutliers | **69.27%** | **13.94%** | **10** | **42.9%** |
| univariate-1cl-nbxp | **72.36%** | **12.67%** | **49** | **28.6%** |
| univariate-2cl-nbxp | **77.52%** | **10.60%** | **155** | **57.1%** |
| bivariate-2cl-nbxp | **72.90%** | **12.94%** | **71** | **42.9%** |
| bivariate-3cl-nbxp | **76.82%** | **11.58%** | **96** | **57.1%** |
| residuals-i1-nbxp | **83.01%** | **9.02%** | **120** | **71.4%** |
| tsoutliers + univariate-2cl-nbx | **77.52%** | **10.60%** | **155** | **57.1%** |
| tsoutliers + residuals-i1-nbxp | **79.34%** | **9.60%** | **96** | **71.4%** |
| bivariate-2cl-n15 | **73.56%** | **12.51%** | **105** | **57.1%** |
| bivariate-2cl-n25 | **74.19%** | **12.16%** | **175** | **57.1%** |
| manual | **80.02%** | **10.31%** | **73** | **71.4%** |
| manual-t2 | **74.51%** | **12.67%** | **37** | **57.1%** |
| bivariate-2cl-nbxp + residuals | **85.48%** | **7.69%** | **172** | **85.7%** |
| residuals2-i1-nbxp | **85.68%** | **8.10%** | **238** | **71.4%** |
| residuals2-ni-5bxp | **83.68%** | **8.68%** | **100** | **100.0%** |
| univariate-outlier-4n | **71.66%** | **13.11%** | **46** | **57.1%** |
| multivariate-3d-mvoutlier | **72.20%** | **11.85%** | **177** | **28.6%** |
| tsoutliers + ba-missing + R2-5b | **81.63%** | **9.04%** | **108** | **100.0%** |
| tsoutliers + ba-missing + R-5b | **81.63%** | **9.04%** | **108** | **100.0%** |
| before-after-missing | **64.13%** | **14.87%** | **18** | **0.0%** |
| getOutlierFunction | **79.30%** | **10.64%** | **44** | **71.4%** |

1. Collection of Data: There are different methods to collect data, data can be stored locally to user’s machine, or they can be accessed through the database.
2. Getting data ready for baseline period (baseline period is the period, during which modeling is performed)
3. Validate percentage of data present during modeling phase, if not enough data are present, there will be error message and modeling will not be performed.
4. In the next step the software checks for the presence of enough data during inspection phase. Inspection phase starts right after modeling phase ends. In this case, we are considering a year of data after the baseline ends as the inspection phase.
5. If there is no enough data in inspection phase, modeling can still be done provided there is more than 85% data available during modeling period: i.e. baseline period.
6. After we have acceptable or sufficient percentage of data available to perform modeling, next we check for weather data. Environment Canada or Weather Underground are source for getting weather data. Weather data is considered for modeling, for majority of clients for Energent. As weather is the main driving factor behind consumption of electricity and gas. Relationship between consumption and temperature is analyzed, for majority of the cases it is observed that with increase in temperature consumption of electricity increases for measuring consumption of electricity with increase in temperature Cooling Degree Day (CDD) driver is used, similarly for measuring increase in consumption in Natural gas with decrease in temperature during winter period, Heating Degree Days is used. First model fitting by earth function in R generates set point for these drivers. In addition to cooling and heating degree drivers. In some cases, more than one heating and cooling loads are calculated.
7. Number of iterations are specified by the user, for each iteration the parameters are optimized to get the best models statistics, if the model statistics are not good enough data undergo next iteration for parameter optimization, in this process data undergo outlier detection and removal process if not more than 25% data have been eliminated.
8. If the maximum percentage of data to be eliminated and the maximum number of iterations are achieved, data don’t undergo for elimination or optimization.
9. After completion of iterations and acceptable exclusions final models results are plotted and exported in excel output file.
10. Final model results are deployed on the website and results are communicated to clients.
11. **Test Results**

1. Analysis of improvements of results with exclusion and without exclusion
2. Improvement in model statistics with inclusion of auto exclusion techniques.
3. **Conclusion**

The work proposed an optimized regression models with advanced outlier detection method. Modeling framework used for this work is R. The method involves collecting data, analyzing them, and excluding outliers, and creating a regression model. Performance of the model during modeling phase is measured by goodness-of-fit values set by ASHRAE and IPMVP. If the modeling is not meeting the standards, modeling process is iterated and exclusions are performed again till we get the best fit values.

1. **References**
2. *www.degreedays.net*
3. “*LEAN Energy Analysis: Using Regression Analysis to Assess Building Energy Performance”,* byM. Donnelly, J. Kummer, K.Drees, Source: Johnson Controls
4. “*Multi-linear Regression Models to Predict the Annual Energy Consumption of an office building with different shapes*” by M. Mottahedi, A. Mohmmadpour, S. Amiri, D.Riley, and S. Asadi, *International Conference on Sustainable Design, Engineering and Construction.*
5. “*Analysis of a Residential Building Energy Consumption Demand Model*”, by W. Yu, B. Li, Y. Lei, and M. Liu, *Molecular Diversity Preservation International* *(MDPI), Energies* 2011, issue: 4, pp: 475-487, DOI: 10.3390/en4030475
6. “*Multiple regression models for energy use in air-conditioned office buildings in different climates*”, by J. C. Lam, K.W. Wan, D. Liu, C. Tsang, *Elsevier Journal on Energy Conversion and Management*, issue: 51, pp: 2692-2697, Year: 2010
7. “*A* *Decision Tree method for building energy demand modeling* ”by Z. Yu, F. Highlight, B. Fung, Hiroshi Yoshino, *Elsevier Journal on Energy and Buildings*, issue: 42, issue: 42, pp: 1637-1646, Year: 2010
8. “*Kernel Regression Real-Time Building Energy Analysis*” by M. Brown, C. Barrington-Leigh, and Z. Brown, *Journal of Building Performance Simulation*, available online, in press.
9. “*Performance Modeling of Building Energy Usage: M&V Just Got Real*” by E. Urbatsch, available online
10. “*Development and Validation of Regression Models to Predict Annual Energy Consumption of Office Buildings in Different Climate Regions in the United States of America*” by S. Amiri, M. Mottahedi, S. Asadi, D. Riley, *5th International Speciality Conference, Vancouver, British Columbia, June 2015.*
11. “*Understanding Industrial Energy Use through Sliding Regression Analysis*”, by C. W. Eger and K. Kissock, *ACEEE Summer Study on Energy Efficiency in Industry,* Year: 2007
12. “*Predicting Energy Usage U*sing *Historical Data and Linear Models*”, by M. Safa, J. Allen and M. Safa, *31st International Symposium on Automation and Robotics in Construction and Mining,* Year:2014
13. “*R Core Team (2015). R: A language and environment for statistical computing. R foundation for*

*Statistical computing, Vienna, Austria*”, URL https:// [www.R-project.org/](http://www.R-project.org/).

1. “*earth: Multivariate Adaptive Regression Splines*”, by S. Milborrow. Derived from mda: mars by T. Hastie and R. Tibshirani. Uses Alan Miller’s Fortran utilities with Thomas Lumley’s leaps wrapper.
2. *R6: Classes with reference semantics*” by W. Chang, *R package version 2.1.1.* *https:// CRAN.R-project.org/package=R6*
3. “*mvoutlier: Multivariate Outlier Detection based on robust methods*” by P. Filzmoser and M. Gschwandtner, <https://CRAN.R-project.org/package=mvoutlier>
4. “*ggplot2: Elegant Graphics for Data Analysis”.*  Springer-verlag New York, 2009
5. “*Dates and Times Made Easy with lubridate. Journal of Statistical Software*”, by G. Grolemund, H. Wickham, 40*(3), 1-25. URL http://www.jstatsoft.org /v40/i03/.*
6. “Modern Applied Statistics with S. Fourth Edition”, by W. Venables and B. Ripley, *Springer, New York, ISBN 0-387-95457-0*
7. “*xts: eXtensible Time Series*”, by A. R. Jeffrey and J. M. Ulrich, *R package version 0.9-7.* [*https://CRAN.R-peoject.org/package=xts*](https://CRAN.R-peoject.org/package=xts)
8. “*xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files*” by A. Dragulescu, *R package Version 0.5.7.* [*https://CRAN.R-project.org/package=xlsx*](https://CRAN.R-project.org/package=xlsx)
9. “*gtable: Arrange grobs in tables*..” by H. Wickham, R package version 0.1.2. <https://CRAN.R-Project.org/package=gtable>
10. “*American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. (2007). ASHRAE standard 90.1, Atlanta, GA*
11. “*International Performance Measurement and Verification Protocol*”, National Renewable Energy Laboratory, [*www.nrel.gov/*](http://www.nrel.gov/)