# Energy Generation and Consumption Data Analysis

# Contents

# Introduction

This analysis of energy data allows for a breakdown of energy generation and consumption. This kind of technical analysis allows energy managers to break down the areas in which they should focus their attention and dive deeper into understanding why there is such a high energy consumption. This analysis is important for generation discussions on how to incorporate changes to your business for reducing costs and managing energy consumption levels.

Renewable energy generation from technologies that are commercially available and developed rapidly today. We analyze the trend and the optimized generation amount to maximize profit in this project.

# Background

The process of deploying predictive analytics in energy generation and consumption is anticipated to detect the energy trend and wisely produced with the hopes of maximizing profit. The merits of this project are both qualitative and quantitative. Which energy should be produced than others, and what is the reasonable amount of energy that should be produced in each month?

# Problem Presentation

In this project, we determine the trend of energy consumption and the average amount of energy used in each month. Therefore, energy companies will be able to build and produce more effectively.

We also identify what the most prospected energy is based on the proportion of energy generation. Moreover, because renewable energy is more likely to be used recently, we can optimize renewable energy to be suitable for the market trend in this project.

# Specification and Design

The project determines the trend of energy consumption by year from 2001 to 2008 and the average amount of energy used in each month. Due to the time-based nature of the data, a time series approach will focus primarily on the monthly amount.

The necessary information for this project must include at least the following: The energy types, energy generation, energy consumption. Additionally, the data must include all the above information in 2001 to 2008 in order to determine any trends in energy.

A naïve model will serve as the baseline with which to compare subsequent models. The results of this project will fall into three groups:

- Descriptive – Summary of the overall data characteristics and trends
- Predictive – Trend of Energy Amount
- Prescriptive – Optimization renewable energy amount

# Data Pre-Processing

Based on the dataset, the preparation is as follows:

1. Total Electric Power Industry was extracted from Type of Producer
2. US total consumption was extracted from State
3. Sum of all four-consumption energy source in one month

# Data Acquisition

The time frame for the project focuses on the energy consumption and generation from 2001 to 2008. The data set also contained the type of energy with the monthly figure.

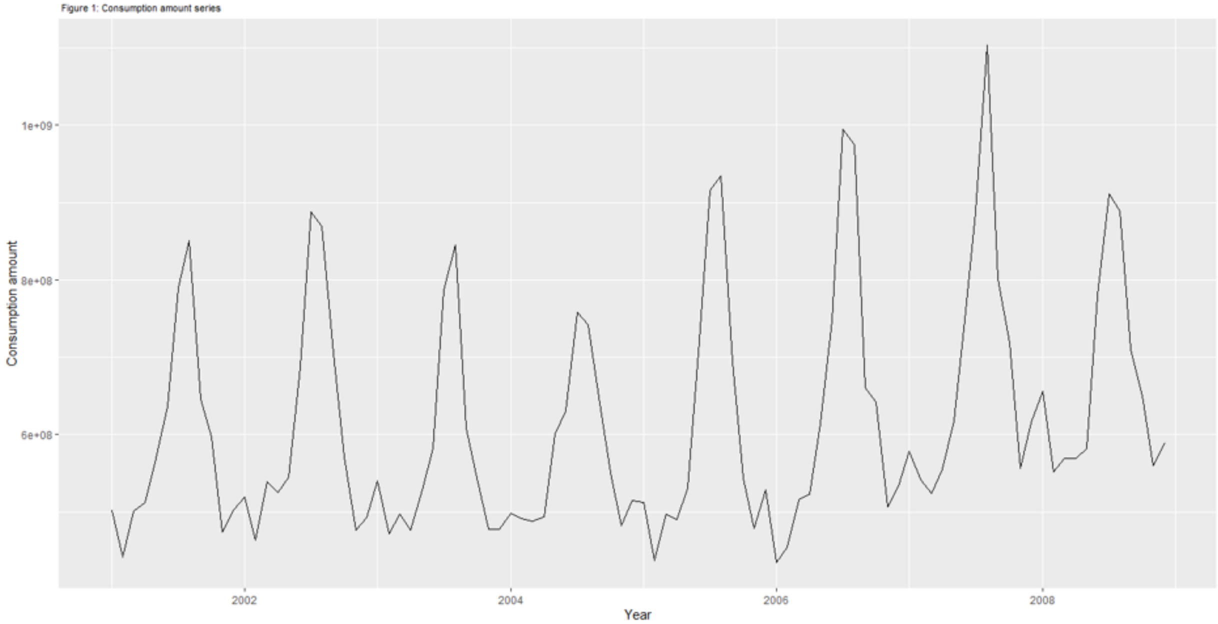
For this project, the raw data that will be used is provided from OpenEI. OpenEI is growing into a global leader in the energy data realm - specifically analyses on renewable energy and energy efficiency.

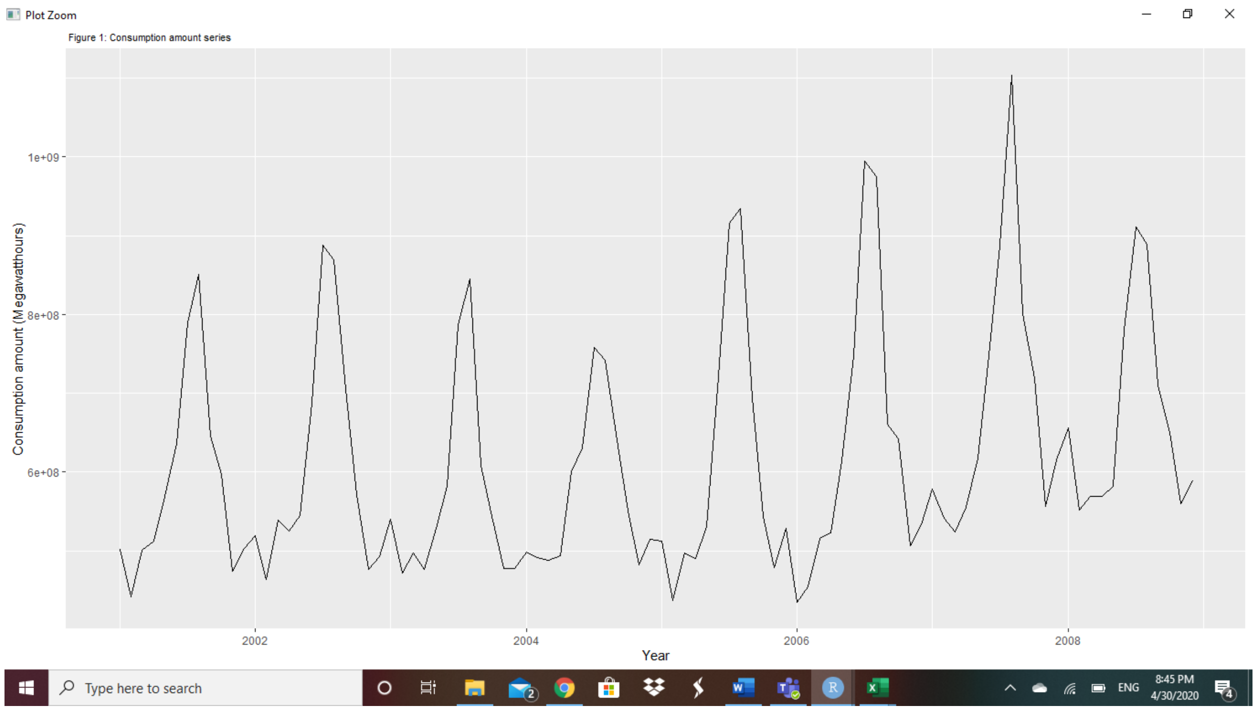
The consolidation of information that is utilized for this project is accessed and made available through https://openei.org/datasets/dataset/electric-power-monthly-monthly-data-tables

# Descriptive Results

ENERGY CONSUMPTION TREND

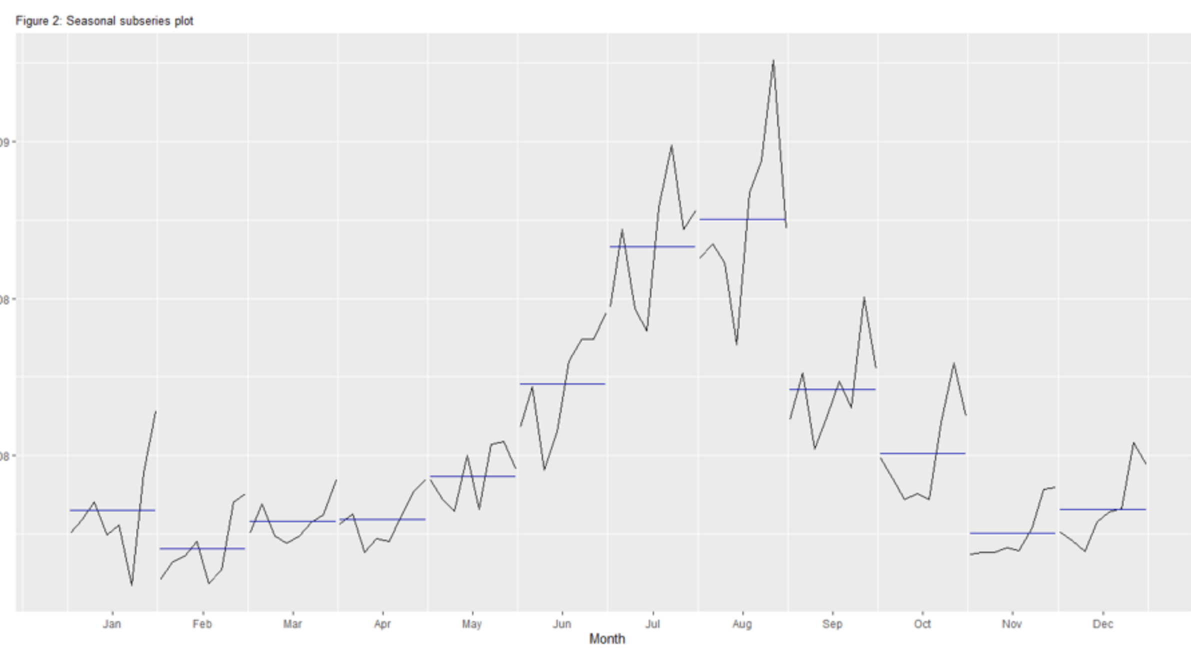
Figure 1: Consumption amount series

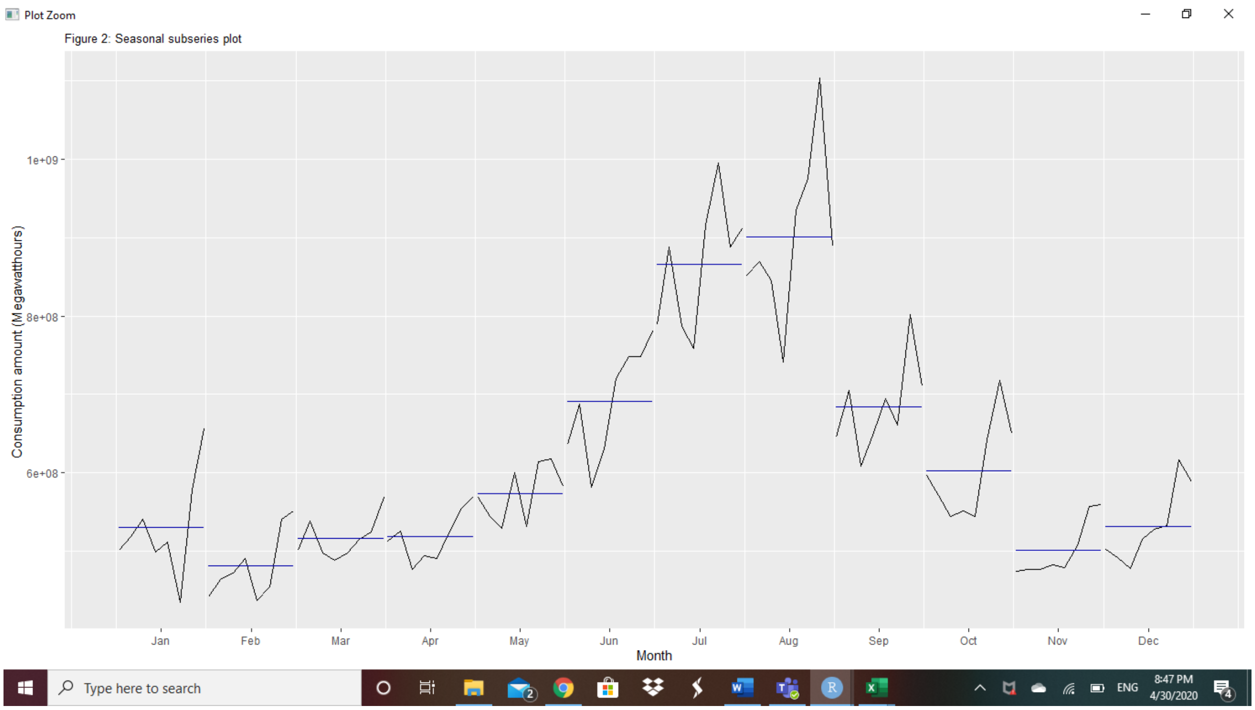





According to the above graph, there is seasonal growth by year with an increasing trend of energy consumption. There is a decrease in the middle of 2004 to the beginning of 2005. However, the consumption amount significantly increases after that.

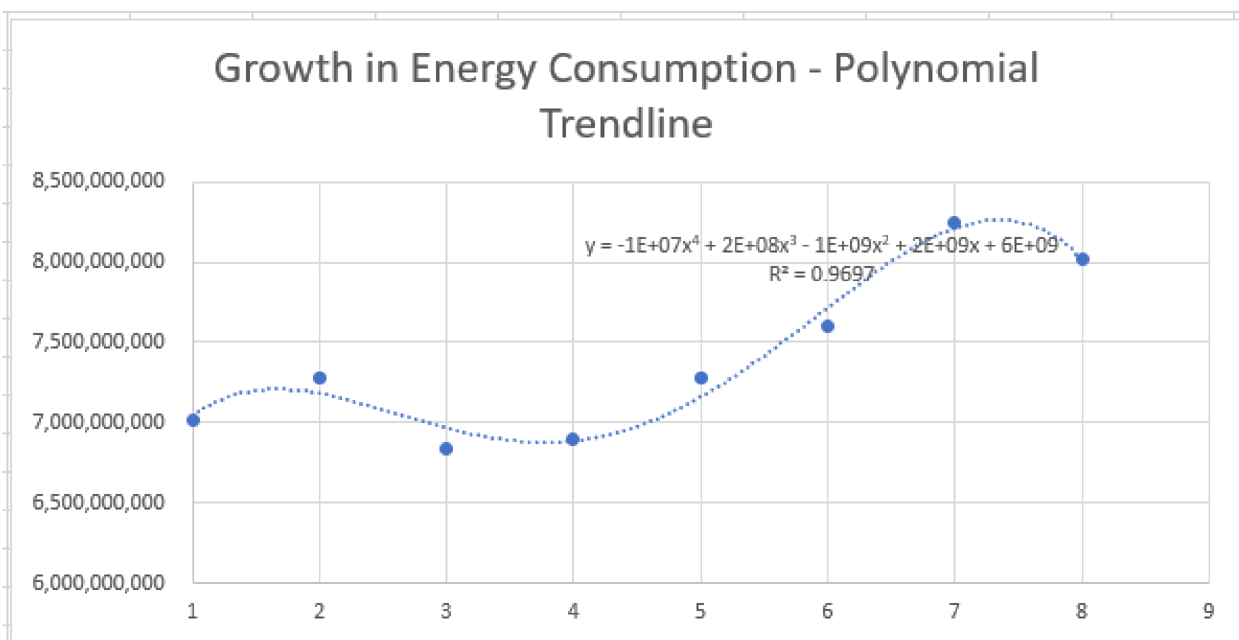
MONTHLY ENERGY CONSUMPTION DISTRIBUTION





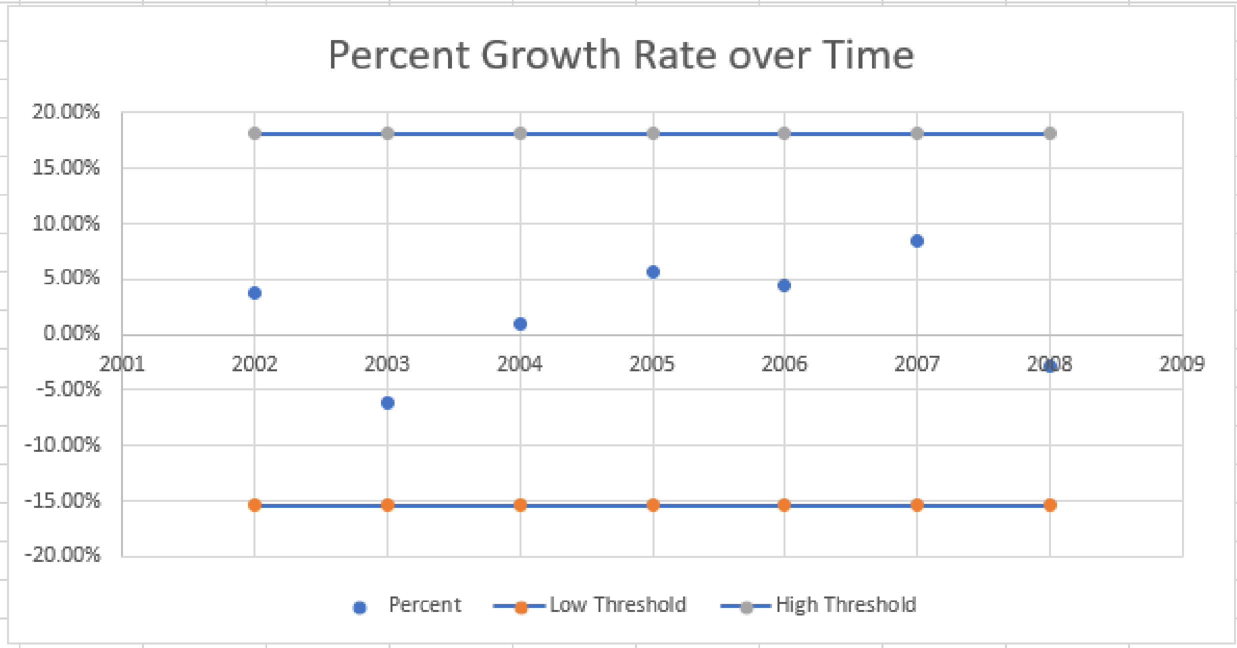

In the specific aspect, energy is used most in July and August. Hot weather can explain this sudden increase. In general, people tend not to use much energy in January and February.

GROWTH IN ENERGY CONSUMPTION




The polynomial trendline is the best formula for explaining the growth in energy consumption from 2001 to 2008.

PERCENT GROWTH RATE OVER TIME




There is no outlier in percent growth over time from 2002 to 2008 of energy consumption based on the above graph. The percentages range from -6 % to around 8%.

PROPORTION OF ENERGY SOURCES GENERATION

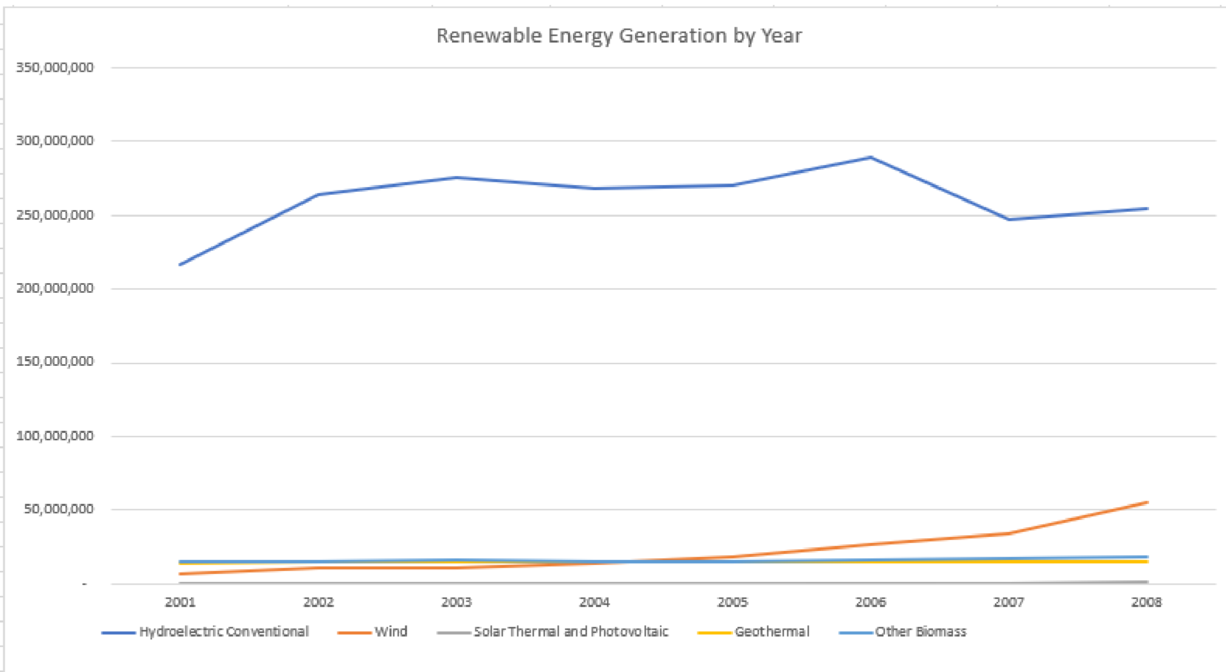
| ENERGY SOURCE | PERCENT |
|---|---|
| Coal | 49.60% |
| Petroleum | 2.38% |
| Natural Gas | 18.99% |
| Other Gases | 0.33% |
| Nuclear | 19.73% |
| Hydroelectric Conventional | 6.55% |
| Wind | 0.55% |
| Solar Thermal and Photovoltaic | 0.01% |
| Wood and Wood Derived Fuels | 0.95% |
| Geothermal | 0.37% |
| Other Biomass | 0.40% |
| Pumped Storage | -0.19% |
| Other | 0.32% |
| Total | 100% |

According to the average amount of energy generation from 2001 to 2008, coal, nuclear, and natural gas are the three greatest sources of energy generation. A fair proportion of renewable energy is produced.

PROPORTION OF RENEWABLE ENERGY SOURCES GENERATION

| RENEWABLE ENERGY SOURCE | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 |
|---|---|---|---|---|---|---|---|---|
| Hydroelectric Conventional | 216,961,045 | 264,328,831 | 275,806,328 | 268,417,307 | 270,321,256 | 289,246,416 | 247,509,975 | 254,831,383 |
| Wind | 6,737,331 | 10,354,278 | 11,187,465 | 14,143,741 | 17,810,551 | 26,589,138 | 34,449,927 | 55,363,101 |
| Solar Thermal and Photovoltaic | 542,755 | 554,831 | 534,001 | 575,157 | 550,294 | 507,704 | 611,793 | 864,315 |
| Geothermal | 13,740,501 | 14,491,309 | 14,424,231 | 14,810,974 | 14,691,743 | 14,568,028 | 14,637,212 | 14,951,347 |
| Other Biomass | 14,548,151 | 15,043,714 | 15,811,991 | 15,420,568 | 15,420,393 | 16,098,523 | 16,524,554 | 17,733,759 |
| TOTAL | 252,529,783 | 304,772,963 | 317,764,016 | 313,367,747 | 318,794,237 | 347,009,809 | 313,733,461 | 343,743,905 |

| RENEWABLE ENERGY SOURCE | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 |
|---|---|---|---|---|---|---|---|---|
| Hydroelectric Conventional | 85.92% | 86.73% | 86.80% | 85.66% | 84.79% | 83.35% | 78.89% | 74.13% |
| Wind | 2.67% | 3.40% | 3.52% | 4.51% | 5.59% | 7.66% | 10.98% | 16.11% |
| Solar Thermal and Photovoltaic | 0.21% | 0.18% | 0.17% | 0.18% | 0.17% | 0.15% | 0.20% | 0.25% |
| Geothermal | 5.44% | 4.75% | 4.54% | 4.73% | 4.61% | 4.20% | 4.67% | 4.35% |
| Other Biomass | 5.76% | 4.94% | 4.98% | 4.92% | 4.84% | 4.64% | 5.27% | 5.16% |




As you can see, there is a stable increase in the amount of renewable energy generation except in 2007. Wind energy sources have seen speedy growth from only 2.67% in 2001 up to 16.11% in 2008.

## Data Transformation

For this project, the data does not need to be transformed. The dataset is consistent from 2001 to 2008 so transformation is not necessary.

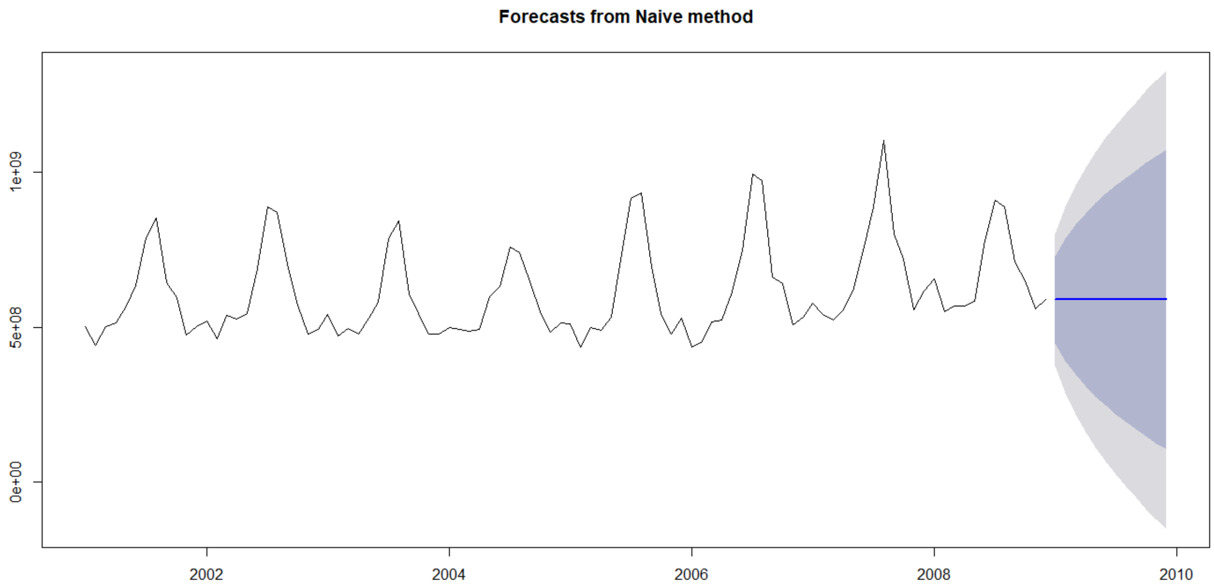
The total amount and year are two usable variables, and there is no multicollinearity between them.

# Predictive Results

Based on the time series consumption, exponential smoothing is a popular forecasting method for short-term predictions. Such forecasts of future values are based on past data whereby the most recent observations are weighted more than less recent observations. Thus, Naïve method, Holt-Winters exponential smoothing, and automated exponential smoothing forecasts are applied for a forecast.

## Naïve method

A Naïve Prediction serves as a baseline for subsequent model comparison. To create this, a rolling average of the previous month's actual energy consumption was used to generate the next month's predictions. The prediction was then fitted to the actual data on the time-series chart showing actual and predicted monthly consumption.

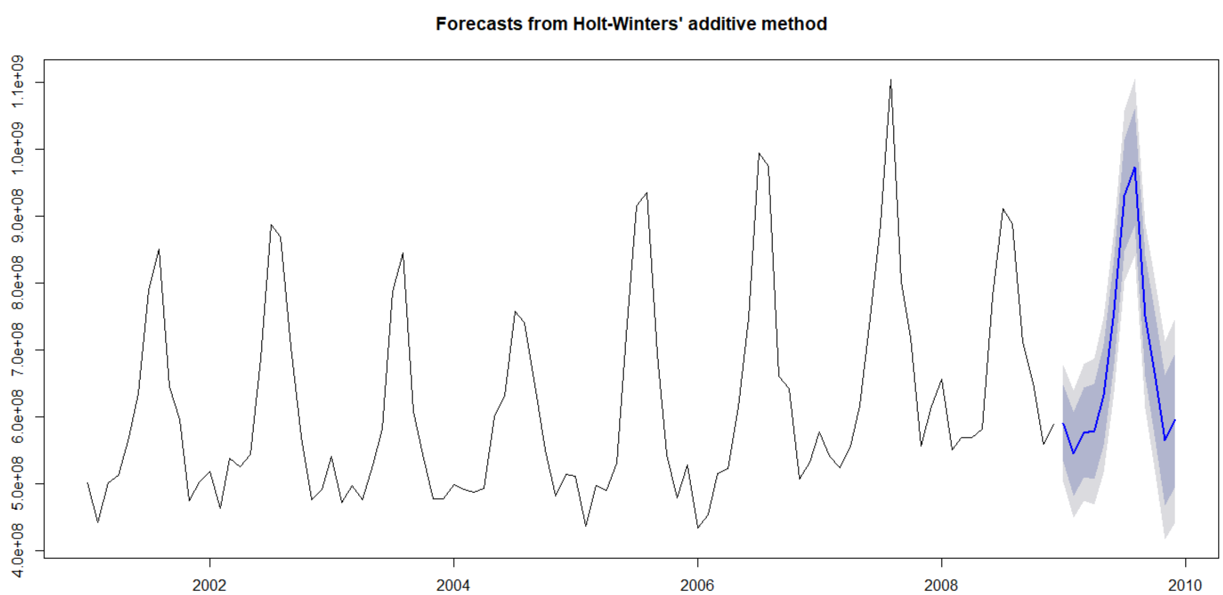


```
Forecast method: Naive method

Model Information:
Call: naive(y = series, h = 12)

Residual sd: 109569621.0189

Error measures:
                   ME      RMSE      MAE       MPE     MAPE     MASE      ACF1
Training set 915068.6 108995254 81758963 -1.062862 12.68544 1.708118 0.2789326

Forecasts:
         Point Forecast     Lo 80      Hi 80      Lo 95      Hi 95
Jan 2009      588381413 448698374  728064452  374754640  802008186
Feb 2009      588381413 390839765  785923061  286267533  890495293
Mar 2009      588381413 346443293  830319534  218368988  958393838
Apr 2009      588381413 309015335  867747491  161127867 1015634960
May 2009      588381413 276040643  900722184  110697426 1066065400
Jun 2009      588381413 246229242  930533584   65104823 1111658003
Jul 2009      588381413 218814830  957947997   23178098 1153584729
Aug 2009      588381413 193298117  983464710  -15846347 1192609173
Sep 2009      588381413 169332296 1007430530  -52498907 1229261733
Oct 2009      588381413 146664859 1030097967  -87165759 1263928586
Nov 2009      588381413 125105183 1051657643 -120138439 1296901265
Dec 2009      588381413 104505172 1072257654 -151643437 1328406263
```

The Naïve method shows the baseline for comparison with Holt-Winter and auto exponential smoothing forecast method.

Lo 80 and Hi 80 are the boundaries of an 80% confidence interval and accordingly, the wider bounds of Lo 95 and Hi 95 represent the 95% confidence interval.

## Holt-Winters exponential smoothing

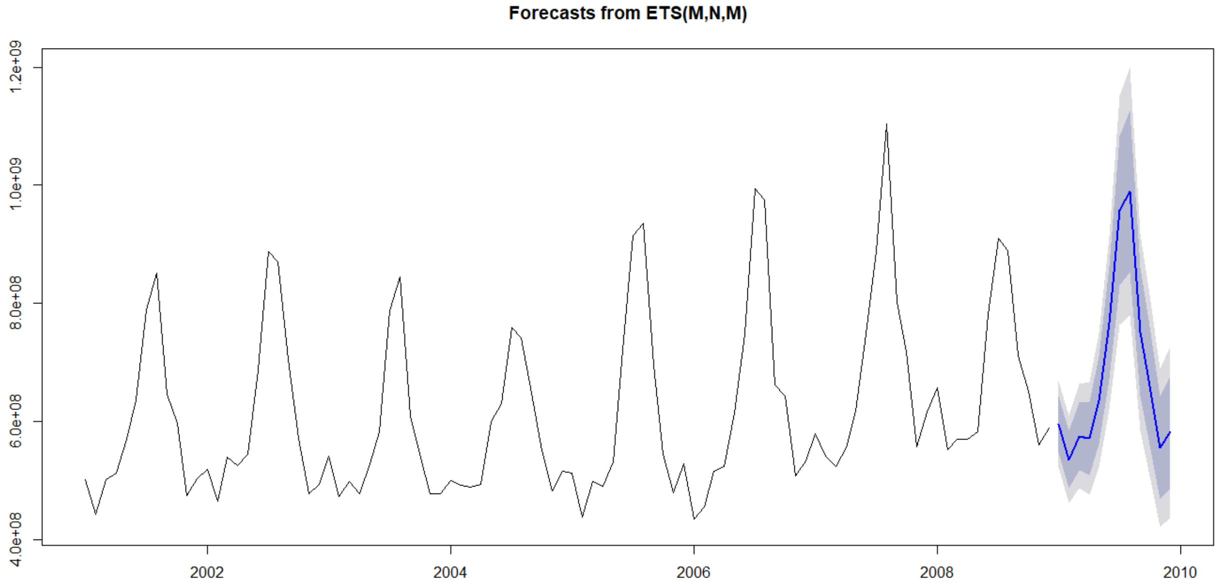


The plot represents the forecasted level for future time periods and visualizes the 80% and 95% confidence intervals. The higher the confidence, the wider the bounds.

```
Call:
 hw(y = series, h = 12)

  Smoothing parameters:
    alpha = 0.4295
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 581088847.6988
    b = 1193358.2298
    s = -91321068 -120673786 -18325325 67615117 291127309 249818842
           78904414 -45346600 -98992376 -99426257 -130179916 -83200353

  sigma:  44696124

     AIC     AICc      BIC
3836.831 3844.677 3880.425

Error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE      ACF1
Training set -544409.4 40801792 29938516 -0.2663198 4.797827 0.625479 0.1065236

Forecasts:
         Point Forecast     Lo 80      Hi 80     Lo 95      Hi 95
Jan 2009      590713375 533432987  647993762 503110582  678316167
Feb 2009      544919404 482578096  607260711 449576602  640262206
Mar 2009      576864213 509840960  643887465 474360993  679367432
Apr 2009      578485885 507085071  649886698 469287764  687684005
May 2009      633319577 557792641  708846512 517811097  748828056
Jun 2009      758752575 679311768  838193383 637258345  880246806
Jul 2009      930859564 847687165 1014031962 803658357 1058060771
Aug 2009      973350405 886605167 1060095642 840685012 1106015797
Sep 2009      751033554 660855359  841211748 613117906  888949201
Oct 2009      666283157 572796474  759769839 523307614  809258699
Nov 2009      565122288 468438808  661805767 417257666  712986909
Dec 2009      595661964 495882642  695441286 443062660  748261269
```

Holt-Winters exponential smoothing is a time series forecasting approach that takes the overall level, trend and seasonality of the underlying dataset into account for its forecast.

## Automated exponential smoothing forecasts




Automated exponential smoothing forecast method is automatically determined whether a time-series component is additive or multiplicative.

There are three letters in the model. The first letter denotes the error type, the second letter denotes the trend type, and the third letter denotes the seasonality type. Our dataset model is "MNM" which fits a model with multiplicative error, without trend and multiplicative seasonality.

```
ETS(M,N,M)

Call:
 ets(y = object, lambda = lambda, biasadj = biasadj, allow.multiplicative.trend = allow.multiplicative.trend)

  Smoothing parameters:
    alpha = 0.5303
    gamma = 3e-04

  Initial states:
    l = 602355344.0399
    s = 0.854 0.8155 0.97 1.1029 1.453 1.406
           1.1252 0.9329 0.8381 0.8435 0.7859 0.873

  sigma:  0.0628

     AIC     AICc      BIC
3802.482 3808.482 3840.947

Error measures:
                  ME     RMSE      MAE         MPE     MAPE      MASE         ACF1
Training set 1647824 38956185 29821303 -0.01000574 4.722918 0.6230302 0.0009716516

Forecasts:
         Point Forecast     Lo 80      Hi 80     Lo 95      Hi 95
Jan 2009      594568893 546723342  642414444 521395437  667742348
Feb 2009      535264712 486488234  584041191 460667525  609861899
Mar 2009      574517646 516681363  632353928 486064685  662970607
Apr 2009      570826313 508382216  633270409 475326308  666326317
May 2009      635439715 560789344  710090086 521271825  749607605
Jun 2009      766413298 670572955  862253641 619838142  912988454
Jul 2009      957641234 831043642 1084238827 764026925 1151255544
Aug 2009      989679266 852129247 1127229285 779314666 1200043867
Sep 2009      751217532 641943750  860491314 584097704  918337360
Oct 2009      660682867 560476870  760888863 507431019  813934714
Nov 2009      555402725 467848688  642956761 421500380  689305069
Dec 2009      581705119 486654978  676755260 436338473  727071765
```

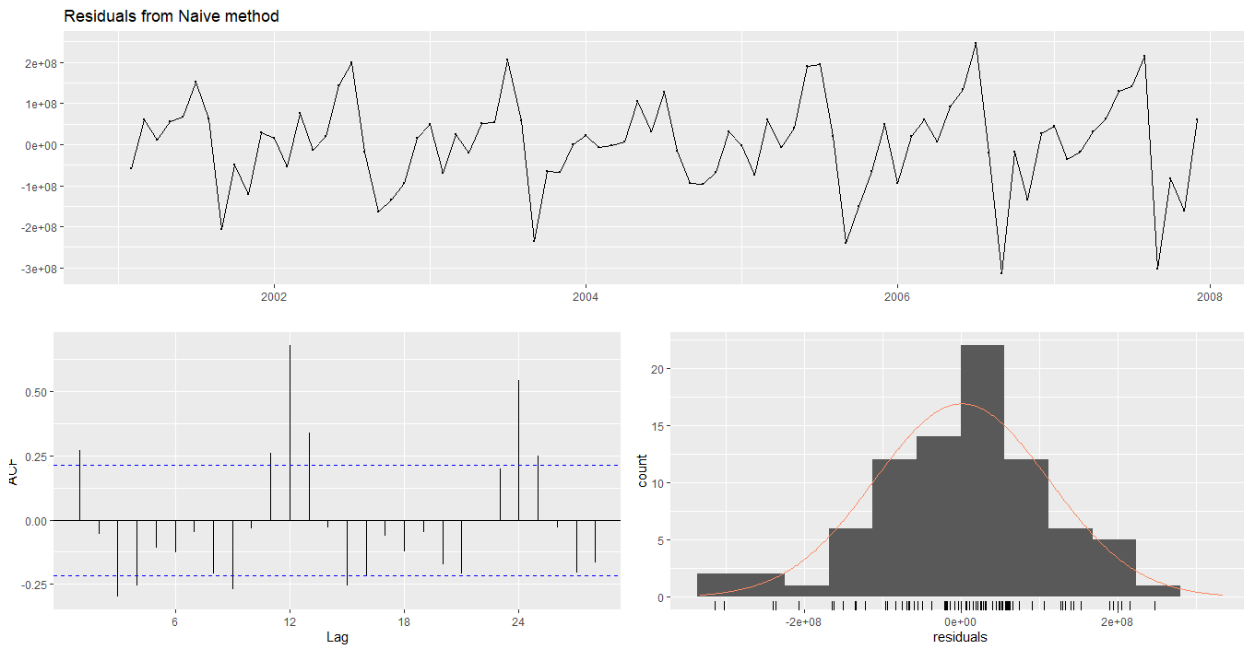
## Accuracy

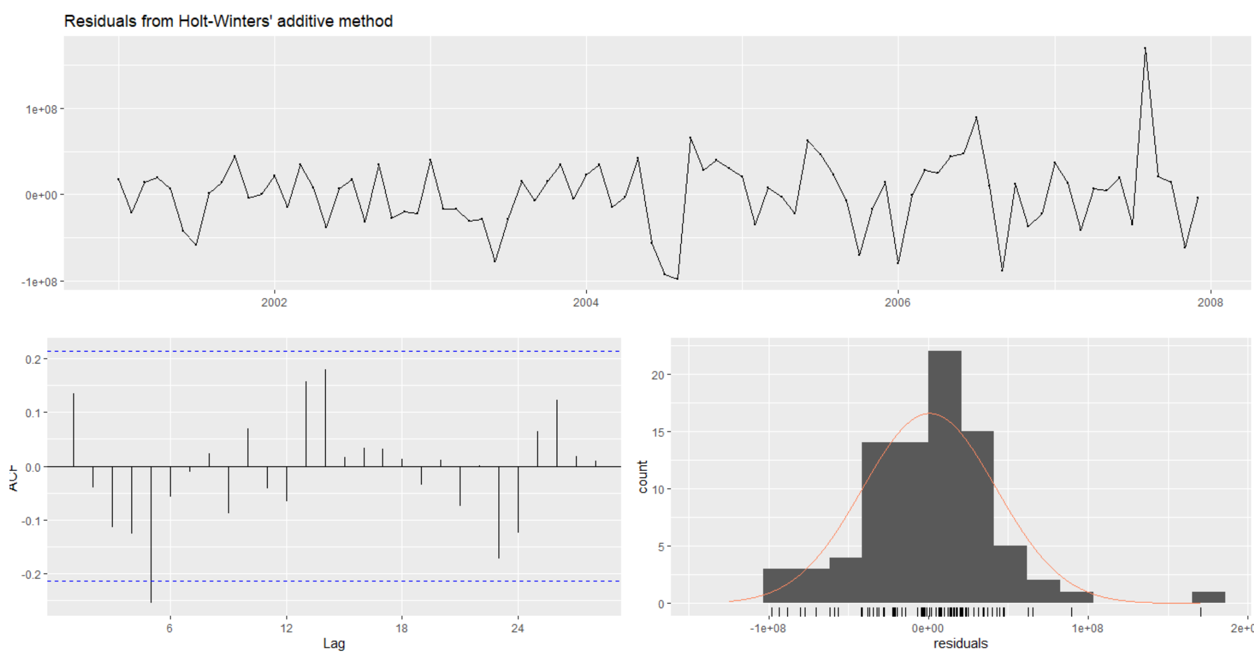
We fit a model to a subset of the data (training dataset), create a forecast for the remaining period (test dataset) and calculate the forecasting error by comparing the original observations with the corresponding fitted values.

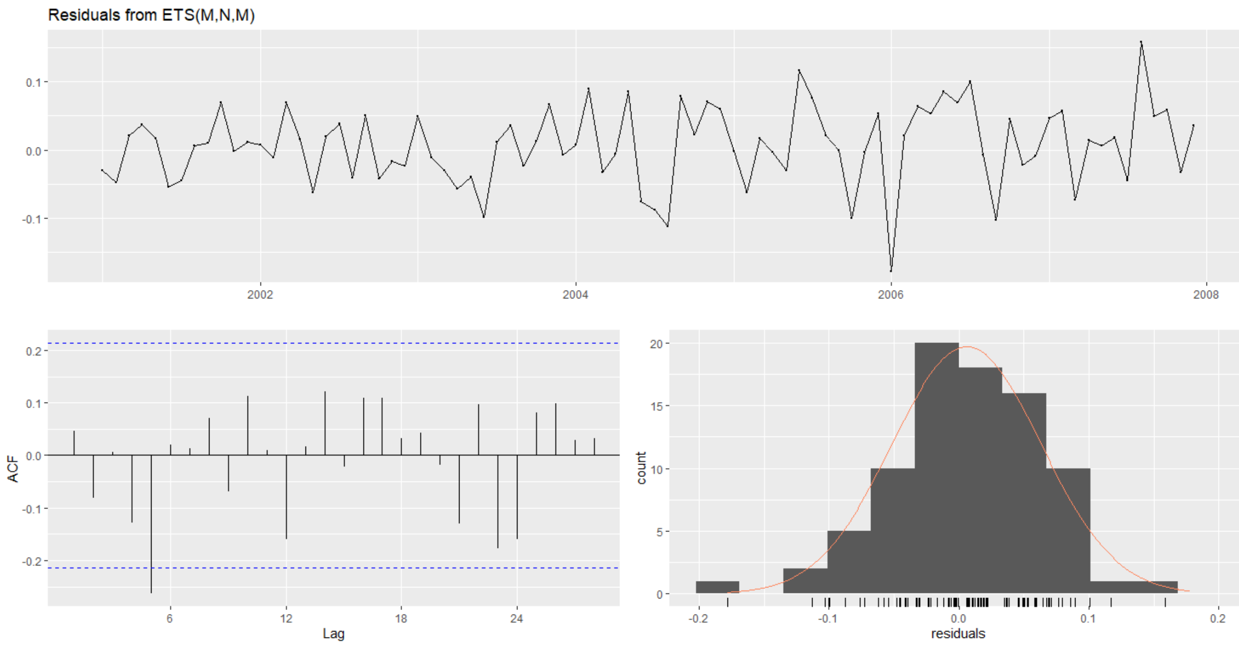
```
> accuracy(forecast(fit_naive_train), test)
                   ME      RMSE       MAE       MPE     MAPE     MASE      ACF1 Theil's U
Training set  1381480 110548894  82921342 -1.024235 12.97524 1.767128 0.2715915        NA
Test set     70991536 144836796 109335551  7.499761 14.28126 2.330039 0.6680329  1.270309
> accuracy(forecast(fit_hw_train), test)
                     ME     RMSE      MAE        MPE     MAPE      MASE       ACF1 Theil's U
Training set   704703.7 41464009 31067927 -0.1883478 5.027158 0.6620854 0.13414998        NA
Test set     -49467219.6 62590220 56679172 -7.4216629 8.520034 1.2078840 0.09282546 0.6235621
> accuracy(forecast(fit_auto_train), test)
                    ME     RMSE      MAE        MPE     MAPE     MASE       ACF1 Theil's U
Training set   3656395 38558642 28309133  0.2672664 4.492770 0.603293 0.07922561        NA
Test set     -52533164 74475164 58271383 -7.3101016 8.184025 1.241815 0.40049722 0.6822535
```

The lower RMSE and MAPE indicate a more accurate forecast. At this point, based on the MAPE value, the automated exponential smoothing model is more accurate.

This can be addressed with time series cross-validation whereby various training datasets are tested against various test datasets. In this analysis, we proceed with our chosen training and test dataset and verify Naïve method, Holt-Winters and automated exponential smoothing forecasting model further by analyzing the residuals in order to select the more suitable model.


Residuals from Naive method
2e+08
1e+08
0e+00
-1e+08
-2e+08
-3e+08
2002
2004
2006
2008
ACF
0.50
0.25
0.00
-0.25
6
12
18
24
Lag
count
20
15
10
5
0
-2e+08
0e+00
2e+08
residuals


Residuals from Holt-Winters' additive method
1e+08
0e+00
-1e+08
2002
2004
2006
2008
ACF
0.2
0.1
0.0
-0.1
-0.2
6
12
18
24
Lag
count
20
15
10
5
0
-1e+08
0e+00
1e+08
residuals

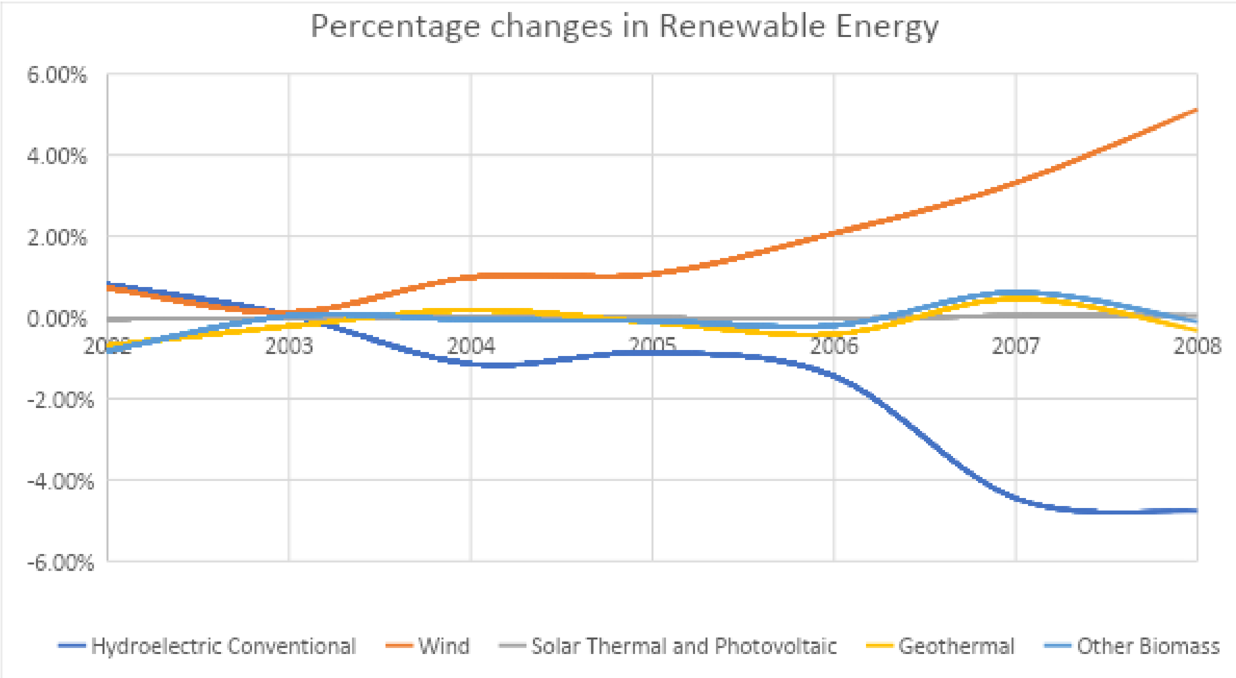



The residuals of the Holt-Winter model and automated exponential smoothing forecasting model are both more autocorrelated than those of the Naïve model. In fact, the autocorrelation values at many lags for Naïve model are way larger than those of the Holt-Winter model and automated exponential smoothing forecasting model. For instance, lag 1,12,13, etc., show respective peaks that go beyond the 95% limits defined by the dash blue lines. That means there is information in the data that the model is not using efficiently. (The Naïve model is not ideal)

# Prescriptive Results

Recent trends show that consumers use more renewable energy. Our project is to optimize the total profits of the renewable energy production mix.

OVERALL RENEWABLE TREND




According to the graph above, the trend of using wind energy increases steadily while hydroelectric tends to decrease from 2006. Wind energy is also widely used. However, consumers use less hydroelectric power now compared to the past. Thus, these trends are reasonable.

1. Decision variables

The decision variables are the generation of renewable energy amounts of Hydroelectric Conventional, Wind, Solar Thermal and Photovoltaic, Geothermal, and Other Biomass.

Generation Renewable Energy Amounts are real values.

2. Constraints

Generation Renewable Energy Amounts must be greater than or equal to the consumption and smaller than or equal to the generation.

Generation Wind Energy is always greater than or equal to Generation Hydroelectric Conventional Energy.

We have the total consumption and generation with the renewable energy percentage as below:

| | |
|---|---|
| Consumption | 30,175,235,820 |
| Generation | 30,815,311,919 |

| Renewable energy | Percentage Generation | Percentage Consumption |
|---|---|---|

| | | |
|---|---|---|
| Hydroelectric Conventional | 7.00% | 6.98% |
| Wind | 7.00% | 6.58% |
| Solar Thermal and Photovoltaic | 2.00% | 1.59% |
| Geothermal | 0.50% | 0.40% |
| Other Biomass | 1.47% | 1.50% |
| Total | 17.97% | 17.05% |

Moreover, the total average generation cost for five kinds of renewable energy must be greater than or equal to the minimum cost.

3. Objective variable

The prescription analysis is to optimize the generation of renewable energy amount while maximizing the profit (total revenue minus total average cost).

4. Sensitivity analysis

The Generation of Wind Energy is always greater than or equal to the Generation of Hydroelectric Conventional Energy today, so we apply sensitivity analysis on the amount of wind energy. We assumed that wind energy's average cost is smaller than or equal to hydroelectric conventional energy's. The production is still the same when applying this assumption because hydroelectric conventional energy must be maintained at a certain rate for not wasting minimum spending.

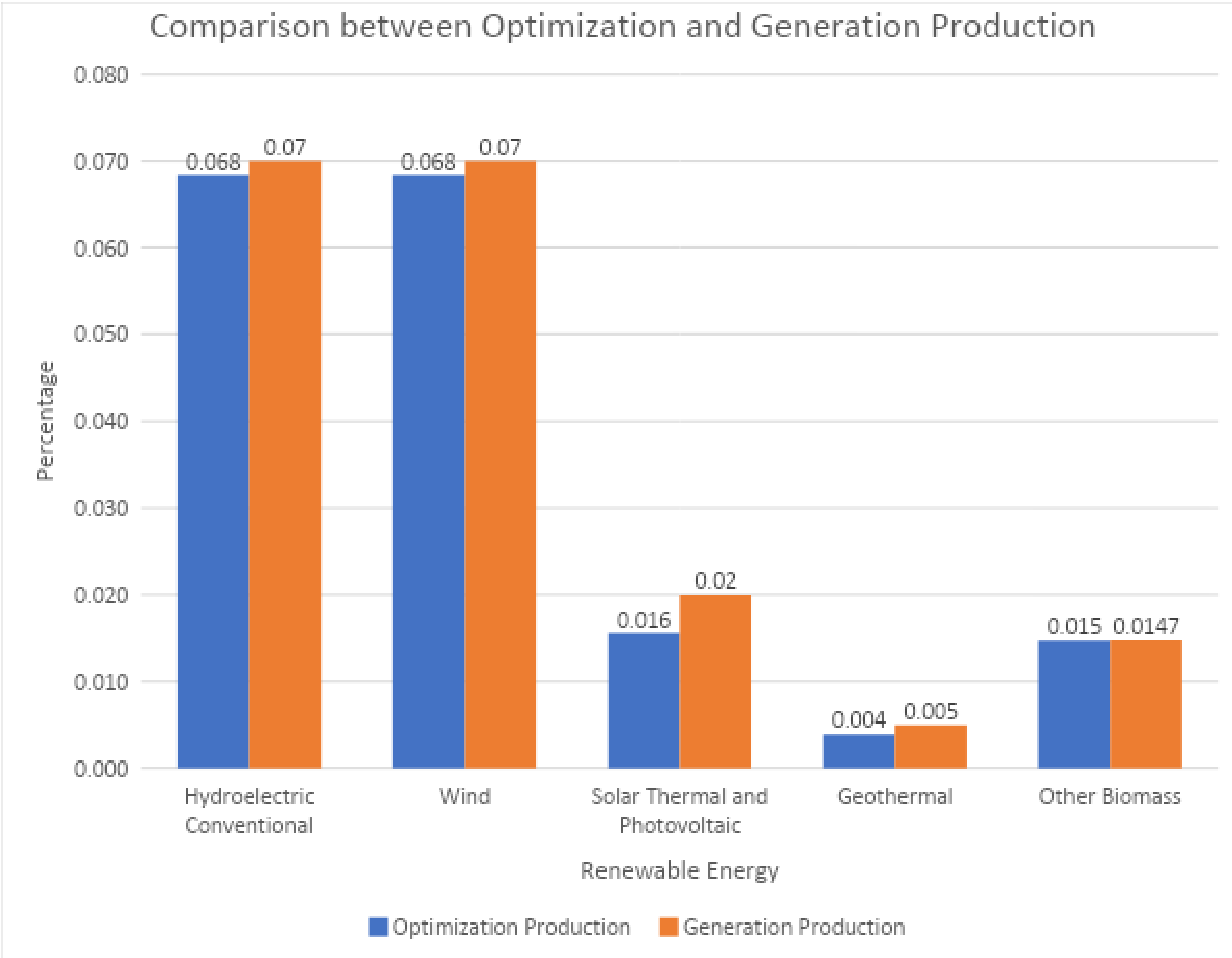
When Generation Wind Energy cost is lower than minimum cost. The company will suffer a loss.

5. Listing of method or methods tried

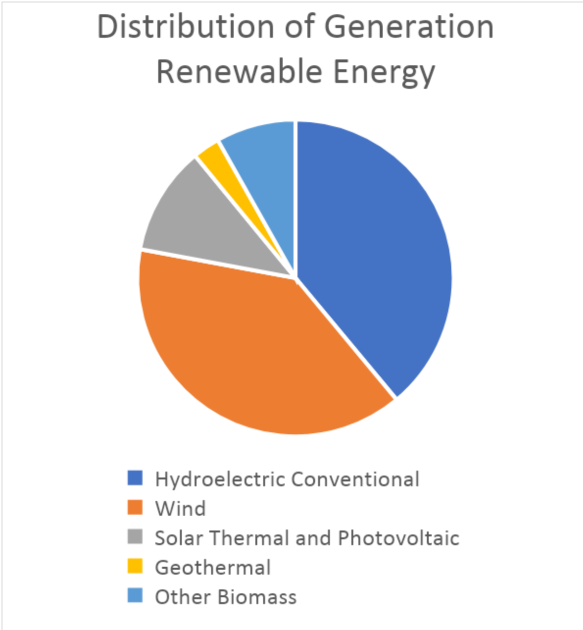
We used SimpleLP on SolverTable in Excel to optimize the project objective.

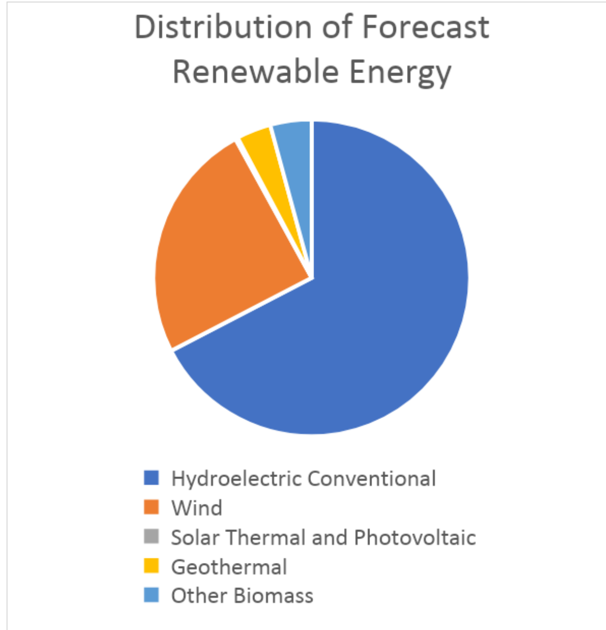
6. Visualization of results

Most of the generation amounts are lower than the actual generation. However, the other biomass energy price is lower than the average, the actual generation of this energy is reasonable.




Our project analyzed and developed a predictive model using the Automated exponential smoothing algorithm. We used the number of energy generation which was recorded based on the month.




We found that our model predicted the distributions of Geothermal and Other Biomass are rather good. However, the dataset is ten years old, so the distributions of Hydroelectric Conventional, Wind, and Solar Thermal and Photovoltaic are not compatible with the current record.

We were able to identify the trend of the generation of renewable energy. We recommend constructing an intervention program to test our model with updated data with the purpose to create harmony between consumption and generation.

# Conclusion

Naïve, Holt-Winter and automated models are compared by their MAPE for their forecast against the test dataset. Further validation is provided by analyzing the residuals and their distribution for three models with the automated exponential smoothing model qualifying as the more suitable forecasting model.

We should consider new technology – one of the most important factors when generating renewable energy. Besides, making procedural changes compatible with minimum cost (fixed cost, variable cost, etc.) will optimize profit. We need follow-up research to measure the improvements.

Our study is limited by the age of data available. There are many changes in technology and consumer demand, but our records are over ten years old which was recorded from 2001 to 2008.