vanhout adviseurs en installateurs

Measurement and validation of energy savings.

Quality of underlaying data: real-time or monthly?

29-05-2019

It's **NOT** possible to measure energy savings
It **IS** possible to quantify energys savings

Jan van Hout

Sander ter Mors

Exerpt from originally content published on the kW Engineering blog.

RETScreen Expert

- Mission: Empower cleaner energy decisions worldwide

World's leading clean energy decision-support tool

- Energy efficiency; renewable energy; cogeneration
- 36 languages covering 2/3rds world's population
- 600,000+ users in all countries, growing at 40,000+ new users each year
- 1,200+ universities & colleges also use for teaching & research
- Well over $8 billion in direct user savings since 1998

- Next generation of software, RETScreen Expert, released in September 2016

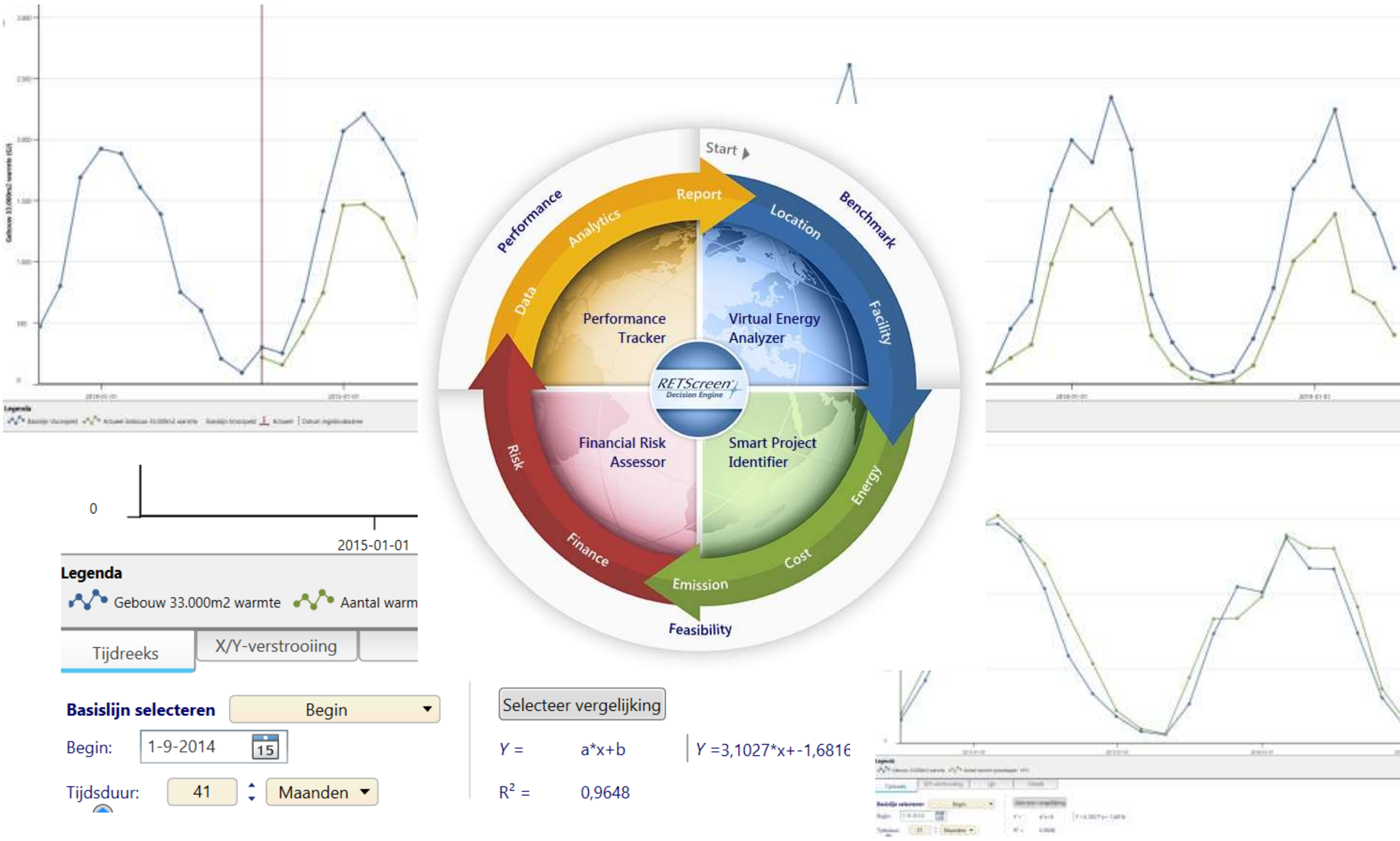
ieso
REEEP
NASA
UNEP
gef

Natural Resources Canada
Ressources naturelles Canada

It’s NOT possible to measure energy savings

It is possible to quantify energys savings

Regression analysis to validate and quantify the results of energy saving modifications to building (heating) systems.


Start
Report
Analytics
Performance
Data
Location
Benchmark
Facility
Performance Tracker
Virtual Energy Analyzer
RETScreen Decision Engine
Financial Risk Assessor
Smart Project Identifier
Risk
Finance
Emission
Cost
Energy
Feasibility
0
2015-01-01
Legenda
Gebouw 33.000m2 warmte
Aantal warm
Tijdreeks
X/Y-verstrooiing
Basislijn selecteren
Begin
Begin:
1-9-2014
15
Tijdsduur:
41
Maanden
Selecteer vergelijking
Y =
a*x+b
Y =3,1027*x+-1,6816
R2 =
0,9648

Social studies use statistics to quantify ‘new/unknown’ relations

In energy (CO_2), statistics is used to quantify known relations.

What is linear regression and why do we use it?

When evaluating the energy saving potential of building retrofits or design, we use linear regression to estimate the dependence of one variable (dependent variable), such as energy consumption, on one or more independent variables such as ambient temperature.

The goal is to determine the relationship existing between them.

Establishing a linear relationship between variables allows us to predict future energy use, and therefore **estimate** future potential energy savings from a project with some degree of accuracy.

Regression analysis is the essence of calculating energy savings.

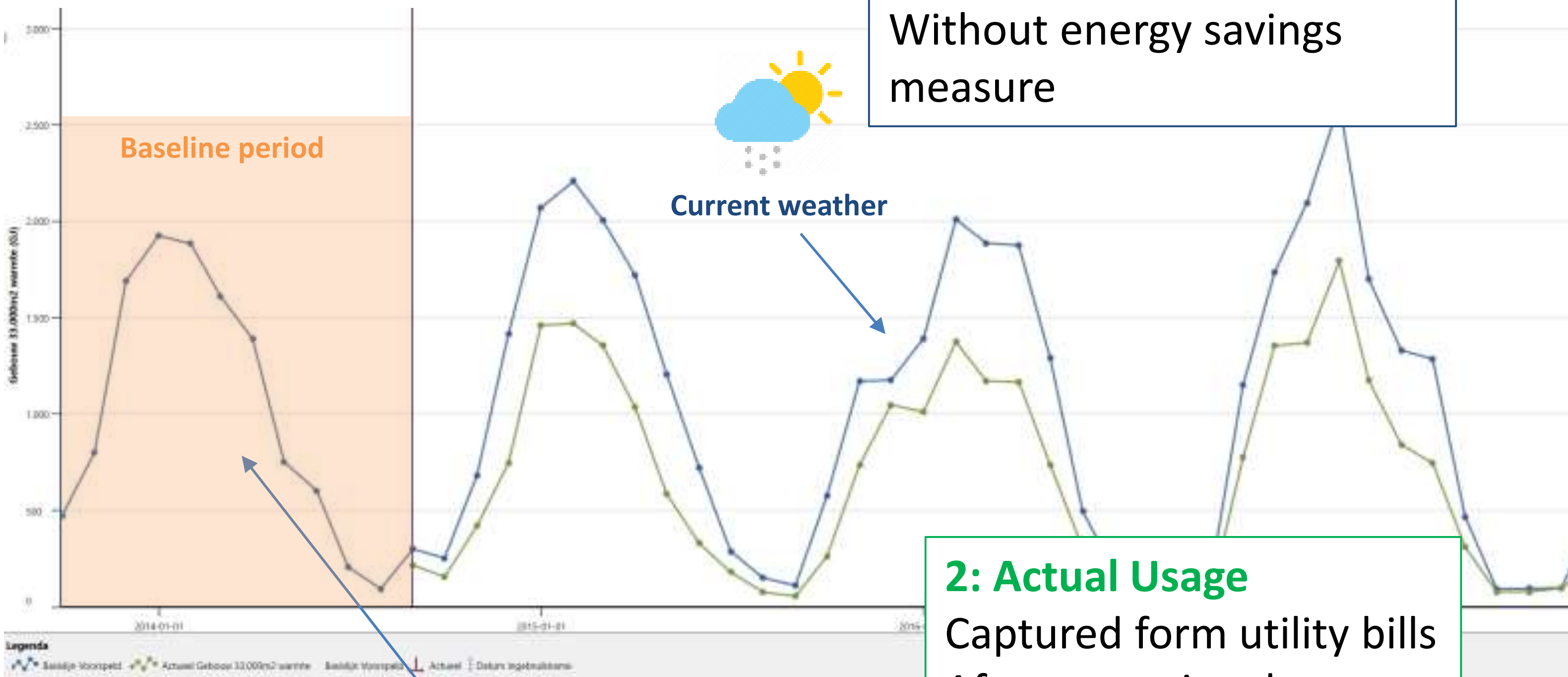
It’s devised of 3 steps:

1. **Creating a baseline for energy usage**
   use utility data preceding any change intended to improve operational efficiency with respect to energy consumption is collected and used as an established baseline
2. **Building a predictive model with energy usage independent variables**
   (for heating energy the independent variable could be weather)
3. **Comparing modeled usage to actual usage**

Energy savings measure operational

today

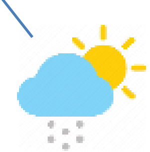
3: Modeled Usage
Of "What would have been"
Without energy savings
measure



Current weather

2: Actual Usage
Captured form utility bills
After operational energy
savings measure

1: Actual usage
Captured from utility bills
Used to build a <u>weather normalized</u> model



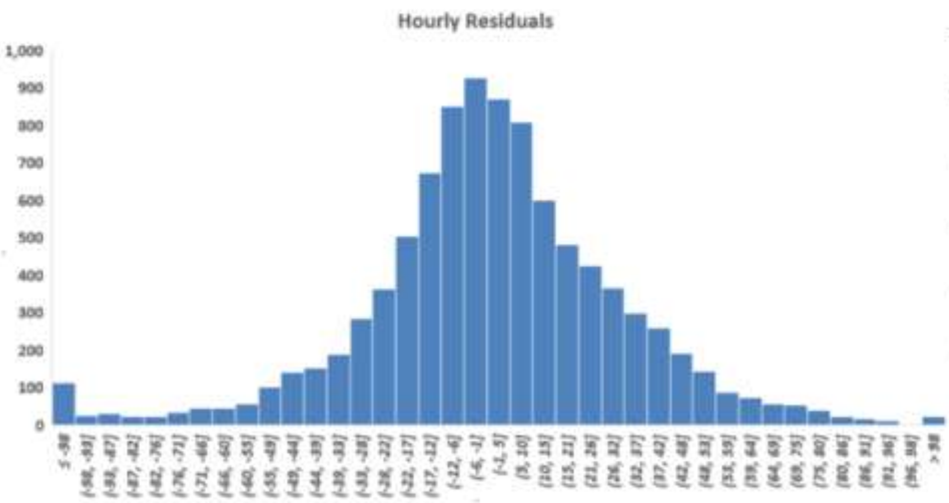
Historical weather

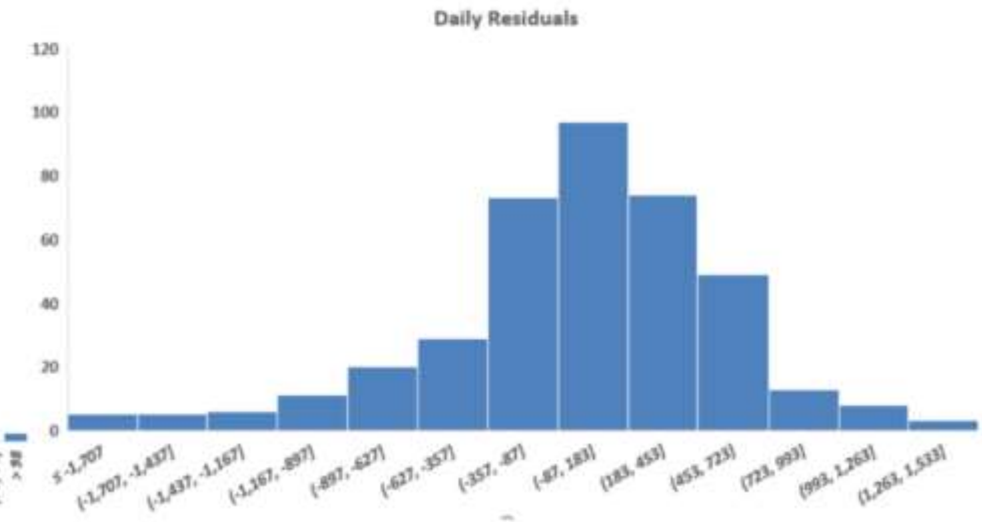
Savings = Modeled - Actual

But what quality of (utility) data should be used?

- Real-time, IoT?
- Hourly
- Daily
- Monthly

What is you’re assumption/first guesses?





Linear regression makes several assumptions about the data at hand. In the past, verifying the validity of these assumptions was not seen as an absolute requirement as the estimation methodology depended primarily on engineering wisdom established over decades.

However, as the industry moves from being a discipline rooted in experience to one based on empirical and fact-based methods, the need for understanding these assumptions and their implications is higher than ever.

Four main assumptions of linear regression are:

Linearity of data: The relationship between the predictor and the predicted is assumed to be linear

Normality of residuals: The residuals are assumed to be normally distributed. Residuals are the difference between the predictor and predicted values.

Homoscedasticity: (homogeneity of residual variance): The residuals are assumed to have a constant variance.

Independence of residuals: The residuals are assumed to be independent of each other.

It is important to note here that time series datasets, such as those used in the energy industry, require significant transformations in order to satisfy all these assumptions. , so we will focus on finding the conditions under which deviation from these assumptions is minimal.

Let’s compare annual hourly and daily datasets of a facility to see how the above listed assumptions hold up. Each dataset contains four variables: time, energy, temperature, and energy predicted.

Linearity of data is checked by plotting the predicted values (predicted energy use) against the predictor values (time, energy, temperature). This method is straightforward when we have only one predictor variable; however, it gets increasingly tedious as the number of predictor variables increase. Another method to check for linearity in data is to plot the residuals against the predicted values.

The data fulfills this linearity assumption if the residuals are scatted randomly around the zero line.

The data fulfills this linearity assumption if the residuals are scatted randomly around the zero line.

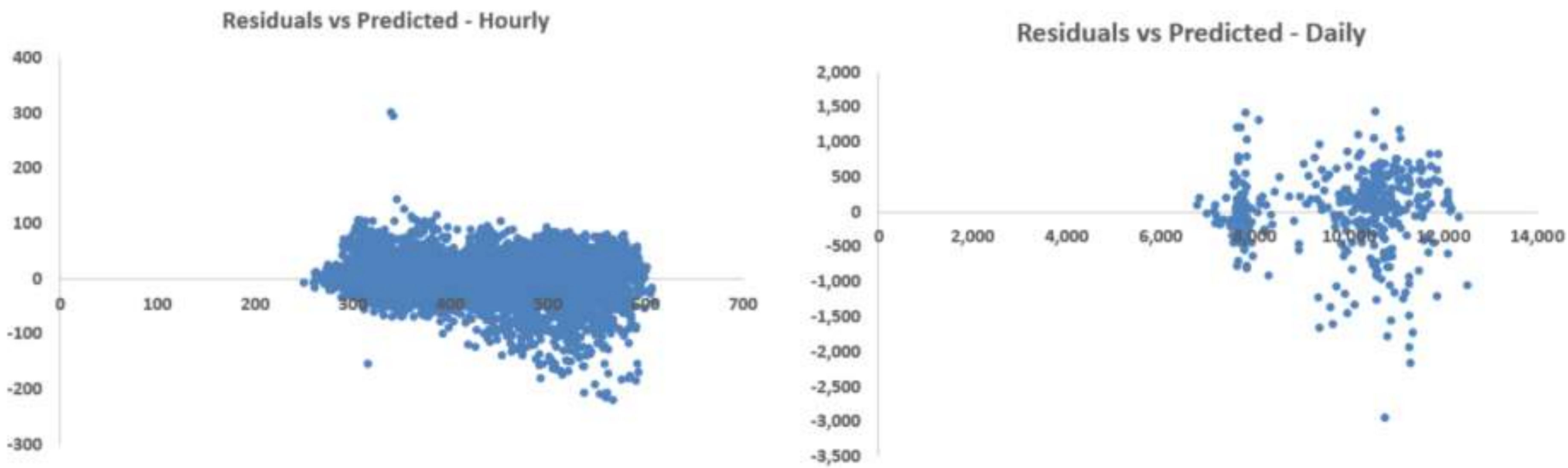


Figure 1: Residuals vs Predicted

A brief glance at Figure 1 shows greater amount of data above the zero line and greater deviations from the zero line on the bottom half of the graphs.

Normality of residuals is tested by plotting their histograms. Figure 2 shows that both the hourly as well as daily residuals are normally distributed.

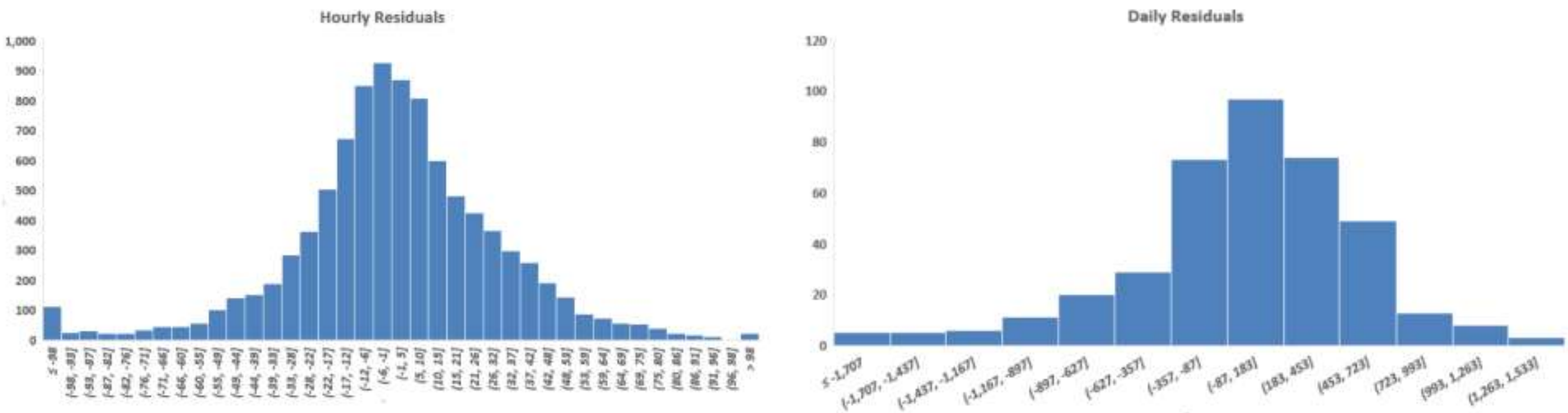


Figure 2: Residuals' Histograms

Homoscedasticity, or homogenous variance of residuals, is tested by plotting the residuals over time. Ideally, we would like to see a constant amount of scatter around the zero line. If the scatter increases or decreases over time, the dataset is in violation of this linearity assumption.

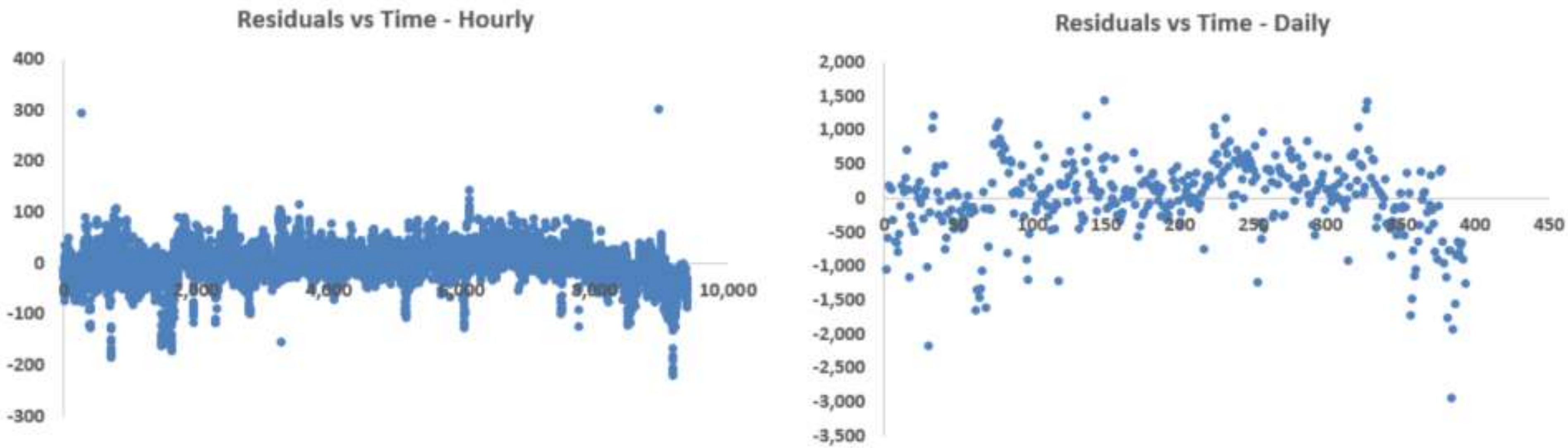


Figure 3: Residuals vs Time Observations

Figure 3 shows that the residuals, of both the hourly and daily models, have approximate constant variance except for the last few observations.

Independence of residuals

When the residuals from different time periods (usually adjacent) are correlated, we say that the residuals are autocorrelated. It occurs in time series data when the errors (residuals) associated with one time period carry over into future time periods. While this does not affect the unbiasedness of the ordinary least-squares (OLS) estimators, it causes their associated standard errors to be smaller than the true standard errors, leading to the conclusion that the estimates are more precise than they really are.

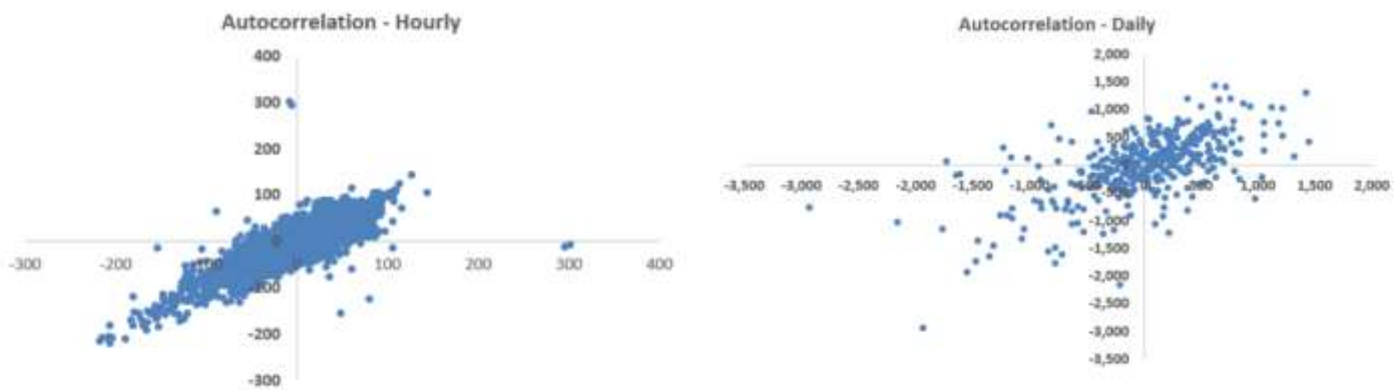


Figure 4: Autocorrelation

We check for autocorrelation by plotting the errors against themselves lagged by one time unit in a scatterplot. The correlation coefficient of the residuals is calculated by taking the square-root of the coefficient of determination (R2).

Independence of residuals

Data Interval	R^2	Correlation Coefficient
Hourly	0.73	0.85
Daily	0.38	0.62

Table-1: Correlation Coefficients

Figure 4 shows that as a residual's value increases (or decreases), the value of its immediate successor also increases (or decreases), implying a serial correlation, i.e. autocorrelation. R2, and its square-rooted value, are a quantification of this serial correlation.

A higher value the correlation coefficient is indicative of a high linear relationship and high correlation between the x and the y variables.

It is clear from Figure 4 and Table 1 that both the hourly as well as the daily residuals are heavily autocorrelated.

However, between the two **the daily dataset** has lesser correlated residuals.

Data Interval	R^2	Correlation Coefficient
Hourly	0.73	0.85
Daily	0.38	0.62

Table-1: Correlation Coefficients

Without transformation, time series datasets rarely satisfy all linearity assumptions.

By comparison of the two datasets on each count, we conclude that the daily data violates the linearity assumptions to a lesser degree than the hourly data, and should, therefore, be used as the basis of empirical energy modeling for this facility.

Does this conclusion also hold for daily data versus monthly data?

-Yes it does, most often!

What quality of (utility) data should be used for measurement and verification of energy savings?

- Start with 12 months, check the validity of the regression analysis. Enhance the resolution if needed
- It's easier to find and process 12 months of data than a mulitude of data
- More data doesn't mean automatically more information ☺

There might be scenarios wherein a granular profile of operation is required for project implementation. For example, demand side management requires an understanding of hourly demand data. In cases like these, while a daily model might provide a more mathematically sound estimation, the hourly model will serve the purpose of visualizing and quantifying hourly demand better. Both hourly as well as daily modeling practices come with their own specific advantages and risks. As a practitioner, it will be up to you to balance the advantages against the risks to determine a data frequency that is most suitable for your purposes.

RETScreen helps with statistical analysis

Equation

Gegevenstabel	dWDM Totaal
Afhankelijke variabele (Y)	Gebouw 33.000m2 warmte (GJ)
Onafhankelijke variabele (x)	Aantal warmte-graaddagen 17°C (°C-d)

Validation results

Heating degree-days 17°C	Pass

Pass

Regression results

Number of observations:	12
Number of iterations:	3
Sum of residuals:	0,0991
Average residual:	0,0083
Residual sum of squares - Absolute:	62,1287
Residual sum of squares - Relative:	62,578
Standard error of the estimate:	2,5016
Coefficient of multiple determination (R^2):	0,989
Coefficient of multiple determination - Adjusted (Ra^2):	0,9878
Root-mean-square error (RMSE):	2,4926
Coefficient of variation of the RMSE:	0,0772
F-test (p-value):	4,0706E-11
Net determination bias error (NDBE):	0,00026

Coefficient results

Name	Value	Standard error	t-ratio	p-value
a	4,6939	0,1569	29,918	4,0706E-11
b	1,9936	1,2397	1,6082	0,1389

Pass

vanhout adviseurs en installateurs

Questions?
We always have openings for student work.
☺