Kommunikations- und Technologieforschung

# SMART 2011 / 0072: Methodology for energy-efficiency measurements applicable to ICT in buildings (eeMeasure)

## D1.2 Non-residential methodology

**Version 2.0 August 2011**

**Author: Gregg Woodall**

**Prepared by:**

empirica Gesellschaft für Kommunikations- und Technologieforschung mbH,

Bonn, Germany (lead contractor/coordinator)

in co-operation with

Clicks and Links Knowledge | People | Technology

Clicks and Links Ltd

GREGG WOODALL ASSOCIATES

**Gregg Woodall Associates**

# Table of Contents

# 1 Introduction

This Measurement and Verification (M&V) methodology is intended to promote good practice and consistency in the reporting of ICT-PSP project results. Use of the methodology should also assist others to more clearly identify significant future energy saving opportunities including the development of local and national policy.

The methodology has been produced as part of the eeMeasure project and should be used in conjunction with the eeMeasure software and its integrated online user guide.

It is assumed that all projects consist of properties where an Intervention is made to reduce Energy Consumption and where Energy Consumption can be determined both with and without the Intervention.

## 1.1 Document structure

The document is structured as follows:

Chapter 1 – this introduction.

Chapter 2 - defines terms, intended to foster a common language between projects.

Chapter 3 - describes the high level M&V methodology.

Chapters 4 – 9 explain each step in the methodology

Appendices – provides additional background information.

## 1.2 Background of this document

This document is based primarily on four sources of information:

1. The Residential Methodology common deliverable published by eSESH[1] in September 2011 with the initial version completed by 3e-HOUSES[2] in September 2010.
2. The EVO International Performance Measurement & Verification Protocol (IPMVP)[3].
3. The experiences of other ICT-PSP residential and non-residential project teams.
4. Statistical techniques used in other situations.

[1] http://esesh.eu/fileadmin/eSESH/download/documents/outputs/CIP_Common_deliverable_eSESH.pdf
[2] http://www.3ehouses.eu/sites/default/files/3e-HOUSES_Deliv_1-2_Definition_of_Methodologies.pdf
[3] http://www.evo-world.org/

# 2 Definition of Terms

A wide variety of terminology is used in energy saving projects and local / European / international standards. The purpose of this glossary is to define a small number of important terms that can be commonly used throughout ICT-PSP energy efficiency projects in order to simplify communication and comparison.

The terminology is based primarily on EN16212 and the IPMVP (EVO International Performance Measurement and Verification Protocol), both of which have extensive glossaries that may be of benefit in standardising further terms.

***Baseline Period***

The time before an Intervention when no Energy Savings are expected.

***Baseline Period Consumption*** (based on EN16212)

Energy Consumption with no Intervention.

UNITS kWh / day or kWh / week etc

***Consumption***

See Energy Consumption.

***Control Group***

Demand Unit(s) where there is no Intervention and where there is a very similar Energy Consumption profile to the Experimental Group – also see Chapter 6.

***Demand Unit***

The location at which energy is consumed. A Demand Unit may be a property or any part of a property.

***eeMeasure***

The online software used to store, analyse and display the results of ICT-PSP projects.

***Energy Consumption***

Energy quantity i.e. the dependent variable.

UNITS ideally kg $CO_2$ per period, where kgCO2 = kWh * emission factor (kgCO2/kWh)

***Energy Consumption Model***

A regression model used to predict Energy Consumption based on Predictor Variables. The model takes into account the Routine and Non-Routine Adjustments specified by the IPMVP.

***Energy Saving*** (based on the IMPVP)

For Type 1 Projects

Energy Consumption = Pre-Intervention Energy Consumption

- Post-Intervention Energy Consumption

For Type 2.1 & 2.2 Projects

Energy Consumption = Average Energy Consumption in the Baseline Period

- Average Energy Consumption in the Test Period

For Type 2.3 Projects

Energy Consumption - Average Energy Consumption in the Control Group

- Average Energy Consumption in the Experimental Group

For Type 3 Projects

Energy Consumption = Av Test Period Energy Consumption as predicted by the Model

- Av Test Period Energy Consumption as measured

UNITS $kgCO_2$ / day or kWh / year etc

***Energy Saving Action***

A deliberate action directly resulting in Energy Consumption Reduction, which is either:

(i) a User action stimulated by a project Intervention.
EXAMPLE - a User turning off lights, stimulated by real time information.

(ii) an action undertaken by the project (in this case the terms Action and Intervention are interchangeable).
EXAMPLE - the project directs changing of lamp types.

***Energy Saving Intervention - ESI***

See Intervention.

***Energy Use*** (EN16212)

Manner or kind of application of energy.

EXAMPLE Lighting, ventilation, heating, processes.

***Experimental Group***

Demand Unit(s) where there is an Intervention.

***Frequency***

The number of measurement points per unit of time.

***Intervention (or Energy Saving Intervention – ESI)***

A measure implemented by the project which is expected to result in an Energy Saving. The measure may be either:

(i) a facilitating measure that stimulates User actions, or
(ii) a direct action taken by the project to reduce Energy Consumption (in this case the terms Action and Intervention are interchangeable).

NOTE the term Intervention or ESI should be used in preference to other similar terms including: energy conservation measure – ECM (IPMVP); energy efficiency improvement measure - EEI measure (EN16212); energy performance indicator – EnPI (ISO15001); energy efficiency measure – EEM; energy conservation opportunity – ECO.

***Measurement and Verification – M&V*** (based on IMPVP)

Determination of Energy Saving requiring both accurate measurement and a replicable methodology.

***Model***

See Energy Consumption Model.

***Monitoring*** (EN16212)

Recording and checking of metered and other data over a period of time.

***Monitoring Granularity***

The level to which Energy Consumption and Predictor Variables are monitored

EXAMPLE sub-metering to separate heating and hot water, or occupancy counting for different zones within a property.

***Predictor Variable***

Any factor that has a significant impact on Energy Consumption. When reference is made to a Predictor Variable, the implication is that it has an impact on demand also see Chapter 6.

NOTE the factors may be classified as Routine or Non-Routine Adjustments according to the IPMVP.

***Pre-Intervention***

The time before an Intervention in Type 1 Projects.

***Post-Intervention***

The time following an Intervention in Type 1 Projects.

***Test Period***

The time following an Intervention when Energy Savings are expected.

***Test Period Consumption*** (based on EN16212)

Energy Consumption with Intervention.

UNITS kWh / day or kWh / week etc

***User*** (based on EN16212)

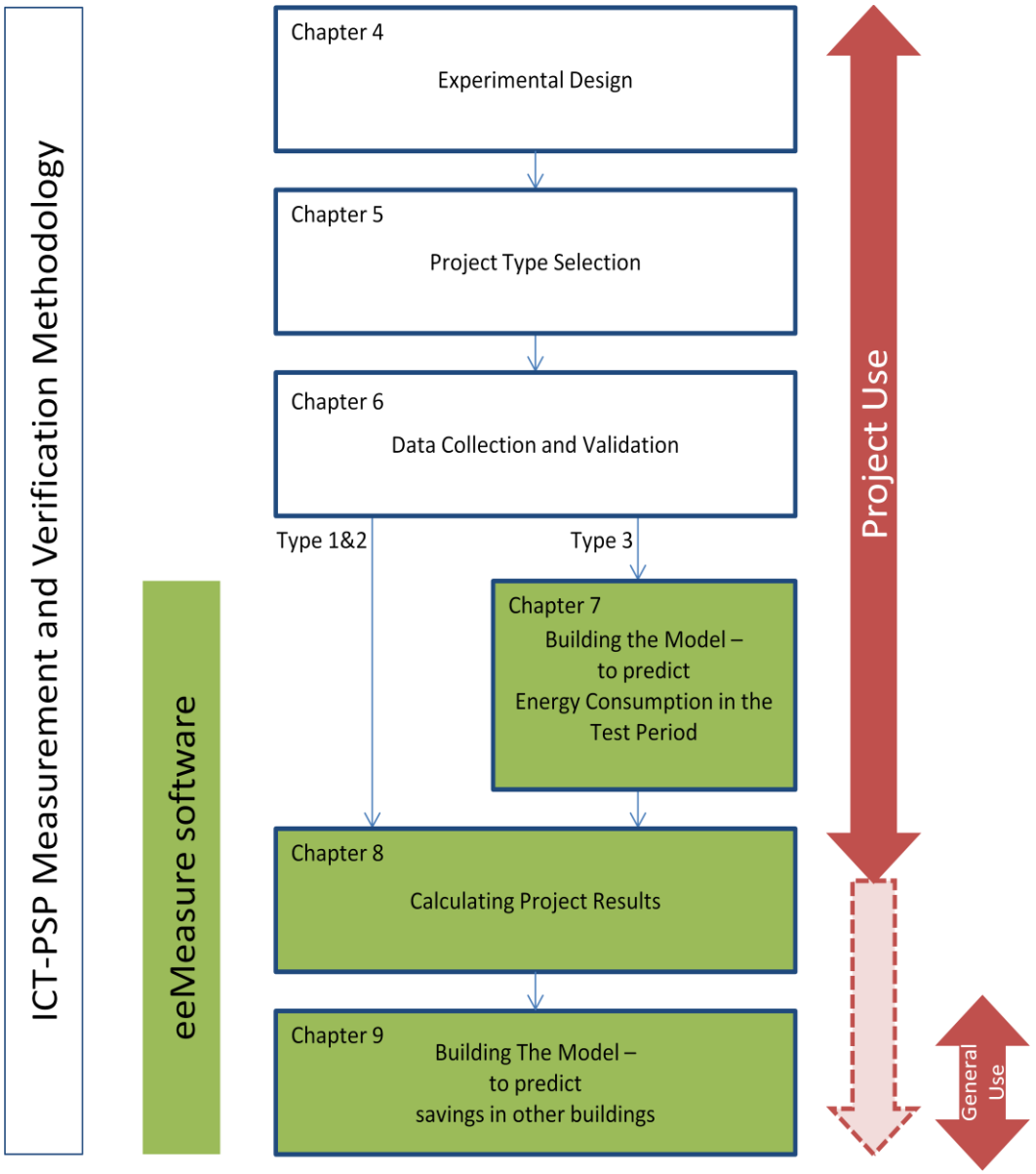
Ultimate person(s) consuming energy for final use.

***User Behaviour Change (or User Behaviour Transformation – UBT)***

Change of User behaviour resulting in reduced Energy Consumption.

# 3 High Level Methodology

The high level process flow shown below is created to be used by ICT-PSP projects. The first five steps are those normally taken by a project to report results. The final step helps policy makers and project managers to determine potential Energy Savings in other properties. Each step is described in the following chapters.




Note that the integrated eeMeasure online user guide provides a more detailed description of the process flow to be used when uploading data, building models and calculating results.

# 4 Experimental Design

The outline design of a project will determine the statistical significance of results and should therefore be considered carefully. In general terms –

- Experiments should be replicated i.e. the sample size should be large enough to identify a normal distribution of results and to identify significant outliers.
- Experimental periods should be long enough to include a representative set of conditions that will impact Consumption.
- Periodic variation of Consumption should be predictable. It is possible to predict variable Consumption patterns by either:
  (i) Creating a Model during a Baseline Period with Predictor Variables as the factors in a regression equation.
  (ii) Using a Control Group with very similar characteristics to the Experimental Group.
  (iii) A combination of (i) and (ii).

It should also be noted that:

a. There may be some Demand Units that are not possible to model and therefore energy savings cannot be calculated.
b. Numerous technical issues could make it impossible to collect reliable data from specific Demand Units.

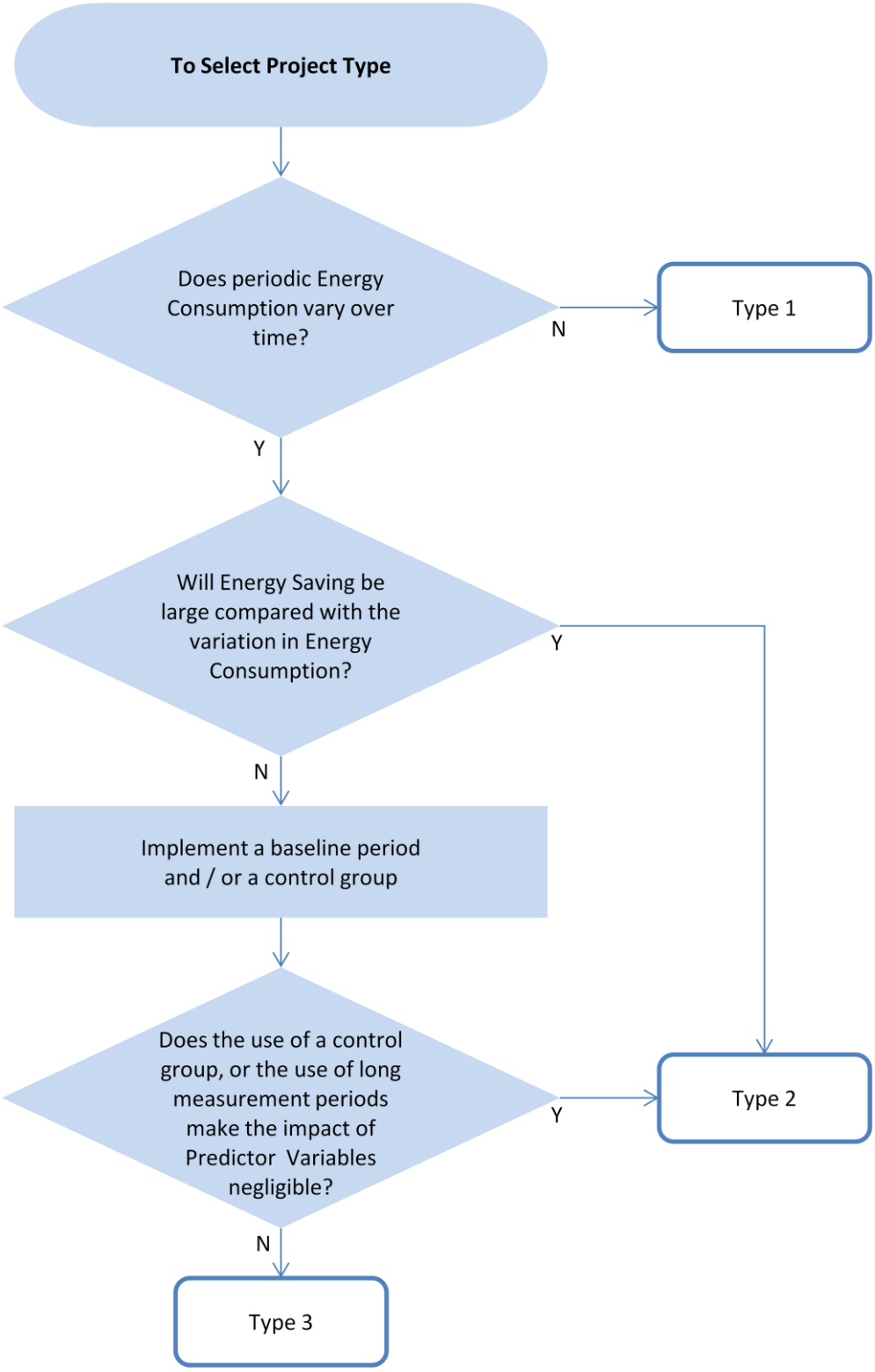
Therefore expect that some Demand Units selected at the start of a project may not be usable in the final results.

**Experimental Design – Notes**

It is critical that projects are designed in a way that facilitates the calculation of accurate and useful energy savings results. The consequence of a poorly designed project is that there can be little confidence in the results.

# 5 Project Type Selection

There are broadly three Project Types and the flow diagram below shows the process to select the appropriate Type:



Project types are categorised as follows:

| Type | Characteristics | Necessary Data |
|---|---|---|
| 1 | Energy Saving can be simply calculated by subtracting the Post-Intervention Energy Consumption from the Pre-Intervention Energy Consumption. In this case the Energy Consumption may be either measured or calculated from trusted device specifications. An example Intervention would be the replacement or reprogramming of an energy consuming device. | Pre-Intervention Energy Consumption,<br>Post-Intervention Energy Consumption. |
| 2.1 | Energy Saving is large compared with the variation in Energy Consumption i.e. the impact of Predictor Variables is negligible.<br>A more accurate result may be obtained by averaging results over a number of measurement periods. | (Average) Baseline Period Consumption,<br>(Average) Test Period Consumption. |
| 2.2 | Variation of Energy Consumption during the Baseline Period is closely replicated by the variation of Energy Consumption during the Test Period and each period includes a representative range of conditions that significantly impact Energy Consumption. | Periodic Baseline Period Consumption,<br>Periodic Test Period Consumption. |
| 2.3 | Variation of Energy Consumption in the Experimental Group is closely replicated by the variation of Energy Consumption in the Control Group. | Periodic Test Period Consumption for the Experimental Group,<br>Periodic Test Period Consumption for the Control Group. |
| 3 | Energy Consumption is significantly impacted by Predictor Variables (e.g. external temperature, daylight hours, occupancy variations, equipment failure) that cannot be fully compensated for by a Control Group.<br>Note that in this case, the Predictor Model can be used in conjunction with a Control Group. | Periodic Baseline Period Consumption and Predictor Variables,<br>*a Predictor Model*,<br>Periodic Test Period Consumption and Predictor Variables. |

**Relationship between Project Types and IPMVP Options**

IPMVP has been developed for a wide range of situations, primarily focused on the definition of contractual terms where a third party is responsible for energy saving. This methodology expands the IPMVP methodology to cater specifically for ICT-PSP energy saving projects.

IPMVP has four calculation options – A,B,C&D

Option A is equivalent to Type 1.

Option B could be described in terms of Type 2 or Type 3.

Option C could be described in terms of Type 2 or Type 3.

Option D is beyond the scope of this methodology.

**Select Project Type – Notes**

eeMeasure does not make any reference to Project Types as the software is able to calculate Energy Savings for all of the three Types.

Project Types are included in this methodology to aid understanding of the data and process required for each Project.

# 6 Data Collection and Validation

The periodic Energy Consumption in any property tends to fluctuate significantly and is dependent on a number of factors (e.g. external temperature, daylight hours, holidays). These influencing factors are referred to as the Predictor Variables and although external temperature often has the greatest influence, there are likely to be other Predictor Variables that significantly impact Energy Consumption.

It is assumed that in all cases Energy Consumption will be metered and the meter reading data may be collected manually or automatically.

**Control Groups**

It is possible that the impact of some Predictor Variables (e.g. external temperature) can be nullified by comparison of the Experimental Group with a Control Group. This methodology is only useful if the Control Group truly reflects the periodic change to Energy Consumption in the Experimental Group. In this case, the Control Group is effectively predicting the no-intervention Consumption in the Experimental Group.

Unfortunately it is often difficult to find an appropriately reflective Control Group. The best type of Control Group is a randomised group selected from a large sample of Demand Units all of which have similar Consumption characteristics. The wider the range of Consumption characteristics, the larger the Experimental and Control groups need to be.

An alternative method to determine the relationship between a Control Group and an Experimental Group is to perform a regression analysis, identifying the extent to which the two groups are correlated.

If a Control Group cannot be used to nullify all significant Predictor Variables, then the calculation of Energy Saving needs to be derived from a Baseline. Note that Control Group and Baseline methodologies can be combined to create a more accurate model in some circumstances.

**Predictor Variables**

A Predictor Variable is any measurable factor that impacts Energy Consumption. Predictor Variables are used in regression analysis as described in Appendix A. It is important to monitor all significant Predictor Variables and the impact of multicollinearity should also be understood (see Appendix A)

It may be possible to automate the collection of Predictor Variables (e.g. visitor numbers) or information may need to be collected manually (e.g. equipment failure / replacement log). All monitored data must have a time stamp with an appropriate accuracy. It is possible that there will be 10's of Predictor Variables, but the first step is to identify the most significant ones and find a way to measure them.

Predictor Variable data may be sourced from:

Commercially available information e.g. HDD, daylight hours

Simple information e.g. school holiday dates

Existing processes e.g. museum visitor numbers

New processes e.g. equipment failure / replacement log

The selection of Predictor Variables is also discussed in Appendix A.

**Fixed Data**

There are some factors which do not change over time, but which do impact Demand Unit Consumption. This "Fixed Data" is not used in a Model to predict no-intervention Consumption, but it can be used along with project results to predict Consumption in properties within the same range. An example of Fixed Data is "floor area". In this case the Energy Saving results from a number of properties can be used, along with the floor area of those properties, to predict the Energy Saving of any similar property with a given floor area.

There are some factors that could be Predictor Variables or Fixed Data. An example of this is occupancy. In the case of a dwelling, day-to-day occupancy could significantly impact Energy Consumption and occupancy would be a Predictor Variable. If occupancy does not vary over time, then "fixed" occupancy data can be used, along with project results, to predict the energy saving of any similar dwelling with a known occupancy.

**Data Frequency**

It can be useful to record and analyse relatively high frequency data (e.g. hourly) to detect the Predictor Variables that significantly impact Energy Consumption. High frequency data may also be useful to analyse the impact of an intervention. However, for the purposes of calculating Energy Saving, the most important issue is to identify the Predictor Variables that have a relationship with Energy Consumption that can be modelled.

It is likely that the correlation between Energy Consumption and Predictor Variables is likely to be higher with low frequency data.

**Sub-Metering**

Sub-metering (Monitoring Granularity) should be determined by the need to create Models based on different Predictor Variables. For example - cooking gas may not be dependent on external temperature and therefore if cooking gas use is significant, then heating and cooking gas should be separately metered.

**Metering Accuracy**

Where the potential for Energy Saving is relatively small (proportional to Energy Consumption) monitoring accuracy must be good. Where an Energy Saving Intervention is likely to yield a very significant saving, monitoring accuracy is less important.

**Corrupt Data**

Corrupt data can often be identified by an automated data collection system and the system may also automatically replace the corrupt data with an interpolated value. If the Energy Consumption data is of a higher frequency than any significant Predictor Variable, or if a Predictor Variable is of a higher frequency than the Energy Consumption data, then regression cannot be used to identify the corrupt data and it may be of benefit to graphically display the data so that any changes to regular patterns can be easily identified. If a corrupt data point is identified then a replacement value should be calculated by interpolation. Ideally all edited values should be identified by some form of tag (an example would be to use a different colour for corrected data in a spreadsheet[4]).

If the Energy Consumption data is of the same Frequency as any significant Predictor Variable, then eeMeasure can be used to identify corrupt data which will have a large residual value when the regression is applied. In this case, no useful information is added if a data point is calculated by interpolation and therefore the corrupt data should be excluded.

**Validation**

There are many potential problems that may occur during the data collection process and whenever possible, data should be validated by an alternative method of measurement e.g. a hand held meter or observation of meter readings. Bad data is worse than no data as it implies a result that is not necessarily correct.

**Waste Detection**

When energy data is analysed, it is often possible to quickly identify significant potential energy savings e.g. by plotting hourly electricity Consumption, it may become obvious that lights are being left on for too long in the evening. It is then very tempting to correct this problem because we are focused on saving energy. However, this correction may seriously change the results of the project and therefore any such correction must be considered very carefully in the context of the Intervention that is being tested.

**Data Collection and Validation – Notes**

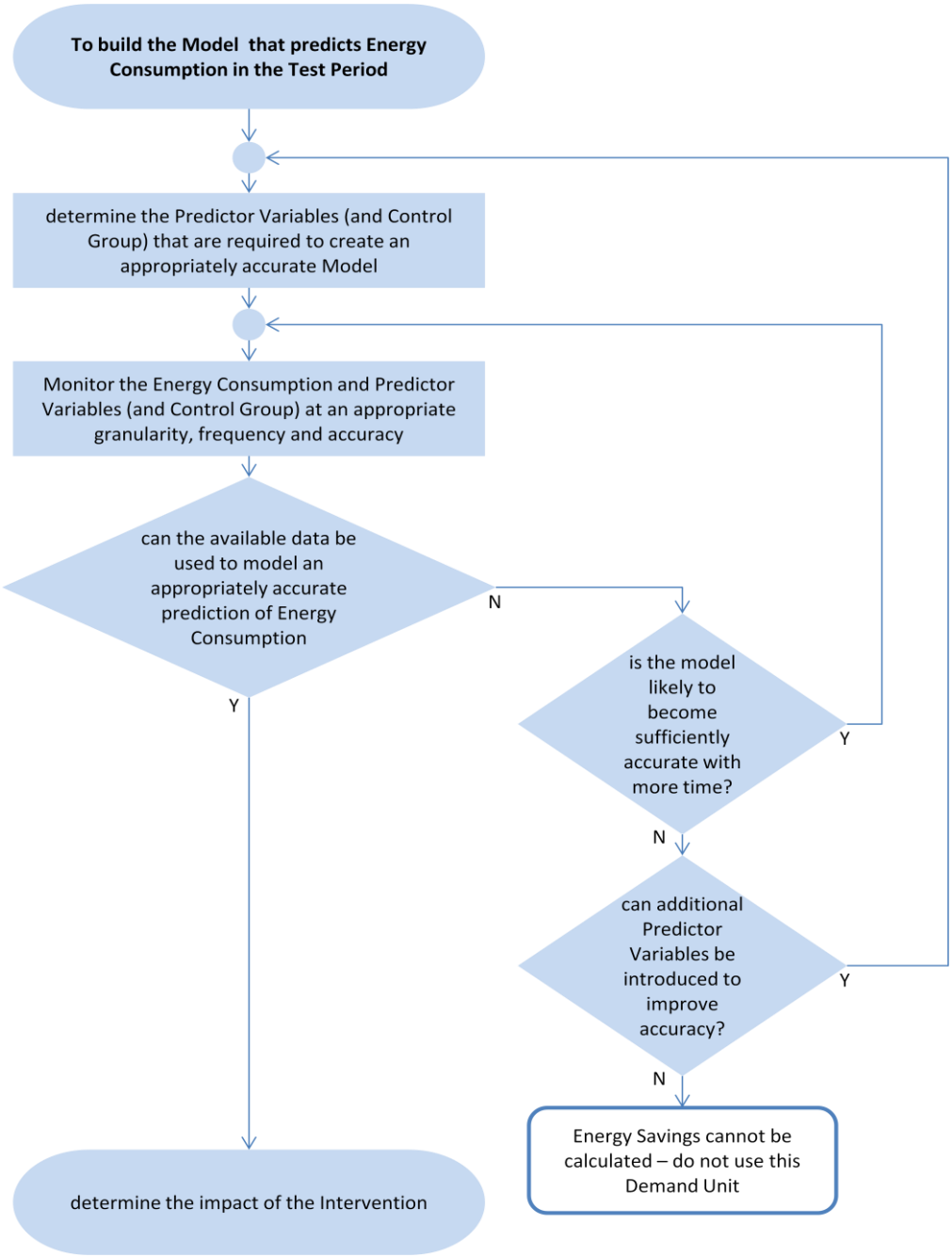
It is essential that appropriate data is collected and validated. Selecting the sources of data may be an iterative process and projects may need to allow time to restart measurement periods if appropriate data was not available from the start.

[4] Note that these tags are not carried through into eeMeasure and are only useful for the project to identify corrected data

# 7 Building The Model – to predict Energy Consumption

The Type 3 Model is used to predict what the Energy Consumption would be during the Test Period if there had been no Intervention. The Model is based on regression analysis as explained in appendix A (note that use of a Control Group is optional).

The objective is to make the model as accurate as possible. If it is not possible to create a sufficiently accurate model, then the uncertainty of results may be huge.



**Building the Model**

**Step 1** - Monitor Energy Consumption and Predictor Variables until sufficient data points are available to determine a correlation between Energy Consumption and Predictor Variables.

**Step 2** – Upload Baseline Period Energy Consumption and Predictor Variable data to eeMeasure. The Frequency for Energy Consumption and Predictor Variables does not need to be the same.

**Step 3** – Use eeMeasure to perform a regression analysis and consider the calculated determination of correlation ($R^2$) and value of residuals.

**Step 4** – Remove any outliers that are either:

(a) Obvious data collection errors, or

(b) Other obvious anomalies (e.g. public holiday)

**Step 5** – Decide whether the model is appropriately accurate to predict future Energy Consumption from any likely combination of Predictor Variables, because:

a. The set of Predictor Variables includes a reasonable range of values (e.g. seasonally high and seasonally low external temperatures; high occupancy and low occupancy).
b. The accuracy with which Energy Consumption can be predicted is good relative to the expected Energy Saving. Note that a high value of $R^2$ is a good starting point, but it should not be the only measure of appropriate accuracy.

If the Model is appropriately accurate, then the Baseline Period can be finished and the Intervention can be implemented. If the Intervention is not ready for implementation then the Baseline Period can continue and accuracy of the Model may be further improved.

**Step 6** - If the Model is not appropriately accurate, then:

a. The Baseline Period should be extended if more data is likely to make the model more accurate.
b. Further Predictor Variables should be added (if available) to make the model more accurate
c. The Baseline Period should be re-started if other Predictor Variables are needed and which have not been monitored to date.
d. The Demand Unit should be abandoned if there appears to be no way of creating an appropriately accurate Model.

If R2 is low, further Predictor Variables must be found to improve predictions. If R2 remains low, only very large Energy Savings will be reliably detected.

**Peak Shaving**

The impact of a "peak shaving" Intervention can be calculated by building a model based only on the peak time during each day. The peak period can be set in eeMeasure.

**Type 3 Model – Notes**

Building the Model is an iterative process and the analysis of different Predictor Variables can be time consuming. A good Model is critical in delivering accurate project results.

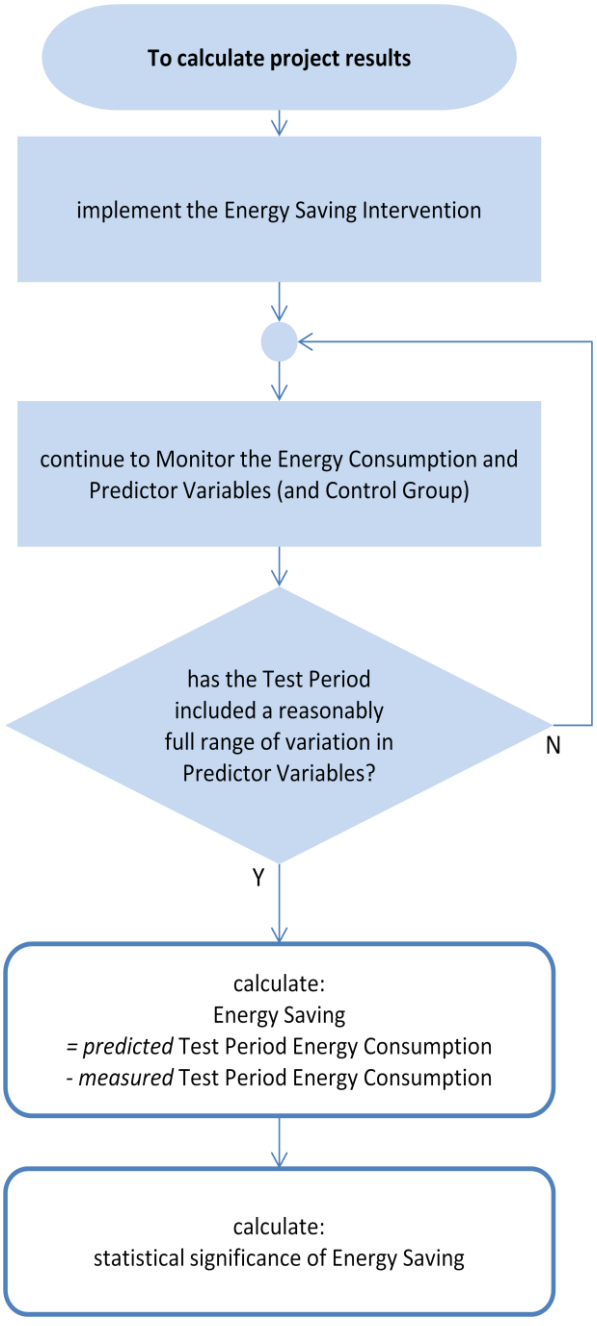
# 8 Calculating Project Results

Type 3 Results are calculated from:

Energy Consumption = Av Test Period Energy Consumption as predicted by the Model

- Av Test Period Energy Consumption as measured

Where the Model has been appropriately defined as per Chapter 7 and where the Test Period is long enough to include an appropriately representative range of Predictor Variables.



**During The Test Period**

**Step 1** - Monitor Energy Consumption and Predictor Variables until sufficient data points are available to provide an indication of Test Period Energy Consumption.

During the data collection process, identify corrupt data and correct by interpolation where appropriate. All corrected data should be tagged if possible (an example would be to use a different colour for corrected data in a spreadsheet[5]).

**Step 2** – Upload Test Period Energy Consumption and Predictor Variable data to eeMeasure. The Frequency for Energy Consumption and Predictor Variables does not need to be the same.

**Step 3** – Use eeMeasure to compare predicted and measured values and assess the range of calculated Energy Savings over time.

**Step 4** – Remove any outliers that are either:

(a) Obvious data collection errors, or

(b) Other obvious anomalies (e.g. property closed for emergency repairs) that were not accounted for by the Predictor Variables.

**Step 5** – eeMeasure calculates the average saving.

Following an Intervention, the calculated Energy may change from week to week with an interesting trend. In some behaviour change projects the initial Energy Savings will be high as the Intervention is greeted with enthusiasm and then savings may reduce over time as enthusiasm diminishes. In other behaviour change projects the initial savings may be low as people get used to working in new ways and then increase over time as the new behaviours become habitual.

The common approach to comparing the impact of Interventions is to use the average Energy Saving over the Test Period. This approach should be considered carefully and it may be more useful to average the Energy Savings only during a period of consistent results.

All final measurement results should be projected to annual savings (percentage, kWh and $kgCO_2$)

**Statistical Significance**

The result of the test of the statistical significance of the Energy Saving of the project is shown in eeMeasure. The significance of the regression coefficients is based on the calculation of Student's t. Further information on significance calculation and the t statistic can be found at http://elsa.berkeley.edu/eml/ra_reader/8-hypo.pdf

[5] Note that these tags are not carried through into eeMeasure and are only useful for the project to identify corrected data

# 9 Building the Model - to Predict Savings in Other Properties

There are fixed factors that affect the Energy Saving in every property (e.g. socio-economic group, property age, occupant age, floor area, heating type, location). If there are project results where the same Intervention has been made in many different properties, then it is possible to model the impact that these factors have on Energy Saving.

This model is of the same form described in chapter 7

Energy Saving = a' + b'V'1 + c'V'2...

Where a' is a constant and b',c'... are co-efficients that are multiplied by the value of their respective Predictor Variable V'1, V'2...

# Appendix A – Basic Statistics

**The Model**

To build a Model it is necessary to understand some basic statistical terminology and concepts. This appendix introduces the basic statistics knowledge that is needed to calculate savings.

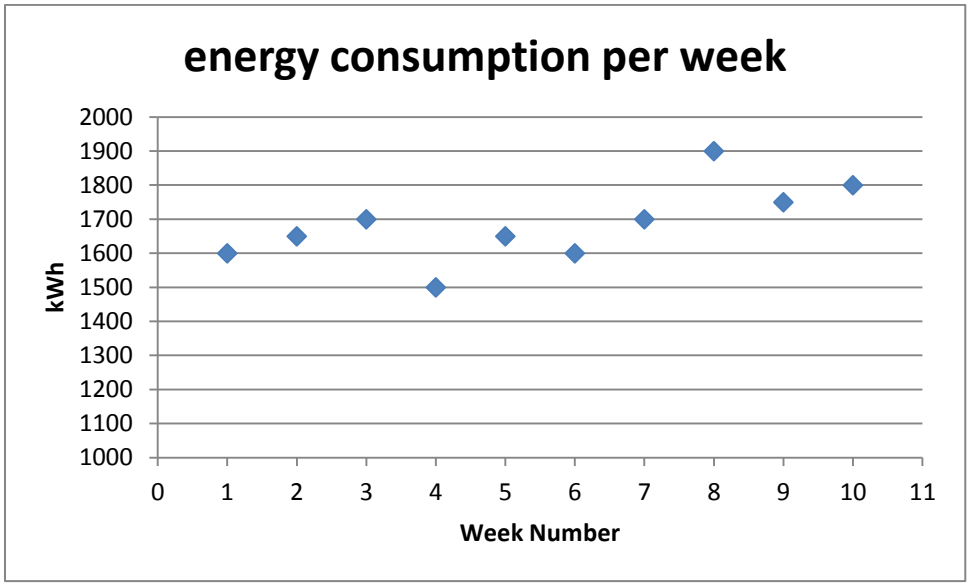
**Variables**

Energy Consumption generally changes as a result of one or more factors that vary over time. These factors are referred to as "Predictor Variables".

**Linear regression**

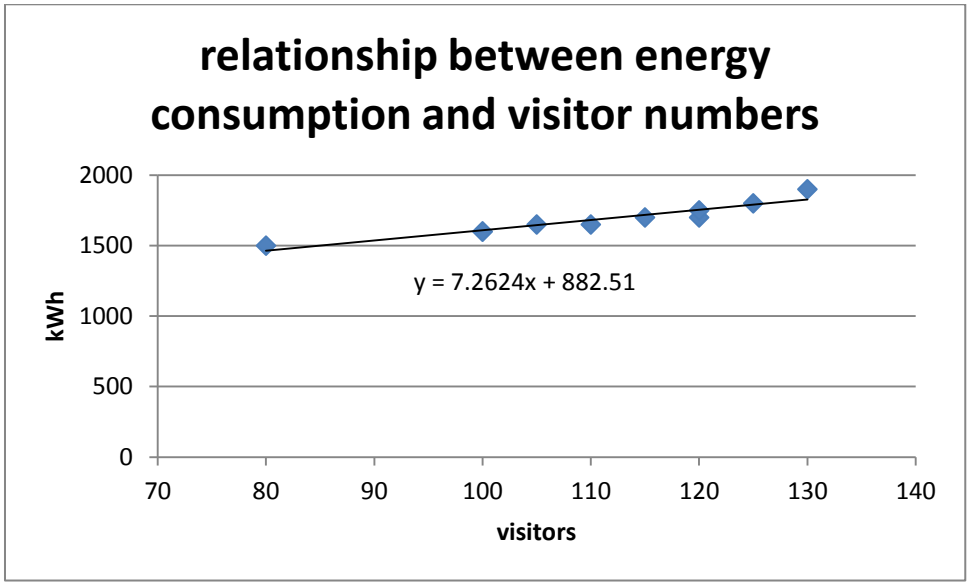
To calculate the Energy Saving that results from an intervention, it is necessary to know what the Energy Consumption would have been if the intervention had not been made. eeMeasure uses linear regression[6] to predict future Energy Consumption and the concept of linear regression is explained here by example.

In simple cases, Energy Consumption is strongly influenced by just one Predictor Variable and in this example it is the number of visitors in a public building. Weekly Energy Consumption over a ten week period is shown here:




[6] Linear regression is used by default, but it is possible to construct non-linear relationships using the eeMeasure Dataset Manipulation function

The relationship between visitor numbers and Energy Consumption can be plotted on a scatter graph:

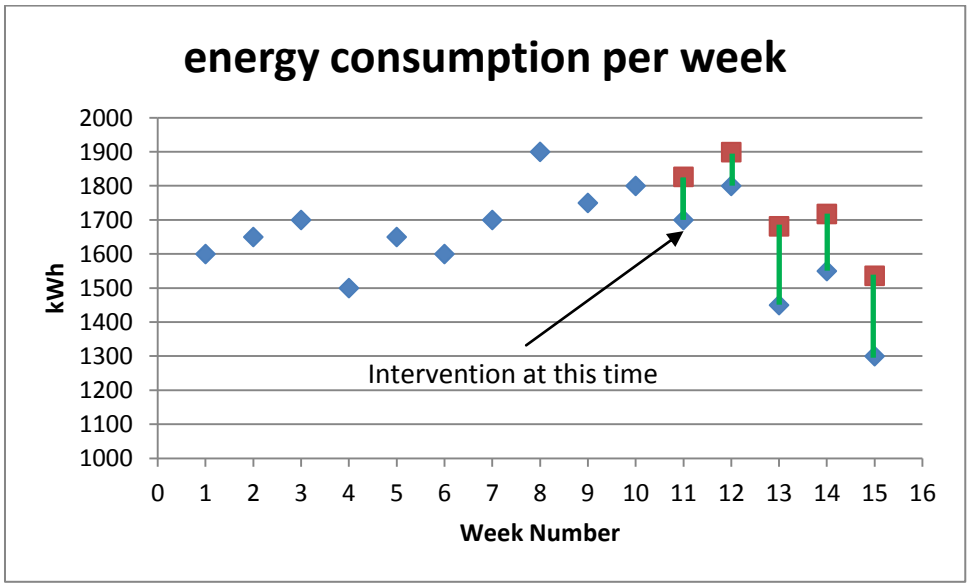



The graph includes a line of best fit and the equation of this line is the regression relationship between Energy Consumption and visitors numbers (the Predictor Variable). The equation is of the form: $y = bx + a$

Where "b" is the correlation coefficient and "a" is the regression constant

The equation can be used to predict Energy Consumption for any number of visitors within reason.

If an Energy Saving Intervention is made at the end of week 10, then the Energy Consumption in weeks 11+ can be predicted by the regression equation. In the graph below, blue diamonds represent actual Energy Consumption measurements and red squares represent the Energy Consumption (predicted by the model created from data in weeks 1-10 and) derived from visitor numbers in weeks 11-16. Weekly savings are represented by the green lines.



**Residuals and Determination of Correlation**

In the example above, the relationship between Energy Consumption and visitor numbers could clearly be approximated to a straight line. The vertical difference between a measurement point and the line is known as the Residual and a measure of how well the line represents the actual measurements is the Determination of Correlation or $R^2$. This number is calculated from the measurement values, the equation of the line and the residuals. $R^2$ always has a value between 0 and 1, where high values indicate good correlation and low values indicate poor correlation.

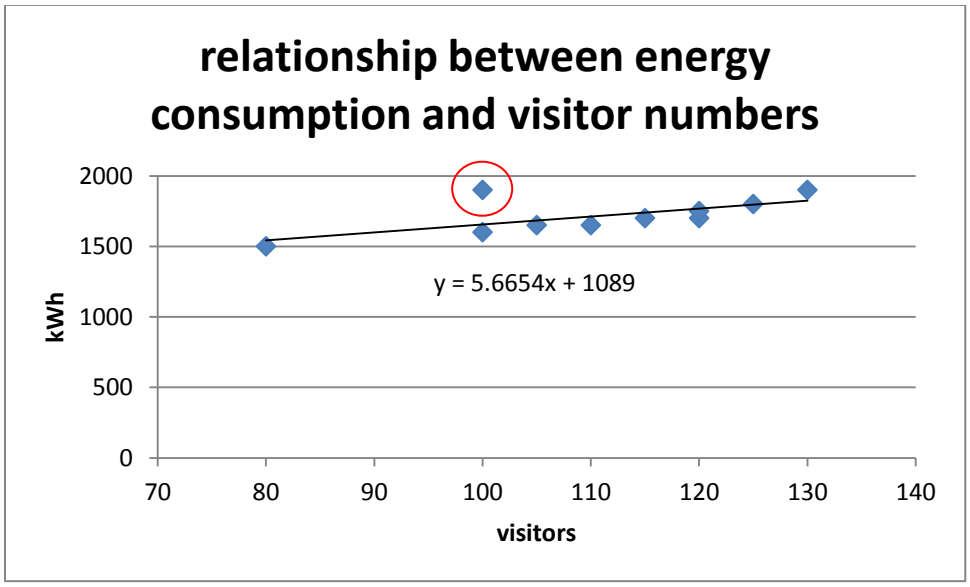
In the example above, the calculated $R^2$ = 0.90 This means that 90% of the variation in Energy Consumption can be attributed to the change in visitor numbers i.e. 10% of the variation is due to other factors that we have not considered.

$R^2$ is a good indicator of correlation, but it does not always provide all of the information required to determine whether the model is sufficiently accurate. This is discussed in later paragraphs.

**Outliers**

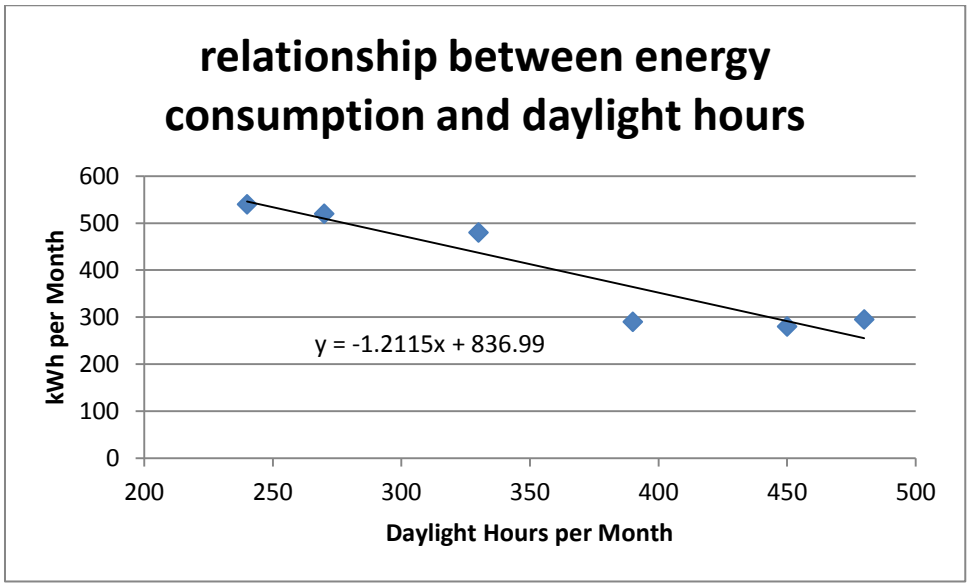
In the example below there is obviously one measurement point that is an outlier i.e. it is further from the line than most other data points. This outlier may be the result of corrupt data, or it may indicate that there is another Predictor Variable that also needs to be considered.

It is essential that all significant outliers are investigated to determine their cause.



**Non-Linear Regression**

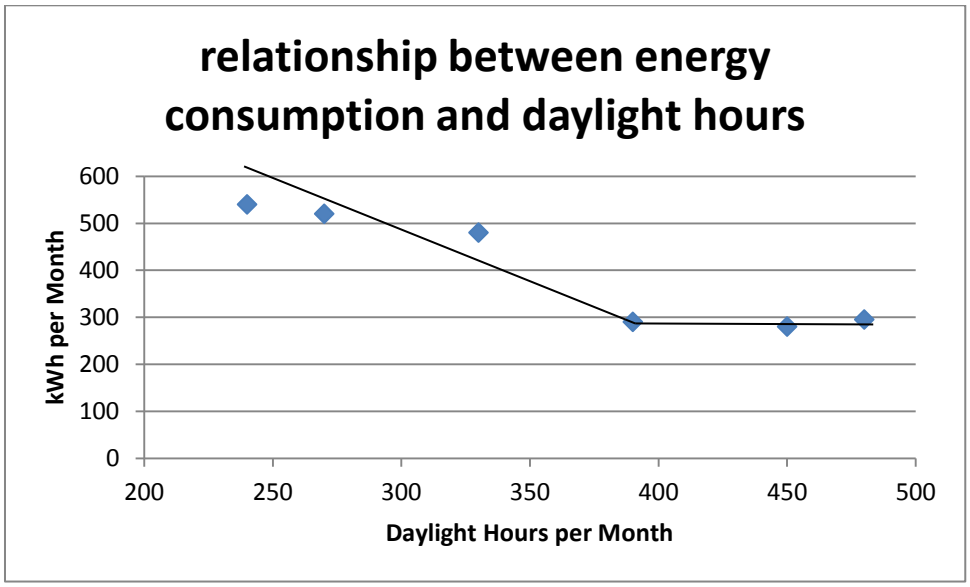
It is possible that the relationship between Energy Consumption and a Predictor Variable is non-linear i.e. the relationship does not approximate to a straight line. The example below shows a scatter graph where the relationship between daylight hours and Energy Consumption is being investigated.




From a quick observation, it looks like the line is a good approximation to the results and that the vertical distance between the points and the line is not great. In fact $R^2$ = 0.88 which seems quite good. However the table of the results shows a different story.

| Month | hours | kWh |
|---|---|---|
| 1 | 240 | 540 |
| 2 | 270 | 520 |
| 3 | 330 | 480 |
| 4 | 390 | 290 |
| 5 | 450 | 280 |
| 6 | 480 | 295 |

It can be seen from the table that when the monthly number of daylight hours reaches approximately 390, then the Energy Consumption no longer decreases. A more accurate representation of reality is a graph with a point of inflection. The same data points are plotted below, but with a different line of best fit.



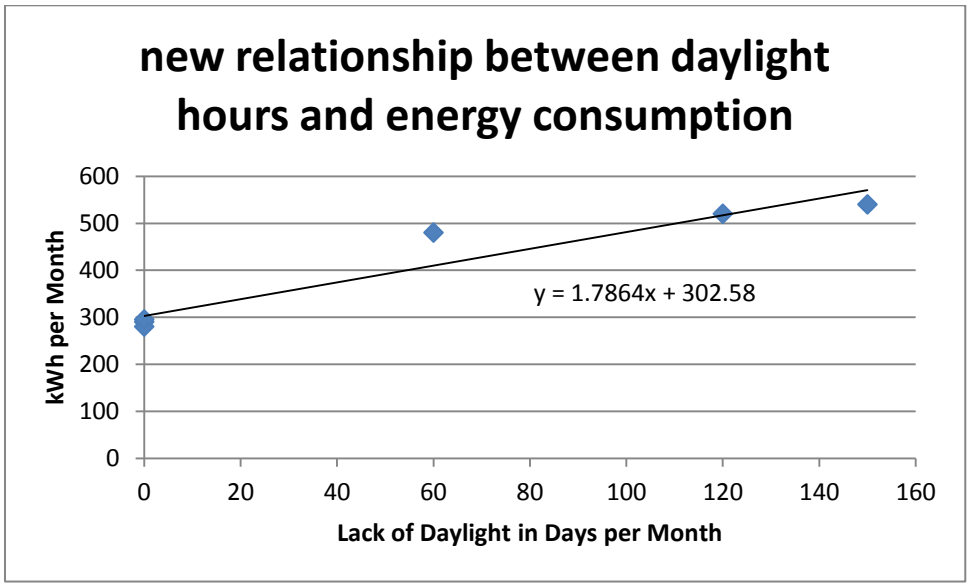

Now we can see that although $R^2$ was high, the regression equation would not be very good at predicting Energy Consumption based on daylight hours per month. It is therefore essential to consider factors as well as $R^2$ when deciding whether linear regression is appropriate.

There is a very similar issue if peak or average external temperature is used as a Predictor Variable. When the temperature reaches a certain point, heating will be switched off and further increases in temperature will not be matched by reduced Energy Consumption. This is why the concept of Heating Degree Days (HDD) is used. HDD is a measure of the days when external temperature is below a threshold value and it is assumed that internal space heating is required. The relationship between Energy Consumption and HDD is therefore linear and can be used to predict future Energy Consumption. Similarly the daylight hours data used above could be converted to values that have a linear relationship with Energy Consumption.

If we know that Energy Consumption is only related to a lack of daylight, when the number of daylight hours is less than 390 per month, then

Lack of Daylight Hours per Month = 390 - Daylight Hours per Month

And if the result is less than zero, then Lack of Daylight Hours per Month = 0

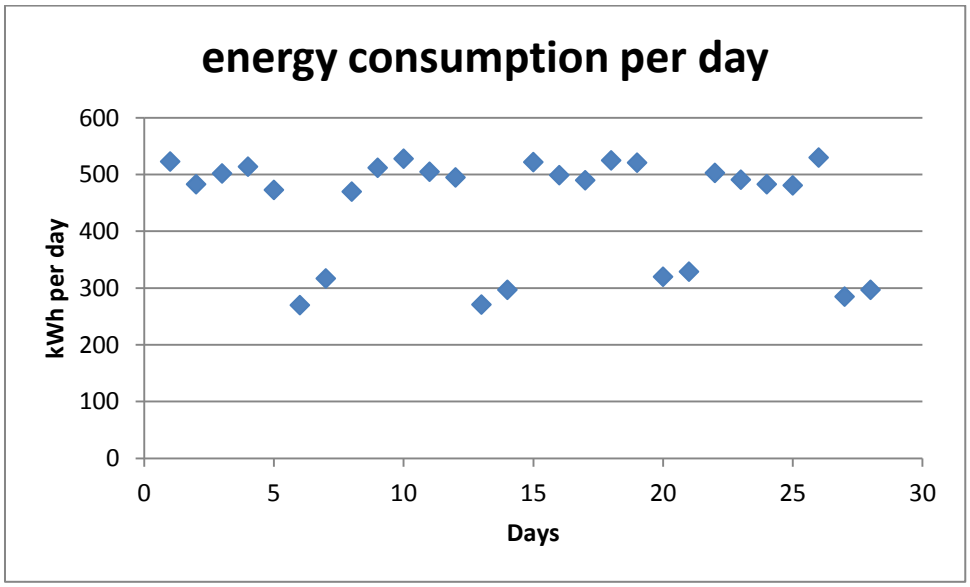



In this case $R^2$ = 0.91 but the small improvement in $R^2$ does not tell the full story of how much better this equation will predict Energy Consumption compared with using actual daylight hours as the Predictor Variable.

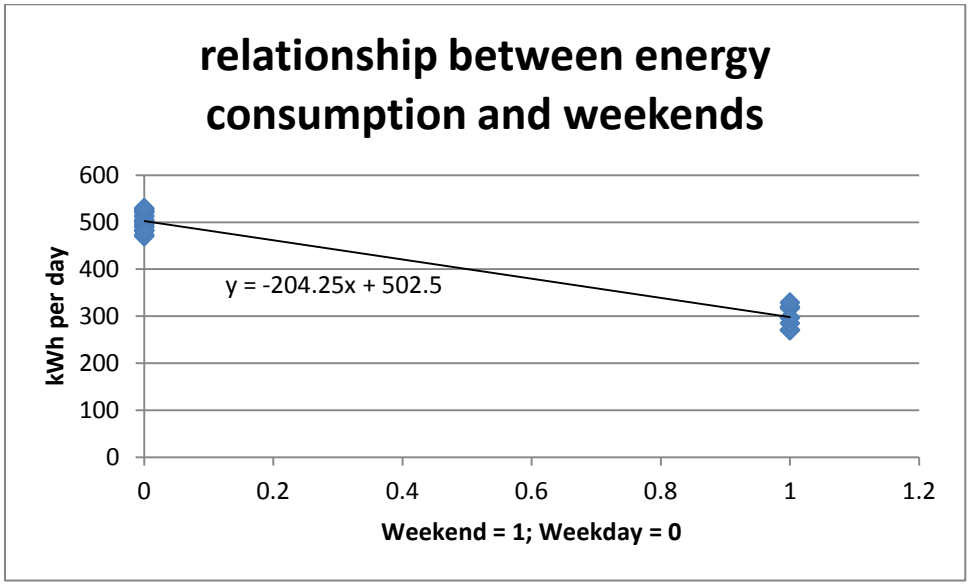
It should be noted that non linear trends can be difficult to identify and some thought is required to determine whether Energy Consumption should be linearly related to all values of the Predictor Variable. Trends may be more obvious if higher frequency data is used (e.g. weekly instead of monthly).

**Boolean Predictor Variables**

In some situations there is a significant correlation between Energy Consumption and a Predictor Variable that only has values of 1 or 0. An example of this is the identification of weekends in daily data, where weekend Energy Consumption is significantly different from weekdays.



The regression relationship between weekends and Energy Consumption is plotted below.




Where there is a consistent difference in Energy Consumption between the 0 and 1 Predictor Variable values, the calculated value of $R^2$ is likely to be very high. In this case $R^2$ = 0.96 i.e. 96% of the variation in Energy Consumption is due to the difference in Consumption between weekdays and weekends.

This high correlation between Energy Consumption and the Boolean Variable is likely to mask the impact of other Predictor Variables. When creating a Model it is therefore suggested that the relationship between Energy Consumption and other Predictor variables is analysed first during one Boolean state (e.g. either during school term time or during school holiday). In this way it should be possible to more accurately determine the impact of each Predictor Variable and also identify significant outliers in the data.

**Multiple Regression**

It is highly likely that the variation in Energy Consumption over time is caused by more than one Predictor Variable. Multiple regression can deal with any number of Predictor Variables and will increase the value of $R^2$ and increase the accuracy with which a model can be used to predict future Energy Consumption.

The "best fit" equation for multiple regression is of the form:

$$Y = a + bV1 + cV2 + ...$$

Where "V1,V2,..." are the Predictor Variables, "b,c,..." are the correlation coefficients and "a" is the regression constant.

Unfortunately multiple regression cannot be represented in a simple graphical form and so there needs to be a reliance on numerical analysis. It is important to verify that the correlation coefficients are within the range of expected values, that $R^2$ is appropriately high and the value of each residual is not unusually high.

It should be noted that if one or more of the Predictor Variables in a multiple regression are Boolean, then the value of $R^2$ is likely to be high and a greater emphasis is needed on analysis of the residuals to determine how accurately the model represents reality.

**Multicollinearity**

It is possible that two or more of the chosen Predictor Variables are strongly correlated. An example of this might be HDD and rainfall. If this happens, then a statistical phenomenon called multicollinearity occurs and needs to be understood. Multicollinearity does not affect the theoretical accuracy of a model, but it may cause the correlation coefficients to change dramatically when one data point is added or removed.

For example, the model:

Daily Energy Consumption = 200 + (30 * Daily Rainfall) - (100 * HDD)

may change to:

Daily Energy Consumption = 200 + (80 * Daily Rainfall) - (50 * HDD)

when one data point is added

If the pattern of rainfall and HDD in the Baseline Period is well matched by the pattern of rainfall and HDD in the Test Period then multicollinearity should not cause a problem. However, if the two patterns are significantly different, then the model is likely to be highly inaccurate.

It is important to detect strong correlation between Predictor Variables and if multicollinearity is likely to be a problem then the less significant Predictor Variable(s) should be excluded from the model.

**Predictor Variable Selection**

Any factor to which a number can be assigned could be considered as a Predictor Variable. Each potential Predictor Variable should be added to and removed from the Model to determine its impact. There are some factors which should only be used with caution – one example of this is indoor temperature:

- If indoor temperature is not a consequence of the Intervention and if the indoor temperature can be set by the User, then indoor temperature could be a useful Predictor Variable.
- If indoor temperature is not a consequence of the Intervention and if the indoor temperature cannot be set by the User, then it is not likely to be a useful Predictor Variable.
- If indoor temperature is a consequence is potentially a result of the intervention then it should not be used as a Predictor Variable.

Climatic changes are the main reason of variability in Energy Consumption profiles and heating degree days (HDD) or cooling degree days (CDD) are the recommended units because they tend to have a linear relationship with Energy Consumption.

There is no simple correlation between high frequency (e.g. hourly) Energy Consumption and high frequency temperature measurements because

(i) heating is not on at night and so there will be vast numbers of data points where temperature is low and Consumption is zero,

(ii) Consumption will generally be high first thing in the morning as the building heats up, irrespective of external temperature,

(iii) heating is not used at all when the external temperature is high