Internal Report 2012–16 August 2012

Universiteit Leiden
Opleiding Informatica

Extracting Information

from an

Energy Expenditure Dataset

Jan Kalmeijer

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Extracting Information from an Energy Expenditure Dataset

Jan Kalmeijer (0954721)

August 20, 2012

Abstract

Aiming for a sustainable future, Rijkswaterstaat is installing more and more smart energy meters. The data generated by these meters can be used to manage energy expenditure. Understanding how the energy expenditure of objects is related to external processes such as the weather, identifying which objects are showing different expenditure patterns from other objects with otherwise similar properties, or identifying outliers in the expenditure patterns of objects, is essential to facilitate management. To these ends KNMI weather data is coupled to the expenditure data, and both hierarchical clustering and a Self-Organizing Map are used to cluster the expenditure data. Although some information is gained in this process, finding a general solution to the problems turns out to be difficult.

1 Introduction

This bachelor thesis is the result of a collaboration between Rijkswaterstaat (RWS) and LIACS, Leiden University. RWS is responsible for managing a lot of objects; not only offices, but also floodgates, pumping stations etc. Since these objects consume a lot of energy and RWS focuses on durability, RWS is looking for ways to reduce this consumption. For this reason, RWS has monitored the quarter-hourly energy expenditure of their largest objects since 2006, and is expecting to be monitoring all objects in the near future. In this thesis only a subset of the data is used, and the expenditure data is coupled to meteorological data collected by the KNMI. The resulting data and the process of coupling the expenditure data to the meteorological data is described in Section 2.

Not counting the dataset description, this thesis consists of three main parts. Each part has an accompanying question. These questions are as follows:

1. How does the energy expenditure relate to the weather conditions?
2. How to identify objects that behave differently from other objects?

3. How to identify outliers in the expenditure pattern of an object?

We attempt to answer each question using a data mining-type approach, rather than a statistical one. Each question is encapsulated in its own section. We attempt to answer the first question in Section 3, using the sample Pearson correlation coefficient. Using the correlation coefficient we attempt to find out if similar objects (e.g., all offices), have similar correlations. We also try to find out how the energy expenditure of a specific region relates to the weather. As the weather effect might be delayed, we examine the correlations on the basis of both the hourly data and the daily averages of the hourly data. Figure 1 captures the process of this section.

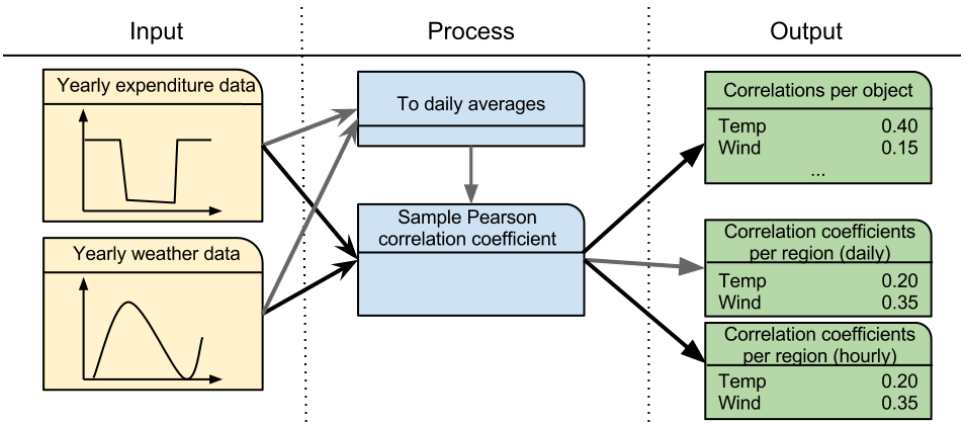


Figure 1: Process of Section 3.

In Section 4, we attempt to identify objects that behave differently from other objects. This is done by clustering the expenditure series using a Self-Organizing Map (SOM). The SOM is applied on both the expenditure data as-is, and a "bag-of-patterns" representation of the data. Ideally, the SOM would output large clusters of similar objects, and small clusters containing the objects that behave differently. Figure 3 captures the process of this section.

We attempt to answer the last question (How to identify outliers in the expenditure pattern of an object?) in Section 5. The expenditure pattern of each object is examined on a day-to-day basis, in order to find days that are different from the norm. This examination is done by an ad-hoc and a hierarchical clustering method. The ad-hoc method tries to find days on which public lighting objects have their lights on unnecessarily. The ad-hoc method is used as a heuristic for the parameter settings of the hierarchical clustering method. The outliers found by both methods are compared, and we try to explain them. The hierarchical clustering method is also used to find outliers in objects that are not classified as public lighting. The process is captured in Figure 2.

The thesis is concluded in Section 6 where we give a summary of the previous sections and discuss further research.

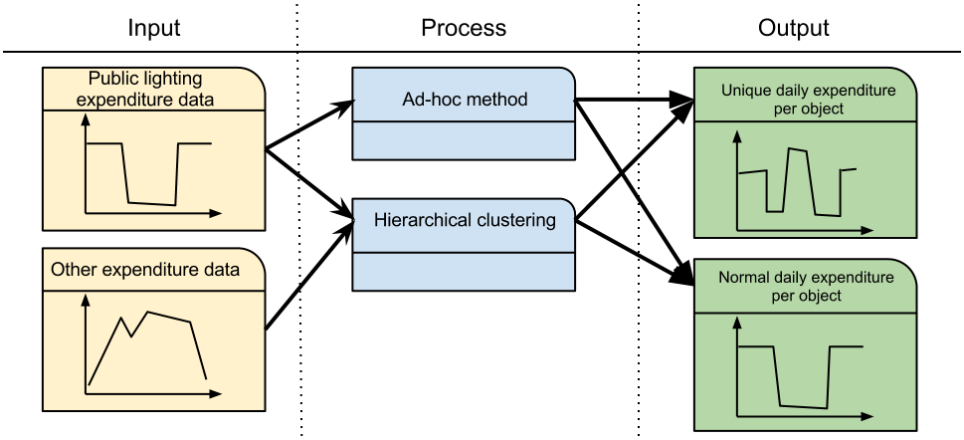


Figure 2: Process of Section 5.

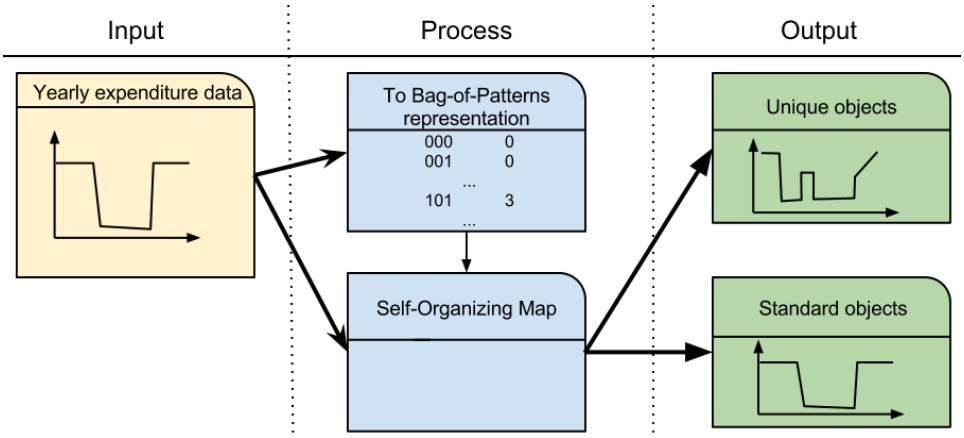


Figure 3: Process of Section 4.

2 The Dataset

During this project the following datasets are used: energy expenditure data of Rijkswaterstaat (RWS), meta information of the energy expenditure data, and meteorological data collected by the KNMI. First the energy expenditure set and meta information are discussed in Section 2.1, followed by the meteorological data in Section 2.2. Lastly we discuss how to combine these datasets for a more complete picture in Section 2.3.

2.1 Energy Expenditure Data

The dataset used in this paper is collected by RWS over the period 2008 up to and including 2010. It contains the quarter-hourly energy expenditure for ±200 objects. The set of objects is diverse; an *object* can range from office building, to tunnel, public lighting etc. Each object is uniquely identified by an 18 digits long string called the *EAN-code*. For reasons of confidentiality the EAN-code is replaced with a 4-digit identifier. However, the identifiers remain consistent throughout the thesis.

Table 1 reports on the properties of the dataset. The set of objects of which the energy expenditure is measured does not remain consistent throughout the years. More surprising is that the data contains some negative expenditure measurements. Negative values are caused by objects which generate energy themselves, and provide this to the energy supplier. This does not occur frequently. The dataset contains no missing values.

Property	2008	2009	2010
Total objects	194	209	209
Objects gained compared to previous year	N/A	32	18
Objects lost compared to previous year	N/A	17	18
negative-measurements	0.11%	0.21%	0.50%

Table 1: Properties of energy expenditure data over 2008–2010.

RWS also provides a classification for most of the objects. These classes are called *expenditure categories* or *categories* for short. See Table 2 for a list of all categories, along with their frequency of appearance in the year 2010. Within one category there can be quite some variation in the types of objects. An example is the category public lighting / traffic regulation installation. The objects in this category vary from only public lighting and only traffic regulation installations to a combination of the two.

In addition to the category, the GPS coordinates of the connection socket of each object are known. This position is used as an approximation of the actual position of the object. Section 2.3 about the enrichment of the data explains how this is done. In conclusion, on a yearly basis the data consists of the following attributes:

EAN-code	E_1	E_2	...	E_{35040}[1]	GPS Coordinates	Category
...6729	12.0	12.0	...	12.0	(51.87, 4.55)	Office

Here E_x depicts a quarter-hourly energy expenditure measurement in *kWh*.

[1]The total measurements per object is $4 \times 24 \times 365$ (= 35040) for the years 2009 and 2010, and 35136 for the leap year 2008.

2.2 Meteorological Data

The KNMI provides compilations of hourly measurements of 35 automatic weather stations on their website [1]. In total there are 20 features measured, ranging from the average temperature over the past hour, to Boolean values indicating whether it snowed at some point during the past hour. See Table 4 for a list of all features and their abbreviations. Not every station measures each feature. As a result there are missing values. There are even features for which during some hour not a single station measured a value. See Table 3 for a list of how often this occurs in the year 2010. In addition to the weather related measurements, the GPS coordinates of each of the 35 stations are also provided.

2.3 Enrichment

To examine the influence of weather conditions on the energy expenditure of an object, a coupling between the meteorological and expenditure data is created. The coupling gives us an approximation under which weather conditions an expenditure measurement occurred. The coupling is achieved

Expenditure categories	Frequency
Public lighting / traffic regulation installation	78
Office	33
Small building	3
Bridge / dam	8
Pumping station	11
Floodgate / weir	40
Tunnel	20
Traffic control center	9
Radar post	2
Measurement station	0
Lighthouse	0
Unknown	5
Total	209

Table 2: Frequency of each category for the year 2010.

Feature	Times missing
WW	7.78%
T10N	47.62%

Table 3: Percentages of the time no station measured the listed features. The percentage is 0 for any omitted features.

Abbreviation	Description
DD	Wind direction (degrees) averaged over the last 10 minutes of the past hour (360=north, 90=east, 180=south, 270=west, 0=calm, 990=variable)
FH	Average wind speed (in 0.1 m/s)
FF	Wind speed (in 0.1 m/s) averaged over the last 10 minutes of the past hour
FX	Strongest squall (in 0.1 m/s) over the past hour
T	Temperature (in 0.1 °C) at 1.50 m altitude
T10N	Minimum temperature (in 0.1 °C) at 10 cm altitude over the past 6 hours
TD	Dewpoint temperature (in 0.1 °C) at 1.50 m altitude
SQ	Duration of sunshine (in 0.1 hours), calculated on basis of global radiation (-1 for < 0.05 hour)
Q	Global radiation (in J/cm^2)
DR	Duration of rainfall (in 0.1 hours)
RH	Amount of rainfall (in 0.1 mm) (-1 if < 0.05 mm)
P	Air pressure (in 0.1 hPa) transformed to sea level
VV	Horizontal sight (0=less than 100m, 1=100–200 m, 2=200–300 m, ..., 49=4900–5000 m, 50=5–6 km, 56=6–7 km, 57=7–8 km, ..., 79=29–30 km, 80=30–35 km, 81=35–40, ..., 89 $\geq$ 70 km)
WW	Weather code (00–99)
N	Cloudiness (coverage of sky in eights, 9=sky imperceptible)
U	Relative humidity (in percentage) at 1.50 m altitude
M	Mist; 0=no occurrence, 1=occurrence
R	Rain; 0=no occurrence, 1=occurrence
S	Snow; 0=no occurrence, 1=occurrence
O	Lightning; 0=no occurrence, 1=occurrence
Y	Icing; 0=no occurrence, 1=occurrence

Table 4: The abbreviations and descriptions of features measured by the KNMI.

by first converting the expenditure data to hourly values. Since the expenditure data is measured in *CET* and the meteorological data in *UT*, the first measurements of the meteorological data of a specific year, corresponds to the second expenditure measurement of that year.

As there are 35 stations, there can be multiple measurements for the same feature at a given time. Since the GPS coordinates of all stations and all objects are known, the distance between an object and a station can be calculated using the haversine formula [6]. Each of the energy expenditure measurements of an object is then coupled to most nearby measurement of each feature. The result is the best approximation of the weather conditions under which each energy expenditure measurement occurred.

3 Influence of External Factors on Energy Expenditure

In this section, we study how the energy expenditure of the objects described in Section 2 is related to the weather. Correlation coefficients are used to examine this relationship. The relationship between the weather and expenditure is studied on the object level, as well as on the regional level. On object level we focus on the relation between weather and the expenditure of a single object. The relation on regional level suggest how the expenditure of a set of objects is affected by the weather, *possibly* indicating whether RWS is meeting its goals related to reducing energy expenditure.

3.1 Pearson Correlation Coefficient

Correlation coefficients give an indication of the strength of dependency between two variables. The correlation coefficient used in this thesis is called the *sample Pearson correlation coefficient*. This coefficient captures the linear dependency between two stochastic variables. The Pearson correlation coefficient, denoted as r, is calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Here X and Y are two stochastic variables and $\overline{X}$ is the mean of X.

The value of r is in the interval $[-1, +1]$. A value of -1 indicates a perfect decreasing linear relationship and a value of $+1$ a perfect increasing relationship. Values close-by 0 indicate there is no *linear* relationship between the variables. However, a value near 0 does not indicate the absence of a relationship, as the relationship can be non-linear. Interpretation of other values is more complex. There are different guidelines to estimate which value ranges indicate weak, moderate or strong correlation. In this thesis we use the guidelines in Table 5.

Correlation	Positive	Negative
Weak	0.0 to 0.3	−0.3 to 0.0
Moderate	0.3 to 0.7	−0.7 to −0.3
Strong	0.7 to 1.0	−1.0 to −0.7

Table 5: Guidelines for interpreting correlation coefficients.

It should be noted that a large (in absolute values) correlation coefficient does not indicate a casual relationship, it only shows two variables are strictly connected. The largeness of the correlation coefficient also does not imply a steep regression line.

3.2 Individual Correlation

In Section 2.3 we described a process to extract the weather conditions under which an energy expenditure measurement occurred. In other words, for some expenditure measurement of some object, most of the weather variables listed in Table 4 are known. The only variables that are not always known are WW and $T10N$, hence we leave them out. For all the other variables the correlations between the variable and the expenditure can be calculated.

Listing all correlation coefficients for every object and variable leads to a long and not fully informative list. Instead we combine the correlation coefficients of objects, and group them together on the basis of their expenditure category. The minimum, maximum and average correlation are shown in Table 6 for those groups where any of the objects in the group has a strong correlation.

Table 6 contains some peculiarities. Four out of six groups show a large difference between their minimum and maximum correlation. For these groups the objects corresponding to the minimum and maximum correlation were retrieved and their expenditure pattern compared to the overall pattern of the group.

The traffic control center group shows significant deviation between the minimum and maximum correlation. Both expenditure patterns do not seem much different from the average, so we are unsure what other category they might fit in.

In the floodgate/weir group the object corresponding to the minimum temperature T is not much different from the average pattern. The expenditure of pattern the object associated with the maximum looks more like the pattern of a tunnel or office, although its expenditure does not show the same "smoothness" the expenditure patterns of those groups typically show. However, the name of the object in the database clearly states that the object is a floodgate.

Bridge/dam

	min	max	average
T	−.8670	−.4007	−.6361
TD	−.8287	−.3440	−.5968

Traffic control center

	min	max	average
T	−.4228	.7671	.2516

Tunnel

	min	max	average
T	−.7660	.5144	−.0607
TD	−.7531	.4671	−.1582
SQ	−.2492	.7551	.3079
Q	−.2912	.8393	.3338

Radar post

	min	max	average
T	−.8386	−.5818	−.7356
TD	−.8259	−.5651	−.7201

Floodgate/weir

	min	max	average
T	−.8963	.4392	−.5339
TD	−.8350	.4013	−.4834

Public lighting/ traffic regulation

	min	max	average
T	−.8236	.5344	−.3508
TD	−.8461	.2757	−.2439
SQ	−.6022	.7391	−.3967
Q	−.6714	.8373	−.4246

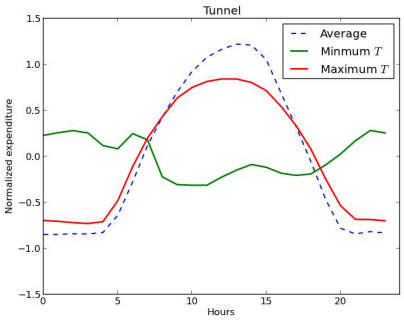
Table 6: Compilation of all strong correlations for each category. Categories omitted did not have any strong correlations.

In the tunnel group the expenditure pattern of the object associated with the minimum T looks more like the pattern of a floodgate/weir, pumping station or bridge/dam. This same object is associated with the minima for TD, SQ and Q. This object is a large tunnel, but this connection belongs to the version for cyclists, which uses different illumination from tunnels for motorized vehicles. The objects corresponding to the maxima for each variable are different, but both their expenditure patterns look like the pattern of a tunnel (see Figure 4a). The high correlations are expected, as a tunnel must produce more light when it is bright outside of the tunnel, to ease the transition when leaving or entering the tunnel.

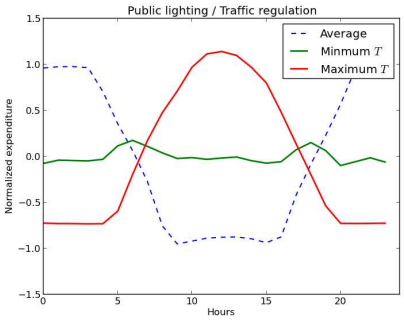
The public lighting/traffic regulation installation group is a tricky one. For the variable T, both the minimum and maximum object seem to correspond to different categories. The objects associated with the maximum looks more like a tunnel or office. The category of the object associated with the minimum is more unclear. The object associated with the minimum values for SQ and Q is public lighting. The object associated with the maximum value is the same object as associated with the maximum value for T. See also Figure 4b.

3.3 Regional Correlation

RWS manages their objects on regional level. Each of these regions is the size of a Dutch province. However, not every region has the same amount of



(a) Average daily expenditures of the tunnel group.



(b) Average daily expenditures of the public lighting/traffic regulation group.

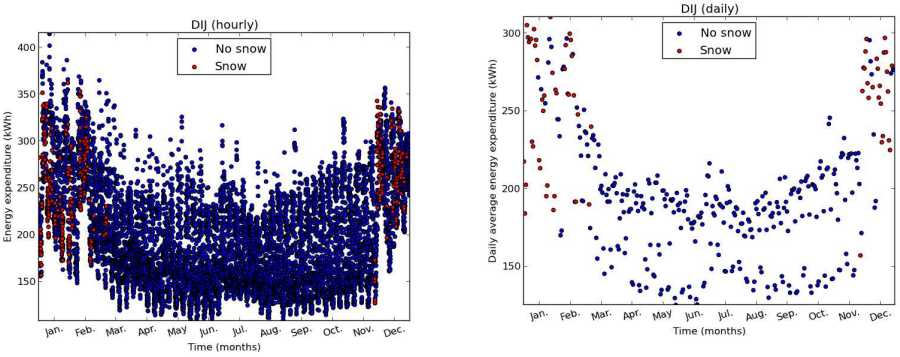
Figure 4: Expenditures of the objects associated with the extreme values of the correlations.

objects. Table 8 in Appendix A lists how many objects are present in each region, and to which expenditure categories the objects in the region belong. For each of these regions we want an indication of the effect of the weather on the expenditure within that region. We again use the sample Pearson correlation coefficient. To construct a coupling between the expenditure measurements in a region and the weather measurements, we do something similar to Section 2.3; the total expenditure of the objects in a region is coupled with the measurements available from the station for which the distance between the station and all objects in the region is minimal.

3.3.1 Daily Versus Hourly Effects

For some weather effects such as global radiation, we expect immediate reactions. The sensor of an object immediately picks up on the global radiation and changes the lighting. Other weather effects, such as temperature, take some time to influence an object's expenditure. In addition to this there is the strong daily pattern that each object shows. These factors might mask the relation between the weather and expenditure. Therefore it makes sense to calculate the correlations between both the hourly values and daily averages of the energy expenditure and weather effects.

The correlation coefficients for each region are shown in Appendix B. The hourly and daily average correlation coefficients are shown in Table 9 and Table 10 respectively. The daily average coefficients are typically larger than the hourly coefficients. This could be due to the daily pattern masking the correlation, or the delayed effect. Another possibility is that a specific event (such as snow) typically only occurs during seasons for which the average expenditure is larger anyway. Figure 5 gives an example of this.



(a) Hourly correlation coefficient: 0.197 **(b)** Daily average correlation coefficient: 0.426

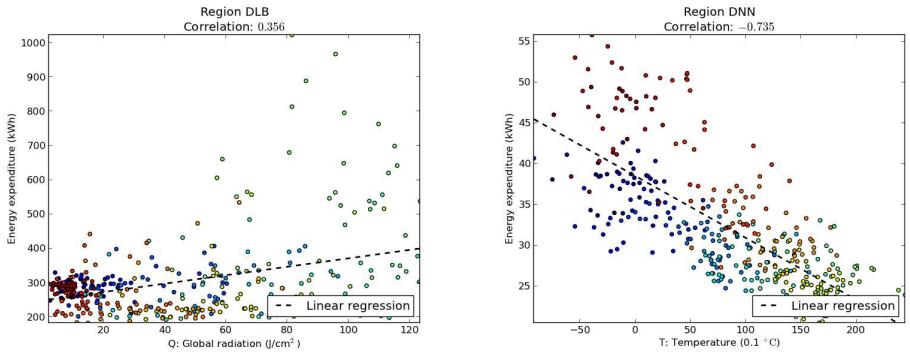
Figure 5: Expenditure in the region DIJ during the year 2010. Measurements which occurred during some form of snow are marked red.

3.4 Conclusions

In Section 3.2 we grouped the objects together on basis of their expenditure category, and inspected the objects with the strongest correlations within their group. Within a category, there are very different relationships between the expenditure and certain weather variables. However, anomaly detection within a group on basis of the correlation coefficients seems ineffective, as objects that showed strange correlation coefficients did in fact belong to that group.

In Section 3.3 we examined the correlation coefficients of each region on the basis of the hourly data and the daily averages thereof. The daily correlations are overall stronger than the hourly correlations, possibly because the daily expenditure pattern masks the effect of the weather on the expenditure, or because the effect the weather has on the expenditure is delayed. The expenditure per region showed quite a few moderate correlations with temperature, global radiation, humidity, snowfall, icing and cloudiness. The daily correlation coefficients even show two strong correlations between the temperature and the energy expenditure of the regions DNN and DZH.

Please note that the correlation coefficients give an indication of how strictly two variables are coupled. It gives no indication of causality. A moderate correlation might seem impressive, but when shown in a scatter plot (Figure 6a) the spread is quite large. This is even true for strong correlations (Figure 6b).



(a) moderate correlation **(b)** strong correlation

Figure 6: Example of the spread at a strong and moderate correlation coefficient.

4 Clustering of the Energy Expenditure Series

There are multiple approaches to time series clustering. In [3] the authors group the approaches into three major categories. The first category operates directly on the raw data, the second on features extracted from raw data and the third on models built from the raw data. In this thesis the focus is on the first two. By raw data the normalized standard scores of the data is meant, as this allows comparison between expenditure series with different amplitudes.

The clustering algorithm used is a Self-Organizing Map. It is briefly described in Section 4.1. The distance measure used is Euclidean distance. The feature based extraction method converts the data to a "bag of patterns" representation, presented in [5]. A summary of the method, and how it is applied in this thesis is found in Section 4.2.

Clustering of the expenditure series is important. It allows for the detection of outliers and detecting groups of objects which behave similarly. However, rating the quality of the found clustering is difficult. The dataset does not have labels for specific expenditure patterns, but there are labels for specific categories of objects (e.g., Office). Under the premise that objects in similar categories behave alike, this allows us to rate the quality of the clustering. Since this premise is not necessarily true and the labeling is imperfect, manual inspection of visualizations of the found clusters is also used to rate the approaches.

Both the Self-Organizing Map and Bag-of-Patterns approach have some parameters to set. Multiple combinations are compared on quality.

4.1 Self-Organizing Maps (SOM)

A Self-Organizing Map (SOM) [2] is a clustering and data visualization technique. The goal is to find centroids (also called *reference vectors* or *neurons*) and to assign each object in the dataset to the centroid that is the best approximation of that object.

The most original aspect of SOM is that it imposes a topographic organization on the centroids. An example of such an organization is a lattice. In this case each centroid is assigned a pair of coordinates (i, j). The amount of centroids is fixed, and the size of the topology depends on the amount of centroids the SOM has. The topology determines how centroids influence each other. Two centroids influence each other more, the closer together they are in the lattice. As a result two centroids which are closer together in the lattice are also more related to each other than to centroids that are farther away.

Clustering using a SOM can be described in a series of steps. First the centroids are initialized. Centroid components can be chosen at random from the value ranges observed in the data, or the initial centroids are set to randomly selected examples from the dataset.

After initialization the algorithm enters a loop which terminates when the SOM has converged. First the algorithm selects an object from the dataset. This object is assigned to the centroid it is most alike, based on some distance metric; Euclidean distance for example. The centroid that is most like this object is also referred to as the *best matching unit* or *BMU*.

Once the BMU for the current object is found, the algorithm executes the update rule. This rule is the most complicated, so it is first described informally. The BMU should become more like the current object. The neighbors of the BMU should also become more like the current object, but the effect of change should be diminished. It is much like a person learning something new. The person shares his newly acquired knowledge with family and friends (his neighborhood), but they will not understand it as well as the person does. SOM uses some additional parameters to enforce convergence; a neighborhood size and a learning rate. These parameters decrease over time. To keep in line with the previous analogy, both a person's ability to learn and his will to explain it to his entire neighborhood diminish over time.

Let $c_1, \ldots, c_k$ be the k centroids. For time step t, let $p(t)$ be the current object and assume the BMU of $p(t)$ is $c_j(t)$; further more, $c_i(t)$ is a centroid in the neighborhood of $c_j(t)$. For each centroid $c_i(t)$ in the neighborhood of $c_j(t)$ (this includes $c_j(t)$), we apply the following rule:

$$c_i(t+1) = c_i(t) + h_{i,j}(t) \times (p(t) - c_i(t))$$

A possible definition of $h_{i,j}(t)$ is as follows:

$$h_{i,j}(t) = \alpha(t) exp\left(\frac{-dist(r_j, r_i)^2}{2\sigma^2(t)}\right)$$

Here $\alpha(t)$ is a learning rate parameter, $0 < \alpha(t) < 1$, which decreases with time and controls the rate of convergence; $dist(r_j, r_i)$ is the Euclidean distance between the grid location of the two centroids c_j and c_i; $\sigma(t)$ is a parameter indicating the neighborhood size, which also decreases with time. In this example the update rule uses a Gaussian function to define neighborhood influence, but others functions may be used.

The strength of SOMs is enforcing neighborhood relationships on the resulting cluster centroids. As neighboring clusters are more closely related, this facilitates the interpretation and visualization of the results. One of the weaknesses is that the user must define a lot of parameters. Section 4.3 describes how the parameters are set during the experiments.

4.2 Feature Extraction Using Bag-of-Patterns

The Bag-of-Patterns (BOP) representation is introduced by Lin and Li in [5]. The concept is to split the time series into subsequences using a sliding window. Each subsequence is then converted into a word using SAX (Symbolic Aggregate approXimation) [4]. This leaves us with a multiset of words. As each word represents a pattern in the time series, the resulting set is called a bag-of-words. As the result is a set, the ordering between the patterns is lost.

To convert a time series to a Bag-of-Patterns, three parameters must be defined, the first being the sliding window. The other two parameters are related to SAX. Within the sliding window, the order of measurements is preserved. As the pattern of the daily expenditure of an object is important, the sliding window is set to one day. In the original algorithm, the window slides one measurement at the time, i.e., every 15 minutes. This is unfit for our data, as the day-to-day alignment is very important. Figure 7 gives an example of clusters formed when the original sliding window is used. As this result is undesirable a jumping window is used instead. Using the jumping window, each day is converted to a single word by the SAX algorithm. The resulting bag-of-words contains 365 (or 366) words when applied to a year long series.

SAX is used to convert each sequence into a word. SAX uses two parameters: α (the size of the alphabet) and w (the size of the words produced). Hence the amount of possible SAX words is α^w. SAX splits a subsequence into w segments of equal length. Each segment has a value, which is the mean of the values in the segment. To construct a word, the segment values are converted to symbols using a breakpoint table. These breakpoints indicate what range of values map to which symbol. There are $\alpha - 1$ breakpoints, and the breakpoints are defined such that all regions have approximately equal probability to be selected, based on a Gaussian distribution. Figure 8 summarizes the process.

The above explains how the Bag-of-Patterns in obtained, but it still has

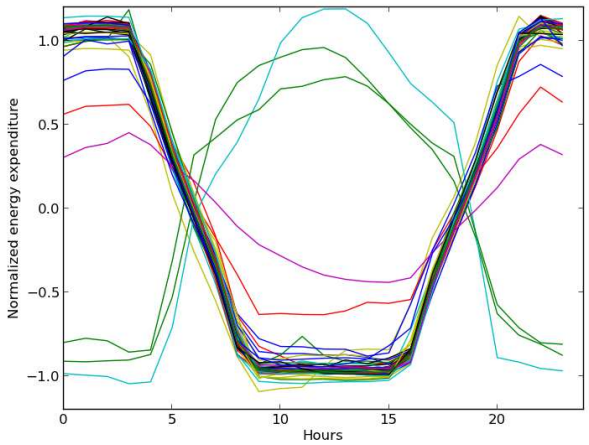


Figure 7: Example of how the sliding window used in BOP is unfit for clustering time series where alignment is important. The incorrectly clustered green and cyan colored expenditure patterns have the same period as the other patterns, but a different alignment.

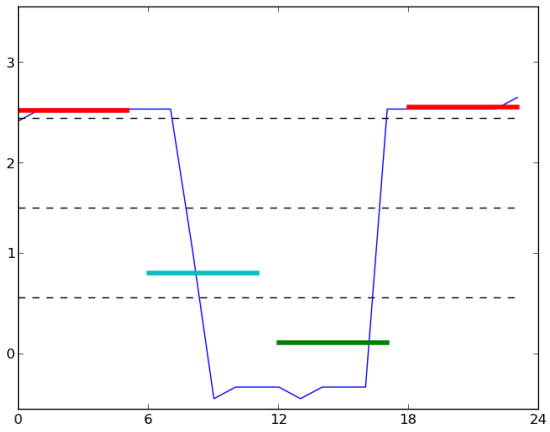


Figure 8: Example of how SAX converts an expenditure pattern with $\alpha = 4$ and $w = 4$. The dashed lines indicate the breakpoints. The resulting word is 3103.

to be explained how it can be used to determine the similarity between two series. Rather than viewing the Bag-of-Patterns as a set, we can represent it as an array A which is of length α^w. The i^{th} element of A then is the frequency of the i^{th} word. For two Bag-of-Patterns, the Euclidean distance between the frequencies of each word can then be used as distance measure. It is worth noting that this representation is not the most efficient, as for repetitive patterns most elements will be 0.

4.3 Experiments

In this section we combine the presented methods and compare their effectiveness. We compare the precision of multiple configurations. For the Self-Organizing Map we present the configuration, and we examine how a different amount and organization of the centroids affect the quality of the clustering. We also examine how the SAX word length w and alphabet size α influence the quality of the found clusters.

The configuration of the SOM used during these experiments is as follows: The reference vectors are initialized randomly and termination occurs when a maximum amount of iterations t_{max} is performed. The initial learning rate, $\alpha(0)$, is set to 0.1. It decreases according to the following scheme: $\alpha(t+1) = \alpha(0) - \psi \times t^2$, where ψ is a constant such that $\alpha(t_{max}) \approx 0$.

The centroids are arranged in a rectangular grid. The neighborhood type used is the Von Neumann neighborhood. The neighborhood of a cell in a Von Neumann neighborhood consist of all cells for which the Manhattan distance to the original cell is less than or equal to the neighborhood size. The Von Neumann neighborhood is depicted in Figure 9. The initial neighborhood size $\sigma(0)$ is set to the Euclidean distance from the center of the grid to one of the corners; i.e. initially the centroid in the center affects all other centroids. It is then decreased by time such that in the end, centroid updates only change the centroid in question. The update rule uses the Gaussian function described in Section 4.1. What remains are the amount and organization of the centroids.

We assume that objects in the same expenditure category (see Table 2) emit similar expenditure behavior. In total there are 12 possible categories, of which there are 10 present in the data of the year 2010. We focus primarily on configurations with 12 centroids, to allow some centroids for outliers. However, we also explore configurations with many centroids to see how this affects the resulting clusters. For each configuration the precision is reported in Table 7. The precision is calculated as follows:

$$precision = \frac{tp}{tp + fp}$$

Here an object is counted as a true positive (tp) if the majority of the examples corresponding to its reference vector have the same class. When this is not the case, the object is a false positive (fp).

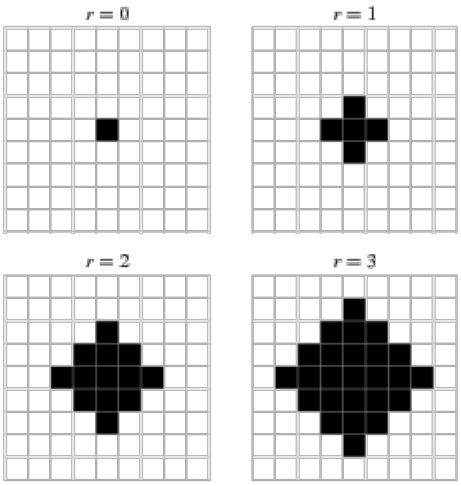


Figure 9: Von Neumann neighborhood for different neighborhood sizes r. Taken from [8].

	Precision	
Topology	Bag-of-Patterns	Raw
1×12	.5749	.6039
2×6	.5217	.6232
3×4	.5314	.5749
1×50	.6329	.7256
2×50	.7150	.7826
3×50	.7633	.8019
4×50	.8406	.8599
8×25	.8164	.8599

Table 7: Precisions corresponding to different SOM configurations. The separator between 3×4 and 1×50 indicates when we start overfitting. The Bag-of-Patterns was generated using $w = 3$ and $\alpha = 3$.

The clustering of raw data is compared to the Bag-of-Patterns in Table 7. For the Bag-of-Patterns approach we chose $w = 3$ and $\alpha = 3$, on the basis of the values in Figure 10. The authors of [5] found empirical evidence that changing α does not critically impact performance. In this case, it seems the same holds for w except for very small values.

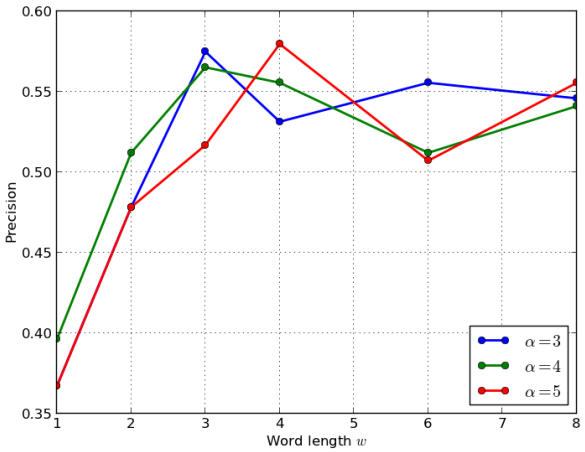


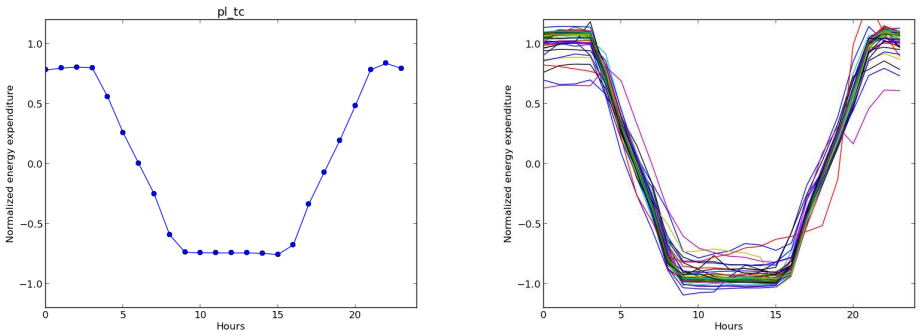
Figure 10: Precision when clustering the Bag-of-Patterns representation generated using different α and w settings. The topology of the SOM is 1×12.

The precision does not tell the whole story. Consider Figure 11 for example. The reference vector shows a clear public lighting pattern, and so do the expenditures of the objects in the cluster. The precision of this cluster is .9608; it contains 49 objects classified as public lighting objects, 1 office and 1 bridge/dam. The average expenditure of public lighting objects and offices are almost opposite, making it likely this object classified as office is in fact public lighting. The same cannot be said for bridge/dam, as the average pattern of that category is very much like those for public lighting.

The above error is fortunately only very small. Less fortunate is that not all clusters are as well-defined as the one shown previously. Figure 12 shows a poor quality cluster and the corresponding reference vector. The cluster contains 23 examples from 7 of the different categories.

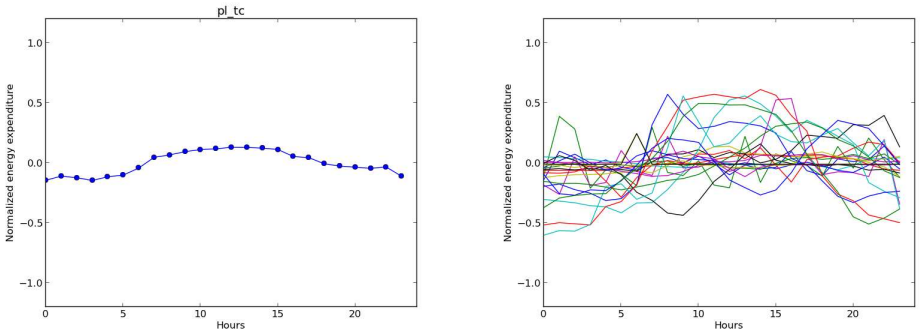
4.3.1 The Effect of Overfitting

The topologies with more than 50 centroids certainly allow the SOM to overfit. The overfitting increases the precision, but surprisingly the precision of 1.0 can not be reached; not even with 200 centroids. This is strange as the dataset contains only 207 objects. In fact, 96 out of 200 centroids do not have an object associated with them.



(a) Reference vector **(b)** Cluster

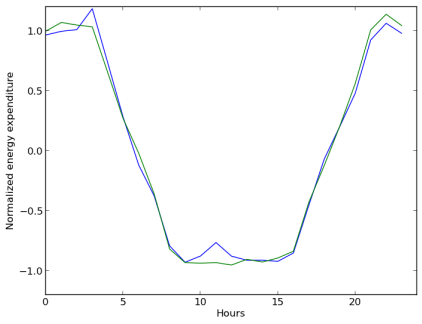
Figure 11: Example of a desirable cluster (precision .9608) found by applying a SOM with a 3×4 topology on the raw data. The expenditures plotted are average daily expenditures.



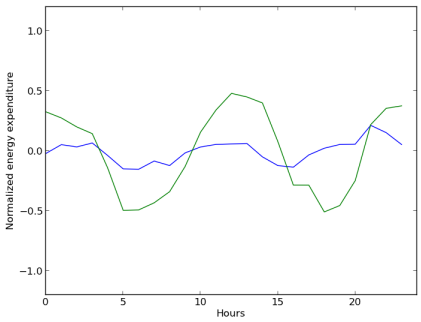
(a) Reference vector **(b)** Cluster

Figure 12: Example of a cluster containing pattern that are significantly different (precision .2174). Found by applying SOM with a 3×4 topology on the raw data. The expenditures plotted are the average daily expenditures.

Most patterns that are clustered together look similar. However, some of these clusters do not score a 100% precision. An example of such a cluster is depicted in Figure 13a. However, precision loss is also caused by strange clustering choices. Figure 13b shows an example of such a cluster. Despite having the freedom to overfit, the algorithm still places them together.



(a) A neatly matched cluster, with only 50% precision according to the categories.



(b) A poorly matched cluster.

Figure 13: Examples of clusters formed when the SOM is allowed to overfit. All examples come from the 25×8 topology.

4.4 Conclusions

In this section we took a look at two techniques to cluster time series using a Self-Organizing Map (SOM); clustering the raw data and clustering the modified Bag-of-Patterns representation of the data. Clustering the raw data is more focussed on local similarity, while the Bag-of-Patterns approach tries to cluster based on some features of the data. The algorithm to convert to the Bag-of-Patterns representation used in this thesis is a modified version of the one in [5], as the original algorithm did not respect alignment of the time series. The modified algorithm respects alignment with regards to days, which turned out to be very important. However, information about seasonal alignment is lost. This information could be important, as the Bag-of-Patterns representation is (slightly) outperformed by the raw data.

The clusters generated by both methods were assigned a precision on the basis of labels given to the objects by RWS. These labels are not perfect, as for some clusters the precision was not a 100%, despite the patterns in the cluster looking very similar. However, the labels still give reasonable insight in the quality of the clusters, as even with the liberty of overfitting, the SOM still made some odd choices, causing further reduction of the precision.

It turned out to be very hard to properly cluster the expenditure series. We were not able to find objects that behaved significantly different from other objects, as the quality of the found clusters is too low. We did find

out that different SOM topologies had a significant effect on the quality of the clusters. We also found out that, at least in this case, overfitting with a SOM is hard.

5 Anomaly Detection within Time Series

In this section two methods for detecting anomalies within a time series are discussed. RWS frequently receives the question why the public lighting is on during daytime, hence the first method is an ad-hoc method for detecting when public lights are turned on during daytime. This method is discussed in Section 5.2. We give some insight on the severity and causes of the found outliers in Section 5.2.1.

The second method, which is described in Section 5.3, is hierarchical clustering. It has been successfully applied on similar data in [7], although not specifically for outlier detection. Both methods are applied on the same dataset, which is a subset of the dataset described in Section 2. How this subset is obtained is described in Section 5.1. The result of both methods are compared in Section 5.4. We also show some of the outliers of other classes, as found by the hierarchical clustering method in Section 5.5.

5.1 The Dataset Sample

The detection methods are applied to a subset of the complete dataset described in Section 2. This sample is not drawn at random. In Section 4 all the expenditure series are clustered. The result is a cluster of decent quality for public lighting objects (see Figure 11). The expenditure series of the objects in this cluster make up the dataset used in this section. The total amount of objects in the subset is 51.

5.2 Ad-hoc "Lights On" Detection Method

Let us take a look at the expenditure in Figure 14. During daytime the lights are off, and the expenditure is somewhere around 0.2. During nighttime the expenditure peaks at about 0.9. Based on this sample of a single day, one could say the base expenditure B of this object is 0.2, and if L is the expenditure when the lights are on, L is 0.9. If the expenditure of the object is around 0.4 during daytime the next day, we could suspect at least part of the lights are on.

Obviously B and L can not be derived from a single day. To get an accurate representation of L, L is defined as the average of the 10% largest expenditure measurements throughout the year. The average is used to correct for measurement errors. We could define B similar to L, except for the smallest values. Unfortunately this leads to problems when the data contains incorrect 0-measurements. For example, if the data contains two

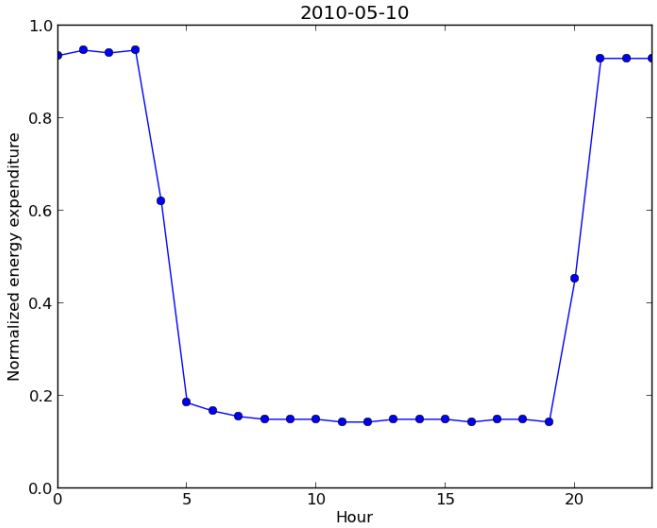


Figure 14: Expenditure over a single day for a public lighting object. The expenditure is normalized by dividing each measurement by the maximum measurement of the year 2010.

weeks ($\approx$ 4% of all measurements) of false 0 measurements, B is already significantly smaller than it should be, as this 4% makes up 40% of the 10% used to calculate B. Instead of all measurements, only the minimum measurement of each day is considered. The base expenditure B then is the average of the 10% smallest of the minimums. Figure 15 gives an example of how this is a better approximation of B.

With L and B defined as above, $L - B$ is the expenditure increase caused by having all the lights on. We suspect that if during some hour during daytime the expenditure exceeds $B + (L - B) \times 0.25$, a significant portion (i.e., at least 25%) of the lights are on. This gives a threshold T ($= B + (L - B) \times 0.25$) that if surpassed during daytime, indicates the lights are on.

But what exactly is daytime? We define it as the range of hours the public lighting should be turned off. Daytime shifts throughout the seasons; it is normal for the lights to be on until 9 AM in the winter, while this is unlikely in the summer. However, it seems reasonable to say that for all seasons, the lights should typically be off during hours in the range 10–15. Hence daytime consists of all hours in the range 10–15. Figure 16 depicts the strategy of the "lights on" detection method. In Appendix C the results of applying this method on the described dataset are listed. The method finds 1014 instances where the expenditure pattern of an object passes the threshold on some day. This is a significant amount, hence the next section is about the severity of these found outliers and what could have caused them to occur.

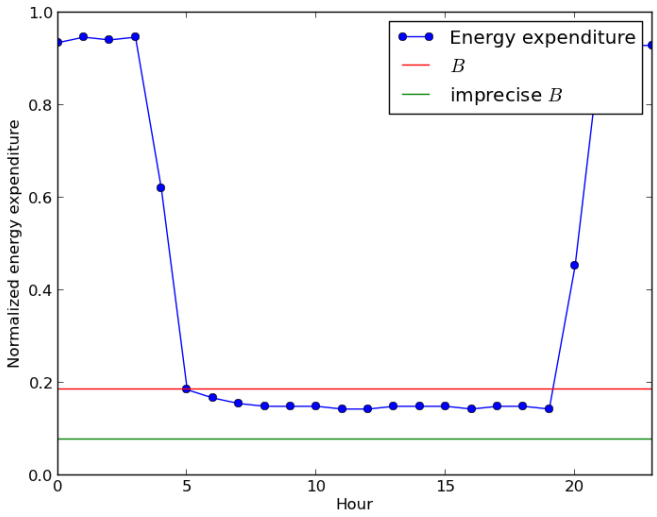


Figure 15: Different methods for calculating B compared. The energy expenditure is from a single day.

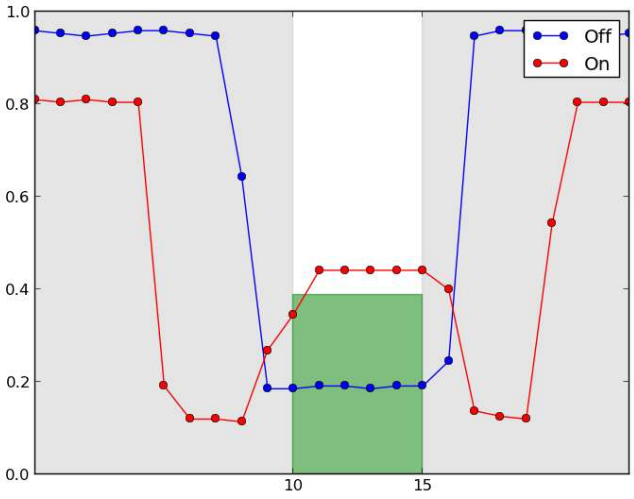


Figure 16: Two days of expenditure measurements for the same object. Any measurements within the white box are considered to be caused by having the lights on.

5.2.1 Severity and Causes of the Found Outliers

When the expenditure passes the threshold in the range 10–15, this is not necessarily bad. There can be cases where the light sensors sense there is little ambient light, hence the public lights are turned on in the 10–15 range. How frequently this occurs is examined empirically. Using the enrichment method described in Section 2.3, each hourly expenditure measurement is coupled to a measurement of the global radiation. This global radiation is plotted against the time, and the measurements made when the threshold was passed are marked.

The outliers for some objects seem related to time, rather than to the global radiation. These outliers occur in a specific period in time. This might be due to construction work at the road where the public lighting is placed. The construction equipment is then connected to the same connection socket at the public lighting. Two examples of this are given in Figure 17.

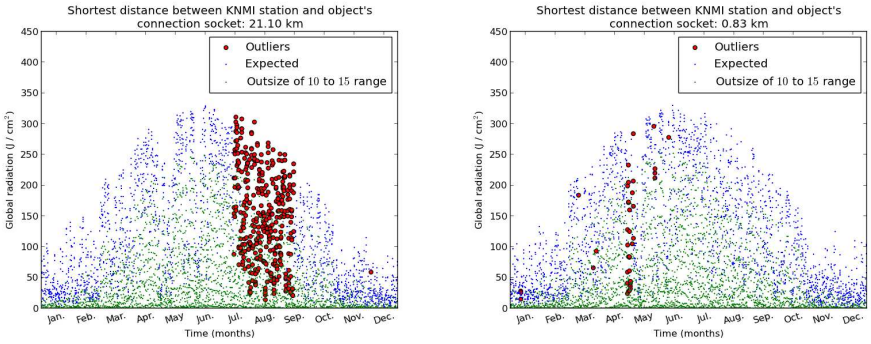


Figure 17: Example of outliers which occur in close proximity in time. Possibly due to construction work.

Outliers of other objects could be related to both time and global radiation. In these cases, the majority of the outliers occur during the darker seasons. This could have multiple explanations, e.g., the object is actually an aggregate object of both an office and a public lighting object. In this case, more energy could be spend on heating the office. The construction argument might also apply. Lack of global radiation does not seem to be the cause, as similar radiation values occur during the brighter seasons, yet do not consistently lead to an outlier. Figure 18 gives an example of these cases.

There are also cases where the outliers occur only during low amount of global radiation (see Figure 19a), and cases where the outliers seem completely unstructured (Figure 19b). Except for this last group, most outliers seem explainable. Despite many detected outliers, if the explanations are correct, not many of them are undesired.

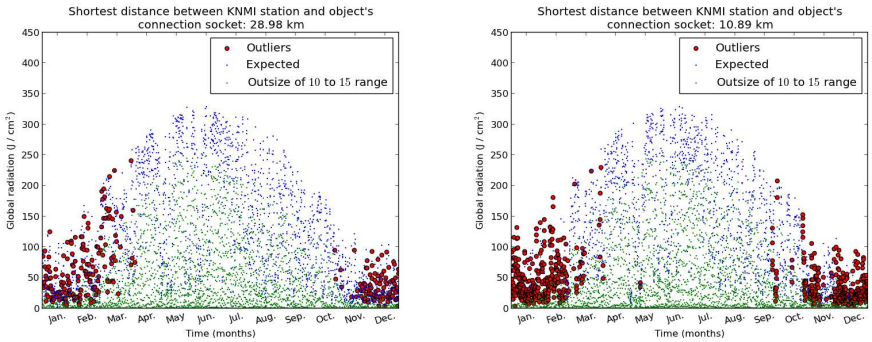
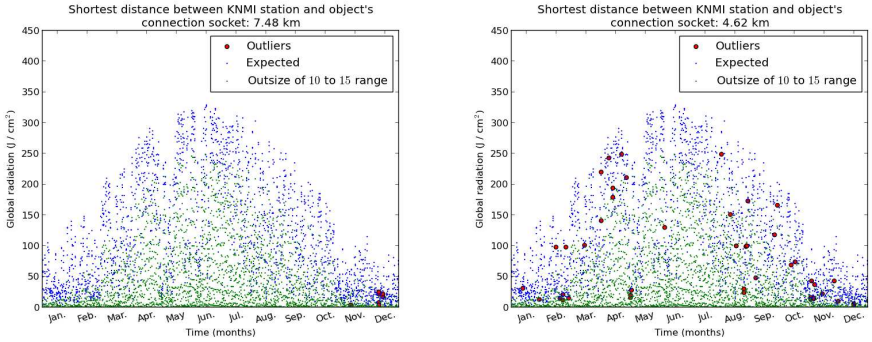


Figure 18: Examples of outliers which occur mainly during the darker seasons.



(a) Example where outliers seem related to the global radiation.

(b) Example where the outliers do not seem related to the global radiation.

Figure 19: Two examples of different behaviour to global radiation.

5.3 Hierarchical Clustering Method

The expenditure patterns of the objects are strongly repetitive when viewed on a daily basis. This is not surprising as the objects are used by humans, who have a strong daily routine. It makes sense to view the expenditure pattern as a series of days. These days can then be clustered together, to see what patterns emerge. As a hierarchical clustering method has been successfully used by van Rijn and van Selow in [7] to mine on similar data, we take a similar approach.

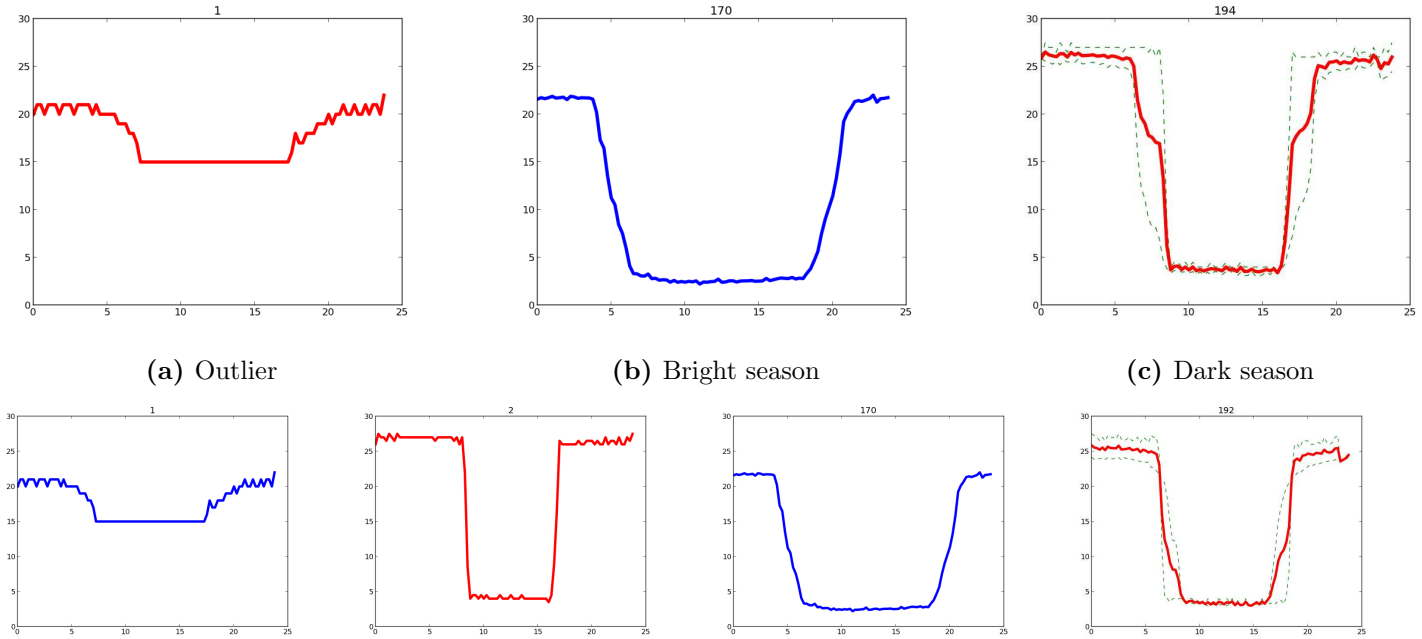
The clustering method used is a bottom up hierarchical clustering method. The quarter-hourly expenditure measurements are sliced into series of one day. Each day starts in its own cluster. Hence we initially start with 365 or 366 clusters. Next, the average distances between all clusters are computed. This is also called *average linkage clustering*. The two clusters with the shortest distance are merged together into a new cluster. Essentially any distance measurement can be used, but the measurement used here is Euclidean distance. The result of the clustering is a binary tree of $2M - 1$ clusters when no stop condition is used, where M is the amount of initial clusters.

When clustering the expenditure series of a public lighting object, we expect two patterns to emerge. The first pattern is the result of the darker seasons, where the lights must be left on longer during the morning, and turned on earlier in the evening. The second pattern is the opposite, where the lights are turned off earlier, and not enabled until later. These clusters are expected to be of almost equal *size* (where size is the amount of days associated with them), as the transition from one to the other happens gradually. Outliers on the other hand, will remain in smaller clusters, until they are finally merged during the later stages.

To see if the "outlier, dark season, bright season"-theory actually holds, the expenditure during 2010 of a random public lighting object was selected. Manual inspection of all 365 days revealed that the expenditure of this object contained no outliers. One day was replaced with an artificial outlier where the lights are on during daytime, to see if the clustering algorithm would identify it; which it did. The average pattern of the resulting clusters during the later stages of the clustering is shown in Figure 20. As the size of cluster the artificial day resides in is 1, the outlier is clearly identified.

The above only works for manual inspection, as there is no stop condition to stop the merging. Stop conditions can be based on the amount of remaining clusters or some distance threshold. Stopping based on the amount of clusters is unfit for this application as the desired amount of remaining clusters is unknown. The other approach is to use a distance criterion; i.e., stop merging when the distance between the clusters surpasses a certain threshold. We call this distance threshold d.

Once the distance threshold d is reached there is a set of clusters remain-



(a) Outlier **(b)** Bright season **(c)** Dark season

(d) Outlier **(e)** Dark season i **(f)** Bright season **(g)** Dark season ii

Figure 20: Example showing how the clustering algorithm detects outliers. For each cluster the average expenditure is plotted on the y-axis, against the time of the day on the x-asis. Red lines mean the cluster is scheduled for merging. The dashed green lines indicate which clusters were merged. The number on top of each graph show the size of the cluster.

ing. Of the clusters in this set, the smaller-sized clusters will most likely be outliers; they make up a small part of the population and are different enough to not be merged. Hence we define an upper bound s. Any cluster with a size smaller or equal to s is considered an outlier. Note that in this case, an outlier is not necessarily negative; some outliers found might have very low expenditure.

The difficulty is deriving s and d. In the following section an algorithm for deriving s and d on basis of the ad-hoc method is described.

5.3.1 Deriving the Distance Threshold and Outlier Cluster Size

In this section we propose an algorithm which tries to mimic the results of the ad-hoc method as closely as possible, in an attempt to find values for the distance threshold d and the upper bound for outlier cluster size s. As d and s are derived from normalized energy expenditure values, the assumption is that the found d and s are also useful for identifying outliers in classes other than public lighting objects.

First, the algorithm applies hierarchical bottom up clustering on normalized expenditure series of each object described in Section 5.1. The clustering method performs a sequence of merges, which we call a *merge sequence*. The merge sequence for each object is described by a list of 364 (365 for leap years) triples of the form $(distance, cluster1, cluster2)$. Here *distance* is the distance between *cluster1* and *cluster2* and *cluster1* and *cluster2* are merged into *cluster1*.

The individual merge sequences are combined into a larger merge sequence, which is a list of quadruples $(distance, cluster1, cluster2, id)$ where the *id* identifies each object on which the merge was applied. This merge sequence is sorted on distance. The merge sequence can then be iterated over, interpreting each quadruple as an instruction for the hierarchical clustering algorithm.

The goal is to get the best approximation of the ad-hoc method. This can be done by performing a step in the merge sequence, and then evaluating for what s the result would be optimal, as s is bounded by 365 (or 366). This evaluation is done by a fitness function:

$$fitness(clusters) = \sum_{c \in clusters} score(c)$$

With:

$$score(c) = \sum_{day \in c} \begin{cases} 1, & \text{if } day \in outliersFoundByAdHoc \text{ and } size(c) \leq s, \\ 1, & \text{if } day \notin outliersFoundByAdHoc \text{ and } size(c) > s, \\ 0, & \text{otherwise.} \end{cases}$$

The fitness function is the sum of all scores of the clusters currently found by the clustering algorithm. A cluster receives one point for each correctly

identified day. A day is said to be correctly identified if it is in a cluster with a size $\leq s$ and the day is an outlier according to the ad-hoc method. If the day is not an outlier, the day is said to be correctly identified if the size of the cluster it is in, is greater than s.

Using the fitness function, the algorithm performs all merge steps and remembers which d and s led to the best fitness. In Section 5.4 the values for d and s are shown, and the outcome of the hierarchical clustering method is compared with the outcome of the ad-hoc method.

5.4 Comparison of the Detection Methods

This subsection outlines the similarities and differences between both detection methods. Using the algorithm described in Section 5.3.1, the parameter settings for the hierarchical clustering method were determined. The value of s, the upper bound for the size of clusters of outliers, was found to be 7. The value of d, the cluster distance to stop merging at, was found to be roughly equal to 7.9157×10^{-8}. Both methods were applied on the dataset sample. The result is a listing of how many of the outliers found by the ad-hoc method are also detected by the hierarchical clustering approach, and how many other outliers the hierarchical clustering method finds. A complete list of the results can be found in Table 11 in Appendix C. In the remainder of this section the most notable differences are outlined through the use of figures. In these figures dashed expenditure series represent normal daily expenditure series. Solid lines represent a daily series of expenditures which the ad-hoc method identified as an outlier. It should also be noted the figures do not always show all clusters, hence the total sizes of the clusters in a figure does not add up to 365.

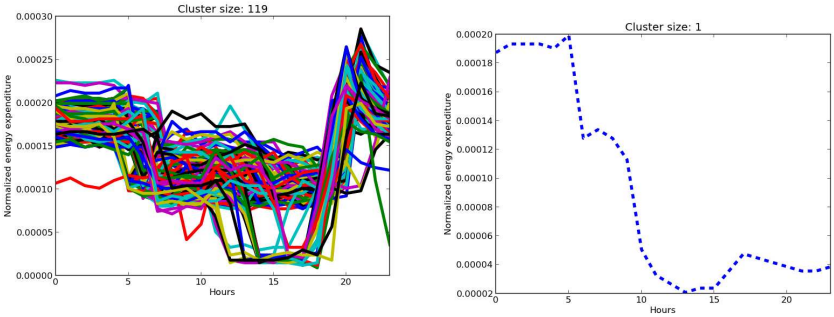
The ad-hoc method identifies 1014 outliers across all 51 objects. Despite deriving the parameters of the hierarchical clustering method using the ad-hoc method, hierarchical clustering identifies only $\approx 25\%$ (254 out of 1014) of the outliers. In addition, it identifies 185 other outliers. The large difference between ad-hoc and hierarchical clustering seems caused primarily by the objects which have ≥ 14 outliers according to the ad-hoc method. In these instances hierarchical clustering identifies only a small portion of the outliers. The problem here may be that the "outliers" occur too frequently. Figure 21a shows that many days identified as an outlier by the ad-hoc method, are similar in their expenditure pattern and end up in the same, large cluster. When applied on the same object's expenditure series, hierarchical clustering identifies an outlier that was unidentified by the ad-hoc method. This outlier is presented in Figure 21b.

There are also cases where the hierarchical clustering predicts many different outliers from the ad-hoc method. Figure 22 shows an example of such a case. In this case the hierarchical clustering method finds 27 out of 28 of the outliers found by the ad-hoc method, yet also finds 84 others. These

84 other daily expenditures are identified as outliers because the object's daily expenditure is very inconsistent throughout the year, which confuses the hierarchical clustering method.

The cases where the ad-hoc method and hierarchical clustering agree on which daily expenditure series are considered outliers, seem to be when the object's expenditure is highly consistent. Figure 23 shows an example where both methods found the same 10 outliers, and the remaining daily expenditure series are highly consistent.

The hierarchical clustering method seems to deliver quite different results from the ad-hoc method. Although it gets confused easier, the hierarchical clustering method is fortunately able to find different types of outliers, e.g., instances where the lights are off during nighttime. In the following section, the hierarchical clustering method is applied to other classes.



(a) Not recognized as outliers by hierarchical clustering

(b) Not found by ad-hoc

Figure 21: Differences between the hierarchical clustering method and the ad-hoc method, when the object's expenditure series contain a large amount of outliers according to the ad-hoc method.

5.5 Outliers in Other Classes

In this subsection we explore how well the parameter settings derived in the previous section generalize. For each of the categories: office, tunnel and floodgate/weir one example of the found clusters is shown. See Figure 24 for the office example, Figure 25 for the tunnel example, and Figure 26 for the floodgate/weir example. Each cluster is annotated with a *possible* explanation for the behaviour in the cluster.

The parameter settings found do not generalize very well. In some cases the hierarchical clustering method groups expenditure series which to the naked eye have little in common (Figure 27). This would mean the distance d at which the clustering algorithm stops merging is too high. Unfortunately, in other cases the clustering method finds many (> 10) different groups of

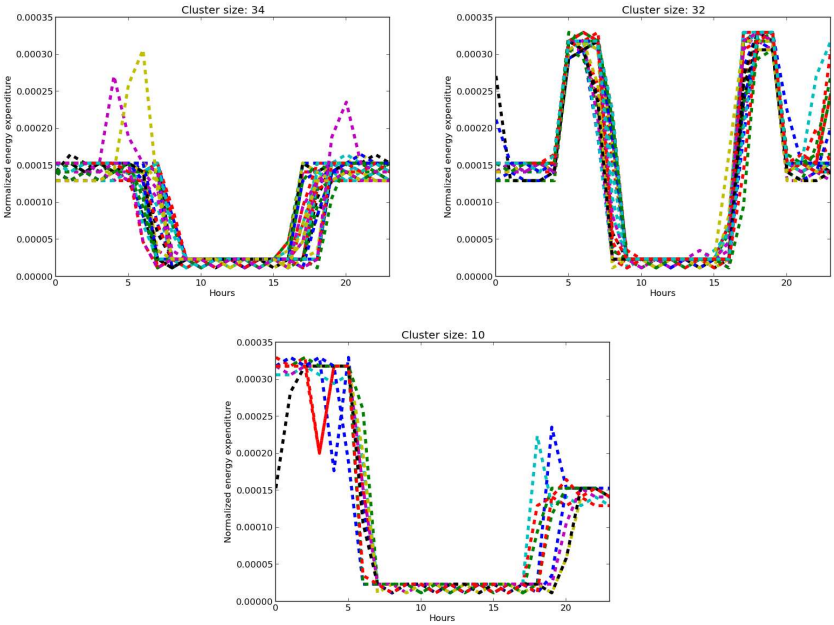


Figure 22: Example of the expenditure of an object showing different behaviour.

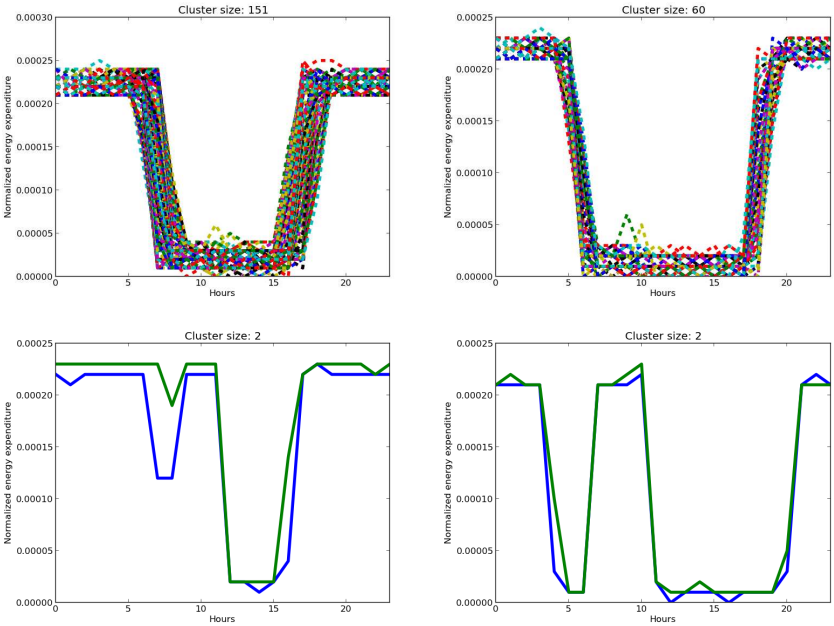
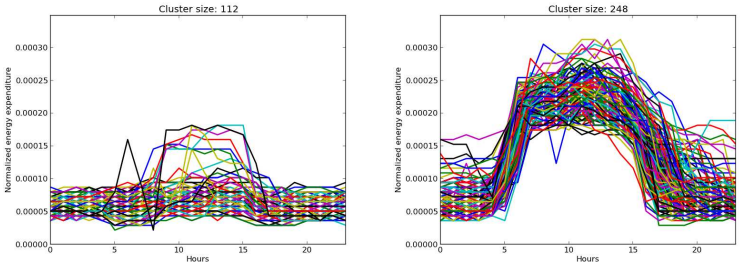
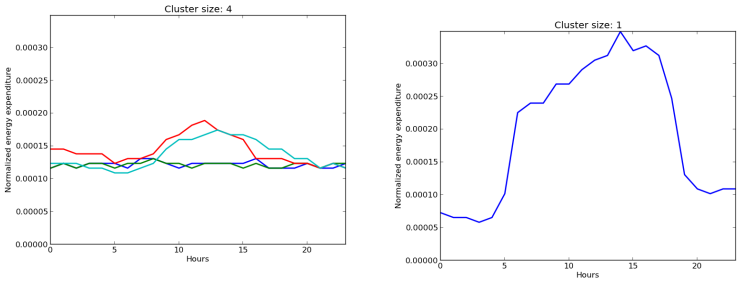


Figure 23: Example of the ad-hoc and hierarchical clustering method agreeing on which daily expenditure series are outliers.



(a) Weekends / holidays

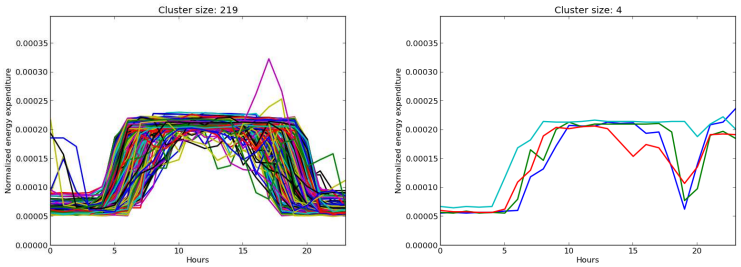
(b) Normal operation



(c) Weekends / holidays with slightly higher expenditure

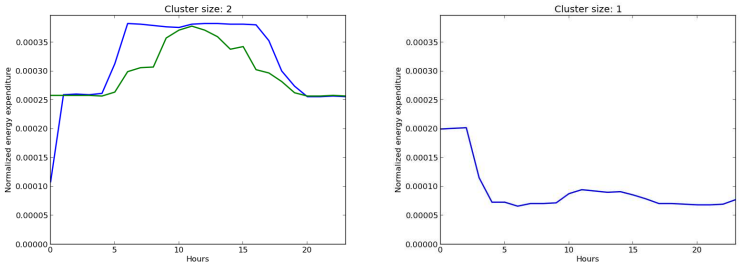
(d) Peak usage

Figure 24: Example of clusters found for an office.



(a) Normal operation

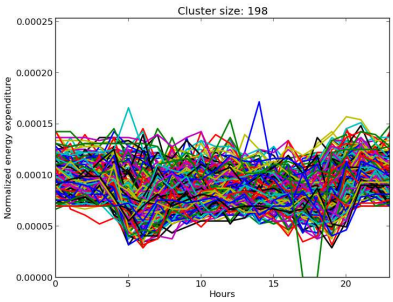
(b) Maintenance during night



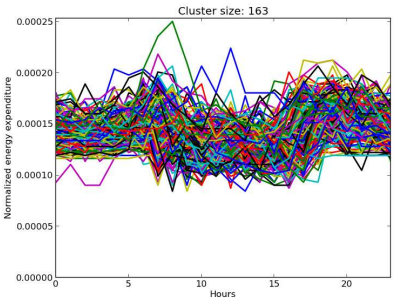
(c) Maintenance during the day

(d) Unexplained

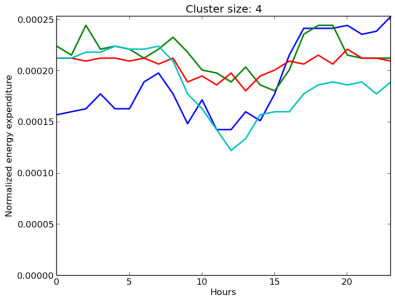
Figure 25: Example of clusters found for a tunnel. Not all clusters are shown.



(a) Little rain



(b) Much rain



(c) Even more rain

Figure 26: Example of clusters found for a floodgate/weir.

outliers, of which some seem reasonably similar. In these cases the value for d seems too low.

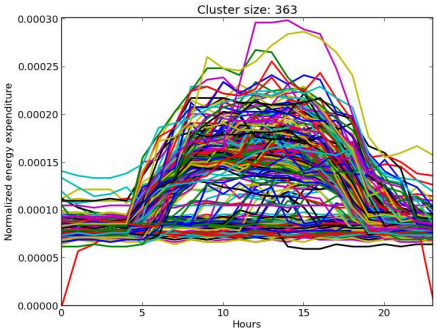


Figure 27: Example of a cluster which seems to contain two different kinds of patterns.

5.6 Conclusion

In this section we took a look at anomaly detection within time series. We have shown two methods for detecting these anomalies. An ad-hoc method which focusses on identifying days where the light is on during daytime, and a generally applicable hierarchical clustering method. The ad-hoc method found a lot of outliers. Fortunately, most found outliers seem explainable with domain knowledge.

The hierarchical clustering method requires specific parameters to be determined; the cluster distance which to stop merging at, and the size at which clusters are considered to contain outliers. These parameter settings were determined using the results from ad-hoc method as a heuristic. The hierarchical clustering method and ad-hoc method did not agree on all outliers, which is partly because the hierarchical clustering method gets confused when the daily expenditure patterns of an object are very different. Another reason they did not agree is because some outliers would appear too frequently for the hierarchical clustering method to detect them.

Using the parameter settings found for the hierarchical clustering method, we took a look at how other classes were clustered. Two problems appeared; the found value for d appeared too high, as some clusters seemed better of if they were split. However, in other cases many different clusters were found. In these cases it seems better to increase the value of d. Using the ad-hoc method to derive the parameter settings leads to decent, but far from perfect results.

6 Summary & Future Work

We started the thesis with the following questions:

1. How does the energy expenditure relate to the weather conditions?
2. How to identify objects that behave differently from other objects?
3. How to identify outliers in the expenditure pattern of an object?

We tried to answer the first question using the sample Pearson correlation coefficient. We learned that the correlations between the energy expenditure and weather variables are larger, if we convert the hourly expenditure and weather measurements to daily averages. This phenomenon might be because certain weather conditions do not affect the expenditure in the same hour, but do have influence over a longer period. Another reason for the increased correlation might be that the daily expenditure pattern obscures some of the influence the weather has.

The expenditure per region showed quite a few moderate correlations with temperature, global radiation, humidity, snowfall, icing and cloudiness.

For the daily averages some regions even show a strong correlation with temperature. However, it should be noted that correlation coefficients only give an indication of how the weather and expenditure are related. It does not allow one to estimate the expenditure on basis of the weather conditions. Future work could try to answer this.

We tried to answer the second question by clustering objects on their expenditure patterns. We used a Self-Organizing Map and clustered both the normalized (but otherwise raw) expenditure data and a Bag-of-Pattern representation of the data. The Bag-of-Pattern representation performed slightly worse than the normalized expenditure data. However, neither performed very well. The only group that was clustered near-perfectly was the public lighting group. Possibly because this group has the most consistent expenditure pattern.

Future work could attempt to improve the quality of the clustering by finding another representation of the data. We learned from the Bag-of-Patterns approach that the daily alignment of the expenditure data matters and should not be discarded. A possible explanation as to why the Bag-of-Patterns representation is outperformed by the raw expenditure data, is because the Bag-of-Patterns representation discards any alignment on a scale larger than daily. It also does not take into account that the strings `0001` and `0002`, are more similar than for example `1110` and `0001`. A better representation would avoid these flaws.

We tried to answer the third and last question by first solving the question for a smaller set of objects: only public lighting objects. This resulted in an ad-hoc method which identified on which days the lights of a public lighting object were on during daytime. Using the answers of the ad-hoc method as a heuristic we identified our parameter settings for a hierarchical clustering algorithm. Both algorithms identified the same outliers as long as the expenditure pattern of the object was very consistent, and there were not too many outliers. Many outliers were not recognized through hierarchical clustering as they occurred too frequently. In other cases the hierarchical clustering method identified many outliers when the expenditure pattern was very inconsistent. The found settings did not generalize perfectly when also considering categories other than public lighting. Perhaps a better representation of the data will also make this problem more trivial.

Overall we have formed a reasonable understanding of the data. The irregularities of the data make it difficult to deal with, but different representations might help. The outlier detection methods are not robust enough to be applied in practice, but the regional correlations might help managing for energy reduction.

7 Acknowledgements

Thanks to Harald Versteeg from Rijkswaterstaat for allowing me to work on this project. Thanks to Rudy van Mierlo for supervising me at Rijkswaterstaat. Also many thanks to dr. Walter Kosters and Frank Takes from LIACS, Leiden University for their supervision and help.

References

[1] Koninklijk Nederlands Meteorologisch Instituut. Uurgegevens van het weer in Nederland. `http://www.knmi.nl/klimatologie/uurgegevens/`, April 2012.

[2] T. Kohonen. The self organizing maps. *Proceedings IEEE*, Vol. 78(9):1464–1480, 1990.

[3] T. Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, Vol. 38:1857–1874, 2005.

[4] J. Lin, E. Keogh, W. Li, and S. Lonardi. Experiencing sax: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery Journal*, pages 4–8, 2007.

[5] J. Lin and Y. Li. Finding structural similarity in time series data using bag-of-patterns representation. *Proceedings of the 21st International Conference on Scientific and Statistical Database Management*, pages 465–468, 2009.

[6] R.W. Sinnott. Virtues of the haversine. *Sky and Telescope*, Vol. 68(2):159, 1984.

[7] J.J van Wijk and E.R. van Selow. Cluster and calendar based visualization of time series data. *Proceedings of IEEE Symposium on Information Visualization*, pages 2–7, 1999.

[8] E.W. Weisstein. von Neumann Neighborhood. `http://mathworld.wolfram.com/vonNeumannNeighborhood.html`, Juli 2012.

A Regional Expenditure Category Composition

	RDU	DLB	DON	DIJ	DNB	DID	DZH	DNN	DZL	DNH	DVS	DNZ
public lighting or traffic control	30	3	4		3		27	2	3	5		1
office	4	1	2	2	4	1	6	1	4	3	1	4
tunnel		2					11	1	2	4		
unknown			1					1	2			1
radarpost								1	1			
pumping station	1	3	1		1		1	1	1	2		
flootgate or weir	5	7	4	5	2		8		6	3		
traffic control center	2		2		1		2		1	1		
bridge or dam					2		3		2	1		
small building			1	1					1			
total	42	16	15	8	13	1	58	7	23	19	1	6

Table 8: List of how many objects are present in each region and what expenditure category the objects belong to.

B Regional Correlation Coefficients

B.1 Hourly Correlation Coefficients

	RDU	DLB	DON	DIJ	DNB	DID	DZH	DNN	DZL	DNH	DVS	DNZ
DD	−0.140						−0.132	−0.138		0.137		
FH			0.111	0.146		0.146	0.114	−0.126		0.215	0.145	0.131
FF			0.116	0.144		0.149	0.119	−0.123		0.209	0.147	0.128
FX			0.120	0.109		0.167		−0.143		0.245	0.164	0.136
T	**−0.518**	0.279		**−0.458**	**0.465**	**0.410**	**−0.548**	**−0.585**	−0.300		**0.353**	**−0.334**
TD	**−0.446**	0.123		**−0.493**	**0.368**	0.291	**−0.620**	**−0.496**	−0.228		0.214	**−0.344**
SQ	**−0.307**	0.201	0.206		0.279	**0.361**		**−0.475**	−0.198		**0.384**	
Q	−0.255	0.253	0.275		**0.393**	**0.511**		**−0.554**	−0.223		**0.508**	
DR										0.146		
RH												
P								−0.110	−0.112	**−0.302**		
VV	**−0.325**	0.157		−0.188	0.208	0.215	−0.155	**−0.318**	−0.188		0.212	
N	0.171	−0.100		0.156			0.104		0.122	0.163		0.137
U	**0.317**	**−0.339**	−0.239		**−0.310**	**−0.378**		**0.461**	0.255	0.109	**−0.405**	
M								0.133			−0.100	
R			−0.103				−0.109			0.209		
S	0.174		0.105	0.197			0.227	0.178	0.144			0.180
O												
Y				0.120			0.113	0.158				

Table 9: Sample Pearson correlation coefficients for the hourly expenditure of each regions and all the weather variables. Coefficients with an absolute value < 0.1 are omitted, values between 0.1 up untill 0.3 are typeset normally, and values ≥ 0.3 are typeset in **bold**.

B.2 Daily Correlation Coefficients

	RDU	DLB	DON	DIJ	DNB	DID	DZH	DNN	DZL	DNH	DVS	DNZ
DD	−0.262		−0.145	−0.244	0.145	0.120	**−0.307**	−0.228	−0.146	0.209		
FH	0.111	−0.169			−0.176				0.148	**0.307**	−0.104	0.183
FF	0.104	−0.163			−0.170				0.151	0.300		0.177
FX		−0.139			−0.131				0.145	**0.343**		0.193
T	**−0.646**	0.299	−0.111	**−0.674**	**0.492**	**0.466**	**−0.762**	**−0.735**	**−0.478**	−0.136	**0.371**	**−0.505**
TD	**−0.584**	0.178	−0.165	**−0.647**	**0.427**	**0.455**	**−0.752**	**−0.689**	**−0.388**		**0.337**	**−0.471**
SQ	**−0.390**	0.288		**−0.338**	0.276	0.130	−0.284	**−0.452**	**−0.370**	−0.289	0.151	**−0.333**
Q	**−0.539**	**0.356**		**−0.483**	**0.401**	0.241	**−0.488**	**−0.654**	**−0.552**	**−0.376**	0.239	**−0.442**
DR		−0.166	−0.138		−0.158				0.174	**0.396**	−0.113	0.162
RH			−0.135	−0.113			−0.143		0.128	**0.349**		
P	−0.133	0.117	0.120	−0.103				−0.168	−0.216	**−0.383**		−0.139
VV	**−0.450**	0.151		**−0.410**	0.220	0.149	**−0.374**	**−0.394**	**−0.327**		0.124	−0.205
N	**0.310**	−0.220		0.243	−0.175		0.194	0.246	**0.311**	0.277		0.227
U	**0.402**	**−0.371**		**0.326**	**−0.306**	−0.183	**0.312**	**0.476**	**0.441**	0.288	−0.223	0.229
M				0.205				0.239	0.148			
R		−0.182	−0.268	−0.120	−0.134		−0.159		0.115	**0.452**		
S	**0.349**		0.225	**0.426**	−0.115	−0.213	**0.440**	**0.357**	**0.401**		−0.154	**0.364**
O	−0.151	0.154		−0.141	0.114		−0.133	−0.140		0.230		−0.106
Y	0.219		0.116	**0.303**			0.286	0.299	0.239			0.154

Table 10: Sample Pearson correlation coefficients for the average daily expenditure of each regions and all the weather variables. Coefficients with an absolute value < 0.1 are omitted, values between 0.1 up untill 0.3 are typeset normally, and values ≥ 0.3 are typeset in **bold**.

C Hierarchical Clustering Compared to Ad-hoc

	Hierarchical Clustering		
Ean	Shared with ad-hoc	Additional	Ad-hoc
0016	0	2	1
0025	0	0	0
0026	1	20	1
0021	0	0	0
0022	2	0	3
0024	0	0	2
0027	2	6	5
0028	0	0	3
0029	33	38	37
0013	0	0	0
0030	0	0	15
0031	1	1	14
0006	0	0	0
0009	3	1	6
0007	4	0	5
0020	0	0	0
0032	6	0	13
0033	1	2	30
0034	2	0	2
0035	0	0	37
0036	0	0	136
0001	1	1	255
0037	5	0	15
0011	1	0	3
0038	0	0	0
0039	0	0	3
0040	5	5	11
0010	0	1	4
0015	0	2	1
0041	0	0	3
0042	0	0	0
0043	0	0	4
0017	3	1	22
0005	1	1	2
0023	0	0	0
0044	0	0	1
0019	2	0	5
0002	10	0	10
0045	106	5	128

Table 11 – continued from previous page

	Hierarchical Clustering		
Ean	Shared with ad-hoc	Additional	Ad-hoc
0008	6	0	9
0046	0	0	61
0047	1	0	6
0048	0	2	0
0014	28	7	48
0049	27	84	28
0050	0	0	2
0012	2	1	4
0003	0	2	14
0004	1	3	2
0051	0	0	2
0018	0	0	61
total	254	185	1014

Table 11: Comparison between the hierarchical clustering method and the ad-hoc method. Showing how many instances both methods predict to have outliers. Of the hierarchical clustering method is shown how many outliers it shares with the ad-hoc method, and how many other outliers it finds.