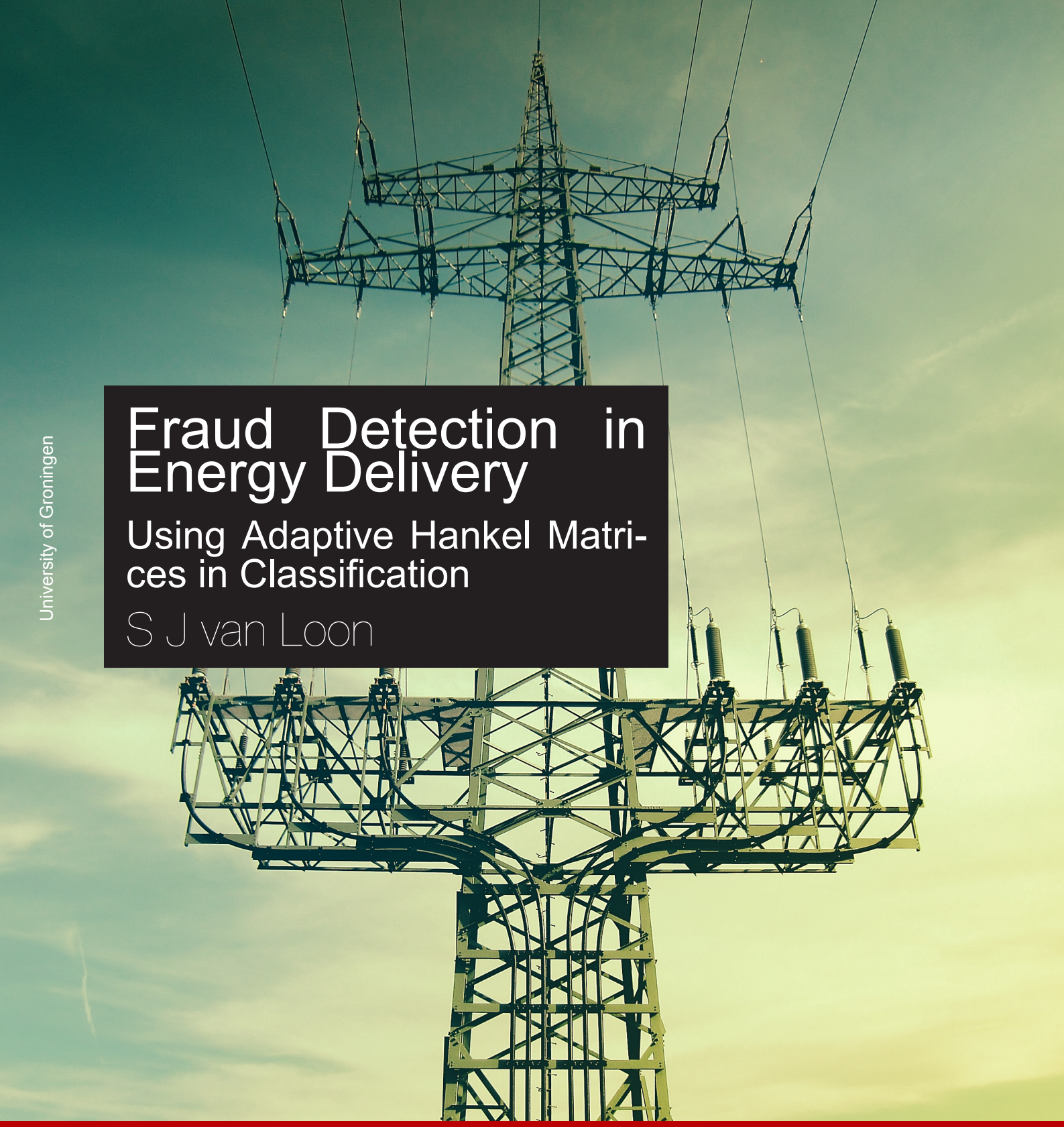


Fraud Detection in Energy Delivery

Using Adaptive Hankel Matrices in Classification

S J van Loon

University of Groningen

university of groningen / faculty of science and engineering

Fraud Detection in Energy Delivery:
Using Adaptive Hankel Matrices in Classification

by

S J van Loon

to obtain the degree of Master of Science
at the University of Groningen.

Student number:	1795813	
First Supervisor:	Prof dr M Biehl	University of Groningen
Second Supervisor:	M Mohammadi MSc	University of Groningen
External Supervisor	R Bosgraaf	ValueaA

Acknowledgments

During the course of the thesis I have enjoyed the help and support of many people. First and foremost I would like to express my gratitude to Professor Michael Biehl, whose continuous support, feedback, supervision and cheerful guidance have made this thesis possible. The same goes for the people at Coteq and ValueA, Richard Bosgraaf, Ad Schellevis, and Marcel van den Berg, who I like to thank for giving me the opportunity of doing this graduation project, their support throughout the process, and for the many interesting conservations we have had at the Theater hotel.

Furthermore I would like to thank my friends at the university, especially Jelle, Laura, and Rick, with whom I shared many thoughts, problems, conversations and cups of coffee. Without their support, input and feedback this thesis would not have been made.

Above all I want to express my deepest gratitude to Marjoke, my family, and my friends whose continues love, support and patience throughout my whole studies made it all possible.

Fraud Detection in Energy Delivery: Using Adaptive Hankel Matrices in Classification

Abstract

In the Netherlands there is a high correlation between theft of electricity and cannabis growing operations. Growing rooms and the related electricity theft pose a risk to the physical safety of the general population due to illegal manipulation of the electricity network, faulty connections, and excessive consumption of electricity. It is estimated that in the Netherlands yearly 200 million euro of losses are caused by electricity theft by cannabis growing operations.

Coteq is a Dutch Distribution Network Operator that tries to dismantle cannabis growing operations in collaboration with local law enforcement. In order to locate growing operations, a software solution provided by ValueA is used for the manual identification of patterns in electricity usage indicating illegal activities.

This thesis provides an extension to the software solution of ValueA by automating the identification of suspicious patterns. In this thesis three dissimilarity measures are compared. Euclidean distance, Hankel based dissimilarity, and Dynamic Time Warping. Of this three measures the dynamic time warping is shown to give the best classification results.

Contents

Acronyms

DNO Distribution Network Operator. 1, 2, 4, 7, 12, 55

DSB Dutch Safety Board. 1

DTW Dynamic Time Warping. 2, 29–31, 38, 45, 51–53, 55, 56

KNN K-Nearest Neighbors. 17, 18

LTI Linear Time-Invariant. 22–26

1 Introduction

On the first of March 2018, the Dutch Safety Board (DSB) [1] published a report [5] on the environmental safety of illegal cannabis grow rooms. In the report the DSB highlights the physical safety risk of illegal cannabis cultivation. Considering the dangers to the public caused by illegal cannabis cultivation the DSB has made the following recommendation to *Netbeheer Nederland*, the affiliation of Dutch Distribution Network Operators (DNOs).

> Ensure the continued development in the short term of an automated measuring system on the distribution network that makes it possible to measure unsafety at home level caused by the illegal manipulation of the electricity network, excessive consumption of electricity, and cannabis-related patterns of consumption. Detect unsafe connections on the electricity network and act against residents who cause cannabis-related dangers in the electricity network.[5]

The recommendations capture the importance of DNOs actively trying to detect and act against illegal cannabis growing operations. In this thesis we will present a system for automatic detection of electricity consumption patterns related to cannabis growing operations.

1.1 Motivation

The motivation behind this thesis is a combination of solving the practical problem of automated detection of cannabis-related patterns in electricity con-

[1] The DSB is an autonomous organization invested in 2005 by the Dutch government to investigate incidents related to public safety.

sumption and the academic motivation for real world application of recent introduced techniques.

1.1.1 Subspace Based Classification

Time-series data consists of sequential measurements ordered over time[2]. Times-series, converse to non-sequential data, suffers from something called the alignment problem. The alignment problem occurs when only a part of a time-series is relevant to a certain task, but is obfuscated by the non-relevant parts. When the relevant part of a times-series A occurs in a different part of the time-series compared to another time-series B, the time-series A and B are misaligned. Due to the alignment problem conventional distance or dissimilarity methods, which work well on non-sequential data, might not work on misaligned times-series data. More information about time-series and the misalignment problem is given in section 4.4.

The alignment problem can be (partly) solved using dissimilarity metrics based on subspace angles between Hankel matrices. In [56] a comparison was done regarding several Hankel based dissimilarity metrics. During this research three dissimilarity metrics were tested on 35 time-series datasets taken from [12]. On average a 10% performance increase[3] compared to Euclidean distance was achieved by using Hankel based dissimilarity measures. In this research the rotational invariant subspace approximation was found to be the best performing over most of the used datasets. In this thesis we will deploy the rotational invariant subspace angle approximation for fraud detection on the electricity grid to validate its performance in real world application. For comparison Euclidean distance and Dynamic Time Warping (DTW) will also be used.

1.1.2 Detecting Cannabis-related Patterns in Electricity Consumption

Coteq is a Dutch DNO licensed to distribute electricity in the municipalities Almelo, Hof van Twente, and Oldenzaal. At Coteq a small team of experts is full time employed to battle electricity fraud, with special attention to electric-

[2] In this thesis only time-series are considered, however the problems relating to time-series and the techniques used to solve them can be applied to most forms of ordered, sequential data such as contours, spectrograms, and images.

[3] In terms of classification rate.

ity theft committed by cannabis growing operations. The fraud detection team at Coteq tries to manually identify patterns in the electricity consumption typical for cannabis growing operations in order to locate and dismantle cannabis growing operations together with local law enforcement. Manual identification of suspicious pattern is done using a software solution provided by ValueA.

Manual detection of cannabis-related electricity consumption patterns is time consuming and tedious work which requires expertise of the typical patterns. Automatic detection of cannabis growing operations would provide several benefits.

1. By shifting from manual to automatic detection the experts would be able to better allocate their time previous spend on the manual examination of electricity usage patterns.
2. By automating the detection process using machine learning, cannabis related electricity consumptions unknown to the experts may be found, resulting in a higher detection rate.
3. Currently the monitoring is limited to one power distribution station at a time. Automation of detection could lead to the upscaling of the monitor process, making it possible to measure several distribution stations simultaneously.

The benefits above show the advantages of an automatic detection system for cannabis related electricity consumption patterns and forms one of the main motivations for this thesis.

1.2 Contribution

In this thesis we address the problem of automatic detection of suspicious electricity consumption patterns associated with illegal cannabis growing operations. The goal of this thesis is to develop tooling for the automatic detection of suspicious pattern in electricity consumption measurements.

To achieve the goal of an automated detection system we present a prototype to automatically detect suspicious energy consumption behavior to supplement the software solutions provided by ValueA and to aid the energy fraud detection team of Coteq. Used techniques include learning vector quantization combined with subspace angle approximation on Hankel matrices as dissimilarity measure. The contribution will be a continuation of the work done

in [60] in which Hankel based methods were explored for time series classification without using a sliding windows approach. In this thesis will test the Hankel based methods in a practical application.

1.3 Related Work

Electricity theft is not a problem unique to the Netherlands, however the strong correlation between electricity theft and cannabis growing operations might be. In [69] it is estimated that in 30% of the known cases of electricity theft in the United Kingdom related to cannabis growing operations[4], compared to 95% of the detected cases in the Netherlands[46].

Other sources which consider electricity theft focus on developing countries [10, 16, 62] where electricity theft is of a different, more domestic nature. As a result, most literature does not focus on detection on electricity usage patterns relating to cannabis growing.

Known solutions for the identification of electricity theft are often based on smart meters readings, such as the solutions proposed in [10, 16, 35, 46, 57, 62, 77]. These solutions are not suitable for Dutch DNOs due to privacy laws protecting customer rights. Section six of the code of conduct of Dutch DNOs [66] states that smart meter data may only be retrieved six times a year and usage of the smart meter data is restricted to certain cases. This means Dutch DNOs cannot rely on smart meter readings for the purpose of fraud detection. This means most related work cannot be used for the practice of fraud detection.

1.3.1 Scope and Limitation

The main aim of this thesis is the development of an automated fraud detection system using the Hankel based techniques in [56] using power usage readings from substations. The performance of the developed system will be partly dependent on how suitable the Hankel based methods are for the problem of fraud detection on the electricity grid.

The application developed for this thesis will be a standalone prototype, but will be developed in such a way that it can be embedded into the existing systems of ValueA with only a few modifications. Focus will be put on the

[4] Note that the relative amount of stolen electricity relating to cannabis growing operations might be higher due to the high electricity usage of cannabis growing operations.

implementations of the various (Hankel based) classifications modules and supporting module needed for data pre-processing and cross validation.

1.4 Organization

This thesis is divided in four parts, all of which have will have a short introduction presenting the topics discussed and their organization. In part I an embedding of the project is given by providing background information regarding cannabis cultivation, electricity theft and the related dangers. Part II gives an overview of the techniques and methods used in this thesis. Part III discusses the technical solution using the methods presented in the previous part. Here we also discuss the classification pipeline, including the aggregation of the data and the validation of the used classification models. Finally the results are presented and discussed in part IV, finalizing with a concluding chapter and an overview of possible future work.

Part I

Cannabis Cultivation and Energy Fraud

Introduction

This thesis presents a solution for the automated detection of cannabis related patterns in electricity consumption. In this part embedding of this thesis is presented by given a culture background into the cultivation and regulation of cannabis growing in the Netherlands in chapter 2. With this background we get a better understanding of cannabis cultivation in the Netherlands, and how it is possible to detect these operations via their electricity consumption patterns.

In the next chapter, chapter 3 the dangers and financial cost of electricity theft are discussed, giving a better insight into the motivation of Dutch DNOs to actively look for cannabis growing operations within their service area.

2

Cannabis Regulation and Cultivation

In this chapter background information is provided into the regulations of cannabis cultivation in the Netherlands and the typical characteristic of a cannabis growing operation. Section 2.1 is pure informative and provide some background information regarding the regulation of cannabis in the Netherlands. In section 2.2 general methods for cultivating cannabis used by professional producers. These methods might provide some background into why it is deemed feasible to detect cannabis growing operations on basis of their electricity usage.

2.1 Cannabis Regulations in the Netherlands

While in most European countries both the cultivation and trade of cannabis are illegal, cannabis legislation in the Netherlands is in a juxtaposition due to its tolerance policy.

In 1976 the *opium wet* (opium law) was introduced in which the distinction between hard and soft drug was made and the tolerance policy was introduced. The tolerance policy regarding *soft drug* entails that the sale of cannabis related products is a criminal offense, but is tolerated by the Public Prosecution Service for some designated parties (coffee shops). The possession and production of cannabis by members of the public is also tolerated for small quantities, cannabis cultivation on a professional level is however prosecuted. Cultivation is considered professional when six or more cannabis plants are grown, or when equipment is used to promote the growth of the cannabis yield[67]. This leads to the curious situation in which the selling of cannabis by coffee shops is allowed and regulated, but the so called *back door*, growing cannabis to sup-

ply coffee shops, is still illegal. As a result, most of the cannabis production occurs in criminal circuits.

2.2 Cannabis Cultivation

In [7] an overview is given about cannabis in the Netherlands, including the history, growing culture, the associated criminality, and the combat against cannabis related criminality. When cannabis became popular around 1960 most cannabis was imported from hot climate countries like Morocco, since the Dutch climate is not very suitable for cannabis cultivation. Some cannabis was (and still is) grown outside, but the colder climate does not allow for successful growth of high grade cannabis on a larger scale. At the start of the eighties growers started growing indoors using artificial lighting causing a switch from imported cannabis to *nederwiet*, Dutch grown cannabis[78].

Most professional growing operations have an extensive setup using hydroculture[1], artificial lighting, air extraction systems, automated watering and fertilizing systems. The general grow period is 71 days; 63 days with a twelve-hour lighting period and eight days with eighteen hours of lighting.

In [72] the effect of electric lighting on floral development and potency of cannabis plants were examined. During the study an experiment was conducted in which different strains of cannabis were grown under various lighting intensities with an artificial day length of twelve hours. The study spanned the range $(400 - 500\mathrm{W\,m^{-2}})$ for a flowering cannabis crop recommended in an assortment of published and on-line growing guides. The majority of illicit cannabis growing operations in the Netherlands were reported to use power levels within this range. There was a significant $p < 0.0001$ correlations found between the power level and total mass of THC produced. This study shows that in general cannabis growers will have a typical pattern of high electricity usage in order to maximize yield.

From [7, 72] we can conclude there is a typical lighting pattern that is often used which maximizes the yield of cannabis production. Assuming the power usage is high enough, it is feasible to distinguish the usage pattern of cannabis growing from other electricity usage patterns. This makes examination of electricity consumption a fruitful method for the detection of cannabis growing operations.

[1] Hydroculture is a method to grow plants without soil.

3

Electricity Theft: Dangers and Losses

When a distribution network operation (DSN) transmits electricity to its customers, a certain percentage of the electricity the DSN distributed is lost; there is a difference between the amount of electricity delivered to the distribution system and the amount of electricity that customers are billed for.

A part of the loss in electricity is caused by the physical properties of the components of the distribution network. This type of loss is known as *technical loss*. An example cause of technical loss is the electrical resistance of the power lines and transmission systems causing electricity to be converted to heat.

A second type of loss is known as *non-technical loss* which is the collective term for all losses not caused by the natural properties of the distribution system. Two of the biggest factors of non-technical losses are customers that do not pay their bills and people committing electricity theft. In the Netherlands 60% of non-technical losses are caused by electricity theft.[23]

Compared to non-paying customers, electricity theft is a difficult problem since one has to distinguish electricity theft from technical loss and offenders are (in general) hard to locate.

In this chapter we discuss types of electricity theft and its relation with cannabis growing in section 3.1. In section 3.2 we give an overview of the economic losses caused by electricity theft. We finalize this chapter by highlighting the dangers related to electricity theft and cannabis growing operations in section 3.3.

3.1 Electricity Theft

Regular electricity consumption occurs via a metered connection between the main power line and the and the consumer. The meter tracks the amount of electricity that is used, and customers are billed based on the meter readings. Electricity theft occurs when people use electricity without their usage being registered. In general people commit electricity theft by one of the following ways[65].

1. Meter tampering: the act of interfering with the meter to corrupt its readings. This is often done to older, disk type meters by slowing down or stopping the disk.
2. Bypassing: the act of bypassing the meter between the connection with the main power line and the main fuse box.
3. Illegal connections: creating a new (illegal) connection with the main power line.

The last two methods, beside being fraudulent, also have major safety risks since they require manipulation of live power lines.

3.1.1 The Relation Between Cannabis Cultivation and Electricity Theft

There is a strong relation between electricity theft and cannabis growing in the Netherlands. In [7] it is estimated that at eight out of ten cannabis growing operations electricity is stolen for the purpose of growing cannabis. Stated reasons for cannabis growers to commit energy theft are to maximize profit and the fear of getting caught by suspiciously high energy bills.

In [64] cannabis growers are stated as being the main offenders of electricity theft. This is also in line with the reports about electricity theft, already in 1995 there were known cases of cannabis growing operations where electricity was stolen[7]. Reports from 2012[23], 2014[29], and 2016[30] showed that yearly around 5000 cannabis growing operations are caught which were involved with electricity theft.

3.2 Economical Losses Caused by Energy Theft

In [64] some figures regarding the cost of electricity theft in the Netherlands are presented. In the Netherlands an estimated 1 billion $\mathrm{kW\,h^{-1}}$ of electricity

Figure 3.1: An electric installation at a busted cannabis growing operation. (Photo courtesy of www.politie.nl, 2016.)

is stolen each year. This amount represent 200 million euro of financial losses, would this electricity be sold to consumers. This amount is based on the following calculations: in the Netherlands there are an estimated 30 thousand illegal cannabis growing operations which, on average, use 35 thousand $\mathrm{kW\,h^{-1}}$ of electricity a year. Of the 200 million euro of financial losses 20% consists of transportation costs, 45% delivery costs, and 35% in taxes. Yearly around 150 million $\mathrm{kW\,h^{-1}}$ is reclaimed from people convicted of electricity theft, this represents about 15% of the amount of the stolen energy.

Dutch DNOs together employ around hundred people full time to find and prosecute people committing electricity theft, bringing addition cost to the raw value of stolen electricity.

3.3 Dangers Related to Cannabis operations

Besides the economic losses, DNOs also clamp down on energy theft due to its related dangers. One of the major safety risks of cannabis growing operations is that of fires due to faulty electrical connections. According to [5] at least 75% of dismantled grow rooms each year can be considered dangerous because of a combination of the high levels of power that are used, the illegal tapping of electricity, and the fact that the equipment is installed and used inexpertly. This is illustrated in fig. 3.1 which shows an electrical installation in a busted grow room.

Due to the illicit nature of cannabis growing operations there is no exact

statistics about cannabis related fires. In 2015 Salvage, a Dutch foundation of fire insurance companies, estimated in [76] that in the period 2012–2014 two percent (2%) of all fire insurance claims in the Netherlands were related to cannabis growing operations. However, [23] states that according to Dutch police above 30% of all fires in residential areas are caused by illegal cannabis growing operations. This discrepancy might be explained by cannabis growers not being eager to file a claim at their insurance company for cannabis related fires.

In [5, 64] additional health hazards of cannabis growing operations, not related to electricity, are noted. The most major health hazards are.

- Carbon monoxide poisoning due to bad functioning gas heaters or generators.
- Carbon dioxide poisoning due to plants using oxygen and releasing carbon dioxide.
- Carbon dioxide poisoning due to fertilizers.
- Water damages due to flooding and excessive moisture (including molt related problems).
- Contamination of water supply and air by chemicals used.

Given the fire hazards and the dangers listed above it is clear that illegal growing operations bring a danger to the public health.

Part II

Pattern Classification

Introduction

This part supplies the theoretical background of this thesis regarding the machine learning methods that are used for the classification of suspicious, cannabis related, patterns of electricity consumption.

First chapter 4 present a general overview of pattern classification, which special attention to time-series classification and dissimilarity methods. We also present our classification model used in this thesis, k-nearest neighbors, as a couple, well known distances metrics. In the following chapter, chapter 5, the Hankel based methods are presented. These methods present a relatively quick way for the comparison of time-series. The Hankel based methods are one of the focus points of this thesis and are therefor explained in detail. The last chapter of this thesis, chapter 6, the dynamic time warping method is discussed.

In this part we will focus solely on the theoretical background. The practical applications of the methods discussed in this part are discussed in the next part.

4

Pattern Classification

Pattern classification is the classic machine learning task of assigning a category, class, to some observation based on information present in the observation.

A single observation is a collection of features, i. e. individual properties of the phenomenons being observed. When all features are numeric, an observation can be represented by a *feature vector*. In this thesis we only consider observation consisting solely of numerical data.

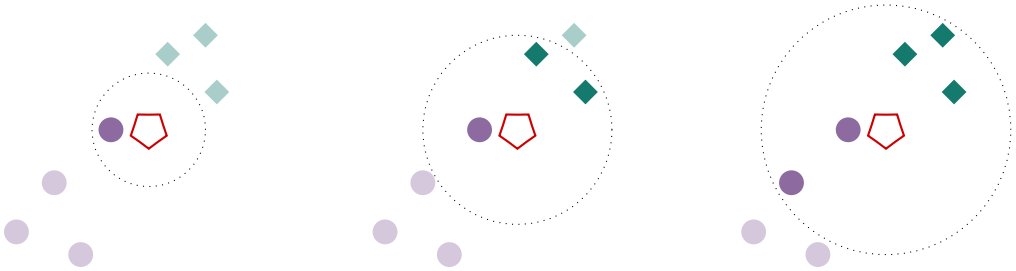
In this chapter general information regarding pattern classification is given, with special attention being paid to dissimilarity measures and time-series classification.

4.1 Classification

When the probabilistic structure of a classification problem is known, classification can often be done using a Bayesian classifier [24]. E. g. a new, unseen observation can be classified by calculating the probability that the observation belongs to a certain category. This is done using prior and class conditional probabilities and evidence that is present in the observation.

Unfortunately for many problems, knowledge about their probabilistic structure is incomplete. In most situation only some general knowledge about the classification problem is available, together with *training data*, a collection of observations.

Classifiers classify a new observation based on its similarity to a set of known observations. More formally stated, given a collection of N observations from



(a) When $k = 1$, the new observation ( ) is classified as B ( ).

(b) When $k = 3$, the new observation ( ) is classified as A ( ).

(c) When $k = 5$, the new observation ( ) is classified as A ( ).

Figure 4.1: An example of the influence of choice of k when using KNN. A new observation ( ) is shown together with known observations from class A ( ) and class B ( ). Lighter color shades are used to indicate elements that do not belong to the closest k observations. To further highlight the k points closest to the new observation ( ), a dashed circle is drawn with its origin in ( ), surrounding the k observations closest to ( ). Note that the dashed circle is solely a visual aid and does not represent any element in the classification algorithm.

C distinct classes the training set $\mathcal{X}$ is defined as

$$\mathcal{X} = \left\{ (\vec{\xi}^{\,\mathrm{i}},\ y_{\mathrm{i}}) \,|\, \vec{\xi}^{\,\mathrm{i}} \in \mathbb{R}^d,\ y_{\mathrm{i}} \in \{1, 2 \cdots, C\}_{i=1}^{N} \right\} \tag{4.1}$$

where

$$\begin{aligned} \vec{\xi}^{\,\mathrm{i}} &= \text{ feature vector of observation } i, \\ y_{\mathrm{i}} &= \text{ class label of observation } i, \\ \mathbb{R}^d &= \text{ feature space of the observations with } d \text{ features} \end{aligned}$$

and a new observation $\vec{\xi}^{\,\mathrm{new}}$, predict the label y_{new} based on similarities between $\vec{\xi}^{\,\mathrm{new}}$ and known observations in $\mathcal{X}$. In the next section the simple but strong classification algorithm K-Nearest Neighbors (KNN) is explained.

4.2 K-Nearest Neighbors

The nearest neighbor classification rule was introduced in [14] and is based on the notion *things that look alike must be alike* [13]. Practically, this means that, given a set of known observations and some measurement of similarity, a new observation $\vec{\xi}^{\,\mathrm{new}}$ is classified based on the known observation which is most similar, i. e. *nearest* to the new observation.

The nearest neighbor rule can be extended to the k-nearest neighbor rule, this rule classifies a new observation $\vec{\xi}^{\,\mathrm{new}}$ by assigning it the label most fre-

quently represented among the k nearest observations. In binary classification problems the choice of k is often an odd number to avoid ties.

A drawback of the KNN classification algorithm are the computational cost of classifying a new observation, as all distances between the known observations and the new observation have to be calculated and sorted.

4.2.1 Choice of k

When KNN is used as classification model the parameter k must be set. What value for k is appropriate depends on the characteristics of the problem. When the value for k is too low, the model can over-fit on the training data and will be sensitive to noise. A too large value might result in an under-fitted model.

An example of the influence of k is shown in fig. 4.1. Comparing the classification results in fig. 4.1 we can see that the choice of k influences the label assigned to a new observation.

The proper order of k is often identified using cross-validation for different values of k, although some heuristics exist in which k is often set proportional to the number of observations in the training-set[17]. More about the choice of k can be found in [34].

4.3 Distance and Dissimilarity Measures

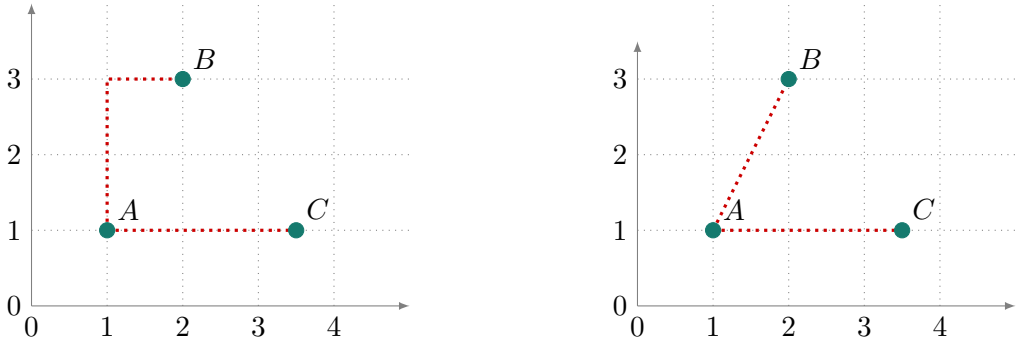
In the introduction of this chapter we stated '*classifiers classify a new observation based on its similarities with a set of known observations*'. Which begs the question, what is similarity and how can we measure it? In most classification algorithms a distance function is used. Most distance functions are of a special type of functions called *metrics*. A metric on a feature space $\mathbb{R}^d$ is a function $d(a, b) : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ adhering to the following properties [24, Chapter 4.6].

Non-negativity $d(a, b) \geq 0$, all distances between two points are greater or equal to zero.

Reflexivity $d(a, b) = 0 \Leftrightarrow a = b$.

Symmetry $d(a, b) = d(b, a)$

Triangle-inequality $d(a, b) + d(b, c) \geq d(a, c)$



(a) Using City Block distance $d_{\mathrm{C}}(A, B) = 3$ and $d_{\mathrm{C}}(A, C) = 2.5$.

(b) Using Euclidean distance $d_{\mathrm{E}}(A, B) = \sqrt{5} \approx 2.2$ and $d_{\mathrm{E}}(A, C) = 2.5$.

Figure 4.2: To illustrate the effect of the distance measure on classification, the closest point to A is calculated using both the City Block in a and Euclidean distances in b. B is the closest to A when Euclidean distance is used, while C is closest when City Block is used.

An often used distance metric is the Euclidean distance which is defined as

$$d_{\mathrm{E}}(\vec{\xi}^{\mathrm{a}}, \vec{\xi}^{\mathrm{b}}) = \left(\sum_{i=0}^{d} \left(\vec{\xi}_{\mathrm{i}}^{\mathrm{a}} - \vec{\xi}_{\mathrm{i}}^{\mathrm{b}} \right)^2 \right)^{\frac{1}{2}}. \tag{4.2}$$

Another well-known distance metric is the City Block distance, also known as the Manhattan distance, given by

$$d_{\mathrm{C}}(\vec{\xi}^{\mathrm{a}}, \vec{\xi}^{\mathrm{b}}) = \sum_{i=0}^{d} \left| \vec{\xi}_{\mathrm{i}}^{\mathrm{a}} - \vec{\xi}_{\mathrm{i}}^{\mathrm{b}} \right|. \tag{4.3}$$

Both the Euclidean and the City Block distance are instances of the more generalized Minkowski distance,

$$d(\vec{\xi}^{\mathrm{a}}, \vec{\xi}^{\mathrm{b}}) = \left(\sum_{i=0}^{d} \left(\vec{\xi}_{\mathrm{i}}^{\mathrm{a}} - \vec{\xi}_{\mathrm{i}}^{\mathrm{b}} \right)^n \right)^{\frac{1}{n}}. \tag{4.4}$$

Choosing n in eq. (4.4) to be 1 or 2 results in the City Block distance or Euclidean distance respectively. The Minkowski distance is equal to the City Block distance for $(n = 1)$ and equal to the Euclidean distance when $(n = 2)$.

The choice of distance function is important in classification problems. This is illustrated in fig. 4.2, which shows that the point closest to A varies for the two functions. Which distance function can be considered best in a classification problem depends on both the problem and the corresponding data. Speed,

simplicity, and sensitivity to noise or outliers can all be reasons to prefer one distance function above others.

4.3.1 Dissimilarity Functions

Consider a dataset which observations contain positions of people relative to the origin. I. e. the observations contains coordinates of people relative to a starting position. Given this dataset, what would be a suitable function if we want to classify people based on the *direction* they have traveled, invariant to the distance traveled. In this case a distance function would not work. Instead, one can apply a function which returns the angle between feature vectors. Note that the angle between two different vectors can be zero if the vectors have the same direction but different magnitudes. This means the angle function is not reflexive and therefor is not a metric.

Distance functions and the angle function from above are members of a group of functions called *dissimilarity functions*. Dissimilarity functions, as the name suggests, give some measurement of dissimilarity between objects which increases as objects become more different. Dissimilarity functions which are not a metric are called *dissimilarity measures*.

The output range of a dissimilarity measures does not have to be in the range $[0, \infty)$, meaning there can be a maximum dissimilarity; this is for example the case with the angle function, which has as range $[0, \pi]$.

4.4 Time-Series Classification

At the start of this chapter the feature vector was introduced. Often the order of measurements in the feature vector does not carry any meaning. When, for example, a feature vector contains measurements like age, height, and weight, order of the features does not matter, as long as it is consistent within the data set. This, however, does not hold for time-series.

Time-series consist of a sequence of measurements that are ordered in time, often sampled using a regular time-interval. Examples of time-series are sound and video recordings, heart-rate data, EEG data, stock exchanges, and tidal fluctuation. In most of the literature the term *time-series* is used as a catch all phrase that denotes any sequential dataset, including non time related data such as character- and DNA-sequences.

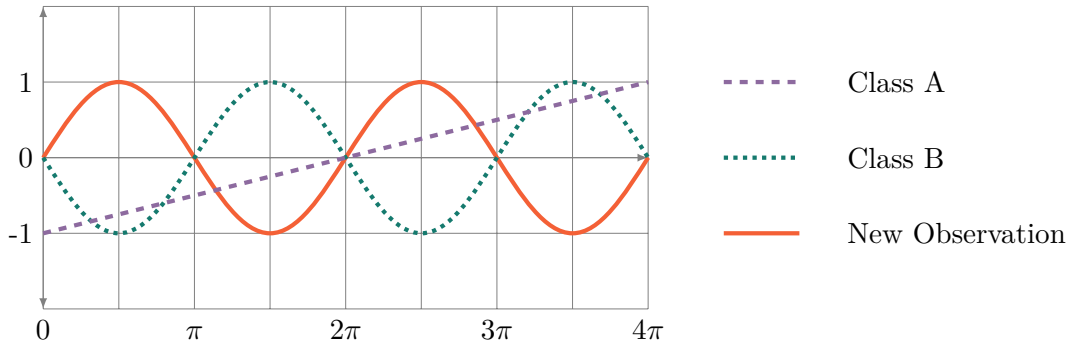


Figure 4.3: Plotted are three observations of time-series, two from which the label is known and a new observation. In the figure we see that the observation of class A contains a sinusoid and the observation of class B contains a linear signal. The new observation also contains a sinusoid, but with a phase shift. If the observations are compared using Euclidean distance, the linear signal from class A is closer to the new observation as the sinusoidal signal from class B, which seems counter intuitive.

4.4.1 The Misalignment Problem

Compared to non-sequential data, the classification of time-series is often more difficult. The main obstacles are caused by irrelevant data obscuring patterns which are highly predictive of a class and patterns which are shifted relative to each other.

Consider for example a sound-bite which starts and ends with background noise. When the irrelevant parts of the data are of different lengths for different elements in the dataset, the relevant parts are misaligned. This is known as the *misalignment problem*. This is illustrated in fig. 4.3. In the figure three time-series are plotted, two with a known label and a new observation. If we classify the new signal, class B would be an obvious choice, since both signals are sinusoids, although both are in a different phase. However, when using Euclidean distance we find $d_{\mathrm{E}}(\mathrm{A}, \mathrm{new}) = 38.0$ and $d_{\mathrm{E}}(\mathrm{B}, \mathrm{new}) = 50.1$. Thus, the new observation is classified as a linear signal, class A.

In the next chapter we present a Hankel based dissimilarity method as a solution to the misalignment problem. Using this method the dissimilarity between the two sinusoids is approximately zero while the dissimilarity between the linear signal and the sinusoid is 0.48.

5

Subspaces and Hankel Matrices

In chapter 4, section 4.4 we discussed the alignment problem with classifying time-series data, showing incorrectly aligned times-series can lead to high inner class distances when the wrong distance function is used. In this chapter we present a Hankel based dissimilarity measure which does not suffer from the alignment problem. We first give some theoretical background on Linear Time-Invariant (LTI) systems in section 5.1. Next we explain the concepts of subspaces in section 5.2 and how to measure dissimilarities between subspaces in section 5.3.

5.1 LTI Systems

Dynamical systems provide a way to model the temporal evolution of a data sequence. This section provides some background regarding LTI systems, which is the simplest dynamical system.

Given an ordered measurement sequence $Y = y_0, y_1, \dots, y_k$ with $y \in \mathbb{R}^d$, the temporal evolution of this sequence is modeled as a function of a low dimensional state vector $x_k \in \mathbb{R}^d$ that changes over time [53]. The LTI system is defined as

$$y_k = \mathbf{C}x_k \tag{5.1}$$

$$x_k = \mathbf{A}x_{k-1} + w_k, \tag{5.2}$$

Here y_k is the output of the linear *measurement equation*, x_k is the output of the linear *state equation*, the matrices $\mathbf{A}$ and $\mathbf{C}$ are constant over time, and w_k represents uncorrelated zero mean Gaussian measurement noise. A visual

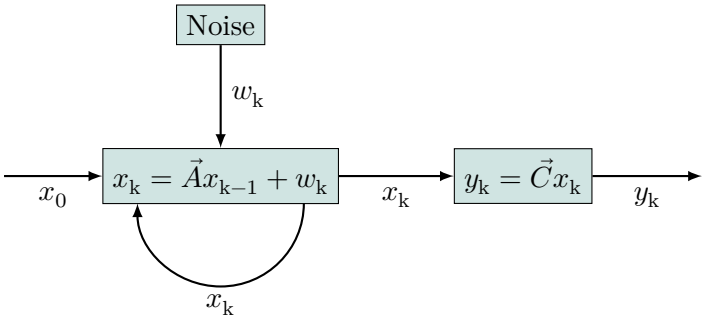


Figure 5.1: A visual representation of a LTI system. The state equation is initialized with x_0 and constant $\vec{A}$. The measurement equation is initilialized with constant $\vec{C}$. At each time-step k, x_k is determined in the state equation using its constant $\vec{A}$ and its previous output, x_{k-1} as input. In the measurement equation y_k is determined using the output of the state equation and its constant $\vec{C}$. The factor w_k represents uncorrelated zero mean Gaussian measurement noise.

representation of a LTI system is presented in fig. 5.1.

Given a measurement Y, one can estimate $\mathbf{A}$, $\mathbf{C}$, and the starting position x_0 to identify the corresponding LTI system for classification purposes. The identification of the triple is however a non convex problem[1] and thus, given a finite measurement y_k, a triple $(\mathbf{A}, \mathbf{C}, x_0)$ which can generate such a measurement is not guaranteed to be unique [53]. Consequently, system identification is computationally expensive and not robust, which makes it unsuitable for classification purposes. These problems can be avoided by using subspace identification on Hankel matrices associated with the measurement signals [71]. Here the key insight is that, given a dynamical system, all output measurements lie on a single subspace, assuming a noiseless output. This means that the subspace spanned by the columns of a Hankel matrix is equivalent to the subspace of the associated LTI system. Therefor the subspace spanning a LTI system can be found and used for classification purposes, without identifying the underlying LTI system.

[1] A non convex problem can have multiple local optimal solutions, making it hard to prove a found solution is the optimal, e. g. global, solution.

5.2 The Hankel Matrix

Given a sequence $Y = y_0, y_1, \dots, y_{\mathrm{m+n}}$ of order m, the associated Hankel matrix is given by

$$H_y = \begin{bmatrix} y_0 & y_1 & y_2 & \cdots & y_{\mathrm{n}} \\ y_1 & y_2 & y_3 & \cdots & y_{\mathrm{n+1}} \\ y_2 & y_3 & y_4 & \cdots & y_{\mathrm{n+2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{\mathrm{m}} & y_{\mathrm{m+1}} & y_{\mathrm{m+2}} & \cdots & y_{\mathrm{m+n}} \end{bmatrix}. \tag{5.3}$$

where m is the maximal order of the LTI system[74] and the columns of H_y contain all the subsequences of Y of length m and the anti-diagonals of H_y are constant.

Given that H_{y} is Hankel matrix associated with the output of the measurement in eq. (5.1) of the LTI system we can, in the absence of noise, $w_{\mathrm{k}} = 0$, rewrite eqs. (5.1) and (5.2) to express y_{k} in terms of the triple $(\mathbf{C}, \mathbf{A}, x_0)$:

$$\begin{aligned} y_{\mathrm{k}} &= \mathbf{C}x_{\mathrm{k}} \\ &= \mathbf{CA}x_{\mathrm{k-1}} \\ &= \mathbf{CA}^2x_{\mathrm{k-2}} \\ &= \cdots \\ &= \mathbf{CA}^t x_0. \end{aligned} \tag{5.4}$$

$$\tag{5.5}$$

Consequently H_y can be rewritten to

$$H_{\mathrm{y}} = \Gamma X, \tag{5.6}$$

where

$$\Gamma = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^m \end{bmatrix} \text{ and } X = \begin{bmatrix} x_0 & x_1 & \cdots & x_{\mathrm{k}} \end{bmatrix}. \tag{5.7}$$

Equation (5.6) shows that the columns of H_{y} and Γ span the same subspace regardless of the initial values of the LTI system. This means that given two measurements from the same LTI system, the smallest principal angle between the subspaces of the Hankel matrices is zero [53]. In other words, the subspace

angles between the Hankel matrices of two output measures can be used to identify whether these outputs could be produced by the same LTI system. Note that, since the relation between output signals and LTI systems is not unique. However, for classification purposes it is fair to assume that two Hankel matrices which share the same subspace belong to the same LTI system, and thus belong to the same class. In the next section methods for both the calculation and approximation of the subspace angles between two Hankel matrices are given.

5.2.1 Hankel Dimensions

In the section above we show how a Hankel matrix is constructed from a sequence $Y = y_0, y_1, \dots, y_{\mathrm{m+n}}$, using m for the number of rows and n for the number of columns. Although [74] states that m must be set to the maximum order of the LTI system, there is not much research done on how to find the right dimension of the Hankel matrices. In [53, 74] the number of rows and columns of the Hankel matrix are chosen such that the Hankel matrix is as square as possible, while in [52] the number of rows is set to be twice the number of columns. In all cases no definite argument is presented as to why the dimensions are chosen as such. In [56] the dimension of the Hankel matrices were choosing empirically based on classification performance using nearest neighbor for 35 different datasets, no direct correlations between the found dimensions and classification results where found.

5.2.2 Normalization of Hankel Matrices

A Hankel matrix H can be normalized with

$$\hat{H} = \frac{H}{\sqrt{\|H \cdot H^T\|_F}}, \tag{5.8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The Frobenius norm of a $m \times n$ matrix A is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{\mathrm{ij}}^2}. \tag{5.9}$$

5.3 Comparison of Subspaces

This section presents the calculations needed to obtain the subspace angles between two Hankel matrices, as well as two methods that give an approximation of the subspace. In line with [52, 73, 74] we assume that all Hankel matrices are normalized according to eq. (5.8).

5.3.1 Subspace Angle

Given two Hankel matrices $\hat{H}_p$ and $\hat{H}_q$ associated with the outputs of unknown LTI systems, we obtain a dissimilarity measure by calculating the subspace angle between the normalized matrices. This process is known as canonical correlations of linear subspaces, as it gives a measurement of the correlation between two subspaces. Given two subspaces F and G which dimensions are constraint such that $p = \dim(F) \geq \dim(G) = q \geq 1$, the smallest principal angle $\theta_1(F, G) \in [0, \frac{1}{2}\pi]$ is defined as

$$\theta_1 = \cos^{-1}\left(\max_{\vec{u}\in F}\max_{\vec{v}\in G} \vec{u}^T\vec{v}\right) \tag{5.10}$$

with $\|\vec{u}\|_2 = \|\vec{v}\|_2 = 1$.

The other principal angles are defined recursively. Let $F^{\perp}_{\vec{u}_k}$, $G^{\perp}_{\vec{v}_k}$ be the orthogonal complement of F and G with respect to $\vec{u}_k$ and $\vec{v}_k$ respectively. Then the principal angle θ_{k+1} is

$$\theta_{k+1} = \cos^{-1}\left(\max_{\vec{u}\in F^{\perp}_{\vec{u}_k}}\max_{\vec{v}\in G^{\perp}_{\vec{v}_k}} \vec{u}^T\vec{v}\right) \tag{5.11}$$

with $\|\vec{u}\|_2 = \|\vec{v}\|_2 = 1$.

This process continues until all principal angles of at least one of the subspaces are represented. A more in depth review on classification using subspace angles is given in [1, 6, 41, 88]. In the rest of the paper we will refer to this method as the *subspace angle method*.

5.3.2 Subspace Angle Approximation

The calculation of the subspace angle between two Hankel matrices is computational intensive. It can therefore be beneficial to use an alternative dissimilarity measure [52, 55, 73]. The following dissimilarity measure approximates

the difference between two subspaces

$$d_{\mathrm{A}}(\hat{H}_p, \hat{H}_q) = 2 - \|\hat{H}_p \cdot \hat{H}_p^T + \hat{H}_q \cdot \hat{H}_q^T\|_F. \tag{5.12}$$

In the rest of the thesis we refer to this method as the *approximation method*. The approximation of the subspace angle is based on the triangle inequality on the matrix norm which captures the alignment between two subspaces [52]. We can use the following geometric interpretation to understand how the triangle inequality works on the matrix norm.Consider a matrix as an operation transforming a space, then the norm of the matrix quantifies the maximum amount a unit vector is be stretched by this transformation. Given two Hankel matrices $\hat{H}_p$ and $\hat{H}_q$ which are normalized using eq. (5.8), we know that

$$\|\hat{H}_p \cdot \hat{H}_p^T\|_F = \|\hat{H}_q \cdot \hat{H}_q^T\|_F = 1.$$

If the subspaces of $\hat{H}_p$ and $\hat{H}_q$ are *aligned* and thus the subspace angle between two matrices is zero, we have

$$\|\hat{H}_p \cdot \hat{H}_p^T + \hat{H}_q \cdot \hat{H}_q^T\|_F = 2,$$

and thus

$$d_{\mathrm{A}}(\hat{H}_p, \hat{H}_q) = 2 - \|\hat{H}_p \cdot \hat{H}_p^T + \hat{H}_q \cdot \hat{H}_q^T\|_F = 0.$$

Similarly, given two Hankel matrices $\hat{H}_r$ and $\hat{H}_s$ whose subspace are not aligned, based on the triangle inequality we get

$$\|\hat{H}_r \cdot \hat{H}_r^T + \hat{H}_s \cdot \hat{H}_s^T\|_F < 2.$$

and thus

$$d_{\mathrm{A}}(\hat{H}_p, \hat{H}_q) = 2 - \|\hat{H}_r \cdot \hat{H}_r^T + \hat{H}_s \cdot \hat{H}_s^T\|_F > 0.$$

5.3.3 Approximation with Rotation

The subspace angle approximation as defined in eq. (5.12) is invariant to the direction in which the state changes due to the $H \cdot H^T$ operation, which results in a symmetric matrix. For some classification tasks, such as gesture modeling [74], the direction of change is of importance. To solve this, [74] introduces a

variant on the approximation method,

$$d(H_p, H_q) = 2 - ||\widetilde{H}_p + \widetilde{H}_q||_F. \tag{5.13}$$

Here $\widetilde{H}$ represents a Hankel matrix normalized according to

$$\widetilde{H} = \frac{H}{||H||_F}. \tag{5.14}$$

This method is referred to as the *rotation method*. Note that similar to the approximation method, the rotation method is also based on the triangle inequality.

6

Dynamic Time Warping

In the previous chapter we introduced Hankel based dissimilarity measures, which work under the presumption all signals are linear. DTW offers a method for the comparison of non-linear signals and is resistant to expansion or contraction of patterns [50]. The dissimilarity measure was originally introduced as a possible solution for speech recognition problems [75, 83], but is now utilized in many areas such as handwriting recognition tasks [8, 68]; the detection of structural damages by analyzing ultrasonic wave signals [22]; the classification of killer whale vocalizations [9]; sewer flow monitoring [25] and fault detection in swine wastewater treatment [45].

In this chapter we discuss the calculations of the DTW in section 6.1. We also discuss some variants in section 6.1.2.

6.1 Classical DTW

Given two observations $\vec{X} \in \mathbb{R}^m$ and $\vec{Y} \in \mathbb{R}^n$ with $X = x_1, x_2, \dots, x_m$ and $Y = y_1, y_2, \dots, y_n$ with the same sampling rate, but not necessary the same length, the fluctuation in timing can be represented as a sequence of tuples containing matched indices of $\vec{X}$ and $\vec{Y}$.

$$F = c_1,\ c_2,\ \dots,\ c_K$$

where c_k contains the indices from $\vec{X}$ and $\vec{Y}$,

$$c_k = (i_k,\ j_k)$$

with

$$i_k \in \{1, 2 \dots, m\}$$
$$j_k \in \{1, 2 \dots, n\}.$$

In other words, sequence F represents a mapping of the time axis of $\vec{X}$ onto that of $\vec{Y}$. This sequence is the *warping path*, also known as the *warping function* or *matching path*.

To be valid a warping path must meet the following three conditions.

1. The **boundary condition** requires alignment of the first and last elements of $\vec{X}$ and $\vec{Y}$. Thus, $c_1 = (1, 1)$ and $c_K = (m, n)$
2. The **monotonicity condition** enforces that matching cannot be done backwards in time. This means that given a match $c_k(i_k, j_k)$, for all matches c_l where $k < l$, $i_k \leq i_l$ and $j_k \leq j_l$ have to hold.
3. The **step size condition** dictates that F has a continuity, meaning that no element from $\vec{X}$ and $\vec{Y}$ can be omitted and all tuples in F must be distinct.

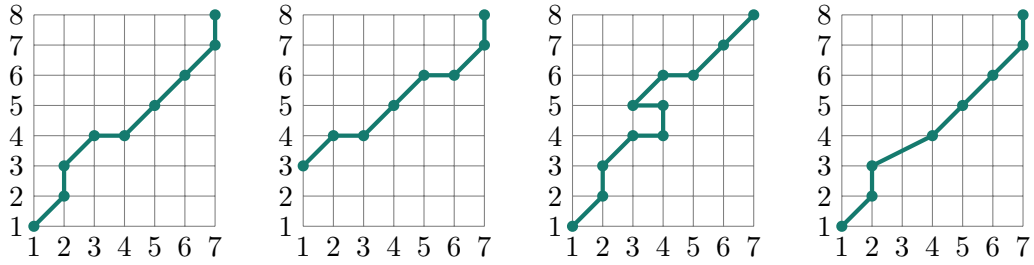
In fig. 6.1 four possible paths are illustrated between two observations. Here fig. 6.1a is a valid warping path, meeting all conditions. All other paths are breaking one of the conditions above. In fig. 6.1b the boundary condition is broken, in fig. 6.1c the monotonicity condition is not met, and fig. 6.1d gives an example of discontinuity.

When the two observations are linear and thus have no timing differences, the warping path is equal to the diagonal line $i = j$. When the timing differences grow, the warping path deviates from the line $i = j$. Thus, the warping path defines how the temporal axis of signals are locally stretched or compressed in order to reduce temporal differences between the observations.

In the next sections we give a dissimilarity measure between two signals for a given warping path and explain how this dissimilarity measure is used to obtain the optimal path.

6.1.1 Warping Path Dissimilarity

The warping path forms the basis of the DTW method and can be used to define a dissimilarity as follows. Given the warping path $F = c_0, c_1, \dots, c_K$



(a) Valid warping path.

(b) Boundary violation.

(c) Monotonicity violation.

(d) Continuity violation.

Figure 6.1: Illustrations on paths of index tuples for some sequences $\vec{X}$ of length 6 and $\vec{Y}$ of length 7. Given is one valid warping path in a and three invalid warping paths. In b the boundary condition is violated by starting at $(0, 2)$. In c monotonicity is broken by going from $(3, 4)$ to $(2, 4)$. In d continuity is broken by going from $(1, 2)$ to $(3, 3)$. Adapted from [26].

between two observations $\vec{X}$ and $\vec{Y}$ we can measure the distance between the tuples $c = (i, j)$:

$$d_{\mathrm{P}}(c) = d_{\mathrm{P}}(i, j) = \|x_i - y_j\|.$$

Using this point-to-point distance, we can calculate the dissimilarity between two signals for a given warping path as the summation of the distances of the warping path F with length K,

$$d_{\mathrm{dtw}}(F) = \sum_{k=0}^{K} d_P(c_k). \tag{6.1}$$

Although this formula is well-defined and symmetric, it is not positive definite and does not satisfy the triangle inequality [26]. Thus, similar to the Hankel based dissimilarity, DTW defines a dissimilarity measure and not a metric.

Alternatives exist from the dissimilarity above. For example in [15, 75] a weighted version is used. In the next section we give the classical method for obtaining the warping path using the dissimilarity function above.

6.1.2 Warping Path

A warping path F identifies the relative stretching/compressing of the temporal axes of observations. The optimal warping path, F^* it the path which gives the lowest dissimilarity using eq. (6.1), by warping the timing axes such that

the correspondence between the two signals is maximized. Note that there can be multiple warping paths with the lowest cost.

In this section we will present the classical method for finding the optimal warping path as given in [75, 83]. This method calculates $d_{\mathrm{DTW}}(F^*)$ using dynamic programming by defining a recursive definition of the shortest path from the tuple (i, j) to $(0, 0)$.

$$
\begin{aligned}
d_{\mathrm{c}}\left(0,0\right) &= d_{\mathrm{p}}\left(0,0\right) \\
d_{\mathrm{c}}(i,j) &= d_{\mathrm{p}}(i,j) + \min \begin{cases} d_{\mathrm{c}}(i-1,\ j-1) \\ d_{\mathrm{c}}(i-1,\ j) \\ d_{\mathrm{c}}(i,\ j-1) \end{cases}
\end{aligned} \tag{6.2}
$$

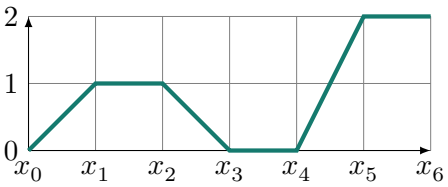
Given some observations $\vec{X}$ with length m and $\vec{Y}$ with length n, for which there exist some optimal path F^* we get $d_{\mathrm{c}}(m,n) = d_{\mathrm{DTW}}(F^*)$.

In fig. 6.2 an example is given of an optimal warping path between two signals, resulting in a dissimilarity of zero between the two observations.

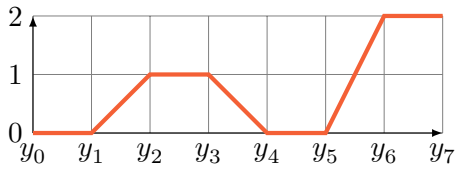
This dynamic approach has as advantage that the computational complexity of finding the optimal path is $\mathcal{O}(m \cdot n)$, while a brute-force approach has an exponential complexity.

Dynamic Time Warping Derivatives

In the section above the original algorithm is presented for finding the warping path. Several other methods exist which mainly differ in the definition of what constitutes a valid warping path, or which penalize deviations from the linear path. Other alternatives may apply an alternative distance measure $c(i, j)$ to calculate the point-to-point distances. In [9] some alternatives are shown for calculating the optimal warping path.



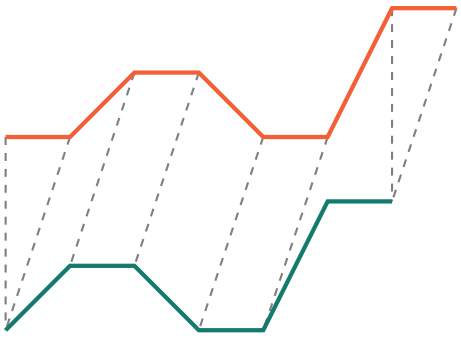
(a) Plot of X.



(b) Plot of Y.

y_7	2	1	1	2	2	0	**0**
y_6	2	1	1	2	2	**0**	0
y_5	0	1	1	0	**0**	2	2
y_4	0	1	1	**0**	0	2	2
y_3	1	0	**0**	1	1	1	1
y_2	1	**0**	0	1	1	1	1
y_1	**0**	1	1	0	0	2	2
y_0	**0**	1	1	0	0	2	2
	x_0	x_1	x_2	x_3	x_4	x_5	x_6

(c) Point-to-point distance matrix between X and Y using city block distance. The resulting warping path is shown using green highlighting.



(d) Visualization of the point-to-point matching between X and Y.

Figure 6.2: Given are two signals X given in a and Y given in b with length of 7 and 8 respectively. In c the point-to-point distance matrix between X and Y is given using City Block Distance. The warping path is shown in the point-to-point matrix by highlighting the corresponding cells. In d the warping is visualized by drawing a dashed line between matched points in the warping path.

Part III

Technical Solution

Introduction

In the previous part the theoretical background of the methods used in this thesis were given. We presented one classification model, k-nearest neighbors, and three different types of methods for calculating dissimilarity between observations. The euclidean distance metric, Hankel based dissimilarity measures, and dynamic time warping.

In this part we present the classification pipeline which will form the technical solution for the detection of cannabis related patterns of electricity consumption. First we will give a general overview of the complete classification pipeline in chapter 7, providing an overview of all steps in the pipeline from data acquisition, up until the validation of the used classification techniques. Next chapter 8 discussed how the energy usage data is recorded, stored and labeled to create a training set. The following chapter, chapter 9 discusses the pre-processing steps, while the last chapter, chapter 10, presents the classification model and methods used for its validation.

7 Classification Pipeline

In this chapter we introduce the classification pipeline constructed for the detection of suspicious signals on the electricity network. The pipeline starts with aggregation of the data at distribution stations and is finalized with cross validation to measure the performance of our classification model. The complete pipeline visualized in fig. 7.1.

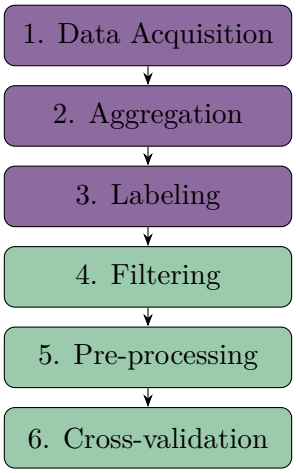


Figure 7.1: The complete classification pipeline. The purple steps, 1 to 3, are out of the scope of this thesis and done by ValueA and Coteq. The green steps, 4 - 6, are developed as part of this thesis.

We distinguish between two phases. In the first phase the training data is created. In the second phase this training data is used to test the various classification algorithms.

The first phase consists of step 1 to 3: data collection, aggregation, and labeling. These steps are outside the scope of this thesis and are done by a collaborative effort of Coteq and ValueA. The end results of the first phase is a

labeled dataset containing observations of suspicious and non suspicious electricity consumption.

The second phase consists of step 4 to 6: filtering, pre-processing, and cross-validation. In this phase the training dataset is cleaned up, normalized, and used to test the different classification techniques discussed in the previous chapters. In this phase the validation is important to achieve a better understanding of which technique provides a viable solution to the problem and which do not. The next chapters discuses the steps of the classification pipeline.

8 Electricity Usage Data

In this chapter we discuss the first three steps from the classification pipeline. Data acquisition, aggregation, and labeling. While these steps were done prior to the start of this thesis, they are discussed in this chapter, providing background information regarding the nature of the data and the labeling.

8.1 Data Acquisition

At Coteq usage data is gathered by an energy monitor located in transformation and distribution substations. The monitor used to gather the data is a 'MET*SyS* Total Demand Monitor'. The monitor allows for concurrent measurements on a maximum of twelve (three phase) connections. More information about the monitor system can be found at [42].

The MET*SyS* monitor system can measure both phases and neutral of the electric potential (volts), electric power (watts), and electric current (amperes). However, since this thesis is an exploratory research we will focus on the measurements of the electric current. This measurement is to be most discriminatory for the detection of cannabis growing operations according to fraud experts at Coteq.

During the monitoring process data is collected over thirty seconds intervals and aggregated using the average. The data is then stored in a PostgreSQL database [33], along with relevant meta-data such as time-stamp, cable number, and phase. During the monitoring process care is taken to ensure a constant sampling rate over all measurements. This is importance for dissimilarity measures as the DTW.

8.2 Data Aggregation

In the previous section we discussed how the electricity usage data is collected. The next step in the process is to split the data into 24 hour intervals, from 00:00 to 24:00. Thus, a single observation in the dataset consists of a measurements over a one-day period on one phase of a cable. Furthermore, the measurements are aggregated by averaging over fifteen, thirty and sixty minute intervals. This has multiple positive effects.

- The signal length is drastically reduced. For example, when an aggregation window of 15 minutes is used, the length of the observations is shrunken from 2880 to 96. This means classification is less expensive both in terms of computational power and memory.
- Averaging signals works as noise suppression by suppressing high peaks and low dips.
- In cases of missing data were the gaps in the time-series fall within the aggregation window, the missing data will be automatically be imputed by the averaging.

In the next section the use of aggregation as a noise reduction technique is further illustrated.

8.2.1 Noise Reduction

By aggregation the observations to a large time windows, noise in the form of high and low peaks are reduced. In fig. 8.1 the effect of aggregation are shown. When the original signal is aggregated using the average over sixty minute intervals, the resulting signal is much smoother, while the important characteristics are still present. Here the resolution, aggregation window, is important. When this window is too small, noise will still be presented. When the aggregation window is too large, defining characteristics of the signal, patterns in the signal that are important for classification, can get lost.

8.3 Manual Data Labeling

After the data aggregation, the data is labeled by the fraud experts at Coteq. This is done in a setup in which single, 24 hour observations are shown to the fraud expert in random order. No identifiers like date and place are shown,

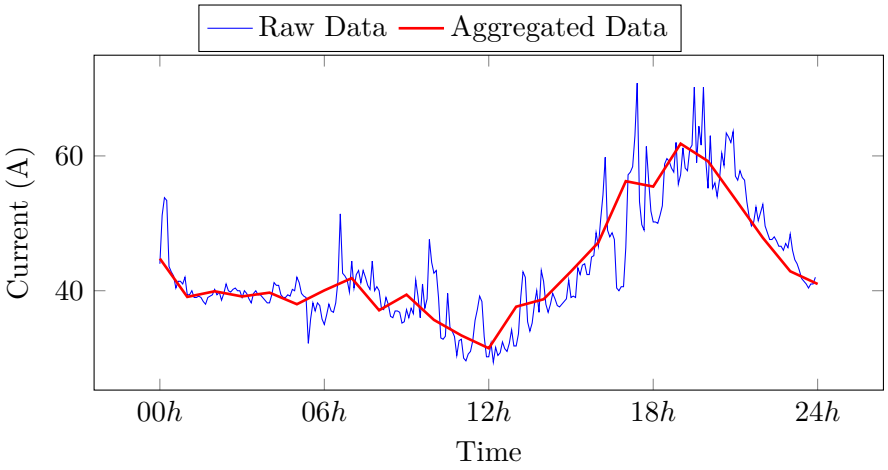


Figure 8.1: To illustrated the use of aggregation as noise reduction, a raw signal is plotted along with a version aggregated using the average over 60 minute intervals. The raw observation is plotted using a thin blue line, while the aggregated signal is plotted using a thicker orange line. We can see the aggregated version is much smoother as the original signal, while still following the overall trend.

but the fraud expert is able to switch between the fifteen, thirty and sixty minutes aggregation modes.

On the basis of the patterns visible in the observation the fraud expert will label an observation **suspicious** or **normal** to the best of his/her knowledge. In case of high uncertainty about the right label the observation may also be labeled **unknown**.

This way of labeling differs from the normal setup in which the fraud expert analyze the electricity usage. In a normal situation fraud experts look at data of multiple days in a row and take meta information like day of the week and location into consideration when classifying measurements. This is done to prevent false positives by normal, non-fraudulent activities. Example settings which can cause false positive are large offices, factories and schools, which shows, similar to growing operations, a pattern of high electricity usage for an extended long period each day. Such patterns are often not visible during weekends. Therefor it is import for classification accuracy to check if a certain patterns is present throughout the week, since offices are often closed during the weekend while cannabis operations are not. This check is unfortunately not possible in the setup used for the labeling of the data.

9 Pre-Processing

In the previous chapter we discussed how the electricity usage data is collected, aggregated, and labeled. These steps are part of the already existing framework of ValueA and Coteq. In this chapter we move into the scope of this thesis and discuss the pre-processing steps taken in the classification pipeline.

Pre-processing is a vital step in any machine learning solution. It ensures the available data is as clean as possible, contains little noise and irrelevant features. In case of missing data, imputation can also be a part of the pre-processing process.

The first pre-processing step in the classification pipeline is already discussed in section 8.2 of the previous chapter. In section 9.1 we discuss filtering of the dataset, removing all observations unsuitable for training purposes. The final pre-processing step, normalization, is discussed in section 9.2. Besides the pre-processing steps discussed in this chapter, principal component analyses and Fourier transformation were tested as ways for dimensionality reduction. Unfortunately these techniques did not work on the given dataset and are therefor not discussed.

9.1 Filtering

The first pre-processing step that is taken is the filtering of observations which are not suitable for classification purposes. There are three cases an observations is considered not suitable for classification.

- When it has too many consecutive missing values.
- When an observation was recorded on a changing day.

- When the fraud experts were not certain about activity in the observation.

All three cases are discussed below.

9.1.1 Missing Values

Some observations in the dataset have missing data. Due to circumstances, for example when a power outage occurs, no measurements were recorded for some period. When the duration of the reading failure is short, the missing data is automatically imputed when aggregating the data. However, when the failure period is larger as the aggregation window, the aggregated signal still contains missing data. In these cases the observations are removed from the dataset.

9.1.2 Changing Days

Since there is only one monitor system at Coteq available, the system is moved to a new distribution substation on a weekly basis. This is done to ensure complete coverage of their service area. During a changing day the MET*SyS* system is disconnected, moved to a new substation, and reconnected. However, due to the way the measurements are stored, aggregated and labeled, the changing procedures result in observations containing measurements from two locations. Therefore, all observations recorded on a changing day also removed from the dataset.

9.1.3 Unknown Labels

In some cases the fraud experts were not certain whether an observation should be considered suspicious or normal. In these cases the observations are given the label 'unknown'. Observations labeled as such are removed from the dataset.

9.1.4 Remaining Data

After removing observations with missing data, unknown labels, and observations recorded on changing days the resulting training set contains 4700 observations. Due to the nature of the dataset there is a large imbalance between the distribution of the classes. Of all remaining observations 89% is labeled

normal and 11% is labeled as suspicious. Note that we not reduce this imbalance by re-sampling the dataset. Since nearest neighbor is used as classification model, multiple occurrences of the same observations would have no effect. Would we use a classification model which requires training such as learning vector quantization, re-sampling is advisable.

9.2 Normalizing

After filtering, the next pre-processing step is the scaling of the data using the z-score. This normalization step is used to scale features such that they have the same properties as a normal distribution, having a zero means and unit variance. In general, given a feature ξ_i with a known mean μ_i and standard divination σ_i, the normalized feature is given by

$$\hat{\xi}_i = \frac{\xi_i - \mu_i}{\sigma_i}. \tag{9.1}$$

In this thesis all used observations consists of measurements of a single feature, electric current. However, in the normalization step we consider at what point of day a measurement is recorded. This is done since the average electricity usage changes with the time of day. In general people use more energy during noon than at midnight. Therefor we scale all measurements taken at time t using the mean and standard deviations of the measurements also recorded at that time.

In the following chapter, section 10.2, we discuss how the methods are validated by splitting the dataset into training and test datasets to simulate the performance in the real world. When we normalize the data we have to ensure that the mean ξ_i and standard deviation σ_i are estimated using only the data in the training set.

10

Classification Model and Validation

In the previous chapter we discussed the steps of the classification pipeline with dealt pre-processing of the data. In this chapter we discuss the remaining steps of the classification pipeline, the classification model and its validation.

10.1 Classification Model

In the previous part we presented one classification model, k-nearest neighbors and three different types of dissimilarity measures, Euclidean distance, Hankel based dissimilarities, and dynamic time warping.

In this thesis we focus on examining which dissimilarity measure is most suitable for the classification of electricity usage data. Therefor one of the simplest classification models is used, nearest neighbors. By using this classification model it is straightforward to compare the different dissimilarity measurements since the classification model does not have to be trained and does not require parameter tweaking.

Below we explain the reasoning behind each choice of dissimilarity measure.

1. Euclidean distance. The first dissimilarity measure that is used is the Euclidean distance metric discussed in section 4.3, eq. (4.2). Although the Euclidean distance is not very suitable for time-series classification, as discussed in section 4.4 of chapter 4, there are still a few reasons the choice was made to implement the distance measure. The Euclidean distance metric is a simple and quick method which is often used as performance baseline. Secondly, little is known about the classification models for electricity consumption patterns. For example, although likely, we are not certain the classification of electricity consumption suffers from the alignment problem. Using the Euclidean distance as a baseline provides insight into the nature of the electricity consumption data.

2. Subspace angle approximation method. This dissimilarity measure is chosen since it was proven successful in [56] when applied to a range of time-series datasets taken from the UCR time-series archive [12]. However, as the maintainers of the archive point out in [4] the series in the datasets are assumed to be of equal length, real valued and have no missing values. This somewhat optimistic assumptions do not always reflect problems in real world situations. Therefore, we like to test the Hankel based dissimilarity measure in a real life application as a succession on the research conducted in [56]. From the three Hankel based dissimilarity measures presented in chapter 5, only the subspace approximation method is used. We assume the classification problem is rotational invariant and since the subspace angle approximation method has a higher performance, in terms of accuracy and speed, compared to the exact subspace angle method, it is the obvious candidate.

3. Dynamic Time Warping dissimilarity. This dissimilarity measure is used as secondary baseline for the following two reasons. First, in contrast to the Hankel based method, the DTW method does not assume the classification problem to be linear. Secondly, the DTW method has been proven successful for many datasets [4] and has also been used as baseline in [4] in which a range of different time series classification techniques are tested.

To validate the performance of the above dissimilarity measure we will use cross validation. This process will be explained in the next section.

10.2 Cross Validation

To test the performance of a certain classification technique, available data is often split into a training set and a test set. The training set is then used to initialize the classification model and to predict the labels of the observations in the test set. By comparing the predicted labels by the known labels, one attains a measurement of performance, by observing how many observations are classified correctly. This validation process can give an indication in how well a model performs in real life.

When the dataset is split, the quantity of data available for training is reduced. Therefor there is less information to be used in classifying, which can make classification more difficult. Furthermore, when the data is split at random, you can not be sure the data in the two sets is representative of the real life situation. For example, when by accident the observations in the test set

are relative easy to classify, the model will have an unrealistic high performance which is not representative for the real life performance.

As a solution to the two problems above, k-fold cross validation is used. Here the dataset is divided into k sets. Next each set is used once as testing set, using the remaining sets as training data. As a result the classification model is validated a total of k times and all observations are classified one time. The performance of the classification model can then be represented by the average performance over the k folds. In this particular instance we use 10 fold cross validation.

When splitting the data in training and test sets one has to pay attention to how the data is distributed. Each set should be representative of the overall nature of the data. This means one has to ensure that the division of the labels is approximately the same over all sets. E. g. all ten sets should consist of approximately 89% normal observations and 11% suspicious observations.

To ensure different observations are well distributed over all test sets, the division of the observations between the sets is done at random. This ensures that all sets are a reflection of the overall dataset. Would the data not be distributed at random, but using the order as it is stored in the dataset, the different split could be a misrepresentation. In our case the observations are stored in chronological order. Would we not distribute the data at random, a single split could contain measurements from a single distribution station.

As mentioned in section 9.2 the normalization of observations should occur in every fold using only the training data of the particular fold to estimate the means and standard deviations.

10.3 Experiment

To test the performance of the nearest neighbor classification in combination with Euclidean distance, section 4.3, subspace angle approximation, section 5.3.2, and dynamic time warping chapter 6, all dissimilarity measures are tested using 10-fold cross validation as explained in section 10.2.

To establish the right aggregation window subspace angle approximation was used with an aggregation window of 15, 30, and 60 minutes. These window sizes correspond to the window sizes that are available to the fraud experts responsible for labeling the data. To establish the right dimensions of the Hankel matrices. A complete parameter sweep is done for all possible val-

ues.

To ensure that the ratio between normal and suspicious observations is similar among all splits, the two classes are divided separately and combined into one split. To ensure the same folds are used over the different experiments, the seed of the random generator is set to a constant.

10.4 Implementation Details

In this section some implementation details about the experiment and classification pipeline are presented. The data is stored in a PostgreSQL database [33]. The classification pipeline is done in Python using the following packages.

- Pandas is used to retrieve data from the database [58].
- Numpy is used for the support of matrix and vector operations [70].
- SciPi is used for a few numerical operations such as the scaling of the data [44].

In the implementation care is taken to ensure compatibility with the existing framework of ValueA. This means that the classification module can be used with the raw measurement data. To further ensure compatibility the API of the classification framework was made equivalent to the machine learning frameworks of the SciPi modules.

The modules for the cross validation, k-nearest neighbors, euclidean distance, and Hankel based distances were developed for this thesis. The DTW module was adapted from [59].

Part IV

Results and Conclusion

11 Results

In this chapter the results of the experiments as discussed in section 10.3 will be presented. We first present the performance of the three dissimilarity measures using the optimal settings for the various parameters in section 11.1. Next we give a more in depth look at the classification behavior when the parameters are varied.

11.1 Classification Accuracy

In the experiment cross validation is used to simulate how a classification model would perform in the real world. By comparing the output predictions of the models by the true labels, as determined by the fraud experts, we get an indication of the real world performance of the different models.

To rate the performance of the classification models we first look at the overall accuracy,

$$\text{Accuracy} = \frac{\#\text{ Correctly classified observations}}{\#\text{Total Observations}} \times 100.$$

The accuracies for the three dissimilarity methods are given in table 11.1.

Dissimilarity Measure	**Accuracy**
Euclidean distance	88 %
Subspace Angle Approximation	90 %
Dynamic Time Warping	94 %

Table 11.1: Best overall accuracy of the dissimilarity measures using nearest neighbor classification on 15 minute aggregation window. Accuracy is calculated by combining all results of the 10-fold cross validation.

		Prediction Labels	
		SUS	**NRM**
Actual Labels	**SUS**	True Positive	False Negative
	NRM	False Positive	True Negative

Table 11.2: The confusion matrix projects the real labels of the observations of the labels as predicted by the model. True Positives (TP) are observation correctly classified as containing suspicious patterns, True Negatives (TN) are observations correctly classified as not containing suspicious patterns. False Positives (FP) and False Negatives (FN) are observations which have gotten an incorrect prediction.

11.1.1 Precision and Recall

The accuracy expresses the performance as the percentage of observations correctly classified. However, if there is a large imbalance in the classes, the accuracy can give a skewed representation of the actual performance. Considering we have a dataset in which 89% of the observations are labeled *normal*, we could achieve an accuracy of 89 % with a classification model which labels all observations as *normal*. It is clear that, despite the high accuracy, such model has no value.

We consider the confusion matrix in table 11.2 to get better insight into the class-wise performance. Using the confusion matrix we define the metrics *precision* and *recall* as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \cdot 100 \tag{11.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \cdot 100 \tag{11.2}$$

The precision denotes the percentage of the signals which are labeled as suspicious which are actually suspicious. The recall denotes the percentage of suspicious signals which are actually detected by the system. The F-score expresses the harmonic average between the precision and recall.

$$\text{F-Score} = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \tag{11.3}$$

In section 8.2 we have discussed the aggregation of the measurement by av-

Dissimilarity Measure	Aggr. Win.	Precision	Recall	F-Score
Euclidean Distance	**15 min.**	46 %	54 %	50 %
	30 min.	45 %	51 %	48 %
	60 min.	47 %	51 %	49 %
Subspace angle approximation	**15 min.**	54 %	56 %	55 %
	30 min.	48 %	52 %	50 %
	60 min.	49 %	51 %	50 %
Dynamic Time Warping	15 min.	72 %	70 %	71 %
	30 min.	76 %	72 %	74 %
	60 min.	75 %	71 %	73 %

Table 11.3: The performance of the dissimilarity methods for three different aggregation windows. Scores are calculated by combining all results of the 10-fold cross validation. For the subspace angle method the optimal Hankel dimensions are found empirically for each aggregation window. The optimal window is highlighted using a bold font and different indentation.

eraging over a certain aggregation window. In table 11.3 the prediction and recall are given for each aggregation window and dissimilarity measure combination. Here we see that the Euclidean distance and subspace angle approximation both work best using a 15 minutes aggregation window, while the optimal window for DTW is 30 minutes. Overall we can see the best performance is reach by using DTW.

11.2 Classification Speed

The classification speeds were determined by taking the average time necessary to classify a new observation. The average timings are given in table 11.4.

Dissimilarity Method	Avg. prediction speed in sec.
Euclidean Distance	0.31
Subspace Angle approximation	0.46
Dynamic Time Warping	1.59

Table 11.4: Average time needed to classify one observation. All timings were recorded in the same environment. Note that timing might differ in relative to the used environment. Therefor speeds should mainly be considered relative to each other.

12 Discussion

In the previous chapter we presented the results of the experiment as discussed in section 10.3. In the experiment we mainly focus on finding the optimal dissimilarity measure for the given problem, classification of suspicious cannabis related patterns of electricity consumption. In table 11.1 the maximum classification accuracies are presented for the Euclidean distance, subspace angle approximation and Dynamic Time Warping. Looking at these results we see that all three methods are pretty good accurate of around 90%, with the DTW method being a positive outlier with an accuracy of 94%. The Euclidean distance has the lowest accuracy of the three methods. This is expected since, as discussed in section 4.4, the Euclidean distance is not shift invariant and thus less suitable for the classification of time-series.

When looking at the accuracies we do have to consider the bias in the dataset towards normal energy usage patterns. The classes in the dataset are very unbalanced with 89% of the observations have the label *normal*. This means that the accuracy alone cannot give a proper indication of performance. As already discussed in section 11.1.1 a model which classifies all observations all normal could achieve a 89% accuracy. This is as high as the accuracy of the Euclidean distance method.

To get a better insight into the performance we look at the precision and recall. By doing so we focus on the performance regarding the classification of *suspicious* signals. This is done since we prefer a false positive over a false negative. This because the cost of missing a suspicious signal is higher compared to the cost of further researching a false positive. For the same reasons we also prefer a high recall above a high precision. Note that there is a limit to choosing false positives above false negatives and the right balance between the two

should be determined in practice.

When we look at the F-scores given in table 11.3 we can see that the performance of the subspace angle approximation is 5% higher as that of the Euclidean distance. In terms of relative improvement the subspace angle approximation gives a 1.1 performance increase compared to the Euclidean distance. This is in line with the results of the previous research in [56] in which an average increase of 1.08 was reached over 35 datasets. Thus, the performance improvement the Hankel based method brings compared to the Euclidean distance is as good as expected. However, this increase is small compared to the improvement that is achieved by using DTW, which performs almost 1.5 times better. Concerning the aggregation window sizes we see that the DTW works better in combination with a larger aggregation size compared to the other two methods. This is positive, since a larger aggregation window results in a smaller signal length and thus a faster classification.

The high difference in performance between the subspace angle approximation and DTW might be explained by the data. As explained in section 5.3.1, the Hankel based methods assume that the patterns in the data are linear, while the DTW method assumes patterns in the data might be warped. The results suggest that the latter is probably a better assumption as the former.

In table 11.4 the average time for the classification of one observation is given. We can see that both the Euclidean distance and the subspace angle approximation have similar classification speed. This is not surprising, since the goal of the Hankel based dissimilarity is to have a quick method for the comparison of time-series. The DTW method is a factor three slower. This is expected, since finding the warping path is a much more involved optimization problem instead of a direct calculation.

12.0.1 Training Data

In the experiments we attempt to make the classification results representative of real world performance by means of cross validation. However, given the origin of the dataset and its labels we have to make a few remarks regarding the validity of the results. Since the data is labeled by the fraud experts of Coteq, we have no real data on the *true-ness* of the labels, e. g. a measurement being suspicious does not guarantee that a cannabis operation is present. We therefor have to remark that the performance of the classification model is mostly representative of how well the model can emulate the decision of the

fraud experts. Given the remark above we have to remember the goals of the classification model, to automate the detection of cannabis related patterns of consumption. This goals can still be reached if the system can be used to replace part of the manual labor.

13

Conclusion And Future Work

In the Netherlands illegal cannabis growing operations are often associated with electricity fraud. To avoid detection and paying high energy bills, growers often steal electricity by circumventing or manipulating power meters. Cannabis related electricity fraud causes a yearly total financial loss of 200 million euro. Furthermore, due to faulty connections and high energy usage, growing rooms bring a danger to the public safety.

At the Dutch DNO Coteq, fraud experts try to detect cannabis related fraudulent behavior. One of the methods used to identify cannabis growing operations is the manual inspection of electricity consumption data. Since the cultivation of cannabis is associated with a typical electricity consumption, it is possible to detect the presence of a growing operation.

With this thesis we present a solution for the automated detection of cannabis related patterns in electricity consumption. The automated detected can be used to lower the workload, reduce the repetitive nature of the manual inspection of electricity usage data. Automation also allows for the upscaling of measurement locations.

In this thesis we use the classification model nearest neighbors together with three dissimilarity measure, Euclidean distance, subspace angle approximation, and Dynamic Time Warping. The results show that the euclidean distance is not suitable for the classification of suspicious signals and should not be used in this context. Improvement can be made by employing the shift-invariant Hankel based dissimilarity subspace angle approximation. This method provides an improvement on the Euclidean distance method while stile maintaining fast classification speed. However, the highest improvement are made by using DTW. Using this dissimilarity measure results in a high overall accuracy

and a sufficiently high detection rate of suspicious signals. Although the DTW warping is computationally more expensive as the Hankel based dissimilarity, the classification improvement is high enough to validate the slower speed.

Overall we can conclude that the automated detection of cannabis related patterns of consumption is possible, with the DTW dissimilarity measure giving the best performance of the tested methods.

13.1 Future Work

This thesis presented a first working prototype for the automatic detection of cannabis related patterns of electricity consumption. As with all prototypes, improvements can be made to improve accuracy, speed and overall performance. In this section we give a few recommendations for future improvements of the classification model.

In this thesis we have tested three dissimilarity methods, of which the DTW method clearly has the best performance. Many variants on the DTW method exists which can provide improvement both in terms of speed an accuracy. In [9] a few alternative methods of determining the warping path are presented.

In [84] a method is presented which creates signal prototypes by combing observations using warp averaging. This could present a feasible way to reduce the number of elements in the dataset, which can greatly reduce the computational power. By creating a prototypical signal one can also gain more insight into what constitutes a suspicious pattern.

In the experiment and results we have made the assumption the labels in the dataset were correct. However, as already discussed in section 8.3 the process used for the manual labeling of the data was not ideal, the trueness of the labels depends on the knowledge of the fraud experts and the process allows for human error in the labeling. We therefor have to consider the possibility that the some observations in the dataset might not have the correct label. Real life performance might be better or worse as the results presented in this thesis. Therefor, when the classification system is deployed, improvement in the labeling can be made by continuously (re-)evaluating the labels of current and new observations. Further improvement can be made by retroactively labeling data as suspicious when a growing operations are caught. This way the trustworthiness of the suspicious labels can be increased. In addition, this can lead to the discovery of cannabis related patterns yet unidentified.

Bibliography

[1] P. A. Absil, A. Edelman, and P. Koev. "On the largest principal angle between random subspaces". In: *Linear Algebra and Its Applications* 414.1 (2006), pp. 288–294. ISSN: 00243795. DOI: 10.1016/j.laa.2005.10.004.

[4] A. Bagnall et al. "The Great Time Series Classification Bake Off: a Review and Experimental Evaluation of Recent Algorithmic Advances". In: *Data Mining and Knowledge Discovery* Online First (2016).

[5] Dutch Safety Board. *Environmental safety of cannabis grow rooms*. Tech. rep. Den Haag: Safety Dutch Board, 2018.

[6] Jeroen Boets et al. "Clustering Time Series, Subspace Identification and Cepstral Distances". In: *Communications in Information & Systems* 05.1 (2005), pp. 69–96.

[7] F Bovenkerk, Willemien I M Hogewind, and Nushin C Milani. *Hennepteelt in Nederland : het probleem van de criminaliteit en haar bestrijding*. Dutch. Dutch, [Cannabis Growing in the Netherlands: the problems of organized crime and the war on drugs]. Willem Pompe Instituut voor Strafrechtswetenschappen, Jan. 2002. ISBN: 9067203076 9789067203074.

[8] Carlo Di Brina et al. "Dynamic time warping: A new method in the study of poor handwriting". In: *Human Movement Science* 27.2 (2008), pp. 242–255. ISSN: 0167-9457. DOI: https://doi.org/10.1016/j.humov.2008.02.012.

[9] Judith C. Brown and Patrick J. O. Miller. "Dynamic time warping for automatic classification of killer whale vocalizations". In: *The Journal of the Acoustical Society of America* 119.5 (2006), pp. 3434–3434. DOI: 10.1121/1.4808856.

[10] Madalina Mihaela Buzau et al. *Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning*. 2018. DOI: 10.1109/TSG.2018.2807925.

[12] Yanping Chen et al. *The UCR Time Series Classification Archive*. www.cs.ucr.edu/~eamonn/time_series_data/, an updated version is also available at http://timeseriesclassification.com/. July 2015.

[13] T. Cover. “This Week’s Citation Classic: Cover T.M. & Hart P.E. Nearest neighbor pattern classification”. In: *Current Contents* 13.1 (1982), p. 20.

[14] T. Cover and P. Hart. “Nearest neighbor pattern classification”. In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967), pp. 21–27. ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1053964.

[15] VB Deecke and Vincent M. Janik. “Automated categorization of bioacoustic signals: avoiding perceptual pitfalls”. English. In: *Journal of the Acoustical Society of America* 119 (Jan. 2006), pp. 645–653. ISSN: 0001-4966. DOI: 10.1121/1.2139067.

[16] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni. “Support vector machine based data classification for detection of electricity theft”. In: *2011 IEEE/PES Power Systems Conference and Exposition*. Mar. 2011, pp. 1–8. DOI: 10.1109/PSCE.2011.5772466.

[17] L. Devroye, L. Györfi, and G. Lugosi. “A Probabilistic Theory of Pattern Recognition”. In: Springer, 1996. Chap. 26, pp. 452–455.

[22] A. C. S. Douglass and J. B. Harley. “Dynamic Time Warping Temperature Compensation for Guided Wave Structural Health Monitoring”. In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 65.5 (May 2018), pp. 851–861. ISSN: 0885-3010. DOI: 10.1109/TUFFC.2018.2813278.

[23] T. van Dril et al. *Energietrends 2012*. Dutch, [Energy Trends 2012]. Nov. 2012.

[24] Richard O. Duda, Peter E. Hart, and David G. Stork. “Pattern Classification (2Nd Edition)”. In: New York, NY, USA: Wiley-Interscience, 2000. Chap. 2. ISBN: 0471056693.

[25] D.J. Dürrenmatt, D. Del Giudice, and J. Rieckermann. “Dynamic time warping improves sewer flow monitoring”. In: *Water Research* 47.11 (2013), pp. 3803–3816. ISSN: 0043-1354. DOI: https://doi.org/10.1016/j.watres.2013.03.051.

[26] “Dynamic Time Warping”. In: *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, 2007. Chap. 4, pp. 69–84. ISBN: 978-3-540-74048-3. DOI: 10.1007/978-3-540-74048-3_4.

[29] J. Gerdes, S. Marbus, and M. Boelhouwer. *Energietrends 2014*. Dutch, [Energy Trends 2014]. Sept. 2014.

[30] J. Gerdes, S. Marbus, and M. Boelhouwer. *Energietrends 2016*. Dutch, [Energy Trends 2016]. Sept. 2016.

[33] The PostgreSQL Global Development Group. *PostgreSQL*. Online; accessed 08-08-2018.

[34] Peter Hall, Byeong U. Park, and Richard J. Samworth. "Choice of neighbor order in nearest-neighbor classification". In: *Annals of Statistics* 36.5 (2008), pp. 2135–2152. ISSN: 00905364. DOI: 10.1214/07-AOS537.

[35] Wenlin Han and Yang Xiao. "NFD : A Practical Scheme to Detect Non-Technical Loss Fraud in Smart Grid". In: (2014), pp. 605–609.

[41] Jiaji Huang, Qiang Qiu, and Robert Calderbank. "The Role of Principal Angles in Subspace Classification". In: *IEEE Transactions on Signal Processing* 64.8 (2016), pp. 1933–1945. ISSN: 1053587X. DOI: 10.1109/TSP.2015.2500889. eprint: 1507.04230.

[42] I-PROSyS. *MEtSyS Total Demand Monitor*. Online; accessed 08-08-2018.

[44] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. Online; accessed 24-07-2018. 2001.

[45] B.H. Jun. "Fault detection using dynamic time warping (DTW) algorithm and discriminant analysis for swine wastewater treatment". In: *Journal of Hazardous Materials* 185.1 (2011), pp. 262–268. ISSN: 0304-3894. DOI: https://doi.org/10.1016/j.jhazmat.2010.09.027.

[46] Petr Kadurek et al. "Theft detection and smart metering practices and expectations in the Netherlands". In: *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)* (Oct. 2010), pp. 1–6. DOI: 10.1109/ISGTEUROPE.2010.5638852.

[50] JB Kruskal and Mark Liberman. "The symmetric time-warping problem: From continuous to discrete". In: *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Jan. 1983. Chap. 4, pp. 125–161.

[52] Binlong Li, Octavia I. Camps, and Mario Sznaier. "Cross-view activity recognition using Hankelets". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2012), pp. 1362–1369. ISSN: 10636919. DOI: 10.1109/CVPR.2012.6247822.

[53] Binlong Li et al. “Activity recognition using dynamic subspace angles”. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (2011), pp. 3193–3200. ISSN: 1063-6919. DOI: 10.1109/CVPR.2011.5995672.

[55] Liliana Lo Presti et al. “Hankelet-based dynamical systems modeling for 3D action recognition”. In: *Image and Vision Computing* 44 (2015), pp. 29–43. ISSN: 02628856. DOI: 10.1016/j.imavis.2015.09.007.

[56] S.J. van Loon. “Comparison of Hankel Based Similarity Metrics for Time-Series Classification”. In: Unpublished Research Internship Paper. Nov. 2017.

[57] Daisuke Mashima and Alvaro A. Cárdenas. “Evaluating Electricity Theft Detectors in Smart Grid Networks”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7462 LNCS. 2012, pp. 210–229. ISBN: 9783642333378. DOI: 10.1007/978-3-642-33338-5_11.

[58] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 51–56.

[59] Alex Minaar. *Time Series Classification and Clustering with Python*. Onlinel accessed 05-08-2018.

[60] Mohammad Mohammadi. “Hankel matrices for use in Learning Vector Quantization”. Master Thesis. Hochschule Mittweida, 2016.

[62] J. Nagi et al. “NTL detection of electricity theft and abnormalities for large power consumers in TNB malaysia”. In: *Proceeding, 2010 IEEE Student Conference on Research and Development - Engineering: Innovation and Beyond, SCOReD 2010* SCOReD (2010), pp. 202–206. DOI: 10.1109/SCORED.2010.5704002.

[64] Netbeheer Nederland. *Factsheet Energiediefstal*. Online. Dutch, [Fact Sheet Energy Theft]. https://www.netbeheernederland.nl/dossiers/energiediefstal-13/documenten, May 2015.

[65] Netbeheer Nederland. *Veiligheidsrisicos Hennepplantages*. Online. Dutch, [Safety Risks of Cannabis Growing Operations]. https://www.netbeheernederland.nl/dossiers/energiediefstal-13/documenten, Jan. 2012.

[66] Netbeheer Nederland. *Gedragscode Verwerking van Persoonsgegevens door Netbeheerders in het kader van Installatie en Beheer van Slimme Meters bij Kleinverbruikers*. Dutch, [Code of Conduct Regarding Processing of Personal Data by Dutch Network Operators in the Context of the Installation and Management of Smart Meters for Consumers]. 2012.

[67] Government of the Netherlands. *Toleration policy regarding soft drugs and coffee shops*. Visited April 2018. Apr. 2018. URL: https://www.government.nl/topics/drugs/toleration-policy-regarding-soft-drugs-and-coffee-shops.

[68] Ralph Niels, Louis Vuurpijl, and Lambert Schomaker. "Automatic Allograph Matching in Forensic Writer Identification." In: *International Journal of Pattern Recognition and Artificial Intelligence* 21 (Feb. 2007), pp. 61–81.

[69] OFGEM. *Tackling Electricity Theft - The way forward*. Tech. rep. 2014, pp. 1–61.

[70] Travis Oliphant. *A guide to NumPy*. 2006.

[71] Peter Van Overschee and Bart De Moor. "Subspace algorithms for the stochastic identification problem". In: *Automatica* 29.3 (1993), pp. 649–660. ISSN: 0005-1098. DOI: http://dx.doi.org/10.1016/0005-1098(93)90061-W.

[72] David J. Potter and Paul Duncombe. "The Effect of Electrical Lighting Power and Irradiance on Indoor-Grown Cannabis Potency and Yield". In: *Journal of Forensic Sciences* 57.3 (2012), pp. 618–622. ISSN: 00221198. DOI: 10.1111/j.1556-4029.2011.02024.1x.

[73] L. Lo Presti and M. La Cascia. "Using Hankel matrices for dynamics-based facial emotion recognition and pain detection". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2015, pp. 26–33. DOI: 10.1109/CVPRW.2015.7301351.

[74] Liliana Lo Presti et al. "Gesture Modeling by Hanklet-based Hidden Markov Model". In: *Accv2014* (2014), pp. 1–17. DOI: 10.1007/978-3-319-16811-1_35.

[75] H. Sakoe and S. Chiba. “Dynamic programming algorithm optimization for spoken word recognition”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (Feb. 1978), pp. 43–49. ISSN: 0096-3518. DOI: 10.1109/TASSP.1978.1163055.

[76] Stichting Salvage. *Hennepteelt oorzaak één op vijftig grote branden*. Dutch, [Cannabis Growing Cause of One out of Fifty Large Fires]. Visited January 2018. Mar. 2015. URL: http://www.stichtingsalvage.nl/stichting-salvage/nieuws-en-persberichten/verbond-hennepteelt/.

[77] Josif V. Spirić, Slobodan S. Stanković, and Miroslav B. Dočić. “Identification of suspicious electricity customers”. In: *International Journal of Electrical Power and Energy Systems* 95.September 2017 (2018), pp. 635–643. ISSN: 01420615. DOI: 10.1016/j.ijepes.2017.09.019.

[78] Marcel Toonen, Simon Ribot, and Jac Thissen. “Yield of illicit indoor cannabis cultivation in The Netherlands”. In: *Journal of Forensic Sciences* 51.5 (2006), pp. 1050–1054. ISSN: 00221198. DOI: 10.1111/j.1556-4029.2006.00228.x.

[83] T. K. Vintsyuk. “Speech discrimination by dynamic programming”. In: *Cybernetics* 4.1 (Jan. 1968), pp. 52–57. ISSN: 1573-8337. DOI: 10.1007/BF01074755.

[84] Kongming Wang, Henri Begleiter, and Bernice Porjesz. “Warp-averaging event-related potentials”. In: *Clinical Neurophysiology* 112.10 (2001), pp. 1917–1924. ISSN: 1388-2457. DOI: https://doi.org/10.1016/S1388-2457(01)00640-X.

[88] P. Zhu and A. V. Knyazev. “Angles between subspaces and their tangents”. In: *Journal of Numerical Mathematics* 21.4 (2013), pp. 325–340. ISSN: 15702820. DOI: 10.1515/jnum-2013-0013.