Booklets

Research & Development & Innovation

25

Canal de Isabel II

A Pattern Recognition System for the Identification of Residential End Uses of Water

Authors

Pedro Luis Peñalver, Pablo García Rubí, Vanesa Pérez Salas
José Antonio Sánchez del Rivero, Roberto Díaz Morales, Julia Lastra García, Sergio García Caso

Project Direction

Juan Carlos Ibáñez Carranza

This documentis an English translation ofthe original study published by Canal de Isabel II under the title:
*"**Sistema de reconocimiento de patrones para identificación de usos finales del agua en consumos domésticos**"*

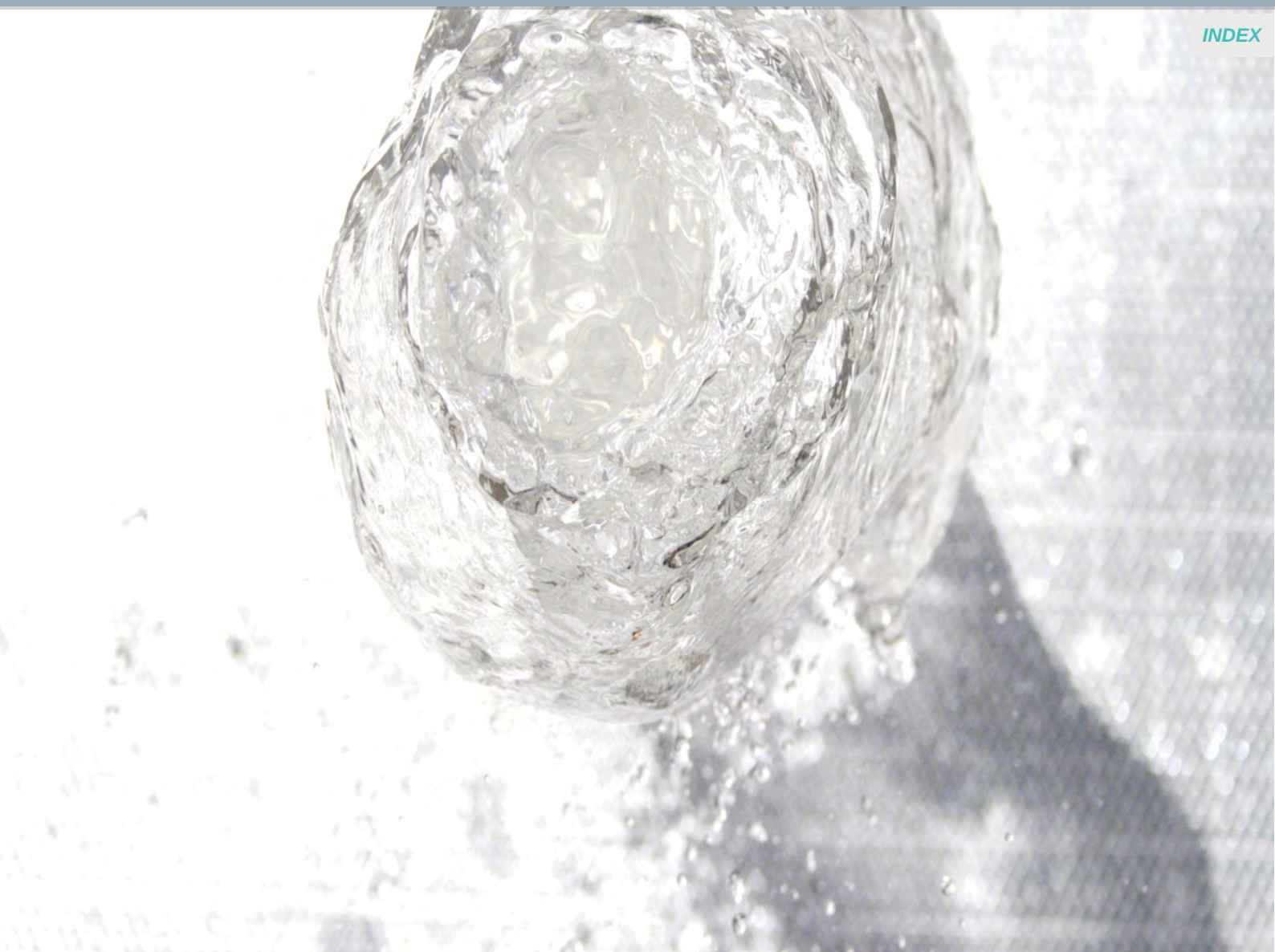
Print ISSN: 2254-8955
Electronic ISSN: 2340-1818
Legal Deposit (Spain): M-28367-2017

Canal de Isabel II

25

A Pattern Recognition System for the Identification of Residential End Uses of Water

INDEX



Liability Exclusion

The statements included in this document reflect the authors' opinions, and not necessarily those of Canal de Isabel II.

Neither Canal de Isabel II nor the authors of this document accept responsibility for any loss or damage that might be suffered by any institution or individual basing their actions on the contents of this document or the opinions expressed by the authors.

INDEX

Presentation

The collection Booklets of Research, Development & Innovation ofCanal de Isabel II are part of the vision of the company's knowledge management and the development of its 2017-2020 R&D&I Strategy.

These Booklets represent an element for diffusion of projects and initiatives developed and promoted by the company, and aim at innovation in areas related to the water services in an urban environment.

They deal with the problems tackled by each project as well as the results obtained. The aim of publishing these Booklets is to share experience and knowledge with the entire water industry sector, with the scientific community and with all those who work in the fields of research and innovation. With these publication, what it is hoped is contribute the improvement and efficiency in water management and, as a result, make it possible to offer a better service to the citizens.

The titles published in the series to date are shown in the following table.

INDEX

BOOKLETS OF RESEARCH, DEVELOPMENT & INNOVATION PUBLISHED

Collection Number	***Research, Development and Innovation Booklets published***
1	Transferences of Water Rights between Urban and Agrarian Demands. The case of the Community of Madrid
2	Identification of Hydrometeorological Runs and Tendencies within the scope of the Canal de Isabel II system
3	Contribution of Canal de Isabel II to the International Demand Management Project (IDMF)
4	Microcomponents and Explanatory Factors on Domestic Water Consumption in the Comunidad de Madrid
5	Virtual Water and Hydrological footprint in the Comunidad de Madrid
6	Study on the saving potential of water for residential uses in the Comunidad de Madrid
7	Potentials of efficiency in using dishwashers in the Comunidad de Madrid
8	Accuracy in the measurement of individual water consumption in the Madrid Region
9	Research project to define and assess the applicability of a Bioassay Test to determine the toxicity of water using Zebra Fish embryos
10	Water Use Efficiency in Gardening in the Region of counited de Madrid
11	Remote sensing techniques and geographical information systems for assessing water demand for outdoor uses in the Comunidad de Madrid
12	Cyanotoxin Dynamics Study in two of the Canal de Isabel II supply reservoirs in the autonomous region of the Comunidad de Madrid
13	Development of a validation, estimation and prediction of hourly consumption by sector, for the distribution network of Canal de Isabel II
14	Monitoring of the consolidation urban development in the Comunidad de Madrid using remote sensing techniques
15	Experiences in the recovery of phosphorus from wastewater, in the form of Struvite, at Canal de Isabel II
16	Integration of weather forecasting in the management modules supply system of Canal de Isabel II, via daily contributions models
17	Improvement in forecast capacity of monthly and seasonal runoff in the scope of Canal de Isabel II
18	Inflow of nutrients from the basin to Pinilla reservoir. Effect on the eutrophication process
19	A new criterion for calculating urban sewage flows
20	Idea Management at Canal de Isabel II Gestión: The GENYAL Experience

INDEX

Collection Number	***Research, Development and Innovation Booklets published***
21	Research on measuring techniques for subsidence related to groundwater exploitation
22	Precipitation patterns in the basins of the Lozoya and adjacent rivers
23	Observability study for hydraulic state estimation of the sectorised supply network
24	Study of failure causes and modes in pipes, service connections, and water meter assemblies in the Comunidad de Madrid

INDEX

TABLE OF CONTENTS

Executive Summary

Technical Data Sheet

Project title	A Pattern Recognition System for the Identification of Residential End Uses of Water
Research line	Assurance of balance between availability and demand
Areas involved at Canal de Isabel II	Subdirección I+D+i
External participation	Exeleria, Treelogic
Aim and justification of the Project	To develop an automatic system to identify the end uses of water in the different residential applications, based on the signals recorded by precision meters, using advanced methodologies of pattern recognition and supervised signal classification, such as Artificial Neural Networks (ANN), statistical methods, and others.
Contribution to the state of the art	It presents its own mathematical algorithm for the automatic transformation of volumetric pulses into instant flows and, the identification of events associated with different residential uses, allowing each one to be characterized and quantified. It develops two methodologies for classifying events, one based on support vector machines (SVM) and another on Artificial Neural Networks (ANN), analyzing and comparing the results obtained with both methods. The developed computer application allows the massive treatment of data from different meters and dates, reducing the operator's intervention to selecting the data to be processed.
Summary of the Project development and outstanding milestones	• Compilation of the readings from 375 volumetric meters with pulse emitter, with a reading precision of 1 and 0.1 litres. • Formulation of a mathematical algorithm for the transformation of pulse readings into time series of flows. • Development of a computer application (VBA on Access) for massive treatment of data from meter readings, for the transformation of pulses into flows. • Creation of a methodology to identify events from the series of flows obtained, and the development of a second computer module in VBA for its automation. • Development of two computer procedures for massive event classification (labelling) according to end uses, one based on Support Vector Machines (SVM) and the other on Artificial Neural Networks (ANN), from the prior labelling of a certain number of operator-made events that allow specific classification models to be created for each meter (individual models). All distinguishing according to the meter precision (1 or, 0.1 litres). • For the two classification procedures, ANN and SVM, both methods have been developed to classify events identified from readings of new installations, in other words meters with no previous labelling, creating general models that make up for the non-existence of individual models for these new installations. • Design of graphic reports of results. • Integration of the different modules in a single computer application that contemplates the whole procedure.

INDEX

Summary of the results obtained	• For 1-litre precision meters, in overall terms considering all the events identified with the readings in all of the analyzed meters, the classification of these events by means of individual models based on ANN presents an overall percentage of hits, in terms of volume, of 86%, as opposed to 63% of those based on SVM. • For 0.1-litre precision meters, the percentage of hits reaches 91% (ANN) and 85% (SVM). • With the general models, the overall precision falls in comparison with the individual models and is 82% for ANN, and 75% for SVM, also in terms of volume.
Research Lines open for continuing the work	• Optimization of the automatic labelling process and application to massive data. • Evaluation of the impact of water-saving campaigns using the automatic labelling tool. • Applicability of the automatic labelling to large patterns of consumption. • Generation of integral control panels linked to automatic labelling.

INDEX

Executive Summary

OBJECTIVE

The main objective of this study is summed up in the title and is none other than to develop an automatic system to identify the end uses of water in the different residential applications, based on the signals recorded by precision meters, using advanced methodologies of pattern recognition and supervised signal classification, such as Artificial Neural Networks (ANN), and methods based on Support Vector Machines (SVM).

The meters used in this work do not measure flows directly, but are rather equipped with a digital pulse emitter which produces a signal (pulse) every time a certain volume is consumed (1 litre or 0.1 litres, depending on the precision of the apparatus). As a result, it was necessary to previously conceive a way to turn said pulse registers into flows.

It must be stressed that the methods developed in this project could also be applied to other kinds of recorders, such as those of flow, when the transformation of pulses into flows does not take place.

All of these procedures have been automated and programmed by means of a computer application that allows the massive treatment of data from a multitude of meters and long periods of time, without the participation of an operator.

METHOD

The initial information of the work presented here includes the data corresponding to the pulses recorded by 375 meters from January 2008 to July 2015, which supposes the processing of the registers of some 20,340 months, approximately, which include more than 34.65 million water use events. These registers were previously classified using an automatic method to identify uses developed by Canal de Isabel II, based on Bayesian Networks. For every meter, there are at least two months of data manually classified by an operator, which served as training data for the Bayesian Networks. This previous labelling (classification), carried out manually, is considered representative of the ground truth for each meter and is what has been used to generate, through a training process, ANN and SVM models based on which new events are automatically classified.

The classes of uses used in this classification or labelling were the following:

- ✓ Faucets
- ✓ Toilets
- ✓ Showers, including bathtubs
- ✓ Clothes washer
- ✓ Dishwasher
- ✓ Swimming pool
- ✓ Irrigation
- ✓ Leaks

Based on this information, the following works were performed, grouped by modules:

- Module 1. Transformation of pulses into flows
- Module 2. Identification of events
- Module 3. Development of the Artificial Neural Network models (ANN)
- Module 4. Development of models based on Support Vector Machines (SVM)
- Module 5. Assignment of end uses
- Module 6. Creation of new models for future installations

As work prior to these modules, a review of the state-of-the-art on pattern recognition and automatic classification of end uses of water was made.

Processes for conversion of pulse records into flow, and events isolation were performed by two separate mathematical algorithms conceived *ad hoc*, so they allow this transformation automatically, and it does not depend on the more or less subjective criteria of an operator.

To generate the ANN and SVM based models, the prior labelling of a certain number of events is needed This dataset is used by the algorithms to "learn" and to build the models that will assign the corresponding label to the remaining events automatically.

Two types of meters were used, depending on their precision: of 1 litre and 0.1 litres. This precision corresponds to the consumed volume that produces a pulse and was borne in mind in drawing up the different mathematical algorithms both turning pulses into flows and producing ANN and SVM models.

These tasks required the development of specific computer applications for each of the modules. For the first two modules - transformation of pulses into flows and identification of events -***Visual Basic for Applications*** (***VBA***) was used under ***Access***[1]; whereas, in the rest of the modules, the production environment used was ***Miniconda***, a small version of ***Anaconda***[2], package distribution and management based on ***Python***[3], which only contains ***Conda*** and ***Python***.

As a colophon of the work done, these computer applications were integrated into a single one developed under ***Access***, which includes the whole procedure necessary for the identification and classification of the end uses of water in residential consumptions.

[1]ACCESS: ©Microsoft
[2] CONDA: Open Source and with new BSD Licence
[3] PYTHON: Programming Language with PSFL Licence

INDEX

This application enables the identified events to be classified from new readings of the meters for which a specific classification model has already been generated (individual model), and the development of new models for meters of other installations, provided there is a set of training data (classified manually). If these training data are not available, the general models built on the basis of the events treated in this study can be used.

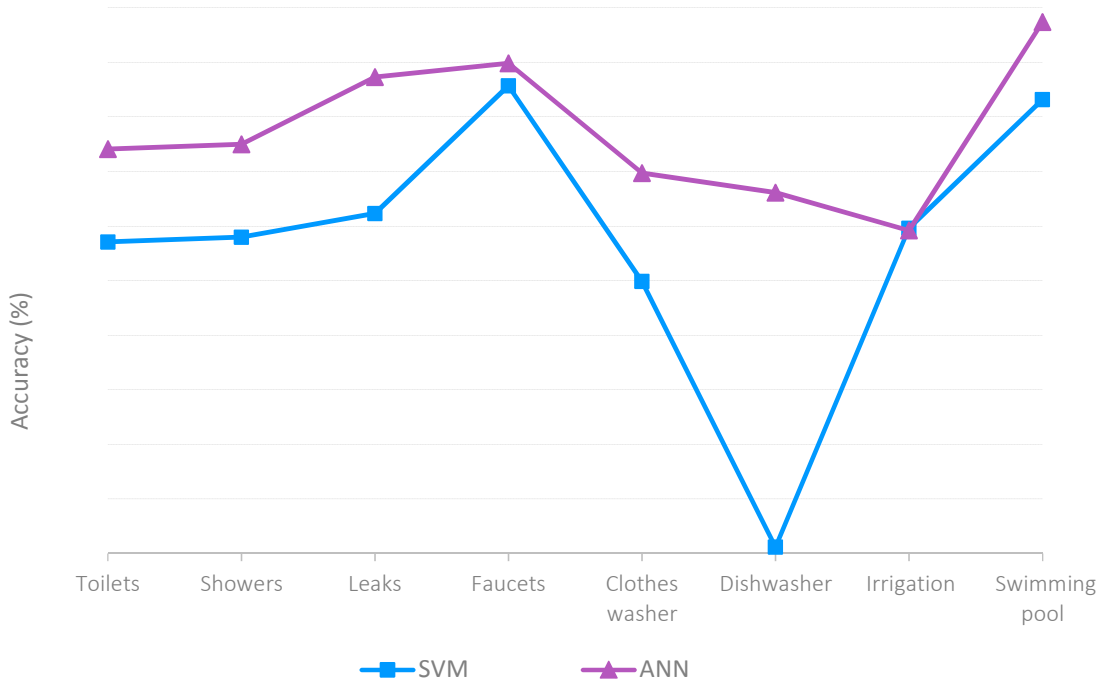
RESULTS AND CONCLUSIONS

For meters with a precision of 1 litre, the results obtained with ANN are significantly better than those obtained with SVM, both in overall terms and on meter level.

In fact, having in mind the results obtained with both classification methods, the accuracy with SVM is 67.41 %, whereas with the ANN it is 81.78 %, referring to the number of events correctly classified. In terms of correctly classified volume of water, the accuracy is very similar, it being 63.41 % for SVM and, 85.76 % for the ANN.

Observing the results of all the meters of this kind and distinguishing by uses, the precision obtained with ANN is always better than that obtained with SVM. Figure 1 shows the difference in the precision considering each use separately.

FIGURE 1. PRECISION OF THE ALGORITHMS USING 1 LITRE METERS



INDEX

Regarding individual meters, itis not possible to find any for which the results of SVM are better than those of ANN.

The number of events correctly classified with ANNs increases by an average 22.35 %, exceeding 40 % on 19 occasions.

Furthermore, for the 19 analysed **meters with a precision of 0.1 litres**, as with the 1 litre meters, the results obtained with ANN are better than those obtained with SVM both in overall terms and on meter level. The average accuracy of SVMs and ANNs is 84.78 % and 91.19 % respectively. Ifthe percentage is quantified as a volume, these values drop to 73.5 % and 85.9 %. This is because most events are *Faucet* type use, with medium-low volumes, which classify very well with both methods.

Figure 2 reflects this very well. The precision in the classification of *Faucets* is over 95 % with both algorithms. As most ofthe events are ofthis kind, many erroneous classifications are produced as the methods recognise the bias in the distribution of events, and tend to classify events of other kinds as *Faucets*. Despite this, a considerable improvementis seen in the precision ofthe classification of *Toilet*, *Shower*, *Clothes washer*, *Dishwasher* and *Irrigation* type events.

FIGURE 2. PRECISION OF THE CLASSIFICATION ALGORITHMS BY USE IN 0.1 LITRE METERS

Similarly, with respectto the meter, itis not possible either to find any eventfor which the results of SVM are significantly better than those of ANN.

The number of events correctly classified with ANNs Network increases by an average 8.5 %, exceeding 20 % on 3 occasions.

For the **general models**, the overall precision ofthe method is 75.18 % with SVM and 82.17 % with ANNs.

INDEX

Comparing the individual models with the general, it is concluded that, as it was to be expected, the results are better when applying the models in individual meters.

The tables and graphs that follow show the results obtained using the two classification methods, ANNs and SVMs, for all the data processed in the period ranging from January 2008 to July 2015.

Specifically, Table 1 and figures 3 to 6 reflect the distribution results by uses of the total consumption in the period from January 2008 to July 2015 for the ANN classification method. Similarly, Table 2 and figures 7 to 10 show the classification results for the SVM method according to total consumption uses throughout the referred period.

TABLE 1. RESULTS OF THE CLASSIFICATION BY ANN, DISTRIBUTION BY TOTAL CONSUMPTION, BY USE IN THE PERIOD FROM JANUARY 2008 TO JULY 2015

Use	Total consumption (m^3)		Monthly average consumption (m^3)	Nº of events total		Nº of events monthly average	Average consumption, per event (L)
Faucets	54,198	41%	595.58	21,028,397	61%	231,081	2.58
Toilets	13,662	10%	150.14	2,614,989	8%	28,736	5.22
Showers	43,323	33%	476.08	2,338,834	7%	25,701	18.52
Clothes washer	9,133	7%	100.36	1,291,398	4%	14,191	7.07
Dishwasher	1,951	1%	21.68	1,883,072	5%	20,923	1.04
Swimming pool	88	0%	0.99	1,608	0%	18	54.56
Irrigation	2,751	2%	30.57	48,657	0%	541	56.54
Leaks	5,543	4%	60.91	5,442,828	16%	59,811	1.02
Total	130,649	100%	1,436.30	34,649,783	100%	381,003	3.77

FIGURE 3. RESULTS OF THE CLASSIFICATION BY ANN, DISTRIBUTION OF TOTAL CONSUMPTION M̂

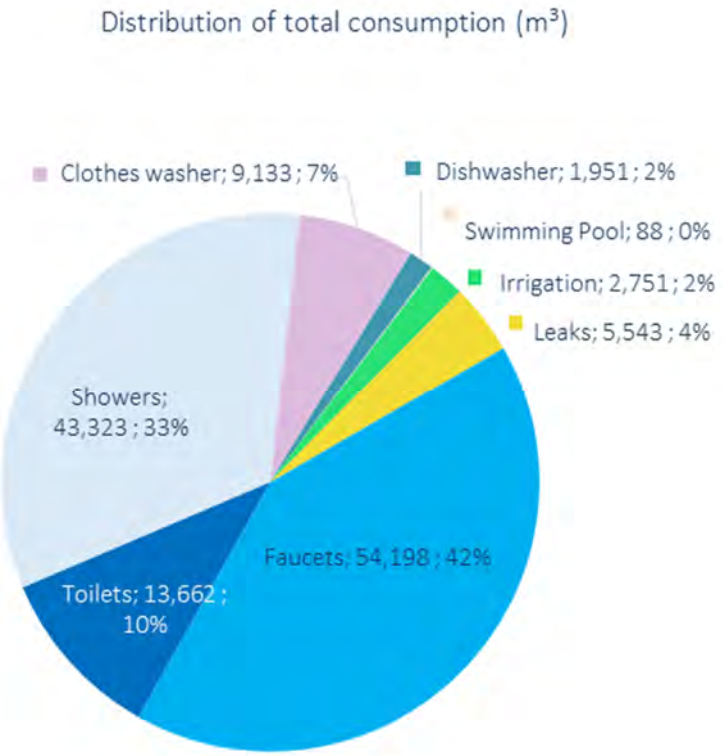
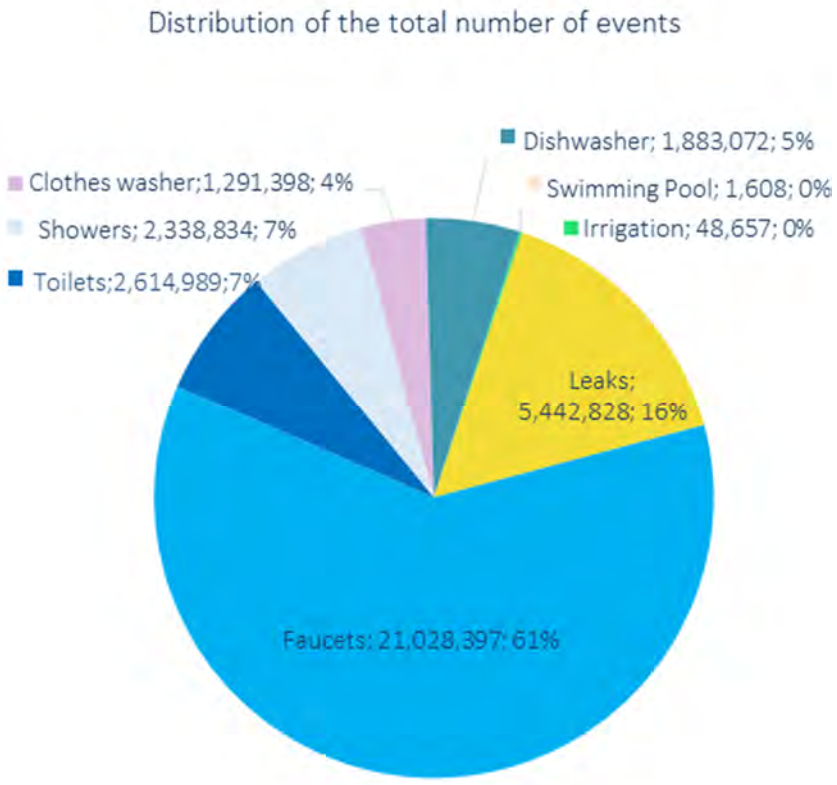


FIGURE 4. RESULTS OF THE CLASSIFICATION BY ANN, DISTRIBUTION OF THE TOTAL NUMBER OF EVENTS



INDEX

FIGURE 5. RESULTS OF THE CLASSIFICATION BY ANN, AVERAGE CONSUMPTION PER EVENT (L)

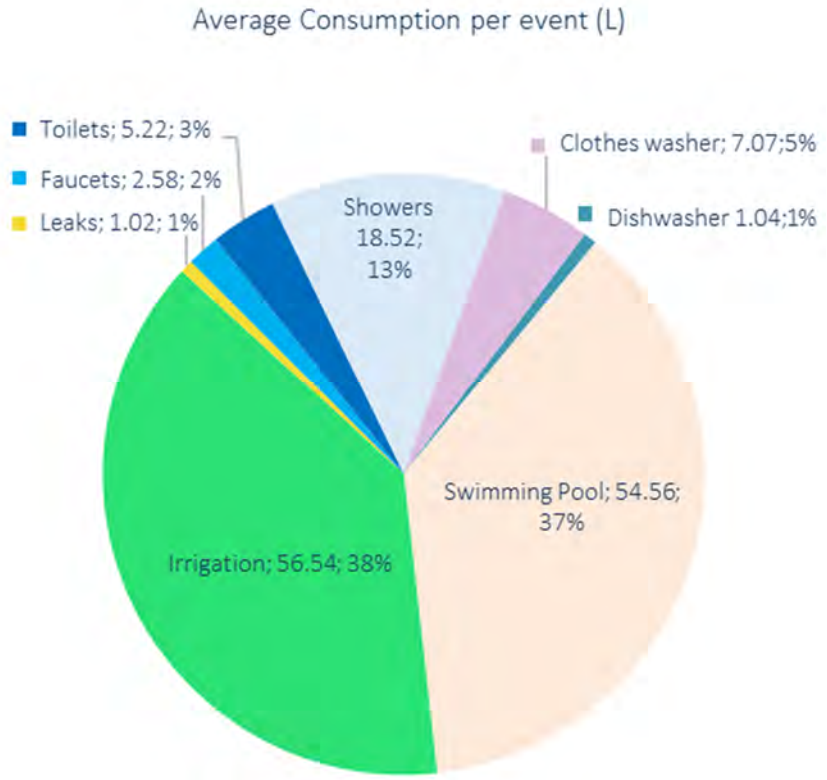
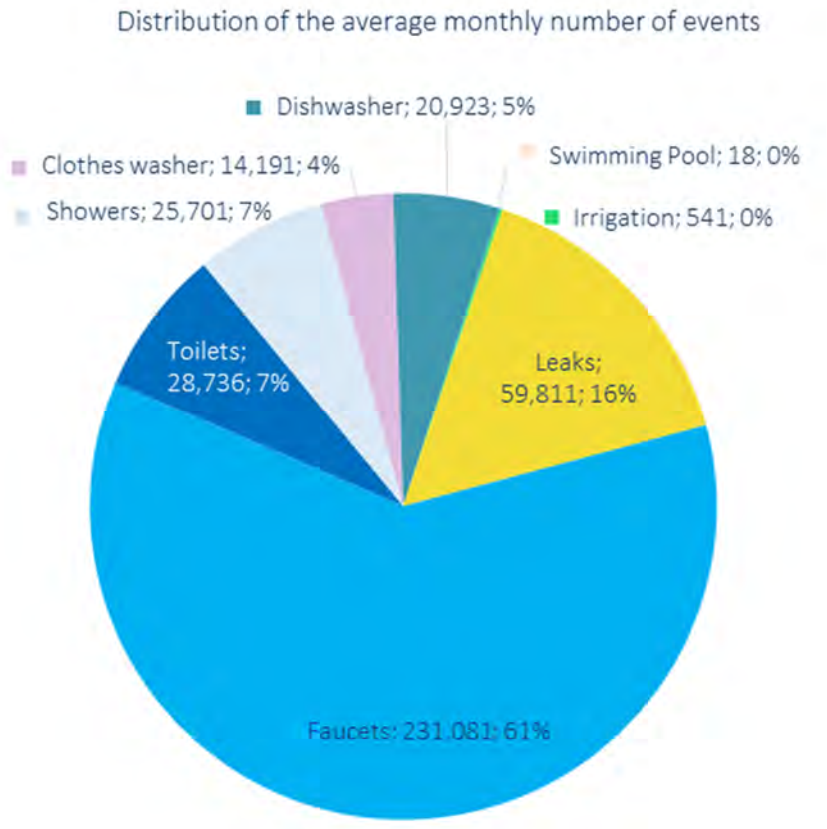


FIGURE 6. RESULTS OF THE CLASSIFICATION BY ANN. DISTRIBUTION OF THE AVERAGE MONTHLY NUMBER OF EVENTS

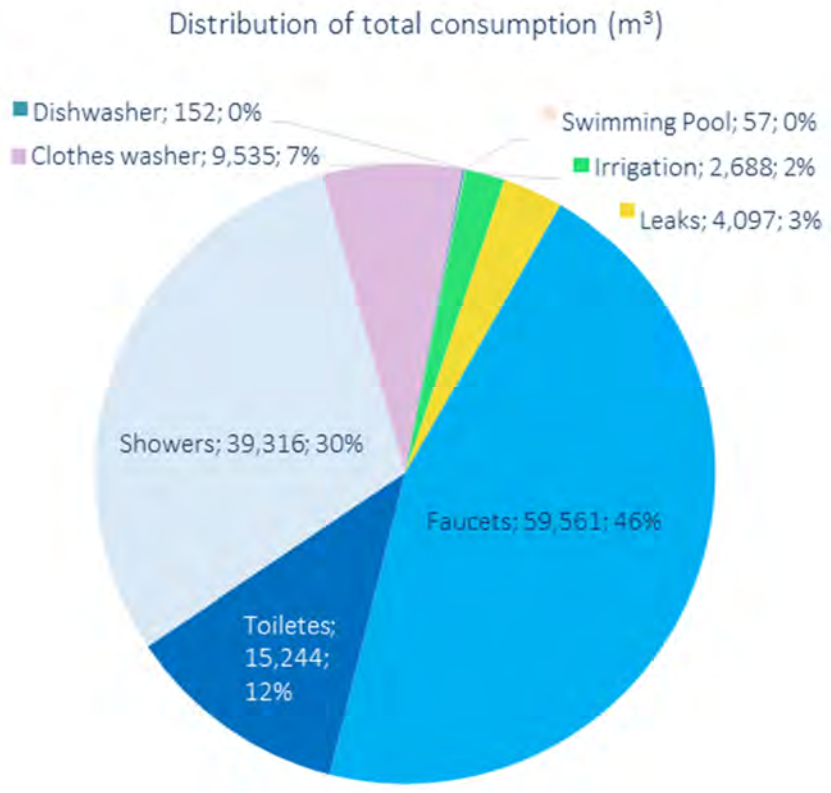


INDEX

TABLE 2. RESULTS OF THE CLASSIFICATION BY SVM. DISTRIBUTION BY TOTAL CONSUMPTION, BY USE IN THE PERIOD FROM JANUARY-2008 TO A JULY-2015

Use	Total consumption (m^3)		Monthly average consumption (m^3)	Nº of events total		Nº of events monthly average	Average consumption, per event (L)
Faucets	59,561	46%	654.51	24,915,468	72%	273,796	2.39
Toilets	15,244	12%	167.52	2,379,525	7%	26,149	6.41
Showers	39,316	30%	432.04	1,900,028	5%	20,879	20.69
Clothes washer	9,535	7%	104.78	1,205,685	3%	13,249	7.91
Dishwasher	152	0%	1.69	78,946	0%	877	1.93
Swimming pool	57	0%	7.11	755	0%	94	75.32
Irrigation	2,688	2%	29.86	50,218	0%	558	53.52
Leaks	4,097	3%	45.02	4,119,158	12%	45,265	0.99
Total	130,649	100%	1,.442.53	34,649,783	100%	380,869	3.77

FIGURE 7. RESULTS OF THE CLASSIFICATION BY SVM. DISTRIBUTION OF TOTAL CONSUMPTION M̂



INDEX

FIGURE 8. RESULTS OF THE CLASSIFICATION BY SVM. DISTRIBUTION OF THE TOTAL NUMBER OF EVENTS

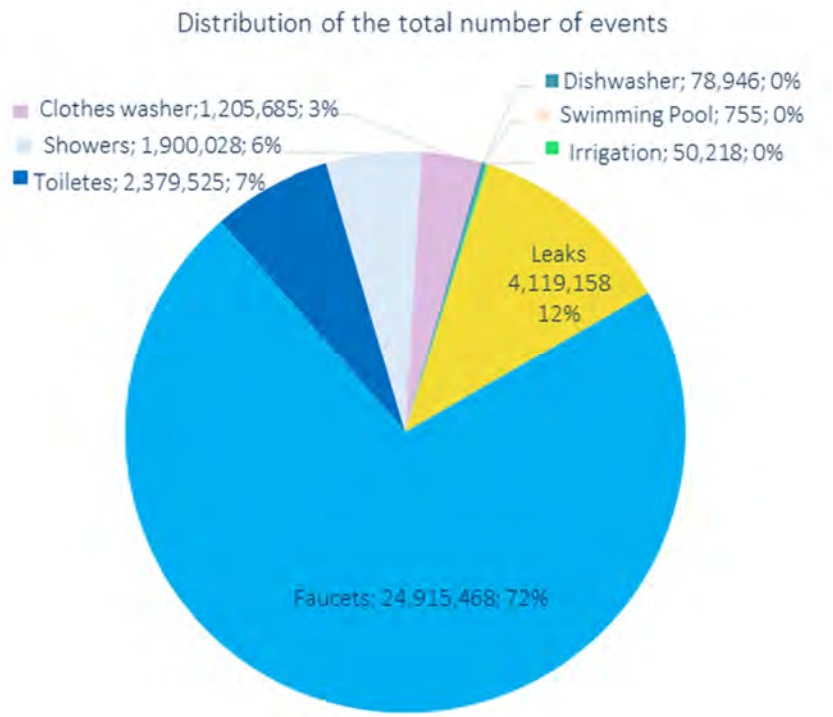
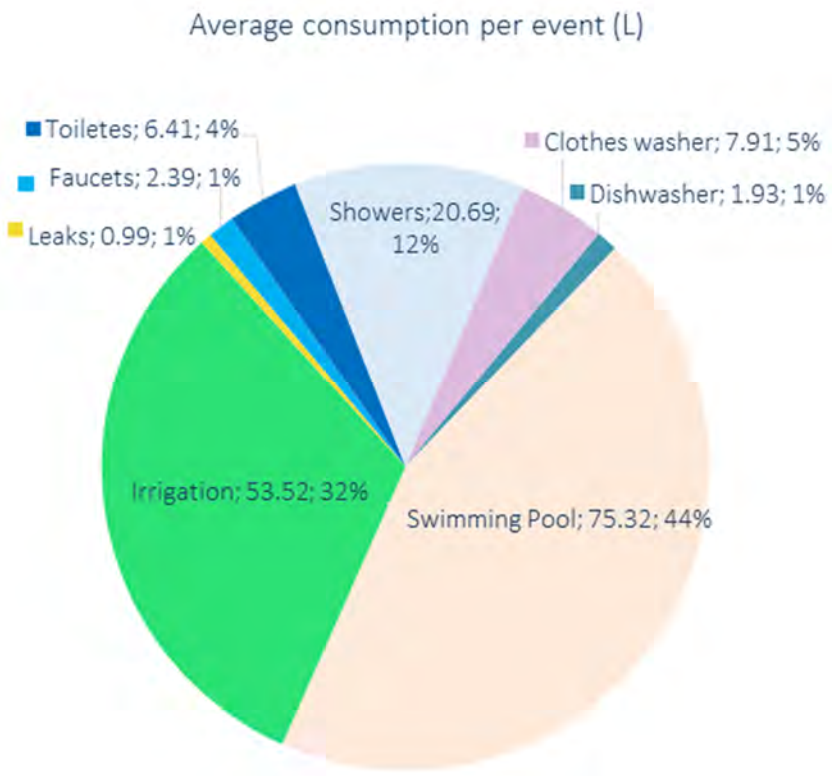
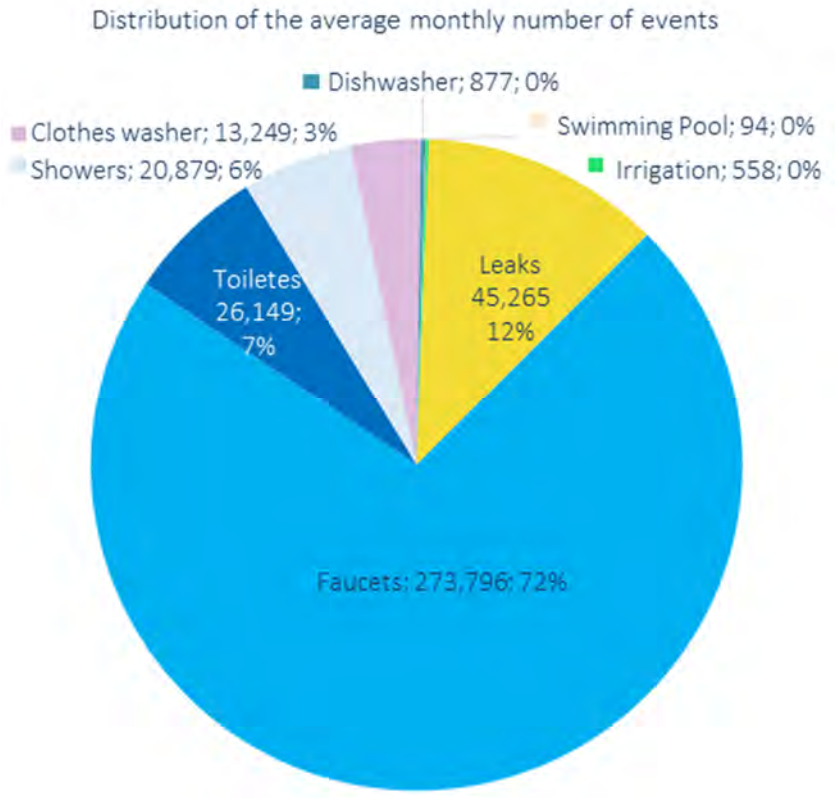


FIGURE 9. RESULTS OF THE CLASSIFICATION BY SVM. AVERAGE CONSUMPTION PER EVENT (L)



INDEX

FIGURE 10. RESULTS OF THE CLASSIFICATION BY SVM. DISTRIBUTION OF THE AVERAGE MONTHLY NUMBER OF EVENTS



INDEX

1. Introduction

Since early 2008, Canal de Isabel II has been monitoring a selection of houses around the region, which began with class **C** volumetric meters with pulse emitter and a precision of one litre; after 2011 the meters installed were class **D** with a precision of one decilitre.

Since the beginning of the project, more than fifteen million hours of water consumption have been monitored, and about 140 million litres of water were accounted for, in a total of 375 different installations.

To identify the end uses using the recorded flows (pulses), *Trace Wizard*® commercial software from *Aquacraft* was used. This application makes a preliminary classification of the different events detected, which must be supervised by an operator who decides the final assignment of uses. This method requires a large number of operator hours, estimated at 1-2 hours for processing a week of consumption data one-dwelling, which makes it practically non-viable for monitoring a sample of a considerable size for several years.

Therefore, in early 2009, an automatic use identification methodology was developed using Bayesian methods and heuristic of minimum entropy to discretise the samples data series. The method is based on the statistical analysis of series of data, lasting two or three months, for each user, processed by operator and used as a training dataset. This statistical analysis consists of calculating the probability that a certain event, given its duration, peak flow and average flow, corresponds to each of the different typified uses, and adopting that with a higher probability.

Therefore today, for each of the 375 installations there is a historical series of at least two months of information processed by operator, and the rest with the Bayesian classifier.

In the project described in this document, the development of more advanced methods for classifying residential water consumption events has been dealt with, which improve the performance of the present statistical techniques, increasing their reliability, versatility and ease of use, and which may be exported to other water supply systems.

The events classification process requires preliminary learning or training to allow the patterns of the different uses to be recognised. This learning can be automatic by applying different models, which are usually classified into five categories, which are:

- neural networks
- learning based on cases or instances
- genetic algorithms
- rule induction
- learning by analogy.

The recent comparisons of which these models have been an object means that today they are all accepted as very similar, at least regarding learning capacity. In specific applications, some may be more efficient than others, but the differences handled a few years ago are in question. It is more and more widely accepted, for instance, that the learning done by the neural networks, which is qualified as sub symbolic learning, is no more sub symbolic than the classification rules, although it is admitted that there are aspects which distinguish them and which mean that each one has specific applications.

There are many applications of **Machine-learning**, **data mining** for classification being one of the most significant. Establishing relations between multiple features is a complicated process and often doomed to failure, which makes it difficult to design solutions for these kinds of problems. Machine-learning strategies offer an efficient solution to this problem. These techniques are seen to be valid for the problem of classification (final assignment of a value) bearing in mind certain features (input values).

Each register in a set of data used by the **Machine-learning** algorithms is represented using the same features. These features may be continuous, discrete or binary. If each register is labelled with its correct output value (label), they are called *"supervised learning strategies*". On the other hand, the strategies that use unlabelled registers are called *"unsupervised learning strategies*".

The classification problem that affects the area of this project focuses on the supervised learning strategies. These strategies are included in the concept of supervised inductive learning. Induction in a broad sense, consists of finding properties common to a finite sub series of elements of a certain domain and considering that these properties can be extended to any element of the domain. A good induction would try to preserve the truth, but it would reveal those elements with errors producing noise in the information they provide, and may even contemplate exceptions to the knowledge drawn.

2. Objectives

The main objective of this study is summed up in the title and it is to develop an automatic system to identify the end uses of water in the different residential applications, based on the signals recorded by precision meters, using advanced methodologies of pattern recognition and supervised signal classification, such as Artificial Neural Networks (ANN), and methods based on Support Vector Machines (SVM). To achieve this, it was necessary to develop different algorithms and procedures allowing the different phases of this process to be automated, which are:

- Convert the pulse readings of the meters into flows
- Identify events
- Generate learning models
- Classify events

The meters used in this work do not measure flows directly, but are rather equipped with a digital pulse emitter that produces a signal (pulse) each time a certain volume is consumed (1 litre or 0.1 litres, depending on the precision of the device), the developed methodology could also be applied to other kinds of precision recordings, such as direct flow measurement, as the pulses in themselves do not form part of the data input of the models, but rather the information drawn from the flows associated with the continuous pulse registers.

The different uses managed in this classification or labelling were the following:

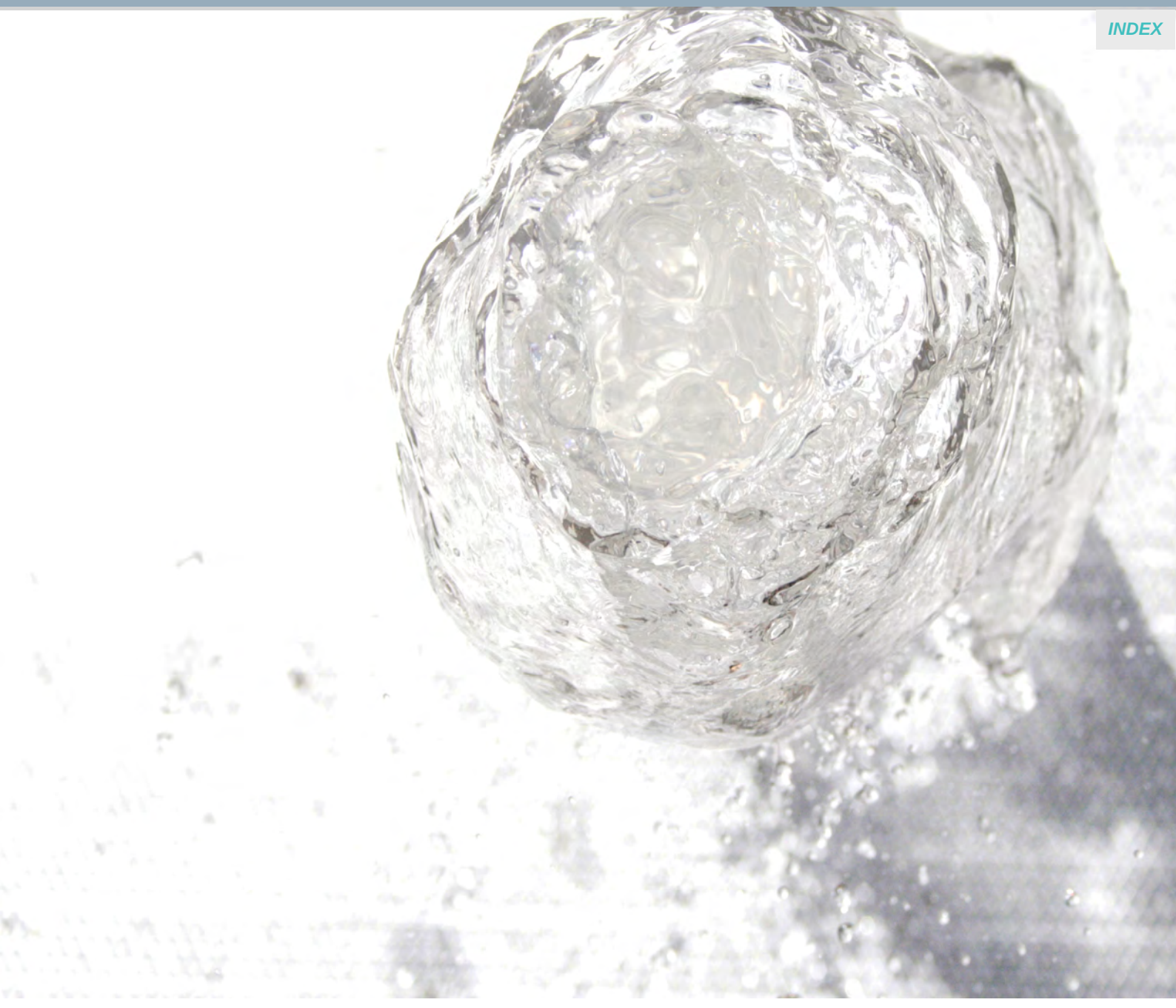
- Faucets
- Toilets
- Showers (including bathtubs)
- Clothes washer
- Dishwasher
- Swimming pool
- Irrigation
- Leaks

Each of these phases gave rise to the corresponding computer modules, which were finally grouped in a single application.

This computer application allows the massive treatment of data from a multitude of meters and long periods of time without the participation of an operator.

This study worked with the data from the pulses recorded by 375 meters from January 2008 to July 2015, which means the processing of the registers over 15 million hours and a total accounted volume of approximately 140 million litres of water, in some 34.65 million water use events.

3. State of the art



The automatic classification of end-uses of water in the residential area is a current problem whose solution consist of drawing up precise schemes of water management and adequate measurement of its consumption. Significant improvements in this measurement and the implementation of data analysis techniques have made it possible to address this problem with positive results. The desire to improve the monitoring and analysis of water consumption has enabled consumption data to be collected via a smart measurement system to be transferred and stored in a repository for later analysis. But to carry out this consumption analysis, filtered and processed information is required, in other words, consumption data broken down by use. Therefore, the key for analytical tool development or to extend water management companies customer's knowledge, it is needed to develop pattern recognition algorithms that are capable of categorising flows and/or events in typical residential uses (clothes washer, dishwasher, etc.).

3.1. STATE-OF-THE-ART IN CONSUMPTION MEASUREMENT SYSTEMS

First references in household water consumption measurement, for the determination of the end uses of water, date from the end of the 20th century. These kinds of studies are still under way today to achieve a greater understanding of the processes of household consumption.

Canal de Isabel II started research work in this field in 2001 with different data collection phases between 2001 and 2003, and throughout 2006 (see Booklet R+D+I 4[4]). Today, since 2008 it has maintained a stable sample of around 300 houses monitored continuously.

22 studies have been located and analysed, of them more than half correspond to work done in Australian cities between 2005 and 2014; four are from cities in the United States, three in California (in 2004 and 2010), and one in Seattle (2000); three in Spain, that of Canal de Isabel II in the Comunidad de Madrid (2001 – 2006), another focused on the city of Zaragoza (2010), and the third in various cities (2002). The two remaining studies were carried out in several cities of New Zealand (2007) and in Abu Dhabi, Arab Emirates (2013). Regarding the number of residences involved, this varied but more commonly ranged from 100 to 300 installations, in some case reaching as many as 384 houses (Zaragoza, 2010), and even 474 (Townsville, Australia, 2007). With respect to the type of meter, in all studies but one, volumetric meters were used with pulse emitters and precisions of, at least, one decilitre, in other words, a pulse is emitted each time 0.1 litres are consumed; there are studies with meters of precision of up to 0.014 litres.

As can be seen, Australia is one of the countries that has most worked on analysing patterns of end use of water using smart meters. All the studies carried out along this line are based on predefined pattern analysis algorithms, and the parameters that characterise these patterns must be checked and adjusted manually in each case, which generates a certain degree of uncertainty mainly when several events overlap.

[4] Cuaderno de I+D+i nº 4, Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid. Canal de Isabel II.

INDEX

To try to complement these methods, the Australian research agency CSIRO is developing a prototype system that automatically determines the end use of residential water without the need for the owners of the dwelling to record events each day. The prototype includes sensors capable of identifying the use of indoor devices by recording the acoustic signals produced during an event. When this information is combined with the flow register by smart meter, the system allows the water flow to be very precisely defined in each of the devices in a disaggregated way.

To test the prototype, a study during 3 years on the end use-of water in dwellings was made by the Griffith University[5]. On it were reviewed the data analysed on a selection of 320 homes in four areas of Australia (Ipswich, Brisbane, Gold Coast and Sunshine Coast). The device was installed in only one dwelling, as part of the general water end use study, and to check the system.

Three types of analysis were therefore obtained in this dwelling for each end use, depending on the origin of the registered data:

- Daily data obtained by the users themselves
- Data obtained by *Trace Wizard*
- CSIRO acoustic sensor

The device uses the signals of acoustic sensors that coincide with flow events of the smart meter to assign the end use. The acoustic sensors add two additional dimensions to the analysis of the end use, which are:

- Spatial identification: this allows the signal to be segmented by the area where the acoustic signal has been produced (kitchen, bathroom ...).

- Classification of use: the features of the acoustic signal can be related to a particular type of device. At the present time, that classification is done manually, although automation is feasible through a more ambitious testing process

The results of the field test showed conclusively that the acoustic sensors can precisely determine the end use, when it is used along with the flow data obtained from the smart meter. The results were more precise than those obtained by the *Trace Wizard* program. However, several problems were detected: data synchronisation errors; frequent loss of communications in the Wi-Fi network or 3G modem. The designed prototype sensors had no internal memory, so data-gaps were produced in the information. The next version of these devices will include a data card and the synchronisation error problem will be solved.

The field test showed that acoustic devices can replace the use of the*Trace Wizard*. The challenge now is to automate the data interpretation process.

[5] Roger O'Halloran, Michael Best and Nigel Goodman, Urban Water Security Research Alliance. Technical Report No. 91. 2012

INDEX

3.2. CHARACTERIZATION OF RESIDENTIAL USES OF WATER

Different research works carried out in the United Estates[6], Spain[7] and Brazil[8y 9] suggest different methodologies to characterise water consumption in households. In these research works, pulse water meters and *dataloggers* were used to acquire data and different methodologies were developed for their analysis.

To enable signals or pulses to be identified, a series of features, of each device, should be firstly known (flow, volume, time of use, etc.). This allows the program to distinguish between the use of a faucet and a toilet tank, for example. If these parameters are not well-adjusted, correct identification cannot be made.

When three, or more events occur simultaneously, it may not be feasible to precisely dissect all the end uses. The usual procedure is to turn the pulses into flows (litre/second) to draw up consumption graphs (flow vs. time). The signals (representing water consumption signals) are correlated with the time used on the information of each device (provided by the users). This correlation allows precise identification of the signals of some of the uses, but generally, due to the poor precision of the information provided, some of the consumptions should be estimated. In the case included in the Booklet of R&D&I nº 4[7] edited by Canal de Isabel II, the information provided by the pulse signals was compared with the previous characterisation of their amplitude and time patterns, which allowed the use to be identified, at all times.

Fernandes[9] and Barreto[10] used a meter on each hydraulic device and, on the input pipe supply, to guarantee good precision in the characterisation phase. Despite the precision of the information on these studies, this methodology is difficult to apply in housing blocks due to the installing work needed in the supply network. This method has no technical or economic viability in most situations.

[6] Mayer, P Water. Energy savings from high efficiency fixtures and appliances in single family homes. USEPA, Combined Retrofit Report 1, 2005.

[7] Cubillo, F., Moreno, T., Ortega, S.: Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid. Cuaderno de I+D+i nº 4. Canal de Isabel II, 2008.

[8] Almeida, G.A., Kiperstok, A., Dias, M., Ludwig, O. *Metodologia para caracterização de consumo de água doméstico por equipamento hidráulico.* Anais do Silubesa/Abes. Figueira da fo. 2006.

[9] Fernandes, B.C.: Construção de um sistema eletrônico de monitoramento de consumo de água residencial. Projeto de graduação apresentado ao departamento de Engenharia Elétrica. P. 65 centro tecnológico da Univ. Federal do Espírito Santo, 2007.

[10] Barreto, D. Perfil do consumo residencial e usos finais da água. Ambiente Construído, Porto Alegre 8(2), 23–40 (2008) ISSN 1678 8621; © 2008, Associação Nacional de Tecnologia do Ambiente Construído, April/June 2008.reto

INDEX

3.3. STATE-OF-THE-ART IN PATTERN RECOGNITION AND WATER CONSUMPTION CLASSIFCATION

An expose is now made of the different techniques referred to in the literature for pattern recognition in each type of device of household water consumption, by using the data from the meters.

3.3.1. Technique 1. Robust linear multi category Classifier

This methodology is inspired in the application of the classification of end uses of residential electricity and gas consumption, although there is a fundamental difference between them: in the classification of the end uses of water, it cannot be supposed that the consumption flow is constant, and the maximum and minimum flows may have a wide range of possible values for a single device, neither there is a regular consumption base as it happens with electricity[11].

An *event* is described by four physical features: volume, duration, maximum flow and most frequent flow. In addition to these features, the analysis instrument uses the start-time, end-time and frequency of the most frequent flow during the *event*. All this information may be drawn from the data quantization given by the smart meter.

Consideration of the problem

The data used to develop this model come from the use of smart meters in 74 households, over two weeks. The flow data were stored every second and with a resolution of 100 and 250 pulses per litre, depending on the type of meter. The data were re-sampled in a resolution of 1 litre. The data with identified end uses contain the following information on each identified *event:* activation of the installation, start time, duration, volume, average-flow and maximum-flow.

The level of quantization, in other words, the resolution taken by the researchers (1 litre), is adapted to the meters currently available on the market. The sampling time (1 second) is also adapted to the quantization and is significantly ahead of the present sampling time used on the market. The flow registered in the households is not very high (for instance the maximum registered flow of all the identified uses is 0.64 litres/second), which implies that:

- In 1 second, only one pulse, at most, may be registered (as the taken resolution is 1 pulse per litre).
- The minimum time between two, non-null, consecutive registers is 2 seconds.

[11] Vasak M., Banjac G., Novak H.: Water Use Disaggregation Based on Classification of Feature Vectors Extracted from Smart Meter Data, Procedia Engineering, 119, 1381 – 1390, 3h Computer Control for Water Industry Conference, CCWI 2015.

INDEX

The main problem with the quantization is the large loss of information: processed data have only one dimension available, i.e. time when a pulse occurs. This means that all the features associated to a single water use can be extracted only from the information contained in timings of the corresponding records. However, mapping smart meter records to a single water use is not a straightforward task. Consider the situation when a distance between two consecutive pulses is 60 seconds. It cannot be asserted that the second pulse is a of a large flow during previous few seconds or of a small flow or leak during last 60 seconds; on the other side, it could be resulted by two consecutive (possibly different) water uses where the volume of first use accounts for less than 1 litre.

Therefore, it is assumed that only one fixture is used at a time, disaggregation of two or more simultaneous water uses may be too difficult with such a small resolution.

In order to map records to a single water use, the researchers define a maximum distance parameter based on which a decision on whether the two consecutive pulses correspond to the same water use is made. If the time distance between two consecutive pulses is less than maximum distance, they are assigned to the same group; only groups containing two or more records are considered. If this parameter is too small, it can result in splitting one water use into two or more. On the other side, if the parameter is too large, two or more water uses can be merged into a single use; if the merged water uses are different, a choice regarding which fixture should be associated to the resulting use should be made. Therefore, the choice of this parameter is not straightforward and a trade-off between these two cases should be found. As a good trade-off, the maximum distance parameter is set to 150 seconds. The long events (dishwasher cycle, clothes washer cycle and shower) are formed by intermittent cycles that are characterised as separate events if the water meter is inactive for more than 150 seconds, so the device programs can be identified. Simultaneous uses are assigned the device of dominant use.

The start and end times estimated with the processed registers do not coincide with those estimated by using high - resolution registers. What's more, the total estimated volume using the processed registers can only be a multiple of 1 litre.

Useful data

In the study data, 29% of the events overlapped in time, 20% of the overlapping time was of different devices and the overlapping added up to 13% of the total water consumption.

There are 12 types of *events* considered in the study: sink, clothes washer, toilet-1, toilet-2, toilet-3, dishwasher, shower, pressure shower, faucets, bath tub, water softener and outdoor faucet. Faucets, clothes washer and toilet (toilet-1, toilet-2 and toilet-3) add up to more than 92% of the total uses, 69% of the duration and 62% of the total volume consumed. Moreover, although they are used very little (2.20% of the total number of uses). The shower adds up to around 19% of the total duration and consumes a lot of water in the households, 23.06% of the total volume. Considering only the labels without joining similar devices, we find that the faucets, the clothes washer, toilet 1, toilet 2 and the shower add up to 93.09% of all identified uses, with 84.6% of the duration and 83.1% of the volume consumed.

Vector of features of End Uses of Water

As has been said before, an *event* can be described by physical features: volume, duration, maximum-flow and most-frequent-flow, as well as other features that may be brought in: initial moment, final moment and frequency of most frequent flow. Two harmonic functions are also considered, *sin* and *cos*, which may be used to represent this time of day.

These functions are continuous, periodic and their values uniquely determine the time of day. Therefore, each end-use is represented by the following feature vector:

$$x = \left[v, d, q_{max}, q_{freq}, N_{freq}, \sin\frac{2\pi t}{T}, \cos\frac{2\pi t}{T}\right] \in \boldsymbol{R}^7$$

Where v is the volume, d is the duration, q_{max} is the maximum flow rate, q_{freq} is the most-common-flow rate, N_{freq} is the number of the most common flow rate occurrences, t is the number of minutes elapsed from midnight and $T = 24 \cdot 60$ is the total minutes in one day.

Each *Feature-Vector* x is accompanied with the end use characterisation and the robust linear classification (M-RLP) can be applied to them.

Classification method

This work uses a linear classifier, which assumes that two types of data can be separated by a linear boundary. These kinds of techniques are not sufficiently descriptive and in fact there may be a non-linear boundary that separates the data better. To make the linear models work in non-linear blocks, the *feature - vector* was mapped to higher dimensions where it is possible to apply non-linear classification techniques. Therefore, the original feature-vector of $x = [x_1, \dots, x_7]$ is mapped to a higher dimension space using the following transformation:

$$\phi: x \longrightarrow [x_1, \dots, x_7, x_1^2, \dots, x_7^2, x_1x_2, \dots, x_1x_7, \dots, x_6x_1, \dots, x_6x_7] \in R^{35},$$

The squares of each feature and the products between the features have been added to the original vector. Furthermore, the complexity of a model of this kind increases very much in terms of the number of necessary parameters in the model.

RESULTS

The end uses corresponding to the different devices are labelled, as has been seen. There are 12 labels in total. The results obtained were not very satisfactory: only 5 out of 12 of the devices had a position other than 0. The characterisation of faucet, pressure shower and water softener had a precision of more than 50%, but the aggregate use of the pressure shower and softener add up to less than 0.5% of all uses, which makes the obtained classifier useless.

INDEX

The results of the classification of the common model with 7 features are not very satisfactory. This is because the problem is not linearly separable with these features and linear classifier is used. Tap usage accounts for more than two thirds of the total uses of water and the precision of its classification is under 5%, while toilet has 0 accuracy. After the features space is extended with additional features, the precision for some uses improves, although the classification for toilet 2 was much worse.

This study shows that when a different model is produced for each household, the precision of the classification improves significantly when we compare it with that obtained using the common classification model.

3.3.2. Technique 2. Adaptive Neuro-based Fuzzy Interference System (*Anfis*)

In the work done in 2008 by Corona *et al.*[12] the automatic identification of the uses of water in a household was suggested by means of an*Anfis* model and fuzzy clustering.

Consideration of the problem

The data used to prepare this model came from a house with three inhabitants and on whose residential water uses, a set of data was drawn up with 1,000 examples, 100 for each of the five classes and 100 for each of the five subclasses. The classes are: WC, faucet, shower, dishwasher and clothes washer. The shower divides into three subclasses, due to the different sharing habits of the three people and, the clothes washer into another two, as result of each user intervention at the beginning of the washing cycle.

The researchers considered that the problem of classifying the residential uses of water varied much from one house to another, as it depends on the number of outputs, hydraulic installations, types of devices used, consumption habits, etc. They therefore used a model for the classifier seeking for easy interpretation by humans and the possibility of adding the knowledge provided by users or experts: neuro-fuzzy model, and particularly the*Anfis* model.

Neuro fuzzy models

The neuro-fuzzy models characterise by the use of best advantages features of neural networks and the fuzzy logic models; on the one hand, they provide the learning and generalisation capacity of the neural networks, and on the other the logical reasoning based on rules of inference.

[12] Corona-Nakamura M. A., Ruelas R., Ojeda-Magaña B., Andina D.: Classification of Domestic Water Consumption using an Anfis model; Conference paper. Automation Congress, 2008.

INDEX

Adaptive Neuro-based Fuzzy Inference System (Anfis)

The neuro-fuzzy architecture used in classifying the end uses of water is an adaptable neuro-fuzzy network, known as *Anfis*. This architecture is equivalent to a fuzzy inference system that can be built from the relations between input and output values of a dataset. In this inference system, *Anfis* tunes the membership functions during the training process of the model. Membership functions and fuzzy rules that should be adapted to the problem, must be defined before training the *Anfis* model. For the initial estimate of these parameters, it is possible to use a clustering algorithm such as the *Fuzzy c - Means* (FCM), the *Mountain Method*, the subtractive method or simply the expert's knowledge. The fuzzy rules are based on the Takagi-Sugeno inference method and the conclusions are polynomial functions.

Subtractive Clustering Method (SCM)

This method is often used when the number of clusters is known, but not their centers. Its aim is precisely to estimate these centers. This is a fast method and is based on a similar idea to the *Mountain Method*, as both divide the space of data through a grid in which the intersections are a set of candidates belonging to a given cluster. The calculation of center of the cluster is based upon density of the data set. Although the *Mountain Method* is simple and effective, its computational cost grows exponentially with the dimension of the problem, as result of the evaluation of the density function evaluation over all the points of the grid.

Classification

Due to the large differences in consumption in different locations and activities, the characteristic space presents large regions without data. This is the reason that unrestricted functions have been chosen, so that they cover all the space, and the classifier has the chance to recognise points beyond the frontier of the data available. The function selected to represent the fuzzy sets of the model is a Gaussian function.

Once the classifier is trained, some membership functions of the *Anfis* model will have to be changed to achieve better recognition of the new data, especially when they are between classes. In this case, a Gaussian is replaced by a bounded Gaussian (Gaussian2), a bell or triangular function.

The number of fuzzy rules from the classification results or *clustering* generally coincides with the number of classes or clusters.

RESULTS

The results obtained using this methodology have been positive. In the worst case, the percentage of success was more than 91%, but as the training was done with data from a single household, it is not possible to guarantee these results over a broader, and more varied sample.

INDEX

3.3.3. Technique 3. Hybrid model of filtering, Artificial Neural Network and Markov hidden model

In the works done by Nguyen, Zhang and Stewart in 2012, a hybrid pattern filtering and recognition model is suggested for the categorisation of residential uses of water[13].

This technique is the most sophisticated and interesting of those gathered.

Consideration of the problem

The database used for the study bade by these authors was very broad, with high-resolution meters, storing 0.014 litres/pulse in intervals of 5 seconds, carried out in more than 500 households over 3 years. The events were taken manually from these data and classified using the*Trace Wizard* software. There were seven possible outcomes of this process of identification: shower, faucet, dishwasher, clothes washer, toilet, bath tub and irrigation.

A sample was used of approximately 100,000 events, 83,000 classified for training, and 16,000 for verification or test.

The pattern recognition techniques used in this study are: Hidden Markov Model (HMM), Artificial Neural Network (ANN) and the Dynamic Time Warping (DTW) algorithm. Hybrid combination of these techniques was finally chosen as most suitable and precise for the pattern recognition system.

Hidden Markov Model (HMM)

This model was used in this study as one of the classifiers for the end use of water, based on the shape of the event. However, the weakness of this technique lies in the fact that it fails to adequately classify those categories highly dependent on user behaviour. These uses are highly variable, so that sometimes display features that are very similar to the features of other categories

The shower, bath tub and irrigation can present alike patterns despite being different categories. It is therefore necessary to include an additional technique that can inspect the physical features of these events: Artificial Neural Network.

Artificial Neural Network (ANN)

The authors and developers of this method used a compensation network with a backpropagation training algorithm as the main technique for learning the typical pattern of each category in terms of physical features (for example volume, duration, maximum-flow, etc.).

[13] Nguyen, Zhang Y Stewart. Analysis of Simultaneous Water End Use Events Using a Hybrid Combination of Filtering and Pattern Recognition Techniques. International Congress on Environmental Modelling and Software (2012).

INDEX

Dynamic Time Warping (DTW)

This is a popular method for measuring the similarity between two-time series of different lengths. This task is generally done by finding an optimal alignment between two-time series with certain restrictions. The series are extended or shortened in the time dimension to determine a measure of their similarity, regardless of certain non-linear variations. The aim is to find a mapping with the shortest possible distance. This technique has often been used in pattern recognition or in word image searching. In the tool developed by the researchers, it was key to find intertwined water cycles related to that precise event, and for mechanised events (clothes washer and dishwasher, mainly) which were not classified correctly with the ***HMM*** or with the ***ANN***. The clothes washer and the dishwasher have water use cycles associated with programs selected by the consumer, which can be recognised with the***DTW***.

RESULTS

For the categories with clearly defined patterns such as the dishwasher and the clothes washer, the precision was higher (around 90%), whereas those more influenced by human behaviour had success rates of around 60%-80%.

3.3.4. Technique 4. Other techniques

In 2011, experimental tests were made in three German cities, including Berlin[14], under controlled conditions to explore different possibilities for preparing different facilities and testing algorithms. The steps for the experimental tests were:

- building an experimental installation
- calibrating the hydraulic system
- acquiring the signals
- storing the signals
- processing the data, and
- developing algorithms for pattern recognition in water consumption.

Digital signal processing tools such as convolution algorithms were used in the data processing, in an attempt to resolve the signal overlapping resulting from the simultaneous use of different devices in the hydraulic system. Techniques were then implemented and tested for the extraction of features and classifiers for identifying the signals of each kind of flow in their respective device. The extraction of features was then proposed, and a pattern recognition classifier to develop the algorithm in accordance with the device studied. These features can be drawn in the scope of time or frequency.

The time or frequency features of the transient reply from the hydraulic system to the activation of hydraulic devices are expected to vary in accordance with the features and position in the hydraulic supply system. This justifies the need to use different measuring devices, although afterwards only the signal of the flow in the pipe supply is used as input in the classifier. The algorithm compares the signal with the prototype generated for each device until the identification process is complete.

[14] Almeida G., Vieira J., Marques J., Cardoso A. Pattern Recognition of The Household Water Consumption Through Signal Analysis. In: Camarinha-Matos L.M. (Eds) Technological Innovation for Sustainability. Doceis 2011. IFIP Advances in Information and Communication Technology, Vol 349. Springer, Berlin, Heidelberg. 2011.

INDEX

4. Methodological approach

INDEX

The developed works have been grouped into the following modules:

- Module of transformation of pulses into flows
- Module of identification of events
- Module of classification of events
- Module of data queries and charts generation

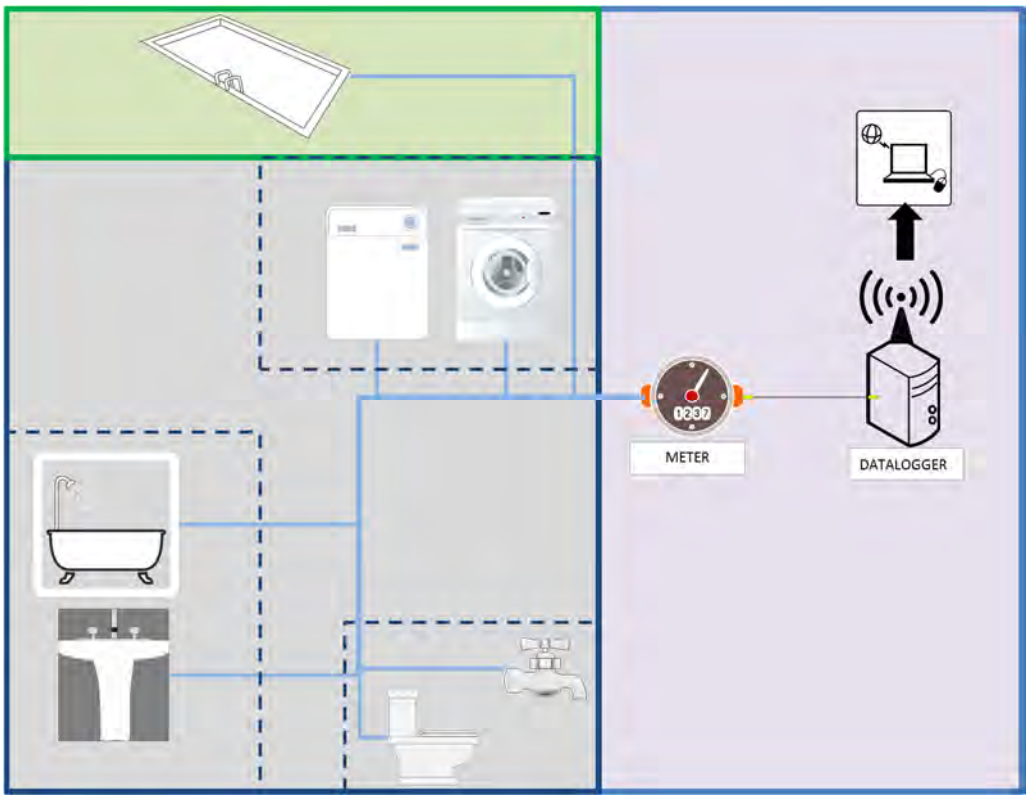
All these modules have been integrated in a computer platform developed in *Visual Basic for Applications (VBA)*, under *Access (©Microsoft)*, which allows massive and automatic treatment of meter readings (pulses), from which, using the methodology described in the following sections, the end uses of the different residential consumptions are determined.

4.1. TRANSFORMATION OF PULSES INTO FLOWS

4.1.1. Initial information

The initial consumption data available come from readings of volumetric meters with pulse emitter (see Figure 11). These kinds of meter emit a pulse every time a certain volume is consumed, which is given by the precision of the meter (1 or 0.1 litres, depending on the type of device), automatically recording the instant when it is produced, to a precision of one second. In meters of greater precision (0.1 litres), two, and even three pulses are relatively frequently produced in a second.

FIGURE 11. RECORDING LAYOUT OF READINGS OF METERS WITH PULSE EMITTER



Source: Own production from images published by pixabay.com, free of copyright under licence of Creative Commons CC0

INDEX

Each meter therefore produces a time series of accumulated pulses.

This entire meter reading information is stored in *Access* type databases, grouped in different folders according to the dates of the registers.

4.1.2. Calculation algorithm. Moving Averages

The transformation of the meter readings (instant-pulses accumulated) into time series (instant-flow) is rather more complicated than it might initially seem, and more so if the aim is to automate using a mathematical algorithm. In fact, an isolated pulse corresponding to a consumption, for instance, of 1 litre, may obey infinite combinations of flows occurring since the recording of the previous pulse, and each of them would give rise to different instant-flow series depending on the adopted criterion. Does this consumed volume respond to a constant flow from the time of the previous pulse? Have there been several uses that have produced an intermittent flow finally totalling 1 litre? How much time have these consumptions lasted? When the time between two pulses is short, in the order of a few seconds, supposing that the litre is distributed constantly in the time elapsed between the two consecutive pulses, may be a simplification that can be assumed to a certain extent, but as the time gets longer such a supposition is ever less plausible; it is not believable that there has been a constant consumption of 0.25 l/h for 4 hours, unless there is a leak, and this is precisely a pattern that allows leaks to be identified. In short, the procedure chosen for this transformation of pulses into flows can produce results that move excessively away from the reality to be reproduced.

To assess the goodwill of the chosen method, the calculated series of flows must be compared with the original series of flows that has produced the series of pulses used in the procedure. After considering different procedures, it was finally chosen to use a mathematical algorithm based on calculating moving averages, as given below:

1. From the series of readings of accumulated pulses, the series of discrete pulses is built by the simple difference of consecutive readings in the meter (ΔL).

2. Construction of a regular time series: As residential water consumption is not regular in time, the time interval between two consecutive registers is not either, and therefore gives a series with uneven intervals. It is therefore a question of drawing up a regular time series with time increases of one second.

3. As a first approach, it is supposed that when the pulse is issued, the volume given by ΔL·P is consumed, in which P is the precision of the meter (equivalent to the volume of each pulse).

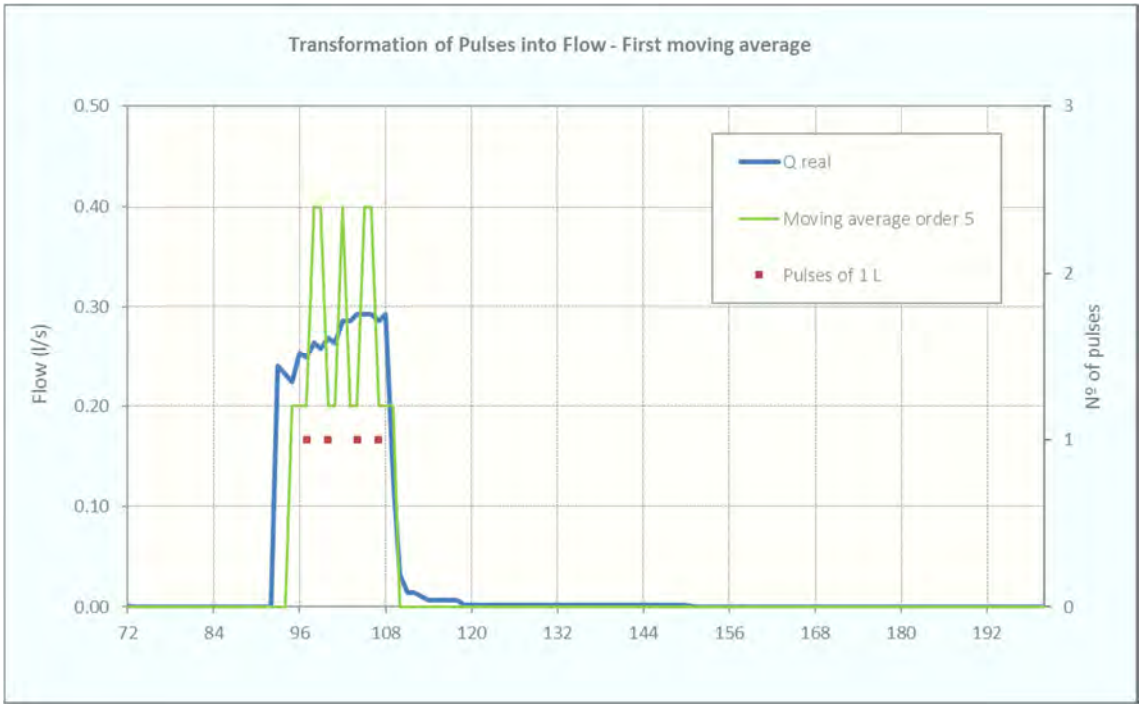
4. Each second is assigned a flow equal to the moving average of the flows calculated before.

5. To get an even better result, a second moving average of the above moving average is recalculated. The order of these moving averages is a parameter to be adjusted, as will be detailed further on.

Figure 12 illustrates the result obtained with this process and the data of a 1-litre pulse emitter meter.

INDEX

FIGURE 12. TRANSFORMATION OF PULSES INTO FLOW. FIRST MOVING AVERAGE



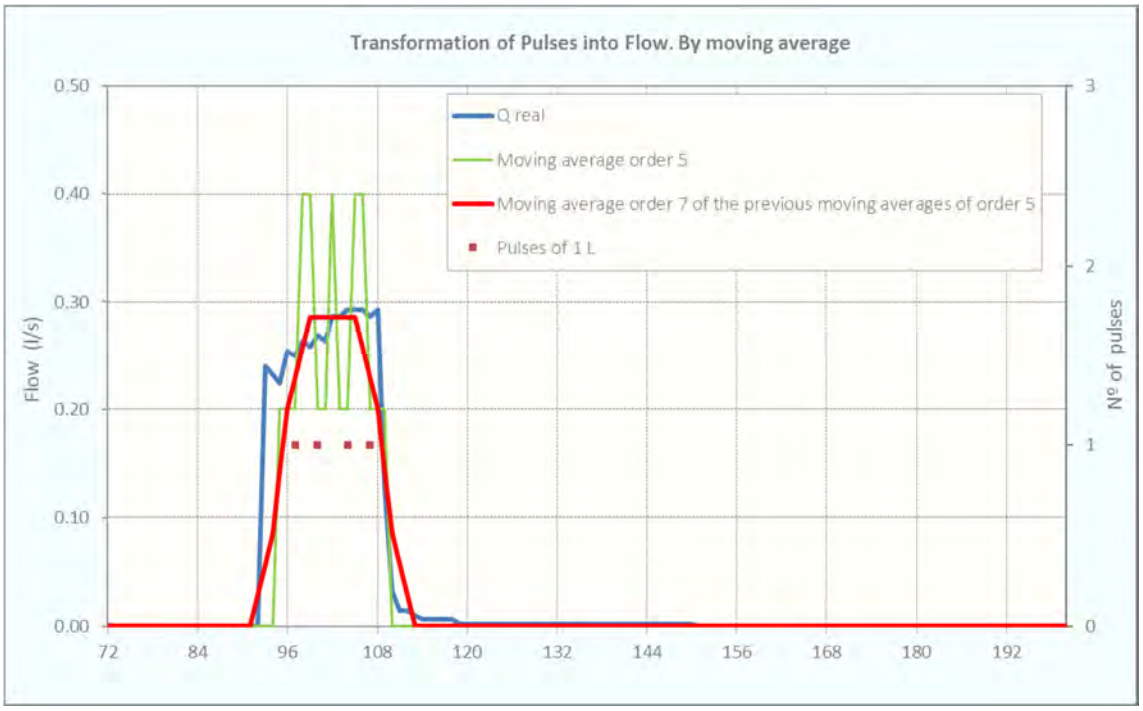
The **blue** colour curve is the real flow that would have been produced by the recorded pulses, indicated with ■, whereas the **green** line represents the flow calculated with the proposed method (which by way of illustration has been represented for a moving average of order 5).

If the process is repeated by calculating a new moving average of the calculated series, the result is much better adjusted, as can be seen in the example attached, this time applying moving averages of order 7 of the series of moving averages of order 5 calculated before.

Figure 13 shows these calculated values, and a much better adjustment is seen between the curve of calculated flows (in **red**) and the curve of real flows (**blue**) than that previously calculated (**green**).

It has been seen that the assumption of one or another value for the order of the moving average has major influence on the quality of the fitting of the results obtained, so it is a parameter that has to be adjusted. The following section describes the procedure to be followed in this.

FIGURE 13. TRANSFORMATION OF PULSES INTO FLOW, BY MOVING AVERAGE



4.1.3. Calculation parameter adjustment

The adjustment of the method used focuses on the determination of the order of the moving average to be applied so the fitting of the curve of calculated flows fits optimally with the curve of real flows.

To assess the goodwill of the results obtained with this methodology, the correlation coefficient and typical error between a curve of flows generated artificially*a priori* is analysed, which will be considered as a curve of "real" flows, and the curve of flows calculated applying the described moving average method.

The proposed adjustment procedure is as follows:

1º) It starts with a known time series of instant flow data from a certain user, which might be real or fictitious.

2º) The time series of pulses that would be generated by said known curve of instant flows is calculated if there were a 1-litre precision pulse meter.

3º) From this series of pulses, the method is applied for different cases of pairs of values corresponding to the orders of the first and second moving average, in order to select the values that give the best results according to the correlation coefficient and typical error.

INDEX

4º) For each pair of values, the different "episodes" of water consumption are identified, an episode being that period of time with a flow other than zero, between one instant with nil flow and the next[15].

5º) The series of real and calculated flows are compared, and the correlation coefficient and typical error are determined.

The correlation coefficient is defined as the ratio between the covariance of the two series of data (X and Y) and the product of the respective standard deviations:

Correlation coefficient:

$$\sigma_{xy} = \frac{Cov\ (X,Y)}{\sigma_x \cdot \sigma_y}$$

With the covariance:

$$Cov\ (X,Y) = \frac{\sum(x-\bar{x})\cdot(y-\bar{y})}{n-1}$$

And the typical deviations of each series:

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})}{n-1}}$$

$$\sigma_y = \sqrt{\frac{\sum(y-\bar{y})}{n-1}}$$

Where $\bar{X}$ and $\bar{y}$ are the respective averages of the two series of data.

The typical error is a measurement of the amount of error in the prognosis of the value and of y for an individual value of x and is given by the following expression:

$$\sqrt{\frac{1}{n-2}\left[\sum(y-\bar{y})^2 - \frac{[\sum(x-\bar{x})(y-\bar{y})]^2}{\sum(x-\bar{x})^2}\right]}$$

6º) Finally, having analysed the results obtained for the different pairs of values, one will be selected to give the optimal combination in relation to the largest correlation coefficient and the smallest typical error.

[15] Do not mistake the term "Episode" for that of "Event" treated in Module 2, and which refers to a period of time in which a certain domestic use occurs. An episode may be the result of a combination of different overlapped events or include a single isolated event. And vice versa, an event may be translated into different flow episodes, as is the case of a clothes washer with several fillings and emptying cycles: each filling produces a different episode.

INDEX

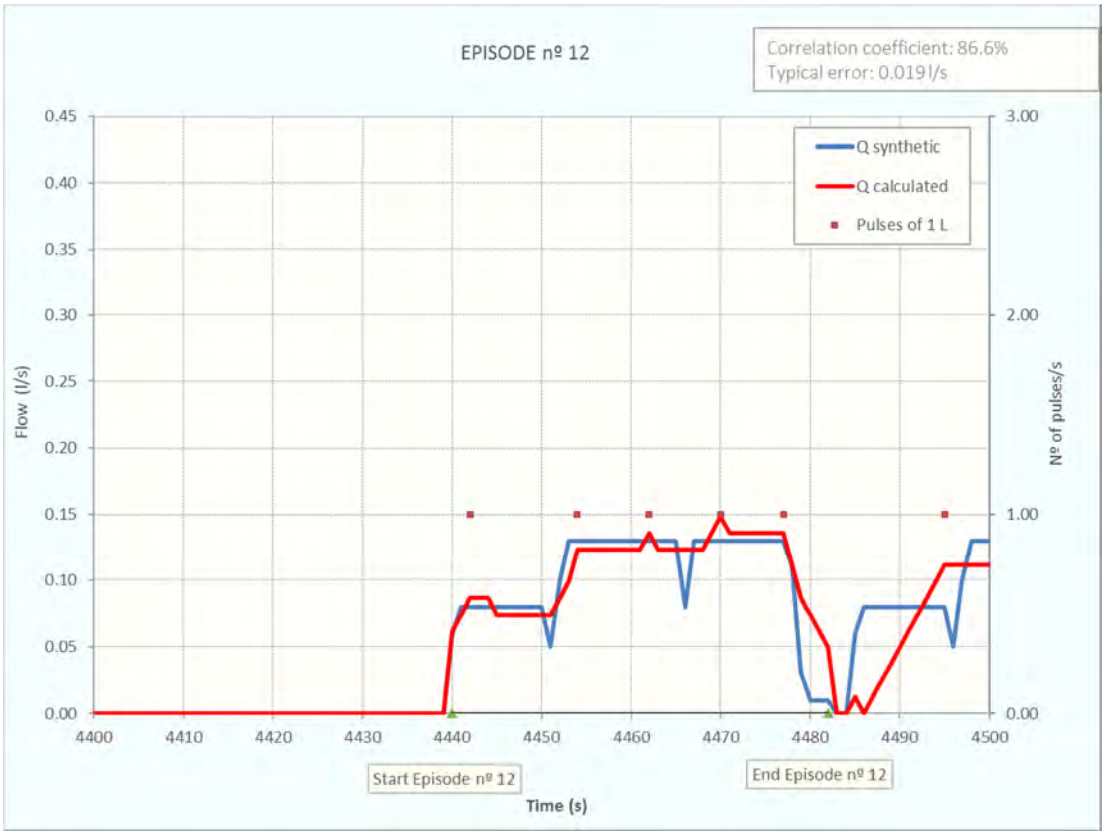
As the two types of meters with different precisions are available (1 and 0.1 litres) and its precision intervenes directly in the method used, the optimal order of the moving average to be applied in either case is also different. The optimal values of this parameter are determined by comparing the results with different pairs of values on real and synthetic series of data.

Finally, the following orders of moving averages are taken:

- 9-9 for meters with a precision of 1 litre
- 3-3 for meters with a precision of 0.1 litre

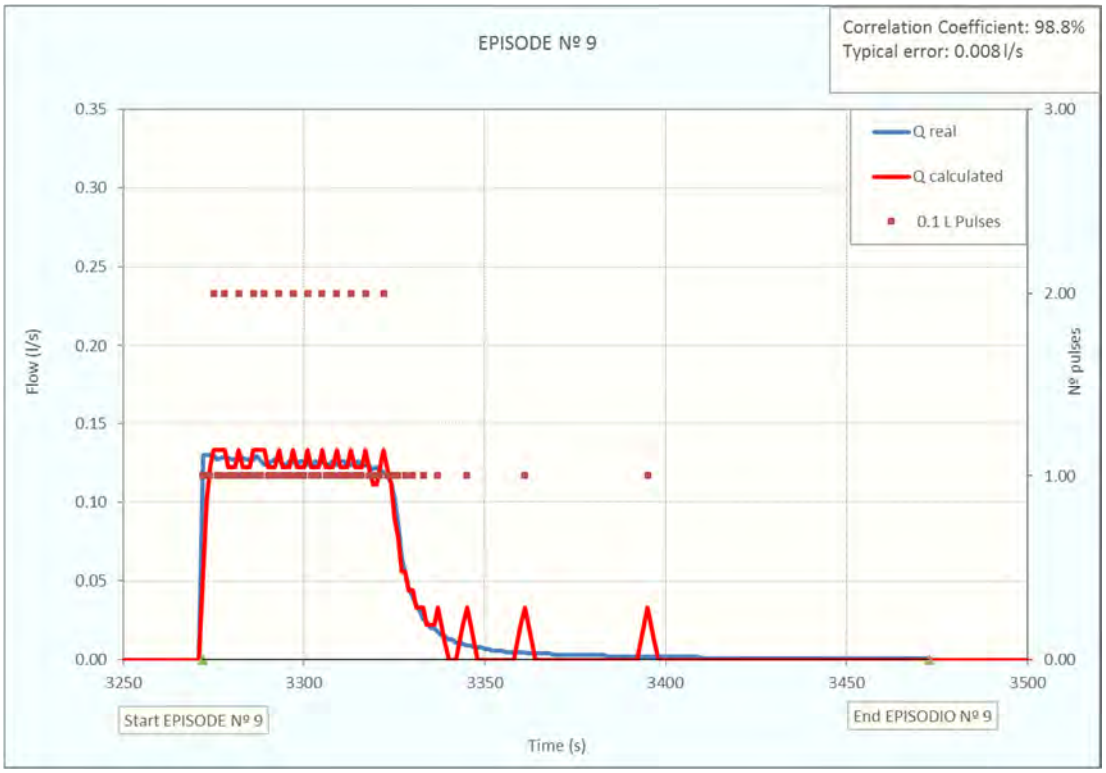
Figures 14 and 15 below illustrate some results of this adjustment.

FIGURE 14. ADJUSTMENT OF THE ORDERS OF MOVING AVERAGES FOR METERS WITH A PRECISION OF 1 LITRE. RESULTS OBTAINED WITH MOVING AVERAGES IN THE ORDER OF 9 AND 9 FOR A SYNTHETIC SERIES OF FLOWS



INDEX

FIGURE 15. ADJUSTMENT OF THE ORDERS OF MOVING AVERAGES FOR METERS WITH A PRECISION OF 0.1 LITRE. RESULTS OBTAINED WITH MOVING AVERAGES IN THE ORDER OF 3 AND 3 FOR A SYNTHETIC SERIES OF FLOWS



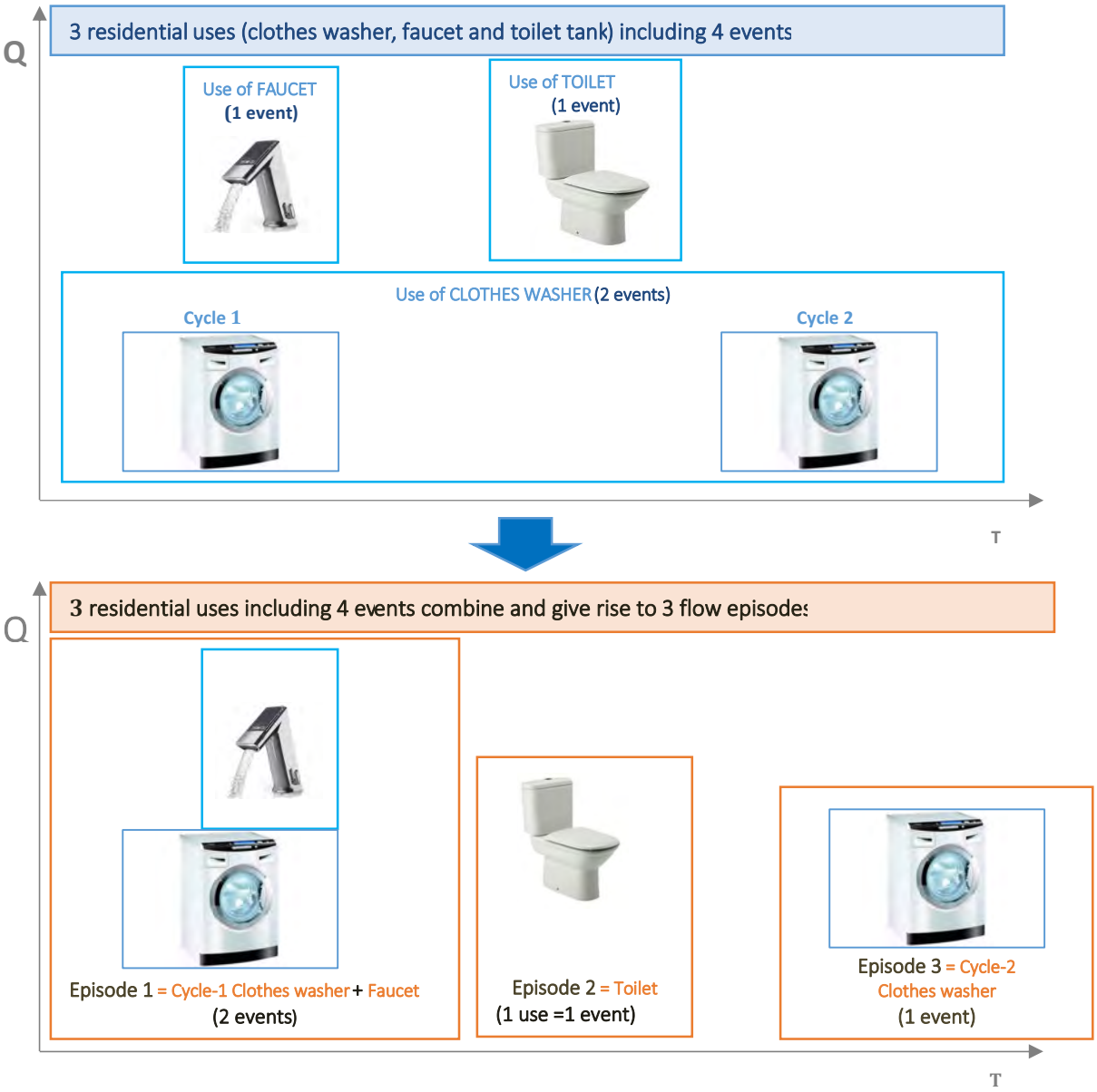
4.2. IDENTIFICATION OF EVENTS

The episodes of flows obtained before respond to different combinations of daily residential uses, such as the opening of a faucet, the starting of a clothes washer or the use of the toilet, whose patterns of consumption could be recognised more easily if they came individually.

Furthermore, each specific use may be translated into one or several elementary units of consumption or events, defined as those periods of sufficient time in which the instant flow remains clearly distinguishable from the rest. Therefore, the use of electrical appliance like a clothes washer, with a programme of several wash and rinse cycles, would produce events that could be distinguished and separated from each other. At a certain time, any of the cycles could coincide in time with the use, for instance of a faucet or a toilet to produce a complex episode (Figure 16).

INDEX

FIGURE 16. FLOW EVENTS EPISODES GENERATED BY DIFFERENT RESIDENTIAL USES



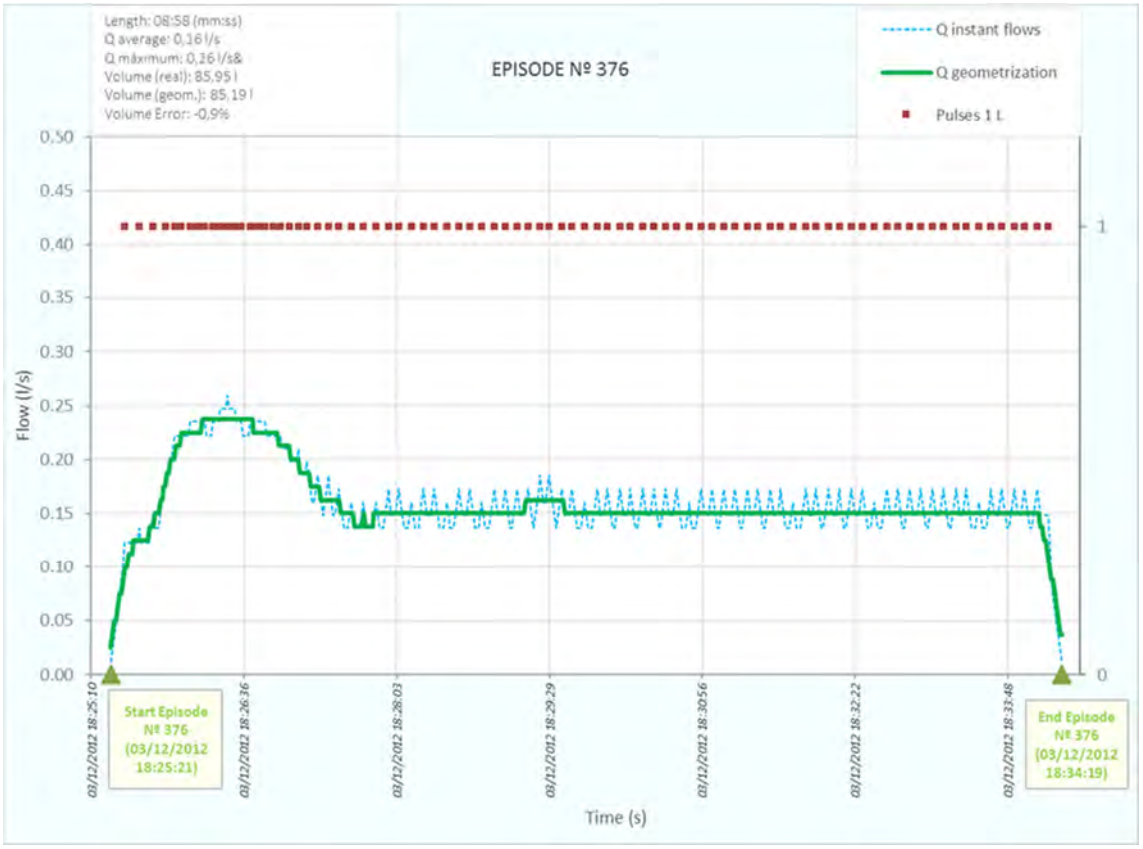
4.2.1. Geometrization of episodes

The episodes obtained with the methodology described above present flow fluctuations, though they are not highly significant *("noise")*, which respond more to the applied method than to a real consumption with frequent opening and closing. These fluctuations make it practically infeasible to achieve an automatic characterisation depending on easily quantifiable parameters.

The proposed methodology simplifies these flow episodes by reducing them to geometric forms consisting on simpler elements *(events)*, assimilable to more or less rectangular or trapezoidal figures. This simplification process has been called ***geometrization*** and its results can be seen in Figure 17, which represents the results obtained from a specific episode (episode nº. 376 of December 2012, and meter 006) by way of example.

INDEX

FIGURE 17. IDENTIFICATION OF EVENTS. GEOMETRIZATION PROCESS



The small flow fluctuations (blue line in Figure 17), the noise produced by the algorithm of transformation of pulses into flows, are eliminated with the proposed*geometrization* (green line in the same figure), allowing a simpler and more automate episode analysis.

The mathematical algorithm that produces this*geometrization* consists of the calculation of the instant flow moving average order of 20 of the instant flows, rounding this measurement to the nearest multiple of 0.0125. Both values, the order of the moving measurement and rounding, have been established experimentally after several tests with different order values to obtain an optimal *geometrization.*

INDEX

4.2.2. Identification of events

As has been said before, an *event* is every elementary unit of consumption occurring in a sufficient period of time in which the instant flow remains clearly distinguishable from the rest. A specific water use may be formed by one or several events, whose possible combination with the events of other uses give rise to a more or less complex flow episode.

It must be said that the *geometrization* of the instant flows is nothing more than a tool conceived to facilitate the identification of the first and last instance of the events forming part of a registered flow episode. The *geometrized* flows are not designed to replace the instant ones, calculated before, but are just an artifice to be able to identify events and assign them a series of parameters, which allow them to be labelled (assigned certain water use) in the development of Module 3.

The episodes are treated as a superimposition of events "stacked" on each other like a staircase, which are identified by the treads on the steps resulting from the*geometrization*.

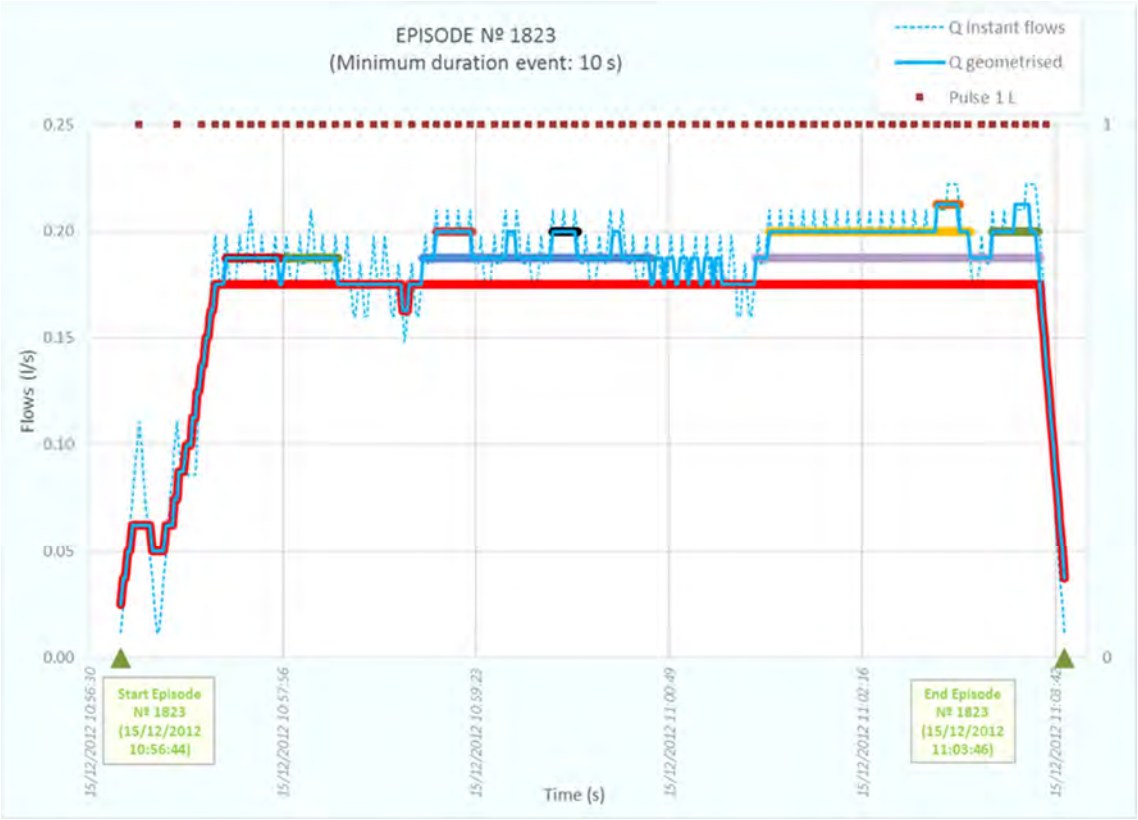
The adopted criteria to discriminating events is the following: it has been considered that if a flow remains constant for a certain time or if a possible flow variation is not maintained (step tread of minimum duration), this will be a unique event. The quantification of this minimum duration has been established empirically by comparing the results for different durations, and having opted for a value of 20 seconds. In other words, a change of event will be considered if a change produces a new flow that is maintained for at least 20 seconds.

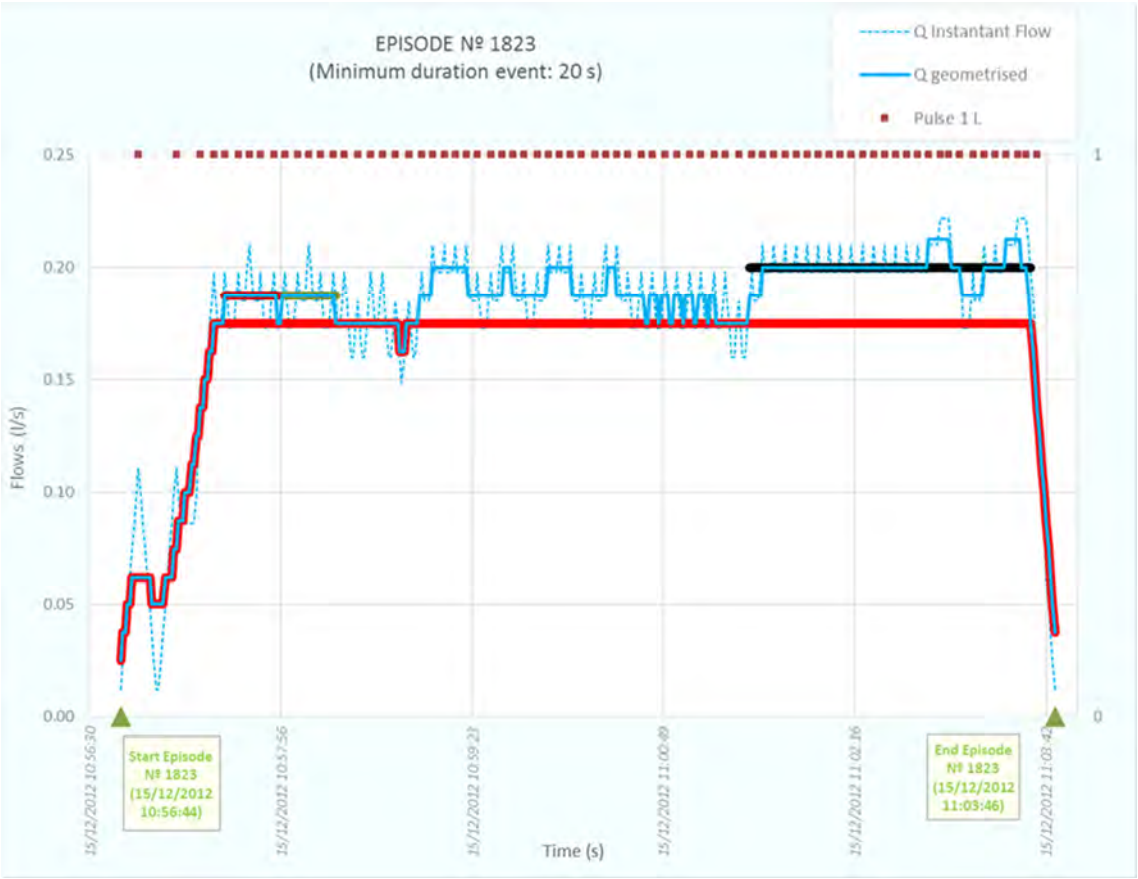
The following figures show the difference between considering a minimum duration of 10 seconds and a minimum duration of 20 seconds. As can be seen in the first case (Figure 18) the episode is split into 10 events, whereas in the second they are reduced to 4 (Figure 19). The results have been compared with those obtained in a similar study, using these same meters but applying another methodology, specifically that of the *Trace Wizard* model having verified that the results obtained by said model are of the same order as those obtained with the duration of 20 seconds.

In short, the division made into events with a minimum event duration of 20 seconds is appropriate.

INDEX

FIGURE 18. IDENTIFICATION OF EVENTS CONSIDERING A MINIMUM DURATION OF 10 AND 20 SECONDS

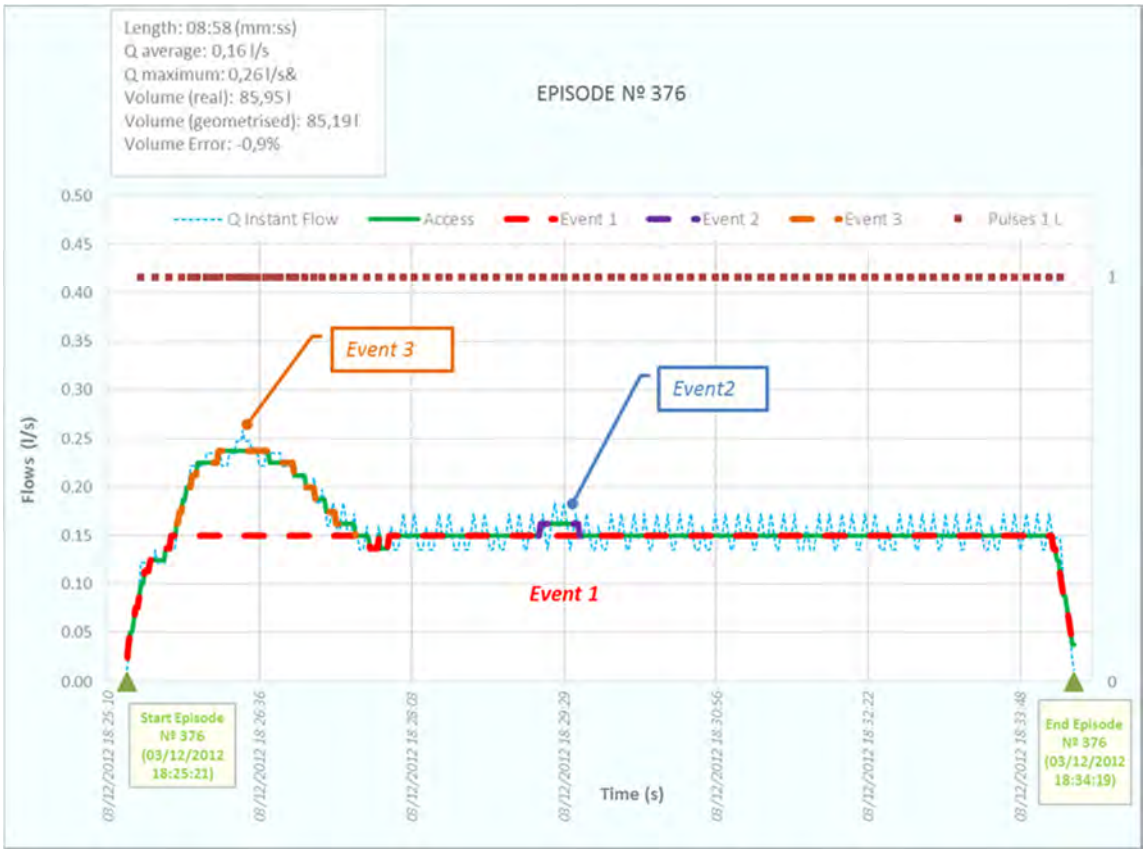




INDEX

Using the Episode nº 376 as an example, Figure 19 clearly marks the flow of 0.15 l/s as a step that defines a first event that can be considered to spread throughout the whole episode. Over it, a much shorter second step is distinguished, but sufficient to be considered an event, and is a flow that is maintained for 22 seconds. And finally, a third event also "supported" by the first, completes the episode.

FIGURE 19. IDENTIFICATION OF EVENTS, EPISODE 376



Once the episodes have been split into elementary events, these should be characterised by the parameters described in the section that follows.

INDEX

4.2.3. Parameters for characterising events

The episode *geometrization* and their decomposition into events allow to identify their first and last instants and so it is possible to define the parameters necessary for their characterisation, which refer both to data from the instant and geometrised flows.

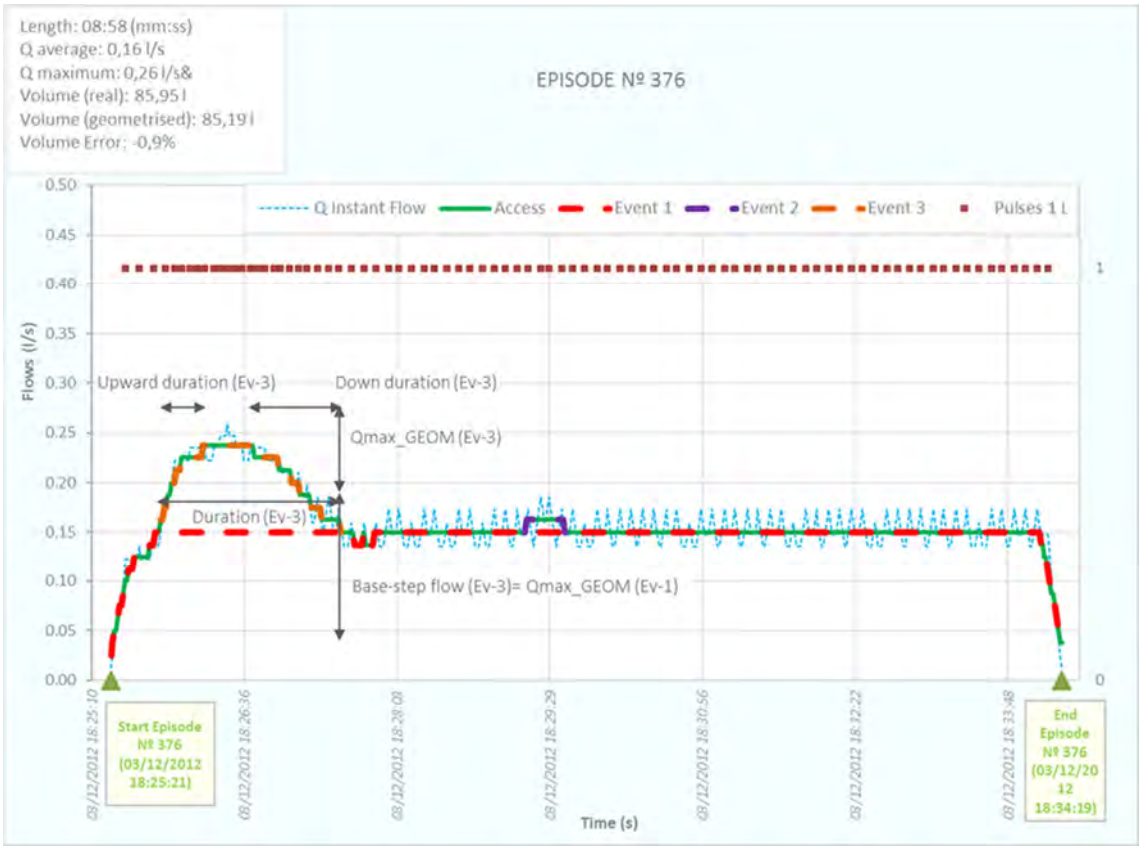
The parameters that have been considered are described below and, by way of example, the values they take the specific case of the above-mentioned***event 3 of episode 376 of December 2012, in meter 006***

- ***IdMeter***: meter identifier.
- ***Month***: the month in which the event occurred.
- ***Episode***: the correlative number identifying the flow episode to which the event belongs.
- ***Event***: the correlative number of the event during the episode.
- ***Start***: the instant when the episode starts.
- ***FinalInstant***: the instant when the episode ends.
- ***Duration***: the time the event lasts.
- ***AscendingDuration***: the duration of the ascending branch.
- ***DescendingDuration***: the duration of the descending branch.
- ***AscendingGradient***: the ratio between the maximum flow and the duration of the ascending branch.
- ***DescendingGradient***: the ratio between the maximum flow and the duration of the descending branch.
- ***Volume***: volume of the event.
- ***Qmax_GEOM***: maximum flow of the geometrized event.
- ***NumSimulEvents***: the number of events registered simultaneously and which lie below the event in question; this indicates the level it occupies on the “stairs of geometrized events”.
- ***StepBaseFlow***: the flow on which the event is “supported”.
- ***Simultaneity***: this indicates the events produced simultaneously along with the event in question.

For a better understanding, Figure 20 graphically shows the parameters of*Duration, AscendingDuration, DescendingDuration StepBaseFlow* and *Qmax_GEOM*, of event 3 of said episode 376.

INDEX

FIGURE 20. PARAMETERS FOR CHARACTERISING EVENTS



4.3. CLASSIFICATION OF EVENTS

Once the events have been identified and characterized, as exposed before, it is time for them to be classified and assigned into the final water use. This "classification" process requires prior learning, which is done by taking already classified events as a reference to serve as a pattern for the different residential uses to be considered.

The events previously classified by operator are those which have been used as a learning pattern for the automatic classification of all the events identified in the previous module.

The classification has been made considering two different methodologies, which depend on the learning method, and are:

- Classification of events by means of *Artificial Neural Networks* with deep learning techniques.
- Classification of events by means of *Support Vector Machines*.

INDEX

The classes of uses used in this classification or labelling were the following:

-Faucets
-Toilets
-Showers, including bathtubs
-Clothes washer
-Dishwasher
-Swimming Pool
-Irrigation
-Leaks

4.3.1. Labelling events by operator

Supervised learning algorithms have been used for automatic event labelling. These kinds of algorithms start with a set of previously classified training data to classify new data by using them on a learning base. To generate this training set, a period has to be processed during which the end use of the events detected are previously known, and the events have to be accurately identified using the methodology described in the previous section.

The periods of time where the events have previously been manually identified by an operator were selected. In parallel, the automatic procedure developed for the identification of events has also been processed.

In order to assign the training data labels, an operator has matched each detected event, with its equivalent label associated to it. The list of events is not unequivocal between platforms, so the event match process was carried out as follows:

1. For each user, each event obtained with the automatic procedure (event identification module) is compared with the operator labelled events that totally or partially coincide in time, in order to select the operator labelled events with a start instant before the end of the event of the automatic procedure, and its final instant after the time when this started.

2. Having selected the coinciding events, their features are analysed to find the most similar one. A method based on two comparisons is used to do this:

   a) The Jaccard index

   b) The volume of the events

   The Jaccard index is a coefficient that measures the degree of similarity between two sets, regardless of the elements they contain. This similarity is measured as the number of elements present in the intersection of both sets divided by the total number, in other words:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

   Therefore, the Jaccard index is calculated as the duration (seconds) in which the event of module 2 and the event labelled by operator coincide in time (seconds) over the period of time from the beginning of the first event to the end of the last event, in other words:

$$J(ev_{TW}, ev_{M2}) = \frac{duration\ (ev_{TW}\ \cap\ ev_{M2})}{duration\ (ev_{TW}\ \cup\ ev_{M2})}$$

INDEX

The event with higher Jaccard index will be selected, although this must be ratified by the comparison of volumes. The comparison of the volume is necessary, because the events are not divided with the same criterion in the events identification module as by the operator, and therefore the time in which the events coincide may not be reliable. This comparison is a concealed tolerance, as it ratifies or undoes the assignment based on the Jaccard index. If either of the two volumes is the triple of the other, in other words if:

$$\frac{Vol(ev_{TW})}{Vol(ev_{M2})} < \frac{1}{3} \quad o \quad \frac{Vol(ev_{TW})}{Vol(ev_{M2})} > 3$$

Then the rest of the candidates are studied to see which has a more similar volume.

4.3.2. Input variables

The set of input data of the algorithm is composed by 37 variables. Some of these variables have been calculated previously by the event identification module (8 variables). The rest of them (29 variables) are obtained in real-time in order to have more useful information when classifying and assigning the water use of each event, thus allowing the creation of predictive models that classify the events correctly.

For certain classification algorithms, a selection of variables must be made due to the algorithmic complexity of having a relatively large set of data (variables and observations). Within the framework of this project, this selection of variables was necessary for the classification based on support vector machines, but it was not necessary for the classification using neural networks, due to its great computing and learning power has allowed to use every variable, and also more observations.

When the amount of data is large enough, the neural networks are capable of automatically selecting the variables that are useful to classify a problem. This is possible as they can multiply each input variable by a weight, and this will be close to zero in the variables that do not provide useful information.

In the case of support vector machines, the right selection of variables is necessary for the correct operation of the algorithm. This selection was made independently for the meters of 0.1 and 1 litres and the procedure used was commonly known as**greedy**, which consists of an iterative process in which the most useful variables are selected by running the algorithm using each variable individually; then, to select the second variable, it is again executed using the winning variable, and each one of the other variables.

Table 3 shows the variables selected for each kind of meter. The Table 3 shows that more variables in the 0.1-litre meters have been selected. This is because, when the 1-litre meters are used, the lack of precision in the flow measurement makes that many variables do not provide useful information.

INDEX

TABLE 3. INPUT VARIABLES FOR EVENT PRE-PROCESSING

Variable	*Description*	*1-litre type meter*	*0.1-litre type meter*
Duration	Duration (seconds) of the event	✓	
AscendingDuration	Duration (seconds) of the ascending branch of the event	✓	✓
DescendingDuration	Duration (seconds) of the descending branch of the event	✓	✓
AscendingGradient	Gradient of the ascending branch	✓	✓
DescendingGradient	Gradient of the descending branch	✓	✓
Volume	Volume of the event in litres		✓
Qmax_GEOM	Maximum flow (litres per second) geometrised	✓	✓
StepBaseFlow	Maximum flow (litres per second) of the event with which it is superimposed	✓	✓
EpisodeTime	Start time of the episode in which the event occurs		✓
EpisodeVolume	Total volume (litres) of the episode in which the event occurs	✓	✓
EpisodeDuration	Total duration (seconds) of the episode in which the event occurs	✓	✓
Volume of the four previous episodes	4 variables representing the volume (litres) of each of the four episodes before the episode in which the event occurs		✓
Duration of the four previous episodes	4 variables representing the duration (seconds) of each of the four episodes before the episode in which the event occurs		✓
Time passed since the four previous episodes	4 variables representing the time passed (seconds) since the end of each of the four previous episodes until the start of the episode in which the event occurs		✓
Distance to the four previous episodes	4 variables representing the time passed (seconds) since the end of the episode in which the event occurs until the start of each of the four previous episodes		✓
Volume of the episode with a volume of more than 1 litre closest in time	Volume (litres) of the closest episode in time of between the 5 previous episodes and 5 later with a volume of over 1 litre		✓
Time passed between the episode with a volume of more than 1 litre closest in time	Time passed (seconds) of the episode in which the event occurs to the closest episode of the 5 previous episodes and 5 later episodes with a volume of over 1 litre		✓

INDEX

4.3.3. Feature Normalization

The input algorithm variables have been submitted to processes of normalization as part of the habitual procedure of data analysis, prior to the development of classification and prediction models.

The normalization consists on transforming the data so that all of the variables are measured on the same scale. It is therefore possible to make comparisons between them that are independent of the unit of measure used in each one and avoiding that some of them could have more weight than others in the models developed later.

There are different techniques to normalize invariant scale variables. In the classification model based on support vector machines developed as part of this project, the chosen transformation converts the input variables into variables with average 0 and standard deviation 1, following a **standard score**:

$$\tilde{X} = \frac{X - \mu}{\sigma}$$

Where $\tilde{X}$ is the X standardized variable, μ y σ are the average and standard deviation of the variable X.

However, for classification with neural networks, the variable normalization turns them into the interval [0- 0.1]. It therefore uses the following expression $X' = X - min(\mathrm{X})$

$$\tilde{X} = \frac{X'}{max(X')} * 0.1$$

Where $\tilde{X}$ is the X normalized variable.

4.3.4. Classification of events by means of Artificial Neural Networks with deep learning techniques

A neural network is a mathematical tool that models the brain's operation in a simplified manner. In simplistic terms, it is a series of mathematical operations on an input vector that results in another output vector.

The perceptron

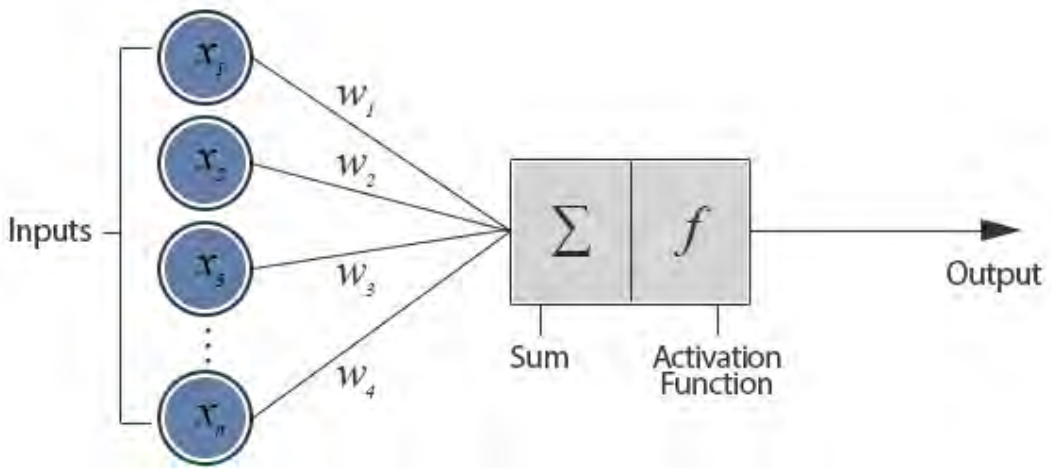
The basic computation element is usually called ***node*** or ***neuron***. Although there are many types of neurons (perceptron, RBF, self-organised maps, recurrent, etc.) the most widespread (as it is general and applicable to all kinds of problems) is the ***perceptron***.

As input, the perceptron receives a series of variables from an external data source. Each input has an associated weight ***w***, that changes throughout the learning process.

Each unit applies an activation function ***f*** over the sum of the inputs weighted using the weights. Figure 21 chose the structure of a perceptron.

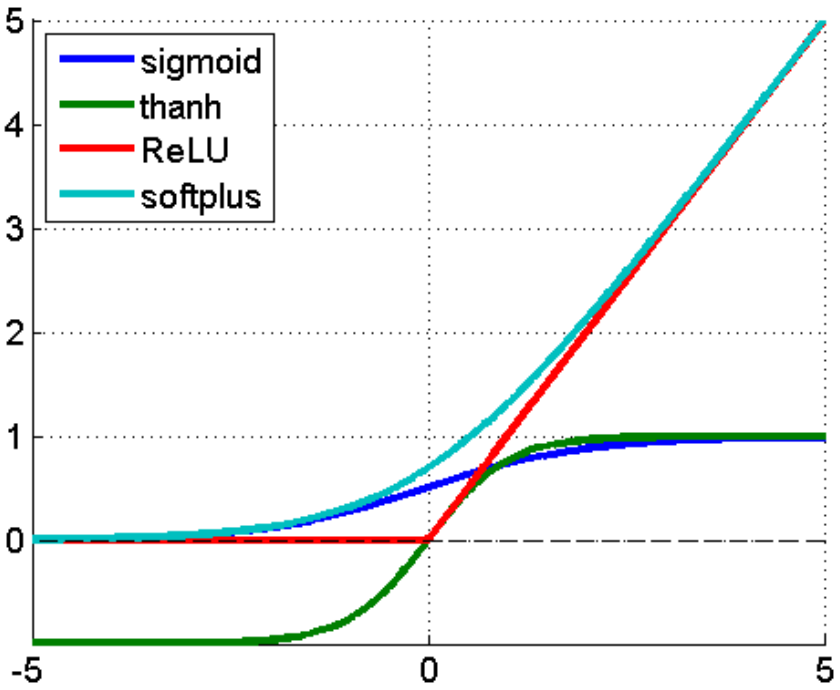
INDEX

FIGURE 21. ARTIFICIAL NEURAL NETWORKS. STRUCTURE OF A PERCEPTRON



Regarding activation functions, there are many alternatives: sigmoid, hyperbolic tangent, rectified linear (ReLU), softplus, etc. Figure 22 shows a few examples of activation function, representing the output value in accordance with the input.

FIGURE 22. ARTIFICIAL NEURAL NETWORKS. ACTIVATION FUNCTION



INDEX

Multilayer perceptron

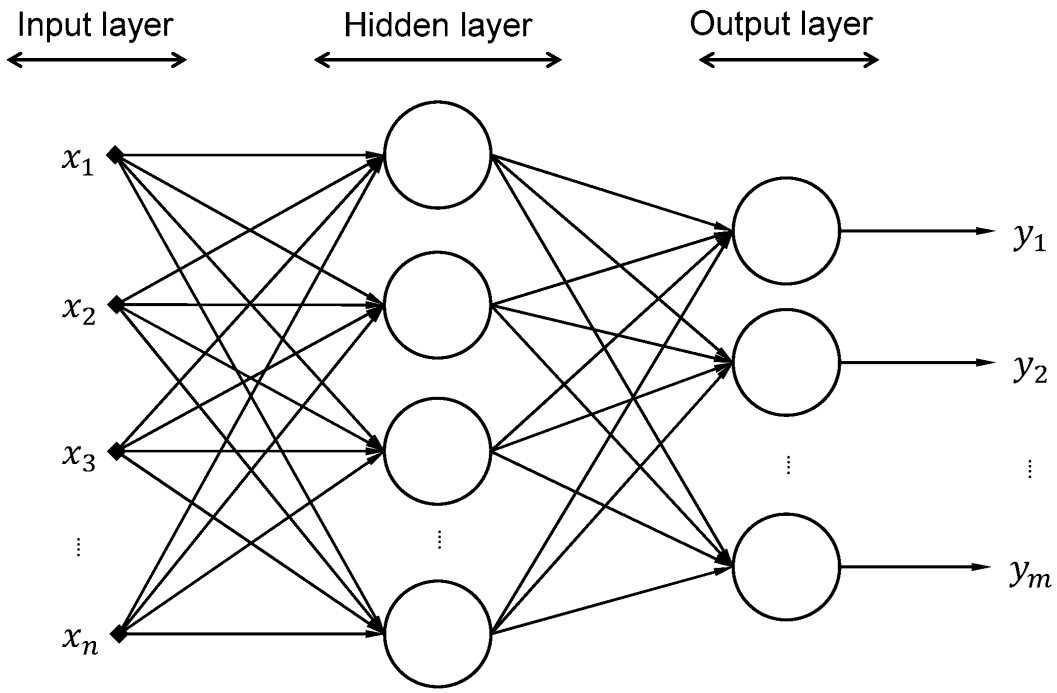
Supposing a perceptron is used to classify data problems, only linearly separable problems in the input feature space will be solved. In order to avoid this limitation, it is possible to create anArtificial Neural Network (ANN) formed by multiple layers[16]. This structure could solve non-linear problems (Figure 23).

When a neural network is built using perceptron, it is called a**multilayer perceptron**, where the output of every neuron, in one layer, are the inputs of every neuron, in the next layer. In accordance with the position of the layers, they can be classified into three types:

- **Input**: they receive the input parameters (input vector).
- **Output**: they produce the result values (output vector).
- **Hidden**: they contain the intermediate calculations.

Figure 23 shows a typical example of a neural network.

FIGURE 23. EXAMPLE OF AN ARTIFICIAL NEURAL NETWORK. MULTILAYER PERCEPTRON



A typical neural network is shaped by an input layer with a number of neurons equal to the number of variables of the dataset, and output layer with a number of neurons equal to the number of categories into which the data can be classified and several hidden layers dealing with processing the data.

To obtain the value of the weights of each neuron of the network, the**backpropagation** algorithm is used.

[16] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer-Verlag New York. Inc. Secaucus, NJ, USA.

INDEX

Backpropagation

Backpropagation is a supervised learning algorithm used to train artificial neural networks.

The algorithm uses a two-phase propagation-updating cycle. Once a pattern has been applied to the network input as a stimulus, it is propagated from the first layer through the upper layers of the network to produce an output. The output signal is compared with the true output and an error signal is calculated for each of the outputs.

The error is backpropagated from the output layer to every neuron of the hidden layer that contributes directly to the output. However, the neurons in the hidden layer only receive a fraction of the total error signal, based roughly on the relative contribution that each neuron has made to the original output. This process is repeated layer by layer until all neurons of the network have received an error signal that describes their relative contribution to the total error.

Using this process, as the network is trained, the neurons of the intermediate layers organise themselves so that different neurons learn to recognise different features of the total input space. After the training, when they are presented with an arbitrary input pattern that contains noise or is incomplete, the neurons of the hidden layer of the network will answer with an active output if the new input contains a pattern similar to those learnt in training.

Acceleration of training

Usually, before the neural network training begins the neurons weights are initialised at random.

One problem with the backpropagation algorithms is that the error is diluted exponentially as it crosses layers on its way to the beginning of the network. This is a problem because in a very deep network (with a lot of hidden layers), only the last layers are trained, whereas the first remain virtually unchanged. Therefore, in the past it was more practical to use networks with a few hidden layers containing a lot of neurons instead of networks with many layers containing few neurons. This was like this until improvements were developed to train networks with many layers.

Stacked autoencoders

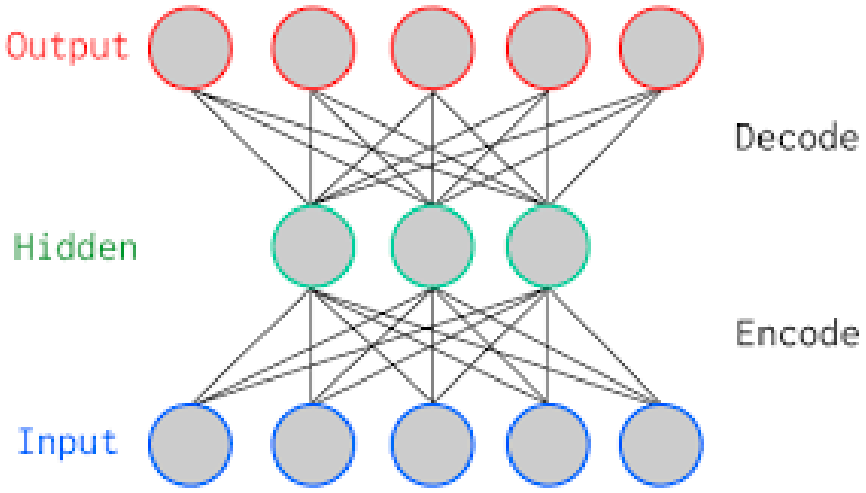
This technique has been used in this project to improve the performances and accelerate the training of the deep neural networks. Stacked autoencoders are a technique used to initialise the weights of the neural networks that make the training faster, thus creating networks with many hidden layers*(Deep Learning)* in a reasonable training time[17].

An **autoencoder** is a neural network with a single hidden layer that learns to produce exactly the same information on the output as it receives on the input. Therefore, the input and output layers must always have the same number of neurons. For instance, if the input layer receives the pixels of an image, the network is expected to learn to produce exactly the same image in its output layer, something which is not useful at first sight. The structure can be seen in Figure 24.

[17] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). *Stacked Denoising Autoencoders: Learning useful representations in a deep network with a local denoising criterion.* Journal of Machine Learning Research, 11 (Dec), 3371-3408

INDEX

FIGURE 24. ARTIFICIAL NEURAL NETWORK. STACKED AUTOENCODERS



The key lies in the hidden layer. Let's suppose we have an autoencoder with fewer neurons in the hidden layer than in the input and output layers. As this network is required to produce the same result on the output as it receives on the input, and the information has to pass through the layer, the network will be forced to find an intermediate representation of the information in its hidden layer using fewer numbers. Therefore, when input values are applied, the hidden layer will have a compressed version of the information, but it will also be a compressed version that can be decompressed again to recover the original version on the output.

In fact, once trained, the network may be divided in two: a first network that uses the hidden layer as an output layer and a second network that uses this hidden layer as an input layer. The first network would be a compressor and the second a decompressor.

Precisely for this reason, these kinds of networks are called autoencoders as they are able to discover a way to encode the information on its hidden layer, for themselves. One of the advantages is that they do not need a supervisor to show them examples of how to code information.

Adding noise to the input data

There is a technique that ensures that the network does not simply copy the information when the hidden layer has more neurons than necessary. This technique, instead of indicating the same value for the inputs and outputs, introduces **noise in the input vector** and leaves the output without noise when the examples are composed. The network is thus forced to generalise because it will have several slightly different examples of inputs that produce the same output.

The intermediate representation in the layer will have to be focused on the common features of all versions of the same data with different noises.

Stacked autoencoders

A single autoencoder can detect fundamental features in the input information. However, to create machines capable to detect more complex concepts, more power will be needed in the machine.

For instance, from raw information without meaning (for example pixels of images) an autoencoder is capable of detecting simple features (lines and curves). If the coded result in this hidden layer is applied another autoencoder, it will be capable of finding more complex features (such as circles, arcs, straight angles, etc.). If this process is repeated, a hierarchy of ever more complex features will be obtained along with a stack of encoders. Following the example of the images, given enough depth and sufficient example images, some neurons will be activated when the image has a face and without any supervisor having to tell the network what a face is like.

The idea of *Deep Learning* via stacked autoencoders is precisely that: using several encoders and training them one by one, using each trained encoder to train the next.

After initialising the neural network with stacked autoencoders, when it is trained through the backpropagation algorithm, the training ends in a much shorter time and with the need for fewer training data.

Ensemble with other algorithms

A technique widely used to accelerate the automatic training of complex algorithms is to add variables which give information that is useful to classify[18]. For example, by training much simpler and faster algorithms and including their output as a new input variable in the neural network.

Within this project, a fast and simple algorithm was used called**gradient boosted trees**[19]. This algorithm builds a chain of decision trees in which each tree tries to solve the errors of the previous.

The prediction of this algorithm is used as a new input in the neural network. The network thus starts with this knowledge, and its efforts focus on learning to classify the more complex problems.

Training and test criteria

In order to correctly validate the performance of the neural networks in these databases, a training and testing process was used to verify the accuracy when processing data that had not been seen during the training procedure.

Before processing all the meters with the definitive parameters, to select the parameters that achieve the highest accuracy, a previous process was carried out.

It was started by making tests to select the parameters indicating the depth of the tree and the number of estimators of the gradient boosted trees. In view of the results, it was decided to maintain the number of estimators at 100 and the depth of the tree at 11 levels.

[18] Zhang, C., & Ma, Y. (Eds.). (2012).*Ensemble Machine Learning: Methods and Applications* Springer Science & Business Media.
[19] Chen, T., & Guestrin, C. (2016). Xgboost:*A Scalable Tree Boosting System.* Proceedings of the KDD 2016 Conference.

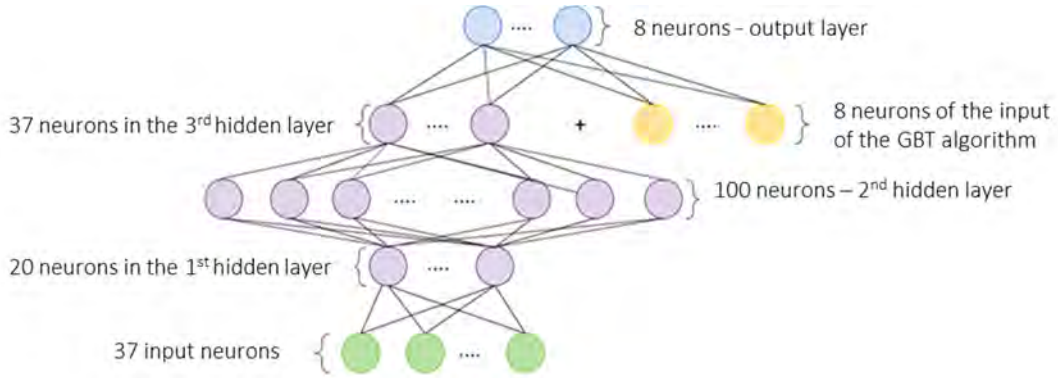
INDEX

The number of training epochs indicates the number of times that the total set of training events is going to be used by the neurone to update the weights, and is one of the parameters of the neural network.

In terms of Neural Networks, all meters have been executed by validating the number of times (with values of 20, 50, 100) and the number of layers and neurones of the neural network. The results showed that the architecture of the neural network was not the most important thing, but that the results improved as the number of training epochs was increased. It was therefore decided to set the number of epochs at 100 and the number of layers and neurons as follows (see Figure 25):

- 37 input neurons. One for each input variable.
- 20 neurons in the first hidden layer.
- 100 neurons in the second hidden layer.
- 20 neurons in the third layer + 8 input neurons of the gradient boosted trees.
- 8 neurons in the output layer. One for each kind of event.

FIGURE 25. ARTIFICIAL NEURAL NETWORKS. TRAINING PROCESS

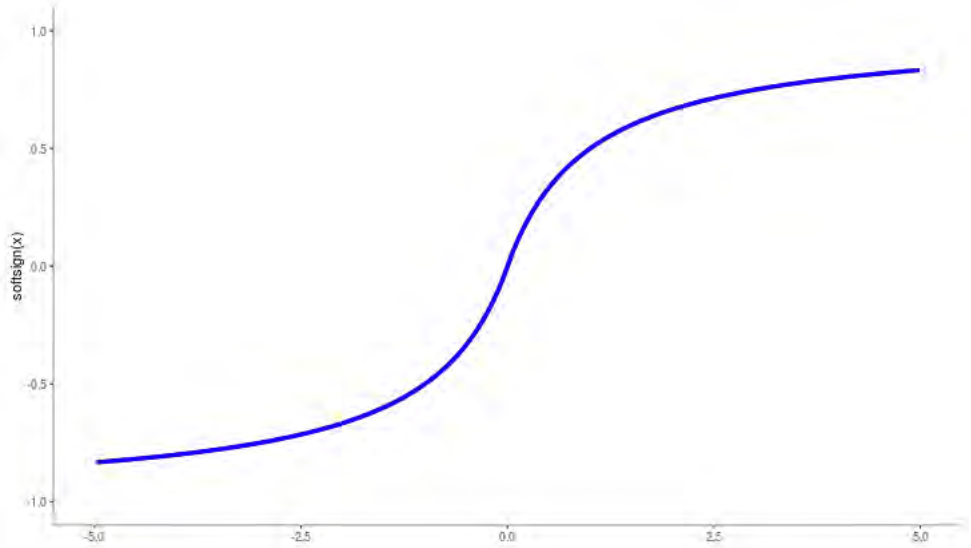


After making tests with several types of activation functions in the hidden layers, the ***softsign*** function was selected, which is represented in Figure 26, as it is the one that achieved the best result:

$$f(x) = \frac{x}{1 + |x|}$$

INDEX

FIGURE 26. FUNCTION OF ACTIVATION IN THE HIDDEN LAYERS: SOFTTSIGN



In the case of output neurons, the activation function was ***softmax***, which is used in multi class problems where the output must only belong to a single category. This function has the expression:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

Where x_i is the output of neuron i in **N** total output neurons.

a) Test Data

When the data to generate the models are loaded, 30% of the data were left out in order to measure the accuracy of the method with events that have not previously been used to train. In other words, a test set was accomplished using 30% of the observations.

With the results of the models trained with the standardised and selected variables the error rate is calculated, as well as the precision rate in litres and the confusion matrix in the assignment of end uses.

b) Training Data

As the neural networks do not have an excessively high computational cost, the number of events used for training and test was limited to 100,000, in other words, in most meters all the events were used. Of the total number of events, 70% are used to train the neural network.

In the first phase of the training, the autoencoders are trained to initialise the weights of the neural network. In each of the input layers of the autoencoders, a Gaussian noise is added with a standard deviation of 0.01.

Then the events are processed with gradient boosted trees. The prediction of this algorithm is added as an input to the neural network.

c) Development and implementation

The described algorithms were developed and implemented in the programming language *Python* 2.7[20], which is software developed on *Open Source* license scheme, that allows its free use and distribution. *Anaconda,*[21] the *Python* installer was used with a set of packages aimed at automatic learning and data analysis, also with *Open Source* licence.

The packages used for developing the neural network models were:

- *Scikit-learn*[22] is an ***Open Source Automatic Learning*** library for *Python* programming environments. It includes the implementation of the most important state-of-the-art algorithms in classification, regression, Bayesian methods, etc. It is highly effective thanks to the use of the *NumPy* and *SciPy* numerical and scientific libraries, strongly optimised for the performance of mathematical calculations. **Version 0.17.1** of this library was the version used.

- *NumPy*[23] is the library entrusted with giving high-level mathematical and vectorial support for operating with vectors and matrices. The **version** used is **1.11.1**.

- *Pyodbc*[24] is the library that manages the connection to the Access database via the use of the ODBC standard. The **version** used is **3.0.10**.

- *Theano*[25] is the *Python* library that allows the definition, optimisation and efficient evaluation of mathematical expressions that involve multidimensional arrays. The **version** used is **0.8.2**.

  The features of *Theano* include:

  - ✓ Integration with *Numpy*.
  - ✓ Use of user transparent GPU. It makes intensive calculations 140 times faster than by using the CPU.
  - ✓ Efficient symbolic differentiation. *Theano* makes derivatives with one or several inputs.
  - ✓ Speed and stability in the optimisations. It obtains the correct result for *log(1+x)* even when the ***x*** is really small.
  - ✓ Dynamic code generation in **C**. It evaluates the expressions faster.
  - ✓ Extensive testing and self-verification unit. It detects and diagnosis several kinds of errors.

[20] https://www.python.org
[21] https://www.continuum.io/anaconda-overview
[22] http://scikit-learn.org/
[23] http://www.numpy.org/
[24] https://pypi.python.org/pypi/pyodbc
[25] http://deeplearning.net/software/theano/

INDEX

- *Keras*[26] is a minimalistic, highly modular neural network library written in*Python* and can be run at the top of *Theano*. It was developed in order to help the researcher to reach the result in the shortest possible time. The **version** used is **1.0.8**.

  The main features of *Keras* are:

  - ✓ It allows the fast and easy creation of prototypes (through the total modularity, minimalism and extendability).
  - ✓ It supports both convolutional networks and recurrent networks, and combinations of both.
  - ✓ It supports different kinds of connections between neurons (including multi-input and the formation of multiple outputs).
  - ✓ It is easily run on the CPU and the GPU.

- *XGBoost*[27] is an open code library compatible with *Python*, which implements GBT. The **version** used is **0.6.**

4.3.5. Classification of events using Support Vector Machines

Support Vector Machines[28] or **SVM**, are a set of automatic learning algorithms that are used to solve classification or regression problems.

To understand the motivation of the SVM, it suffices to imagine a set of points where each one belongs to a single class (if, for instance, there are two classes: red and blue, a single point may either be in the red class or in the blue class). The SVM will deal with finding the function that correctly separates both categories data (black line in Figure 27).

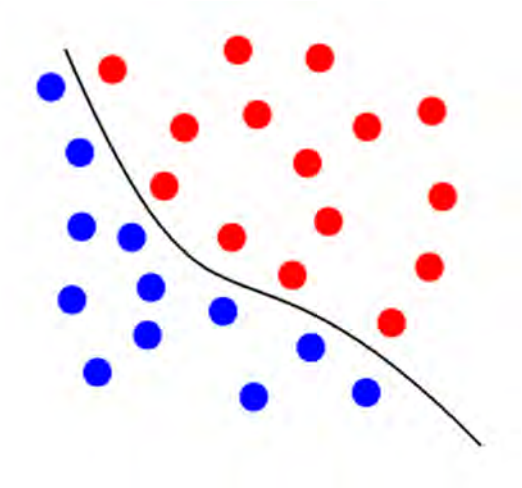
Something as simple to explain conceptually is truly complex in practice, as the boundaries between classes are not as clear as in Figure 27. Very often it is not possible to find the boundary correctly separating the classes, or simply having more than two classes adds great complexity to the problem.

In this section, a technical explanation of the SVM formulation and the algorithm chosen to solve the classification problem will be explained.

[26] https://keras.io/
[27] http://xgboost.readthedocs.io/
[28] Vapnik, V. *The nature of statistical learning theory*. Springer Science & Business Media, 2013

INDEX

FIGURE 27. MOTIVATION OF THE SVM



Support Vector Machines for binary classification

SVMs were originally designed to solve binary classification problems (problems in which the data can belong to two different classes or categories): one is considered positive ($y = 1$) and the other negative ($y = -1$).

There will therefore be a training set $\mathcal{D}$ composed by n observations in a space of p variables

$$\mathcal{D} = \{(\boldsymbol{x_i}, y_i) | \boldsymbol{x_i} \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

Where $\boldsymbol{y_i}$ is the class of the observation $\boldsymbol{x_i}$. (the note in bold $\boldsymbol{x_i}$ indicates that it is a vector).

a) Maximum margin classifier

The SVM originally stem from the maximum margin classifier. An alternative when there is a problem that is perfectly classifiable by a plane is to find the plane with the largest distance to the data.

The formulation[29] of the problem to be solved is as follows: find a hyper-plane separating the labelled examples, in other words,

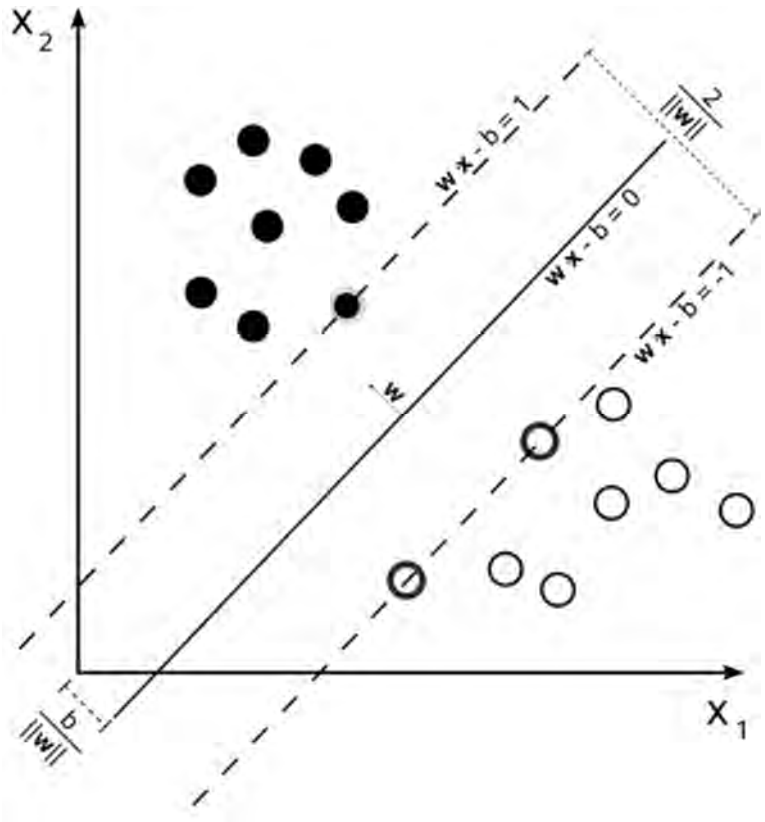
$$Maximise: \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$Subject\ to: y_i(\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1, i = 1,2,\dots,n$$

[29] Vapnik, V. *Pattern recognition using generalized portrait method* Automation and remote control, 1963, 24:774–780

INDEX

The plane that separates both classes is $\boldsymbol{wx_i} + b = 0$ and the parallel planes with maximum margin are $\boldsymbol{wx_i} + b = \pm 1$ respectively, so the problem comes down to seeking the values of $\boldsymbol{W}$ and $\boldsymbol{b}$ which maximise the margin (distance between the parallel planes) and which guarantee that all the examples of the training set are correctly classified (see Figure 28).

FIGURE 28. SUPPORT VECTOR MACHINES. MAXIMUM MARGIN CLASSIFIER



b) Linear soft margin SVM

As most problems cannot be separated by a plane, this classifier was improved by including what it is called "soft margin" which allows the existence of certain data which do not meet the conditions of being separated by the maximum margin classifier.

For each observation $(\boldsymbol{x_i}, \boldsymbol{y_i})$, a loss or a gap is allowed $\boldsymbol{\xi i}$, and the restrictions of what is considered a *well classified* example are relaxed. The training example $(\boldsymbol{x_i}, \boldsymbol{y_i}) \in \boldsymbol{D}$ is considered well classified if it checks:

$$w \cdot x_i + b \geq +1 - \xi_i \text{ for } y_i = +1$$

$$w \cdot x_i + b \leq -1 + \xi_i \quad for\ y_i = -1$$

$$with\ \xi_i \geq 0\ \forall i \in \{1, \ldots, p\}$$

INDEX

The amount of losses on the training set is adapted with the parameter C:

- With a value of high $\boldsymbol{C}$ the algorithm will strive to correctly classify the observations of the training set, but will run the risk of not being able to generalise and correctly classify new observations.

- On the other hand, for low values of $\boldsymbol{C}$ the system will obtain general solutions that will work both for observations seen during the training and for the new, but there will be a risk of obtaining a very simple or low accurate solution.

By adding these new conditions, the formula of the problem becomes:

$$Maximise: \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$Subject\ to: \begin{matrix} y_i(\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1, i = 1,2,\dots,n \\ \xi_i \geq 0, i = 1,2,\dots,n \end{matrix}$$

This is a problem of convex optimisation that guarantees a single solution.

Sometimes the complexity in the formula of a problem requires a search for an alternative formulation with known resolution. This is what is known as a dual formulation. In this case, it may be done as a Lagrange formula and may be solved by quadratic programming:

$$Maximise\ F(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j \boldsymbol{x}_i\boldsymbol{x}_j$$
$$Subject\ to: \begin{cases} \sum_{i=1}^{n}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases}$$

Where α_i is the Lagrange multiplier associated with the sample $\boldsymbol{x}_i$.

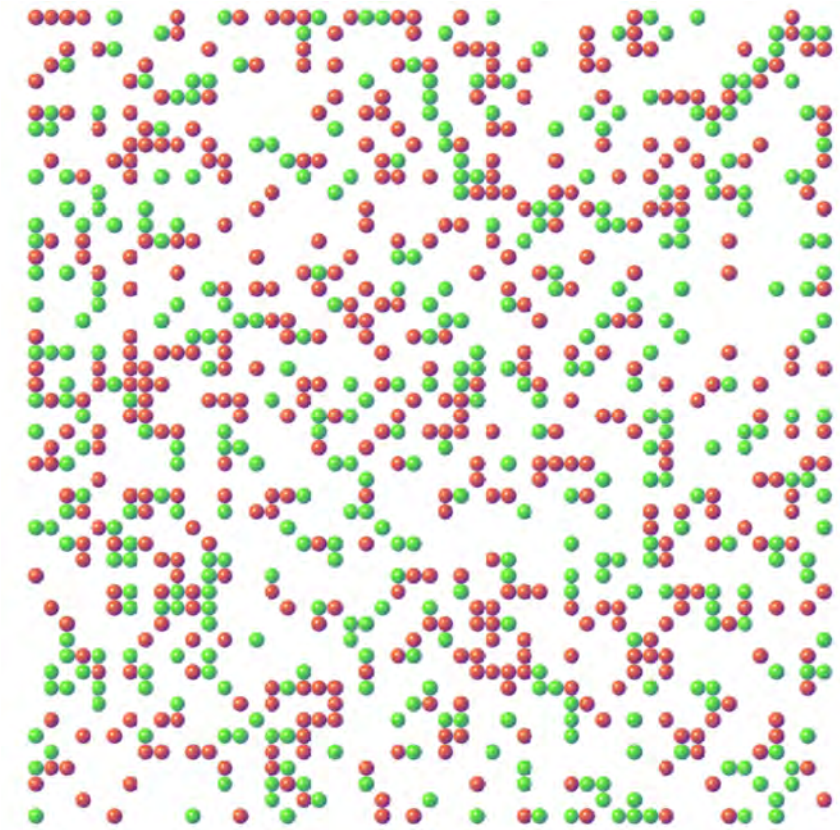
SVM for non-linear classification

When classifying, the use of simple plane to separate both kinds of data is not the best option, as in real life there are problems that are highly complex in nature. It is therefore necessary to find methods to obtain non-linear separators, as illustrated in Figure 29.

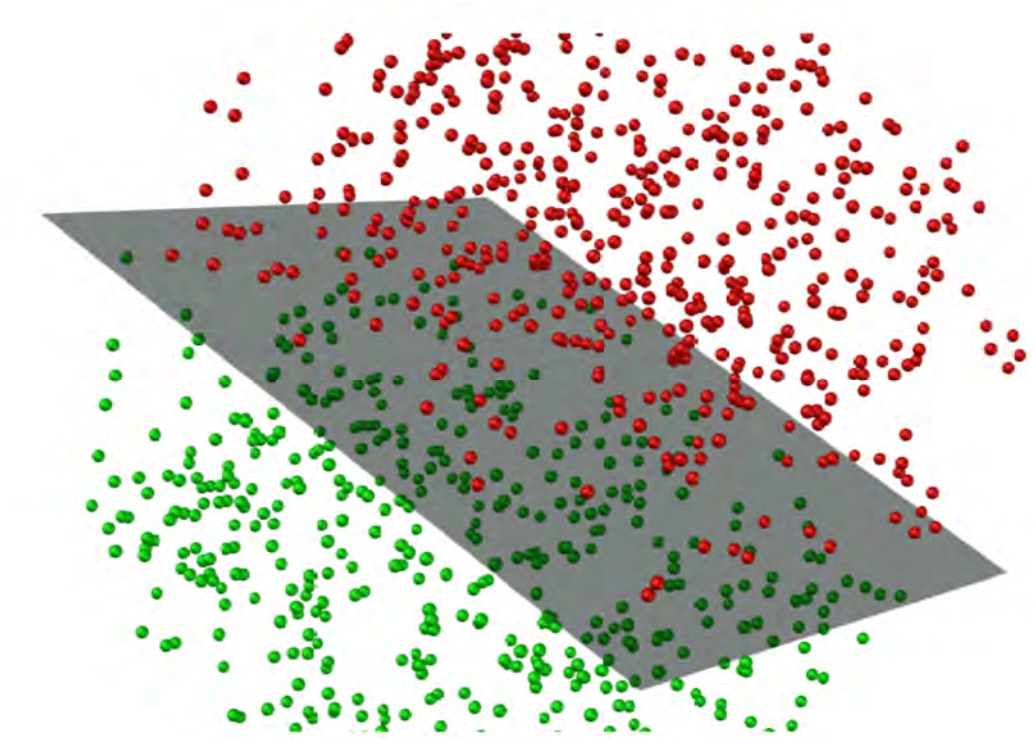
INDEX

FIGURE 29. SUPPORT VECTOR MACHINES. NON-LINEAR SEPARATORS AND CHARACTERISATION SPACING

a. Non-Linear Separators



b. Characteristics space



INDEX

The underlying idea behind these methods is mapping the input space into a high dimension vectorial one ($N >> n$) (called **characteristics space**), where linear separation is possible. This is known as the **kernel trick**[30], which replaces the inner products of the formulation with a non-linear function in the original input space:

$$Maximise\ F(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

The most widely used **kernel function** (functions that make the transformation, illustrated in Figure 29 *b.*) are:

- **Radial base function** (Gaussian): $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}$
- **Polynomial:** $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \boldsymbol{x}_j + 1)^p$
- **Hyperbolic tangent:** $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(k\boldsymbol{x}_i \boldsymbol{x}_j + c)$

Formulation of the SVM for multi class problems

As was mentioned at the beginning of this section, the support vector machines were originally created as binary classifiers. In practice, the number of categories in which an observation may be classified is often more than two. To deal with a classification of more than two classes, a method has to be found that turns a multi class classification problem into multiple binary classification problems.

1. **One-versus-all** method. It builds k binary classification models, where k is the number of classes. The m^{th} model is trained with all of the elements of the class m assigned to the positive class and the rest to the negative. Once this problem is solved, there are as many decision functions as classes.

When classifying, the k models are used and the class whose model has obtained a greater score is assigned.

2. **One-versus-one** method. It builds $k(k-1)/2$ classifiers and each one is trained with the data belonging to two different classes, in other words, given the training data of the classes i and j, the following problem is resolved:

[30] Boser et al., 1992] Boser, B. E., Guyon, I. M., And Vapnik, V. N. (1992).*A Training Algorithm for Optimal Margin Classifiers*. In Proceedings of The Fifth Annual Workshop on Computational Learning Theory, 1992, Pages 144–152. Acm

INDEX

$$Minimise: \frac{1}{2}\|w^{ij}\|^2 + C\sum_{t=1}^{n} \xi_t^{ij}$$
$$\left(w^{ij}\right)^T \phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \qquad si\ y_t = i$$
$$\left(w^{ij}\right)^T \phi(x_t) + b^{ij} \leq -1 + \xi_t^{ij}, \qquad si\ y_t = j$$
$$\xi_t^{ij} \geq 0$$

Once the classifiers have been built to choose the class of a new data, voting strategies are used: if the element $\boldsymbol{x}$ belongs to the class $\boldsymbol{i}$ one is added to the score of a class $\boldsymbol{i}$. If, on the other hand, it belongs to $\boldsymbol{j}$,it is the score of this class that increases by one point. In the end, the class with the highest value is chosen and if there is a tie between several of them, the one with the lower index is chosen. This is known as the Max Wins strategy.

The strategy to be used in this project is **one-versus-one**, as the training time is shorter than **one-versus-all**. The Support Vector Machine training time depends on the number of elements in the training set. Using **one-versus-one**, more models are trained but with fewer data than using **one-versus-all**[31].

SVM resolution: SMO algorithm

To solve the formulation of the SVM seen in the previous section, an**optimization problem** has to be solved (finding the maximum of a function) with restrictions (conditions the data must meet). There are numerous alternatives for this, including:

- ***Sequential Minimal Optimization***, *SMO*[32],
- ***Interior Point Method***, *IPM*[33] y
- ***Iterative Re-Weighted Least Squares***, *IRWLS*[34] y [35].

SMO was chosen for this module as, on the one hand, it does not require a large amount of RAM memory and may be run on a mid-range PC, and on the other, it is the most widespread and widely used method[36].

[31] Hsu, C. W., & Lin, C. J. *A comparison of methods for multiclass support vector machines* IEEE transactions on Neural Networks, 2002, 13(2), 415-425.
[32] Platt, J. Et al. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization* Advances in Kernel Methods-Support Vector Learning, 1999, 3.
[33] Karmarkar, N. A *New Polynomial-Time Algorithm for Linear Programming* In Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing, 1984, Pages 302–311. ACM
[34] Pérez-Cruz, F., Bousoño-Calzón, C., and Artés-Rodriguez, A. Convergence of the IRWLS procedure to the support vector machine solution. Neural Computation, 2005, 17(1):7–18.
[35] Díaz-Morales, R. and Navia-Vázquez, A. *Improving the efficiency of IRWLS SVMS using parallel Cholesky factorization* Pattern Recognition Letters.
[36] Chang, C.-C. and Lin, C.-J. (2011). *Libsvm: A Library for Support Vector Machines*. ACM Transactions on intelligent systems and technology (TIST), 2011, 2(3):27.

INDEX

SMO is an algorithm based on *Microsoft Research*, and breaks this problem into a series of smallest possible sub-problems. To do this, in each iteration it chooses two of the coefficients α_i that appeared in the dual formulation explained the previous section "b) Soft margin linear SVM" and updates them to optimise the cost function. By working only with two variables each time, it can solve the problem analytically in a simple way.

Validation and test criteria

In order to correctly validate the accuracy of the SVM in these databases, a validation and testing process was used to evaluate the performance when processing data never used to train the SVM.

a) Test

The data set of data used to create the models was split into two: training set (composed by 70% of the observations) and test set (formed by the remaining 30%). A test set is needed to measure the accuracy of the method in events that have not previously been used for training.

The final result of the performance of the algorithm is calculated by evaluation on this test set.

b) Training

Due to the computational cost of the SVM, with a complexity of $O(n^3)$ (every time the number of training data is doubled, the run time is multiplied by 8) and, also due to the large number of meters to be processed, it is desirable to limit the maximum number of training data. This value has been set by default at 10,000 events.

The aim of the first phase is to obtain the value of the SVM parameters:

- C: parameter of the cost function
- γ: parameter of the core function of the radial base function type

To obtain the best value of these parameters, 75% of the data were used to train the model and the remaining 25% to validate the result. The parameters achieving the best validation accuracy were selected.

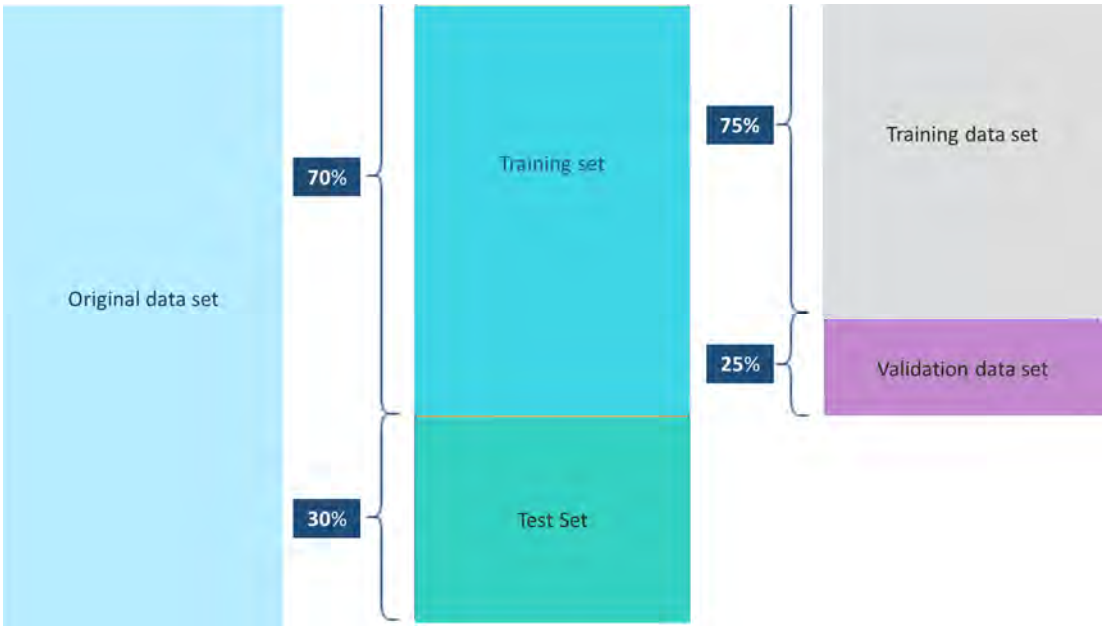
An initial test in which parameter values of 10^n were tested, **n** took whole values of between -4 and 4, it was seen that for the different meters the values were always C equal to 100 or 1,000 and of γ equal to 10, 100 or 1,000. These are the values validated individually for each meter.

After obtaining the values of these parameters, training is performed with all the dataset (limited to 10,000 data, as we have discussed before).

Figure 30 shows the division of the original dataset into data for training, validation and testing.

INDEX

FIGURE 30 DIVISION OF THE SET OF DATA IN TRAINING, VALIDATION AND TEST



c) Development and implementation

As for the development of the neural networks, *Python 2.7* programming language with the *Anaconda* was used for the SVM.

The packages used were *Scikit-learn*, *NumPy* and *Pyodbc*, described in the previous section.

5. Results

5.1. METER MODELS

This section summarises and compares the results of the classification made with different models: individual models (1 litre and 0.1 litre meters) and general models with the two methodologies used.

The general models were trained using the information from all the meters, whereas the individual models only bear in mind the information from the meter for which they are going to predict. At the end of this section, a brief comparison is made of the methods developed in this project and the statistical methods used up to now by Canal de Isabel II.

General models are developed because in the future there may not be training data individualised by dwelling, and classifiers may be trained with data from other dwellings.

As stated above, the number of data of the SVM was limited to 10,000, and of the neural networks to 100,000, and 30% of said data were used in the evaluation tests.

5.1.1. 1-Litre meter models

To compare the results in this section, a total of 239 1-litre meters were considered. The classification algorithms were trained on 70% of the log of each meter and the test was made on the remaining 30%.

The results obtained with the Neural Network are significantly better than those obtained with the SVM, both in overall terms and on meter level.

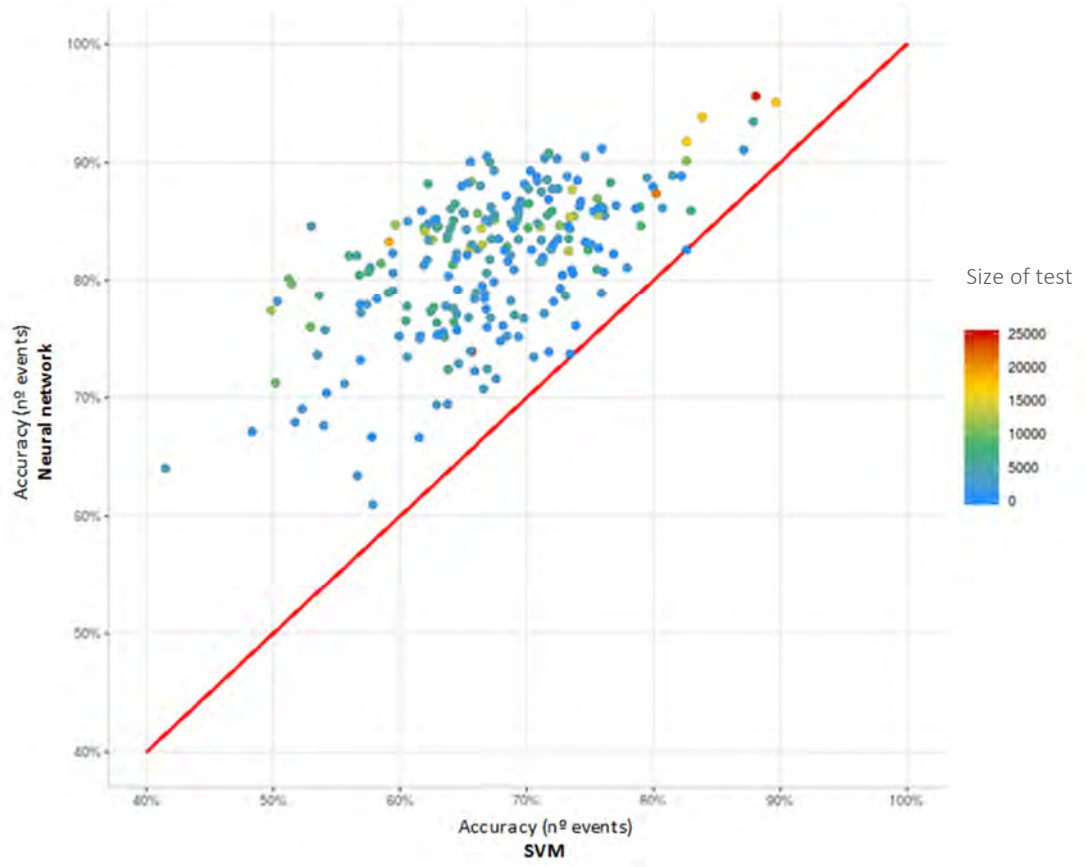
Considering the total results obtained with both classification methods, in other words, if all the cases classified in all the test sets are added up (i.e., all the meters of this kind) the accuracy of the SVM is 67.41%, whereas in the case of the neural network is 81.78%. This precision has been calculated bearing in mind the number of correctly classified events. The average accuracy in terms of correctly classified litres is very similar, and is 63.41% for the SVM and 85.76% for the neural network.

If the results are broken down on the level of meters, it cannot be found any meter where the results of the SVM are better than those of the Neural Network. In Figure 31, each point represents one of the analysed meters and, its position on the plane is the result of confronting the precision obtained with the two methods. Therefore, if the point is above the red straight, it means that the precision obtained for this meter with the Neural Network is larger than that obtained with the SVM. If it is on the straight, both precisions are equal. If it were below the straight, the SVM would have produced better results than the Neural Network.

The colour of each point indicates the number of test events. Most of the test sets exceed 500 events, although 11 meters were found (approximately 5% of all 1-litre meters) whose test sets are of a smaller size. At this point, the hypothesis could be considered that the larger the test set, the greater the precision, but it would be necessary to check it with a larger sample and with another size distribution.

INDEX

FIGURE 31. COMPARISON OF THE PRECISION OF THE CLASSIFIERS. 1-LITRE PRECISION METERS



The number of events correctly classified with the Neural Network increases by an average 22.35%, exceeding 40% on 19 occasions.

5.1.2. 0.1-Litre meter models

19 meters of 0.1-litre type were used in this comparison, from the total number of meters. The classification algorithms were trained using 70% of the historical records and the test was made on the remaining 30%.

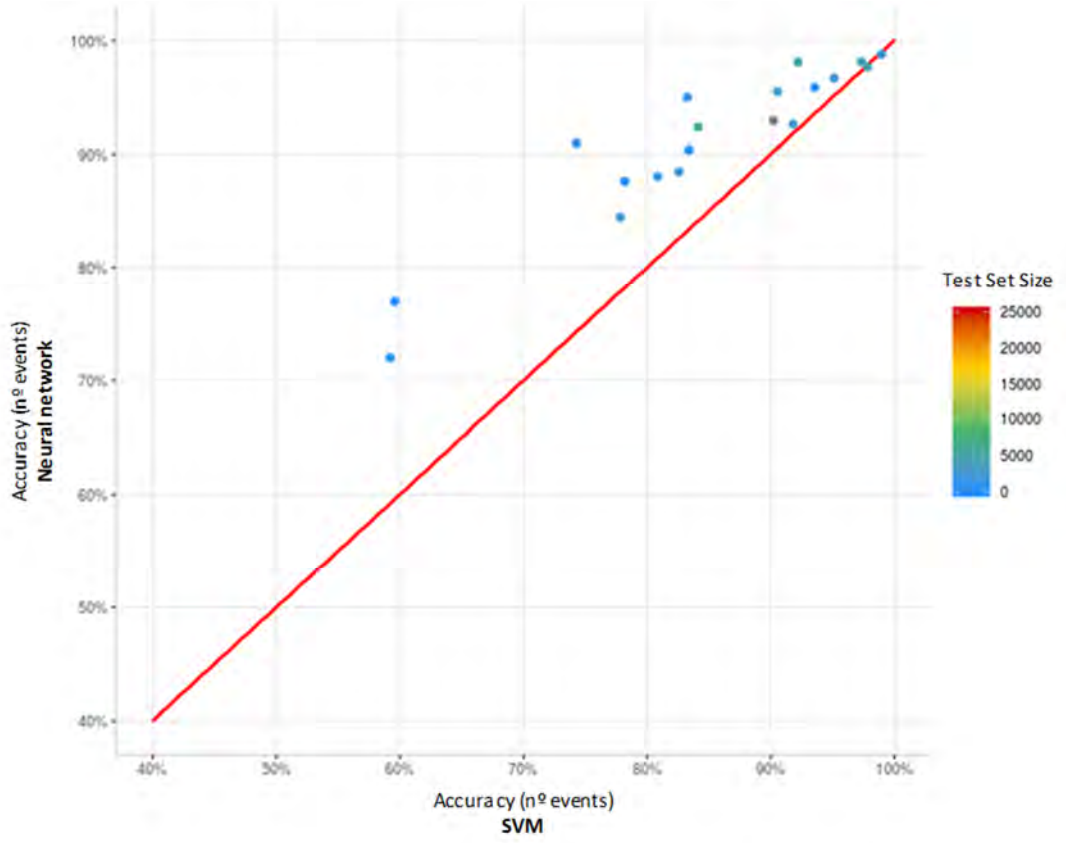
As it happened with the 1-litre meters, the results obtained with the Neural Network are better than those obtained with the SVM, both in overall terms and on meter level.

The accuracy of the SVM and of the Neural Network is 84.78% and 91.19%, respectively. If the percentage is quantified in correctly classified litres, these fall to 73.5 % and 85.9%.

Breaking down the results on the level of meters, it is not possible to find any for which the results of the SVM are significantly better than those of the Neural Network.

In Figure 32, each point represents one of the analysed meters and, its position on the plane is the result of confronting the precision obtained with the two methods. Therefore, if the point is above the red straight, it means that the precision obtained for this meter with the Neural Network is larger than that obtained with the SVM. If it is on the straight, both precisions are equal. If it were below the straight, the SVM would have produced better results than the Neural Network.

FIGURE 32. COMPARISON OF THE PRECISION OF THE CLASSIFIERS. 0.1-LITRE PRECISION METERS



The number of events correctly classified with the Neural Network increases by an average 8.5%, exceeding 20% on 3 occasions.

5.1.3. General models

General models are those that have been trained with data from every meter, distinguishing whether they are of a precision of 1 litre or 0.1 litres. These models have been tested to see whether it was possible to train with data from other different meters without losing precision in the results.

The limit on the inputs for training these models is 100,000 events in the case of the Neural Network, and 10,000 in the case of the SVM.

Only meters with, at least, 4 uses of a different nature in their historical records (e.g., showers, faucets, leaks and clothes washer) have been considered.

The overall accuracy is 75.18% using the SVM and 82.17% with the Neural Network.

INDEX

Comparing the individual models with the general, the results are better when applying the models in individual meters.

5.1.4. Comparison with statistical models

This section contains a comparative analysis of the classification methods implemented in this project and the previous automatic statistical classification method.

It was not possible to make an analysis of the precision of the methods in the line followed in previous sections (giving a correct classification percentage in the number of events and litres), as the classification was not available that was made by an operator in the periods in which it was classified using the Bayesian statistical method.

Given the different nature of the uses available for all users, a selection of results was chosen from the three compared methods (Bayesian, SVM and ANN) for the overall evaluation of the operation from the volume distribution assigned to each kind of use. In other words, a sample was selected and the total litres were quantified that were assigned to each kind of use to compare it with the distribution of the volume in the data labelled by operator.

NB:

*The sample for the overall comparison was taken on the **1-litre** type meters to avoid the bias of **faucet** events detected in 0.1-litre type meters. The events that are not present in the majority of the users (swimming pools and irrigation) were also eliminated from the sample.*

Table 4 and Figure 33 show the volume distribution according to the classification method. The SVM have problems in detecting *Leaks* and *Dishwasher*.

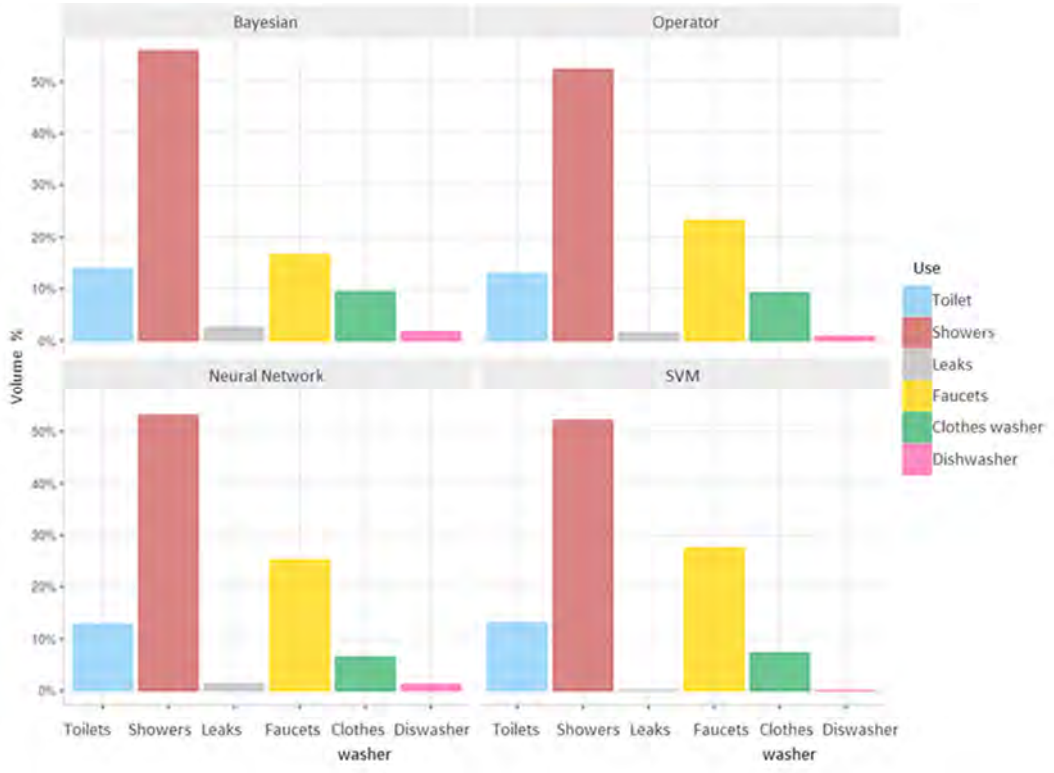
TABLE 4. DISTRIBUTION OF THE CONSUMED VOLUME ACCORDING TO THE CLASSIFICATION METHOD USED

Use	*(%) Operator volume*	*(%) Bayesian volume*	*(%) SVM volume*	*(%) Volume ANN*	*Average Vol. (L)*	*Volume typical deviation (L)*
Toilets	12.97%	13.82%	13.07%	12.79%	6.15	1.84
Showers	52.37%	55.96%	52.14%	53.14%	21.33	32.87
Leaks	1.47%	2.51%	0.08%	1.34%	1	0.02
Faucets	23.21%	16.61%	27.43%	25.12%	3.06	3.83
Clothes washer	9.21%	9.42%	7.28%	6.47%	3.75	3.40
Dishwasher	0.77%	1.69%	0.01%	1.14%	1	0

The differences seen when comparing the methods with the operator must be interpreted with care, as they were not performed in the same periods, so part of the variation could be explained by changes in daily habits among users (increase or decrease of the number of inhabitants in the dwelling, changes in working hours, etc.).

In general terms, the Bayesian method tends to mistake*shower* events with *faucet* events, whereas the Neural Network classifies these events precisely.

FIGURE 33. DISTRIBUTION OF VOLUME, BY TYPE OF USE, ACCORDING TO THE DIFFERENT CLASSIFICATION METHODS



5.2. COMPUTER APPLICATION

The implemented application integrates the different computer packages developed in a single computer application developed with VBA for *Access*, and includes the whole necessary procedure for identifying the end use of water in residential consumption. The appearance of the main menu of this application is shown in Figure 34.

For a certain meter, the procedure is as follows:

1) **Transformation of pulses into flows and event identification** From the accumulated pulse readings, the time series of flows are calculated with intervals of one second. It generates an *Access* file with these flow time series (flow episodes). Starting with these flows, the different events that form the flow episodes are identified and characterised, generating a new*Access* database that includes the feature parameters of each of the identified events.

2) **Initial training labelling**. Each event is assigned the (end use) label which was assigned to each event identified by operator. The methodology used was developed in Modules 3 and 4. The result of this labelling is specified in a new field that is added to the event database.

INDEX

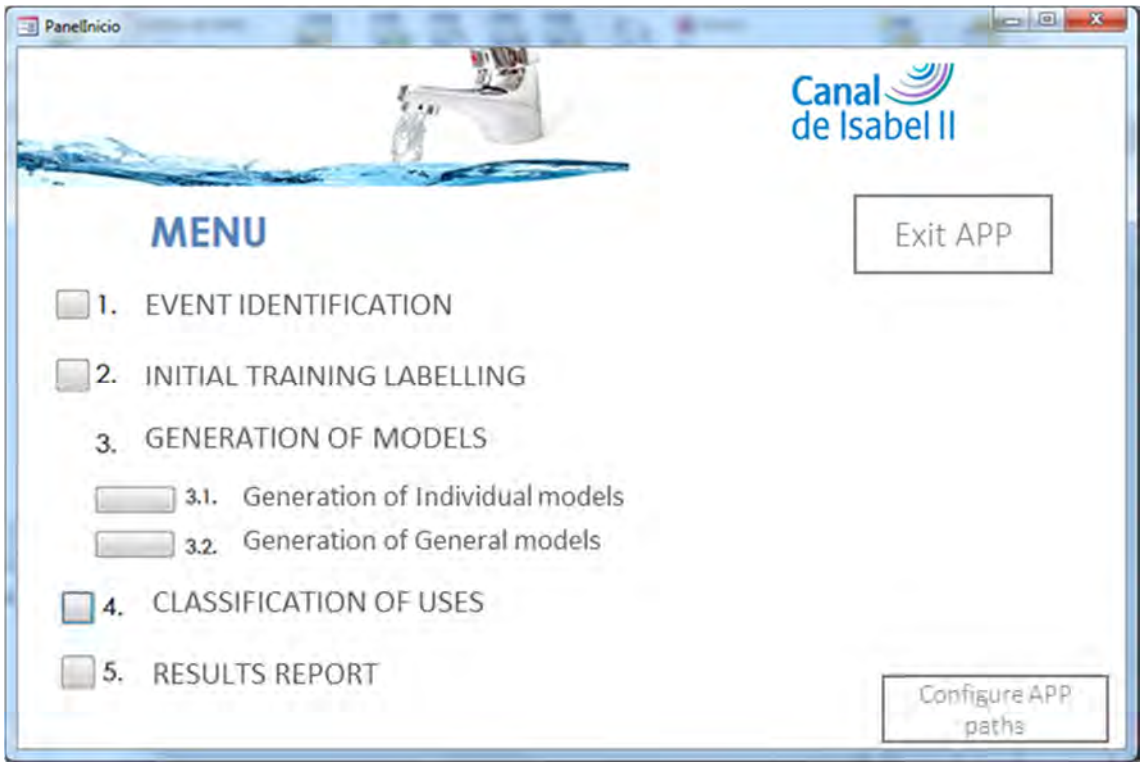
3) **Generation of models**. The initial training labelling allows the generation of a model for a specific meter. Given that two methodologies have been developed, one based on Artificial Neural Networks (ANN) and the other on support vector machines (SVM), the application allows the generation of both types of models.

   For new installations where there is not initial operator labelling, a model variant has been developed that uses the initial labels of the meters that did have this prior labelling to generate the so-called General Models, one for each methodology (ANN or SVM). In addition to the methodology, these general models depend on the precision of the meters (1 or 0.1 litres), giving four types of models, one for each methodology and meter precision.

4) **Classification of uses**. By using one of the generated models, the events identified in the first step are classified and assigned to a new label.

5) **Results report**. Finally, a report is presented on an *Excel* spreadsheet with tables and charts presenting the most relevant data of the classification.

FIGURE 34. COMPUTER APPLICATION. GENERAL MENU



INDEX

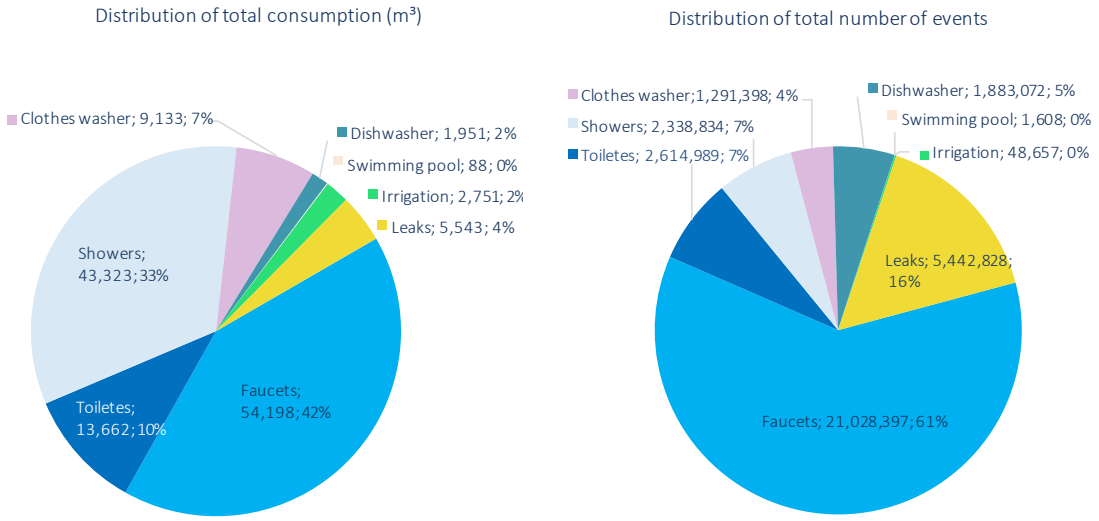
5.3. RESULTS OF THE CLASSIFICATION

The tables and charts included below refer to overall results drawn up using all the data recorded from January 2008 to July 2015.

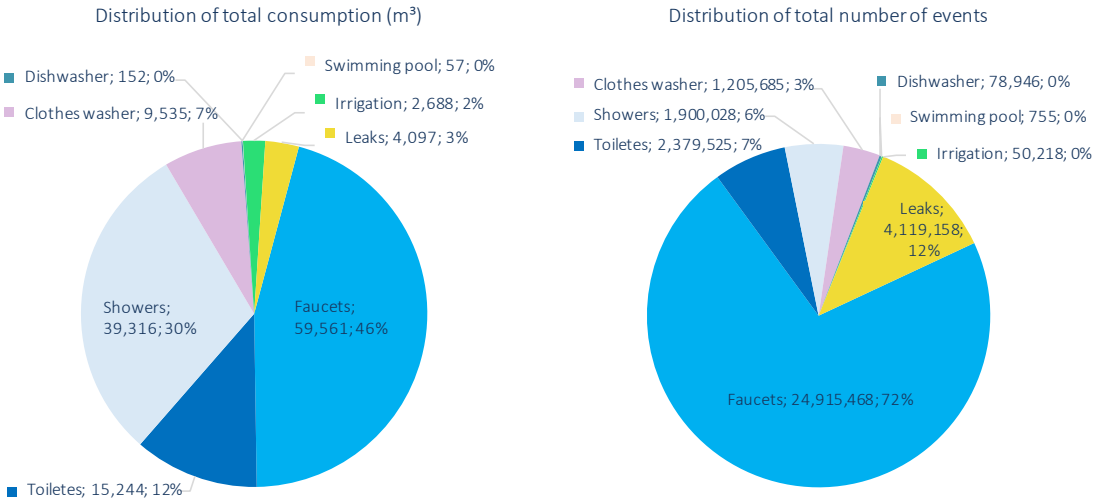
First of all, there is a comparison of the results obtained in the classifications achieved with both methodologies, ANN and SVM, and certain differences are seen between both methods. Therefore, the most common uses, *toilets tanks, showers* and *faucets* represents the 10%, 33% and 42% according to the ANN methodology, respectively; whereas according to SVM methodology the values for the same uses are 12%, 30% and 46%, as it can be seen in Figure 35.

FIGURE 35 COMPARISON. CLASSIFICATION BASED ON ANN AND SVM OF DISTRIBUTION ACCORDING TO USES OF TOTAL CONSUMPTION, IN THE PERIOD 2008/01 TO 2015/07. OVERALL RESULTS OF ALL PROCESSED DATA

Classification based on **ANN** methodology



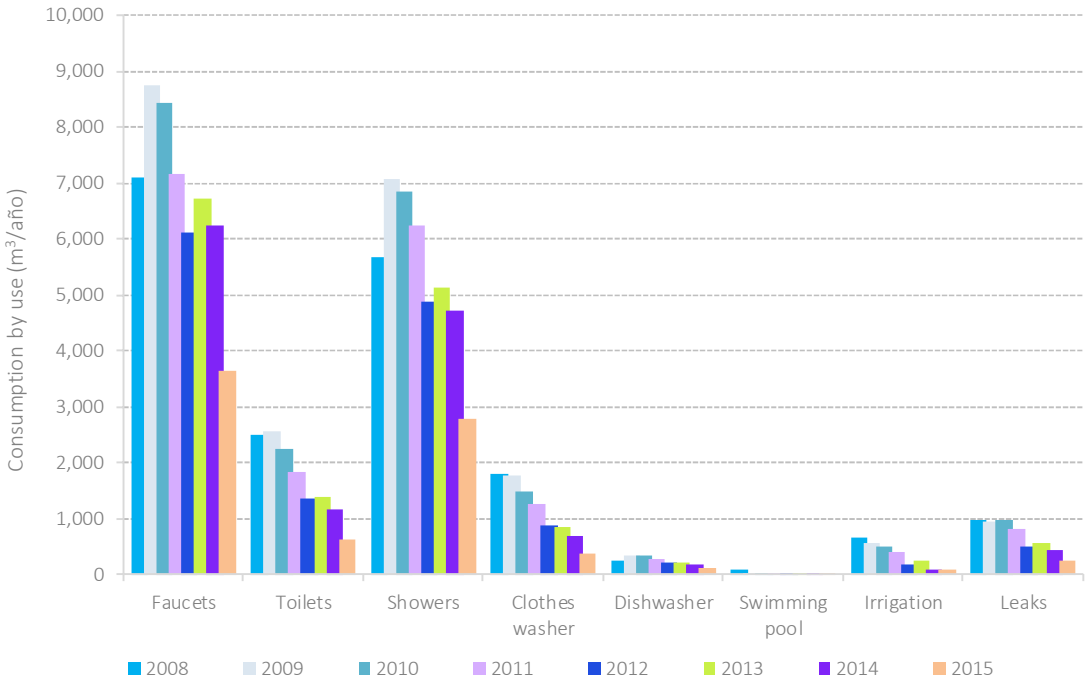
Classification based on **SVN** methodology



INDEX

Figure 36 displays the evolution of the consumptions over the years of study according to the uses, and shows that the consumptions have a tendency to drop.

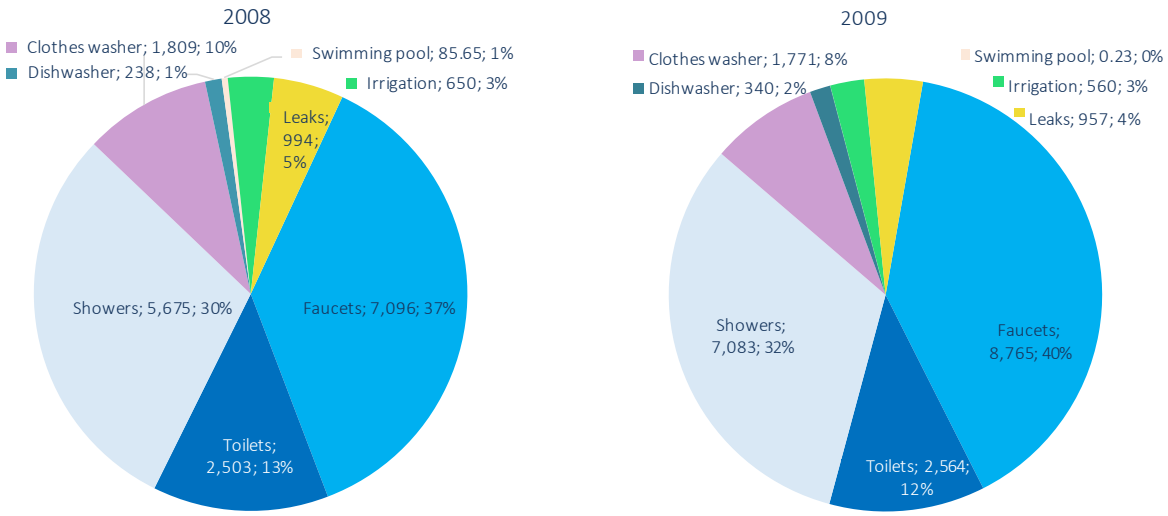
FIGURE 36. RESULTS OF THE CLASSIFICATION. EVOLUTION OF CONSUMPTION IN THE PERIOD 2008 TO 2015 (M³/YEAR)



NB: The year 2015 only includes consumption up to the month of July

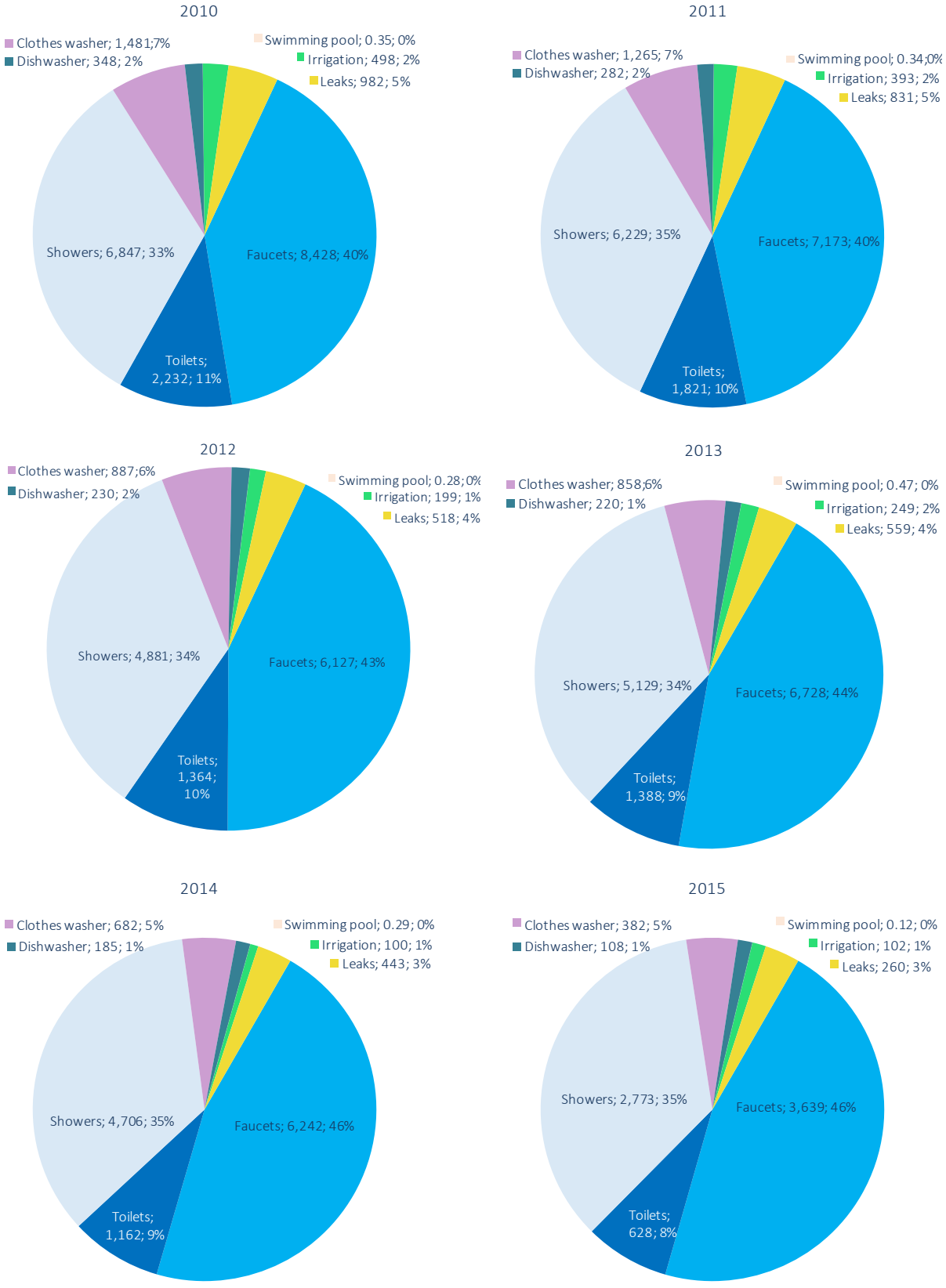
Figures 37 and 38 represent the same results year by year, indicating the percentage represented by each use in the total annual consumption.

FIGURE 37. RESULTS OF THE CLASSIFICATION. DISTRIBUTION BY USES OF THE TOTAL VOLUME CONSUMED IN 2008 AND 2009 (M³)



INDEX

FIGURE 38. RESULTS OF THE CLASSIFICATION. DISTRIBUTION BY USES OF THE TOTAL VOLUME CONSUMED IN 2010 TO 2015 (M^3)



INDEX

6. Summary and conclusions

This document has presented the works developed to automate, on the one hand, the process of identifying water use events in residential consumptions from the readings of meters with pulse emitter and a precision of 1 and 0.1 litres and, on the other hand, the classification of these events in accordance with the end use of water.

The event identification starts by developing a mathematical procedure that transforms the pulses registered by meters into consumption flows. This procedure is based on the calculation of moving averages with different orders by which the volume recorded at a certain instant (pulse) can be distributed over time. The flow episodes obtained in this way may respond to a single event or to a combination of two or more events. Using a second mathematical algorithm, a methodology has been developed that allows those episodes to be discretised in events, in order to characterise them according to different parameters (volume, maximum flow, total duration, duration of the ascending branch, of the descending branch, etc.)

Using previous events (labelled by an operator) recorded in a large number of installations (375 meters) and, by comparison with the events identified with the mathematical algorithm, these events are labelled so that they might serve as a pattern of learning for generating models based on two different techniques: **ANN** (Artificial Neural Networks) and **SVM** (Support Vector Machines). With the models generated for each measuring installation, it is possible to label new events identified and characterised according to the conceived mathematical algorithms from new readings of these meters.

The developed methodology allows the creation of neural networks or support vector machine models for different installations than those analysed in this work, as long as previously classified period of time (manually, by an operator) is provided.

For other installations, if there is no previous operator labelled set of events, general models have been created from the events that have been previously labelled, so that by applying these general models it is possible to classify the events registered by these new installations. The precision obtained with these general models is logically inferior to that of the models developed specifically for each meter, and presumably this precision may drop considerably if these general models are transferred to another context of residential water use very different from Comunidad de Madrid, the region where they have been trained.

For 1-litre precision meters, the results obtained with the Neural Network are significantly better than those obtained with the SVM, both in overall and meter terms levels. In fact, if all the results obtained with both classification methods are borne in mind, the accuracy with the SVM is 67.41%, whereas with the Neural Network it is 81.78%. This accuracy has been calculated bearing in mind the number of correctly classified volume of water. In terms of correctly classified volume the figures are very similar, being 63.41% for the SVM and 85.76% for the Neural Network.

If the results are broken down on the level of meters, it is not possible to find any meter whose results achieved with SVMs are better than obtained with Neural Network.

The number of events correctly classified with the Neural Network increases by an average 22.35%, exceeding 40% on 19 occasions.

Furthermore, for the 19 analysed meters with 0.1 litre precision, as with the 1-litre meters, the results obtained with the Neural Network are better than those obtained with the SVM both in overall terms and on meter level. The accuracy of SVM (Support Vector Machines) and Neural Network is 84.78% and 91.19% respectively. Ifthe percentage is quantified as a volume these values drop to 73.5% and 85.9%.

Likewise, on the level of meters, itis not possible to find any for which the results ofthe SVM are significantly better than those ofthe Neural Network.

The number of events correctly classified with the Neural Network increases by an average 8.5%, exceeding 20% on 3 occasions.

For the general models, the overall accuracy is 75.18% with SVMs and 82.17% with Neural Network.

When comparing the individual models with the general ones, the results are better when applying the models in individual meters, as was to be expected.

Every process has been programmed and compiled in a computer application developed to the effect that allows massive data treatment. The results ofthe classification are presented in the form oftables and charts that summarise the most significant values regarding event volumes and durations and their monthly and hourly distributions. All this discriminating according to the type of use.

7. Next steps

INDEX

The following research lines are suggested to continue with the works presented in this document:

- Optimisation of the automatic labelling process and application to massive data. The computer application has been developed in VBA on Access. To achieve greater agility and processing capacity, an adaptation to another system with better performance, like SQL Server or Oracle is suggested.

- An impact evaluation of water-saving campaigns using the automatic labelling tool, allowing an objective analysis of the campaign repercussion on every water use.

- Applicability of the automatic labelling to large patterns of consumption for the analysis of consumption patterns by sectors or districts, as a tool for updating hydraulic network models.

- Generation of integral control panels associated to automatic labelling, to obtain automated management reports by end uses of water.

- Development of an automatic labelling system as a tool for pre-locating leaks. Development of an early warning system for leaks inside dwellings as an element of water efficiency and provision of value for end users, given the economic savings that this system brings.

APPENDICES

APPENDIX 1. BIBLIOGRAPHIC REFERENCES

Almeida, G.A., Kiperstok, A., Dias, M., Ludwig, O.

Metodologia para caracterização de consumo de água doméstico por equipamento hidráulico. Anais do Silubesa/ Abes. Figueira da Fo. 2006

Almeida G., Vieira J., Marques J., Cardoso A.

Pattern recognition of the household water consumption through signal analysis. Camarinha-Matos L.M. (eds) Technological Innovation for Sustainability. DoCEIS 2011. IFIP Advances in Information and Communication Technology, vol. 349. Springer, Berlin, Heidelberg. 2011

Barreto, D.

Perfil do consumo residencial e usos finais da água. Ambiente Construído, Porto Alegre 8(2), 23–40 (2008) ISSN 1678-8621; © 2008, Associação Nacional de Tecnologia do Ambiente Construído, April/June 2008

Bishop, C. M.

Pattern recognition and machine learning. Information Science and Statistics, Springer-Verlag New York. Inc. (2006) Secaucus, NJ, USA

Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N.

A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, 1992, pages 144–152. ACM

Chang, C.-C. and Lin, C.-J.

(2011). Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3):27

Chen, T., & Guestrin, C.

(2016). Xgboost: A scalable tree boosting system. Proceedings of the KDD 2016 conference

Corona-Nakamura M. A., Ruelas R., Ojeda-Magaña B., Andina D.

Classification of domestic water consumption using an*Anfis* model. Conference Paper Automation Congress, 2008

Cubillo, F., Moreno, T., Ortega, S.

Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid. Cuaderno de I+D+i nº 4. Canal de Isabel II, 2008, Madrid

Díaz-Morales, R. and Navia-Vázquez, A

Improving the efficiency of IRWLS SVMs using parallel Cholesky factorization. Pattern Recognition Letters

Fernandes, B.C.

Construção de um Sistema Eletrônico de Monitoramento de Consumo de Água Residencial. Projeto de Graduação apresentado ao Departamento de Engenharia Elétrica. p. 65 Centro Tecnológico da Univ. Federal do Espírito Santo, 2007

Hsu, C. W., & Lin, C. J.

A comparison of methods for multiclass support vector machines. IEEE transactions on Neural Networks, 2002, 13(2), 415-425

Karmarkar, N.

A new polynomial-time algorithm for linear programming. Proceedings of the sixteenth annual ACM symposium on Theory of computing, 1984, pages 302–311. ACM

Mayer, P.

Water energy savings from high efficiency fixtures and appliances in single family homes. USEPA — Combined Retrofit Report 1, 2005

Nguyen, Zhang, Stewart

Analysis of simultaneous water end use events using a hybrid combination of filtering and pattern recognition techniques. International Congress on Environmental Modelling and Software (2012)

Platt, J. et al.

Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods support vector learning, 1999, 3

Pérez-Cruz, F., Bousoño-Calzón, C., and Artés-Rodriguez, A.

Convergence of the IRWLS procedure to the support vector machine solution. Neural Computation, 2005, 17(1):7–18

Roger O'Halloran, Michael Best y Nigel Goodman

Urban Water Security Research Alliance. Technical Report No. 91. 2012

Vapnik, V.

The nature of statistical learning theory. Springer Science & Business Media, 2013

Vapnik, V.

Pattern recognition using generalized portrait method. Automation & remote control, 1963, 24:774- 780

Vasak M., Banjac G., Novak H.

Water use disaggregation based on classification of feature vectors extracted from smart meter data. Procedia Engineering, 119, 1381 – 1390, 3th Computer Control for Water Industry Conference, CCWI 2015

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A

Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11 (Dec. 2010), 3371-3408

Zhang, C., & Ma, Y.

Ensemble machine learning: methods and applications. Springer Science & Business Media. (Eds. 2012)

APPENDIX 2. INDEX OF FIGURES

INDEX

APPENDIX 3. INDEX OF TABLES

Santa Engracia, 125. 28003 Madrid
www.canaldeisabelsegunda.es