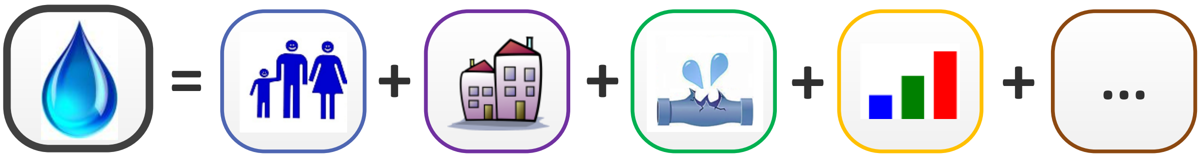


PROFILING CONSUMPTION PATTERNS USING EXTENSIVE MEASUREMENTS

A spatial and temporal forecasting approach for water distribution systems

Aisha Zulquifal Mamade

Thesis to obtain the Master of Science Degree in

CIVIL ENGINEERING

Examination Commitee

Chairperson: Prof. António Jorge Silva Guerreiro Monteiro

Supervisor: Prof. Dídia Isabel Cameira Covas

Supervisor: Doutora Dália Susana dos Santos da Cruz Loureiro

Member of the Committee: Prof. Maria da Conceição Esperança Amado

Member of the Committee: Prof. Helena Margarida Machado da Silva Ramos

March 2013

ABSTRACT

The goal of the current work is to develop a spatial and temporal forecasting for profiling consumption patterns in water distribution systems. This profiling is based on the socio-demographic, billing and infrastructure analysis of each water sector.

Water distribution systems are typically designed based on a set of considerations (e.g. average per capita water consumption, percentage of water losses, peak factors), instead of using information related to the actual characteristics of each water sector. There are few systematic studies that allow the profiling of domestic water consumption in terms of consumption variables (*e.g.* minimum consumption, instantaneous peaking factor) or in terms of the daily average behaviour for a certain weekday or season. This research aims at contributing to filling this gap.

A modified methodology based on Loureiro (2010) is proposed using extensive flow measurements provided by water utilities in Portugal. New methods for calculating socio-demographic indexes and for detecting and classifying outlying events are developed. Socio-demographic indexes are calculated using a geoprocessing tool initially developed by Loureiro (2010) in a Geographic Information System (GIS) that has been significantly improved in the present work. Indexes are calculated using recently published census data.

Concerning the main outputs of this work, a socio-demographic, billing and infrastructure framework is presented and the profiling of network flows is carried out at the District Metered Area (DMA) level.

Consumption profiling provides average demand patterns for different consumer profiles (*e.g.* young families, elder families), empiric relations for estimating design parameters (*e.g.* peaking factors, average consumption) and operational parameters (*e.g.* minimum night flow) for different seasonal scenarios (*e.g.* Summer, Winter) and daily scenarios (*e.g.* weekdays, Saturdays, Sundays and holidays). Results are coherent with previous studies. This contribution considerably reduces the uncertainty in planning and operation of water distribution systems, thereby improving the efficiency and sustainability of these systems.

Keywords: Consumption profiling, District Metering Area (DMA), multivariate exploratory techniques, outlier analysis, spatial analysis, water distribution systems.

RESUMO

O objetivo do presente trabalho consiste em tipificar padrões de consumo, tendo em conta uma previsão temporal e espacial de consumos. Esta tipificação assenta numa análise sociodemográfica, de infraestrutura e de faturação de cada setor de rede.

Os sistemas de distribuição de água têm sido dimensionados com base em critérios gerais, sem atender às características locais dos sectores de água. Existem ainda poucos estudos sistemáticos que permitam tipificar o consumo doméstico de água, quer em termos de um conjunto de variáveis (e.g., consumo mínimo, fator de ponta instantâneo), quer em termos do comportamento médio diário para um dado dia da semana ou época do ano. Este estudo pretende contribuir para a melhoria do conhecimento nas áreas referidas.

Propõe-se uma metodologia adaptada de Loureiro (2010) e aplica-se a um conjunto alargado de Zonas de Medição de Controlo (ZMC) fornecidas por entidades gestoras em Portugal. Em termos da metodologia, este estudo contribui com novas abordagens para o cálculo de índices sociodemográficos e para a detecção e classificação de eventos anómalos nas séries de caudal. As variáveis sociodemográficas são calculadas usando uma ferramenta de geoprocessamento inicialmente desenvolvida por Loureiro (2010) num Sistema de Informação Geográfica. (SIG). Esta ferramenta foi consideravelmente melhorada no presente estudo e usa os dados dos Censos de 2011, recentemente publicados.

Relativamente aos principais resultados, destaca-se o perfil sociodemográfico, de facturação e infraestrutura das ZMCs estudadas, bem como a tipificação de caudais de rede ao nível da ZMC.

Da tipificação resultam padrões de consumo para diferentes perfis de cliente (*e.g.* famílias jovens, famílias idosas), relações empíricas que permitem estimar parâmetros que apoiam diretamente o dimensionamento (*e.g.* factores de ponta, caudal médio) e a operação de sistemas de distribuição de água (*e.g.* consumo mínimo nocturno). Consideraram-se diferentes cenários sazonais (*e.g.* Inverno, Verão) e diários (e.g. dias da semana, sábados, domingos e feriados). Os resultados são coerentes com estudos anteriores. Esta contribuição pretende reduzir consideravelmente a incerteza no planeamento e operação, aumentando assim a eficiência e sustentabilidade dos sistemas.

Keywords: Análise de *outliers*, análise espacial, técnicas exploratórias de análise multivariada, tipificação de consumis, sistemas de distribuição de água, Zona de Medição e Controlo (ZMC).

AKNOWLEDGEMENTS

First of all, I would like to thank Professor Dídia Covas and Doctor Dália Loureiro for the opportunity of developing and learning with this thesis.

Thank you Professor Dídia for your availability, your focus and the creative ideas transmitted. Thank you Dália for your constant concern regarding the investigation purposes of this thesis and your attention to detail.

I likewise want to thank all the members of the Urban Water department in LNEC for their warm welcome and all the opinions shared over this work during the past months. Thank you Sérgio T. Coelho for the inside view into all the other department´s activities.

Also, this work would not have been possible without the water utilities, who actively contributed with their data. So thank you AGS, AdS and SMAS-OA. A particular, thank you to Rita Almeida (AGS) and Catarina Sousa (AdS) not only for the data, but also for your time in sorting out all my doubts.

I would also like to thank Alexandre Santos for his crucial help when exploring the GIS world. Margarida Rebelo, thank you for your help in analysing the socio-demographic side of water.

For all the laughs and relaxing moments I would like to thank all my friends.

Finally but more importantly, I would like to thank my family. I could not have made it without your constant support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF NOTATION

Abbreviation	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AWE	Alliance for Water Efficiency
AWWA	American Water Works Association
cl	Client
DL	Decree-Law 23/95 Article nº19
DMA	District Metered Area
EPA	US Environmental Protection Agency
EU	European Union
FL	Fuzzy Logic
GHG	Greenhouse Gases
GIS	Geographic Information System
HDPE	High-density polyethylene
ID	Identification code
INE	Instituto Nacional de Estatística (Portuguese National Statistics Institute)
IRAR	Instituto Regulador de Águas e Resíduos (Portuguese Water and Waste Regulator Institute)
KMO	Kaiser-Meyer-Olkin
LNEC	Laboratório Nacional de Engenharia Civil (Portuguese National Laboratory of Civil Engineering)
MLR	Multiple Linear Regression
MRSP	Metropolitan Region of São Paulo
OECD	Organisation for Economic Co-operation and Development
OWC	Otawwa West Centre
PCA	Principal Components Analysis
PM	Pressure Management
PMA	Pressure Management Area

PVC	Polyvinyl Chloride
S	Summer season scenario
S&H	Sunday and holidays
sc	Service connection
SCADA	Supervisory Control And Data Acquisition
SD	Standard deviation
SH	Summer, Sunday and holidays scenario
SS	Summer, Saturday scenario
SSE	Statistic Sub-Section
SW	Summer, working day scenario
TS	Time series
UK	United Kingdom
USA	United States of America
W	Winter season scenario
WH	Winter, Sunday and holidays scenario
WMS	Work Management System
WS	Winter, Saturday scenario
WW	Winter, working day scenario

LIST OF SYMBOLOGY

Symbol	Meaning	Units
$\boldsymbol{W_{ij}}$	Weighting ratio for DMA *i* and SSE *j*	[-]
$\boldsymbol{X_{ij}}$	Number of domestic clients (or area, or billed consumption) for DMA *i* and SSE *j*	[-]
$\sum \boldsymbol{X_j}$	Number of domestic clients (or area, or billed consumption) for SSE *j*	[-]
$\boldsymbol{OTL}$	Outlier value in the data series	[m^3/h]
$\boldsymbol{MED}$	Median of a set of previous observations defined by the user	[m^3/h]
$\boldsymbol{c}$	Threshold value to be defined by the user	[-]
$\boldsymbol{Qn}$	Robust standard deviation of the observations based on the Qn scale	[m^3/h]
$\boldsymbol{y_i}$	Dependent consumption variable *i*	[*e.g.* l/cl·day]
$\boldsymbol{\beta_p}$	Regression coefficient related with independent variable *p*	
$\boldsymbol{x_{ip}}$	Independent variable *p*, related with dependent variable *i*	

1. INTRODUCTION

1.1 Context and motivation

"Since the 1960s, water resource management has shifted towards water conservation and demand-side management, and collaborations between engineers, economists, sociologists and urban planners are now common. Advances in technology now permit finer grained measurement of water consumption and new household appliances have far greater water efficiency. Residential water demand models have increased in complexity, with faster, more powerful computational programs, and demand side management tools have been introduced that decrease water demand (Tanverakul and Lee, 2012). "

The development of approaches to profile water consumption in terms of minimum consumption, instantaneous peaking factors and daily peaking factors or in terms of daily average behaviour for a certain weekday or season are essential for a better management and operation of water distribution systems.

For instance, the forecasting of the minimum consumption is important to control water losses; the forecasting of peaking factors is important when designing water infrastructures and measurement equipment such as water meters; an improved knowledge on the daily average consumption, intrinsically associated with social behaviours, is relevant for planning sustainable water distribution systems, since it allows a better management of stored water volumes and of pump operation, reducing energy costs and taking advantage of lower tariffs.

Currently, there are many water utilities with their networks divided in District Metered Areas (DMAs). Building a database with real (and validated) flow data and infrastructure information for these DMAs enables the improvement of forecasting and classification models. This database is of upmost importance for planning and designing new water sectors as well as for its daily operation and management.

The lack of systematic studies combining the various driving factors that lay behind water consumption is the motivation of the development of the current research work.

1.2 Objectives and methodology

The goal of the present work is to improve the methodology proposed by Loureiro (2010) for profiling consumption patterns in water distribution systems, using extensive measurements. This profiling includes empiric relations for estimating design parameters (*e.g.* peaking factor, average consumption) and operational parameters (*e.g.* minimum night flow). These parameters support the planning, designing, operation and maintenance of water distribution systems.

To achieve the proposed goal, the followed approach for the development of the current work was:

(i) Review on the state of the art for consumption analysis and profiling.

(ii) Consumption data collection provided by from several water utilities in Portugal and other types of data such as infrastructure, billing and census data.

(iii) Processing and analysing the collected data using tools and applications developed by LNEC or under its supervision.

(iv) Profiling water consumption variables and patterns and comparison with Loureiro (2010).

Results should provide accurate estimates of parameters for the designing and operation of new water sectors (*e.g.* peaking factors, water losses) and for estimating consumptions in non-metered areas.

1.3 Contents

In order to achieve the specified objectives, this work is organized in eight chapters.

- Chapter 1 introduces the scope of the analysis, the objectives and methodology of the present work and its structure.
- Chapter 2 provides an overview on the existing methods for characterizing and profiling consumption.
- Chapter 3 summarizes the methodology adopted for water consumption analysis and presents a tool and two applications mainly used to enhance data processing.
- Chapter 4 consists of building a socio-demographic, billing and infrastructure framework according to the values from legislation or previous research studies.
- Chapter 5 detects outliers among the flow data series, aims at classifying events resulting from the outlier detection and removes them from the series. Additionally, it characterizes the clean data series.
- Chapter 6 uses the clean series to develop a consumption analysis that includes identifying monthly and weekly scenarios in the flow data series, after which consumption variables and average daily consumption pattern are determined.
- Chapter 7 uses multivariate analysis techniques to identify the main factors behind domestic water consumption in the studied DMA's. Following that, consumption variables and patterns are analysed against the referred factors using correlation and regression analysis. These analyses enable profiling consumption patterns in water distribution systems using extensive data.
- Chapter 8 provides a summary with the final considerations of the developed research work.

2. STATE-OF-THE-ART

2.1 Introduction

"Decades of research have revealed that residential water consumption is a function of various factors including population demographics, water pricing, house characteristics, weather, and emotional and psychological drivers in consumer habits and attitudes (Tanverakul and Lee, 2012). Future water resource management must meet water consumption flexibly by implementing both conservation and sustainable practices. Accurate forecasts are vital for designing water distribution and treatment systems, identifying appropriate water sources, and operating water distribution networks (Tanverakul and Lee, 2012). Historically, conservation efforts have largely been implemented separately across the water and energy domains even though there are numerous benefits to coordinate conservation efforts and exploit water-energy linkages (Abdallah and Rosenberg, 2012)."

The current chapter aims at presenting a state-of-the-art review on water and energy consumption. It includes a set of guidelines for efficient management in water distribution systems, a review on the main driving factors of water and energy consumption, as well as consumption forecasting techniques used in previous research works. Additionally, the potential of the water and energy nexus is explored. A summary of the review is presented at the end of the chapter.

2.2 Best practices for efficient management of water distribution systems

2.2.1 Zoning and DMA / PMA implementation

Water leakage accounts for a significant amount of non-revenue water in many cities of the world. It varies from 3% of the input water into the distribution systems in well managed systems, or recently constructed systems, to over 50% in poorly managed systems (Mutikanga *et al.*, 2011).

Water utilities, particularly in the developing countries, are still grappling with challenges of high water losses due to leakage. A recent World Bank study estimated that over 32 billion cubic meters of treated water are annually lost in leakage from urban water supply systems around the world and half of these losses occur in developing countries (Mutikanga *et al.*, 2011). The same report estimates the full cost of water losses from urban water utilities in developing countries to be as much as US$5 billion per year. In light of global pressures of growing demand and increasing water scarcity, water utilities need to operate more efficiently for sustainable service delivery (Mutikanga *et al.*, 2011).

Water utilities generally first obtain awareness of bursts through customer reports, via call centres where the details of various service problems are collected such as complaints of low pressure, discolouration and possibly signs of visible surface water. Although large bursts in a water supply system are usually found and repaired quite rapidly, not all leaks and bursts result in visible surface water and, therefore, remain undetected for many months, if not longer.(Mounce *et al.*, 2010)

A widely adopted procedure to facilitate leakage management is the sub-division of water distribution systems into District Meter Areas (DMAs). A DMA typically has flow measured at the inlets and outlets of the system. In some cases, pressure measurements are carried out at some locations; tese provide additional information for a more proactive leakage management tool of the network (Mounce *et al.*, 2010; Mutikanga *et al.*, 2011).

District Management Areas (DMAs)

A DMA is an area of a distribution system which is specifically defined (*e.g.* by the closure of valves) and in which the quantities of water entering and leaving the area are continuously metered (Morrison, 2004).

The concept of District Metered Areas (DMAs) was first introduced to the UK at the start of the 1980s by the UK Water Authorities Association. Until that time, most water distribution systems were not metered or measurements were carried out in groups for all commercial, industrial, public (including fire hydrants) or residential use (Tanverakul and Lee, 2012). Figure 1 exemplifies the division of a distribution network into DMAs.

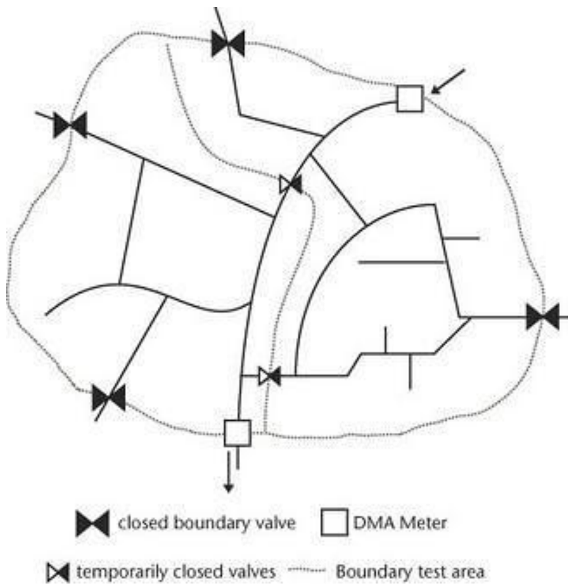


Figure 1 – Division of distribution networks into DMAs (Morrison, 2004)

DMA size

When contemplating the DMA plan, the question often arises: 'What is the optimum size of a DMA for the effective water losses assessment? (Brothers, 2011)

DMA implementation and the corresponding metering, installation, maintenance, calibration, and volume of information requiring analysis, can be too complex and costly to sustain which is why a search for the optimum size is needed (Brothers, 2011).

Farley and Trow (2003) recommend DMAs with a number of service connections between 500 and 3000 in order to guarantee efficiency in pinpointing critical areas with high water losses. Other authors recommend smaller DMAs (with 250 to 600 connections) for a better control of bursts and leaks (Jankovic´-Nisic´ et al., 2004).

Nevertheless, over recent years, there have been significant improvements in the accuracy and robust construction of electromagnetic flow meters for nominal pipe sizes of 100 up to 250 mm (and greater). Supervisory Control and Data Acquisition (SCADA) systems also have significantly improved over the years (Brothers, 2011).

The technology available today with high resolution SCADA systems and corresponding high quality electromagnetic flow meters provides an opportunity to significantly increase the size of DMAs from 30-50km with 3000 service connections, to much larger areas with 80-100km of pipes and up to 10,000 service connections. This outcome enables the implementation of fewer and larger DMAs, that are permanently monitored, with substantial lower initial capital cost , as well as long-term operating cost savings (Brothers, 2011). Table 1 summarizes DMA recommended sizes in terms of service connection range for the studied authors.

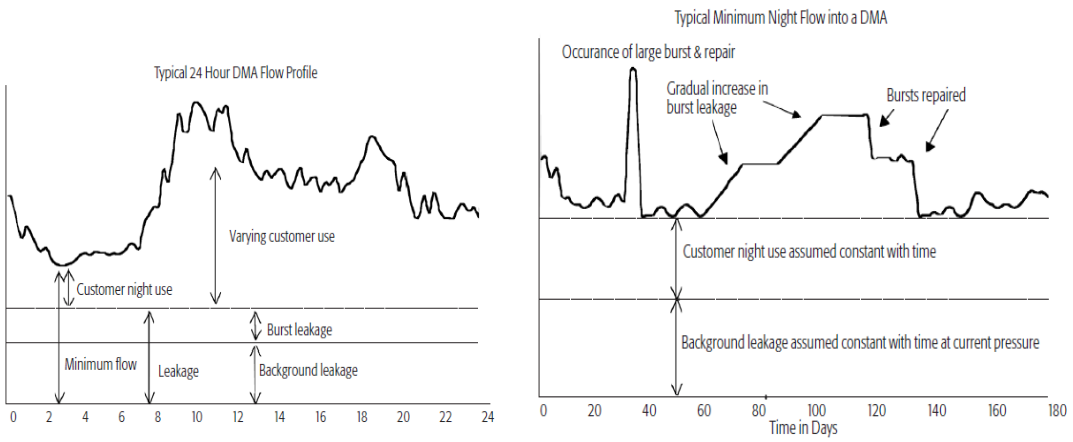
Table 1 – DMA sizes in terms of service connection range

Author	Service connections
Farley and Trow (2003)	500-3000
Jankovic´-Nisic´ et al. (2004)	250-500
Brothers (2011)	3000-1000

Estimating leakage

Best practice analysis of DMA flows requires the estimation of leakage when the flow into the DMA is minimum. This typically occurs at night when customer demand is minimum and, therefore, the leakage component represents the largest percentage of the demand. Techniques are now available to analyse the minimum night flow to estimate the level of leakage and additionally to split this estimate into background leakage and burst leakage as shown in Figure 2 a. The analysis of leakage is based on the minimum night flow, which can be continuously recorded and analysed night after night with the use of data loggers and appropriate software. This analysis enables the monitoring of the DMA or groups of DMAs for the occurrence of new bursts and their subsequent repair, as shown in Figure 2 b.

Based on this analysis, summary reports of estimated leakage from bursts in individual DMAs can be developed to provide the leakage practitioner with a schedule of leakage that can be reduced. This reduction can be represented as a volume of water, a potential number of bursts that can be found, an estimate of the cost of leakage that is being lost, or a ranking system developed to suit local conditions. When fully developed, this analysis will enable the monitoring of a large number of DMAs more effectively and focusing the analysis in key DMAs, which will result on a higher benefit from leak reduction. The above analysis of minimum night flow represents the best practice for leak assessment (Morrison, 2004).



(a) (b)

Figure 2 – Typical flow profile: a) 24 hour profile b) minimum night flow (Morrison, 2004)

Pressure Management and Pressure Management areas (PM and PMAs)

Historically, most networks have had no active pressure control and there is scope to reduce leakage through the reduction of excess pressure (Mutikanga *et al.*, 2011).

The first step in pressure management and DMA management is the definition of the area of the network, the closure of boundaries and the measurement of the inflows and outflows - whether for DMA analysis or to control inlet pressures. Where the topography is favourable, the planning of Pressure Management Areas (PMAs) and DMAs should be undertaken as one overall concept, although implementation of one step may come before the other (Morrison, 2004).

As a final remark, PM not only reduces leakage but also extends useful life of the overall infrastructure, reduces operation and maintenance costs through reduced frequency of main breaks and energy consumption, improves customer service as a result of reduced water supply interruptions and is a demand management tool (Lambert and Fantozzi, 2010).

2.2.2 Outlier detection

Until recently, the state of the art for the management and derivation of meaningful information from water distribution system data has been limited, time consuming and of inadequate accuracy. This is primarily due to a reliance on human data analysis and interpretation, which is unfeasible and inefficient for the growing volume and complexity of data involved (Mounce *et al.*, 2010).

Real time leakage estimation is now increasingly receiving more attention as water utilities, especially in the UK, are set ever increasing targets for regulatory compliance and standards of service delivery. Whenever one analyses a large data set, such as flow data series originated by the SCADA systems, outliers are an inevitable concern that needs to be identified and dealt with (Mounce *et al.*, 2010).

An *outlier* or an *outlying observation* is an observation, or a sequence of consecutive observations, in the time series that significantly deviate from other members of the sample in which it occurs (Grubbs, 1969). In the particular case of flow metered data, outliers can be errors resulting from data acquisition and processing as well as flow peaks associated with pipe bursts, or abnormal consumptions (e.g. fire fighting).

Multiple outliers can contaminate models designed for burst and leak location. Faster detection of bursts saves water, minimises the inconvenience of interruption to customers and decreases the damaging consequences to infrastructure. Flow monitoring techniques are used by water service providers to monitor leakage, generally through offline application of mass balance type calculations, manual observations of change in night line values or filters with empiric relations based on data comparison (Kim *et al.*, 2006; Loureiro, 2010).

There has been recent interest in on-line monitoring of water distribution sensors and corresponding event detection. Event detection algorithms should work in near real-time by obtaining SCADA data, performing some analysis and then returning a binary classification (i.e. sounding the alarm or not). The existing state of the art operational systems in the UK are implemented using flat line alarm levels on key monitoring sites in the control room and allow near real-time identification of large bursts.

To implement on-line process monitoring techniques, it is necessary to extract data associated with the normal operating conditions. Historical databases contain data from normal operating conditions, faulty conditions, various operating modes, start-up periods, and shutdown periods. The presence of outliers further complicates the task of locating data associated with normal operating conditions.

In the UK,Mounce *et al.* (2010) developed an Artificial Intelligence (AI) model that applies Artificial Neural Network (ANN) and Fuzzy Logic (FL) technology for automated online analysis of DMA flow data:

- The ANN model, a mixture density network, was trained using a continually updated historic database that constructed a probability density model of the future flow profile.

- A Fuzzy Inference System was used for classification, comparing the latest observed flow values with predicted flows over time windows such that abnormal flow conditions generate alerts (Mounce et al. 2006).

In order to evaluate the alerts from the AI system, sources of information for correlating alerts were integrated, such as:

- Daily emailed briefings on major events
- Customer contact database (customer reports of visible leakage)
- Work Management System (WMS) record of repairs database

From the probability density functions of predicted flows the fuzzy inference system provides confidence intervals for each detection. Figure 3 illustrates a situation in which the AI system correctly produced an alert for a burst which subsequently resulted in customer reports of leakage, and then was located and repaired. The alert was received at 15:15 on the 13th of December of , with a confidence of 95% and a size estimate of 0.4 l/s. As can be seen from Figure 3, flow was abnormal in the 12 hour window up to this time.

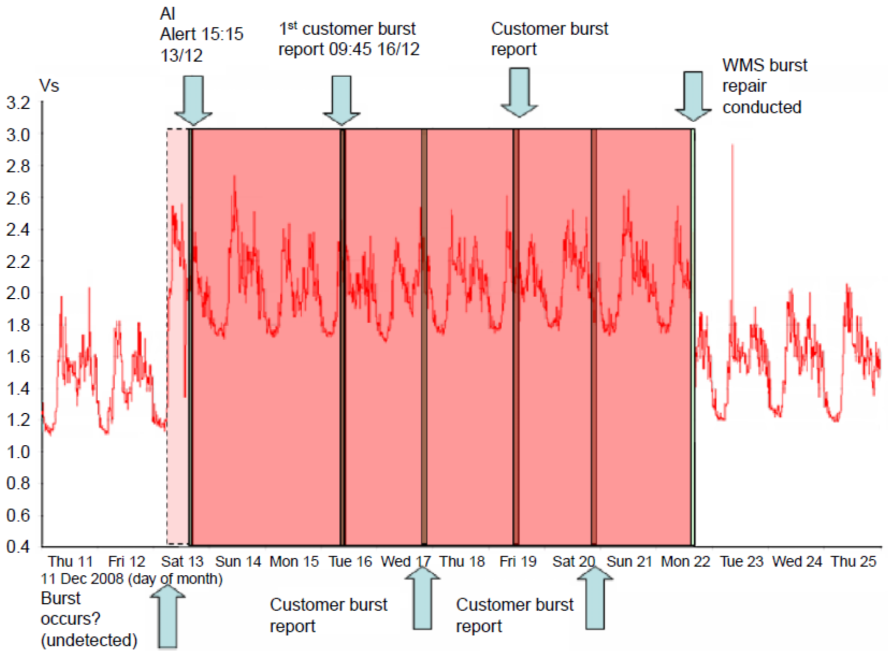


Figure 3 – Example of alert provided by the AI system with additional information (Mounce *et al.*, 2010)

When summarizing the total of alerts produced in one year, seven groups were distinguished: alerts correlated with burst repair, alerts correlated with burst reports, on-going leakage or rezoning, known industrial events, abnormal, data issues, unknown/ghost.

In Portugal, Loureiro *et al.* (2013) have been trying to locate normal operating conditions using two robust estimators: the median and the Qn.

The median of a sample can be described as the numerical value separating the higher half of a sample from the lower half. Qn is the robust standard deviation of the observations based on the Qn scale (Rousseeuw and Croux, 1993).

- Median values are calculated to estimate water consumption for all instants of the flow data series. The median of a certain instant is calculated based on a window of 20 previous days with observations in the same instant.
- Median and Qn are combined in a mathematical expression to define the region of normal operation conditions. Observations outside this region are considered outliers.

This method only needs a database with 20 days of observations and the range of normal operation conditions is calculated without the need of flow data from other DMAs. This method is still under development. One the algorithms will be presented and tested in Chapter 2 of the current research work.

2.3 Understanding driving factors of water and energy consumption

The efficient use of water and energy is becoming a great issue of concern. Many regions of the world are facing a severe drought after years of continued lower than average rainfall. For this reason, as well as the addition of high population growth and strong economic development, water and its use must be managed very carefully. In an attempt to improve water use, many government authorities have imposed a number of water restrictions and water saving measures to ensure the conscious use of water across the residential, commercial and industrial sectors.

The residential (or domestic) water consumption is usually the most important use in an urban areas (OECD, 1999). Due to greater social awareness, people are beginning to value water as a precious resource. Behaviour and attitudes toward both potable and recycled water have forever changed, thus requiring renewed understanding on the link between these factors and water end use (Beal *et al.*, 2010).

Along with water concerns, the demand for energy has been exponentially increasing as a result of the industrialization and globalization. It is predicted that if the current global energy consumption pattern continues, the world energy consumption will increase by over 50% before 2030 (Suganthi and Samuel, 2012). Residential energy use represents about 35% of global energy use and it, therefore, plays a key role in global energy-related environmental problems, such as climate change and resource scarcity (Agency, 2004)

Throughout the EU, there has been a move towards smarter electricity networks, where increased control over electricity generation and consumption has been achieved with improvements in new technologies, such as Advanced Metering Infrastructure. In particular, residential smart meters have been installed in a number of countries around the world such as: Australia, Canada, Italy, Netherlands, Northern Ireland, Portugal and Sweden.

For the above mentioned reasons, managing water and energy resources in an optimal manner has become imperative among water and energy utilities and policy makers.

A number of key water and energy drivers play a role in their residential use. Table 2 presents some of the key drivers recently used for water and energy consumption forecasting.

Table 2 – Water and energy drivers used in recent studies

Category	Examples of drivers	Research works on water forecasting	Research works on energy forecasting
Socio-demographic and economic	Education, income, dwelling type, household makeup, household size, number of rooms, outdoor uses	Beal and Stewart (2011), Polebitski and Palmer (2010),	McLoughlin et al. (2012), Daioglou et al. (2012)
Billing	Volumetric consumption, consumption type, water tariff	Loureiro (2010), Grafton et al. (2011), Tanverakul and Lee (2012)	Shi et al. (2012), Bernstein and Madlener (2011)
Climate and environmental concerns	Temperature, precipitation, efficient devices, conservation attitudes	Beal and Stewart (2011), Adamowski (2008), Grafton et al. (2011)	Daioglou et al. (2012), Suganthi and Samuel (2012)

2.3.1 Socio-demographic, economic and land use drivers

Water

Regarding demographic variables, the most important driver of consumption appears to be the number of persons living in the household (household size) (March *et al.*, 2010). Because of economies of scale, larger households tend to spend less water per capita than smaller households (March Corbella and Saurí i Pujol, 2009).

Additionally, a review of Hunter Water's customer database that younger properties (<15 years old) had a higher consumption than older properties (40–50%), which could indicate a culture change in consumption (White *et al.*, 2007). The gender of the residents is known to influence residential consumption. Makki *et al.* (2011) verified in their study the intuitive notion that women are more likely to consume higher water volumes than men. The same study indicates that the number of teenagers, males and children aged 3 years or less would explain shower end consumption that represents the majority of residential water consumption.

In an attempt to categorize water end uses in Australia, a Southeast Queensland Residential End Use Study was commissioned in 2010. The analysis of 252 homes in four distinct regions allowed the better understanding of the main water end uses. As shown in Figure 4, bathroom (toilet and shower) and laundry activities represent the majority of the residential indoor demand of potable water (Beal and Stewart, 2011).

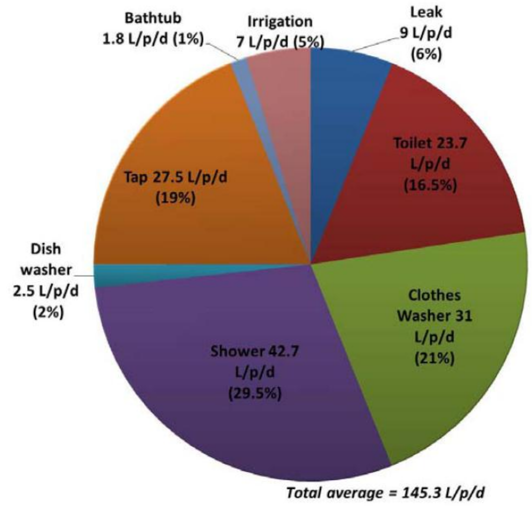


Figure 4 – Average daily per capita water end use breakdown (Beal and Stewart, 2011)

Water consumption analyses generally fail to take into account that the type of residential building affects water consumption. In the US, there are approximately 30 million multifamily buildings that usually have a master water meter and the costs are divided between residents. Consequently, individual bills are only indirectly related to the amount of volume used, which do not encourage individual users to save water. It was observed a 15% decrease in water consumption in multifamily apartment units after the installation of individual water meters (Tanverakul and Lee, 2012).

Furthermore, March *et al.* (2010) have detected a certain neglect of variables such as ageing and immigration. Nonetheless, there are some pioneer studies arguing that immigrants from developing countries and the elderly may observe patterns of more frugal water consumption than the population in general. This may be partially related to the fact that elderly people tend to live in small apartments/flats and have few water-using appliances. Olmstead *et al.* (2003) found that the elderly tended to live in older buildings with few household appliances, and were connected to the urban water supply by means of smaller pipes. The same may be true for immigrants arriving from the developing countries who, at least in Europe, tend to live in older, crowded, and more degraded buildings than average local citizens (March *et al.*, 2010) .

Generally, most studies focus on comparing the total per capita consumption values, showing that individuals that are wealthier, and that live in new and larger homes consume more (Beal and Stewart, 2011). The reason for wealthier families consuming more water is, to some extent, associated with outdoor purposes such as irrigation and due to having swimming pools. This fact is also clear in Figure 5, where irrigation values are clearly higher for the middle to higher socio economic region. Another aspect affecting residential water consumption is occupation and education level, though it is considered as a secondary factor (Schleich and Hillenbrand, 2009).

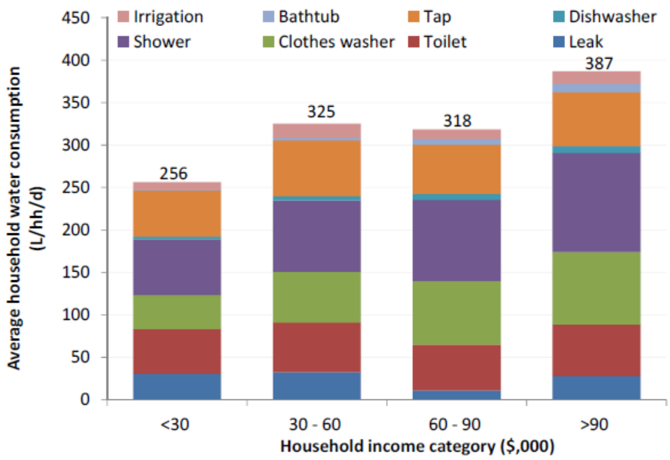


Figure 5 – Impact of socio-economic region on end use water consumption (Beal and Stewart, 2011)

Urban land uses and concretely the density of houses (compact or dispersed) also have an important effect on domestic water consumption (March *et al.*, 2010). There are still few studies linking water consumption and urban planning (Shandas and Hossein Parandvash, 2010). Concerning residential form, for instance, there is still a strong disparity of domestic water consumption between Australia and the United States and Australia (400 and up l/day, on the average), and, Europe (around 170 l/day). This difference is strongly attributed to the wider presence of outdoor uses (gardens and swimming pools) in Anglo-American contexts. In turn, these uses are made possible by the low density, disperse urban form characteristic of Australia and North America as opposed to the higher densities found in Europe. To the extent that low density housing is gaining terrain in countries such as France, Portugal or Spain, one may expect an increase in outdoor uses in these countries and, therefore, of water consumption as well (March Corbella and Saurí i Pujol, 2009).

Water consumption research typically focuses on a micro level (specific use within a household) or at an aggregate level (water use within a city or region). Little research has focused on water use at the multihouse, census tract or sub-section level, the appropriate level for many water planning decisions. The spatial scale of a census tract has many desirable properties, including the availability of data, appropriate frequency of observation, and a spatial resolution similar to many types of demographic data. Using census tracts as a spatial unit provides detailed information concerning water consumption patterns useful to planners and engineers for evaluating conservation programs and land use policy. Utilities interested in spatial pricing schemes may also find this approach useful (Polebitski and Palmer, 2010).

Energy

Concerning socio-demography, past literature has identified dwelling type, appliance holdings and number of occupants to be key factors influencing electricity consumption. However, it is important to note that the frequent occurrence of certain variables may also be a consequence of the ease with which data were collected. For instance, data relating to the top four variables referred can be obtained from national census and household budget surveys with relative ease. Other variables such as floor area may be overlooked due to the difficulty with which this information is gathered, particularly for large sample sizes. Electricity consumption patterns for domestic dwellings are highly stochastic, often changing considerably between customers (McLoughlin *et al.*, 2012).

Yohanis *et al.* (2008) analysed patterns of electricity consumption in 27 representative dwellings in Northern Ireland. Electricity load profiles were characterised based on dwelling type, floor area, number of occupants, number of bedrooms, tenure, occupant age and household income. In particular, the authors found a significant relationship between domestic electricity consumption and floor area.

Daioglou *et al.* (2012) developed a model to analyse possible future developments of residential energy use in five developing world regions: India, China, South East Asia, South Africa and Brazil. As presented in Figure 6, they found in each of these regions that cooking is currently the main end-use function, but other functions, such as space heating, cooling and appliances are becoming important. Daioglou et al. (2012) also examined the influence of dwelling and occupant characteristics on domestic electricity consumption patterns by analysing data obtained from a smart metering survey of a representative cross section of approximately 4200 domestic Irish dwellings. Dwelling type, number of bedrooms, head of household (HoH) age, household composition, social class, water heating and cooking type all had a significant influence over total domestic electricity consumption. Time of use for maximum electricity consumption was found to be strongly influenced by occupant characteristics, HoH age and household composition. Moreover, younger HoH were more inclined to use electricity later in the evening than older occupants.

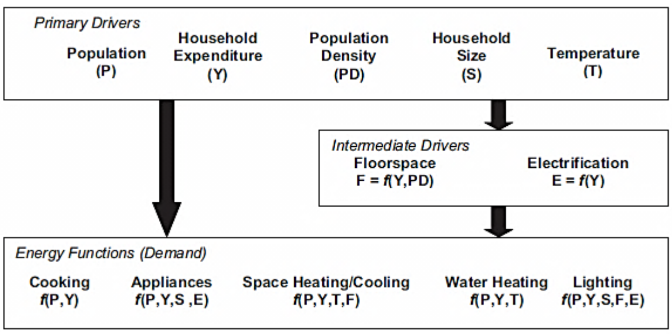


Figure 6 – Example of relationships between energy drivers and functions (Daioglou *et al.*, 2012)

In respect to economic drivers, Santamouris *et al.* (2007) found a significant relationship between income groups and domestic energy consumption. In their study, social, financial, energy and technical data from about 1110 households have been collected during 2004 in the greater Athens area. The sample has been divided into seven income groups and a detailed analysis has been carried out. One of the main conclusions is that there is an almost linear relation between the annual expenses for electricity and the family income. High income families pay almost 160% higher annual electricity costs than the low income ones. In what concerns the annual electricity cost per unit of area and per person it has been found that the lower the income the higher the cost of electricity per person and per unit of area. Low income people pay almost 67% higher electricity cost per person and square meter than high income people. Although middle and high income people use more air conditioning, the relative cost of comfort during the summer period is much higher for people on a lower income. This may be explained by the fact that low income people live in buildings with limited thermal protection and also because low income housing is located in areas of Athens where the heat island has maximum intensity.

2.3.2 Billing drivers

Water

In general, water utilities and water pricing regulatory authorities have eschewed the use of price as the primary method of controlling residential water demand and have, instead, opted for a variety of non-price approaches (Grafton *et al.*, 2011).

Price structures vary widely between regions, generally falling into one of three categories: flat rate, increasing block or decreasing block rates, each with an accompanying fixed water service fee. Flat rate pricing imposes a single volumetric price for all users, regardless of the use, while block rates charge different prices according to the amount used after a set base level. Increasing block rates increase with increasing volume use, whereas the unit cost for decreasing rates goes down as the volume used increases. Generally, a flat rate structure offers little incentive for water saving. Increasing block rates lower water demand by charging customers more based on higher water use.

A 1996 survey by the American Water Works Association (AWWA) found the residential rate structure distribution in the US to be: 39% uniform; 33% declining block: 22% increasing block; and 4% flat, even though the increasing block rate is known to be most effective at targeting high volumes and discretionary use, such as gardening during peak demand period (Tanverakul and Lee, 2012).

In economics, the *demand* for a product is a decreasing function of price: as the price goes up, demand will go down (Beattie and Foster, 1980). The amount demand decreases with increasing price is termed *price elasticity*.

Concerning price elasticity, the following conclusions from past studies were taken by Grafton et al. (2011) and Tanverakul and Lee (2012):

- Although many studies confirm that demand reduction is associated to price rising, such reduction is only moderately significant. Two meta-analysis studies of water demand find that residential consumption does respond to price changes, but is price inelastic.
- When separating into water demand into indoor and outdoor components, outdoor use was found to be consistently more sensitive to price increases, with the use of water sprinkler being highly influenced by price.
- Households are more responsive to price changes the more time they have to adapt to price increases.
- Price elasticity of demand can be greater in the long run, which can be especially important for water authorities and utilities in evaluating the effectiveness of raising the volumetric price of water on water consumption.
- High-income households appear to be less price elastic in terms of their water consumption than low-income households.

The fact that residential consumption might be responsive to price changes opens the path for improving pricing structures. A conducted study in the Metropolitan Region of São Paulo (MRSP) using a combined regressive-progressive block price system indicates that poorer consumers spend nearly 4.5 % of their income on water whereas the richer pay nearly 0.5% of their income but consume more than twice as much (Ruijs et al., 2008).

This suggests that an income-based pricing would be useful to a more equitable system, especially in the fast growing mega-cities of the world that suffer from water shortage – MRSP's population represents 50% of São Paulo state and only occupies 2.7% of its territory.

Another similar study has been conducted in Beijing, showing the potential of adopting the block pricing system in achieving the water resource conservation. In these studies, the common denominator is that the distributional effects of pricing policies cannot be analysed without taking into account regional, economic and cultural backgrounds (Chen and Yang, 2009).

Grafton et al. (2011) surveyed 10 countries to test the importance of price and non-price factors on residential water consumption. As depicted in Figure 7, Mexico has the highest median level of annual water consumption (250 kL year^{-1}) and also has the lowest median of average water price (0.31 € kL^{-1}) where this price is calculated as the ratio of household water expenditures to household water consumption.

France has the lowest median level of water consumption (100 kL yr^{-1}) and the highest median of average water price (2.82 € kL^{-1}). Figure 7 also illustrates the striking and negative relationship between the mean of volumetric price of water (€ kL^{-1}) and the mean of per capita residential water consumption (kL yr^{-1}) among the 10 countries.

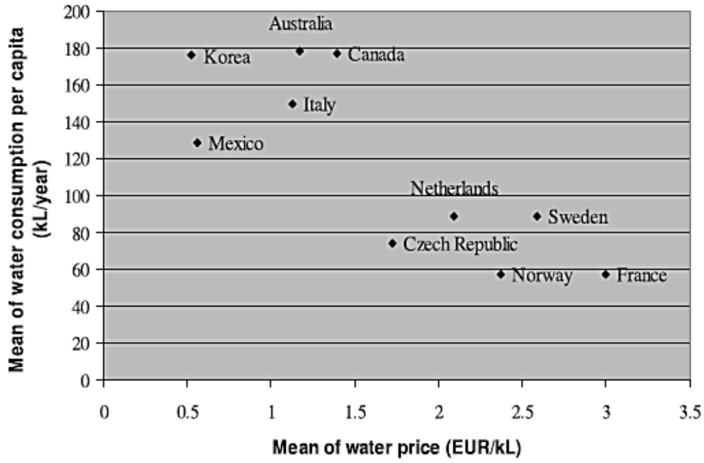


Figure 7 – Residential water consumption per capita plotted against the calculated mean water price in OECD (Grafton *et al.*, 2011)

Energy

Concerning energy,Mounce *et al.* (2010) conducted a study on the responsiveness of residential electricity demand for a set of OECD countries. Electricity demand was found to be inelastic with approximately –0.4 in the long run. In the short run results indicate a price elasticity of approximately –0.1. Results imply that the steering effect of tax-induced price increases on residential electricity demand has a very limited potential for energy conservation, and hence a reduction of Greenhouse Gas (GHG) emissions. In contrast, Shi *et al.* (2012) estimated residential demand for electricity in China as a function of local electricity price, household income, and a number of social-economic variables at household level. They found that the residential demand for electricity responds rather sensitively to its own price in China, which implies that there is significant potential to use the price instrument to conserve electricity consumption. Electricity elasticity across different heterogeneous household groups (e.g., rich versus poor and rural versus urban) were also estimated. The results show that the high income group is more price elastic than the low income group, while rural families are more price elastic than urban families. These results have important policy implications for designing an increasing block tariff.

2.3.3 Climate drivers

Water

Water consumption forecasts have typically not considered the potential impacts of climate change on water consumption. Climate change will impact water supply by decreasing reservoir inflows for many mountainous regions due to shifted spring stream flows, decreasing summer stream flows, and loss of snowpack. Changing temperature and precipitation patterns will also change water use patterns. Climate change may also affect the possibility of increased drought severity, drought duration and water consumption (Polebitski and Palmer, 2010).

Referring to climate variables, Grafton *et al.* (2011) show that two climate variables – precipitation and average summer temperature – help explaining the differences in household water consumption.

Adamowski (2008) studied the Ottawa West Centre (OWC) zone and shows that water consumption can be expected to be high on consecutive dry days with high temperatures and low on rainy days. Additionally, he found that the occurrence of rainfall was a more significant variable than the amount of rainfall itself in the modelling of short-term water consumption. This can be attributed to the fact that people may not want to water their lawns or gardens on a rainy day regardless of the amount of rainfall.

Figure 8 shows the diurnal water consumption pattern for the day of the highest peak consumption for the entire OWC zone record period for the summers between 1993 and 2002. It can be seen that the consumption rises significantly after 4 p.m., reaches a peak just before 9 p.m., and then gradually decreases. For comparative purposes, the diurnal pattern of a typical rainy summer day is also shown. It can be seen that on such a day, when there is likely little or no outdoor water use, there is no significant rise in consumption between 4 and 11 p.m. This illustrates the significance of outdoor water use on peak water consumption in the summer.

Examples of climatic variables that could be investigated in future studies are: evaporation; evapotranspiration; wind speed; relative humidity; cloud amount; and sunshine amount. Unfortunately, not all of the above data are readily available, and often do not exist at all (Adamowski, 2008).

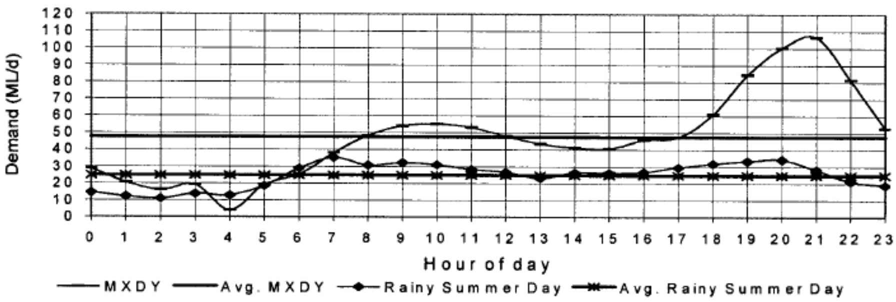


Figure 8 – Typical diurnal summer water usage in Ottawa West Centre zone (MXDY=day of highest peak consumption for entire 1992-2002 record; average MXDY=average peak consumption for entire record) (Adamowski, 2008)

Energy

Hart and de Dear (2004) used regression analysis to determine a relationship between external temperature and household electricity consumption in New South Wales, Australia. Figure 9 shows the diurnal distribution of air-conditioner energy consumption and corresponding outdoor air temperature for the heating and cooling seasons (a) and (b), respectively. The salient feature of the seasonal comparison in Figure 9 is the mean daily energy peak in winter is twice that for summer. During the cooling season (Figure 9 b), energy consumption begins to rise at 9 a.m., peaking at 4 p.m. and then rapidly decreasing to a minimum at 6 a.m., closely tracking the diurnal outdoor temperature cycle. Assuming the causal link between summer temperature and energy consumption extended to the heating season one might expect the winter diurnal load profile to be a mirror image of the summer's, but that appears not be the case (see Figure 9). Winter heating energy consumption has two peaks: one at 8 a.m. and the other at 9 p.m., the periods at which houses are most likely to be occupied.

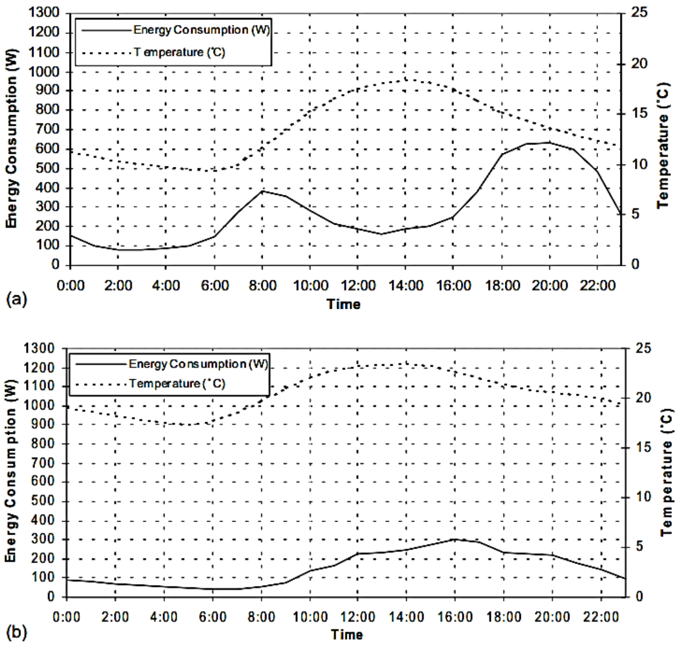


Figure 9 – The diurnal distribution of mean hourly air-conditioner energy consumption and mean hourly outdoor temperature for (a) the heating season (winter) and (b) the cooling season (summer). Energy consumption is averaged across 47 appliances for the cooling season and 41 appliances for the heating season (Hart and de Dear, 2004)

During the heating season the daily minimum temperature occurs between 6 a.m. and 7 a.m., but this coincides with the time of minimum, not maximum heating energy consumption. Heating season load profiles for room heaters produced a similar twin-peak pattern to that of air-conditioners during the heating season.

The previous analyses involved aggregation and averaging across a sample of 47 air-conditioners in the cooling season and 41 during the heating season, thereby masking the intricacies of individual householder behaviour.

Despite the referred fact, the research concluded that there was a significant relationship between external temperature and electricity consumption and that this tended to be stronger during periods of cooler weather. However, it is important to note that the preceding study was carried out in hot climates where electricity is commonly used to heat and cool homes, something which is not replicated in more temperate climates such as in the United Kingdom and in Ireland.

More recently, Águas (2013) suggested the introduction of indoor thermometers to examine the influence of indoor temperature on residential electricity consumption.

2.3.4 Water efficient devices and conservation attitudes

Many water authorities have been promoting the installation of water-saving devices, such as efficient toilets and showerheads. While it seems intuitive that water-saving devices should reduce household consumption, this may not necessarily be true in all cases.

The empirical evidence is actually mixed. For example, a study of low-flow showerhead retrofits in Colorado found no significant influence on consumption, while studies in California and Florida found modest savings (Grafton et al., 2011).

Beal and Stewart (2011) indicate that shower retrofitting has low payback periods (less than 1 year) when taking into account energy savings due to heating the water used in the shower. Concerning the investment in high efficient washing machines, the same study accounts for payback periods near to 6.5 years.

Grafton et al. (2011) indicates that the presence of a water efficient toilet reduces household water consumption by about 25%. By contrast, they found that neither efficient shower heads nor rainwater tanks have a statistically significant effect on household water consumption.

Grafton et al. (2011) also look at the effectiveness of five types of non-price conservation programs in reducing residential water consumption: public information programs, education (school) programs, retrofit programs, permanent ordinances and regulations, and temporary ordinances and regulations, reductions in residential consumption ranging from 1.1 to 4.0% were found.

2.4 Water and energy forecasting techniques

2.4.1 Introduction on forecasting techniques

Although forecasting is not a new discipline, its application in the water sector for consumption estimation has not yet been successful, probably due to the nature and quality of available data, the numerous variables that are hypothesized to affect water consumption (Arbués *et al.*, 2003) and the multiplicity of forecast horizons and periodicities involved. These characteristics have engendered a plethora of studies in an attempt to improve forecast reliability (Donkor *et al.*, 2012).

Reliable urban water consumption forecasting provides the basis for making operational, tactical and strategic decisions for water utilities and is critical for several reasons. Forecast horizon, mainly short-term operation and long-term planning of water resources are two of the most important reasons for forecasting water consumption (Polebitski and Palmer, 2010). For instance, utilities need to know what the water consumption for today and tomorrow will be in order to operate their treatment plants and wells appropriately to meet these consumptions. Utilities also need to predict the water consumption 20-30 years in the future in order to develop new water sources and/or expand their infrastructures (Donkor et al., 2012).

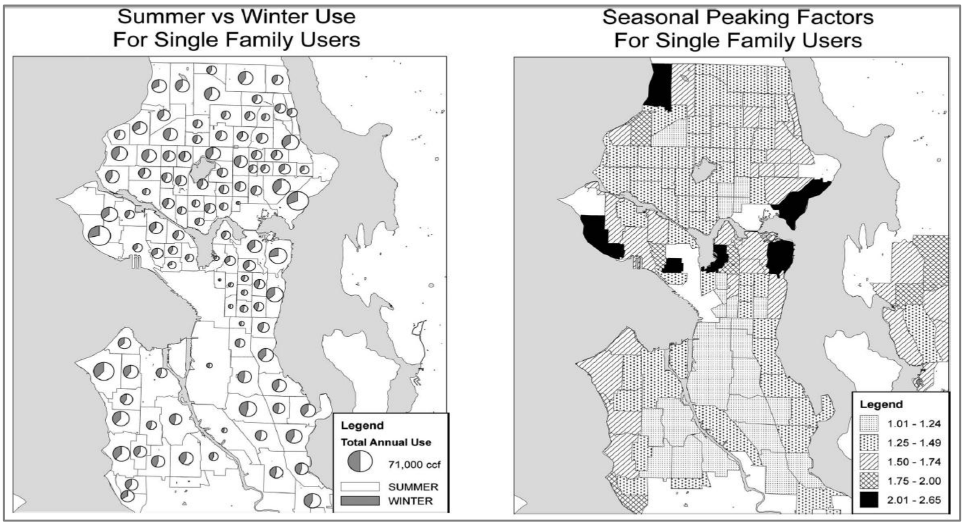
Forecasting can also be divided into a temporal and spatial component. The poor characterization of each can have significant impacts on the forecasting models. Typically the temporal component has been more developed with models reflecting the seasonality of water consumption (McPherson and Witkowski, 2005).

Seasonality in time series refers to periodic fluctuations around the trend line. For flow time series in particular, seasonality is associated with clients' change of behaviours in certain periods of time, for instance, due to changes in temperature (*e.g.* winter, summer).

Concerning temporal forecasting, few forecasting studies determine elasticity for variables on a seasonal (or intra-annual) basis, despite the fact that large and important differences may exist between winter and summer behaviours, for instance (Polebitski and Palmer, 2010).

Polebitski and Palmer (2010) explores temporal and spatial components. To exemplify, Figure 10 a plots an average of annual total consumption in Seattle divided into winter and summer usage for each census tract, with the size of the circle representing total consumption relative to other tracts. For most census tracts, outdoor usage is highly seasonal, driven primarily by weather as with most major mid-latitude cities. Seasonal peaking is spatially variable and dependent on demographic characteristics. The study refers to the seasonal peaking factor as the ratio of average summer consumption to average winter consumptions within a census tract (Figure 10 a. There is considerable spatial variation in single-family seasonal peaking factors (Figure 10 b).

Studies have demonstrated that socioeconomic variables are responsible for the long-term effects on water consumption, while climatic variables are mainly responsible for short-term seasonal variations in water consumption (Adamowski, 2008).



(a) (b)

Figure 10 – Temporal and spatial forecasting: a) Average winter and summer single-family usage; b) Single-family seasonal peaking factors by census tract for Seattle (Polebitski and Palmer, 2010)

Aside from monthly scenarios, weekly scenarios should also be explored given the differences between working days and weekends (Arbués *et al.*, 2003; Loureiro, 2010; Palau *et al.*, 2012; Widén and Wäckelgård, 2010). In particular, Palau et al. (2012) studied water inflows into DMAs of urban networks, finding that differences are particularly important for morning and afternoon models, as depicted in Figure 11.

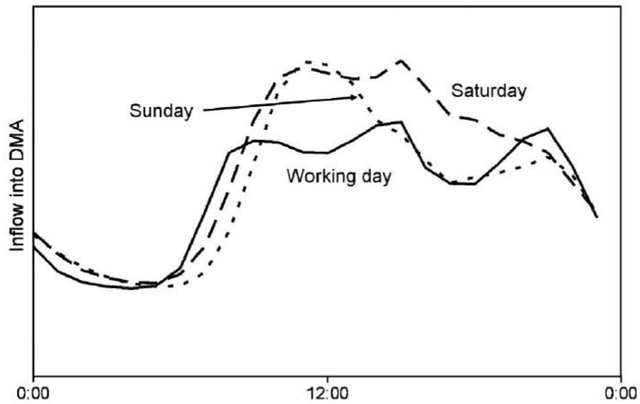


Figure 11 – Average water flow pattern during working days and weekends (Palau *et al.*, 2012).

In respect to water variables used in forecasting models, results of a survey by the AWWA portray (in decreasing order of frequency) the kind of variables that are of interest to water utilities in urban water consumption forecasting: peak day (73.9%), daily total system consumption (65.9%), monthly total system consumption (65.6%), annual per capita consumption (65.4%), annual consumption by customer class (58.0%)

and revenue (57.9%) (Donkor *et al.*, 2012). These results clearly show that urban water consumption forecasting can involve different variables measured at different periodicities (Adamowski, 2008).

Some examples of forecasting techniques will be presented in the following sections along with their main advantages and disadvantages.

Time Series

Time Series (TS) analysis is a statistical model that forecasts future water consumption on the basis of past observations and associated error terms. The referred model is technically known as extrapolation forecasts in the water industry.

This class of models does not account for the effect of exogenous variables such as weather or price, but rather rely on the assumption that past trends will be repeated in the future.

Their failure to take into consideration the effects of changes in demographic, economic and technological variables, as well as water consumption management strategies (such as public awareness campaigns and/or price adjustments) in influencing future water consumption is the main criticism (Donkor *et al.*, 2012).

Nevertheless, an advantage of time series analysis is the capability of revealing the autocorrelation structure of a short-term water consumption pattern over time (Adamowski, 2008).

Multiple Linear Regression

Multiple Linear Regression (MLR) analysis is a statistical model used to test whether a dependent variable, in this case water consumption, is related to more than one independent variable (*e.g.* consumers' behavioural, socioeconomic and technological characteristics)

MLR models can only identify relationships of a pre-specified functional form, and they may not always be sufficient to accurately predict the nonlinear function of the variables involved (Adamowski, 2008). Notwithstanding, MLR is one of the most widely used regression techniques for water consumption. Most researchers have applied factors such as the price of water, number of people in the household, household income, lot size, whether they had a swimming pool, and some climatic data. Additionally, some studies link water use consumption patterns to urban land use patterns in Finland by applying multiple regression analysis to geographic information system and water consumption data (Donkor *et al.*, 2012).

The influence of economic variables on the annual electricity consumption in Northern Cyprus is examined using multiple regression analyses and the relationship between energy consumption, the number of customers, the price of electricity and the number of tourists is determined (Suganthi and Samuel, 2012).

Artificial Neural Networks

Artificial Neural Networks (ANN) uses a mathematical model of biological networks to simulate consumption in a dwelling. It models input determinant variables as a series of neurons. Each neuron can interact with others through a feedback mechanism.

When applied to water consumption forecasting, ANN models a complex number of input parameters that affect consumption at home as well as the influence of each parameter on each other. Their self-learning capabilities can result in an accurate means of modelling consumption within the home. Another advantage is that ANNs make no assumptions about the nature of the relationships between input and output variables (Adamowski, 2008). However, neural networks can also suffer from multi-collinearity issues where high levels of appliance saturation exist, as reported by McLoughlin *et al.* (2012).

Artificial neural networks have recently begun to be used for short-term water consumption forecasting. Currently, the application of neural network models for water consumption forecasting in the literature normally involves a comparative assessment of the performance between different neural network models and conventional regression models or with time series models. An example will be provided in the following section.

In the electricity level, ANNs have been used to forecast electricity consumption at a utility level; however, they have also been applied at a domestic level. Domestic application examples include a neural network to model electricity consumption for domestic appliances, lighting, space cooling, space and domestic hot water heating. Results showed that electricity consumption mostly was influenced by appliance ownership and usage, income, dwelling type and household composition ANNs is also used to model the energy consumption of appliances, lighting, and space-cooling in Canadian residential sector (McLoughlin *et al.*, 2012).

2.4.2 Peak water consumption estimation example

Adamowski (2008) compares multiple linear regression, time series analysis, and artificial neural networks as techniques for peak daily summer water consumption forecast modelling. Peak daily water consumption forecasts are required for the cost-effective and sustainable management and expansion of urban water supply infrastructure. It has been shown that the peak summer water consumption process is stochastic and nonlinear because outdoor water use, the major component of peak summer water consumption, depends on the duration and intensity of rainfall and the characteristics of temperature. As such, the forecasting of peak summer water consumption is complex, and thus the use of artificial neural networks (ANNs), which are capable of modelling nonlinear systems, needs to be explored.

This study used climatic variables, past water consumption and population. More specifically, the data used in this study consisted of daily total rainfall (mm), maximum daily temperature (°C), peak daily water consumption (Ml/day) and population. The peak daily water consumption for a specific day was the peak hour water consumption for that day.

The analysis was carried out on 10 years of peak daily water consumption data and meteorological variables (maximum daily temperature and daily rainfall) for the summer months of May to August and for an area of high outdoor water usage in the city of Ottawa, Canada.

Thirty-nine multiple linear regression models, nine time series models, and 39 ANN models were developed and their relative performance was compared.

From the results of this study, the following was concluded:

- The use of artificial neural networks for use in forecasting peak daily water consumption in the summer months in an area of high outdoor water consumption is marginally better than multiple linear regression and time series analysis.
- Peak daily water consumption is better correlated with the rainfall occurrence rather than the rainfall amount itself.
- Assigning a weighting system to the antecedent days of no rainfall does not result in more accurate models.

2.5 Understanding water-energy nexus

Energy and water uses are interconnected, as schematically presented in Figure 12. The energy sector uses water for mining, refining, and processing liquid fuels and for cooling thermoelectric power plants, while the water sector uses energy for collecting, treating, distributing, and heating water and for collecting, treating and discharging wastewater. This water-energy nexus presents opportunities for effective management and use of both energy and water through various methods. Therefore, saving water and saving energy are intrinsically related.

Knowing and quantifying the energy inputs to water systems is important to understand where acting can be more beneficial. Concerning the energy inputs to water systems, there are four basic elements to consider (Wilkinson, 2007):

1) Primary water extraction and supply delivery: Moving water from source to treatment facilities and tehn to service areas is often energy intensive. In some places this is the largest energy input, particularly when there are pumping stations.

2) Treatment and distribution within service areas: Within local service areas, water is treated, pumped, and pressurized for distribution. Local conditions and sources determine both the treatment requirements and the energy required for pumping and pressurization.

3) On-site water pumping, treatment, and thermal inputs: Water users require energy to further treat water supplies (e.g. softeners, filters, etc.), circulate and pressurize water supplies (e.g. building circulation pumps), and heat and cool water for various purposes.

4) Wastewater collection, treatment, and discharge: Finally, wastewater is collected, treated, and discharged. Wastewater is sometimes pumped to treatment facilities where gravity flow is not possible, and standard treatment processes require energy for pumping, aeration, and discharge.

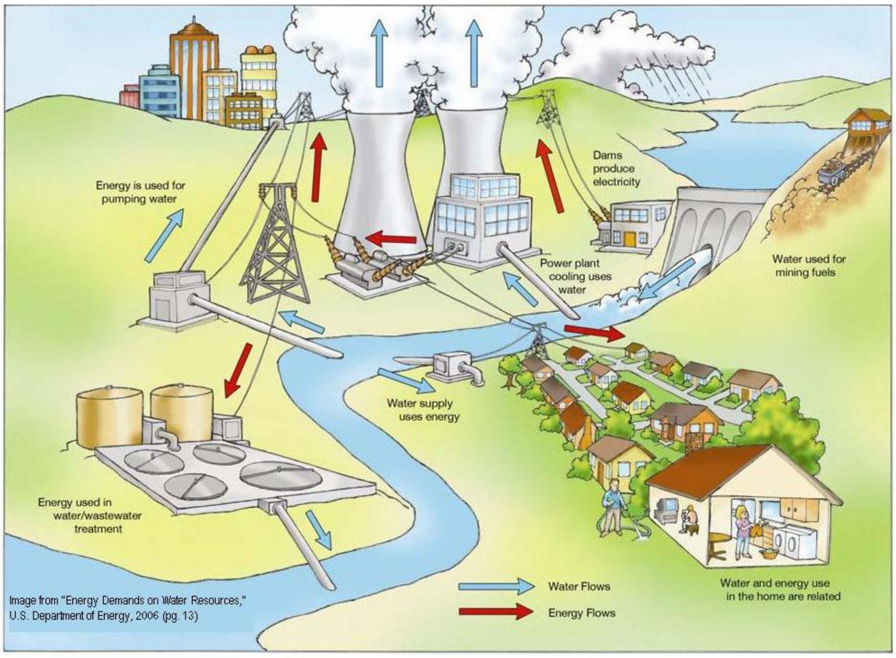


Figure 12 – Water-energy nexus on a macro level (Energy, 2006)

Most of the electricity use in water systems is for pumping, so reduced volumes of flow result in energy savings. Energy management opportunities also exist in improved equipment and operational control systems. Examples include the use of high efficiency motors and .adjustable speed drives, efficient pumps, and effective instrumentation and controls (Wilkinson, 2007)

Historically, US conservation efforts have largely been implemented separately across the water and energy domains even though there are numerous benefits to coordinate conservation efforts and exploit water-energy linkages. Understanding water-energy linkages is necessary for water and energy managers to better plan and promote collaborative water and energy conservation actions (Abdallah and Rosenberg, 2012).

In the residential sector, for instance, the most important link between of water and energy is the energy needed to heat water. Energy expenditures to heat water comprise 17% of total household energy consumption and exclude the energy embedded to deliver potable water to households. Over the past two decades, U.S federal agencies mandated several water and energy efficiency standards for household appliances which have changed water and energy consumption patterns (Abdallah and Rosenberg, 2012)..

The first US energy efficiency standard for residential water heaters took effect in 1990 and was updated in 2004. In 1992, the US Environmental Protection Agency (EPA) and the US Department of Energy established the Energy Star Efficiency Program which helped improve water and energy technical performance efficiency in appliances like water heaters, clothes-washers, and dishwashers (Abdallah and Rosenberg, 2012).

The Energy Policy Act of 1992 (EPAct), was the first federal water-use efficiency standard which mandated all toilets, showerheads, and faucets to operate below a maximum volume or flow rate per use. Later, in 2006, EPA launched Water-Sense, a voluntary program to promote stricter water efficiency levels that reduce volume or flow rates by 20% from EPAct standards (EPA-WaterSense 2010). As a result of these energy and water efficiency programs, indoor appliance efficiencies have significantly improved over the past two decades.

Examples of synergistic energy-water models include the Watergy model (Abdallah and Rosenberg, 2012), which estimates water, direct and embedded energy saved in a hypothetical federal facility through adopting higher efficient water use appliances (e.g., faucets and toilets). The model uses average behaviour values (appliance use frequency) and technologic values (appliance use flow rate or volume), and energy-use factors (e.g., water heater efficiency, intake and set point temperatures) to estimate annual water and energy savings. Watergy also estimates expected payback periods of adopted high-efficient appliances from saving water and energy.

The Alliance for Water Efficiency (AWE) model estimates water use and water and energy savings for utilities that target water conservation actions to sub-populations of their customer bases. Conservation actions include installing high-efficient toilets, faucets, and clothes-washers. The model uses average values for demographic, behavioural, and technological water use factors to estimate program-wide water savings as well as average lump-sum energy parameters to estimate energy use for each appliance per volume of water used.

Most other water and energy models use deterministic (average) approaches to estimate water consumptions by and savings from water and energy conservation actions for major appliances in typical or representative and homogenous household populations (Abdallah and Rosenberg, 2012). These deterministic approaches assume households behave similarly and have similar appliances and conditions. In actuality, residential water and energy uses are heterogeneous and vary significantly among households with demographic (household size), behavioural (use frequency or duration), technological (appliance use volume or flow rate, water heater intake and set point temperatures, heater energy source, heater efficiency), and geographic (climate and water availability) factors contributing to variations among users. When multiplied together to estimate water and energy uses, the uncertainties associated with these factors multiply rather than cancel. Thus, there is a need for more holistic, integrated and heterogeneous approaches to estimate household water and energy linkages and identify targeted, joint, water-energy conservation opportunities.

Complementarily to water use conservation for energy saving, it has been proven that water reuse (*e.g.* recycled water) is far less energy intensive (Wilkinson, 2007).

Stillwell and Webber (2010) estimated for the first time the potential of water reuse in Texas. According to their findings, Texas could save 908 x 10^6 m^3 of water and 330 to 800 x 10^6 kWh of electricity annually. Incorporating water reuse as 12% of total water consumption, Texas could save 2200 10^6 m^3 of potable water annually. This water reuse requires an energy investment of 710 x 10^6 kWh annually for additional treatment and distribution of reclaimed water, yet saves energy for potable water treatment, estimated at 800 to 1040 x 10^6 kWh annually. This energy accounting shows a net energy savings of 73 to 310 x 10^6 kWh annually.

Additionally, Stillwell and Webber (2010) found that water reuse operations are best suited for municipalities currently using advanced wastewater treatment processes, thereby decreasing the energy investment needed for reclaimed water treatment.

There are still few studies addressing the water-energy nexus. A vital step towards approaching the nexus is to develop robust analytical tools, conceptual models, appropriate and validated algorithms, and robust datasets that can provide reliable information on the future use of energy and water (Bazilian *et al.*, 2011).

More approaches are also needed to evaluate the potential of water and energy savings taking into account the energy consumption in water systems.

2.6 Summary and conclusions

The implementation of District Metered Areas (DMA) and Pressure Management Areas (PMA) improves the efficiency of distribution systems, not only reducing leakage and energy consumption, and consequently operation costs, but also improving the level of service provided. It can also be an example to encourage other utilities to improve their performance.

Flow data measurements have inevitably outliers. Outliers can be errors resulting from data acquisition and processing as well as flow peaks associated with pipe bursts, or abnormal consumptions. Faster detection of bursts saves water, minimises the inconvenience of interruption to customers and decreases the damaging consequences to the infrastructure. Two representative methods for detecting outliers have been presented in this state-of-the art review.

Concerning the driving factors of water and energy consumption, socio-demography, income, land use, billing and climate play a key role in explaining consumption. Most of the literature focuses on income, with few studies addressing drivers related with immigration (*e.g.* number of immigrants from developing countries), ageing (*e.g.* number of elderly), urban land uses (*e.g.* density of houses) and climate (*e.g.* evapotranspiration). Moreover, little research has focused on water use at the sub-section level, although this is the preferable spatial unit for planners and engineers to evaluate water saving programs, land use policy and spatial pricing schemes.

In terms of consumption forecasting, the most appropriate scale was found to be the census tract level, although most of the studies examine consumption at a more aggregated level. In terms of forecasting techniques, the majority of the studies use regression models. Residential water consumption models have progressed from simple extrapolations to complex models requiring sophisticated computer programs. All models suffer from trade-offs, generally concerning the availability of data. Another remark is that models should be as parsimonious as possible without compromising on structural integrity and forecast quality.

Recent studies forecast short-term consumptions using artificial neural networks with better results when comparing to regression models. Nevertheless, the question regarding which method to use for consumption forecasting cannot be adequately answered without specifying the forecast variable and horizon. In terms of water consumption variables, for instance, despite the relative importance of peak daily water consumption, limited detailed research has been devoted to this topic, including factors driving peak daily water consumption and forecasting methods (Adamowski, 2008).

With respect to the water-energy nexus, the existing studies generally focus on the potentials of water conservation and water reuse due to energy inputs in water systems. Apart from the American initiatives in promoting water and energy efficient equipment, there is clearly a lack of studies combining both areas, particularly in terms of domestic consumptions. The closest link at the domestic level appears to be in terms of the energy needed to heat water. More approaches are also needed to evaluate the potential of water and energy savings taking into account the energy consumption in water systems.

3. METHODOLOGY AND TOOLS FOR CONSUMPTION ANALYSIS

3.1 Introduction

A modified methodology and the main tools and applications used for the consumption analysis – both adapted from Loureiro (2010) – are presented. The general methodology is composed of four modules namely: data collection, data processing, data characterization and profiling water consumption.

A geoprocessing tool that calculates socio-demographic indexes based on Census data are presented, along with two commercial applications: Monitor and Profiler. The first is used to collect reliable flow data and the second is used for data processing.

3.2 General methodology

The general methodology proposed in this work is applied to domestic water consumption and is presented in Figure 13. The main differences between the current work and the previous methodology developed by Loureiro (2010) are highlighted in bold and are marked with a star.

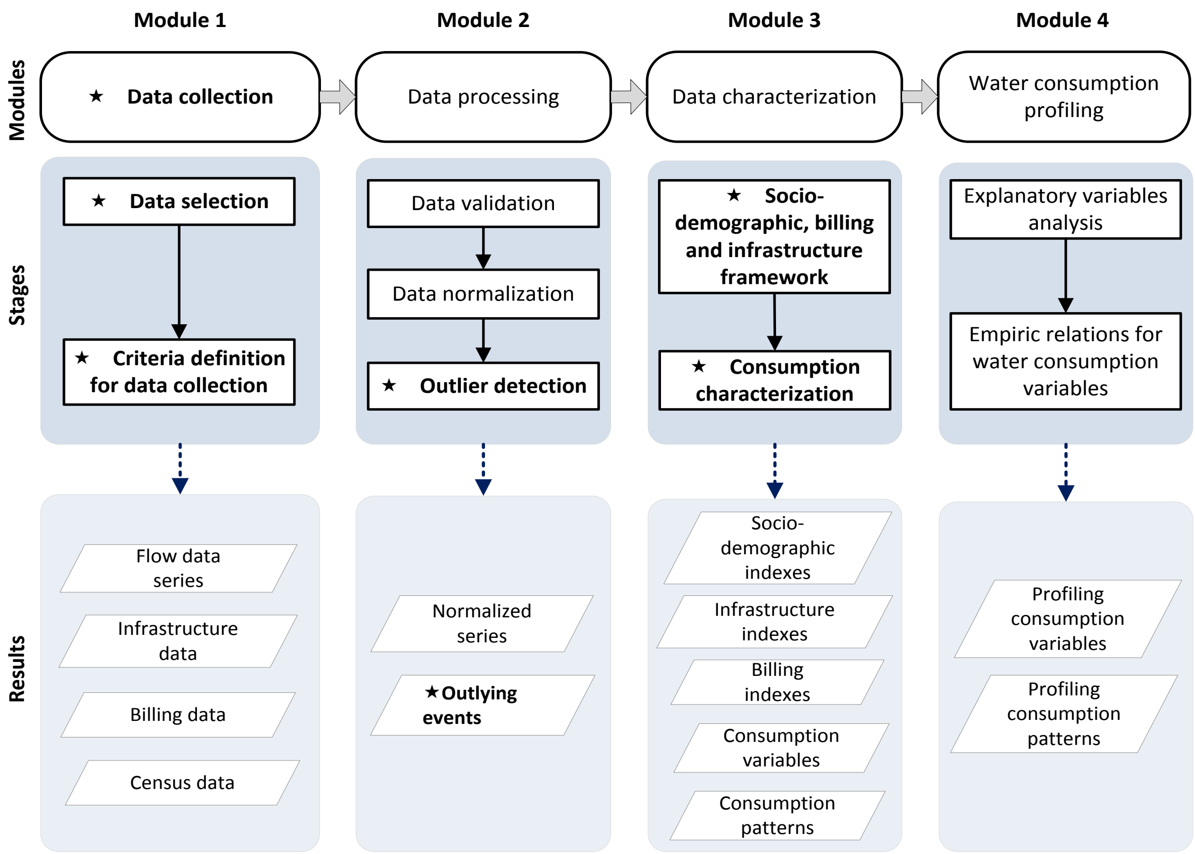


Figure 13 – General methodology for water consumption analysis (adapted from Loureiro (2010))

Module 1 as defined by Loureiro (2010) focused on monitoring planning and implementation, whereas in the current work, it aims at collecting extensive flow data from the supervisory control and data acquisition (SCADA) systems of different water utilities. Thus, this module includes selecting the types of data to be collected, as well as some criteria based on the scope of the analysis. Criteria are defined in 5.2. Flow data series, infrastructure, billing and census data are the results of this module.

Module 2 focuses on validating the collected data, normalizing the time step of the flow data series into a fixed interval of 15 minutes and removing existing outliers. In this module, a significant contribution has been made in terms of the outlier detection, not only by applying and testing a new procedure but also by the intensive work on classifying and characterizing outlying events. This module is developed in Chapter 6.

Module 3 consists of building a framework where socio-demography, infrastructure and billing data enter as input and the outputs include indexes on the same fields, as explored in Chapter 5. New indexes were introduced in addition to the ones analysed by Loureiro (2010). This module also involves describing water consumption among the studied DMAs, which is made building consumption scenarios and defining a list of consumption variables to analyse, along with consumption patterns. Scenario building is carried out using cluster analysis (see Chapter 7). New scenarios are studied comparatively to the work developed by Loureiro (2010).

Module 4 uses the outputs from the previous modules in order to profile water consumption. This profiling is carried out by identifying explanatory variables. These are independent variables which are linear combinations of the socio-demographic, infrastructure and billing indexes previously calculated. This is possible using multivariate statistical analysis. Furthermore, consumption variables and patterns are defined as dependent variables. Then, dependent variables are set against the independent ones using correlation and regression analyses. The result is a profile on consumption patterns in water distribution systems using extensive measurements. This module is developed in Chapter 8.

3.3 Tools and applications

3.3.1 Geoprocessing tool

The geoprocessing tool initially developed by Loureiro (2010) and Rebelo *et al.* (2008) calculates socio-demographic indexes on the categories of Buildings, Dwellings, Families and Individuals at the DMA level. Census data are available at the statistics subsection (SSE) level – see Figure 14.

Statistics subsection (SSE) level is the smallest territorial unit for statistical use available in Portugal and corresponds to the smallest homogeneous building and living area existing inside a statistics section (it corresponds to one city block in urban areas).

Figure 14 – Representation of the statistics subsection level (SSE)

The tool was developed using ArcGIS® and uses a weighting method to convert statistic data at the SSE level into data at the DMA level. The original weighting method is based on the influence areas of each domestic service connections. This tool was improved in the current work, in order to add new weighting methods and the direct calculation of an expanded group of socio-demographic indexes. The we weighting methods of the geoprocessing tool can be summarised as follows:

1) Influence areas: weights according to the influence area of each domestic service connection by calculating the Thiessen´s polygons of each service connection. Assumes that the principle of homogeneity applies, which means assuming that client distribution and water consumption is homogeneous in a certain SSE. This principle cannot be applied in some areas such as the periphery of urban areas.

2) Domestic clients: weights proportionally according to the number of domestic clients in each DMA.

3) Billed consumption: weights proportionally according to the billed consumption in each DMA.

Due to the disadvantages pointed over the 1st method, in case of having georeferenced data on domestic clients and its consumption, it is preferable to use the alternative weighting methods in order to obtain more reliable data. Notwithstanding, it should be noted that when databases belong to very distant periods of time (*e.g.* Census 2001 and clients database from 2012) results cannot be accurate with either model. Census data are only updated with a decade´s periodicity (INE, 2012), which means the best time to analyse the socio-demographics is when census data are released.

The main steps of the developed geoprocessing tool are presented in Figure 15 focusing on the weighting method based on the number of domestic clients. In the first step, SSE with domestic service connections are selected. The data processing that involves calculating socio-demographic indexes for all SSE in the country would be computationally heavy and is not necessary for this analysis purpose.

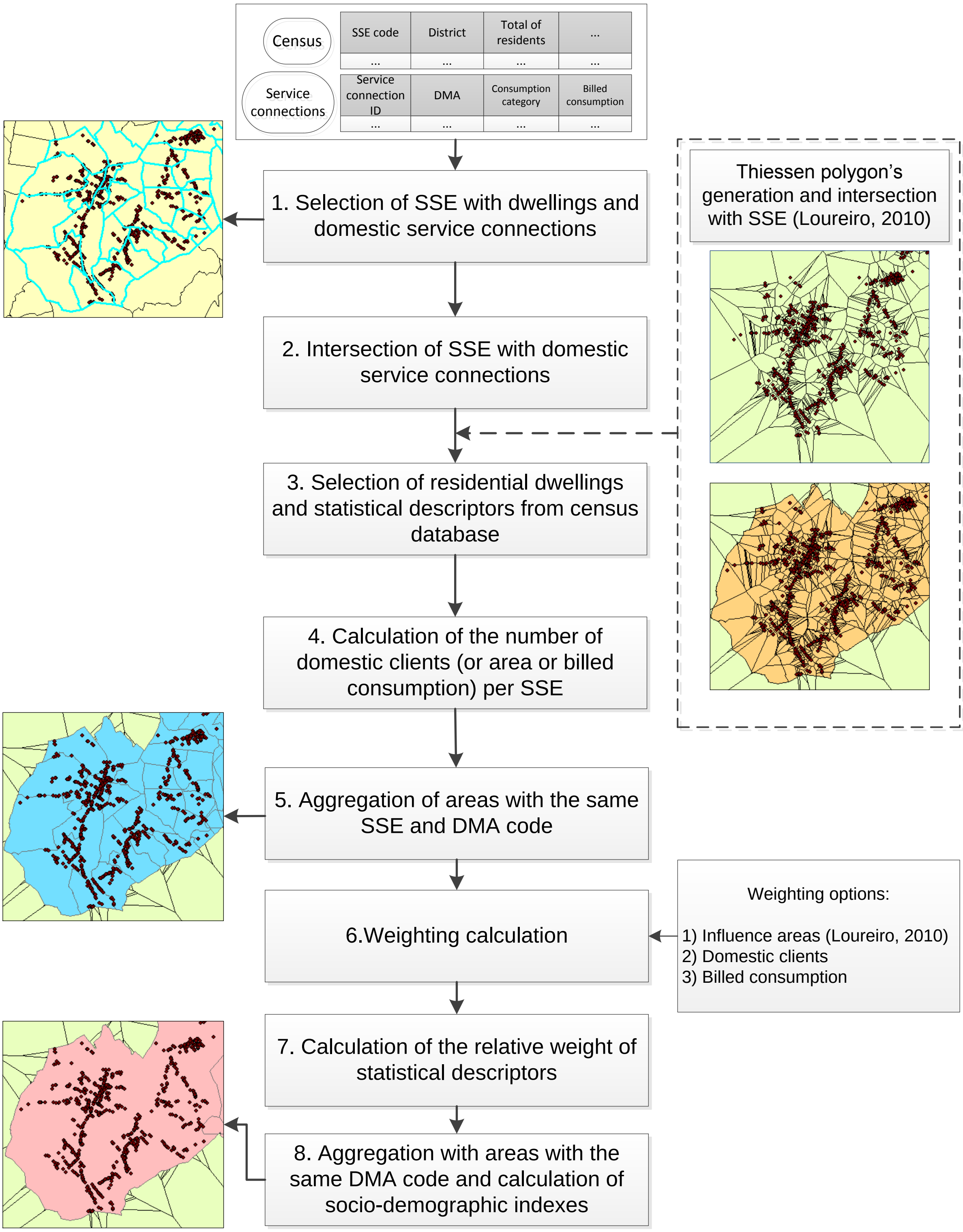


Figure 15 – Main steps of geoprocessing tool that calculates socio-demographic indexes

The first step requires a pre-processing of the infrastructure and billing database. The billing database includes a list of clients, its type, its yearly consumption[1] and the service connection code to which every client is connected. Since water utilities typically use different names for classifying the same type of clients, this pre-processing also includes standardizing the different types into four main categories, as shown in Table 3.

Table 3 – Consumption category standardization (Step 2: Pre-processing)

Consumption category	Examples from water utilities
Domestic	Domestic (or habitation), large families, (5 or more elements), apartment complex, temporary connection, social, low resources
Commerce and industry	Commerce, industry, works, factory, agriculture
Collective	Public institutions, local administration, municipalities, non-profit institutions
Public	Autarchy, fire network, irrigation, garden, general services

Whenever a service connection has more than one client type, the service connection type is classified as domestic if more than 50% of the consumption is domestic; otherwise it is classified with the type of client that represents the majority of the consumption.

In the second step, information on the domestic service connections is intersected with the SSE selected in the first step. In the third step residential dwellings and a set of 55 statistical descriptors are selected from the census database. In the next step, the number of domestic clients per SSE (or area or billed consumption), depending on the weighting method, is calculated. In the fifth step a new field that aggregates SSE and DMA codes is created and the number of consumers that belong simultaneously to the same SSE and DMA is also calculated. This field is very useful since it allows the weighting of every SSE in the DMAs under analysis. Hence, the sixth step consists of calculating the SSE's weight, by dividing the number of domestic clients (or area or billed consumption) within the same SSE and DMA code and the number of domestic clients per SSE (or billed consumption or area), as follows:

[1] Billing data is usually computed by the water utilities with a monthly or bimonthly periodicity. In this work the yearly consumption was calculated for analysis.

$$W_{ij} = \frac{X_{ij}}{\sum X_j} \quad (1)$$

in which:

W_{ij}: Weighting ratio for DMA*i* and SSE*j* [-]

X_{ij}: Number of domestic clients (or area, or billed consumption) for DMA*i* and SSE*j* [-]

$\sum X_j$: Number of domestic clients (or area, or billed consumption) for SSE*j* [-]

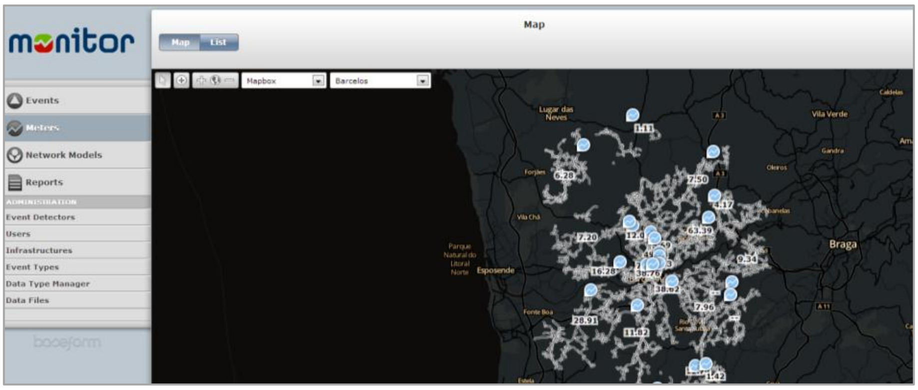
After weighting calculation, the seventh step consists of calculating the relative weight of each statistical descriptor selected in the third step. In the eighth step DMAs are aggregated and socio-demographic indexes are calculated. Indexes are characterized in Chapter 4 and its calculation is presented in Appendix A.

The weighting method that uses influence areas (Loureiro, 2010) requires the generation of the Thiessen polygons' and its intersection with the SSE instead of the second step previously described.

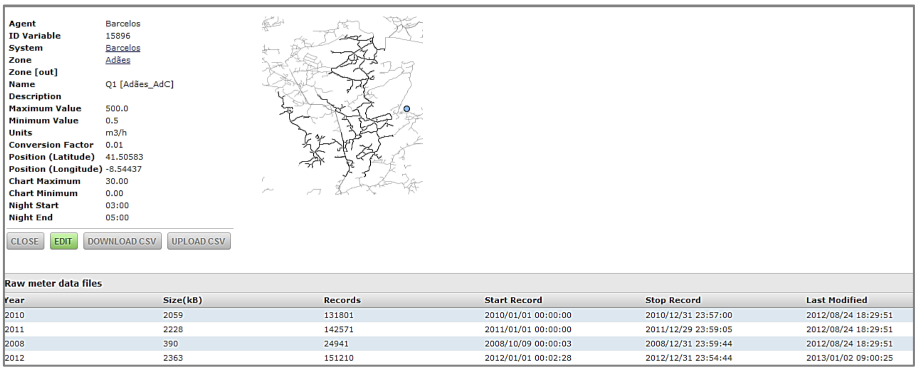
Loureiro (2010) did not include in her study indexes related with the Dwellings but, since this analysis uses Census 2011 data and there might have been alterations at this level that may influence water consumption, indexes related with Dwellings were also calculated.

3.3.2 Monitor

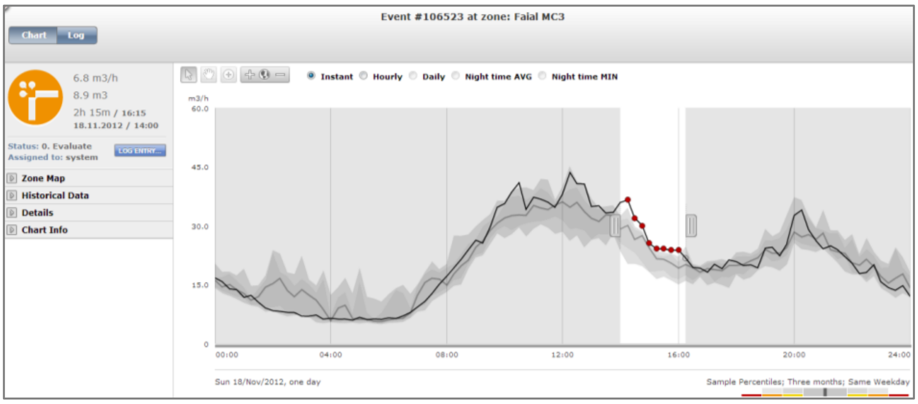
Monitor is a commercial application for real-time monitoring and analysis of water network data (www.baseform.org). It was used in this work for downloading flow data series of some water sectors and for visualizing the location of water meters – Module 1 of the general methodology. This application has several other applications such as a method for identifying outliers in the flow data series. Figure 16 shows the map environment with the location of water meters, the working environment regarding flow data availability ready for download, as well as an example of an outlying event detected in the water sector.



(a)



(b)



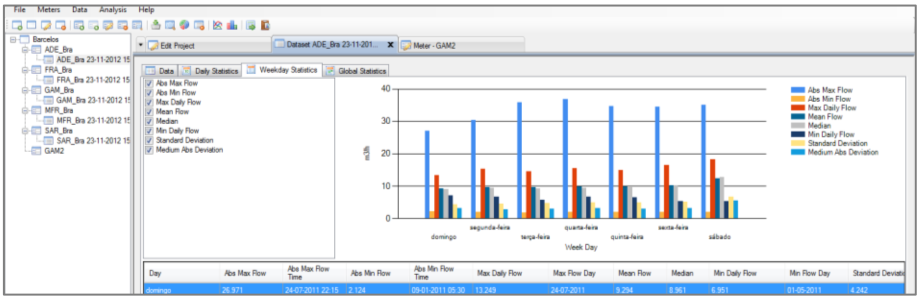
(c)

Figure 16 – Examples of the Monitor application: a) Map environment and meter location for a defined water sector b) Flow data availability for download c) Detected outlying event

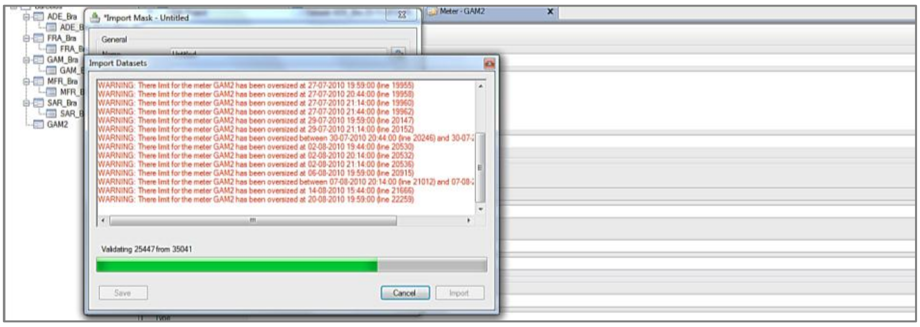
3.3.3 Profiler

Profiler is an application developed by Loureiro (2010) and is used in the context of the present work to validate and normalize flow data series – Module 2 of the general methodology. This application is also used to understand the general behaviour of the series and to search for seasonality. Figure 17 shows examples of the working environment of this application, the validation process and the series' general behaviour.

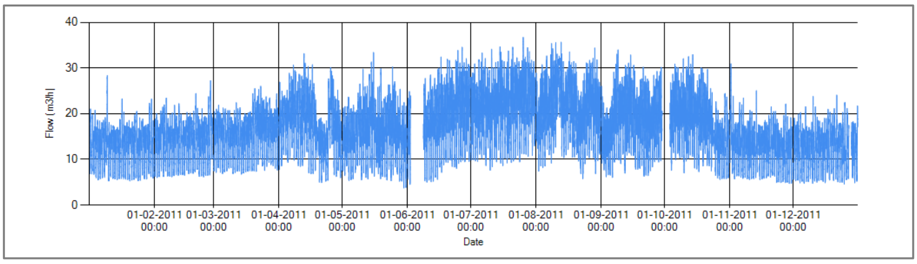
Profiler can also be used to calculate the average statistics of the raw series and the normalized ones, as well as to compute consumption patterns for different types of days. These capabilities have been validated during the development of the present work, contributions were made in the users´ manual and several bugs of this application have been reported and corrected, therefore improving its performance.



(a)



(b)



(c)

Figure 17 – Examples of the Profiler application: a) working environment showing general statistics menu; b) raw series validation process; c) flow series behaviour

3.4 Summary and conclusions

In this chapter the general methodology applied in the present work is presented, along with the geoprocessing tool and two applications that were used in this work. The methodology consists of four modules: *Data collection, Data processing, Data characterization and Water consumption profiling*. This work uses extensive flow data measurements, reason why the use of tools and applications for data collection and processing was extremely important. *Flow data collection* (Module 1) is partially carried out using the Monitor application. *Data processing* (Module 2) is carried out using the Profiler application, mainly in the validation and normalization steps. The geoprocessing tool presented and developed in a GIS environment is crucial for the socio-demographic analysis (part of Module 3) since it allows determining the socio-demographic trends of the water sectors based on Census data on the SSE level.

4. SOCIO-DEMOGRAPHIC, BILLING AND INFRASTRUCTURE ANALYSIS

4.1 Introduction

The present chapter aims at presenting the socio-demographic, billing and infrastructure analysis of nearly 100 water distribution sectors in Portugal. This chapter is part of Module 3 from the general methodology presented in section 3.2.

Socio-demographic, billing and infrastructure indexes were selected as being driving factors of water consumption (Loureiro, 2010; McPherson and Witkowski, 2005; Polebitski and Palmer, 2010) and are calculated to characterize the consumers' behaviour. These indexes will be used in Chapter 8 as inputs for a water profiling model, since they provide essential information for forecasting consumptions and planning new water sectors.

The methodology for socio-demographic, billing and infrastructure analysis and the geoprocessing tool that calculates a series of socio-demographic indexes in a Geographic Information System (GIS), both developed by Loureiro (2010), were improved in this study. The geoprocessing tool uses recently published census data.

4.2 Methodology

The general methodology for socio-demographic, infrastructure and billing analysis of water sectors involves five steps, as depicted in Figure 18. The first step is "Data collection" and consists of collecting georeferenced data from census, infrastructure and billing. Census data were collected directly from the National Statistics Institute website (www.ine.pt) at the census tract level and includes a series of 122 statistical descriptors from which 48 were selected.

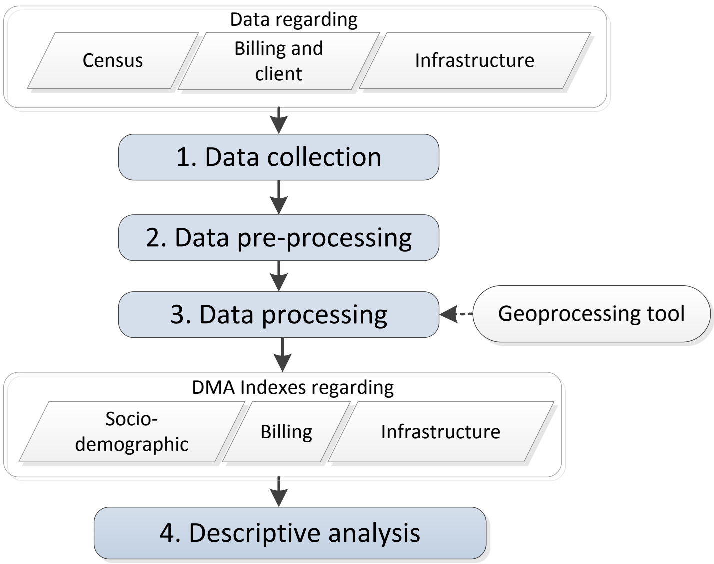


Figure 18 – Methodology for socio-demographic, billing and infrastructure analysis

Infrastructure and billing data were provided by three water utilities and include information on pipes, service connections and types of clients (*i.e.* domestic, collective) and its annual consumption.

Since water utilities usually have information organized in different formats and using various terminologies, a pre-processing is necessary. Therefore, the second step is "Data pre-processing" and includes standardizing fields and combining data (if necessary) in order to have databases with only the required information for the processing step (Table 4).

Table 4 – Field standardization according to analysis type and database (Step 2: Pre-processing)

Analysis type	Database	Standardized fields
Socio-demographic and billing	Service connections combined with clients	DMA, Service connection ID, Client ID, Consumption category, Annual billed consumption
Infrastructure	Pipes	DMA, Year, Material, Diameter, Length
	Service connections	DMA, Service connection ID, Annual billed consumption

The third step refers to "Data processing". For socio-demographic data processing, a geoprocessing tool initially developed by Loureiro (2010) was improved and used in this research study, as previously presented in section 3.3.1. As for the remaining analysis (Infrastructure and billing), processing was carried out by using Microsoft Excel®. As a result of the processing step, socio-demographic, billing and infrastructure indexes are calculated, and summarised in Table 5.

Table 5 – Categories and number of indexes calculated for socio-demographic, billing and infrastructure analysis

Analysis type	Index categories	Number of calculated indexes
Socio-demographic	Buildings, Dwellings, Families and Individuals	22
Billing	Domestic consumption and Non-domestic consumption	18
Infrastructure	Pipes and Service connections	9

The forth step is "DMA selection" and consists of selecting DMA which fulfil a set of eligibility criteria which are defined according to the purpose of analysis. The criteria used for each analysis is presented in Table 6.

Table 6 – Criteria used for DMA validation (socio-demographic, billing and infrastructure analysis)

Analysis type	Criteria
Socio-demographic and infrastructure	1) Number of services connections: 150-6000 2) Percentage of domestic clients: >85%
Billing	1) Daily consumption per client and per day < 668l/client·day

The number of service connections was established in order to include DMA with different dimensions and taking into account the recommendations from previous studies (Farley and Trow, 2003; Jankovic´-Nisic´ *et al.*, 2004). The percentage of domestic clients was defined in order to guarantee that the selected DMA were mostly domestic which is in agreement with Loureiro (2010) analysis. The average daily consumption per client and per day was set considering that Portuguese average daily consumption is 167 l/(inhabitant·day) (Eurostat, 2007) and according to Census 2011 the maximum number of inhabitants per water client is 4, which gives an average daily consumption per client of 668 l/client.

Finally the fifth step consists of a "Descriptive analysis" of the selected DMA using the socio-demographic, billing and infrastructure indexes obtained in the third step.

4.3 Case-studies description

The socio-demographic, billing and infrastructure analyses that are carried out in the current chapter involved the collection and validation of data from 150 DMAs belonging to six different Portuguese water distribution systems distributed along the country. Data were collected at the DMA level and each DMA is identified with a code that has an abbreviation of the DMA name, an underscore and an abbreviation of the district name. For instance, MFR_Bra, which refers to a DMA in Braga district. The validation criteria presented earlier in Table 7 was applied to the 150 DMAs. The number of DMAs validated for each analysis is presented in Table 8.

Table 7 – Number of DMAs validated for socio-demographic, billing and o.

Analysis type	N^{o}. of DMAs validated for analysis
Socio-demographic	96
Billing	101
Infrastructure	96

As depicted in Figure 19, the validated DMAs are located in the districts of Oporto, Braga, Lisbon and Setúbal, were abbreviated as “Por”, “Bra”, “Lis” and “Set”, respectively. DMAs’ general characteristics are also presented in Table 8.

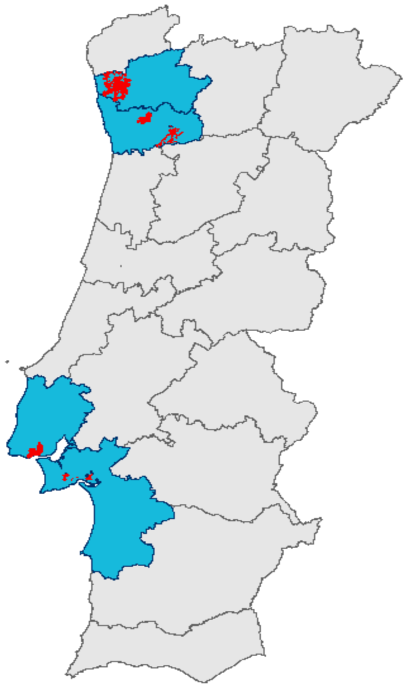


Figure 19 – Portugal map and districts with analysed DMAs. Braga, Oporto, Lisbon and Setúbal (from the top)

Table 8 – General characteristics of DMAs selected for socio-demographic, billing and infrastructure analysis (≈100 DMAs)

Characteristic	Interval	Average value	Median value
Average diameter [mm]	40 – 630	109	100
Network length [km]	4.0 – 95.0	25	18
N.º service connections	151 – 5863	784	587
N.º of clients	93 – 23527	1926	1649
N.º of domestic clients	86 – 21778	1810	1575
N.º of inhabitants	46 – 12778	1789	1140

4.4 Results

4.4.1 Socio-demographic analysis

The methodology defined in 5.2 was applied to 96 DMAs that fulfilled the validation criteria established in Table 6. Socio-demographic indexes were obtained using Census 2011 data with statistical descriptors of

266310 SSE, twice the number of SSE in 2001. Before the availability of Census 2011 data, a preliminary analysis was carried out using Census 2001 data and presented in Mamade *et al.* (2012).

Mamade *et al.* (2012) refers another validation criteria which is the ratio between the number of inhabitants from census data and the number of domestic clients from billing data. If every domestic client is connected to the water distribution system, then the number of domestic clients and the number of households should be approximatley the same. Using the number of inhabitants provided by census (INE, 2012), it was found that the ratio between the number of inhabitants and households (or domestic clients) is approximately 3.

However, results show that in some DMAs the ratio between the number of inhabitants and domestic clients is higher than 3 (see Table 39 in Appendix B) using data from the same period of time. This result might be due to the existence of households that are not connected to the water distribution system (*e.g.* households use groundwater from private wells). For instance, Figure 20 presents a DMA where there are no service connections associated to the majority of households (red circle). This fact has been reported to the water utility and in response, it was found that only 53% of the clients in this DMA were connected to the network.

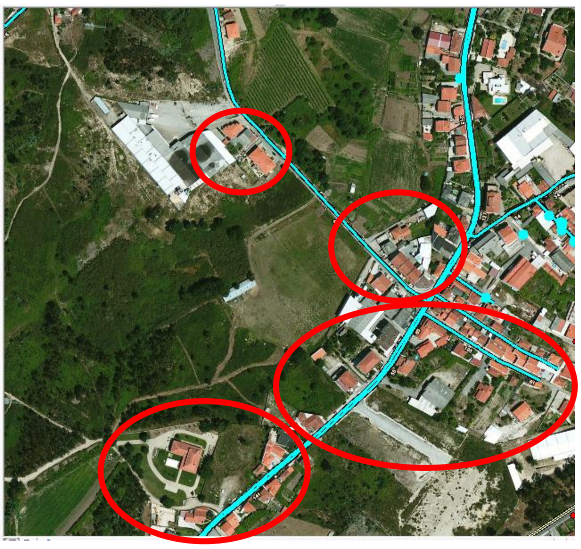


Figure 20 – Dwellings not connected to the water distribution system

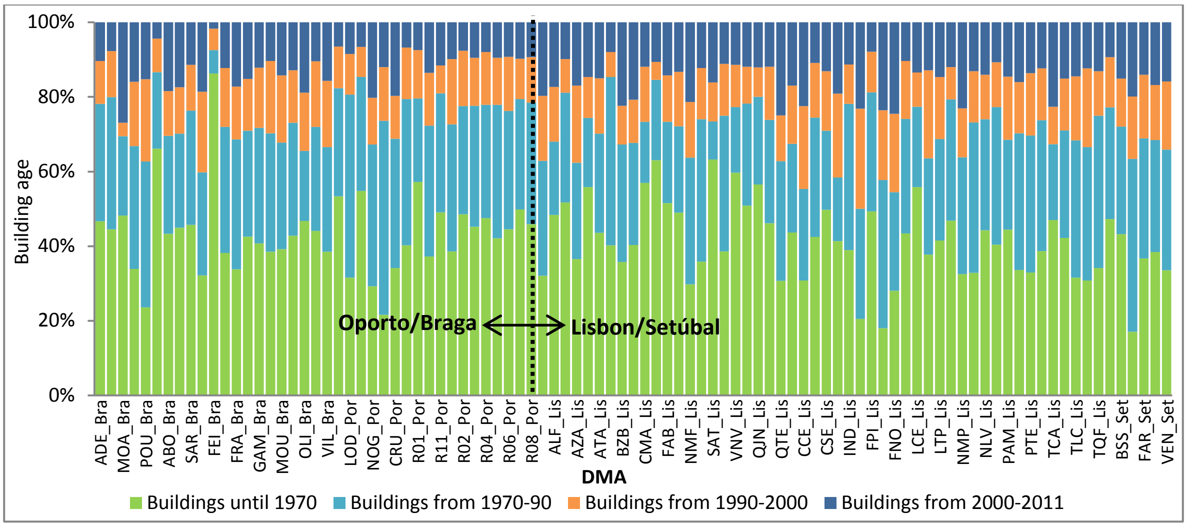
The socio-demographic indexes used for the analysis are presented in Table 9 and its calculation is defined in Appendix A.

Table 9 – Categories and socio-demographic indexes used in the analysis

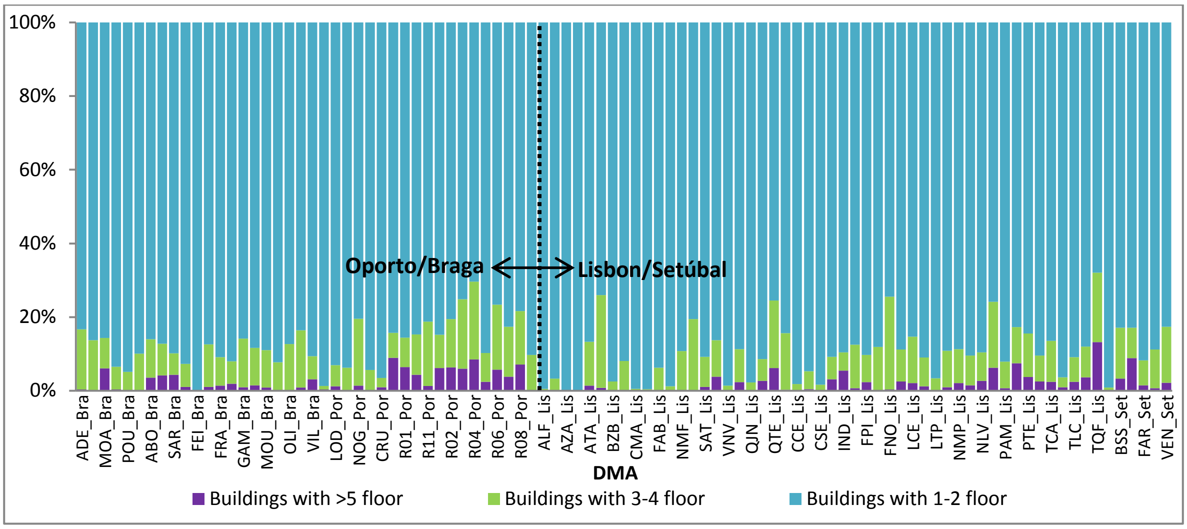
Category	Socio-demographic indexes	Units
Building	Buildings until 1970, 1980, 1990, 2000 and 2011 Buildings with 1-2 , 3-4 and ≥ 5 floor	[%]
Dwelling	Residential immobility Rented dwellings Vacant dwellings	
Family	Families with adolescents Families with elderly Families with unemployed Small families (1-2 elements) Medium families (3-4 elements) Large families (≥5 elements)	
Individual	Population above age 65 years Inactive workers University graduates Economic mobility Active population mobility Population with 12 years of education,	

(b)

Figure 21In what concerns the age of buildings, (Figure 21a), 42% of buildings has more than 40 years (buildings until 1970), 30% were built from 1970 to 1990 and the remaining buildings are from the last two decades. DMA's belonging to the districts of Lisbon and Setúbal have more recent buildings than the ones belonging to the districts of Oporto and Braga (15% against 12%, respectively). As for the number of floors (Figure 21b), 89% of the buildings have 1-2 floors, 9% have 3-4 floors and nearly 2% have 5 or more floors. The fact that the majority of buildings have 1-2 floors indicates the presence of villas. This type of building usually has outdoor uses, such as gardens and lawns that can have irrigation systems connected to the public water supply system (Adamowski, 2008) This fact can help explaining increases of the water consumption in DMAs with this characteristics, as well as explain behaviour changes in the Summer season (Vieira *et al.*, 2002).



(a)



(b)

Figure 21 – Socio-demographic indexes concerning: (a) age of the buildings; (b) number of floors

In respect to the dwellings (Figure 22), an average of 66% of the dwellings are occupied and are used as the main residence of at least one family (Residential immobility) and an average of 15% of them are rented. As for the vacant dwellings an average of 14% of the dwellings is free for buying, renting, demolition or other purposes (INE, 2012). Comparing these results with 2001, the number of vacant dwellings has risen 35% (INE, 2012).

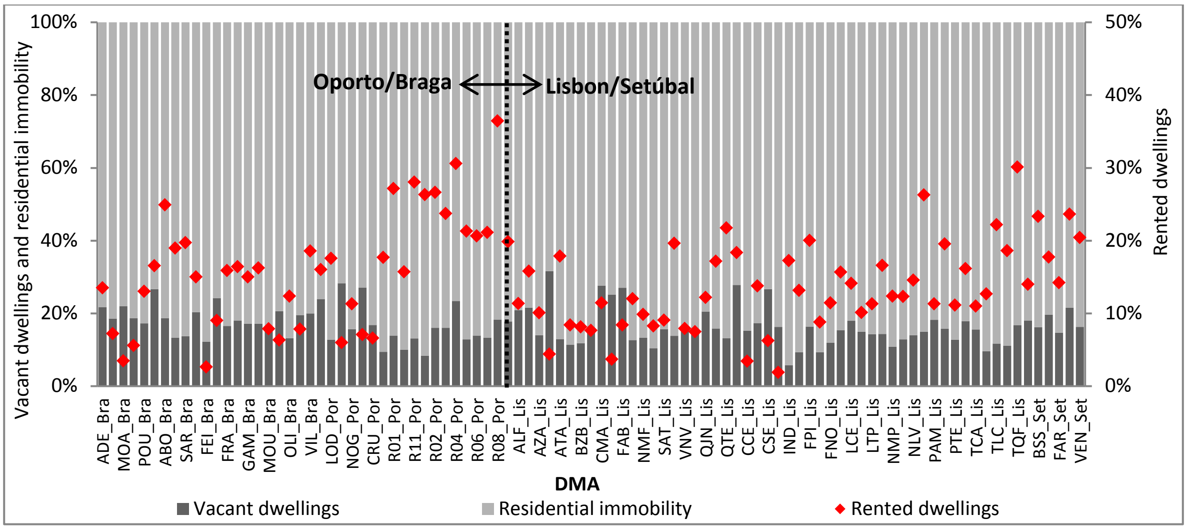
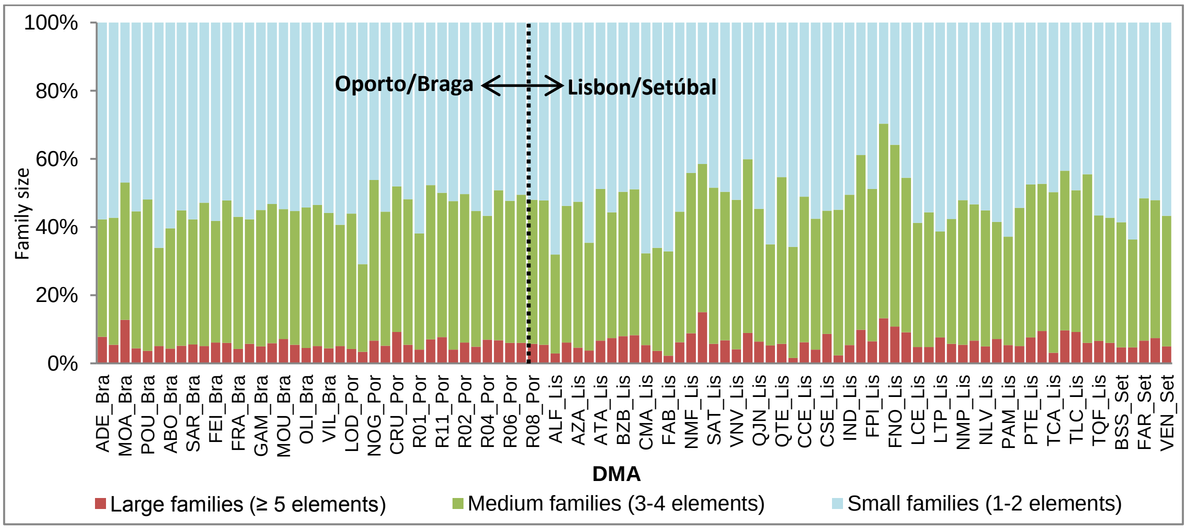
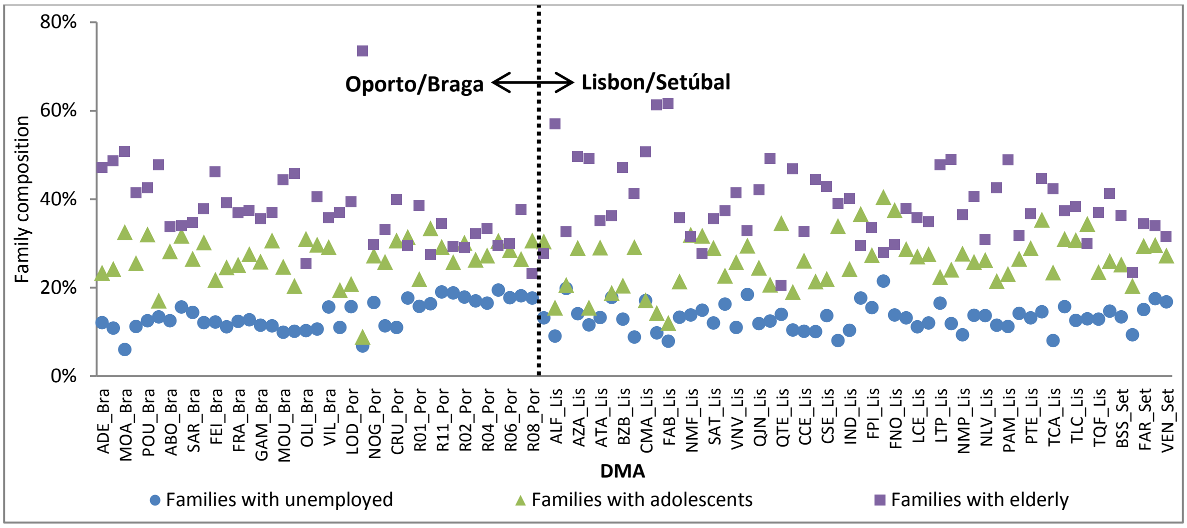


Figure 22 – Socio-demographic indexes concerning dwellings

In what concerns the Family category, Figure 23 shows that there are approximately 54% of small families (composed by 1-2 elements), 40% of medium families (composed by 3-4 elements) and only 6% of large families (composed by 5 or more elements). According to INE (2012) the majority of small families is composed by elderly people (>65 years). In the DMAs under analysis, approximately 38 % of the families have elderly people, although in Oporto´s district, this index has a median value of 32%, thereby suggesting younger families is this district. Families with adolescents (<15 years) represent nearly 30% of the total number of families. This average value decreases for approximately 20% in some of the DMAs belonging to the district of Lisbon.



(a)



(b)

Figure 23 – Socio-demographic indexes concerning families: (a) family size; (b) family composition

In respect to families with unemployed, an average of 14% of the total number of families have unemployed, although the median value of this index for DMAs belonging to the districts of Oporto rises for 17%. When comparing with the percentage of university graduates (Figure 24), in some DMAs belonging to the districts of Lisbon the low percentage of people with a university degree (7.5%) may be correlated with the unemployment rate.

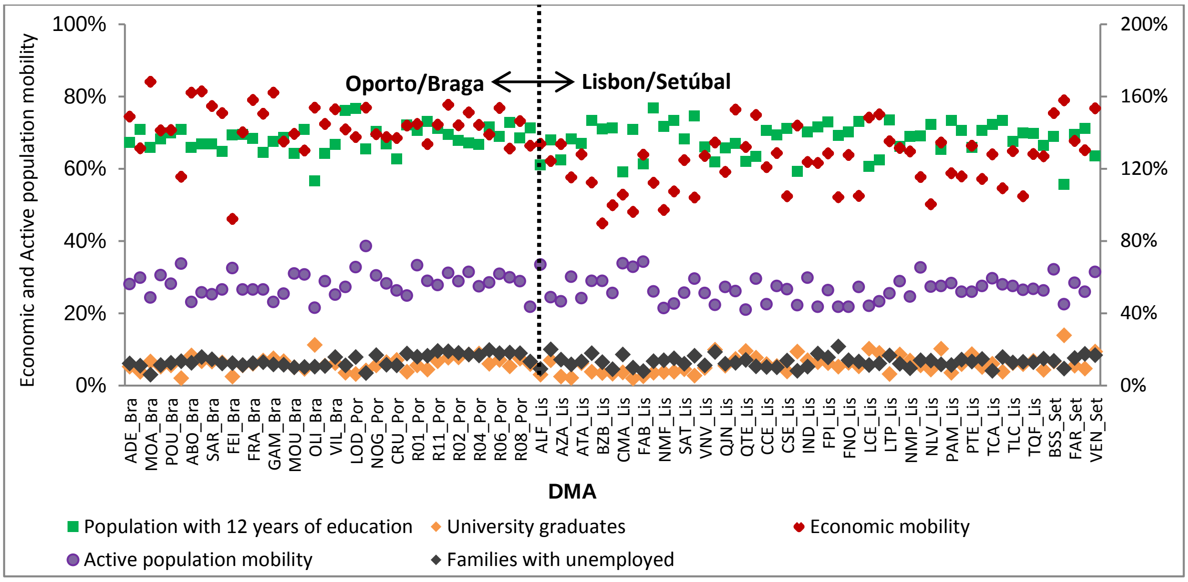


Figure 24 – Socio-demographic indexes concerning individuals and families with unemployed

For the remaining DMAs the average percentage of university graduates is 11%. The percentage of people with at least 12 years of education, which corresponds to the Portuguese mandatory education, is averagely 68%. In terms of economic mobility, which is related with the number workers employed in Trade, Tourism, and Services (the tertiary sector of activity), DMAs belonging to the districts of Lisbon and Setúbal have an average percentage of 64%. In DMAs belonging to the districts of Oporto and Braga, this index increases to an average of 72%. In the same group of DMAs, the number of residents that work or study outside their municipalities (Active population mobility) is approximately 57%, against 53% in DMAs belonging to the districts of Lisbon and Setúbal. High percentages of this index may suggest lower water consumption levels during the day time since residents spend less time at home. Along with the referred, since residents work or study outside their municipalities, the peak factor in the morning period may occurs earlier since residents in these DMAs probably have to get up earlier than residents who work or study in their municipalities (Loureiro, 2010).

As for the other indexes related to Individuals, Figure 25 shows that there are averagely 22% of residents with more than 65 years and 40% of inactive workers, with some DMAs where this index rises to nearly 60%. Inactive workers includes residents with less than 15 years, students, people working as domestics in their homes, retirees and other residents with no employment (Census 2011). High percentages of inactive workers suggest more water consumption during the day (Loureiro, 2010).

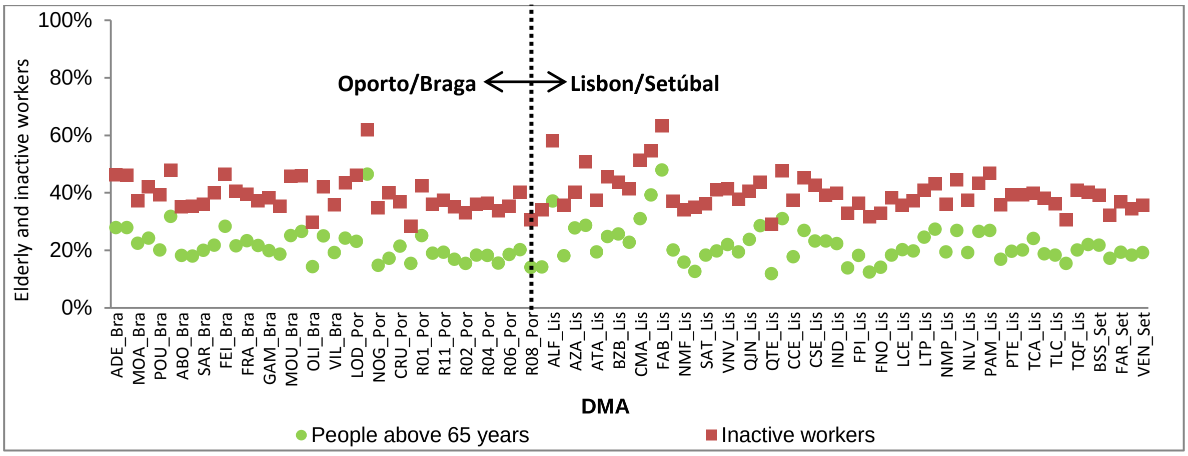


Figure 25 – Socio-demographic indexes concerning individuals: people above 65 years and inactive workers

4.4.2 Billing analysis

The methodology defined in 5.2 was applied to 101 DMAs that fulfilled the defined validation criteria (see Table 6). Billing indexes (Table 10) include information on domestic and non-domestic consumption and were calculated according to Loureiro (2010). Under the domestic consumption category, monthly consumption is analysed and divided in four groups according to Table 11. Groups are defined taking into consideration the IRAR Recommendation n. º 1/2009 (IRAR, 2009).

Table 10 – Categories and billing indexes used in the analysis

Category	Billing indexes	Billing indexes
Domestic	Average domestic consumption per client	[l/(cl.day)]
	Domestic consumption Domestic consumption within 1^{st} – 4^{th} level	[%]
Non-domestic	Commerce-Industry consumption Collective consumption Public consumption	

Table 11 – Domestic consumption level intervals (IRAR, 2009)

Domestic consumption level	Monthly consumption interval [m^3]
1^{st}	[0,5[
2^{nd}	[5,15[
3^{rd}	[15,25[
4^{th}	[25,∞[

In what concerns the consumption categories, as expected, the majority of the consumption is domestic, with an average of 87% in DMAs belonging to the districts of Oporto and Braga and 78% in DMAs belonging to the districts of Lisbon and Setúbal, as shown in Figure 26.

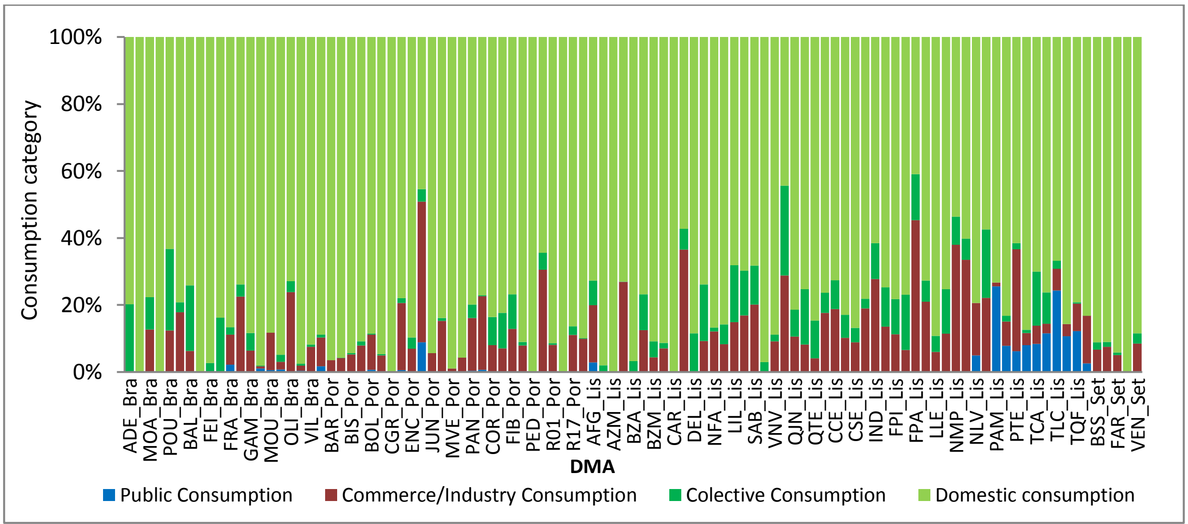


Figure 26 – Billing indexes concerning water consumption categories

Regardless of the DMA district, the second major category is commerce and industry with an average of 11% in all DMAs. Collective consumption accounts averagely for 5% of the total consumption and public consumption has an average of 2%, although some DMAs belonging to the district of Lisbon can reach a maximum of 26% of public consumption.

As for the domestic consumption level (Figure 27), the average billed domestic consumption per client is 266 l/(client·day) in DMAs belonging to the districts of Oporto and Braga and 308 l/(client·day) in DMAs belonging to the districts of Lisbon and Setúbal. In what concerns the average consumption level in all DMAs, the first and third level of consumption represent nearly 14% of the total consumption, each. In contrast, the second level represents averagely 51% in DMAs belonging to the districts of Oporto and Braga and 44% in DMAs belonging to the districts of Lisbon and Setúbal. The fourth level represents averagely 22% in DMAs belonging to the districts of Oporto and Braga and 30% in DMAs belonging to the districts of Lisbon and Setúbal. As expected, DMAs with higher consumptions in the fourth level have higher domestic consumptions per client.

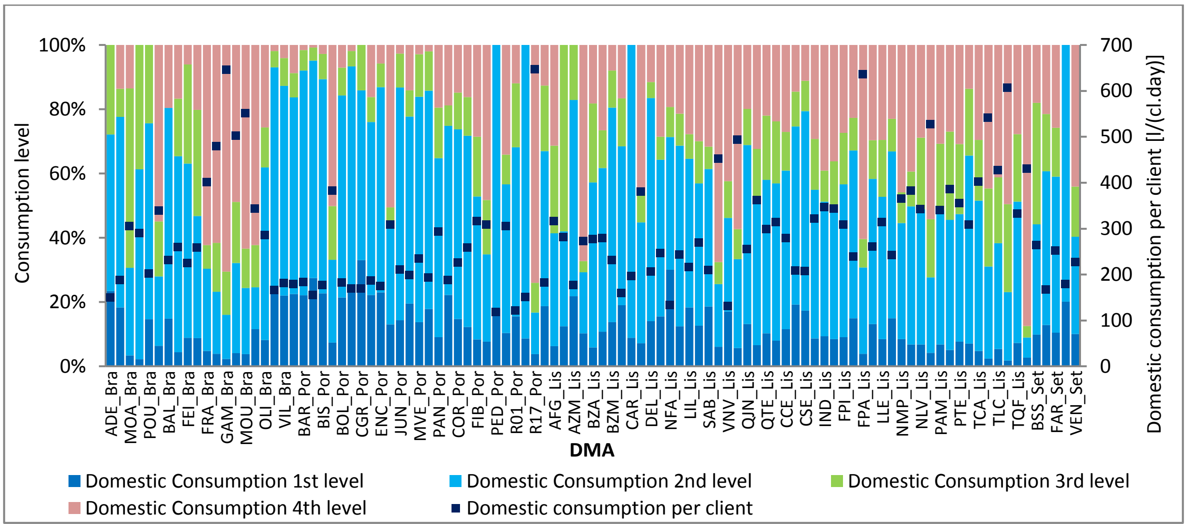


Figure 27 – Billing indexes concerning domestic water consumption

4.4.3 Infrastructure analysis

The methodology defined in 5.2 was applied to 96 DMAs that fulfilled the defined validation criteria (see Table 6). Infrastructure indexes (Table 12) include information on pipes and service connections and were calculated according to Loureiro 2010.

Table 12 – Categories and infrastructure indexes used in the analysis

Category	Sub-category	Infrastructure indexes	Units
Pipe	Pipe installation year	Average installation year (weighted by the pipes' length)	[year]
		Pipes installed within years]1926,1952],]1952,1976],]1976,2000],]2000,∞]	[%]
	Pipe diameter	Average diameter (weighted by the pipes' length)	[mm]

		Pipes with diameters within]0,110],]110,310] and]310,∞]	
	Pipe material	Stainless steel pipes rate Grey iron pipes rate Asbestos cement pipes rate Plastic pipes rate Unknown material pipes rate	[%]
Service connection	-	Billed Consumption per service connection	1. [m3/sc[2]]
		Service connection density	[number of sc /km]
		Average service connection length	[m]

In what concerns pipes diameters, Figure 28 shows that the average pipe diameter is 104 mm, typical of water distribution systems (Vidigal *et al.*, 2008). There are approximately 79% of pipes with diameters within]0,110] mm, 19% of pipes with diameters within]110.310] mm and 2% of pipes with diameters within more than 310 mm.

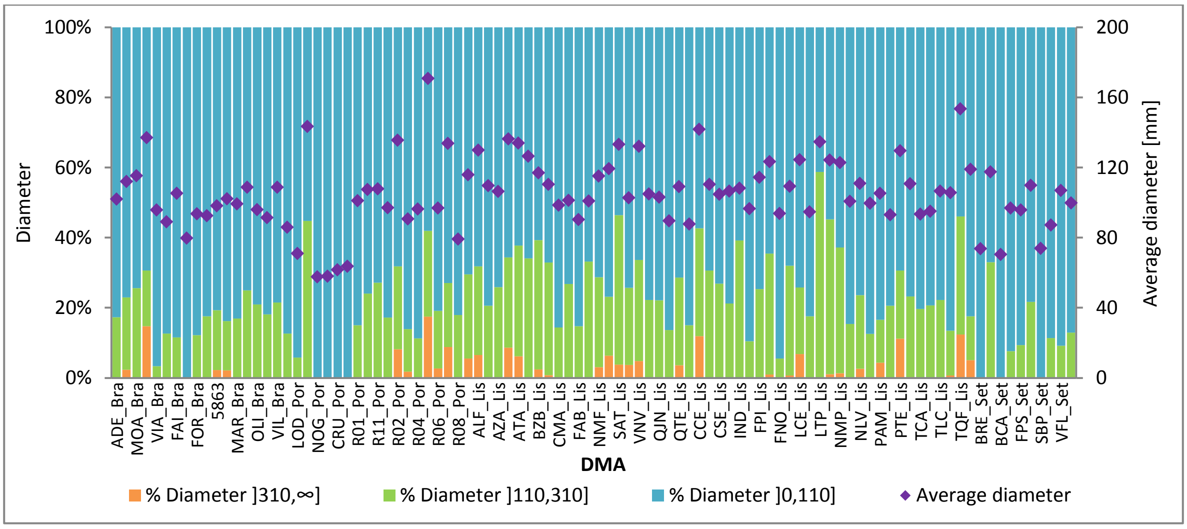


Figure 28 – Infrastructure indexes concerning diameters

In respect to the average installation year (Figure 29), DMAs belonging to the districts of Oporto and Braga are more recent with a median value of 2003 whereas DMAs belonging to the districts of Lisbon and Setúbal have pipes with a median year of 1993. Consequently, DMAs belonging to the districts of Oporto and Braga have most pipes built after 2000 (55%) and after 1976 (45%), whereas DMAs belonging to the districts of Lisbon and Setúbal have the majority of pipes built before 2000 (58%).

[2] sc=service connection

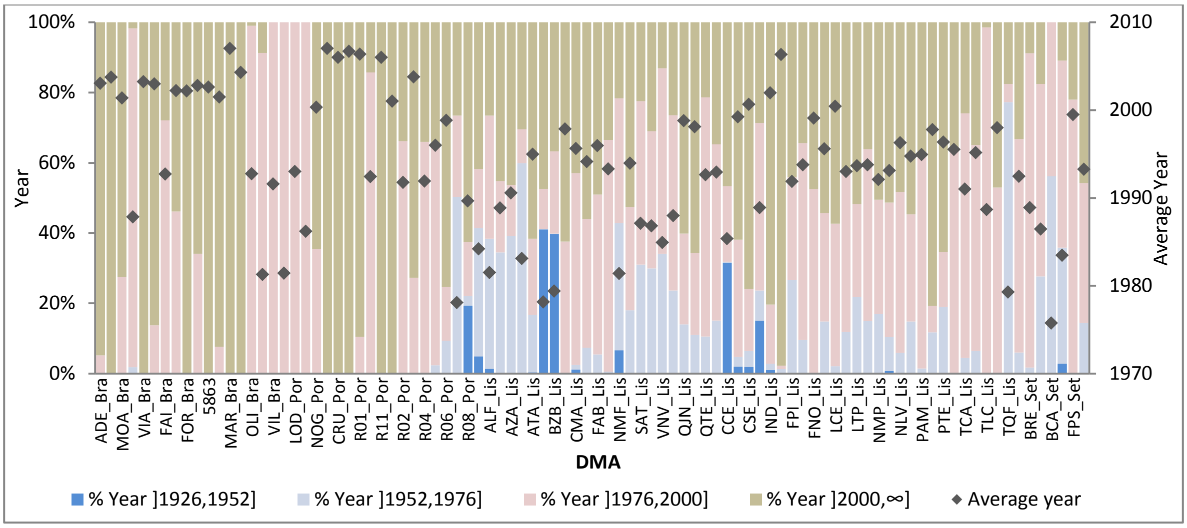


Figure 29 – Infrastructure indexes concerning pipes' age

Pipes' materials are usually related to pipe´s installation year. In particular, before the introduction of plastic pipes, pipes used to be built using metals such as grey iron and asbestos cement. As depicted in Figure 30 the presence of asbestos cement indicates older pipes whereas recent pipes are made with plastic, generally. DMAs belonging to the districts of Oporto and Braga are mostly in plastic (87%), with only 7% of iron pipes and 5% of asbestos cement pipes. As for DMAs belonging to the districts of Lisbon and Setúbal, 75% are plastic pipes, 21% are asbestos cement pipes and 4% are iron pipes. As water distribution networks become older, pipes degrade and experience frequent bursts and leaks, which lead to unacceptable levels of water losses and, too frequently, undesirable service disruptions. Other times, a deficient technical operation of the water network is the reason for a poor infrastructure condition (Palau *et al.*, 2012)

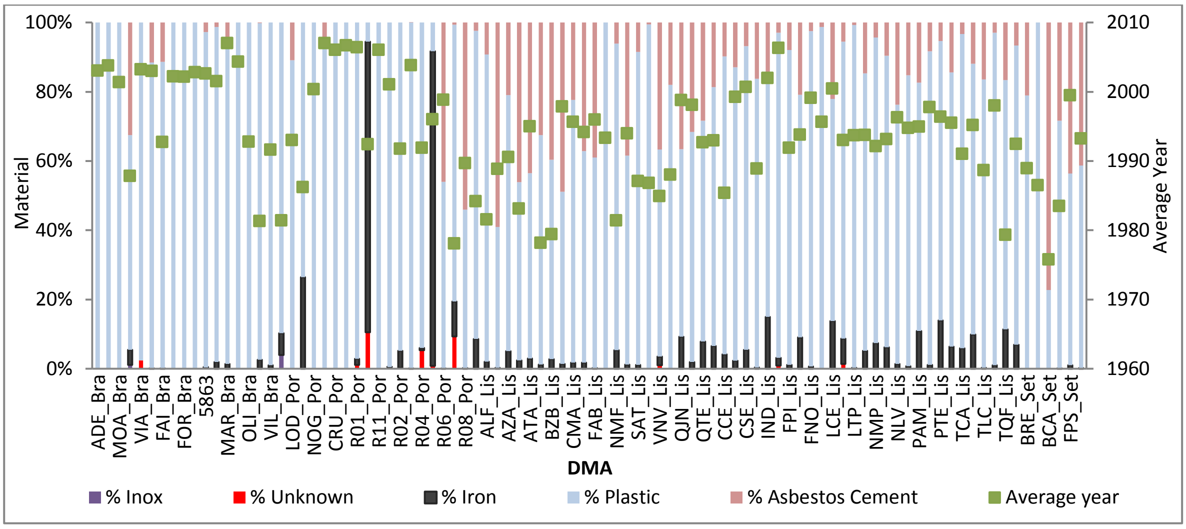


Figure 30 – Infrastructure indexes concerning pipes' material

In what concerns service connections, as shown in Figure 31, service connection density is very regular in all analysed DMAs with an average of 34 service connections per km, which corresponds to a typical value in urban systems.

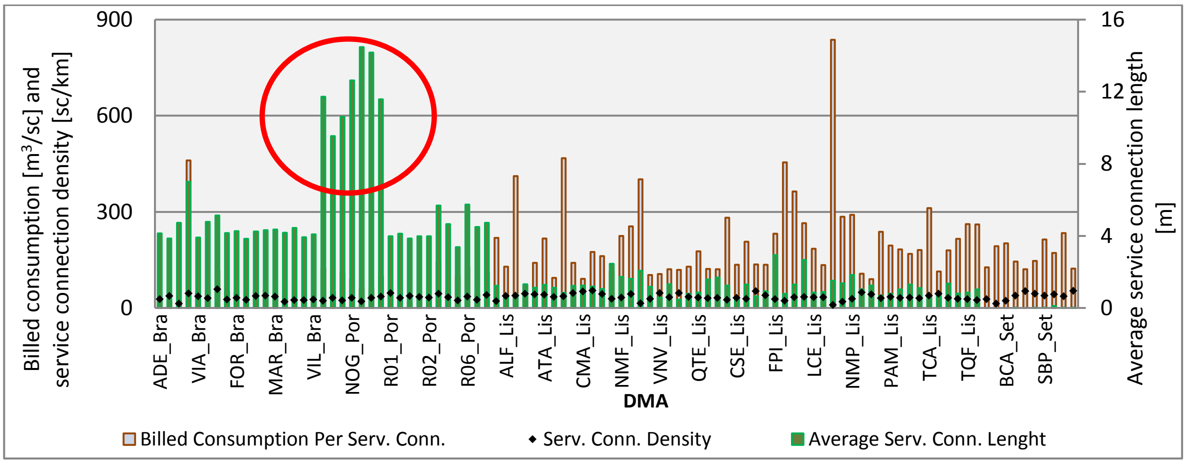
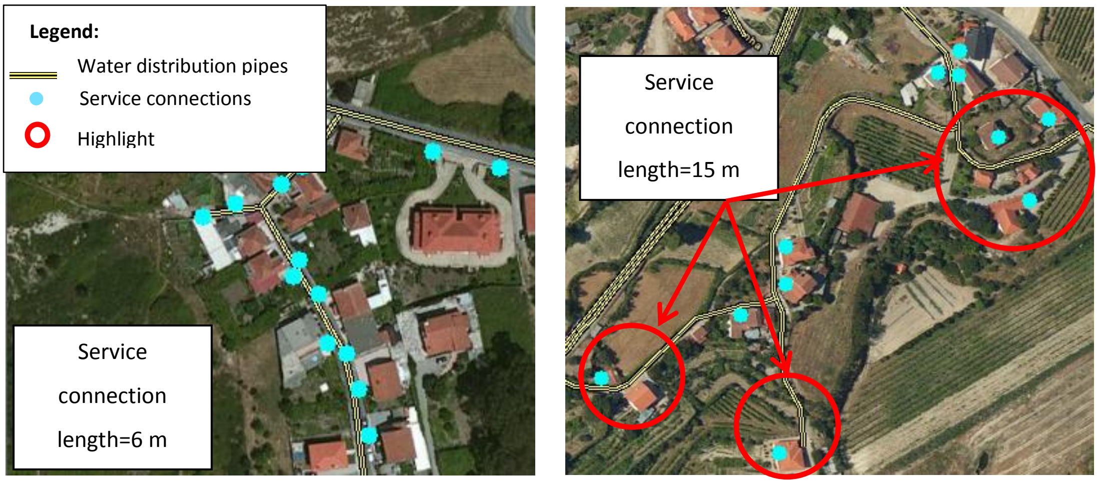


Figure 31 – Infrastructure indexes concerning service connections

In terms of service connection length, DMAs belonging to the districts of Oporto and Braga have an average service connection length of 6 m (Figure 32a), although some DMAs in Oporto´s district show higher values, with a maximum of 15 m, as highlighted in Figure 32b.



(a) (b)

Figure 32 – Service connection length for DMAs belonging to the districts of: (a) Oporto and Braga (b) Oporto (maximum values)

DMAs belonging to the districts of Lisbon have an average connection length of 1 m (Figure 33c) whereas DMAs belonging to the districts of Setúbal have an average connection length of 0.01 m (Figure 33 d).

In terms of billed consumption, DMAs belonging to the districts of Oporto and Braga have lower billed consumptions per service connection per year with an average of 72 m^3 per connection whereas DMAs belonging to the districts of Lisbon and Setúbal have more than twice the consumption, with an average consumption of 204 m^3 per connection. Since this index is calculated based on all types of consumption (*e.g.* domestic, non-domestic), this difference can be attributed particularly to the fact that DMAs belonging to the districts of Lisbon and Setúbal have a higher non-domestic consumption (21%), when comparing to the DMAs belonging to the districts of Oporto and Braga (13%). Additionally, the higher billed consumption in DMAs belonging to the district of Lisbon can be due to buildings with more than 1-2 floors, as presented in Figure 33 c. The higher consumption in DMAs belonging to the district of Setúbal can be associated with higher outdoor uses, such as pool filling, as highlighted in Figure 33d.

(c) (d)

Figure 33 – Possible reasons for differences in billed consumption in DMAs in the districts of (a) Lisbon; (b) Setúbal

4.5 Summary and conclusions

This chapter focused on the socio-demographic, billing and infrastructure analysis of DMA's based on the calculation of 49 indexes for the total analysis categories. Specific criteria were used to validate DMAs for analysis, and respectively 96, 101 and 96 DMAs were selected for the socio-demographic, billing and infrastructure analysis.

A geoprocessing tool that calculates socio-demographic indexes was improved based on an earlier tool developed by Loureiro (2010). Indexes were selected according to previous works where the socio-demographic, billing and infrastructure framework has proven to have a significant influence on water consumption (Donkor *et al.*, 2012; Eurostat, 2007; Grafton *et al.*, 2011; Loureiro, 2010)

The socio-demographic analysis was carried out using Census 2011 data and billing and infrastructure data provided by water utilities in 2010-2011. The use of data belonging to similar time steps significantly increases the results' accuracy.

In general, the socio-demographic analysis carried out for 96 DMAs shows that although belonging to different districts, all the analysed DMAs have a similar socio-demographic profile, since the socio-demographic indexes have similar values. This tendency applies to all socio-demographic indexes except Economic mobility (individuals that work in the tertiary sector) and Active population mobility (residents that study or work outside their municipalities), where DMAs belonging to the districts of Oporto and Braga have higher values. Approximately 57% of the residents living in DMAs belonging to the districts of Oporto and Braga work or study outside their municipalities. This fact can explain water consumption patterns with earlier peak factors in the morning period since residents probably have to get up earlier, as well as a lower consumption rate during the day in DMAs where this index is higher, since residents probably spend less time at home. In opposition, it was found that there are approximately 40% of inactive workers in all DMAs, and there are cases where this index rises for 60%, which can help explaining higher consumption rates during the day.

Concerning the billing analysis that was carried out for 101 DMAs, results also show a homogeneous pattern of domestic and non-domestic consumption among all DMAs. As expected, the subset of analysis the majority of the consumption is domestic (83%), followed by commerce and industry (11%), collective (5%) and public consumption (1%). Domestic consumption per client is approximately 15% higher in DMAs belonging to the districts of Lisbon and Setúbal (308 l/(client·day)) in comparison to DMAs belonging to the districts of Oporto and Braga (266 l/(client·day)). Overall, 50% of monthly consumption ranges from 5 to 15 m^3 (2^{nd} consumption level), though some DMAs belonging to the district of Braga have 60% of monthly consumption higher than 25 m^3 (4^{th} consumption level).

As for the infrastructure analysis that was developed using 96 DMAs, and contrarily to the socio-demographic and billing analysis, calculated indexes show significant differences between networks of DMAs in different districts. Aside from diameter and service connection density, the remaining indexes account for older networks in DMAs belonging to the districts of Lisbon and Setubal. The average diameter of all DMAs is 104 mm and the majority of diameters have less than 110 mm, which is typical in water distribution networks. Average service connection density is 34 connections per km. The presence of asbestos cement pipes was correlated with older networks, whereas the presence of plastic was correlated with more recent ones. On the one hand, DMAs belonging to the districts of Lisbon and Setúbal are typically from 1993 and have 20% of asbestos cement pipes. On the other hand, DMAs belonging to the districts of Oporto and Braga are typically from 2003 and have 87% of plastic pipes, against 5% of asbestos cement pipes.

The socio-demographic, billing and infrastructure analysis developed in this chapter is innovative when comparing to the methodology developed by Loureiro (2010), as presented in Table 13.

Table 13 – Comparison between the current research work and the one developed by Loureiro (2010) in terms of socio-demographic, billing and infrastructure analysis

	Current research study	Previous research works (Loureiro, 2010)
Number of analysed DMAs	100 (approximately)	22
Methods for socio-demographic indexes calculation in the geoprocessing tool (step 3)	Three alternatives	One alternative
Categories of socio-demographic indexes (step 3)	Buildings, dwellings, families and individuals	Buildings, families and individuals
Sub-categories of infrastructure indexes (step 3)	Pipe installation year, pipe diameter and pipe material	None

As a final remark, since recent census database includes more statistics on socio-demographic data, the socio-demographic indexes used in the present work should be reviewed and new indexes could be added. Multivariate analysis techniques will be carried out in Chapter 8, in order to determine from the 49 calculated indexes, which are the most relevant and the driving factors of water consumption for the present case study.

5. OUTLIER ANALYSIS

5.1 Introduction

The current chapter focuses on the description and application of a proposed methodology for the detection and removal of outliers from flow time series. This chapter is the Module 2 from the general methodology presented in section 3.2. The aim is to obtain a more coherent and robust time series to be able to calculate representative and reliable statistics of flow data, as well as to obtain consumption patterns that describe the typical behaviour of domestic clients of a particular type.

The methodology proposed for the detection and removal of the outliers of the raw data series is a five-step procedure and includes flow data collection, data validation, data normalization, outlier detection and finally data cleaning. An algorithm for outlier detection developed by Loureiro *et al.* (2013) was used and tested in the present work for a large number of data series.

This chapter includes the methodology description, the presentation of the main results of the outlier detection and the summary and conclusions.

5.2 Methodology

The methodology used for outlier detection of the flow data series consists of five steps as depicted in Figure 34.

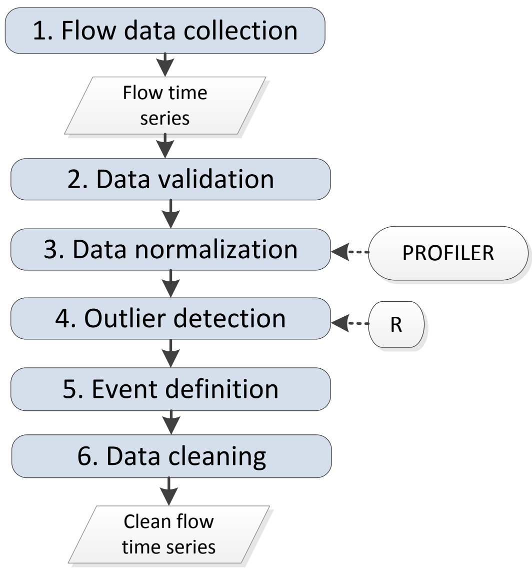


Figure 34 – Methodology used for outliers' removal

The first step refers to "Flow data collection". Flow data series are typically provided by the SCADA systems of water utilities. For the outlier detection, water utilities were asked to provide flow data series that fulfilled the criteria defined in Table 14.

Table 14 – Criteria used for DMA validation (Outlier analysis)

Analysis type	Criteria
Outlier analysis	1) Flow data series with a large period of records from 2010 or 2011 (preferably one year) 2) Flow data series with small time intervals (preferably, a regular time step of 15 minutes)

Flow data from DMAs can be directly measured by one meter at the entrance of the DMA (direct measurement) or indirectly if DMA flow data is obtained by calculations between meters (indirect measurement). The two types of measurements referred are schematically presented in Figure 35. Appendix B shows the list of DMA analysed herein as well as the type of measurement carried out, along eith other characteristics.

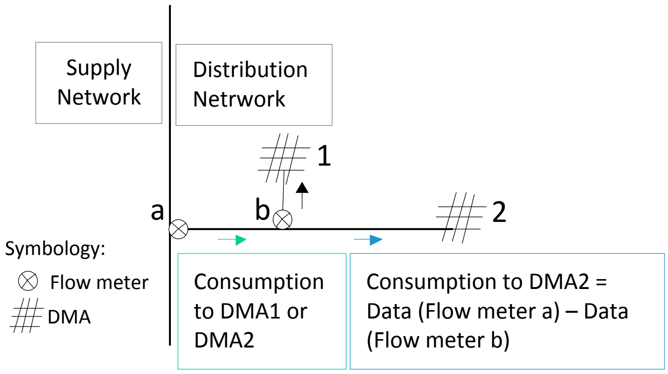


Figure 35 – Examples of direct (1) and indirect (2) flow data measurement

The second step consists of "Data validation" which requires a pre-processing step. The pre-processing involves the removal of repeated values, the correction of inversions in the time series and the conversion of units of flow data, for instance.

In the third step, "Data normalization", the Profiler application developed by (Loureiro, 2010) is used to normalize the data series. This process consists of regularizing the time step of every data series into a fixed value to simplify the further analysis. In case some gaps are observed in the series, PROFILER completes the series by interpolating observations; this is possible as long as the gap does not exceed a certain size (one hour is the predefined value). The normalization is different for instantaneous and for average values of flow data (Loureiro, 2010). As such, the user should specify the data type accordingly.

In the fourth step, "Outlier detection" is carried out by using the "Symmetric method" developed by Loureiro *et al.* (2013). The method consists of an algorithm written in R programming language that uses two robust statistics: these are the median (MED) and the Qn which is the robust standard deviation of the observations based on the Qn scale (Rousseeuw and Croux, 1993). To detect outlying observations to be excluded from the flow data series outliers are detected according to:

$$\mathbf{OTL} \geq \mathbf{MED} + \mathbf{c} \cdot \mathbf{Qn} \vee \mathbf{OTL} \leq \mathbf{MED} - \mathbf{c} \cdot \mathbf{Qn} \qquad (2)$$

in which:

$\boldsymbol{OTL}$:	Outlier value in the data series	[m^3/h]
$\boldsymbol{MED}$:	Median of a set of previous observations defined by the user	[m^3/h]
$\boldsymbol{c}$:	Threshold value to be defined by the user (with c > 0)	[-]
$\boldsymbol{Qn}$:	Robust standard deviation of the observations based on the Qn scale	[m^3/h]

The algorithm is set to calculate medians as long as there are at least four previous days of the same type (*i.e.* working days, Saturdays or Sundays and bank holidays) with non-blank observations in the instant where the median is to be calculated. In addition, the algorithm eliminates every group of four or more observations whose variance is less than 0.001 since they represent constant flow periods which are atypical in domestic consumption (Loureiro, 2010) and may be related with equipment malfunction. Before running the algorithm, the user should specify the number of previous observations that are used to calculate the median value and a threshold value, c, which influences the probability of an observation being detected as an outlier. In the current study, a window of twenty observations was defined as it is considered that it represents a good compromise between having a representative number of observations and avoiding outlier detection to be influenced by seasonal effects of domestic water consumption (Loureiro *et al.*, 2013). As for the threshold value, it was adopted c=1.58 based on Loureiro *et al.* (2013). Assuming that the flow data follows a normal distribution, this value is associated to an outliers' region (or rejection region) of 5.7% as shown in Figure 36.

As consumption behaviour varies with the day of the week, differing from working days to weekends (Arbués *et al.*, 2003; Palau *et al.*, 2012), the median is calculated for different types of the days, namely for "working days" and "Saturdays" and "Sundays and bank holidays".

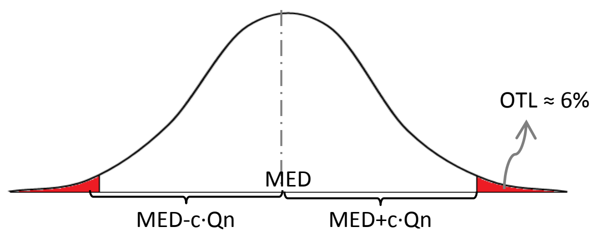


Figure 36 – Normal distribution and outlier detection scheme

After the outlier detection, the next step is the transformation of the outliers in events – step 5. For defining as event the following assumption is considered:

In this study, an *event* is defined when, in every list of five observations, at least three were classified as outliers.

This definition was applied after analysing three DMA with different characteristics and means that detected events have at least 45 minutes duration. However, events can be, for instances, a sudden flow increase or decrease with less than 45 minutes duration. In these cases, an additional condition was defined for a preliminary classification, as presented below:

If a certain *observation* was detected as outlier and there was a period of constant flow before, after or both, or there was a lack of observations before, after or both, then this outlier event is defined as “Sudden variation” since it is probably due to meter errors.

After the events’ definition, a visual analysis was carried out in the three data series in order to exclude events that did not have an abnormal behaviour. Isolated outlying events related with behaviour change occurring before or after bank holidays were also excluded from the outliers’ event list.

A preliminary classification was also carried out based on the events detected for three DMAs with different sizes (see Table 15). This classification depends on basic statistics of the events, such as: average event duration and average flow of the event were calculated.

In this study, the *average flow of the event* is defined as the average value that results from subtracting the flow observations to the median for every instant of the event. Negative values are related with flow decreases.

Table 15 – DMA sizes and characteristics used for outliers' preliminary analysis

DMA	Domestic clients	Average flow [m^3/h]	Minimum flow [m^3/h]
Small	400	3.00	0.50
Medium	2500	50.00	22.00
Large	5000	100.00	30.00

The criteria defined for the preliminary event classification is defined in Table 16. Based on the events detected for the three DMAs with different sizes, two event categories were defined:

- **Data acquisition problems:** this category aims at including events that are probably related with problems in data acquisition from the SCADA systems. It includes "Sudden variation" events, previously defined, and "Constant flow" events, which refer to periods with constant flow that are removed by the "Symmetric method".

- **Pipe bursts, leaks and abnormal consumptions:** this category aims at including events that are probably related with pipe bursts, leaks and abnormal consumptions. This category includes three event types: "Flow increase", "Flow decrease and "Long duration".

  "Flow increase" or "Flow decrease" refer to events where the flow observations are above or below the median value, respectively. Flow increases can be related with outdoor activities, such as pool filling and irrigation (Arbués *et al.*, 2003). Flow decreases can be due to network operations, such as valve maneuvering or connection to another DMA (Loureiro *et al.*, 2013). The flow limit from which an event is considered to be an increase or a decrease was defined according to the average value for irrigating a medium size garden that is 9.0 m^3/h approximately (Loureiro, 2010). The duration for flow increases/decreases was established by the analysis of the list of events avoiding the inclusion of events with a long duration, (*i.e.* higher than 3 hours). For those, a "Long Duration" event type was created, and it includes all flow increases with more than three hour duration. This event type is supposed to include events like pipe bursts.

The different types of events were separated in case of happening exclusively in the night period with the aim of understanding the most common event types in the night period. Taking into account the period of lower flow rates, the night period was defined from 00h00 to 06h00. Figure 37 shows examples of different types of detected events.

Table 16 – Criteria used for classification of events with 45 minutes duration or more

Event Category	Event type	Duration [h]	Average flow [m^3/h]
Data acquisition problems	Sudden variation	0.25	any
	Constant flow	> 1	Variance < 0.001
Pipe bursts, leaks and abnormal consumptions	Flow increase	≤ 3	≥ 9
	Flow decrease	≤ 3	≤ 9
	Long Duration	> 3	> 0

Moreover, water utilities provided work orders which consist of the list of interventions in the system detected by the utility during the period of analysis. However, these data files had very few records and there was a lack of standardization in terms of the classification of the events in categories. For these reasons, very few of the detected events could be correlated with the working orders provided.

Finally, the sixth step of outlier detection is “Data cleaning”, where the detected events of the previous step are deleted from the time series. A comparison is also made using both raw and cleaned data in order to evaluate the differences in basic statistics (average, minimum and maximum flow).

The cleaned series that result from this step are the input for the following chapter – water consumption analysis.

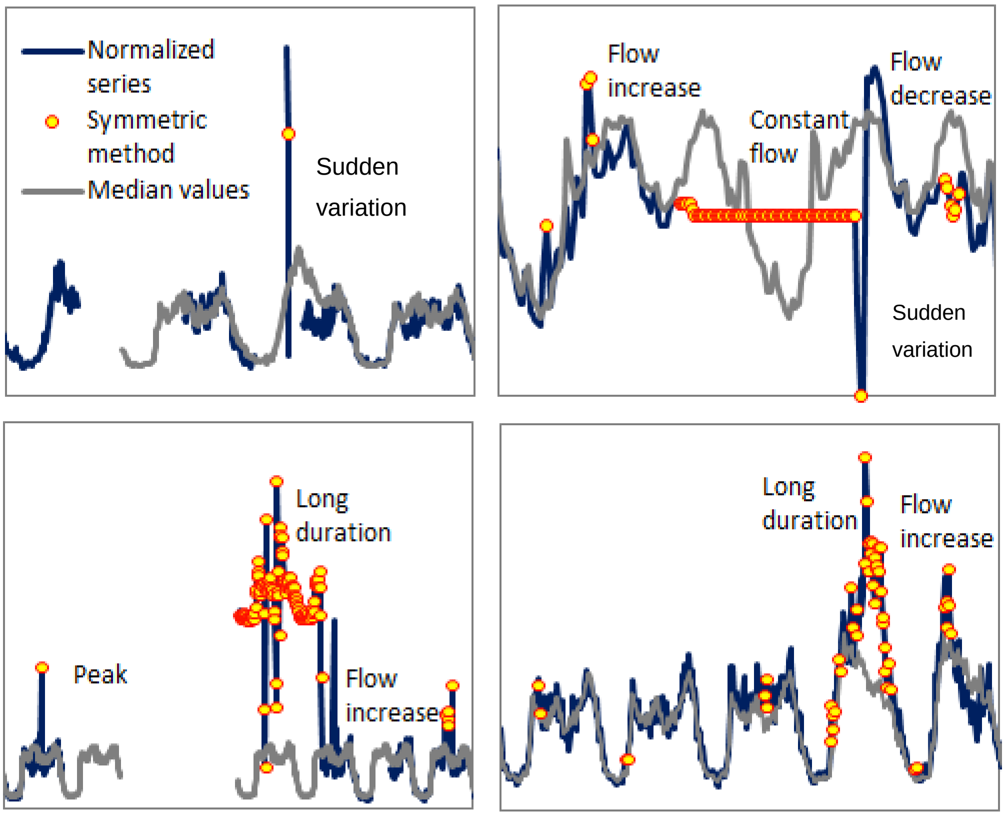


Figure 37 – Examples of different types of events.

5.3 Case-studies description

From the total of 150 DMAs with billing and infrastructure data (see 4.3), only 37 of them had flow data series that fulfilled the criteria defined in Table 17. The reason for the reduction of data availability may be due to the following:

- Some water sectors are referred as DMAs are not being metered yet, and therefore there is no data available.
- Water utilities store large amounts of metered data and sometimes the quality of such data is not assured.
- This study aims at profiling domestic water consumption. Only DMAs with a majority of domestic consumption are selected for analysis.

DMAs with flow data series are located in the districts of Oporto, Braga, Lisbon and Setúbal (see district location in Figure 19, section 4.3). DMAs' general characteristics are presented in Table 17.

Table 17 – General characteristics of DMAs selected for outlier analysis (37 DMAs)

Characteristic	Interval	Average value	Median value
Average diameter [mm]	69 – 145	97	97
Network length [km]	4.0 – 95.0	28	20
N.º service connections	158 – 3698	1110	717
N.º of clients	164 – 5185	1555	892
N.º of domestic clients	121 – 4514	1430	803
N.º of inhabitants	46 – 12778	2981	2038

5.4 Results

5.4.1 Data collection, validation and normalization

As referred, flow data series were collected from 37 DMA with a time step of 15 minutes. The majority of the series have a civil year of history, as presented in Appendix 1, covering the summer and winter seasons. Since the raw data series have outliers, the average, minimum and maximum flow statistics were not calculated to characterize these series. Instead, the median, 10th percentile and 90th percentile statistics were calculated for each series since these are less affected by the presence of outliers as depicted in Figure 38.

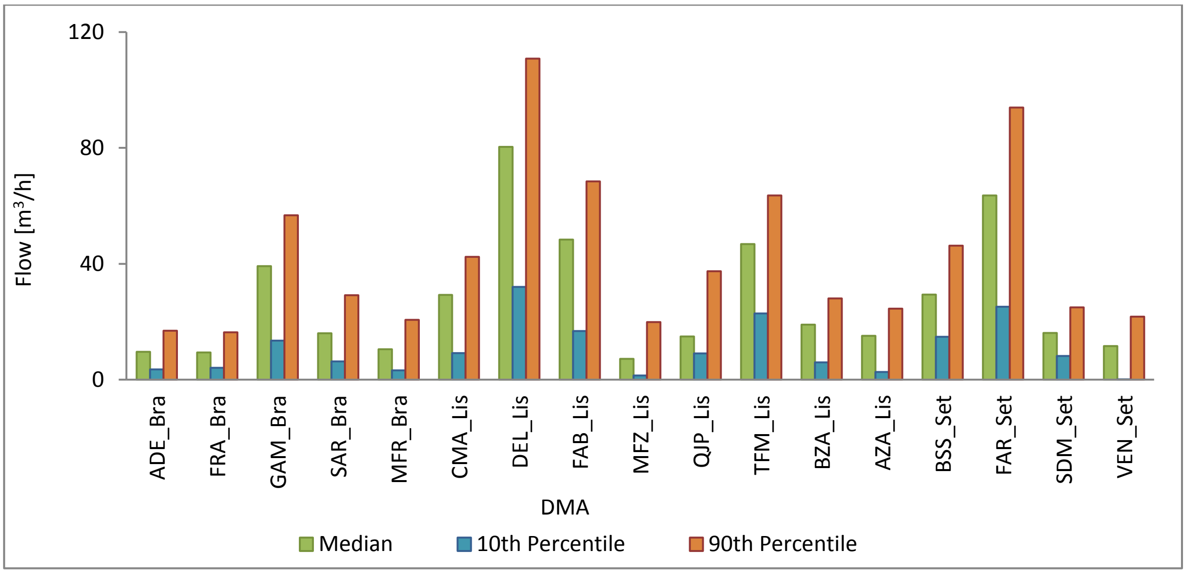


Figure 38 – Median and 10th and 90th Percentiles statistics of flow data series

The figure above suggests that the collected flow data have a good diversity of flow ranges, since medians vary from 1 to 80 m^3/h. Information on the total of domestic clients and service connection and density of connections was also collected and is presented (see Figure 39).

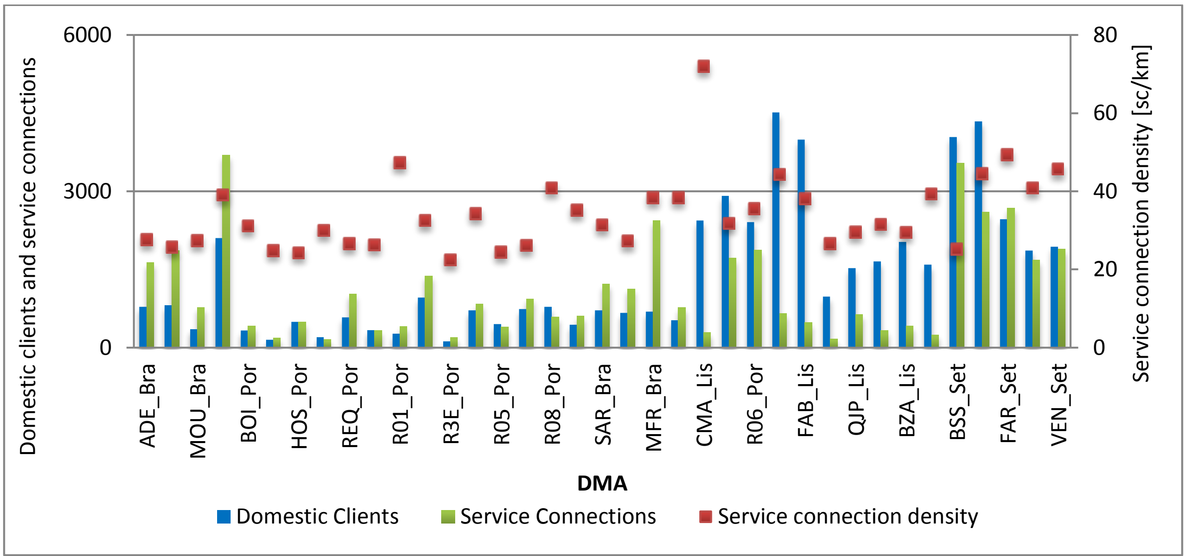


Figure 39 – Total of domestic clients and service connections.

Concerning domestic clients, the analysed DMAs have between 120 to 4500 domestic clients and 35 to 13000 inhabitants according to Census 2011 database. As for the number of service connections, it varies from 160 to 3700 and service connection density ranges from 22 to 70 service connections/km, thereby suggesting once more the presence of diversity in this group of DMA. For this extensive set of data, a cluster analysis has been carried out using the median, the 10th and 90th percentile values of the flow data and the number of domestic clients in order to update the intervals for the definition of DMA sizes previously presented in section 5.2. The intervals for DMA size classification are presented in Table 18.

Table 18 – Updated DMA sizes and characteristics used for outliers' analysis

DMA	Domestic clients	Median flow [m^3/h]	10th percentile flow [m^3/h]
Small	550	6.00	2.00
Medium	2000	32.00	13.00
Large	4200	50.00	20.00

In Portugal, Loureiro (2010) has worked with 22 DMA with median flows varying from 25 to 114 m^3/h and inhabitants ranging from 1956 to 13386. In Spain, Herrera *et al.* (2010) has worked with flow series varying from 5 to 50 m^3/h and 5000 inhabitants, approximately.

Despite the presented 37 DMAs appearing to be suitable for the present study, some problems were detected when analysing the flow data series. Errors, such as sudden peaks of the flow series or flat readings during large time intervals, are presented in Figure 40. These have been reported to the water utilities that agreed that data acquisition problems had occurred. Some of the reported flat readings were due to electricity current peaks that block the SCADA system.

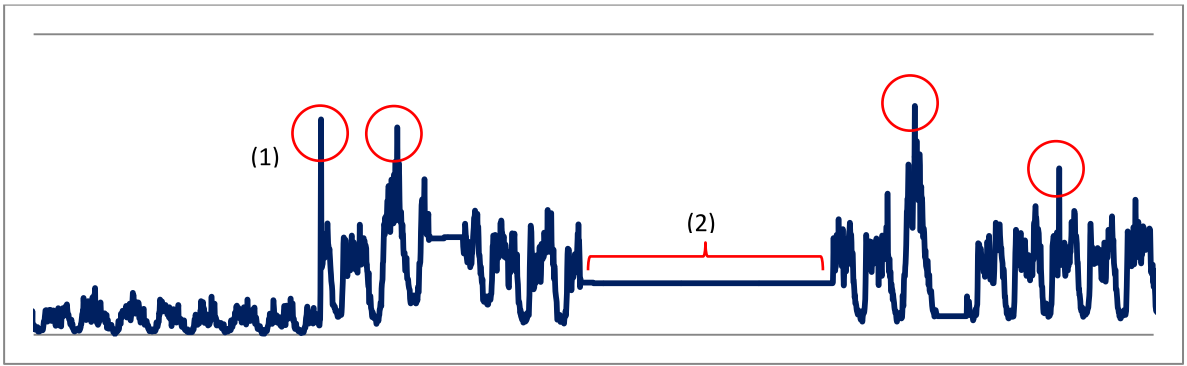


Figure 40 – Examples of data acquisition problems: Sudden increasement (1) and flat readings (2).

The referred issues reduced the number DMAs available for analysis from 37 to 17 DMAs. Notwithstanding the reduction of the number of flow series available for analysis, this new validated subset still presents a good range of variability, as presented in Table 19:

Table 19 – Validated and normalized flow data series characteristics

Characteristic	Interval
Median flow range [m^3/h]	7 – 80
N.º of service connections	170 – 3700
Density of connections [sc[3]/km]	26-70
N.º of domestic clients	700 – 4500
N.º inhabitants	230 – 13000

For the validated subset, data availability was calculated (ratio between number of time steps with suitable data and the total number of time steps in the analysis period) as presented in Table 20 along with the series temporal dimension. For the majority of DMAs, data availability ranges from 80% to 100%, which indicates a reduced number of periods without records, except for VEN_Set DMA, which had a long period without records

[3] sc=service connections

(4 months). Periods without records are usually distributed along the series, with special incidence in the months of February, June and July.

For the normalization of the flow data series, a fixed time step of 15 minutes was defined and the maximum limit for the interpolation was set to one hour.

Table 20 – Flow series parameters

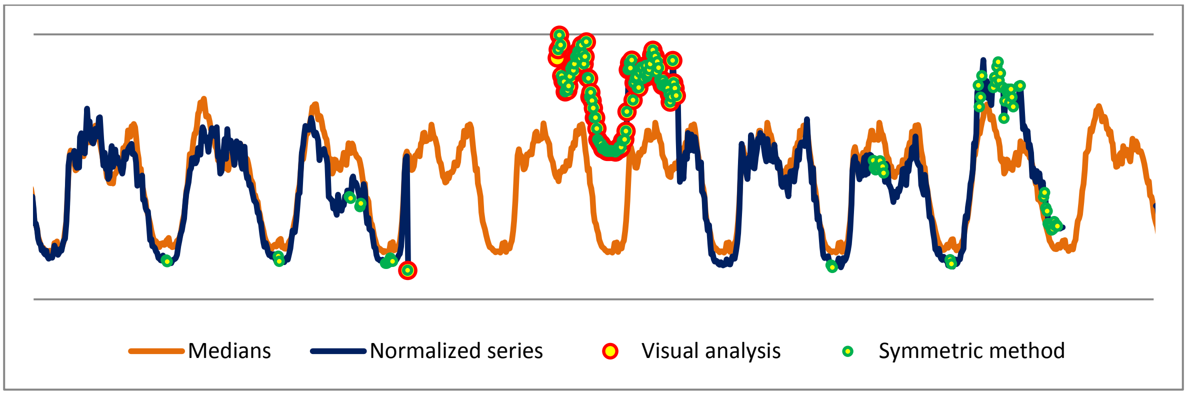
DMA N°.	DMA ID	Time start	Time end	N°. of Days	N°. of time steps of available data	Data availability ratio
1	ADE_Bra	01-01-2011 00:30	31-12-2011 23:45	365	31183	89.0%
2	FRA_Bra	01-01-2010 00:30	31-12-2010 23:45	365	27068	77.3%
3	GAM_Bra	01-01-2010 00:30	31-12-2010 23:45	365	27540	78.6%
4	SAR_Bra	01-01-2010 00:30	31-12-2010 23:45	365	26917	76.8%
5	MFR_Bra	01-01-2011 00:30	31-12-2011 23:45	365	28546	81.5%
6	CMA_Lis	01-01-2011 00:00	31-12-2011 23:45	365	34683	99.0%
7	DEL_Lis	01-01-2011 00:00	31-12-2011 23:45	365	34768	99.2%
8	FAB_Lis	01-01-2011 00:00	31-12-2011 23:45	365	34542	98.6%
9	MFZ_Lis	14-06-2011 08:00	31-12-2011 23:45	201	18281	94.9%
10	QJP_Lis	01-07-2011 00:30	31-12-2011 23:45	184	17537	99.3%
11	TFM_lis	01-05-2011 00:15	31-12-2011 23:45	245	23386	99.4%
12	BZA_Lis	01-10-2011 00:30	31-03-2012 23:45	183	17543	99.9%
13	AZA_lis	01-12-2011 00:30	31-05-2012 23:45	183	17543	99.9%
14	BSS_Lis	01-01-2011 00:15	31-12-2011 23:45	365	28400	81.3%
15	FAR_Set	01-01-2011 00:15	31-12-2011 23:45	365	33873	96.9%
16	SDM_Set	01-01-2011 00:15	31-12-2011 23:45	365	33138	94.7%
17	VEN_Set	01-01-2011 00:15	31-12-2011 23:45	365	20028	57.3%

5.4.2 Outlier detection

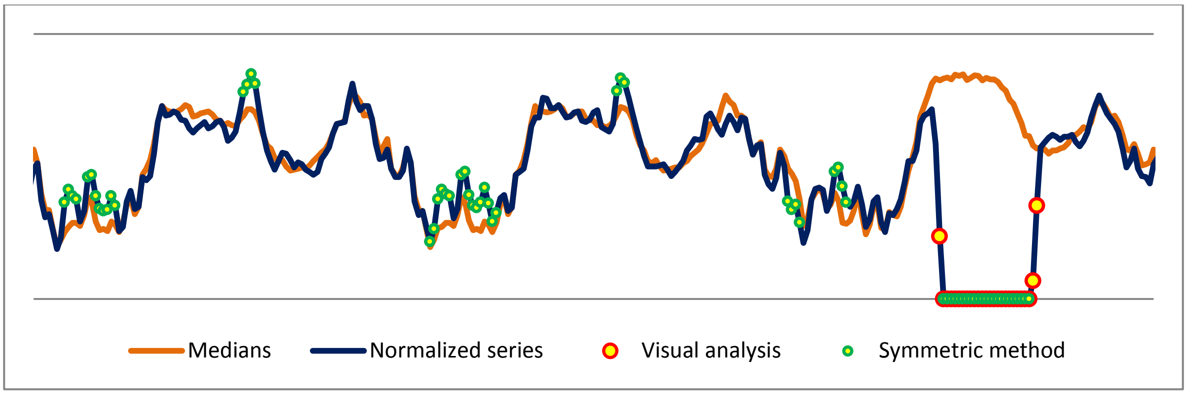
Outlying events were detected for the 17 DMA by using the "Symmetric method" developed by Loureiro *et al.* (2013). Afterwards, a visual analysis of the detected events was carried out, where the type of detected events was identified. It was found that only an average of 11% of the detected events should to be removed from the series, since most were associated to changes of consumption behaviour due to school holidays, bank holidays or seasonal effects. Consequently, it is recommended:

- A better calibration of the threshold value with regards to the flow series behaviour.
- An improvement of the work orders provided by the water utilities to compare real network events with detected outliers.

Concerning the well-detected events, in the one hand, the threshold value adopted (c=1.58) seems to be correctly detecting the beginning and the end of some events such as long events, as shown by Figure 41a when visual analysis is coincident with symmetric method. On the other hand, the same figure shows that the adopted threshold value generates events which are not outliers, but instead represent sudden behaviour changes. Moreover, there are still events where its beginning and end detection needs improvement, as shown for example in Figure 41b, where flat readings are detected but not the immediate observations before and after it.



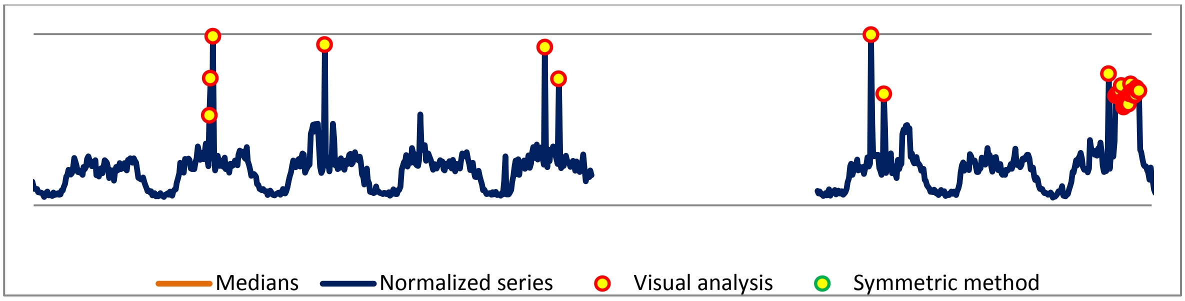
(a)



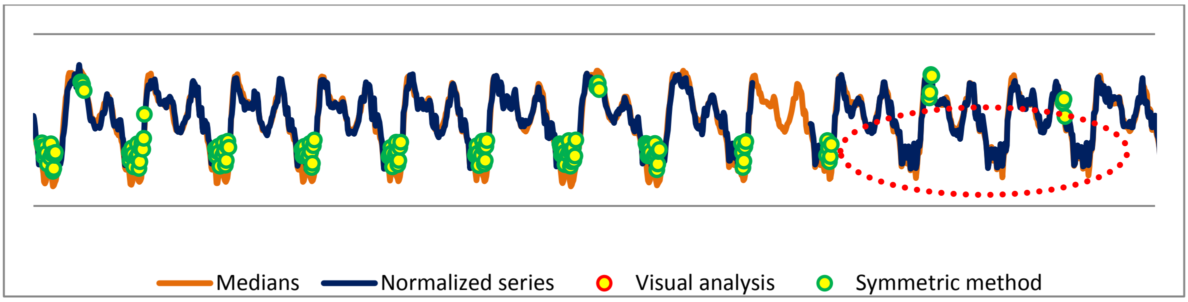
(b)

Figure 41 – Example of well detected events and behaviour changes: (a) long duration; (b) constant flow

In addition, since medians are only calculated after 20 observations this fact restricts the identification of events when medians are not available, which is mainly at the beginning of the data series (see Figure 42a). Furthermore, the fact that medians are calculated based on a window with 20 previous observations means that when a behaviour change occurs during an extensive period, medians are adapted to it. The problem is the window size and would occur with any other statistical variable. This means that previously detected outliers are considered as normal behaviour if they persist for 20 observations, as highlighted in Figure 42b.



(a)



(b)

Figure 42 – Window size related issues: (a) no outlier detection in the beginning of the series; (b) adaptation after 20 observations of behaviour change

5.4.3 Event definition

In what concerns the type of events detected, the classification based on the set of three DMAs with different sizes previously presented in the methodology was followed with two modifications in the limit of the average flow for "Flow increase" and "Flow decrease". These modifications were made so that the criteria for event classification could be adopted for all 17 DMAs with flow series that were analysed in this step.

It was found that, although some "Flow increase" and "Flow decrease" events did not have an average flow of $9m^3/h$, they were still not representative of domestic consumptions. Hence, the average flow criteria for these event types were updated and defined as positive (> 0) or negative (< 0) [4], respectively. Table 21 shows the updated criteria for event classification.

[4] Negative average flows (<0) refer to flows below the median value (see 5.2)

Table 21 – Updated criteria used for event classification

Category type	Event type	Duration [h]	Average flow [m^3/h]
Data acquisition problems	Sudden variation	0.25	any
	Constant flow	> 0	Variance < 0.001
Pipe bursts, leaks and abnormal consumptions	Flow increase	≤ 3	> 0
	Flow decrease	≤ 3	< 0
	Long Duration	> 3	> 0

Following the updated criteria, a total of 512 events were detected, the majority of them belonging to the category of "Sudden variation" and "Flow increase", as presented in Table 22. It was also found that most events take place during the day period, and that there are no "Long Duration" events happening during the night period (00h00-06h00). Most events detected at night are "Sudden variation" and "Constant flow" events.

Table 22 – Total of detected events according to the period of the day

Event	Sudden variation		Flow increase		Flow Decrease		Long duration		Constant flow	
Period	Day	Night	Day	Night	Day	Night	Day	Night	Day	Night
N. events	221	15	132	5	19	3	45	0	57	15
Total	236		137		22		45		72	

A further analysis of the detected events concerning its average flows and durations is depicted in Figure 43. As presented, very few events have a duration that exceeds 12 hours and average event flows higher than 60 m^3/h. In terms of differences between the day and night period, events happening in the day period have higher average event duration and average event flows.

Concerning the two defined category types:

1) Data acquisition problems: "Sudden variation" events that can be associated with registry errors due to improper calibration of flow meters have a fixed duration of 0.25 hours and an average flow that ranges from -20 to 40 m^3/h, approximately. This wide range of variability highlights the importance of monitoring the data that comes from the SCADA systems.

"Constant flow" events can be linked to failures in the SCADA systems that may be due to electricity current peaks. The analysis suggests events of this type have higher durations in the day than in the night period. The high durations of "Constant flow" events consists, once again, of alerts to the water utilities that record DMA flows.

2) Pipe bursts, leaks and abnormal consumptions: "Flow increase" and "Long duration" events are more probably associated with pipe bursts and abnormal consumptions, whereas "Flow decrease" events can be more linked with pipe leaks. Leaks are usually associated with low flow rates. Therefore, it is understandable that "Flow increase" and "Long duration" events have higher average event duration and flow than "Flow decrease" events.

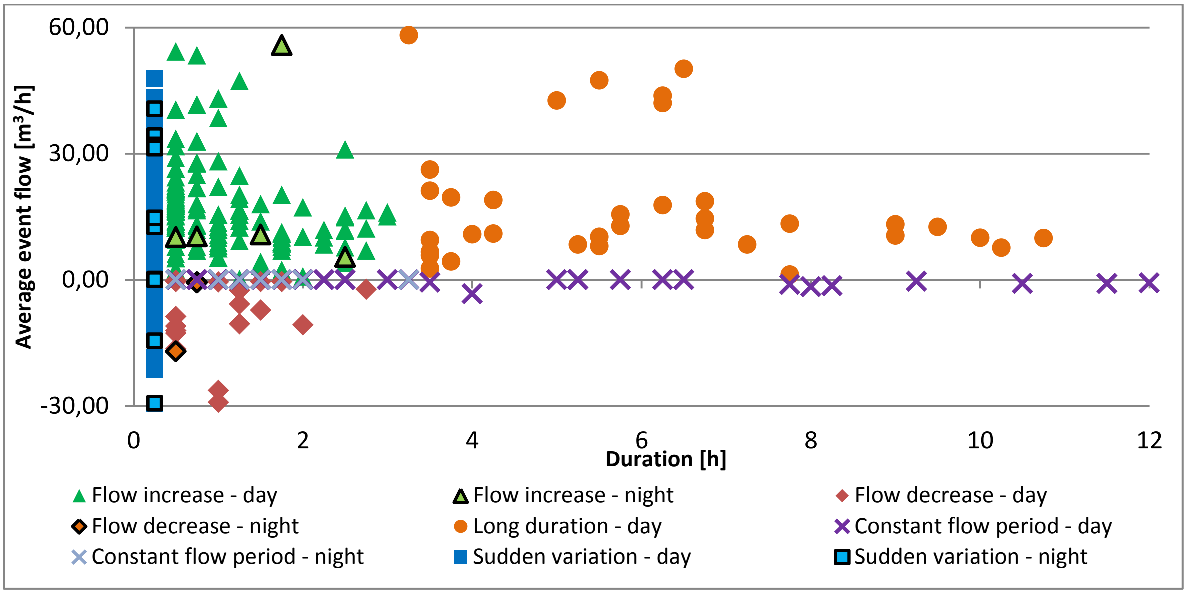


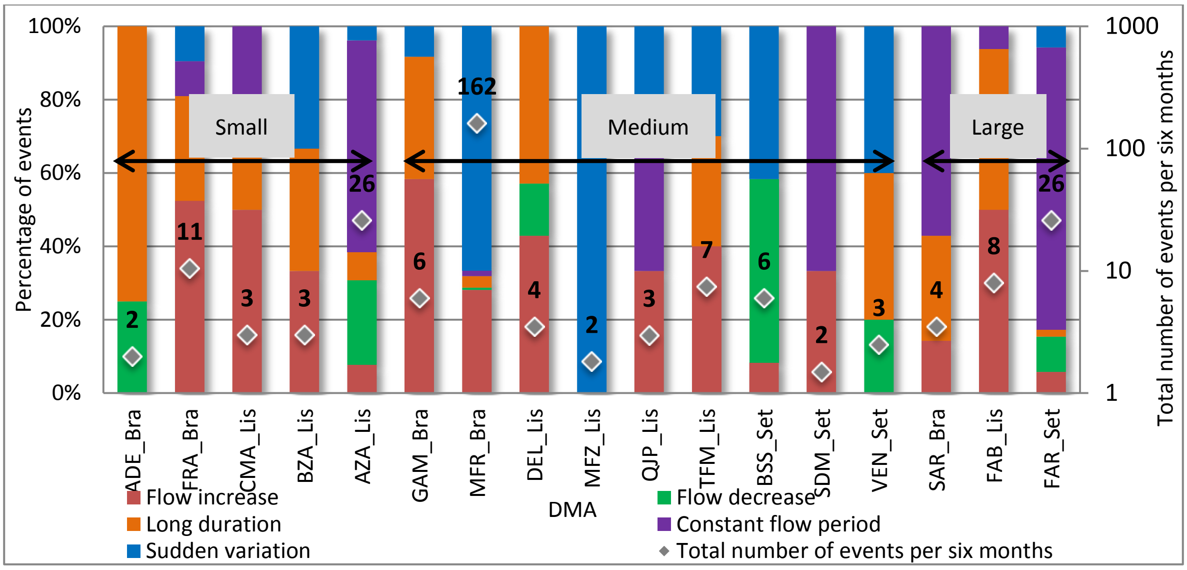
Figure 43 – Detailed information on events' characteristics: (a) general distribution of events in terms of average event duration and average event flow; (b) average event duration; (c) average event flow

Cluster analyses have been carried out in order to identify relations between the occurrence of outlying events and DMA characteristics. For instance, a cluster analysis combining DMA average flow, number of clients, percentage of domestic consumption and average duration of events was carried out, but no relevant relations were found. Afterwards, event categories were organized in terms of the updated DMA sizes presented in Table 18 with the objective of relating outlier categories and DMA sizes. Figure 44a shows the relation between events and DMA size. According to the analysis, it was verified that:

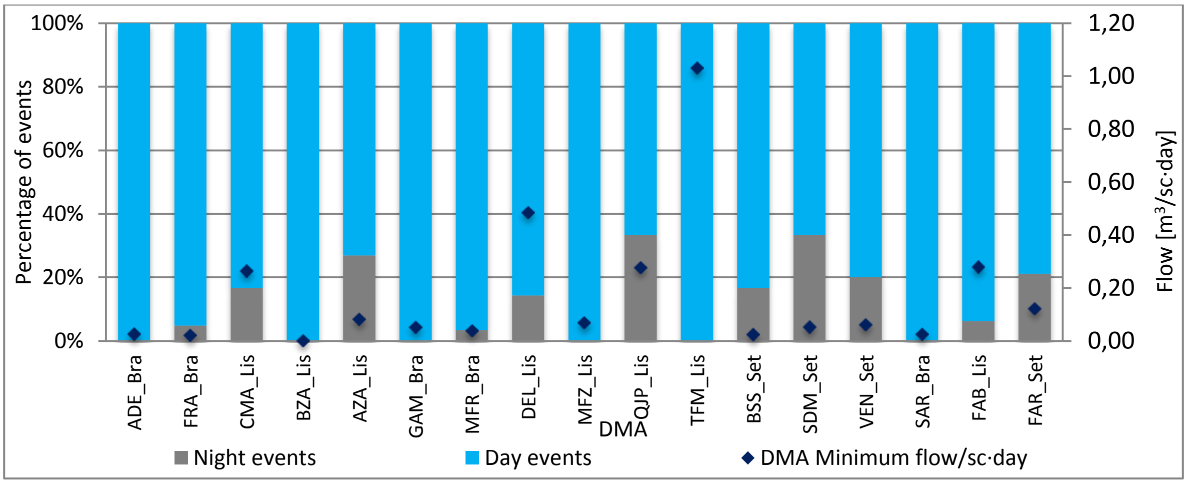
- All DMA sizes have the same range in terms of number of events and have most event categories, although medium-size DMAs have more "Sudden variation" events. "Constant flow" events seem to be more likely to happen in small and large-size DMAs.
- Few events were detected in the subset of the cluster analysis. Only eight DMA's have more than five events: FRA_Bra, GAM_Bra, TFM_Lis and FAB_Lis DMA, which have different sizes and an average of eight events detected, have mainly "Flow increase" events, usually followed by "Long duration" and "Sudden variation" events; BSS_Set has only six events detected, mostly "Flow decrease" events; AZA_Lis and FAR_Set DMA have exactly 26 detected events, most of them being "Constant flow" along

with the other event categories, with an emphasis on "Flow decrease" events and MFR_Bra DMA has 162 events, the majority of them being "Sudden variation" events followed by "Flow increase" events.

An attempt has also been made in order to search for correlations between the DMA minimum flow and the occurrence of night events. As depicted in Figure 44 b with the current subset, the presence of higher DMA minimum flows is not necessarily related with more night events. Four DMAs have no events in the night period, even though one (TFM_Lis) appears to have the highest minimum flow. Only two DMAs have simultaneously more than 20% of night events and more than five detected events. For these two particular cases, flow series were analysed and it was detected a large number of "Constant flow" and "Flow increase" events during short periods of time (less than 20 days) occurring in the night period.



(a)

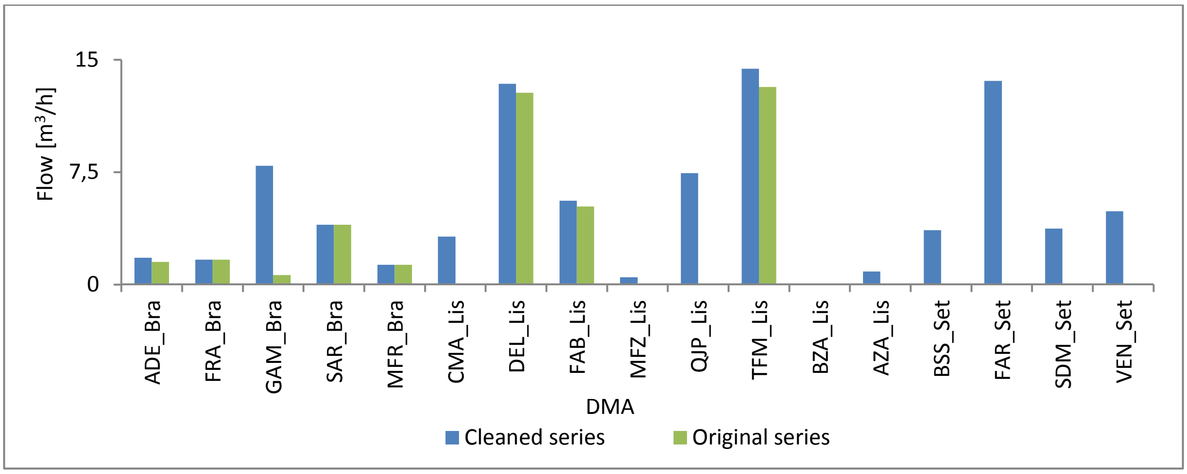


(b)

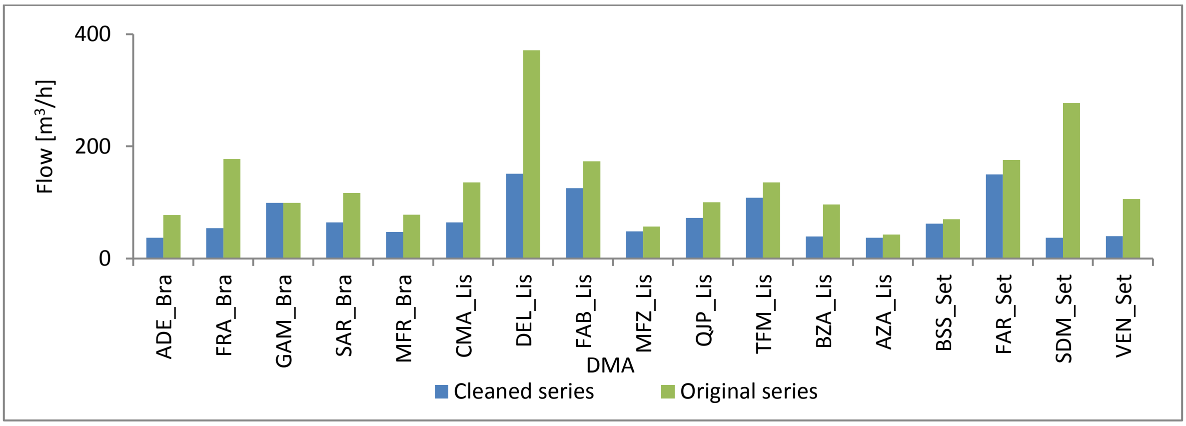
Figure 44 – Relation between events and DMA: a) Event categories and DMA size; b) night events and DMA minimum flow

5.4.4 Data cleaning

Once identified, outlier events were removed from the flow data series. To show the impact of data cleaning, basic statistics such as the minimum and the maximum flow value were calculated and compared with the original data series. As presented in Figure 45, the minimum values increase averagely 60% since events with low average flows, namely the “Constant flow” events, were detected and removed from the original series. As for the maximum values, an average decrease of 40% was observed, mainly as a consequence of removing events with high instantaneous flows, namely the “Sudden variation” events. Other statistics such as the average, 10^{th} and 90^{th} percentile were also calculated but there were not observed significant changes in comparison to the initial data series.



(a)



(b)

Figure 45 – Effects of outlier removal in basic flow statistics: (a) minimum; (b) maximum

5.5 Summary and conclusions

In this chapter a methodology was developed with the purpose of validating and normalizing flow data followed by outlier detection, classification and cleaning for calculating representative and reliable statistics of flow data as well as to obtain consumption patterns that describe the behaviour of domestic clients of a particular type.

The first step of the methodology is related with flow data collection. The case-study initially included 37 DMA’s but since data acquisition problems were identified in the second step – data validation – only a subset of 17 DMAs were validated for the subsequent analysis. The validated subset proved to have a good range of variability in terms of median flow, domestic clients, service connections and inhabitants.

The third step uses the PROFILER application developed by Loureiro (2010) to normalize the flow data series.

In the fourth step outlier detection has been made using the “Symmetric method” developed by Loureiro (2012), which is based on two robust statistics: the median values of a set of previous instants and the standard deviation of the observations based on the Qn scale. The method requires the specification of the number of instants where the medians should be calculated and the definition of a threshold value which is related to the probability of an observation being an outlier. Results show that the threshold value adopted in the present work tended to overestimate the number of outliers for flow series with regular behaviour, thereby suggesting the removal of important information such as domestic behaviours in the night period. Hence, the detected events were visually analysed and only an average of 11% of the events detected by the symmetric method was actually removed from the series.

An attempt to classify the outlying events has also been made. Two statistics related with the detected events were defined: average event duration and average event flow. Five categories were defined with specific criteria: “Sudden variation”, “Flow increase”, “Flow decrease”, “Long duration” and “Constant flow”. The different types of events were separated in case of happening exclusively during the night period with the aim of understanding the most common event types in this period. Results have shown that the most common events in this period are “Sudden variation” and “Constant flow” events. The night period was defined from 00h00 to 06h00, taking into account the period of lower flow rates (Loureiro, 2010).

Outlying events were also compared with work orders provided by water utilities, and it was found that they had very few records and that there was a lack of standardization in terms of events’ classification categories and its characteristics.

Cluster analyses have been carried out in order to identify some relations between the occurrence of outlying events and DMA characteristics. No relevant relations were found.

Afterwards, event categories were organized in terms of the DMA sizes. All DMA sizes appear to have the same range in terms of number of events. Generally, few events were detected and most of them happen in the day period. Moreover, all DMA sizes have all event categories, although medium-size DMAs have more “Sudden variation” events and “Constant flow” events are more likely to happen in small and large-size DMAs.

An attempt has also been made in order to search for correlations between the DMA minimum flow and the occurrence of night events. For the analysed subset, the presence of higher DMA minimum flows is not necessarily related with more night events.

As for the impact of the data cleaning – sixth step of the methodology – minimum values increase averagely 60% since events with low average flows, namely the “Constant flow” events, were observed and removed from the original series. As for the maximum values, an average decrease of 40% was observed, mainly as a consequence of removing events with high instantaneous flows, in particular the “Sudden variation” events.

The outlier detection methodology developed and applied herein is based on the one presented by Loureiro (2010), having the novel contributions presented in Table 23.

Table 23 – Comparison between the current research work and the one developed by Loureiro (2010) in terms of outlier detection

	Current research study	Previous research works (Loureiro, 2010)
Method used for outlier detection (step 4)	Uses robust statistics and is almost automatic	Depends on event categories and is based setting appropriate filters
Event definition (step 5)	After outlier detection (five categories)	Before outlier detection (three categories)
Criteria for event definition (step 5)	Average event duration and flow	Qualitative tendency and average duration
Exploring relations between DMA size and detected events (step 5)	Yes	No
Data cleaning impact evaluated (step 6)	Yes	No

As a final remark, the proposed methodology should be expanded using other types of flow data such as pressure and more case-studies should be validated for obtaining more reliable results.

6. WATER CONSUMPTION ANALYSIS

6.1 Introduction

Following the outlier detection and removal (Chapter 5), this chapter is the second part of Module 3 from the general methodology presented in section 3.2. and aims at determining consumption variables and patterns for different scenarios based on 17 DMAs with the clean flow data series.

Scenarios reveal trends in the consumption behaviour regarding seasonal effects at the year level (*e.g.* winter and summer seasons) and at the week level (*e.g.* working days and weekends). Scenarios are obtained using cluster analyses and are, therefore, used to calculate consumption variables related to maximum, average and minimum consumptions – variables defined according to the literature review (Donkor *et al.*, 2012; Loureiro, 2010) – and consumption patterns referring to the average daily consumption of water at the domestic level.

Scenario exploration provides a significant contribution for a further understanding on the influences beneath the different trends of consumption – a subject to be analysed in Chapter 7.

This chapter includes methodology, results and summary and conclusions of water consumption analysis.

6.2 Methodology

The general methodology for water consumption analysis involves two steps, as depicted in Figure 46. As previously referred, the flow data series have time step intervals of 15 minutes, which means 96 values per day. Clean series obtained in Chapter 5 enter as inputs of the “Scenario Building” (Step 1). In this step, the aim is to find seasonality among the flow data series.

Scenario Building includes two phases. The first one consists of determining the average daily consumption in every month, after which a cluster analysis is carried out.

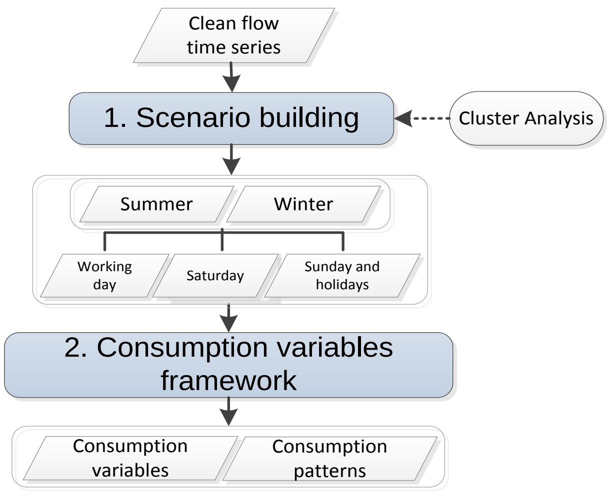


Figure 46 – Methodology used for consumption analysis

The cluster analysis searches for seasonality among groups of months, especially groups that characterize the winter and summer seasons. The second phase consists of determining the average daily consumption for working days, Saturdays and Sundays and bank holidays for the groups of months (*i.e.* winter and summer) previously selected. At this point, another cluster analysis is carried out in order to check if the differences between the referred types of days are significant.

The second step is "Consumption variables framework", which consists of setting up a framework of consumption variables to be calculated for the scenarios identified in step 1. The consumption variables' list was defined according to Donkor *et al.* (2012) and Loureiro (2010). Consumption patterns were also calculated at this step and consist of the average daily consumption, for the identified scenarios. Since the time step of the flow data is 15 minutes, each pattern is composed of 96 values that consist of the average consumption for the days in the scenario that is being calculated.

6.3 Case studies

As previously referred, in terms of water consumption data, utilities were asked to provide flow data series with one year of records (from 2010 or 2011) and a regular time step of 15 minutes. From the total of 150 DMAs with billing and infrastructure data, only 37 had flow data series, and from this subset, only 17 were validated for the outlier analysis. Notwithstanding this data reduction, this 17 DMAs are still representative, as verified in section 5.4.1. This new subset includes DMAs from Braga, Lisbon and Setúbal districts (see district location in Figure 19, section 4.3). DMAs' general characteristics after the data cleaning are presented in Table 24.

Table 24 – General characteristics of DMAs selected for outlier analysis (17 DMAs)

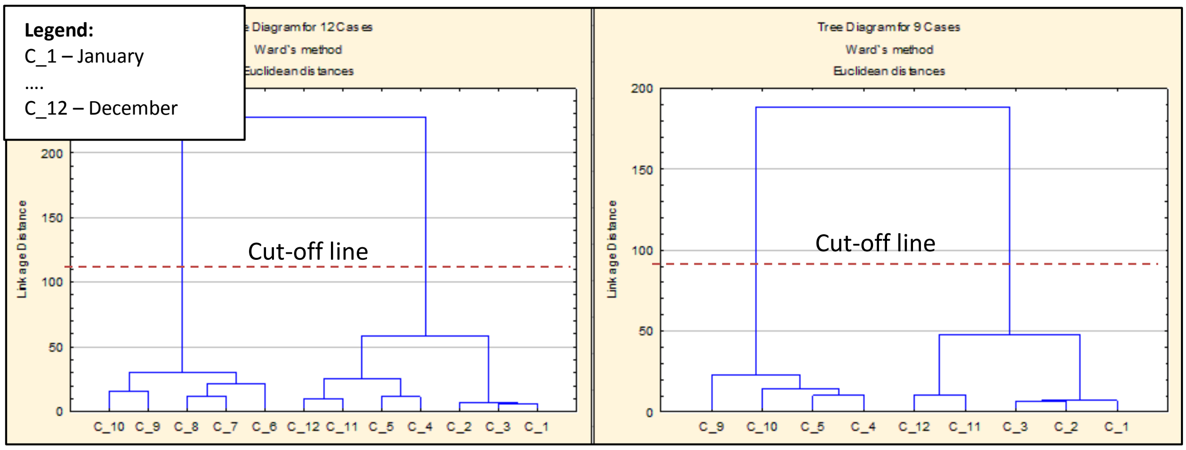
Characteristic	Interval	Average value	Median value
Diameter [mm]	75 – 131	100	102
Network length [km]	4.0 – 95.0	32	21
N.º service connections	250 – 3698	1462	1221
N.º of clients	742 – 5185	2183	2100
N.º of domestic clients	667 – 4514	1992	1863
N.º of inhabitants	46 – 12778	3311	2349

6.4 Results

6.4.1 Scenario building

Following the first step of the methodology presented in 7.2. the average monthly behaviour was determined for all 17 flow data series and groups of months with similar behaviours were identified. For instance, Figure 47 shows dendrograms resulting from the cluster analysis for two different DMAs. The analysis was carried out using the Ward´s method and Euclidean distances and the clusters are obtained by selecting a cut-off line.

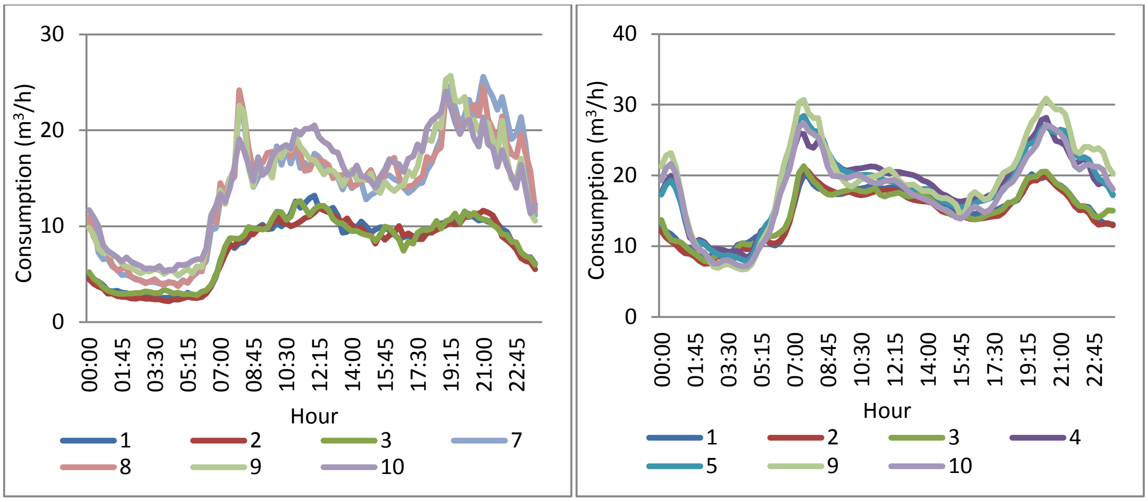
The cut-off line was selected in order to identify the main groups of months, avoiding a high level of disaggregation.



(a) (b)

Figure 47 – Dendrograms resulting from the cluster analysis at the month level: a) MFR_Bra DMA b) VEN_Set DMA

After the cluster analysis, DMAs behaviour was analysed according to the clusters obtained and two groups of months were selected. Generally, one group corresponded to the winter season and the other one corresponded to the summer season. Therefore, monthly scenarios were designated as “Winter scenario” or “Summer scenario”. As an example, Figure 48 shows the average daily behaviour for both scenarios. The winter scenario (Figure 48 a) includes January, February and March for both DMAs, whereas the summer scenario (Figure 48 b) includes July, August, September and October in MFR_Bra DMA and April, May, September and October for VEN_Set DMA. As depicted in the same figure, in DMAs with seasonality effects, the behaviours between two seasons can be significantly different. On the one hand, in MFR_Bra DMA the average daily consumption increases in all periods of the day and there are more irregularities such as peaks and falls in the consumption levels. On the other hand, in VEN_Set DMA, there are some periods of the day where the behaviours between the two scenarios are similar.



(a) (b)

Figure 48 – Average daily consumption for winter and summer scenarios a) MFR_Bra DMA b) VEN_Set DMA

The winter and summer scenarios for the 17 DMAs are depicted in Table 25.

Table 25 – Winter (Blue) and summer (red) scenarios for each analysed DMA

DMA \ Month	1	2	3	4	5	6	7	8	9	10	11	12
ADE_Bra												
FRA_Bra												
GAM_Bra												
SAR_Bra												
MFR_Bra												
CMA_Lis												
DEL_Lis												
FAB_Lis												
MFZ_Lis												
QJP_Lis												
TFM_Lis												
BZA_Lis												
AZA_Lis												
BSS_Set												
FAR_Set												
SDM_Set												
VEN_Set												

Typically, Summer scenarios occur from July to September, whereas winter scenarios occur from November to February. There are two DMAs (BZA_Lis and AZA_Lis) where the summer scenario could not be detected due to the lack of data. In general, there is an increase in the average daily consumption for the summer scenario. This can be due to the following:

- Irrigation of green spaces, that usually occurs from May to October (dry season).
- Pool filling, cleaning and maintenance.

Having explored the winter and summer scenarios, the next step is to understand whether the behaviours between days of the week are different for each scenario. Therefore, another cluster analysis was carried out using the seven days of the week, in the search for differences between working day, Saturdays and Sundays. Bank holidays were associated with Sundays. The cluster analysis was followed by a visual analysis, where the behaviour differences are evident for the majority of DMAs. As an example, Figure 49 shows the behaviours of MFR_Bra and VEN_Set DMAs. Since the majority of the DMAs have different behaviours depending on the day type, weekly scenarios were divided into three categories: working days, Saturdays and Sundays and bank holidays. Recent studies reveal clear distinction between Saturdays and Sundays (Loureiro, 2010; Palau *et al.*, 2012; Widén and Wäckelgård, 2010).

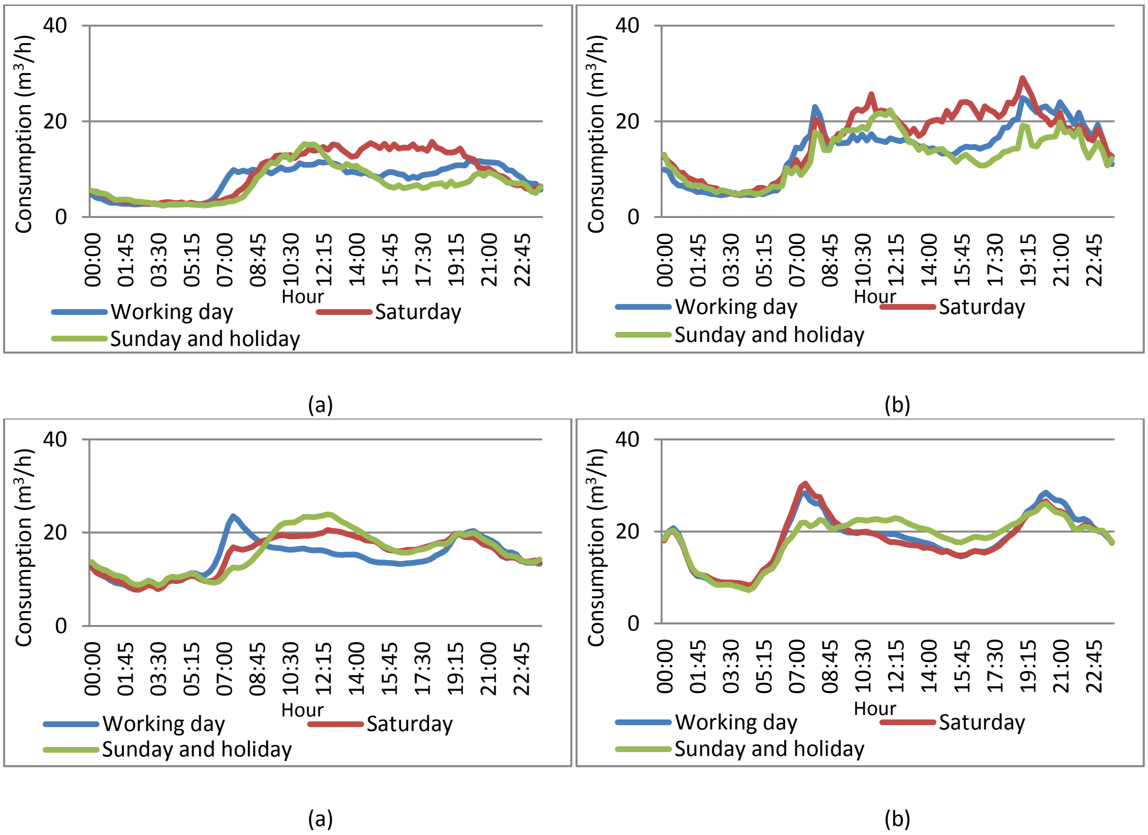


Figure 49 – Average daily consumption for weekly scenarios: a) winter MFR_Bra DMA; b) summer MFR_Bra DMA; c) winter VEN_Set DMA; d) summer VEN_Set DMA

The list of scenarios built up in this step is summarized in Table 26 along with the abbreviations that will be used from now on.

Table 26 – List of analysed scenarios and its abbreviations

Monthly Scenario	Weekly scenario
Winter (W)	Working day (WW)
	Saturday (WS)
	Sunday and holiday (WH)
Summer (S)	Working day (SH)
	Saturday (SS)
	Sunday and holidays (SH)

6.4.2 Consumption variables framework

After identifying groups of months and day types with different behaviours, a list of consumption variables was defined according to Donkor *et al.* (2012) and Loureiro (2010), as presented in Table 27. Consumption variables were calculated for the whole series and for certain scenarios, also presented in Table 27. Before the variables' calculation, the absolute minimum flow happening from [01:00 to 06:00[is removed from each value in the flow series. This procedure aims at separating consumption from real losses and is important because otherwise consumption variables, and consequently its profiling, would include a component that is not related with domestic and non-domestic consumption. In this particular case, since many scenarios were studied, minimums were removed according to the minimum values for each scenario. This procedure is based on the following hypotheses:

- Real losses can be estimated as the absolute minimum flow during the night period. This period was defined as [01:00 to 06:00[, based on the results of Loureiro (2010).
- Real losses are constant for each scenario.

Table 27 – List of quantitative consumption variables per type and category and analysed scenarios

Consumption variable type	Consumption variable category	Analysed scenarios	Units
Peaking	Instantaneous peaking factor (Ipf)	All series, W, S, WW, WS, WH, SW, SS, SH	[-]
	Daily peaking factor (Dpf)	All series, W, S	
	Monthly peaking factor (Mpf)	All series	
Average	Average daily consumption per client (Ave/cl)	All series, W, S, WW, WS, WH, SW, SS, SH	[l/cl·day]
	Average daily consumption per service connection (Ave/sc)	All series, W, S, WW, WS, WH, SW, SS, SH	[l/sc·day]
Night	Night consumption per client (Nig/cl)	All series, W, S, WW, WS, WH, SW, SS, SH	[l/(cl·h)]
	Night consumption per service connection (Nig/sc)	All series, W, S, WW, WS, WH, SW, SS, SH	[l/(sc·h)]
Minimum	Minimum consumption per client (Min/cl)	All series, W, S	[l/cl·day]
	Minimum consumption per service connection (Min/sc)	All series, W, S	[l/sc·day]

The analysis of some consumption variables for the 17 DMAs with clean flow data series is presented below.

Instantaneous peaking factor (Ipf)

Concerning the instantaneous peaking factor, the Portuguese Decree-Law nº 23/95, and Article n. º 19 (abbreviated herein as DL 23/95) recommends that when consumption data is unavailable, the instantaneous peaking factor should be calculated according to the following expression:

$$Ipf = 2 + 70 \cdot P^{-0.5} \tag{3}$$

Being:

P – The population to serve [Inh]

Instantaneous peaking factors were calculated for the whole series and using DL 23/95 expression. Results show that the expression from DL 23/95 provides good estimations in some cases, as depicted in Figure 50.

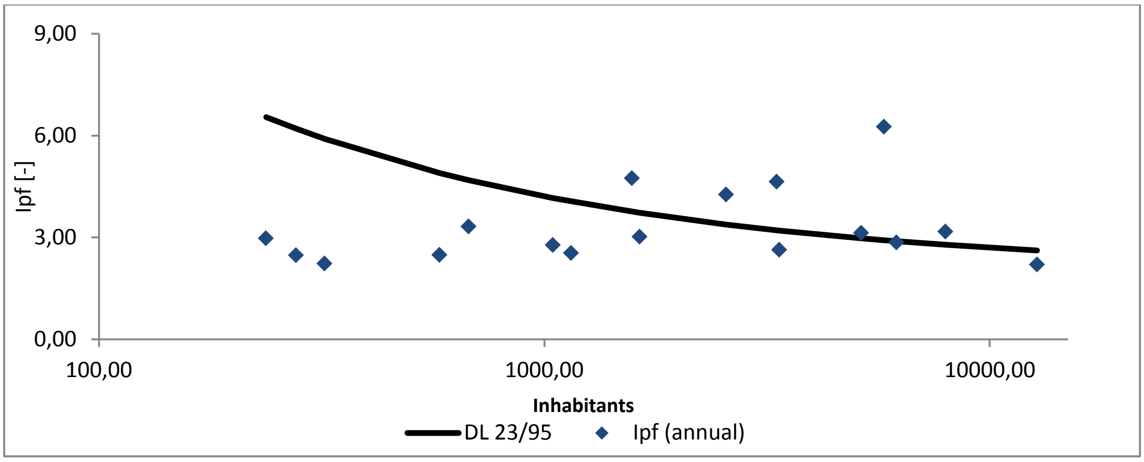
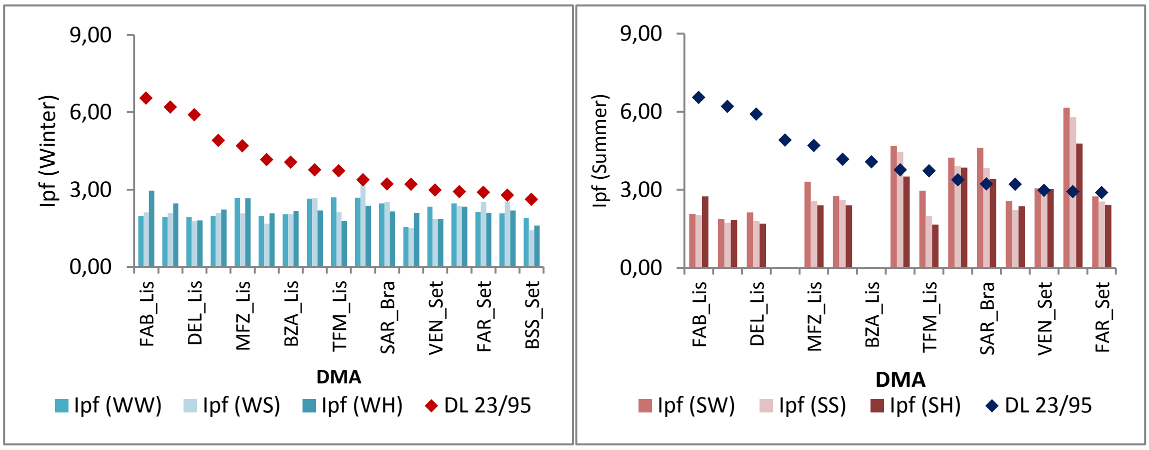


Figure 50 – Comparison between calculated peaking factors (Ipf) and DL 23/95

Concerning the winter and summer scenarios, peaking factors calculated for the winter are typically lower than the ones provided by the DL 23/95 (Figure 51), thereby suggesting over-designed mains.



(a) (b)

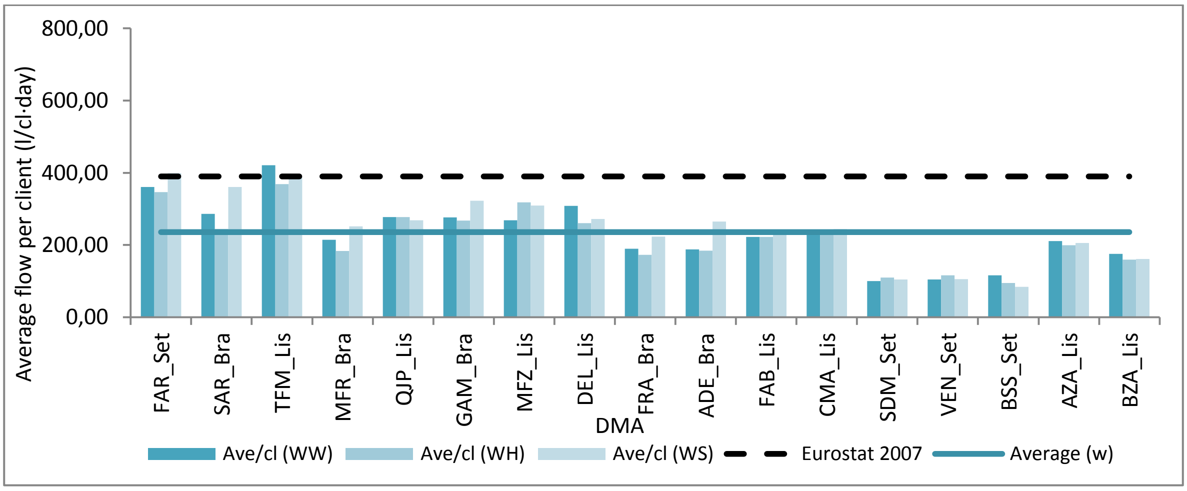
Figure 51 – Peaking factors for monthly scenarios: a) winter; b) summer

Some of the calculated peaking factors in the summer season are higher than the ones from DL 23/95, as depicted in Figure 51. It should also be noted that the summer season is characterized by a general increase in most of the flow series. These results highlight the importance of a more detailed analysis of peaking factors for the design of networks.

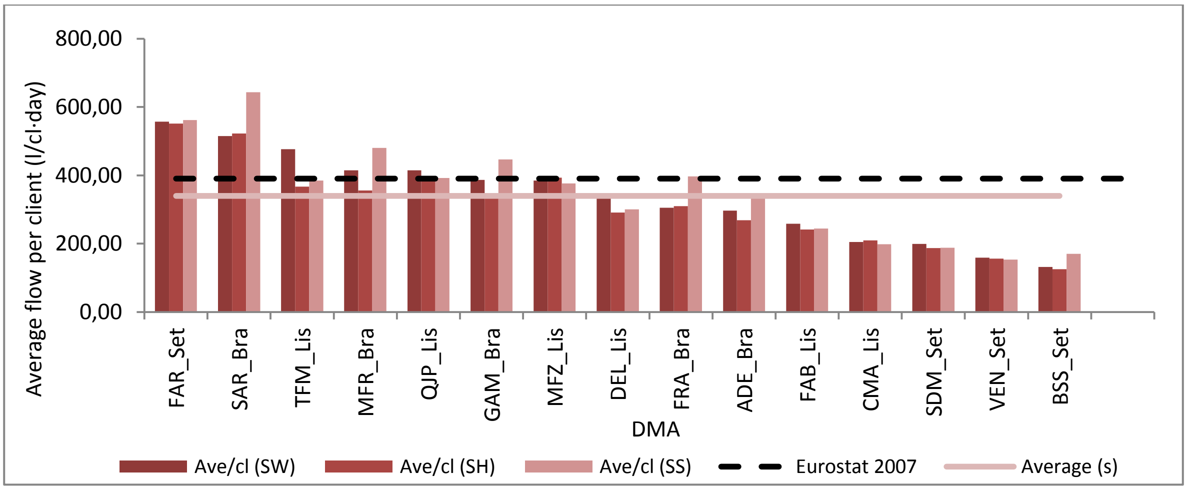
Average daily consumption per client (Ave/cl)

Concerning the average daily consumption per client, calculated values were compared with Eurostat 2007 statistics, which provide the total supply to the domestic sector per inhabitant. This value was converted to express this parameter per client, by multiplying for the average number of inhabitants per client. Hence, the

Eurostat reference value used was 390 l/cl·day. Figure 52 presents the average daily consumption per client for winter and summer scenarios.



(a)



(b)

Figure 52 – Average flow per client: a) winter; b) summer

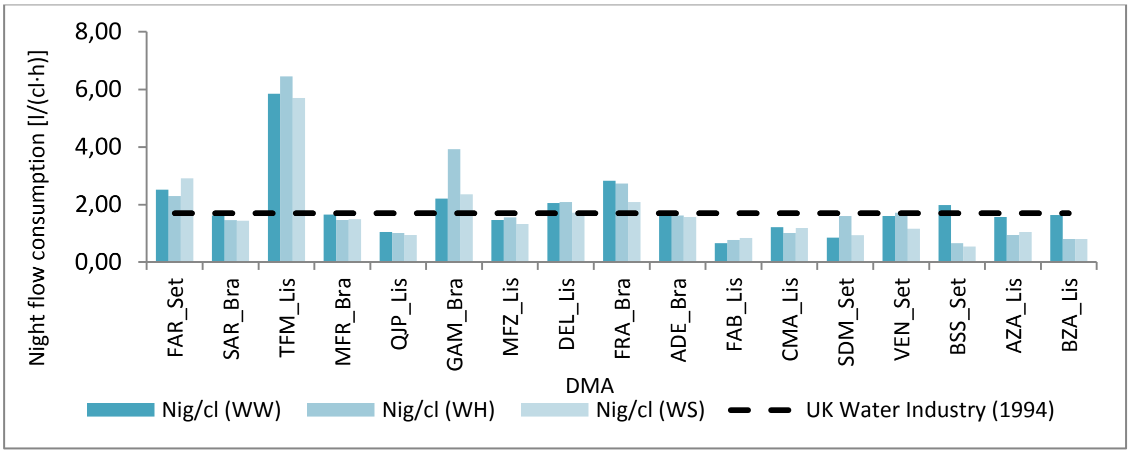
The winter scenario (Figure 52 a) exhibits an average daily consumption of nearly 236 l/cl·day, whereas the summer scenario (Figure 52 b) shows higher values, with an average of 340 l/cl·day. Comparing with the Eurostat statistics, winter values are below the statistics (except in one case). For the summer scenario, more DMAs show higher consumption than the statistics. Moreover, in the summer season is more common to have higher consumptions in the Saturday.

Night consumption (Nig)

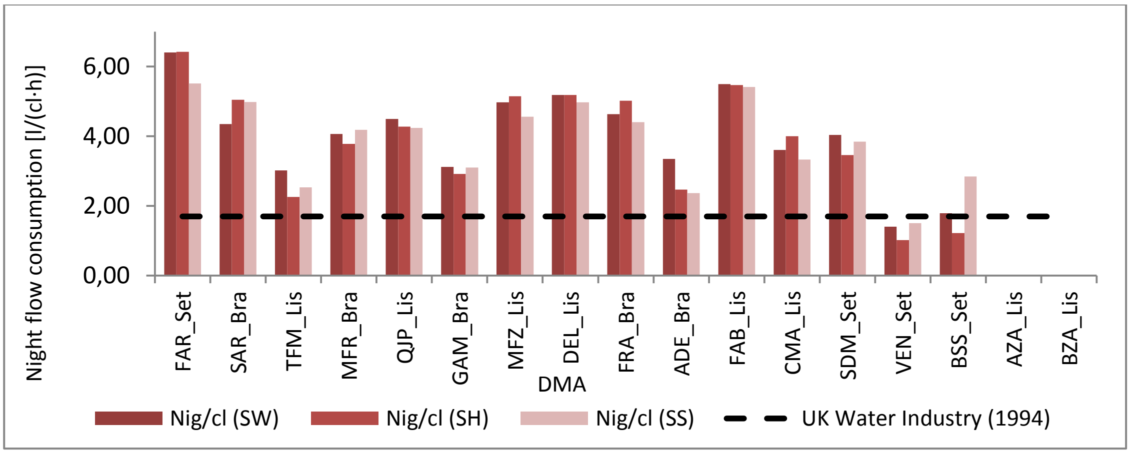
When analysing night consumption – the average flow from [03:00 to 05:00[– seasonal differences are evident. The period for analysing night flows was defined according to Loureiro (2010) and represents the period of minimum night consumption.

It is important to evaluate night flows because water use is at a minimum and it is easier to identify and subtract the legitimate flows. If the night flow minus the legitimate flow is close to zero, the leakage must also be close to zero (Mounce *et al.*, 2010).

Night flows were compared with an indicative value provided by the UK Water Industry (1994) – 1.70 l/cl·h. Results show that although the night flows during the winter season (Figure 53 a) are very near 1.70l/cl·h, values for the summer season are considerably above the UK Water Industry (1994) indication (Figure 53 b). There is an exception in the winter for TFM_Lis DMA, which is probably related with a non-household use.



(a)



(b)

Figure 53 – Night flow per client: a) winter; b) summer

Minimum consumption (Min)

Concerning the minimum consumption defined in the present work as the minimum consumption from [01:00 to 06:00[. Previous studies, particularly Lambert and Hirner (2000), point to an average minimum consumption of 100 l/sc.day. As presented in Figure 54, approximately half of the DMAs are near the referred limit. TFM_Lis and DEL_Lis have a minimum consumption that is 10 times higher than the limit. This fact may be due to the existence of non-household consumptions.

During the winter season the minimum consumption is more directly linked with real losses. In opposition, minimum flows in the summer season are generally higher. This may be related with outdoor uses such as the irrigation of green spaces which usually takes place during the summer season or other outdoor activities such as pool filling, cleaning and maintenance.

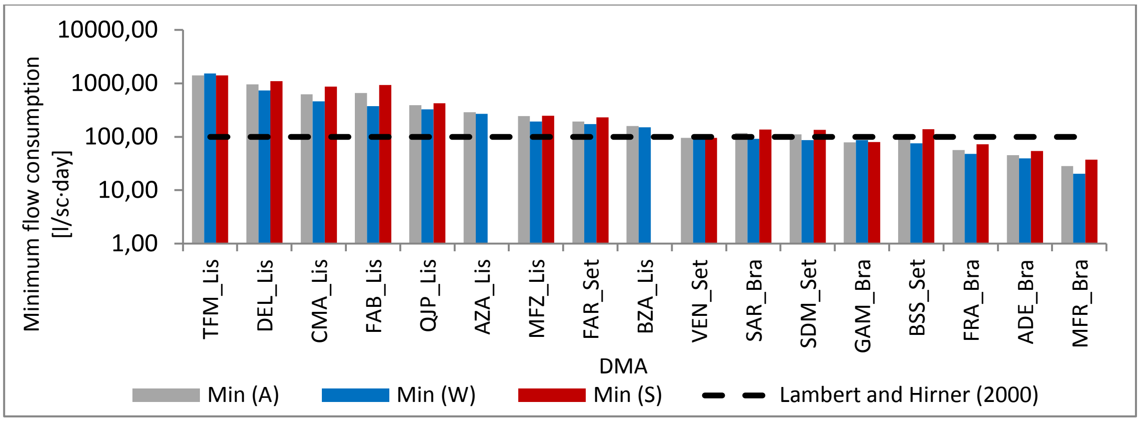


Figure 54 – Minimum flow per client: a) winter; b) summer

6.5 Summary and conclusions

This chapter focused on presenting a consumption analysis that follows a scenario building. Scenarios are defined at a monthly level (*i.e.* winter and summer) and at a weekly level (*i.e.* working days, Saturdays and Sundays and bank holidays).

Results have shown that the actual expression for calculating peaking factors based on DL 23/95 should be updated taking into account more recent studies. Two new expressions were proposed in case of DMAs with more or less than 2500 inhabitants. Comparing the results with DL 23/95 for each season, peaking factors calculated for the winter are lower than the ones provided by DL 23/95, thereby suggesting over-designed mains. However, some of the calculated peaking factors in the summer season are higher than the ones from DL 23/95.

The tendency of increased consumption in the summer season also reflects on the remaining consumption variables. For instance, when analysing night consumption and minimum flows, the seasonal differences are evident. This increase is probably related with outdoor uses such as the irrigation of green spaces or pool filling, cleaning and maintenance.

Furthermore, and as an additional contribution, consumption variables were compared with indicative values from the country regulations and previous studies. Generally, reasonable values were obtained for the studied DMAs.

The methodology for consumption analysis followed in this chapter is innovative when comparing to the one developed by Loureiro (2010) since it does an attempt to remove real losses from the flow series before calculating consumption variables.

7. WATER CONSUMPTION PROFILING

7.1 Introduction

The current chapter aims at profiling water consumptions in DMAs based on the previous analysis of the socio-demographic, billing and infrastructure framework (Chapter 5) of nearly 100 DMAs and consumption analysis of 17 DMAs (Chapter 7). This chapter is the Module 4 from the general methodology presented in section 3.2.

Profiling is made by setting consumption variables and patterns as dependent variables, and socio-demographic, billing and infrastructure indexes as independent variables followed by the application of exploratory statistical techniques to assess correlations between these two groups of variables and obtain empirical relations to quantify them.

In this study, profiling consists of forecasting consumption at the spatial and temporal level. This can be particularly valuable for water to forecast short-term consumption in non-metered areas where knowledge on socio-demographic, infrastructure and other types of data are available. This chapter includes methodology, results and summary and conclusions.

7.2 Methodology

The methodology used for water consumption forecasting is a two-step procedure, as depicted in Figure 55.

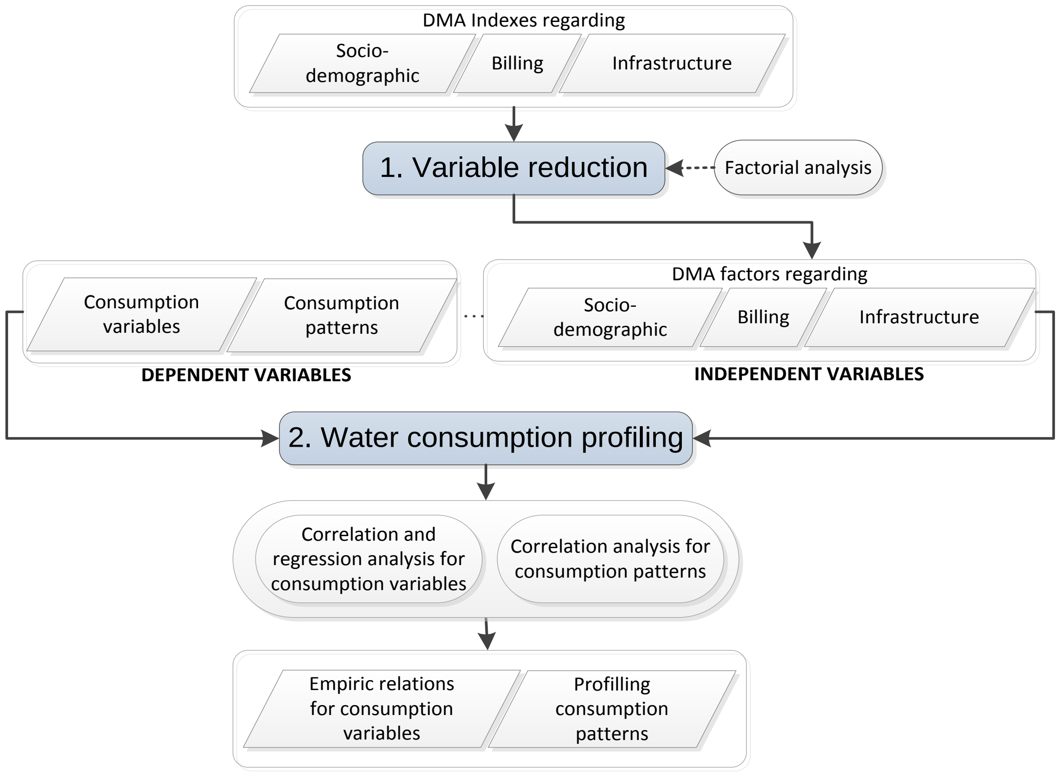


Figure 55 – Methodology for water consumption forecasting

Among the 49 indexes concerning socio-demographic, billing and infrastructure analysis (Chapter 4), the first step of the methodology is “Variable reduction”. Based on the principle of parsimony, a model should depend on the fewest number of independent variables (Donkor *et al.*, 2012). Therefore, a multivariate analysis technique is applied in order to reduce the number of independent variables for the forecasting model. Particularly, a Principal Components Analysis (PCA) using varimax rotation is used. This analysis aims at explaining the correlation between variables and assumes that there are a reduced number of non-measured components, or factors that express what is common between the original variables (Pestana and Gageiro, 2003).

The PCA is carried out by category and follows an iterative procedure to understand the most relevant indexes. The user starts with all indexes from a certain category (*e.g.* socio-demography) and runs the analysis. In the current work the PCA was carried out using SPSS®. Each iteration delivers the following results:

- Factors: the analysis combines all indexes in a reduced number of factors which are a linear combination of the original indexes. Each index has a score on each factor that varies between -1 and 1 and that represents the index weight in the factor (absolute values near 1 indicate higher weights).
- Explained variance: a percentage that, as suggested by the name, refers to the variance that is explained by each factor.
- Kaiser-Meyer-Olkin (KMO) test: A coefficient that varies between 0 and 1 and evaluates the correlation between the variables that are being used in the PCA. In other words, it indicates whether the sample of items is adequate (values near 1) for the PCA.

After every iteration, the user needs to evaluate the quality of the results and decide whether another iteration is needed. This can be done by looking at the percentage of explained variance and the KMO coefficient. If the results are not acceptable (e.g. lower percentage of explained and/or KMO coefficient), the user should do a new iteration removing indexes with lower scores (e.g. scores with absolute value <0.4). After the new run, the quality of the results should be re-evaluated. The user should perform as much iteration as needed in order to obtain acceptable values in both parameters: percentage of explained variance and KMO coefficient. The limits for assuming acceptable values should be defined according to the sample and the analysis.

The second step is “Water consumption profiling”, where factors resulting from the previous step are the input for the forecasting model, entering as independent variables. Qualitative variables such as the consumption patterns obtained in Chapter 6 are converted into dummy variables and are used as dependent variables, along with the other consumption variables calculated in Chapter 6.

A dummy variable (also known as an indicator variable) is one that takes the values 0 or 1 to indicate the absence or the presence of some categorical effect. In this study, 17 dummy variables (or dummies) were created as a consequence of existing 17 different consumption patterns (Pestana and Gageiro, 2003).

For instance, a DMA with a certain consumption pattern would be assigned with 1 in the dummy associated with that pattern, to indicate its presence. A DMA that did have that pattern would be assigned with 0 in the same dummy, to indicate the absence of that same pattern.

At this point, due to the large number of dependent variables and before the regression analysis, a correlation matrix is calculated so as to identify the strongest correlations between each consumption variable (dependent variables) and each factor (independent variables). Then, the strongest correlations (given by matrix coefficients greater than a defined limit) are explored in a Multiple Linear Regression (MLR) analysis.

MLR analysis explores the relation between a dependent variable and a set of independent variables. This analysis requires that variables belong to an interval and that the relation between them is linear and additive. Nevertheless, this condition is not an absolute one, since qualitative variables can be transformed into dummies and added in the model, as it could be done with the consumption patterns. MLR is used since 1970s and earlier to test the relation between water consumption and independent factors (Tanverakul and Lee, 2012)

For profiling consumption patterns, a correlation matrix using socio-demographic indexes as independent variables is calculated and consumers' behaviour is analysed for different types of days (i.e. working days, Saturdays and Sundays and holidays). Only socio-demographic indexes are used, since these were found to be the most important for the analysis of daily consumption patterns (Loureiro, 2010).

7.3 Results

7.3.1 Variable reduction

A PCA using Principal Factors was carried out with the purpose of reducing the number of socio-demographic, billing and infrastructure indexes into factors. Each factor is a linear combination of the most relevant original indexes. Generally three iterations were sufficient to obtain good explained variances and KMO coefficients, as presented in Table 28. For all analyses, two factors were obtained. It was defined that for the current analysis, explained variances should be higher than 80% in accordance with Loureiro (2010) and KMO coefficients should be higher than 0.70 (this condition could not be applied in the infrastructure analysis). The PCA was carried out for the number of DMAs also shown in Table 28 in order to obtain factors based on a robust sample of DMAs with socio-demographic, billing and infrastructure indexes. Then, the PCA was repeated using only the 17 DMAs with consumption variables to assure that the same factors were obtained.

Table 28 – Input data and results of the PCA

Analysis type	Number of DMAs	Number of iterations	Explained variance	KMO coefficient
Socio-demographic	96	3	82.44%	0.72
Billing	101	3	84.54 %	0.73
Infrastructure	96	5	92.03%	0.65

Factors resulting from the PCA can be represented in a Cartesian axis, where the horizontal axis represents the highest data variability and the second one (vertical axis) represents the second highest data variability. The most representative factors are the ones nearer to -1 or 1, as previously explained.

Results obtained for each analysis type are presented below and compared with the ones obtained by Loureiro (2010).

Socio-demographic factors

As a result of the socio-demographic PCA (Figure 56), two factors are extracted – *Family structure* (factor 1) and *Building structure* (factor 2).

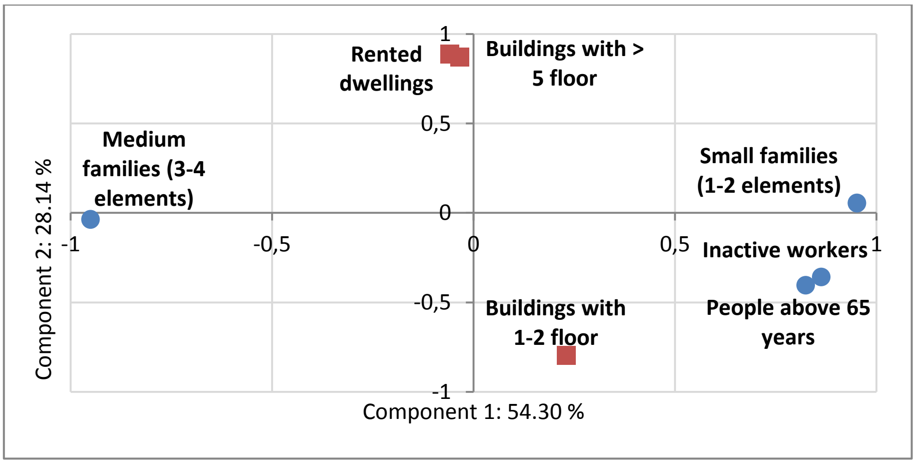


Figure 56 – Projection of socio-demographic indexes in the Cartesian plan 1-2

Based on the two factors and its projection, the following interpretation can be given:

- *Family structure* (factor 1) – represents 54% of data variability and is named as "Family structure" since it reflects the opposition between medium families (composed of 3-4 elements) and small families (composed of 1-2 elements). The percentage of inactive workers and people above 65 years is also correlated with small families. In other words, this means that many small families have inactive workers and elderly people.

- *Building structure* (factor 2) – explains 28% of data variability and essentially expresses the contrast between Buildings with 1-2 floors which are usually villas and Buildings with more than 5 floors. Rented dwellings' score is very close to the latter type of buildings, which means a positive correlation between the two indexes. The designation adopted for this factor is due to the different building structures identified in the studied DMAs.

Concerning the socio-demographic analysis, the *Family structure* factor is also present in Loureiro (2010) work. However, instead of *Building structure*, Loureiro (2010) found *Social stratification*, which included social and economic mobility and People with 12 years of education to be more relevant. Differences can be associated with the fact that the current work uses Census 2011 data and Loureiro (2010) used Census 2001 data. Important socio-demographic changes may have occurred in the past 10 years, which may explain the slight differences obtained.

Billing factors

PCA carried out over billing indexes allowed the distinction of two factors: *Domestic level* (factor 1) and *Consumption type* (factor 2). The projection of billing indexes in the factorial plan 1-2 is shown in Figure 57.

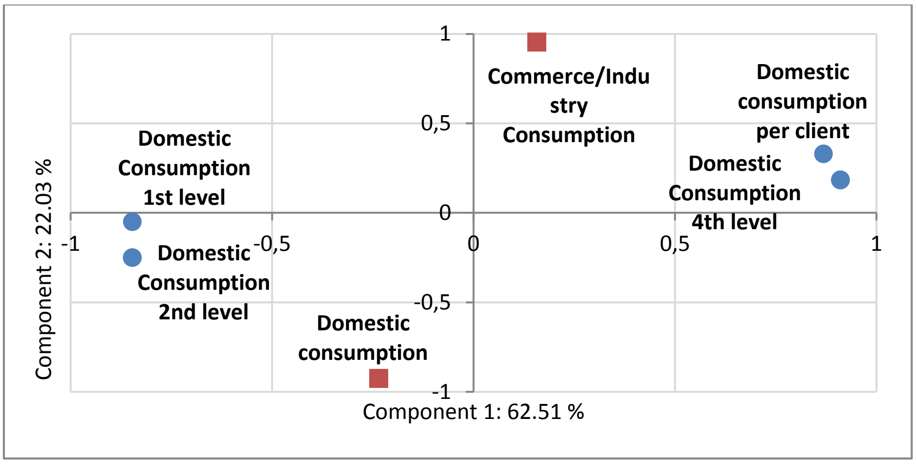


Figure 57 – Projection of billing indexes in the Cartesian plan 1-2

The following interpretation can be given:

- *Consumption level* (factor 1) – considering billing indexes that represent consumption levels according to IRAR (2009), this factor explains nearly 63% of data variability. This factor opposes Domestic consumption in the 1st and 2nd level (*i.e.* monthly consumption below 15 m^3) with Domestic consumption in the 4th level (*i.e.* monthly consumption above 25 m^3). Domestic consumption per client is positively correlated with consumption in the 4th level, which indicates that more consumption in this level leads to higher average consumption per client, as expected. This factor was designated as “Consumption level” since it implies the contrasts between the different consumption levels.

- *Consumption category* (factor 2) – represents 22% of data variability and expresses the contrast between commerce and industry consumption and domestic consumption. Since differences between two types of consumptions were detected, this factor was designated as “Consumption category”.

Concerning factor 1, Loureiro, 2010 also found a *Consumption level* factor that revealed the differences between consumption levels. As for factor 2, Loureiro, 2010, directly selected public and collective consumptions as factors. In this works, however, the billing analysis included all indexes and results showed one factor pinpointing the differences between only two consumption categories: commerce and industry and domestic. Thus, and contrarily to Loureiro, 2010, indexes concerning public and collective consumption were not relevant in this study. In fact, they were dropped out right after the first iteration due to having lower scores.

Infrastructure factors

The results of the PCA carried out over infrastructure indexes are presented in Figure 58. After the analysis, two factors that explain 92% of data variability were revealed: *Pipe diameter* (factor 1) and *Pipe age* (factor 2).

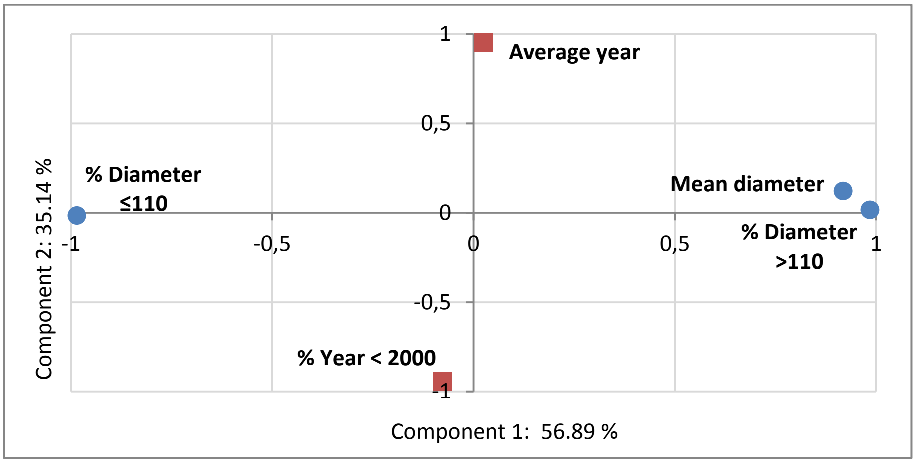


Figure 58 – Projection of infrastructure indexes in the Cartesian plan1-2.

In the light of the results, the following interpretation can be given:

- *Pipe size* (factor 1) – explains nearly 57% of the data variability and separates pipes with diameters above and below 110 mm. The *Mean diameter index* is positively correlated with pipes with diameters above 100 mm, thereby suggesting that higher mean diameters mean a higher percentage of pipes with more than 110 mm. Since the diameter range is exposed in this factor, it was designated as *Pipe size*.

- *Pipe age* (factor 2) – represents 35% of the data variability and it addresses the pipe's age. On the one side of the axis is the *Average year* which is associated with recent pipes (*i.e.* constructed after 2000) and on the other is the percentage of pipes installed after 2000. Although not being present in the current analysis, in Chapter 4 an approach was made relating pipe´s age and material. It was concluded that the majority of recent pipes are made of plastic (PVC and HDPE).

As for the infrastructure analysis, Loureiro (2010) found *Pipe size* and *Pipe material* to be the most relevant factors. *Pipe diameter* included the same indexes as the ones presented herein, whereas *Pipe material* included pipes in plastic, asbestos cement and pipe´s installation year. For the current analysis, *Pipe age* was more relevant, thought it was proven to be intrinsically linked with pipe´s material.

A summary of the factors obtained by the PCA in Principal Factors is presented in Table 29. Comparing to the work developed by Loureiro (2010), the same table shows that in every category, at least one factor is always present in both analysis.

Table 29 – Summary of factors resulting from the PCA

Category	**Factor (Explained variance)**	**Factor designation**	**Symbology**	**Original factors (scores)**	**Loureiro (2010)**
Socio-demography	Factor 1 (54.30%)	Family structure		Small families (0.95), Medium families (-0.95), People above 65 years (0.86), Inactive workers (0.82)	X
	Factor 2 (28.14%)	Building structure		Buildings with >5 floor (0.89), Rented dwellings (0.87), Buildings with 1-2 floor (-0.80)	
Billing	Factor 1 (62.51%)	Consumption level		Domestic consumption per client (0.91), Domestic Consumption 4th level (0.87), Domestic Consumption 2nd level (-0.85), Domestic Consumption 1st level (-0.85)	X
	Factor 2 (22.03%)	Consumption category		Commerce/Industry Consumption (0.95), Domestic consumption (-0.92)	
Infrastructure	Factor 1 (56.89%)	Pipe size		% Diameter >110 (0.98), % Diameter ≤110 (-0.98), Mean diameter (0.92)	X
	Factor 2 (35.14%)	Pipe age		Average year (0.95), % Year <2000 (-0.95)	

7.3.2 Correlation analysis for consumption variables

Having reduced the number of independent variables from 49 to 6, the next step is to analyse which relations between dependent and independent variables should be explored. Since the PCA has been carried out for nearly 100 DMAs and there are only 17 DMAs with consumption variables (*i.e.* there are only 17 cases with dependent variables), a control needs to be made in order to assure that the variable reduction is still valid. This validation has been made by repeating the PCA for the 17 DMAs only and results showed the same factors.

A correlation analysis is then carried out by calculating the correlation matrix between the 55 dependent variables associated with consumption variables and the 6 reduced independent variables associated with socio-demographic, billing and infrastructure indexes. In what concerns consumption patterns, they were transformed into dummy variables, as explained in 8.2 and were also analysed using a correlation matrix. Consumption patterns and their relation with the independent variables will be discussed in section 7.3.4.

Each cell in the correlation matrix has a coefficient that varies from -1 to 1 and absolute values closer to 1 indicate better correlations between certain dependent and independent variable. Correlation coefficients with absolute values higher than 0.5 were selected for analysis.

Table 30 presents a schematic representation of the correlation matrix for consumption variables. For a better understanding of the correlation matrix, only selected variables (coefficient >0.5) are shown. Whenever the factors are positively correlated with the dependent variables a (+) is placed below the symbol and when the correlation is negative a (-) is placed.

Consumption variables were separated into categories (rows) and scenarios (columns), using the abbreviations already used in Chapter 6.

Table 30 – Schematic representation of the correlation matrix analysed

Variables		Scenarios								
Type	**Category**	**A** **Annual**	**W** **Winter**	**S** **Summer**	**WW** **Winter working day**	**WS** **Winter Saturday**	**WH** **Winter Sunday and holiday**	**SW** **Summer working day**	**SS** **Summer Saturday**	**SH** **Summer Sunday and holiday**
Peaking	**Instantaneous peaking factor**	+	+	+	+ +	+		- +	- +	+
	Daily peaking factor	-		-						
	Monthly peaking factor									
Average	**Average flow per client**		- +		- +	+	- +			
	Average flow per service connection	-	-	- + +	-	- +	- +	- + +	- + +	- + +
Night	**Night flow per client**		-		-	-	-	-		
	Night flow per service connection	- +	-	+ - +	-	-	-	+ - +	+ - +	+ - +
Minimum	**Minimum flow per client**		-							
	Minimum flow per service connection	-	-	-						
Symbology		- Family structure - Building structure - Consumption level - Pipe size - Pipe age								

For example, peaking factors are linearly linked with *Pipe age* ( ), *Consumption level* ( ) *and Pipe size* ( ), because these symbols appear in the correlation matrix for *Peaking* type consumption variables. In particular, since a (+) signal is placed below *Pipe age* ( ) *and Pipe size* ( ), this means that peaking factors are positively correlated with these factors. In other words, it means that peaking factors are higher for recent pipes and pipes with diameters above 100 mm. On the other hand, since a (-) signal is placed below *Consumption level* ( ), this means that peaking factors are higher for clients that have monthly consumptions in the 1st and 2nd level (*i.e.* below 15 m^3). This last inference is probably just related with the method for calculating peaking factors. Peaking factors are calculated as a ratio between the maximum and the average consumption. This means that consumptions 1st and 2nd level, that represent lower average consumptions, lead to higher peaking factors due to the calculation method.

The fact that peaking consumption is higher for recent pipes and higher diameters is probably related to a higher variability in consumption in recent networks and with higher diameters.

Following the same logic, for average flow per client or per service connections, factors such as *Building structure* ( ), *Consumption level* ( ) *and Pipe size* ( ) are important. In particular, average flows are generally higher for buildings with 1-2 floors, monthly consumptions in the 4th level (*i.e.* above 25 m^3) and pipes with diameters above 100 mm. Higher average flows can be directly linked with buildings with 1-2 floors since these are the ones that probably have more consumption due to outdoor uses such as irrigation and pool filling.

As for the night flows per client or per service connections, Family structure ( ), *Building structure* ( ), *Consumption level* ( ) are relevant, with differences between seasons. In particular, night consumption in the winter season is higher for families with 3-4 elements, buildings with 1-2 floors. For the summer season, night consumption is higher for families with 1-2 elements, buildings with 1-2 floors and monthly consumptions in the 4th level (*i.e.* above 25 m^3). Since the summer season is characterized by higher consumptions it is reasonable that night flows suffer an increase that can be due to higher outdoor uses. Hence, the association with consumptions in the 4th level (*i.e*. higher water volumes) and buildings with 1-2 floors (*i.e.* higher outdoor uses) is highly probable.

For minimum flows per client or per service connections Family structure ( ) and *Building structure* ( ) are considerably related. After analysing the correlation matrix, higher minimum consumptions are associated with buildings with 1-2 floors and families with 3-4 elements. Higher minimum flows in buildings with 1-2 floors can be related with higher water consumptions that are typical in this type of buildings, due to outdoor uses (as discussed for night consumption). Additionally, since this type of buildings has generally large areas, it probably has more water appliances, which increases the probability of having higher water losses (*e.g.* by having more toilet flushes, for instance).

7.3.3 Regression analysis for consumption variables

A multiple linear regression (MLR) analysis has been conducted on the dependent variables that were correlated with more than one independent variable, based on the correlation matrix previously presented (Table 30). Results are showcased per types of flow variables. The general

Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$ of *n* statistical units, a linear regression model takes the form of:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_I \qquad i = 1, \dots, n \tag{3}$$

in which:

y_i: Dependent consumption variable *i*

β_P: Regression coefficient related with independent variable *p*

x_{ip}: Independent variable *p*, related with dependent variable *i*

ε_I: Random error at case *i*

In every regression, the regression coefficients $\beta_0 \dots \beta_p$ Represent an increase (positive value) or decrease (negative value) in the dependent variable, associated with each independent variable. The expected value of the dependent variable is equal to β_0 when the remaining regression coefficients are null. In order to evaluate the quality of the results, the standard deviation associated with each regression coefficient is shown, as well as the adjusted correlation coefficient r_a^2. This last coefficient is called "adjusted" since it reflects the number of independent variables as well as the sample size. Additionally, the *p* value of the F-test for each regression model is shown.

Results from regression analysis are compared with previous studies about water or energy consumption. This is due to the fact that part of energy consumption is related with domestic water consumption, as explained in section 2.5.

Instantaneous peaking factor

In what concerns the peaking factors, results show a good adjustment with an average r_a^2 of 0.55, which means that 55% of the instantaneous peak factor is explained by the coefficients in Table 31. Generally, instantaneous peak factors are higher for recent pipes (all studied scenarios), large diameters (winter working day scenario) and consumption in the 1st and 2nd level (summer working day and Saturday scenarios).

Loureiro (2010) could not obtain good adjustment results for the peaking factors. However, the fact that the presented results are not correlated with social characteristics corroborate the conclusions of several other studies that found that including socioeconomic variables such as population has a negligible impact on short-term water consumption forecasting. This is understandable since it is highly unlikely that overall zone population would have an effect on daily fluctuations in water consumption (Adamowski, 2008; Alegre *et al.*, 1992).

On the electricity side, McLoughlin et al. (2012) found Household composition, number of bedrooms, water heating and cooking type were the most significant variables to influence maximum electricity consumption.

Table 31 – Regression coefficients resulting from the MLR analysis for Instantaneous peaking factors

Dependent variable	Scenario	Explaining factor	Symbol	Regression coefficient	Estimation	SD	*p*-value	r_a^2	$\overline{r_a^2}$
Instantaneous peaking factor [-]	Winter working day	Constant		β_0	2.24	0.07	0.00111	0.57	0.55
		Pipe size		β_1	0.19	0.089			
		Pipe year		β_2	0.74	0.05			
	Summer working day	Constant		β_0	3.04	0.22	0.00248	0.57	
		Consumption level		β_1	-0.29	0.26			
		Pipe year		β_2	0.64	0.19			
	Summer Saturday	Constant		β_0	2.71	0.22	0.00566	0.51	
		Consumption level		β_1	-0.31	0.26			
		Pipe year		β_2	0.59	0.120			

E.g. *The instantaneous peaking factor (Ipf) for a working day in the winter season can be given as:*

Ipf = 2.24 + 0.19 + 0.74

The empiric relation above means that instantaneous peaking factors for the specified scenario are higher for recent pipes and pipes with diameters above 110 mm (see ***indexes in the both positive axes of*** *Figure 58).*

Average flow

In what concerns the average flow, results using the average flow per client do not have acceptable adjustment values (average r_a^2=0.13) although results using the average flow per service connection show a good adjustment with an average r_a^2 of 0.62. In other words, an average of 62% of the average flow per service connection is explained by the coefficients in Table 32. Generally, average flows per service connection are higher for buildings with 1-2 floors and consumption in the 4th level (all studied scenarios).

Since buildings with 1-2 floors are more likely to have outdoor uses, such as irrigation and pool filling (Arbués *et al.*, 2003), they are also more likely to have higher consumptions (*i.e.* consumption in the 3rd and 4th level), which explains the dependence of the higher average flows on these factors. Loureiro (2010) also found that consumption level, mainly in the 3rd and 4th level, had a positive influence in the average flow.

Since the average consumption is higher for higher consumption levels, and water consumption generally increases in the summer due to outdoor uses (Aksela and Aksela, 2010; Hof and Schmitt, 2011) developing a seasonal pricing scheme would probably promote a more efficient use of water.

Concerning electricity consumption, McLoughlin et al. (2012) found that apartments had significantly lower total electricity consumption than all other dwelling types, a result of their smaller size and fewer number of occupants.

Table 32 – Coefficients resulting from the MLR analysis for Average flow

Dependent variable	Scenario	Explaining factor	Symbol	Regression coefficient	Estimation	SD	*p*-value	r_a^2	$\overline{r_a^2}$
Average flow per client [l/cl·day]	Winter	Constant		β_0	7.77	29.110	0.16793	0.11	0.13
		Building structure		β_1	-0.47	33.597			
		Pipe size		β_2	0.01	37.011			
	Winter working day	Constant		β_0	220.05	29.193	0.14678	0.13	
		Building structure		β_1	-0.49	33.692			
		Pipe size		β_2	-0.01	37.117			
	Winter Sunday and holiday	Constant		β_0	211.47	26.700	0.12920	0.15	
		Building structure		β_1	-0.49	30.815			
		Pipe size		β_2	0.03	33.947			
Average flow per service connection [l/sc·day]	Winter working day	Constant		β_0	821.00	362.142	0.00333	0.61	0.62
		Building structure		β_1	-0.71	351.593			
		Pipe size		β_2	-0.04	445.587			
		Consumption level		β_3	0.31	183.929			
	Winter Saturday	Constant		β_0	730.18	127.045	0.00021	0.66	
		Building structure		β_1	-0.69	164.280			
		Consumption level		β_2	0.33	140.816			
	Winter Sunday and holiday	Constant		β_0	707.54	125.710	0.00024	0.65	
		Building structure		β_1	-0.68	162.553			
		Consumption level		β_2	0.34	139.336			
	Summer	Constant		β_0	821.46	369.162		0.62	

	working day	Building structure		β_1	-0.72	358.409	0.00332		
		Consumption level		β_2	0.31	187.495			
		Pipe size		β_3	-0.04	454.226			
	Summer Saturday	Constant		β_0	330.76	330.76	0.00382	0.61	
		Building structure		β_1	-0.68	321.12			
		Consumption level		β_2	0.33	167.99			
		Pipe size		β_3	-0.01	406.97			
	Summer Sunday and holiday	Constant		β_0	826.59	343.37	0.00450	0.59	
		Building structure		β_1	-0.65	333.37			
		Consumption level		β_2	0.33	174.40			
		Pipe size		β_3	0.01	422.489			

E.g. *The average flow per client (Ave/cl) for the winter season can be given as:*

***Ave/cl* = 7.77 + -0.47 + 0.01**

The empiric relation above means that average flows per client for the specified scenario are higher for buildings with 1-2 floors (see indexes in the negative vertical axis of ***Figure 56)*** *and pipes with diameters above 110 mm (see indexes in the positive horizontal axis of* ***Figure 58****).*

Night flow

In what concerns night flow consumption, results show a good adjustment with an average r_a^2 of 0.76, which means that 76% of the night flow per service connection is explained by the coefficients in Table 33. For the annual scenario, night flows per service connection are higher for buildings with 1-2 floors and consumption in the 4th level (all studied scenarios). For the studied summer scenarios (summer, summer working day and summer Saturday), besides the referred factors, night flows tend to be higher for small families and families with people above 65 years. This is probably due to the fact that elder people have a more regular consumption during the year, when comparing to younger consumers that are more likely to spend less time at home(Mamade *et al.*, 2012).

Table 33 – Coefficients resulting from the MLR analysis for Night flow

Dependent variable	Scenario	Explaining factor	Symbol	Regression coefficient	Estimation	SD	p-value	r_a^2	$\overline{r_a^2}$
Night consumption per service connection [l/(cl·h)]	**Annual**	Constant		β_0	10.04	1.72	0.00029	0.64	0.76
		Building structure		β_1	-0.68	2.22			
		Consumption level		β_2	0.34	1.90			
	Summer	Constant		β_0	9.41	2.28	0.00013	0.79	
		Building structure		β_1	0.42	1.83			
		Pipe size		β_2	-0.50	2.80			
		Consumption level		β_3	0.33	2.18			
	Summer working day	Constant		β_0	9.13	2.21	0.00012	0.80	
		Family structure		β_1	0.41	1.77			
		Building structure		β_2	-0.51	2.71			
		Consumption level		β_3	0.34	2.11			
	Summer Saturday	Constant		β_0	8.54	2.07	0.00009	0.80	
		Family structure		β_1	0.44	1.66			
		Building structure		β_2	-0.47	2.54			
		Consumption level		β_3	0.35	1.98			

E.g. *The night consumption per service connection (Nig/sc) for the annual scenario can be given as:*

Nig/sc = 10.04 + (- 0.68) + 0.34

The empiric relation above means that the night consumptions per service connection for the specified scenario are higher for buildings with 1-2 floors (see indexes in the negative vertical axis of ***Figure 56****) and monthly consumptions above 25 m****3*** *(see indexes in the positive horizontal axis of* ***Figure 57****)*

7.3.4 Consumption patterns

Daily consumption patterns determined in Chapter 7 for working days, Saturdays and Sundays and bank holidays for the winter and summer seasons were grouped using a cluster analysis, which allowed the distinction of 2 pattern groups for the winter working day scenario and 3 pattern groups for the remainder scenarios.

An attempt has been made to use MLR techniques to analyse daily consumption patterns, but no relevant results were obtained. Therefore, patterns were analysed using a correlation matrix. Other techniques such as Decision trees could have been explored.

Taking into account the results of Loureiro, 2010, the correlation analysis was carried out focusing on socio-demographic indexes. For the current chapter´s purposes, the average of each pattern group was calculated in order to simplify the analysis.

The analysis of the correlation matrix revealed two different types of profile: Young families (type 1). Elder families (type 2). In the light of the results, the following interpretation can be given:

- *Young families* (type 1) – This profile includes medium families with 3-4 elements who usually live in buildings with 5 or more floors and whose elements spend most of the time outside their homes (*e.g.* work or study outside their municipalities).
- *Elder families* (type 2) – This profile includes small families with 1-2 elements, typically people with more than 65 years or inactive workers who usually live in buildings with 1-2 floors and spend most of the time at home.

Each profile will be analysed according to its behaviour in three types of days: working days, Saturdays and Sundays and bank holidays. For each case, differences and similarities between the winter and summer season will be pinpointed and a characterization shall be given with reference to the patterns’ behaviours in four instants of the day: consumption in the night period (from 00h00 to 06h00), consumption in the morning peak, consumption during the day (from 15h00 to 18h30) and consumption in the night peak.

Working days

Concerning working days, consumption in the night period (from 00h00 to 06h00) is very similar for both young (Figure 59 a) and elder families (Figure 59 b), with an average consumption factor of 0.4. However, for young families a peak with a consumption factor of 1.2 is observed at 00h30. In terms of the consumption in the morning peak, young families have a morning consumption peak from 07h30 to 08h30 with a consumption factor of 2.0. On the other side, the consumption pattern for elder families does not show a morning peak, but instead has a flat consumption level from 07h15 to 13h15 with a factor of 1.5.

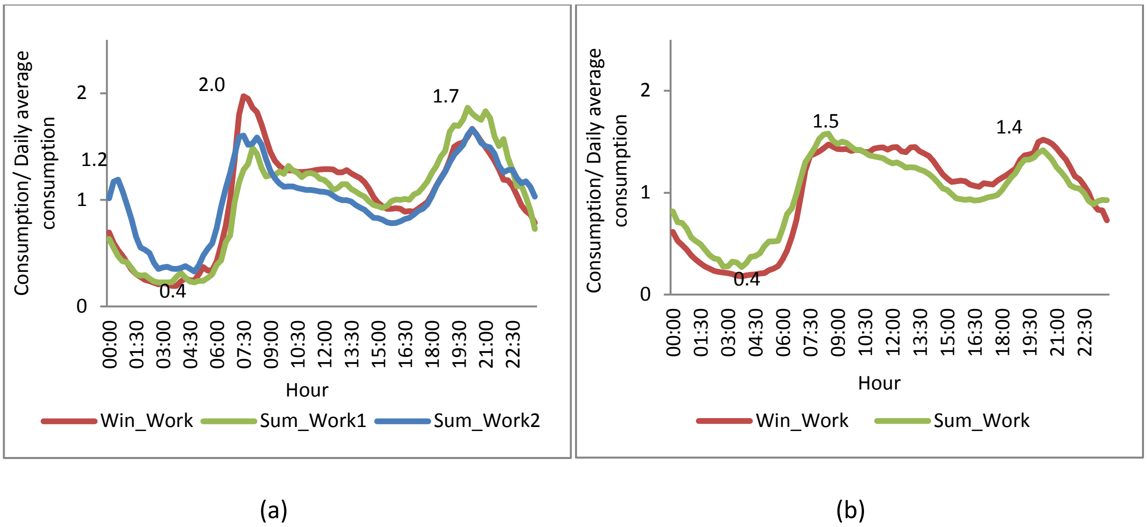


Figure 59 – Daily consumption pattern for working days: a) young families b) elder families

As for consumption during the day (from 15h00 to 18h30), both patterns indicate similar consumption, although consumption in the elder is slightly higher. McLoughlin *et al.* (2012) also found in their study that elder families (above 36 years) had higher electricity consumption patterns during the day, which could be attributed to an increased occupancy at home for longer periods of the day during the day, when comparing to younger families. Finally, concerning the night peak, the maximum consumption is registered approximately at 20h15 with an average consumption factor of 1.7 for young families and 1.4 for elder families. There are no significant differences between winter and summer scenarios, apart from the fact that the morning peak is higher during the winter season for young families and minimum flows in the night period are lower in the same season for elder families.

Saturdays

Concerning Saturdays, consumption in the night period (from 00h00 to 06h00) is very similar for both young (Figure 60 a) and elder families (Figure 60 b), with an average consumption factor of 0.4. Nevertheless, Sum_Sat2 pattern (young families) stands out for having an average consumption factor of 0.7 and a peak with a consumption factor of 1.3 at 00h15. In terms of the consumption in the morning peak, Win_Sat1 (young families) does not show a morning peak but a steady consumption from 11h15 to 18h15 with a consumption factor of 1.7. For the remainder patterns for young families, Sum_Sat1 has a morning peak at 11h00 with a consumption factor of 1.5; Win_Sat2 registers a peak at 11h15 with a factor of 1.6, whereas Sum_Sat2 has two morning peaks: one at 06h45 and the other at 08h00 with consumption factors of 1.3 and 1.2, respectively. Since these latter peaks happen very early for a Saturday pattern, they might not be related with families but with network operations instead. As for elder families, flat consumption levels occur from 09h00 to 14h30 with a factor of 1.7 in the winter season and a factor of 1.5 in the summer season.

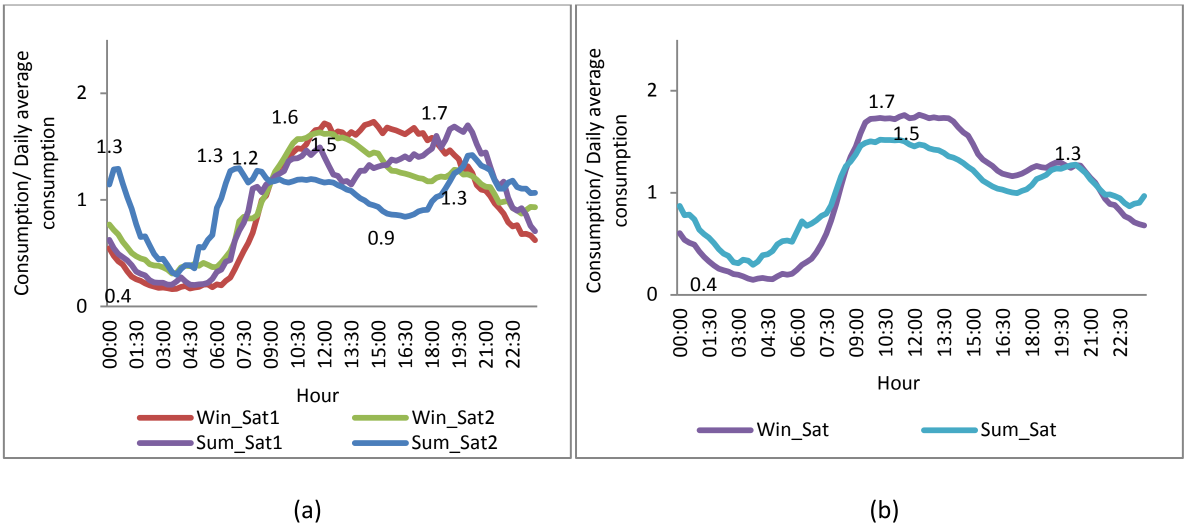


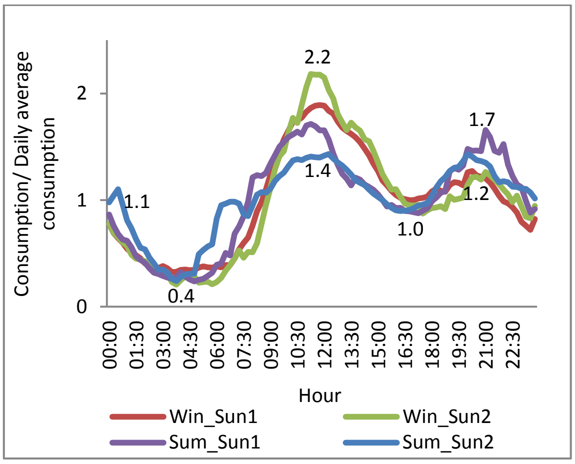
Figure 60 – Daily consumption pattern for Saturdays: a) young families b) elder families

Concerning consumption during the day (from 15h00 to 18h30), young families have a consumption factor that varies from 0.9 to 1.7, whereas elder ones have lower consumptions (average factor of 1.2). This difference might be due to the fact that since young families wake up later, their water usage is delayed. Additionally, it can also be related with the fact that, since young families probably work during the week, they might use more water in the weekends for household purposes such as washing clothes. As for the night peak, the maximum consumption is registered from 19h15 to 20h15, with consumption factors between 1.3 and 1.7 for young families. Elder families have a consumption factor of 1.3 in the night peak. Comparing to working days' behaviour, not only is the peak time extended, but it also happens earlier and with lower consumption factors. Since this peak is usually associated with the dinner meal, the reduction in the peaking factors can be due to families going out for dinner.

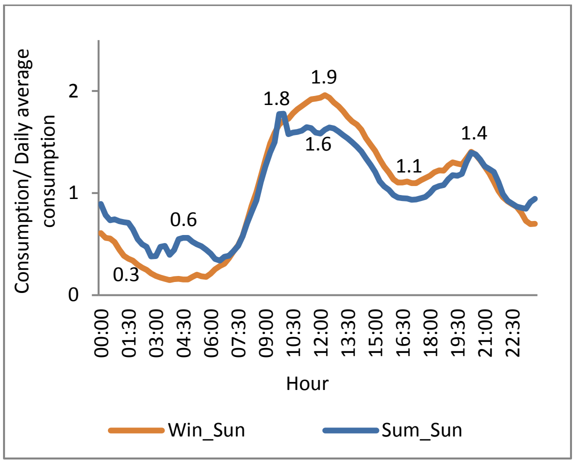
Comparing the winter and summer scenarios, summer scenarios are very variable in young families, as depicted in Figure 60 a. As for the elder, once again minimum flows in the night period are lower for the winter season and the morning consumption peak which in this particular case was described as a flat consumption is higher for the same season, when comparing with the summer season.

Sundays and holidays

Concerning Sundays and bank holidays, consumption in the night period (from 00h00 to 06h00) has an average consumption factor of 0.4 for young families (Figure 61 a). However, Sum_Sun2 pattern (young families) has a peak with a consumption factor of 1.1, registered at 00h15. For elder families (Figure 61 b), the average night consumption factor ranges from 0.3 in the winter and 0.6 in the summer. In terms of the consumption in the morning peak, young families have a peak at 11h15 that varies form 1.4 in the summer to 2.2 in the winter. As for elder families, the morning peak occurs at 09h45 and has a consumption factor of 1.8 in the summer whereas in the winter the peak occurs at 11h30 and has a factor of 1.9. Notwithstanding the fact that there is a morning peak in the summer for elder families, right after the peak, a flat consumption is registered from 10h00 to 12h30, with a factor of 1.6.



(a)



(b)

Figure 61 – Daily consumption pattern for Sundays and holidays: a) young families b) elder families

With respect to the consumption during the day (from 15h00 to 18h30), consumption factors are very similar, with an average of 1.00 to young families and 1.1 to elder ones. Concerning the night peak, the maximum consumption is registered from 20h00 to 21h00 for young families, varying from 1.2 to 1.7, respectively in the winter and summer season. On the other side, for elder families, night peaks occur from 20h00 to 20h15 to elder families with a consumption factor of 1.4 for both seasons.

In this case, it is clear for both consumer types that night consumptions are higher in the summer season and morning peaks and consumption during the day are higher in the winter season.

7.4 Summary and conclusions

In this chapter, domestic water profiling has been carried out concerning to two fields: consumption variables (quantitative variables) and daily consumption patterns (qualitative variables).

The profile of consumption variables is only possible after the reduction of socio-demography, billing and infrastructure indexes into factors (step 1 from the methodology) – which was carried out using Principal factors analysis for 100 DMAs. A correlation analysis is then carried out setting the calculated factors as independent variables and the consumption variables calculated in Chapter 7 for 17 DMAs (a subset of the 100 DMAs) as dependent variables. After analysing which factors were more correlated with the consumption variables, a multiple linear regression was carried out for dependent variables that were correlated with more than one factor. In the light of the results of both analyses, the main conclusions are:

- Instantaneous peak factors are higher for recent pipes, and in some cases, higher diameters and monthly consumptions in the 1st and 2nd level (*i.e.* below 15 m^3).
- Average flow per service connection is higher for buildings with 1-2 floors and in some cases for higher diameters and monthly consumptions in the 1st and 2nd level.
- Night flow per service connection is generally higher for buildings with 1-2 floors. For the summer season in particular, night flow is also higher for families with 3-4 elements, monthly consumptions in the 4th a level (*i.e.* above 25 m^3).
- Minimum flow per service connection client is higher for buildings with 1-2 floors.
- Consumption variables per service connection give better correlation results

Consumption profiling of daily patterns is made taking into account the results of Loureiro (2010) which showed that the daily average consumption behaviour is strictly linked with socio-demographic characteristics. Hence, a correlation matrix based on socio-demographic indexes was calculated and allowed the distinction of two types of consumers: young consumers (type 1) and elder consumers (type 2). Patterns were then analysed in terms of the consumer type, day type (*i.e.* working day, Saturdays and Sundays and holidays) and season (*i.e.* winter and summer) with the aim of profiling consumption in four instants of the day: consumption in the night period (from 00h00 to 06h00), consumption in the morning peak, consumption during the day (from 15h00 to 18h30) and consumption in the night peak. In the light of the results the main conclusions are:

- Instead of morning consumption peaks, elder consumers usually have periods of constant consumption generally lasting 4 hours.
- Young consumers have higher consumption factors in the morning and night peaks, whereas elder consumers have higher consumption factors during the day.
- Elder consumers show a very regular behavior in all patterns contrarily to young consumers' behavior, especially in the Saturday.
- Morning peaks usually occur later in Saturdays, Sundays and holidays, whereas the night peak is less variable with the day type, for all studied scenarios.
- Summer patterns for elder consumers have higher consumptions in the night period and lower consumptions in the morning periods of constant flow, when comparing with the winter season.

The water consumption profiling developed in the present chapter is innovative when comparing to the research work of Loureiro (2010) in the aspects presented in Table 34.

Table 34 – Comparison between the current research work and the one developed by Loureiro (2010) in terms of the water consumption profiling

	Current research study	Previous research works (Loureiro, 2010)
Number of DMAs used in Data reduction – Step 1 of the methodology	Nearly 100	22
Usage of KMO parameter to control the Step 1 results 'quality	Yes	No
Pattern differences between winter and summer – Step 2 of the methodology	Yes	No
Pattern differences between day types (*i.e.* working days, Saturdays, Sundays and holidays)– Step 2 of the methodology	Yes	No
Pattern differences in night consumption– Step 2 of the methodology	Yes	No

8. CONCLUSIONS

The goal of the current work was to develop a spatial and temporal forecasting for profiling consumption patterns in water distribution systems. This goal was successfully achieved and significant improvements have been made in terms of:

- **The geoprocessing tool for socio-demographic analysis:** two new weighting methods were developed and the direct calculation of an expanded group of socio-demographic indexes, using recently published census data, was included. The new weighting methods are considerably better than the original one, since they take into account the number of clients or its consumption, instead of considering a single geometric parameter that does not consider the spatial heterogeneity of consumption.
- **The socio-demographic, billing and infrastructure framework of nearly 100 DMAs:** this database is composed by nearly 49 indexes from DMAs belonging to six different Portuguese water distribution systems distributed along the country. This framework is ready to be available online and to be complemented with additional data about other DMAs.
- **The outlier detection carried out for cleaning the flow time series:** a new method for automatic detection using robust statistics was evaluated. The method allowed the detection of the most relevant types of outlying events. Also, an attempt has been made to classify outlying events and compare with work orders. Although this was less fruitful, it alerted water utilities for the importance of properly classifying this types of events and ensure its connection with work orders.
- **The scenario building for the consumption profiling:** seasonal (*i.e.* Winter and Summer) and daily scenarios (*i.e.* working days, Saturdays and Sundays and holidays) were identified to describe behavioural changes of water consumption throughout time (temporal forecasting component).
- **The calculation of empiric relations for network flows:** empiric relations for estimating design parameters (*e.g.* peaking factors, average consumption) and operational parameters (*e.g.* minimum night flow) were obtained using extensive data from DMAs. This spatial consumption forecasting component allows a substantial reduction in the uncertainty in planning and operation of water distribution systems.
- **The estimation of daily consumption patterns:** two different consumption profiles have been identified (*i.e.* young and elder families) and daily consumption patterns are presented and discussed for each profile and each scenario. The spatial forecasting of daily consumption patterns is a key instrument to improve the operation of water distribution systems.

Furthermore, the methodology used for water consumption profiling consolidates and extends the work developed by Loureiro (2010), since:

- **It uses extensive measurements:** Loureiro (2010) used data from 22 DMAs and this work used data from 100 DMAs, approximately.
- **The PCA has proven to be effective in variable reduction:** 49 indexes concerning the socio-demographic, billing and infrastructure analysis were reduced in only 6 factors that explain more than 80% of data variability and are still an adequate sample for analysis.
- **The water consumption profiling allowed the estimation of empiric relations and daily consumption patterns:** empiric relations were obtained using multiple linear regression analysis with better quality adjustments than Loureiro (2010) in most cases. The results obtained are coherent with Loureiro (2010) and with other research studies.

Future developments should include:

- Testing the profiling in other DMAs with flow data to evaluate the accuracy of the empiric relations obtained.
- Testing new forecasting techniques to the analysis, as explored in the state-of-the-art review (*e.g.* artificial neural networks).
- Adding new water consumption drivers to the analysis, as explored in the state-of-the art-review (*e.g.* concerning climate, income, immigration).
- Adding pressure data to improve the outlier classification analysis.
- Collecting energy data, integrate it in the analysis and compare with previous works.

REFERENCES

Abdallah, A. and Rosenberg, D. (2012) *Heterogeneous Residential Water and Energy Linkages and Implications for Conservation and Management*, Journal of Water Resources Planning and Management.

Adamowski, J. F. (2008) *Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks*, Journal of Water Resources Planning and Management, 134(2), 119-128.

Agency, I. E. (2004) *Oil Crises and Climate Challenges: 30 Years of Energy Use in IEA Countries*, OECD Publishing,

Águas, M. (2013) *Smart Meter – Poupança de Electricidade nas Famílias*, Lisboa e-nova.

Aksela, K. and Aksela, M. (2010) *Demand Estimation with Automated Meter Reading in a Distribution Network*, Journal of Water Resources Planning and Management, 137(5), 456-467.

Alegre, H., Machado, P., Craveiro, P. and Coelho, S. T. (1992) *Caracterização dos consumos domésticos de água na cidade de Lisboa*, Relatório 197/92, Laboratório Nacional de Engenharia Civil, Lisboa.

Arbués, F., García-Valiñas, M. A. and Martínez-Espiñeira, R. (2003) *Estimation of residential water demand: a state-of-the-art review*, Journal of Socio-Economics, 32(1), 81 -102.

Bazilian, M., Rogner, H., Howells, M., Hermann, S., Arent, D., Gielen, D., Steduto, P., Mueller, A., Komor, P. and Tol, R. S. (2011) *Considering the energy, water and food nexus: Towards an integrated modelling approach*, Energy Policy.

Beal, C. and Stewart, R. A. (2011) *South East Queensland Residential End Use Study: Final Report*, Griffith University.

Beal, C., Stewart, R. A. and Huang, A. (2010) *South East Queensland Residential End Use Study: Baseline Results - Winter 2010*, Griffith University.

Beattie, B. R. and Foster, H. S. (1980) *Can prices tame the inflation tiger?*, J. American Works Association, 441-445.

Bernstein, R. and Madlener, R. (2011) *Responsiveness of residential electricity demand in OECD countries : a panel cointegation and causality analysis*.

Brothers, K. (2011) *Large district metered areas: the right option for you?*, Water 21, October 2011.

Chen, H. and Yang, Z. F. (2009) *Residential water demand model under block rate pricing: A case study of Beijing, China*, Communications in Nonlinear Science and Numerical Simulation, 14(5), 2462-2468.

Daioglou, V., van Ruijven, B. J. and van Vuuren, D. P. (2012) *Model projections for household energy use in developing countries*, Energy, 37(1), 601-615.

Donkor, E., Mazzuchi, T., Soyer, R. and Roberson, J. (2012) *Urban Water Demand Forecasting: A Review of Methods and Models*, Journal of Water Resources Planning and Management.

Energy, U. S. D. o. (2006) *Energy Demands on Water Resources Report to Congress on the Interdependency of Energy and Water*.

Eurostat. (2007) *Water use by supply category and user*, Eurostat.

Farley, M. and Trow, S. (2003) *Losses in water distribution networks. A practitioner's guide to assessment, monitoring and control*, IWA Publishing, Londres, ISBN 1900 222 116.

Grafton, R. Q., Ward, M. B., To, H. and Kompas, T. (2011) *Determinants of residential water consumption: Evidence and analysis from a 10-country household survey*, Water Resources Research, 47(8).

Grubbs, F. (1969) *Procedures for Detecting Outlying Observations in Samples*, Technometrics, 11(1), 1-21.

Hart, M. and de Dear, R. (2004) *Weather sensitivity in household appliance energy end-use*, Energy and Buildings, 36(2), 161-174.

Herrera, M., Torgo, L., Izquierdo, J. and Pérez-García, R. (2010) *Predictive models for forecasting hourly urban water demand*, Journal of Hydrology, 387(1–2), 141-150.

Hof, A. and Schmitt, T. (2011) *Urban and tourist land use patterns and water consumption: Evidence from Mallorca, Balearic Islands*, Land Use Policy, 28(4), 792-804.

INE. (2012) *Censos 2011 Resultados Definitivos - Portugal*, Instituto Nacional de Estatística, Lisboa, 978-989-25-0181-9.

IRAR (2009) *Formação de tarifários aplicáveis aos utilizadores finais dos serviços públicos de abastecimento de água para consumo humano, de saneamento de águas residuais urbanas e de gestão de resíduos urbanos*, http://www.ersar.pt

Jankovic´-Nisic´, B., Maksimovic´, C., Butler, B. and Graham, N. J. D. (2004) *Use of flow meters for managing water supply networks*, Journal of Water Resources Planning and Management, 130 (2), 171 - 179.

Kim, J., Lee, D. and Jung, N. (2006) *Analysis of household water demand patterns by meters recording*, 8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA, August 27-30, 2006.

Lambert, A. and Fantozzi, M. (2010) *Recent Developments in Pressure Management*, IWA Specialized Water Loss Conference, IWA, Sao Paulo, Brazil.

Lambert, A. and Hirner, W. (2000) *Losses from water supply systems: standard terminology and recomended performance measurements*, (Assessed in 18-10-2011).

Loureiro, D. (2010) *Consumption analysis methodologies for the efficient management of water distribution systems (in portuguese)*, Technical University of Lisbon, Lisbon, Portugal.

Loureiro, D., Amado, C., Martins, A., Coelho, S. T. and Mamade, A. (2013) *Outlier detection in water distribution systems – a simple approach (in preparation)*

Makki, A. A., Stewart, R. A., Panuwatwanich, K. and Beal, C. (2011) *Revealing the determinants of shower water end use consumption: enabling better targeted urban water conservation strategies*, Journal of Cleaner Production(0).

Mamade, A., Loureiro, D., Rebelo, M., Santos, A., Covas, D. and Coelho, S. T. (2012) *Análise sociodemográfica de setores de sistemas de distribuição de água: Uma abordagem sistematizada para caracterização de consumos domésticos*, Encontro de Engenharia Sanitária e Ambiental (ENaSB), Évora.

March Corbella, H. and Saurí i Pujol, D. (2009) *What lies behind domestic water use?: a review essay on the drivers of domestic water consumption*, Boletín de la Asociación de Geógrafos Españoles(50), 297-314.

March, H., Perarnau, J. and Saurí, D. (2010) *Exploring the Links between Immigration, Ageing and Domestic Water Consumption: The Case of the Metropolitan Area of Barcelona*, Regional Studies, 46(2), 229-244.

McLoughlin, F., Duffy, A. and Conlon, M. (2012) *Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study*, Energy and Buildings, 48(0), 240-248.

McPherson, T. N. and Witkowski, M. (2005) *Modeling Urban Water Demand within a GIS Using a Population Mobility Model*, Impacts of Global Climate Change, 1-8.

Morrison, J. (2004) *Managing leakage by District Metered Areas: a practical approach*, Water 21, February 2004.

Mounce, S. R., Boxall, J. B. and Machell, J. (2010) *Development and Verification of an Online Artificial Intelligence System for Detection of Bursts and Other Abnormal Flows*, Journal of Water Resources Planning and Management-Asce, 136(3), 309-318.

Mutikanga, H. E., Vairavamoorthy, K., Sharma, S. K. and Akita, C. S. (2011) *Operational Tools for Decision Support in Leakage Control*, Water Practice & Technology, 6(3).

OECD. (1999) *Household water pricing in OECD countries*, ENV/EPOC/GEEI(98)12/FINAL, OECD, Paris, unclassified document ENV/EPOC/GEEI(98)12/FINAL, Paris.

Olmstead, S. M., Hanemann, W. M. and Stavins, R. N. (2003) *Does price structure matter? Household water demand under increasing-block and uniform prices*, New Haven: School of Forestry and Environmental Studies, Yale University, Working Paper.

Palau, C. V., Arregui, F. J. and Carlos, M. (2012) *Burst Detection in Water Networks Using Principal Component Analysis*, Journal of Water Resources Planning and Management-Asce, 138(1), 47-54.

Pestana, M. H. and Gageiro, J. N. (2003) *Análise de dados para ciências sociais - a complementaridade ao SPSS*, Edições Sílabo, Lisboa, 972-618-297-2.

Polebitski, A. S. and Palmer, R. N. (2010) *Seasonal Residential Water Demand Forecasting for Census Tracts*, Journal of Water Resources Planning and Management, 136(1), 27-36.

Rebelo, M., Loureiro, D., Santos, A., Coelho, S. T., Alegre, H. and Machado, P. (2008) *Caracterização de sectores de rede em sistemas de distribuição de água - o contributo das ciências sociais apoiado num sistema de informação geográfica*, 13º Encontro Nacional de Saneamento Básico, 14-17 de Outubro, Covilhã.

Rousseeuw, P. J. and Croux, C. (1993) *Alternatives to the median absolute deviation*, Journal of the American Statistical Association, 88(424), 1273-1283.

Ruijs, A., Zimmermann, A. and van den Berg, M. (2008) *Demand and distributional effects of water pricing policies*, Ecological Economics, 66(2–3), 506-516.

Santamouris, M., Kapsis, K., Korres, D., Livada, I., Pavlou, C. and Assimakopoulos, M. N. (2007) *On the relation between the energy and social characteristics of the residential sector*, Energy and Buildings, 39(8), 893-905.

Schleich, J. and Hillenbrand, T. (2009) *Determinants of residential water demand in Germany*, Ecological Economics, 68, 1756-1769.

Shandas, V. and Hossein Parandvash, G. (2010) *Integrating urban form and demographics in water-demand management: an empirical case study of Portland, Oregon*, Environment and planning. B, Planning & design, 37(1), 112.

Shi, G., Zheng, X. and Song, F. (2012) *Estimating Elasticity for Residential Electricity Demand in China*, The Scientific World Journal, 2012, 6.

Stillwell, A. S. and Webber, M. E. (2010) *Water Conservation and Reuse: A Case Study of the Energy-Water Nexus in Texas*, World Environmental and Water Resources Congress.

Suganthi, L. and Samuel, A. A. (2012) *Energy models for demand forecasting—A review*, Renewable and Sustainable Energy Reviews, 16(2), 1223-1240.

Tanverakul, S. A. and Lee, J. (2012) *Historical Review of U.S. Residential Water Demand*, World Environmental and Water Resources Congress 2012, 3122-3136.

UK Water Industry. (1994) *Managing leakage: interpreting measured night flows*, Relatório e, WRc, Reino Unido.

Vidigal, P., Covas, D. and Loureiro, D. (2008) *Análise alargada de parâmetros tecnicohidráulicos e avaliação de desempenho de sistemas de distribuição de água*, Master thesis, Instituto Superior Técnico, Lisbon.

Vieira, P., Moura e Silva, A., Melo Baptista, J., Almeida, M. C. and Ribeiro, R. (2002) *Inquérito aos hábitos de utilização e consumos de água na habitação*, 10º Encontro Nacional de Saneamento Básico, 16 a 19 de Setembro, Braga, Portugal.

White, S., Robinson, J., Cordell, D., Jha, M. and Milne, G. (2007) *Urban Water Demand Forecasting and Demand Management: Research Needs Review and Recommendations*.

Widén, J. and Wäckelgård, E. (2010) *A high-resolution stochastic model of domestic activity patterns and electricity demand*, Applied Energy, 87(6), 1880-1892.

Wilkinson, R. (2007) *Integrating water and energy resource management: progress and opportunities*, World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat.

Yohanis, Y. G., Mondol, J. D., Wright, A. and Norton, B. (2008) *Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use*, Energy and Buildings, 40(6), 1053-1059.

APPENDIX A – SOCIO-DEMOGRAPHIC INDEXES

Purpose:

This appendix describes the socio-demographic indexes calculated along with the data required from Census 2011.

Table 35 – Socio-demographic indexes: buildings

Index [%]	Expression	Census data
Buildings until 1970	$\frac{E1919+E1945+E1960+E1970}{TTEC}$	TTEC[5] – Nº of classic buildings E1919 – Nº of buildings constructed until 1919 E1945 – Nº of buildings constructed between 1920 and 1945 E1960 – Nº of buildings constructed between 1946 and 1960 E1970 – Nº of buildings constructed between 1961 and 1970
Buildings until 1990	$\frac{E1980+E1990}{TTEC}$	E1980 – Nº of buildings constructed between 1971 and 1980 E1990 – Nº of buildings constructed between 1981 and 1990
Buildings until 2000	$\frac{E1995+E2000}{TTEC}$	E1995 – Nº of buildings constructed between 1991 and 1995 E2000 – Nº of buildings constructed between 1996 and 2000
Buildings until 2011	$\frac{E2005+E2011}{TTEC}$	E2005 – Nº of buildings constructed between 2001 and 2005 E2011 – Nº of buildings constructed between 2006 and 2011
Buildings with 1-2 floors	$\frac{PV2}{TTE}$	PV2 – Nº of buildings with 1-2 floors
Buildings with 3-4 floors	$\frac{PV4}{TTE}$	PV4 – Nº of buildings with 3-4 floors
Buildings with ≥ 5 floors	$\frac{PV5}{TTE}$	PV2 – Nº of buildings with ≥ 5 floors

[5] Classic building definition: Permanent construction, with independent access, limited by external walls and built for people (see http://www.ine.pt/bddXplorer/htdocs/minfo.jsp?var_cd=0005970)

Table 36 – Socio-demographic indexes: dwellings

Index [%]	Expression	Census data
Residential immobility	$\frac{AFRH}{AF}$	AFRH – Nº of households used as 1st residence AF – Nº of households
Rented dwellings	$\frac{ACRHARR}{AFRH}$	ACRHARR – Nº of rented households
Vacant dwellings	$\frac{AFV}{AF}$	AFV – Nº of vacant households

Table 37 – Socio-demographic indexes: families

Index [%]	Expression	Census data
Families with adolescents	$\frac{FCPME15}{TTFC}$	TTFC – Nº of classic families[6] FCPME15 – Nº of classic families with dependent children with less than 15 years
Families with elderly	$\frac{FCPMA65}{TTFC}$	FCPMA65 – Nº of classic families with people with ≥ 65 years
Families with unemployed	$1 - \frac{FCD\_0}{TTFC}$	FCD_0 – Nº of classic families with no unemployed
Small families (1-2 elements)	$\frac{FCR1\_2}{TTFC}$	FCR1_2 – Nº of classic families with 1-2 elements
Medium families (3-4 elements)	$\frac{FCR3\_4}{TTFC}$	FCR3_4 – Nº of classic families with 3-4 elements
Large families (≥5 elements)	$1\text{-}\left(\frac{FCR1\_2}{TTFC} + \frac{FCR3\_4}{TTFC}\right)$	-

[6] Classic families definition: aggregation of people who live in the same dwelling and share any familiar bound (or civil union), or any independent person who occupies a dwelling (or a fraction of it).

Table 38 – Socio-demographic indexes: individuals

Index [%]	Expression	Census data
Population above age 65 years	$\frac{HR65 + MR65}{TTR}$	HR65 – Nº of men with ≥ 65 years MR65 – Nº of women with ≥ 65 years
Inactive workers	$\frac{IR\_SAC}{TTR}$	IR_SAC – Nº of residents with no economic activity
University graduates	$\frac{IRQA\_300 + IRQA\_400}{TTR}$	IRQA_300 – Nº of individuals with a post-secondary course IRQA_400 – Nº of university graduates
Economic mobility	$\frac{IR\_ST}{IR\_EP}$	IR_ST – Nº of residents employed in the tertiary sector IR_EP – Nº of residents employed
Active population mobility	$1\text{-}\left(\frac{IR\_TCR}{TTR} + \frac{IR\_ECR}{TTR}\right)$	IRP_TCR – Nº of residents working in their municipality IRP_ECR – Nº of residents studying in their municipality
Population with 12 years of education	$\frac{IRQA\_110 + IRQA\_120 + IRQA\_130 + IRQA\_200}{TTR}$	IRQA_110 – Nº of residents who finished the 1st level of primary school IRQA_120 – Nº of residents who finished the 2nd level of primary school IRQA_130 – Nº of residents who finished the 3rd level of primary school IRQA_200 – Nº of residents who finished secondary school

APPENDIX B – DMA characteristics for outlier analysis

Purpose:

This appendix presents the basic characteristics of DMAs used for the outlier analysis. It also includes data from the socio-demographic analysis.

Table 39 – DMA characteristics for outlier analysis

Number	DMA	Measurement type	Diameter [mm]	Network length [km]	N.º service connections	N.º of clients	N.º of domestic clients	N.º of inhabitants (Census 2011)	Ratio Inh/cl
1	ADE_Bra	Direct	97	59	1632	880	788	2560	3,2
2	FRA_Bra	Direct	89	72	1861	904	817	5789	**7,1**
3	GAM_Bra	Direct	98	95	3698	2379	2105	7955	3,8
4	MOU_Bra	Direct	107	28	772	405	355	1284	3,6
5	SAR_Bra	Direct	82	39	1221	816	718	3329	**4,6**
6	BOI_Por	Direct	78	13	417	407	330	1827	**5,5**
7	CRU_Por	Direct	69	8	190	163	156	470	3,0
8	HOS_Por	Direct	72	20	493	532	494	474	1,0
9	MCI_Por	Direct	73	54	1717	3608	2913	2249	0,8
10	MTA_Por	Direct	77	5	158	210	204	35	0,2
11	REQ_Por	Direct	79	39	1033	688	580	2722	**4,7**
12	VAL_Por	Direct	77	13	330	362	339	612	1,8
13	R01_Por	Direct	113	9	407	270	270	1300	**4,8**
14	R02_Por	Indirect	101	42	1381	964	964	8020	**8,3**
15	R03_Por	Direct	90	59	2601	4361	4346	10812	2,5
16	R3E_Por	Direct	145	9	194	351	121	356	2,9
17	R04_Por	Direct	91	25	843	717	717	5556	**7,7**
18	R05_Por	Indirect	115	16	399	456	455	2732	**6,0**
19	R06_Por	Direct	102	53	1873	2414	2414	6134	2,5
20	R07_Por	Direct	99	36	939	741	740	2491	3,4
21	R08_Por	Direct	73	14	591	782	782	2606	3,3
22	R12_Por	Direct	97	17	608	441	440	689	1,6
23	MFR_Bra	Direct	105	64	2441	771	696	1573	2,3
24	VIL_Bra	Direct	102	41	1130	742	667	2349	3,5
25	CMA_Lis	Direct	107	4	290	2644	2444	277	0,1
26	DEL_Lis	Direct	112	15	662	5185	4514	321	0,1
27	FAB_Lis	Direct	110	13	481	4400	3997	237	0,1
28	MFZ_Lis	Direct	113	6	168	1006	983	676	0,7

29	QJP_Lis	Direct	102	22	643	1529	1528	1045	0,7
30	TFM_Lis	Direct	110	11	335	1789	1658	1637	1,0
31	BSS_Set	Indirect	85	35	3541	4380	4043	12778	3,2
32	FAR_Set	Direct	88	18	2676	2690	2469	6182	2,5
33	SDM_Set	Indirect	75	20	1679	2100	1863	3372	1,8
34	VEN_SET	Indirect	87	21	1893	2124	1941	5151	2,7
35	BZA_Lis	Direct	123	14	422	2118	2031	1149	0,6
36	AZA_Lis	Direct	131	6	250	1652	1592	582	0,4