Contributions to Industrial Statistics

Sara Fontdecaba Rigat

PhD Thesis directed by
Dr. Xavier Tort-Martorell Llabrés

PhD Thesis co-directed by
Dr. Pere Grima Cintas

Department of Statistics and Operations Research
Technical University of Catalonia

Thesis presented in partial fulfillment of the requirements for the
Degree of Doctor by the Technical University of Catalonia

This thesis is presented as a compendium of the published articles:

1. A Methodology to Model Water Demand based on the Identification of Homogenous Client Segments. Application to the City of Barcelona
   (2012, Water Resource Management, 26, 499-516)
   DOI: 10.1007/s11269-011-9928-5

2. An Approach to Disaggregating Total Household Water Consumption into Major End-Uses
   (2013, Water Resource Management, 27, 2155-2177)
   DOI: 10.1007/s11269-013-0281-8

3. Proposal of a Single Critical Value for the Lenth Method
   (2015,Journal Quality Technology & Quantitative Management, 12(1), 41-51)
   "A Tribute to George Box - Statistical Methodologies and Applications"

4. Analyzing DOE with Statistical Software Packages: Controversies and Proposals
   (2014,The American Statistician, 68(3), 205-211)
   DOI: 10.1080/00031305.2014.923784

in accordance with the regulations on official university courses defined by Royal Decree 99/2011 and applied to the doctoral degree courses offered at the Technical University of Catalonia.

Per als meus pares, Jordi i Marta
Per a la meva germana, Eva

Acknowledgments

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I thank my academic advisor, Dr. Xavier Tort-Martorell, a talented teacher and passionate statistician who contributed to a rewarding experience by supporting my own ideas and my research directions. His patience, motivation, enthusiasms and immense knowledge had been the guidance all the time of the research and writing of this thesis.

I wish to sincerely thank Dr.Pere Grima Cintas, not only for his time and extreme patience, but for his helpful contributions to my development as a young researcher. He gave me intellectual freedom in my work, supported my attendance at various conferences and engaged me in new ideas, demanding a high quality of work in all my endeavors.

I must also thank Dr. Shirley Coleman, for the excellent opportunity of my stay to the University of Newcastle. Her insightful comments both in my work and in this thesis had been very generous; specially for her inspirational and timely advice and constant encouragement.

A very special thanks goes out all my Department colleagues, for the stimulating discussions, for their encouragement and for all the fun we have had in the last years.

Finally, but not least, I would also like to thank my family for the support they provided me through my entire life and in particular, I must acknowledge my best friend, Alex, without whose love, encouragement and editing assistance, I would not have finished this thesis. And, of course, thank you both for your constant support through the ups and downs of my academic career. It has been bumpy at times, but your confidence in me has enhanced my ability to get through it all and succeed in the end.

To all of you, I cannot adequately express how thankful I am.

This research has been possible with the financial assistance of Graduate Research Scholarships FI-DGR according to the IUE/2365/2009, of 29th of July (DOGC num.5456 of 02.09.2009).

Abstract

This thesis is about statistics' contributions to industry. It is an article compendium comprising four articles divided in two blocks: (i) two contributions for a water supply company, and (ii) significance of the effects in Design of Experiments.

In the first block, great emphasis is placed on how the research design and statistics can be applied to various real problems that a water company raises and it aims to convince water management companies that statistics can be very useful to improve their services.

The article *A methodology to model water demand based on the identification of homogeneous client segments. Application to the city of Barcelona*, makes a comprehensive review of all the steps carried out for developing a mathematical model to forecast future water demand. It pays attention on how to know more about the influence of socioeconomic factors on customer's consumption in order to detect segments of customers with homogenous habits to objectively explain the behavior of the demand.

The second article -related to water demand management, *An Approach to disaggregating total household water consumption into major end-uses* describes the procedure to assign water consumption to microcomponents (taps, showers, cisterns, washer machines and dishwashers) on the basis of the readings of water consumption of the water meter. The main idea to accomplish this is, to determine which of the devices has caused the consumption, to treat the consumption of each device as a stochastic process.

In the second block of the thesis, a better way to judge the significance of effects in unreplicated factorial experiments is described.

The article *Proposal of a Single Critical Value for the Lenth Method* analyzes the many analytical procedures that have been proposed for identifying significant effects in not replicated two level factorial designs. Many of them are based on the original Lenth Method and explain and try to overcome the problems that it presents. The article proposes a new strategy to choose the critical values to better differentiate the inert from the active factors.

The last article *Analysing DOE with Statistical Software Packages: Controversies and Proposals* review the most important and commonly used in industry statistical software with DOE capabilities: *JMP*, *Minitab*, *SigmaXL*, *StatGraphics* and *Statistica* and evaluates how well they resolve the problem of analyzing the significance of effects in unreplicated factorial designs.

Contents

Glossary of Acronyms

ACF	Auto Correlation Function
ARIMA	Auto Regressive Integrated Moving Average
CART	Classification and Regression Trees
DoE	Design of Experiments
HNPP	Half-Normal Probability Plot
ML	Maximum Likelihood
NPP	Normal Probability Plot
PACF	Partial Auto Correlation Function
PCA	Principal Componenet Analysis
PSE	Pseudo Standard Error
SARIMA	Seasonal Auto Regressive Integrated Moving Average

List of Figures

List of Tables

Part I
Report

Chapter

1 Introduction

This thesis is about statistics' contributions to industry. Industrial statistics can be defined as the practical work involved in collecting, processing, and analyzing industrial data in order to evaluate the implementation of stated plans and to describe the development of industrial production and its economic efficiency. As it is well known, statistics plays a critical role in modern use of technology in science and industry. Statistical concepts and methods are developed and applied in industries for various problems -for example, in order to monitor the quality of products, to plan effective and efficient experiments, to improve standards or to test, analyze and improve the quality of products and services. The increased attention paid to these problems, and accompanying new statistical methodologies, has created an active and valuable new area of research and application: industrial statistics, recently also called statistical engineering where this thesis is focused on.

The thesis is built as an article compendium preceded by this introduction that presents the set of articles, and closed by the conclusions and results derived from them.

Section 1.1 of this introduction establishes the two blocks which the thesis is divided. In Section 1.2 , the objectives are described by article. Finally, Section 1.3 sets out an outline of the thesis and the article compendium.

1.1 Industrial Statistics

1.1.1 Contributions to domestic water demand management

The water industry is a complex one, with multiple actors playing important roles at different phases throughout the water life cycle. At each step along the way, large and growing data volumes could be created by industrial equipment that can be leveraged to improve efficiencies and ultimately develop new services for the consumers.

In the city of Barcelona, the main water company is *Aigües de Barcelona* founded in 1867, a public-private company which manages the entire water cycle, from catchment, water treatment, transport and distribution to sanitation and waste water treatment for their return to nature or reuse. The company offers service to close to 3 million people in the municipalities of the Metropolitan Area of Barcelona. Nowadays, the company is developing a management policy focused on customer proximity; excellence in service delivery, commitment to innovation and developing the talent of its professionals. It also fosters collaborations with other companies, organizations and public authorities to create value and to develop a sustainable model as a strategic axis.

The work presented in the first block of this thesis is an example of these collaborations. In it we propose two statistical contributions in order to develop mathematical models able to predict domestic water demand. The two projects were part of the strategy of the company to understand customer behavior -something not done until then- and to use this information to put customers at the center of its business and create value for them.

Water demand management has become a requirement, set both internally and externally, for water companies; externally by public organizations and internally to provide a cost effective and efficient service. At present, they have a very limited understanding of factors affecting short-term domestic water demand and rely upon crude methods of forecasting. Gaining this understanding and being able to undertake short term forecasting will undoubtedly have several benefits.

The review of approaches to estimating domestic water demand conducted by Beecher (1996) concluded that most estimates have generally relied upon cross-sectional analysis. It is utilized to estimate the quantity of water used by a cross-section of several residential areas at a given time. The data are derived using zona metering studies or the participation of individual household studies. Large proportions of these studies use price as one of the key independent variables affecting demand. In the area treated in the thesis, water demand is unrelated to price because the majority of households are unmetered and, therefore in any given household, the variation in demand is unrelated to the (constant) price.

The opportunities to improve water demand management are plentiful. The block Contributions to domestic water demand management includes two articles that illustrate the potential of using statistics to improve water utilities processes.

1.1.2 Significance of effects in Design of Experiments

An important way to gain knowledge about processes and products in industry is by experimenting. Experiments are also a fundamental part of research. However, conducting experiments in industry is normally expensive and thus they have to be carefully planned taking into account all available knowledge that may be provided by other sources such as historical process operation or people close to the operations. This is normally considered the first step of what is known as sequential strategy: plan small experiments, analyze them to gain knowledge and use the lessons to plan new experiments. This has proved to be the best and more efficient way to gain knowledge in industrial processes.

According to Montgomery, an experiment can be defined as “a test or a series of tests in which purposeful changes are made to the input variables of a process or system so that we may observe and identify the reasons for changes that may be observed in the output response”. The process under experimentation can have one or many responses (Y). The purpose of an experimentation is to measure the effects that the controllable factors (X’s) have on the output response.

Experiments are normally conducted on controlled systems. That is, the important features of the investigated materials, the nature of the studied manipulations of the system and the measurement procedures are all determined by the experimenter. By contrast, in ”observational studies” the investigator does not control all of these features although the objective of the two types of studies may be identical. Hence, experiments make it possible to verify causality between experimental factors and process responses something extremely important in industry and almost impossible to achieve through an observational study.

The costs of an experiment always makes worth the effort to maximize the information output while at the same time minimize the resources required for producing this information. According to Ye and Hamada, Design of Experiments (DoE) can be viewed as:

> ‘‘a body of knowledge and techniques that enables an investigator to conduct better experiments, analyze data efficiently, and make the connections between the conclusions from the analysis and the original objectives of the investigation.”

Consequently, DoE is useful for an experimenter who wants to understand and improve a product or a process and wants to do it effectively and efficiently.

Two of the pioneers of the DoE field were Ronald A. Fisher and Frank Yates, they worked on problems in agriculture and biology at Rothamsted Experimental Station in the 1920s and 1930s (Montgomery [1980]). Some of Fisher’s many contributions of statistical insight into the DoE field were to point out the importance of randomizing the experimental treatments, of blocking them to achieve better homogeneity, the use of the analysis of variance to judge the significance of effects, and not least factorial designs to achieve separation among them of the studied effects. The essence of factorial designs is that several experimental factors are studied simultaneously instead of one at a time.

Many developments have occurred within the DoE field since the 1930s and the method since its introduction in the industrial world by George Box when he worked at Imperial Chemical industries has become widely used in all types of industries. In fact nowadays, DoE is also used extensively in many areas of science and engineering.

Ye and Hamada, and many other authors, for example, Box Hunter and Hunter, use the concept Experimental design to label the body of knowledge which is also often referred to as Design of Experiments.

DoE contains many statistical methods and therefore knowledge about statistics is a central part of understanding how the methods work. From a theoretical point of view, many efforts have been made in order to determine, when there are no replicates, which of the factors from the experimentation are significant, this is, which ones we are reasonably sure that affect the response. The most common analytical procedure is Lenth's method (Lenth [1989]). The principal advantage of this method is that it uses a simple and sound procedure (given the scarcity of information available) to estimate the effects standard deviation. However, the critical values given by the method originally proposed are not the most appropriate and in many cases produce results that graphical methods (a representation in NPP) and especially simulations show that are clearly incorrect.

The improvement of Lenth's method and how different statistical software address the analysis of DoE and the significance of effects are the principal aspects covered by the two articles of the second block, Significance of effects in Design of Experiments.

1.2 Objectives

The previous section presented a brief introduction of two areas in which the application of industrial statistics is a challenge. The first is to convince water management companies that statistics can be very useful to improve their services and the second to devise a better way to judge the significance of effects in unreplicated factorial experiments. In a more detailed way the objectives are:

Block I: Contributions to domestic water demand management

Article 1, **A Methodology to Model Water Demand based on the Identification of Homogenous Client Segments. Application to the City of Barcelona**, makes a comprehensive review of all the steps carried out for developing a mathematical model to forecast future water demand. The article pays attention on how to know more about the influence of socioeconomic factors on customer's consumption in order to detect segments of customers with homogenous habits to objectively explain the behavior of the demand. The objectives are:

- collect digital data of water uses events in regular customer households and build a unique database containing socio economical and consumption data,
- group customers according to the values of socio economical data,
- model the water demand consumption to forecast short, mediun and long term water demand,
- find the possible relationship between consumption and socio economical information.

Article 2, **An Approach to Disaggregating Total Household Water Consumption into Major End-Uses**, describes the procedure to assign water consumption to microcomponents (taps, showers, cisterns, washer machines and dishwashers) on the basis of the readings of water consumption of the water meter. The main idea to accomplish this is, to determine which of the devices has caused the consumption, to treat the consumption of each device as a stochastic process. The objectives are:

- determine what kind of information can distinguish one appliance from another and characterize them,
- design an algorithm to characterize the water consumption of each device in order to assign automatically the meter consumption to each device.

Block II: Significance of the effects in Design of Experiments

Article 3, **Proposal of a Single Critical Value for the Lenth Method**, analyzes the many analytical procedures that have been proposed for identifying significant effects in not replicated two level factorial designs. Many of them are based on the original Lenth Method and explain and try to overcome the problems that it presents. The article proposes a new strategy to choose the critical values to better differentiate the inert from the active factors. The objectives are:

- to emphasize Box's [1992] well known idea that Experimental Design should be about learning and not about testing and based on it provides a simple approximate, but better than the original method, solution to identify active affects,
- use simulation and comparisons with other methods to propose a simple and a better solution for any 2^k or 2^{k-p}.

Article 4, **Analysing DOE with Statistical Software Packages: Controversies and Proposals** review the most important and commonly used in industry statistical software with DOE capabilities: *JMP*, *Minitab*, *SigmaXL*, *StatGraphics* and *Statistica* and evaluates how well they resolve the problem of analyzing the significance of effects in unreplicated factorial designs. The main objectives are:

- compare methods that are used to run significance tests, and normal probability plots (NPPs) and half normal plots (HNPs), among the five statistical software packages.
- present suggestions (regarding formal significance tests and graphical analysis) to make DOE analysis more effective and easier to understand,
- discuss capabilities and limitations of each statistical software packages.

1.3 Thesis Outline

The results of the present investigation are comprised within four main published articles in several journals. They are divided in two blocks as shown in Section 1.2. It is important to point out that the results of the thesis are presented in the four core articles composing Chapter 2, Chapter3, Chapter 4 and Chapter 5, respectively. The conclusions (Chapter 6) are derived, from the core articles as well, including the global results and future lines of research.

Block I: Contributions to domestic water demand management

- A Methodology to Model Water Demand based on the Identification of Homogenous Client Segments. Application to the City of Barcelona
  (2012, Water Resource Management, 26, 499-516)
  DOI: 10.1007/s11269-011-9928-5

- An Approach to Disaggregating Total Household Water Consumption into Major End-Uses
  (2013, Water Resource Management, 27, 2155-2177)
  DOI: 10.1007/s11269-013-0281-8

Block II: Significance of the effects in Design of Experiments

- Proposal of a Single Critical Value for the Lenth Method
  (2015, An International Journal Quality Technology & Quantitative Management, 12(1), 41-51)
  "A Tribute to George Box - Statistical Methodologies and Applications"

- Analyzing DOE with Statistical Software Packages: Controversies and Proposals
  (2014, The American Statistician, 68(3), 205-211)
  DOI: 10.1080/00031305.2014.923784

Part II
Published Articles

Chapter

2

Article 1

A Methodology to Model Water Demand based on the Identification of Homogenous Client Segments. Application to the City of Barcelona

Sara Fontdecaba, Pere Grima, Lluís Marco, Lourdes Rodero, José A. Sánchez-Espigares, Ignasi Solé, Xavier Tort-Martorell, Dominique Demessence, Víctor Martínez De Pablo, Jordi Zubelzu

Received: 15 July 2010 / Accepted: 29 September 2011 / Published online: 20 October 2011

Abstract. Water management has become a vital concern for both water supply companies and public administrations due to the importance of water for life and current scarcity in many areas. Studies exist that attempt to explain which factors influence water demand. In general, these studies are based on a small sample of consumers and they predict domestic water consumption using ordinary least squares regression models with a small number of socioeconomic variables as predictors, usually: price, population, population density, age, and nationality. We have followed a different approach in two ways; one, in the scope of the study: we have included in the study all consumers of the Barcelona area and as many socioeconomic variables as possible (all the available data from official statistics institutions); and also in the methodology: first, we have segmented clients into homogeneous socioeconomic groups that, as we show later in the Barcelona case, also have homogeneous water consumption habits. This allows for a better understanding of water consumption behaviours and also for better predictions through modeling water consumption in each segment. This is so because the segments' inner variability is smaller than the general one; thus, the models have a smaller residual variance and allow for more accurate forecasts of water consumption. The methodology was applied to the Barcelona metropolitan area, where it was possible to construct a database including both water consumption and socioeconomic information with more

than one million observations. Data quality was a primary concern, and thus a careful exploratory data analysis procedure led to a careful treatment of missing observations and to the detection and correction or removal of anomalies. This has resulted in a stable division of the one million water consumers into 6 homogeneous groups and models for each of the groups. Although the methodology has been developed and applied to the Barcelona area, it is general and thus can be applied to any other region or metropolitan area.

Keywords: Water demand forecast, segmentation, cluster analysis, regression modelling, scenario planning, SARIMA models.

2.1 Introduction

Increasing concerns about the availability of water in adequate quantities and qualities have made more urgent the need to advance towards a sustainable approach to water resources management (Brooks [2006]; Butler and Memon [2006]; Gleick [2003]). In this sense, many international organizations, including the United Nations and the European Union, call for the application of Integrated Water Resources Management (IWRM), which attempts to combine supply and, especially, demand measures (European Commission; ICWE (International Conference on Water and Environment) [1992]). In particular, governments, water companies and consumers must engage in water conservation practices or otherwise expect growing difficulties in the provision of drinkable water, all probably exacerbated by climate change (Corral-Verdugo et al. [2002]; Guy [1996]; Postel [1992]).

In this context, water demand management –which requires an in-depth knowledge of the behaviour of users in relation to consumption (Dziegielewski [1993]; Montini [2006]; Stephenson [1999])– should be a basic component of any water company's strategy (Beecher [1996]). Different behaviours, influenced in turn by a variety of factors, explain for instance the high variations in per capita consumption found between different countries, different regions and different cities and even between different neighbourhoods in the same cities.

The study of residential water demand, its characteristics, and its drivers has received substantial attention in the academic and professional literature (Arbués et al. [2003]; Babel et al. [2007]; Babel and Shinde [2011]; Baumann et al. [1998]; Duke et al. [2002]; Hanke and de Mare [1982]; Kanakoudis [2002]; Opaluch [1982]). Different studies and econometric models show that residential water demand is correlated with: income (Arbués and Villanua [2006]; Arbués et al. [2003]; Baumann et al. [1998]; Renzetti [2002]); population density (Lavière and Lafrance [1999]); age distribution of people in a household (Murdock et al. [1991]); religious and cultural characteristics (Smith and Ali [2006]); the number of people living in a household (Hamilton [1983]; Nauges and Thomas [2003]; Renwick and Green [2000]; Zhang and Brown [2005]); the characteristics of a city or town (Hellegers et al. [2010]; Hasse and Nuiss [2007]; Kahn [2000]; Liu et al. [2003]; temperature and rainfall (Griffin and Chang [1991]) and the presence of water-saving technology in the form of efficient appliances (United States Environmental

Protection Agency (USEPA)).

With all this in mind, the final objective of the project was to define a methodology for forecasting future water demand, one which benefited from already known facts –that is, from the correlations mentioned above. The idea was to segment clients into homogeneous socioeconomic groups through cluster analysis, then check if these groups were also homogeneous with respect to water consumption (as expected, given the extensive literature mentioned in the previous paragraph) and, finally, use these segments to model water consumption and forecast future demand. Because of the segments' homogeneity we expected the models, one for each segment, to have smaller residual variances and thus allow for more accurate predictions. The project was conducted on behalf of and with full support from Aigües de Barcelona, which was interested in learning about the behaviour of their clients and having accurate water consumption forecasts; thus, we had access to water consumption information for over one million Barcelona area residents. In addition we were able to obtain, through official statistics, socioeconomic information for the same population so we could refine and test the devised methodology.

This paper is organized in the following manner: Section 2.2 briefly reviews the scope and data used to carry out the study in the area of Barcelona; Section 2.3 explains the methodology followed in the segmentation and modelling processes and Section 2.4 highlights and discusses the most relevant results obtained in the Barcelona case; and finally, Section 2.5 presents some conclusions.

2.2 Case Study and Data Used

Aigües de Barcelona supplies water to nearly 1,570,000 households belonging to 23 municipalities of the Metropolitan Area of Barcelona.

To carry out the study as outlined, two types of data from two different sources were needed: household water consumption, available from Aigües de Barcelona; and sociological variables, obtainable from official statistical institutions. Obviously, in order to detect different customer profiles and group them by similarity, the study required the use of data that was as disaggregated as possible. The consumption data was available at the household level –Aigües de Barcelona meters the consumption of each household– and the minimum territorial division with official statistics data served as the so called "census tracts". The 23 municipalities included in the study contain a total of 2,358 census tracts; they have a population of between 1,500 and 2,000 people and are designed to be homogeneous with respect to population characteristics, economic status, and living conditions. Their spatial size varies widely, depending on population density, and they are fairly stable along time, which allows for statistical comparisons from census to census.

In order to relate census tract socioeconomic data with water consumption data, all supply pipes (distribution pipes connected to several accounts) were assigned to census tracts –a difficult task, since it had to be done "by hand," based on the addresses; however, it was a task that will be extremely useful not only for this but for many other future studies. Only a negligible number of supply pipes were serving more than one census tract or were impossible to locate.

2.2.1 Socioeconomic Data

After an extensive literature review, the most relevant articles for this purpose are listed in the introduction section. Six factors were identified as significant in explaining domestic water consumption: socio-demographic, behavioural, territorial, cultural, climatic and technological. In addition, different variables were defined for each of these factors as possible drivers that influence domestic water demand (Table 2.1).

Factors influencing water demand	
1) Price	Economic variables
2) Income	
3) Population and population growth	Socio-demographic variables
4) Household size	
5) Age structure of the population	
6) Cultural values	Behavioural variables
7) Population density	Territorial variables
8) Household size and type	
9) Climate	Climatic variables
10) Technology (water saving appliances in the household)	Technological variables

Table 2.1: Factors influencing water demand.

Of those factors, "cultural values" and "technology"[1] were not considered in the study, due to a lack of data for the Metropolitan Area of Barcelona. "Price" was also disregarded because it is the same for the whole area. Finally metering, a variable related to price mentioned in several studies, could not be considered for the same reason, in Barcelona all households are metered.

The available data, relative to the rest of the factors, comes from two sources: the census, conducted every 10 years (the most recent available is from 2001); and variables, collected annually by the Instituto Nacional de Estadística (INE), regarding population structure (age, sex and nationality). The most recent data available for the latter is from 2006. Variables related to the "Climate" factor were collected from meteorological stations.

All the variables have been collected at the census tract level except for "Income", which is only available at a more aggregate level (the municipality, or, for Barcelona, the so-called Zones de Recerca Petites,[2] which are aggregations of census tracts). To increase the precision of the information and to complement the incomplete variable "Income", we introduced the variable "level of studies" into the database, which is highly correlated with income. The following list presented in Table 2.2 shows the 27 variables considered

[1]Partly as a consequence of this study, a survey to gather data on technological variables is scheduled to be conducted.

[2]Small research zones. They are aggregations of somewhat homogeneous census tracts done for the purposes of sociological studies.

in the study, together with a brief description and the units used. They will be referenced in the rest of the paper.

Variables included in the study	
Territorial	
Area: Total surface of the census tract (km^2)	Density: Population (2006) / Area (inhabitants/km^2)
Population: Inhabitants registered in the census tract	
Sex	
%Men: Percentage of men in the census tract	%Women: Percentage of women in the census tract
Age Structure	
%0-14: *Children*. Percentage of the population between 0 and 14 years	%25-65: *Active population*. Percentage of the population between 25 and 64 years
%15-24: *Teenagers*. Percentage of the population between 15 and 24 years	%>65: *Senior population*. Percentage of the population over 64 years
Nationality	
%Spanish: Percentage of Spanish population	%America: Percentage of American population
%Community: Percentage of the population that are Europeans belonging to community countries	%Africa: Percentage of African population
%EU Non community: Percentage of the population that are Europeans belonging to non community countries	%Asia: Percentage of Asian population
Household	
%Principal_Household: Percentage of principal households	%Sec_Or_Empty_Household: Percentage of secondary or empty households. (Secondary households are those destined to be occupied only occasionally, i.e. holiday homes. Empty households are those that remain empty without being occupied.)
%Small: Percentage of small households (under 60 m^2)	%Medium: Percentage of medium households (between 60 m^2 and 90 m^2)
%Big: Percentage of big households (over 90 m^2)	Inhabitants_per_Household: Population / # households
Income	
Income: Per capita Income	
%Without_studies: Percentage of population without studies	%Secondary_studies: Percentage of population with secondary studies
%Primary_studies: Percentage of population with primary studies	%Higher_studies: Percentage of population with higher studies
Population Growth	
Increase_Population: The slope of the linear regression of the population from 2003 to 2006.	

Table 2.2: Socioeconomic variables included in the study which could potentially affect water demand.

2.2.2 Consumption Data

The main source of information about water consumption comes from the Aigües de Barcelona billing department. Every three months the water consumption of each household account is usually recorded for billing purposes. There were three problems to be solved: estimating the monthly water consumption, getting rid of seasonal effects and aggregating the households that belong to the same census tract. The average daily consumption was estimated over the metered three-month period; then the aggregate consumption of households over the 4 year period was calculated so that its longitudinal evolution (seasonality) could be studied and characterized, and its effects eliminated from each household's monthly consumption data; and finally the estimates of each household were aggregated by supply pipe and census tract.

2.2.3 Exploratory Data Analysis and Data Cleaning

A descriptive analysis of the data gave a preliminary idea of the characteristics of census tracts and of consumer behaviour. At the same time it pointed to some outliers. Specifically, 14 census tracts (0.5% of the 2.358 included in the study) were eliminated due to three different reasons: (I) Lack of sociological information: Newly created census tracts that did not exist in the year 2007. (II) Lack of consumption information: Missing data related to the water consumption of the census tract. (III) Atypical information: Atypical sociological information for a census tract in Barcelona.

2.3 Methodology

The methodology followed can be divided into two steps. The first one is aimed at obtaining groups of clients. In the Barcelona case, this means groups of census tracts that have characteristics which are homogenous to the socioeconomic variables that are susceptible to affect water consumption. It is then verified whether or not the groups obtained are also homogeneous in their water consumption. The second step is to find explanatory models and predictive models for the water consumption of each segment. In this section we briefly outline the statistical techniques used in each step.

2.3.1 Segmentation

Cluster analysis (Johnson and Wichern [2002]; Lebart et al. [2006]), also called segmentation analysis, seeks to identify homogeneous groups of cases in a population. That is, its objective is to sort cases into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. It may reveal associations and structures in data which, though not previously evident, are sensible and useful once discovered. The first step in cluster analysis is the establishment of the similarity or distance matrix. This matrix is a table in which both the rows and columns are the units of analysis and the cell entries are a measure of similarity or distance between any pair of cases; we used the most common one, the Euclidean distance.

For this study, a combination of the two basic families of cluster algorithms (hierarchical and partitional) was used. Hierarchical clustering builds a hierarchy of clusters and is usually represented as a tree diagram (called dendogram) that illustrates the arrangement of the clusters, with individual elements at one end and a single cluster containing every element at the other; cutting the dendogram at a given height gives a clustering at a selected precision. Then, two partitional algorithms were used. First, the K-means, which aims to partition all the observations into k clusters, in which each observation belongs to the cluster with the nearest mean. Second, the CART (Classification And Regression Trees), which is a binary recursive partitioning procedure that generates a regression tree.

The steps followed were:

- The socioeconomic variables –omitting consumption– were standardized (subtracted the mean and divided by the standard deviation) in order to give the same weight to all of them, subsequently giving them equal impact on the computation of distances.
- A preliminary classification was obtained using the hierarchical method with the Euclidean distance –because of its interpretability–, and the Ward criterion –in order to achieve a large quantity of groups with a small size. This analysis provided an aggregation tree or dendogram.
- The different classifications (containing different numbers of groups) were used as a starting point for the hierarchical method using the k-means algorithm.
- This two-step cluster analysis procedure was repeated including the consumption variables. The same clusters with only minor changes of census tracts appeared, which means that the groups obtained are not only homogeneous both in their structural composition and in their water consumption behavior, but that they are different from each other as well. Therefore, it is reasonable to develop a model for predicting the water consumption of each group.

2.3.2 Modelization

Once the groups were established, the next step was to model water consumption within each group. Because of group homogeneity, it was anticipated that the models obtained would be very good for prediction purposes (the predicted values would have a low uncertainty). In addition, they would be useful for explaining which variables affect the particular pattern of consumption for each segment.

Two types of models were derived: 1) linear models relating the socioeconomic variables to water consumption useful, as mentioned, for understanding client behavior and rough long-term scenario planning; and 2) predictive time series models (using the SARIMA methodology) for the purposes of accurate short-term predictions.

Linear models (Draper and Smith [1998]; Peña [2010]) allow explanation of the behavior of a numerical variable (Y) based on values of different variables (Xs). The procedure for finding the linear models is detailed below:

- First, the stepwise algorithm (a semi-automated process of building a model by successively adding or removing variables based on the t-statistics of their estimated coefficients) was used, step by step, to find the best model for explaining water consumption.

- Then, the model found through stepwise regression was fitted and the standardized residuals were obtained to study the goodness of fit and the outliers. In general, a model is considered a good one when the residuals: (a) are normally distributed with mean 0, (b) have a constant variance, (c) are independent. In our study, there were no problems with homoscedasticity or independence. Regarding normality, it was expected to have around 5% outliers (data points poorly explained by the model); that is, points having a residual larger than $-2|$. In our case, and because of the amount and nature of the data, we moved the bounds from $-2|$ to $-4|$, and considered outlier points with residuals larger than $-4|$; these points were studied and removed from the model if no good explanation was found for them. Cook's distance was used to estimate the influence of each data point. Points with Cook's distance larger than 1 were studied and removed if appropriate. In total, six data points were removed, a very small number given the size and type of the study.

2.3.3 Time series

The time series models used for short term prediction purposes were developed using the SARIMA methodology (Box et al. [1994]; Peña [2005]). This methodology uses the autocorrelation function to measure the relationship between observations within the water consumption series and to derive predictive models. The idea is to describe how any given observation (x_t) is related to previous observations ($x_{t-1}, x_{t-2}, ...$). This model is then used to forecast the future values of the variable. To define the model for each group, we carried out a three-stage procedure:

- Stage 1: Identification. The estimated Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were calculated and, based on them, the SARIMA model –whose theoretical ACF and PACF most closely resemble the ones estimated from the data– is chosen as the most adequate model.

- Stage 2: Estimation. Via maximum likelihood (ML), we obtained precise estimates of the coefficients of the model chosen in the identification stage.

- Stage 3: Diagnostic Checking. By checking the residuals, we were able to see the suitability of the model fitted for each segment and to detect other effects that influence the behaviour of the series. For example, holiday periods relating to a calendar effect (in our case the Easter period was especially relevant).

Once we have a model for each group, it can be used to make predictions for that group or, by adding the predictions for all the groups, we can obtain a global prediction that is more precise (with less variability) than if we had obtained it from a global SARIMA model. This is so because the groups' inner variability is smaller than the general variability and, thus, the models have a smaller residual variance that allows for more accurate forecasts of water consumption.

2.4 Results and Discussion

This section shows the results of applying the methodology to the Barcelona case (data explained in Section 2.2). The results are presented in three blocks: segmentation, modelling and time series. We dedicate more attention to the segmentation and modelling results than to the short term time series forecasting. In the block on segmentation, we explain the groups of customers with homogeneous water consumption habits found in Barcelona while the block on modelling presents the socioeconomic variables that influence water consumption for each group. More attention is paid to these first two parts because they present new developments: segmentation and the use of segmentation for modelling socioeconomic influences. The third part, on the other hand, merely applies an already well known methodology; the only novelty is its application to homogeneous groups, followed by adding the results so that the total estimate has a smaller variance.

2.4.1 Segmentation of Clients Regarding Water Consumption and Socioeconomic Variables

As stated in Section 2.3, the point of departure was a hierarchical cluster analysis, conducted using the 29 socioeconomic and water consumption variables. The resulting dendogram is shown in Fig. 2.1. If an imaginary horizontal line is drawn in the upper part of Fig. 2.1, and it is slowly moved down along the similarity axes, it will cross a different number of vertical lines. The number of vertical lines indicates the number of clusters (groups) into which the data can be classified for a given similarity (shown on the vertical axis). The main doubt arose when deciding between 6 (dotted line) and 7 (hyphenated line) groups. Finally, 6 groups were chosen, both because of their interpretability and because the CART misclassification error rate was very small (slightly above 15%). Furthermore, the six groups were very easy to explain; they "made sense".

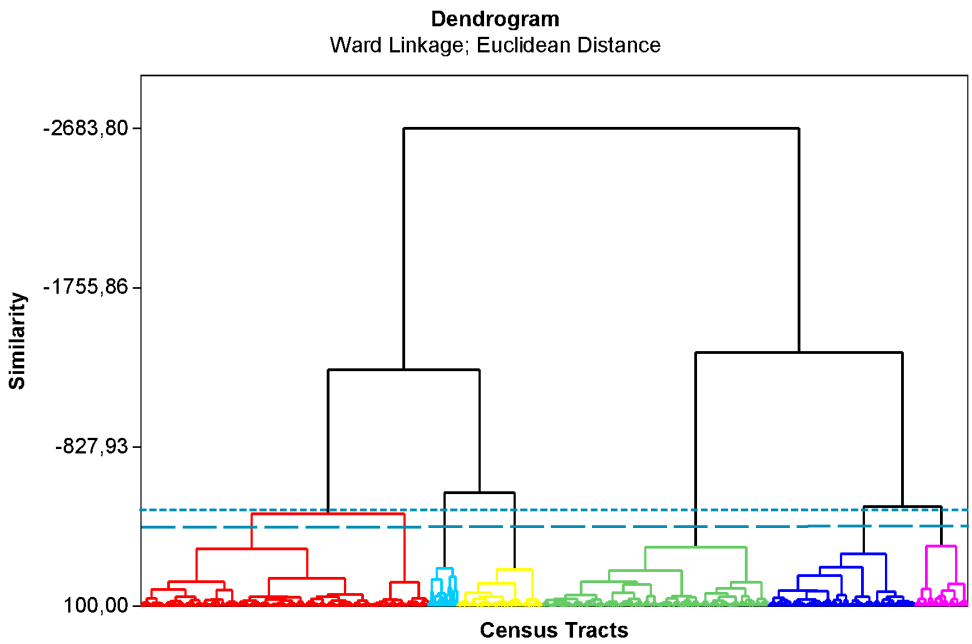


Figure 2.1: Dendogram.

Once the 6 groups were established, we characterized them using two different resources. First, using descriptive statistics of the variables to determine the main differences between groups; and second, painting their position on a spatial map to identify the distribution of each group.

The first step is to use descriptive statistics and graphics to get an initial idea of the characteristics of consumers in each group and the relationships between the socioeconomic variables and water consumption.

Figure 2.2 shows *Boxplots* stratified by group, showing the socioeconomic variables with the biggest influence on group formations. *Boxplots* are a convenient tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between the different groups defined. The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile) and therefore contains the middle half of the scores in the distribution. The median is shown as a line across the box. Therefore 1/4 of the distribution is between this line and the top of the box and 1/4 of the distribution is between this line and the bottom of the box. There are two adjacent values: the largest value below the upper inner fence and the smallest value above the lower inner fence. Outside values, usually considered possible outliers, are indicated by small circles.

The boxplots of figures. 2.2 and 2.3, together with other graphic and data exploratory techniques, allow an interpretation of the socioeconomic characteristics of the groups found. Table 2.3, below, summarizes the main socioeconomic and water consumption traits of each group.

In our case, since groups are tied to a geographical zone, it is also very instructive to paint them on a map. Figure 2.4 shows a map of the Barcelona Metropolitan Area with the six groups in different colors. For those familiar with Barcelona's development history and economic structure, it will be easy to recognize the soundness of the clusters found.

Some general trends can be derived from the group characterization:

- As can be seen on the map, groups are distributed roughly in circles from the center. As mentioned previously, it represents a very reasonable pattern, given the history and physical and socioeconomic structure of Barcelona.
- The middle class groups are the largest in the population (which seems quite logical).
- Groups with lower per capita income are the ones with more immigration and less water consumption.
- Groups with higher per capita income have higher water consumption.
- Group number 5, mostly composed of families with a high level of studies, is quite peculiar. It has very high water consumption (gardens) and the highest decrement, but not a very high per capita income (young people starting their career).

Group 1. Low income, 40% immigration 118 census tracts • The lowest water consumption (93 l pers./day) and the lowest decrement. • The lowest per capita income. • The highest percentage of people without studies or with primary studies. • The highest population density. • The highest percentage of immigration (in order, Asians, Americans and Africans). • The highest percentage of men. • A significant population increment. • 90% small and middle-sized houses.	Group 2. Low income, 20% immigration 411 census tracts • Low water consumption (97 l pers./day) and low decrement. • Low rate of secondary or empty houses. • High population density. • 20% immigration (mainly Americans). • 96% small and middle houses.
Group 3. Lower middle class 634 census tracts • Middle water consumption (105 l pers./day) and decrement in the consumption. • 92% Spanish. • The only group with a decrement in population. • 85% main houses (more than half mid-size).	Group 4. Upper middle class 816 census tracts • Middle water consumption (113 l pers./day). • Quite high per capita income. • High percentage of elderly people. • More women than men. • 40% with secondary or higher studies and only 11% without studies.
Group 5. Young families 73 census tracts • High water consumption (140 l pers./day). • The highest water consumption decrement. • Very low population density and the highest population increment (expanding areas). • The highest percentage of children between 0 and 14 years. • The highest amount of people per household: 4.2. • Low percentage of people without studies. • Mainly middle size and big houses.	Group 6. Wealthy 292 census tracts • The highest water consumption (150 l pers./day). • The second highest water consumption decrement. • The highest per capita income. • High percentage of elderly people. • The highest percentage of women. • Mainly big houses (58%). • The highest percentage of secondary or empty houses. • The highest percentage of people with secondary and higher studies. • Almost no increment in population.

Table 2.3: Summary of the socioeconomic characteristics and water consumption per group.

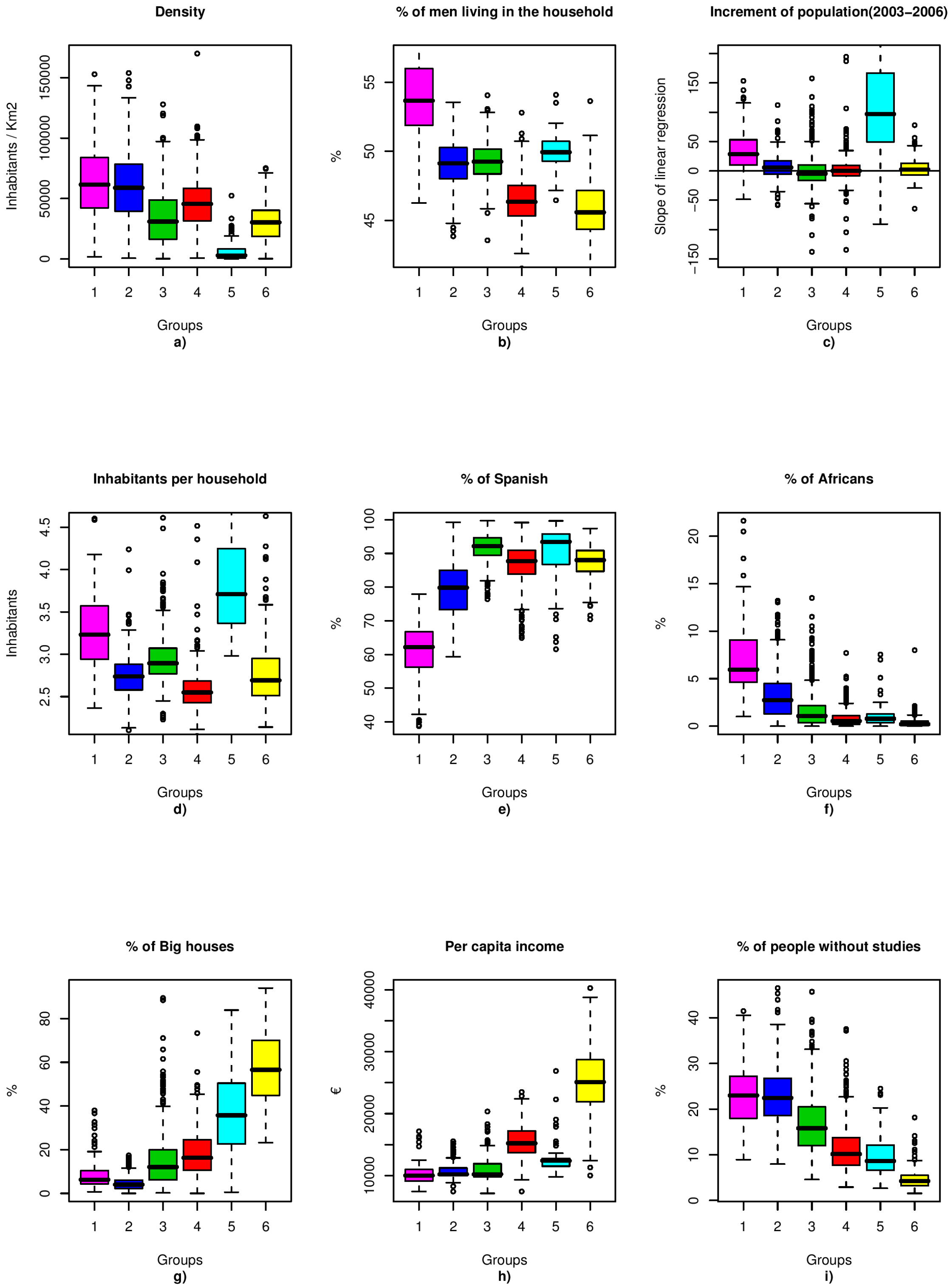


Figure 2.2: *Boxplots* of influential socioeconomic variables (**a-i**) stratified by group.

- The higher decrements in water consumption come from the groups with higher consumption.

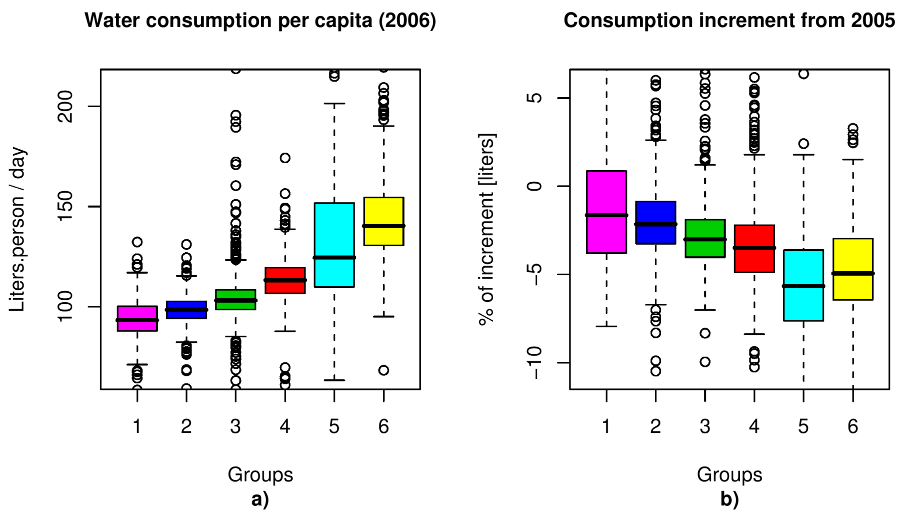


Figure 2.3: *Boxplots* of water consumption (**a**) and consumption increment (**b**), stratified by group.

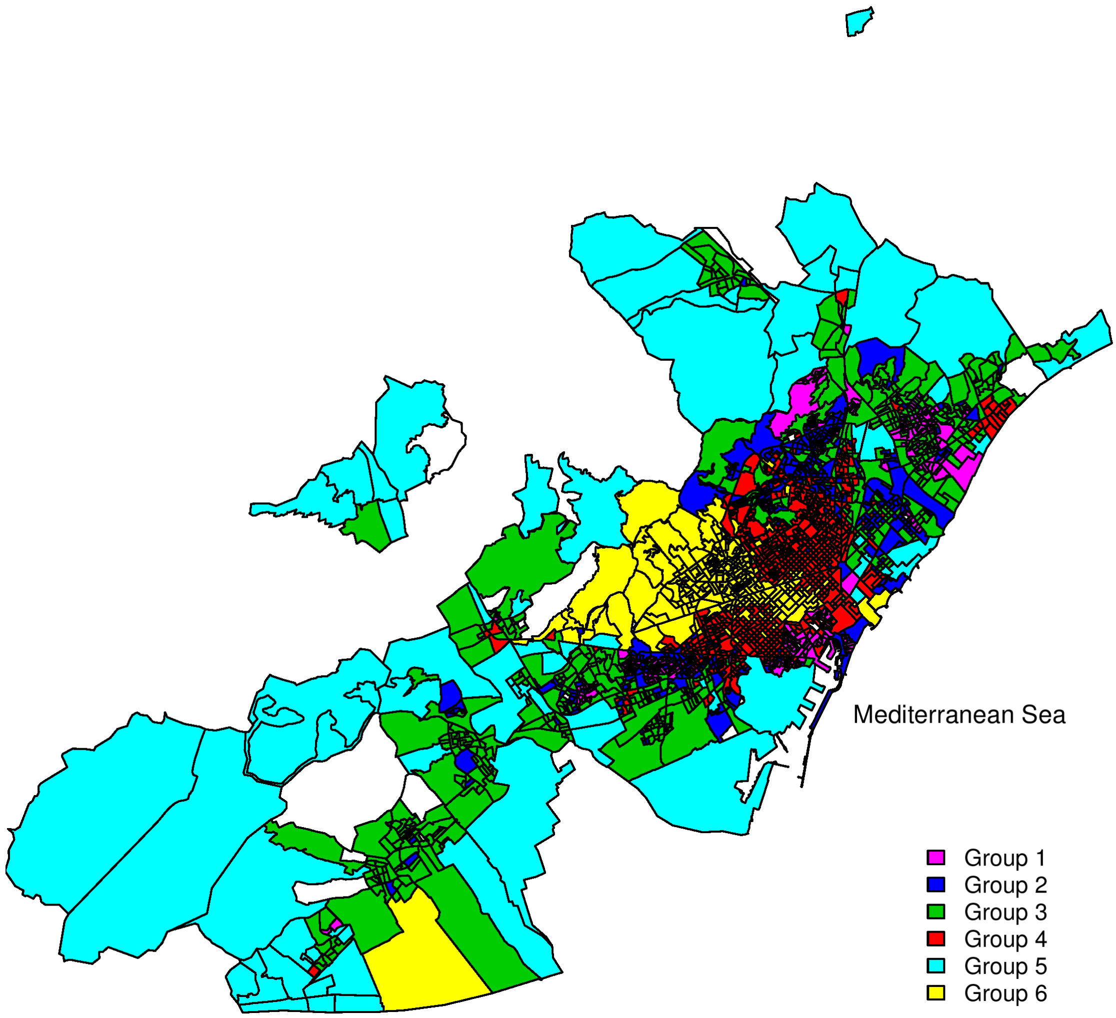


Figure 2.4: Geographical distribution of groups in the Barcelona Metropolitan Area.

2.4.2 Modelling

As referred to in Section 2.3.2, linear models make it possible to explain the behaviour of a numerical variable (Y), called the dependent variable, based on values of different variables (X), called the independent or explanatory variables. In our case, the dependent variable was the water consumption per capita in 2006-2007; and the explanatory variables, considered at the census tract level, were the 27 socioeconomic variables used to define the segmentation analysis explained in Table 2.2, plus information about the type of contracts. We have created six variables, defined as percentages of customers having a type of contract.[3]

A linear model was fitted for each of the six groups of consumers identified in the segmentation stage. Thus, for each group a model relating water consumption with socioeconomic variables was obtained. The models are presented in a table form (Table 2.4) to allow for an easy comparison between groups. The significant variables appear with the actual value of the equation coefficient, the non-significant ones have an asterisk. Thus, the model for Group 1, reproduced from the table, is:

$$\begin{aligned}
&\text{Water consumption} = 121.55 - 0.0000696 \cdot \text{Density} + 42.95 \cdot \%15 - 24\\
&+ 68.08 \cdot \%\text{Community} + 16.431 \cdot \%\text{America} - 81.28 \cdot \%\text{Africa}\\
&- 11.743 \cdot \%\text{Principal_Household} - 6.477 \cdot \%\text{Small} + 16.912 \cdot \%\text{Big}\\
&- 7.9821 \cdot \text{InhabitantsperHousehold} + 0.000218 \cdot \text{Income}\\
&+ 16.991 \cdot \%\text{Secondary_studies} + 42.379 \cdot \%\text{Higher_studies} + 1.983 \cdot \%\text{B}\,.
\end{aligned}$$

[3] The different contract types reflect, very broadly, the number of water points in the household.

	Model Group 1	**Model Group 2**	**Model Group 3**	**Model Group 4**	**Model Group 5**	**Model Group 6**
Constant	121.55	113.64	120.81	117.89	122.53	537.49
Density	-0.0000696	-0.0000696	-0.0000696	-0.000649	-0.0000696	-0.0000696
Area	*	*	*	*	*	*
%Women	*	*	*	*	*	*
%15-24	42.95	42.95	42.95	42.95	42.95	42.95
%25-64	*	*	*	*	*	*
%>65	*	*	*	152.87	-130.55	*
%Spanish	*	*	*	*	38.339	-59.07
%Community	68.08	68.08	-15.02	972.76	68.08	68.08
%EU Non community	*	*	*	375.8	*	*
%America	16.431	16.431	16.431	-316.069	16.431	16.431
%Africa	-81.28	-81.28	11.94	-81.28	-17.06	-81.28
%Principal_Household	-11.743	-11.743	-11.743	-11,743	-11.743	-34.673
%Small	-6.477	-6.477	-6.477	-45.627	-6.477	-6.477
%Big	16.912	16.912	16.912	16.912	16.912	35.716
Inhabitants_per_Household	-7.9821	-1.1421	-7.9821	0.3729	-7.9821	-7.9821
Income	0.000218	-0.000793	0.000218	-0.001956	0.000218	0.000218
%Primary_studies	*	*	*	*	*	*
%Secondary_studies	16.991	16.991	16.991	16.991	16.991	16.991
%Higher_studies	42.379	42.379	42.379	42.379	-65.411	-17.511
%A	*	*	*	*	*	-181.27
%B	1.983	1.983	1,983	1.983	1.983	-331.577
%C	*	*	*	21.313	*	-354.31
%D	*	*	20.09	*	*	-339.88
%E	*	*	*	*	*	-337.06
	$R^2 = 42\%$	$R^2 = 23\%$	$R^2 = 20\%$	$R^2 = 88\%$	$R^2 = 40\%$	$R^2 = 73\%$

Table 2.4: Regression coefficients for the models.

It is important to point out that the socioeconomic variables that do not appear in Table 2.4 are the variables considered as the base level; thus, the coefficient measures the difference in respect to the base level, which is the usual approach when working with compositional data. Several reduced models with approximately the same R^2 were found, so that the final choice was based on knowledge of the data and parsimony of the model.

R^2 measures the percentage of variability of the per capita water consumption as explained by the variables considered. The explained variability ranges from 20% to 88%, depending on the characteristics and homogeneity of the group. A lower value of R^2 does not mean a poor adjustment of the estimated model (in fact, all models are the "best" ones for the available data); it means that the explanatory variables included in the study can only explain a small part of the water consumption. Notice that groups 4 and 6 are very homogeneous and well explained by the model, while groups 2 and 3 have a low R^2, meaning that the variability within these groups is high and only partially explained by the variables in the models.

The interpretation and use of these linear models is not straightforward, and has to be done carefully due to the high correlation between the explicative variables and, therefore, between the regression coefficients.

A first and simple use of the models is for pointing out the relative importance of each variable in explaining water consumption and also in estimating water consumption when small changes of the explanatory variables are considered.

A more sophisticated use is scenario planning. That is, to use them in order to have an idea of changes in water consumption when the socioeconomic characteristics change more drastically; for example, when a specific group experiences an increase in per capita income. Because the socioeconomic variables are not independent, it won't be realistic to set a scenario increasing the per capita income and decreasing the level of studies. To help set realistic scenarios, Principal Component Analysis (PCA) is used. PCA is a multivariate statistical technique (Johnson and Wichern [2002]) that reduces the dimensionality of a database and shows how variables are related.

Basically, the technique creates new variables (called components) that are linear combinations of the original variables. The components are ranked, and usually just a few of them capture a large amount of variability from the database. That is, only a few components (say, the first two components) have almost the same information as the whole database.

A common representation is the scatter plot showing the weights of the original variables in the first two components of a PCA. In such a graph, the longer the projection of the line for a variable into an axis, the more important the contribution of that variable is for what that axis represents. Figure 2.5 shows the graph for the principal component analysis made with data from all consumers. Although we do not aim to interpret this graph from a socioeconomic point of view, notice that the horizontal axis contains variables related to economic data (income, level of studies), whereas the vertical axis contains variables related to immigration.

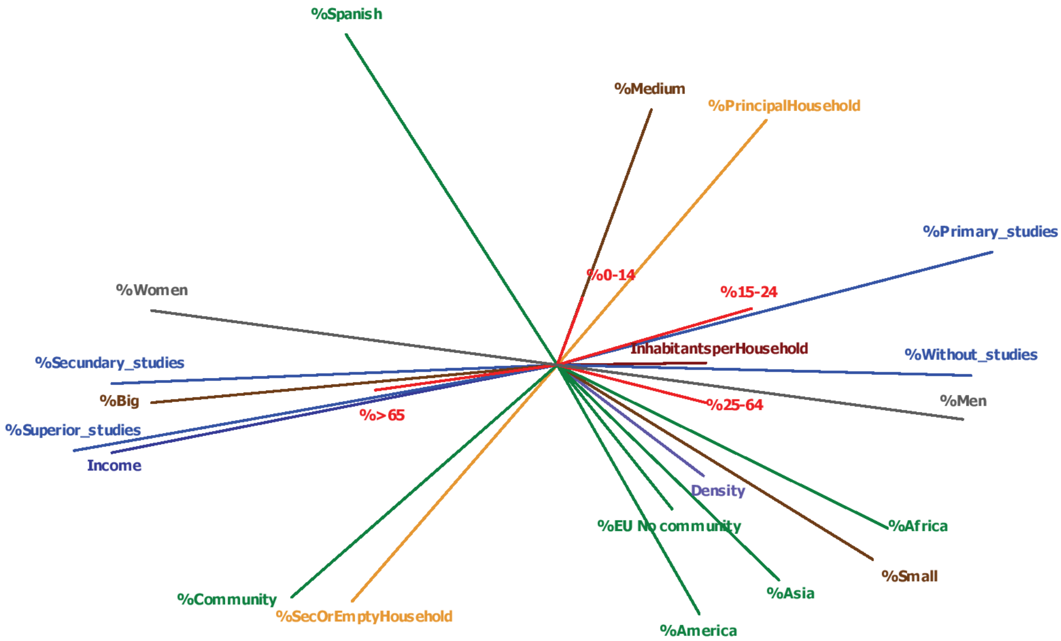


Figure 2.5: Socioeconomic map from a principal component analysis (PCA).

The way to use the PCA graph for scenario planning is the following:

- Variables that are close in the graph should be changed together and in the same direction. Example: Income, % of Secondary_studies and % of Superior_studies.
- Variables that are opposite (at 180°, more or less) should be changed together, but in the opposite direction (if one is increased the other should be decreased). Example: % of men and women.
- Variables that are at 90°are basically independent and thus, can be changed independently. Example: the % of community immigrants (%Community) and % of men and women.

As we have just seen, PCA is a big help in planning scenarios, but still it is very important to use sociological knowledge and common sense. In addition it must be kept in mind that the models reflect the relationship between the socioeconomic variables and water consumption, but not necessarily a cause and effect relationship.

2.4.3 Time Series

As stated earlier, for short term forecasting a time series SARIMA model was fitted to each group's overall consumption. This kind of model is especially suited to short-term forecasting because, in general, it relies heavily on the recent past. In our case, we worked with monthly data. It is reasonable to forecast 6-8 months ahead, up to a maximum of 12. It is also worth noting that these predictions are basically based on past consumption, they don't use socioeconomic data; in our case we forced the models to take into account meteorological data: temperature and rainfall.

The first step was to characterize the profile of monthly consumption –expressed in liters per inhabitant and day– for each group from June, 2003 to December, 2007. We found that the profiles behaved in a similar way with a strong seasonal component, but with different levels of consumption (Fig. 2.6). Then the SARIMA models were fitted (according to the maximum likelihood criterion) taking into account the calendar and weather effects as they might have an influence.

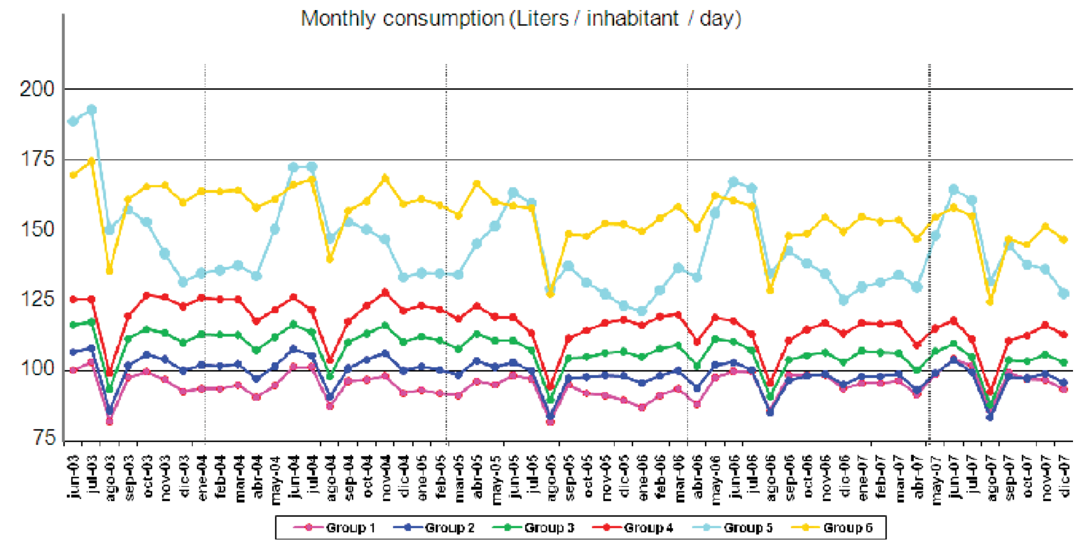


Figure 2.6: Monthly consumption profiles (liter/inhabitant/day) by group from June, 2003 to July, 2007.

An example of a calendar effect is the Easter effect. It has to do with the fact that the month of the year that Easter falls in, April or March, has a lower than expected water consumption in the Barcelona area because people leave the city for a few days on vacation. Another calendar effect is due to the fact that the data vary according to the number of times that each day of the week occurs in each month (four or five times); many people also leave the city on weekends. To take into account the impact of climate variables, we imposed that the effects of temperature and rainfall appear in the model.

Then we used Box & Jenkins methodology (Box et al. [1994]) to obtain and validate a model for each group. The details of the SARIMA models obtained are omitted since they are not of general interest and the methodology used is well known and has been used in other occasions for water consumption prediction (Molino et al. [1996]).

Obviously, from the prediction of consumption for each group, the global monthly water consumption for the Barcelona area can be predicted. The procedure is simple, it has two steps: first, predict the monthly water consumption for each segment; and second, add the results.

What is interesting to note is that the forecast obtained using this method is more reliable and has narrower confidence bands than the one that would have been obtained by modelling the aggregated consumption data directly. As explained in section 2.3.3, this is because it has a smaller variance. Table 2.5 shows a comparison of the forecasts for 2007 using the method based on the groups and the method modelling the aggregated consumption.

Forecast using segmentation (sum of groups forecasts)				**Forecast using the aggregated (global) consumption**			
Total Consumption (m^3)				**Total Consumption (m^3)**			
Forecast	**Standard Deviation**	**Lower Limit (95%)**	**Upper Limit (95%)**	**Forecast**	**Standard Deviation**	**Lower Limit (95%)**	**Upper Limit (95%)**
9,904,222	99,405	9,705,412	10,103,032	9,899,908	211,577	9,476,753	10,323,063
8,872,175	99,128	8,673,919	9,070,431	8,860,713	210,443	8,439,828	9,281,599
9,884,941	119,199	9,646,544	10,123,338	9,870,851	252,594	9,365,662	10,376,039
8,969,167	123,825	8,721,517	9,216,816	8,945,840	262,048	8,421,745	9,469,935
10,051,117	136,144	9,778,829	10,323,406	10,028,858	287,824	9,453,210	10,604,506
9,761,781	139,229	9,483,323	10,040,240	9,729,310	294,108	9,141,094	10,317,525
10,015,510	151,202	9,713,105	10,317,914	9,991,192	319,189	9,352,815	10,629,570
8,146,244	158,195	7,829,854	8,462,633	8,119,114	333,768	7,451,579	8,786,649
9,204,700	159,572	8,885,555	9,523,844	9,181,362	336,519	8,508,325	9,854,399
9,596,037	171,326	9,253,385	9,938,688	9,557,126	361,164	8,834,797	10,279,454
9,512,924	171,801	9,169,322	9,856,525	9,477,400	362,043	8,753,314	10,201,485
9,432,299	183,520	9,065,260	9,799,338	9,397,845	386,624	8,624,596	10,171,094

Table 2.5: Comparison of forecast using the aggregated models (based on groups) and global consumption.

2.5 Concluding Remarks

The paper presents a methodology to study, understand, model and forecast water consumption based on socioeconomic data available from official statistics. The statistical methodology used is based on cluster analysis to identify groups with similar consumption habits and socioeconomic status and then on regression analysis and time series SARIMA models to predict and understand water consumption.

The paper confirms, in the Barcelona case, the initial idea that there is a relationship between water consumption habits and socioeconomic characteristics and that this relationship can be used to detect groups of clients with similar consumption patterns and then use these groups for different purposes: a more accurate short term prediction of water consumption, long term scenario planning and a better understanding of consumer habits. The benefits of having the customers segmented into homogeneous groups does not stop here; another important benefit is the ability to take samples of clients in a stratified way –sampling within each cluster– which allows for more reliable results with small sample sizes.

The results from the Barcelona Metropolitan area are the six clusters that were found, used throughout the article, to show the methodology, the difficulties encountered and the results obtained.

The proposed methodology can be used in other areas with only small adaptations. The only requirements are: having socioeconomic data on a census tract level, water consumption at the consumer (or other disaggregate) level and being able to situate the consumers in census tracts. While writing this article we have been in contact with other water companies, which belong to the AGBAR group, in order to apply the methodology.

Acknowledgements. The authors are grateful to Montserrat Termes from CETAQUA for very useful comments and suggestions during the preparation of this manuscript.
The authors are grateful to R + I Alliance for the financial support that made it possible to develop this project.

Bibliography

Arbués, F. and Villanua, I. Potential for pricing policies in water resource management: Estimation of urban residential water demand in Zaragoza, Spain. *Urban Studies*, 43: 2421–2442, 2006.

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

Babel, M. S. and Shinde, V. Identifying prominent explanatory variables for water demand prediction using artificial neural networks: A case study of Bangkok. *Water Resource Management*, 25(6):1653–1676, 2011.

Babel, M. S., Das-Gupta, A., and Pradhan, P. A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal. *Water Resource Management*, 21(3):573–589, 2007.

Baumann, D. D., Boland, J., and Hanemann, W. M. *Urban water demand management and planning.* McGraw-Hill, New York, 1998.

Beecher, J. Integrated resources planning for water utilities. *Water Resources Update*, 104, 1996.

Box, G. E. P., Jenkins, M., and Reinsel, G. *Time series analysis: forecasting and control.* Prentice-Hall, 3rd edition, 1994.

Brooks, D. B. An operational definition of water demand management. *Water Resources Development*, 22:521–528, 2006.

Butler, D. and Memon, F. Water demand management. *International Water Association Publishing (IWAP)*, 2006.

Corral-Verdugo, V., Frías-Armenta, M., Pérez-Urias, F., Orduña-Cabrera, V., and Espinoza-Gallego, N. Residential water consumption, motivation for conserving water and the continuing tragedy of the commons. *Environ. Manag.*, 30:527–535, 2002.

Draper, N. and Smith, W. *Applied regression analysis.* Wiley, 3rd edition, 1998.

Duke, J. M., Ehemann, R. W., and Mackenzie, J. The distributional effects of water quantity management strategies: a spatial analysis. *Rev. Reg. Stud.*, 32(1):19–35, 2002.

Dziegielewski, B. Management of Water Demand: Unresolved Issues. *J. Water Resour. Update*, 114:1–7, 1993.

European Commission. EU Water Framework Directive. Directive 2000/60/EC.

Gleick, P. H. Water use. *Annu. Rev. Environ. Resour.*, 28:275–314, 2003.

Griffin, R. C. and Chang, C. Seasonality in community water demand. *West. J. Agric. Econ.*, 16(2):207–217, 1991.

Guy, S. Managing water stress: the logic of demand side infrastructure planning. *J. Environ. Plan. Manag.*, 39:123–130, 1996.

Hamilton, L. Saving water: a causal model of household conservation. *Sociological Perspectives*, 26:355–374, 1983.

Hanke, S. and de Mare, L. Residential water demand: a pooled, time series, cross section study of Malmö, Sweden. *J. Am. Water Resour. As.*, 18(4):621–626, 1982.

Hasse, D. and Nuiss, H. Does urban sprawl drive changes in the water balance and policy? The case of Leipzig (Germany). *Landscape and Urban Planning*, 80:1–13, 2007.

Hellegers, P., Soppe, R., Perry, C., and Bastiaanssen, W. Remote sensing and economic indicators for supporting water resources management decisions. *Water Resour. Manag.*, 24(11):2419–2436, 2010.

ICWE (International Conference on Water and Environment). Dublin, 1992.

Johnson, R. A. and Wichern, D. W. *Applied multivariate statistical analysis*. Prentice Hall, 2002.

Kahn, M. E. The Environmental Impact of Suburbanization. *J. Policy. Anal. Manage.*, 19:569–586, 2000.

Kanakoudis, V. K. Urban water use conservation measure. *Journal on Water Supply Research and Technology: AQUA*, 51(3):153–159, 2002.

Lavière, I. and Lafrance, G. Modelling the electricity consumption of cities: effect of urban density. *Energy Econ.*, 21:53–66, 1999.

Lebart, L., Morineau, A., and Piron, M. *Statistique exploratoire multidimensionnelle*. In Dunod, 4th edition, 2006.

Liu, J., Daily, G. C., Ehrlich, P. C., and Luck, G. W. Effects of households dynamics on resource consumption and biodiversity. *Nature*, 421:530–533, 2003.

Molino, B., Rasulo, G., and Tagliatela, L. Forecast model of water consumption for Naples. *Water Resour. Manag.*, 10(4):321–332, 1996.

Montini, A. M. M. The determinants of residential water demand: empirical evidence for a panel of Italian municipalities. *Appl. Econ. Lett.*, 13:107–111, 2006.

Murdock, S. H., Albrecht, D. E., Hamm, R. R., and Backman, K. Role of sociodemographic characteristics in projections of water use. *J. Water Resour. Plann. Manag.*, 117:235–251, 1991.

Nauges, C. and Thomas, A. Long-run study of residential water consumption. *Environmental & Resource Economics, European Association of Environmental and Resource Economists*, 26(1):25–43, 2003.

Opaluch, J. J. Urban residential demand for water in the United Status: Further discussion. *Land. Econ.*, 58:225–227, 1982.

Peña, D. *Procesos de media móvil y ARMA*. Alianza. Análisis de series temporales, Madrid, 1st edition, 2005.

Peña, D. *El modelo general de regresión*. Alianza. Regresión y diseño de experimentos, Madrid, 2nd edition, 2010.

Postel, S. *The last oasis. Facing water scarcity*. London: W W Norton & Co Inc, 1992.

Renwick, M. E. and Green. Do residential water demand side management policies measure up? An analysis of eight California water agencies. *J. Environ. Econ. Manag.*, 40:37–55, 2000.

Renzetti, S. *The economics of water demand*. Kluwer Academic Publishers, Boston, 2002.

Smith, A. and Ali, M. Understanding the impact of cultural and religious water use. *Water. Environ. J.*, 20:203–209, 2006.

Stephenson, D. Demand management theory. *Water S.A*, 25:115–122, 1999.

United States Environmental Protection Agency (USEPA). *Water and energy savings from high efficiency fixtures and appliances in single family homes*. USEPA, Washington.

Zhang, H. and Brown, D. Understanding urban residential water use in Beijing and Tianjin, China. *Habitat International*, 29:469–491, 2005.

Chapter

3

Article 2

An Approach to Disaggregating Total Household Water Consumption into Major End-Uses

Sara Fontdecaba, José A. Sánchez-Espigares, Lluís Marco-Almagro, Xavier Tort-Martorell, Francesc Cabrespina, Jordi Zubelzu

Received: 26 September 2012 / Accepted: 16 January 2013 / Published online: 27 January 2013

Abstract. The aim of this project is to assign domestic water consumption to different devices based on the information provided by the water meter. We monitored a sample of Barcelona and Murcia with flow switches that recorded when a particular device was in use. In addition, the water meter readings were recorded every 5 and 1 s, respectively, in Barcelona and Murcia. The initial work used Barcelona data, and the method was later verified and adjusted with the Murcia data. The proposed method employs an algorithm that characterizes the water consumption of each device, using Barcelona to establish the initial parameters which, afterwards, provide information for adjusting the parameters of each household studied. Once the parameters have been adjusted, the algorithm assigns the consumption to each device. The efficacy of the assignation process is summarized in terms of: sensitivity and specificity. The algorithm provides a correct identification rate of between 70% and 80%; sometimes even higher, depending on how well the chosen parameters reflect household consumption patterns. Considering the high variability of the patterns and the fact that use is characterized by only the aggregate consumption that the water meter provides, the results are quite satisfactory.

Keywords: Water pattern recognition, use of water, domestic consumption, water profile, water disaggregation.

3.1 Introduction

Many efforts have been made to understand the factors and behaviors that cause high variations in per capita water consumption among various regions, cities and people. With this in mind, Aigües de Barcelona conducted a study (Fontdecaba et al. [2012]) to compare the consumption of clients within homogeneous socioeconomic groups. One significant application of this project was higher accuracy in forecasting future water demand. Realizing that a relationship exists between socioeconomic characteristics and water consumption habits was a step forward because, in general, water companies know very little about the patterns and ways their customers use water. Obviously, knowing how clients use the water would be even better as it would allow the companies to, for example, quantify the impact of water efficiency measures or new policy initiatives, as well as facilitating more accurate forecasts. Thus, it was logical (and challenging) for Aigües de Barcelona to seek this information by researching how to disaggregate the whole domestic water consumption into the different major end-uses.

The problem is difficult because individual metering of water devices in homes is awkward and expensive. Further, the absence of pertinent data makes it unfeasible to identify and quantify water consumption by devices with the same accuracy that has been achieved for gas (Yamagami and Nakamura [1996]) and electricity (Farinaccio and Zmeureanu [1999]) consumption.

The objective of the work presented here is to show that whole-house meter data can provide quite accurate end-use water consumption data. The project was conducted on behalf of and with full support from Aigües de Barcelona, which was interested in learning about their clients' consumption habits. The first stage of the project was the analysis of water consumption patterns in the city of Barcelona by monitoring device consumption in different homes from different socioeconomic segments. Two types of information were recorded: the water consumption from the whole-house water meter (read every 5 s) and synchronized information from the devices consuming the water. Thus, the monitored group provided baseline information for tackling the problem of identifying the devices by using only data from the meter. The problem was difficult because of the scarcity of the information on which to base the assignation and also because consumption habits varies highly among people, households and even individuals within the same household.

This paper is organized in the following manner: Section 3.2 briefly reviews the scope and data used to carry out the study in the area of Barcelona; Section 3.3 provides the most relevant results of the data analysis; Section 3.4 explains the methodology followed in uses recognition patterns; Section 3.5 highlights and discusses the most relevant results obtained in the Barcelona case; and finally, Section 3.6 presents some conclusions.

3.2 Case Study and Data Used

3.2.1 Data Sources

In order to ensure that the sample structure was as robust as possible, the property sample structure was designed by taking into account all major factors that were known to influence household water consumption. In the Barcelona area, clients can be grouped into segments (six segments) with homogeneous consumer habits, which are known to be related to socioeconomic status (Fontdecaba et al. [2012]). That was the reason behind considering the variability between households as an important source to consider when designing the sample; accordingly, we wanted to guarantee that households from all 6 segments were included. These segments are distributed on the Barcelona map in a rather predictable pattern; so to avoid geographical bias, the location of the eight households used in this study was also taken into account.

To monitor each of the households, we designed a system which monitored two complementary blocks of data: meter traffic flow and each of the different water consumption points inside the household (kitchen basins, cistern, washing machine, etc.). [1] The data collected was consolidated and stored on a remote server for later analysis.

The first source of data (at the household level) was not difficult to obtain, since it is already collected by the metering system of Aigües de Barcelona. Monitoring water use within the household was the main problem, and we opted for an integrated solution which involved placing a set of non-intrusive flow sensors (called flow switches) at different points on the pipe section leading to a single consumption.

The sample frequency and the water consumption level of detail were key questions regarding the assignation, and they were decided according to the main objective of the project. Thus, we considered the minimum time period that the devices could be in use and could be logged. In the end, the two sources of data were collected every 5 s over a period of 3 months, between December 2009 and February 2010.

3.2.2 Data Management

Information from the meter was collected every 5 s with a 0.1 liter resolution, using a concentrator that sent periodic requests to the electronic meter, which registered, dated and saved the totalized flow and its characteristic attributes. If no flow variation occurred when compared to the previous request, it was ruled out.

The flow switches were permanently aware and attached directly to wireless terminals that, every 5 s, reported to a network master whether water passed through (1) or not (0). The master acted as a coordinator at the household level and also as a gateway to an on-site storage facility that compiled and registered all received data via GPRS to the remote server. This approach allowed the data to be received every day through GPRS and avoided the use of data loggers, which would have been more complex and expensive.

[1] From now on we will use the word "device" to refer to the different things causing consumption weather they are appliances, showers, taps or cisterns.

To sum up briefly, the monitoring process in the electronic meter gave us progressive information about the meter counter and the flow over time. The flow switches provided information about where and when water was used in the household (washing machine, shower, kitchen basin, etc.). Considering the two together offers a picture of household consumption, such as in Fig. 3.1.

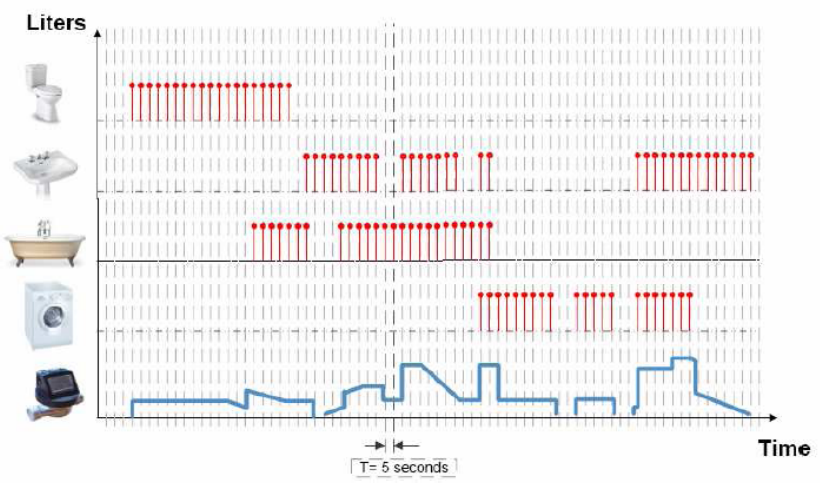


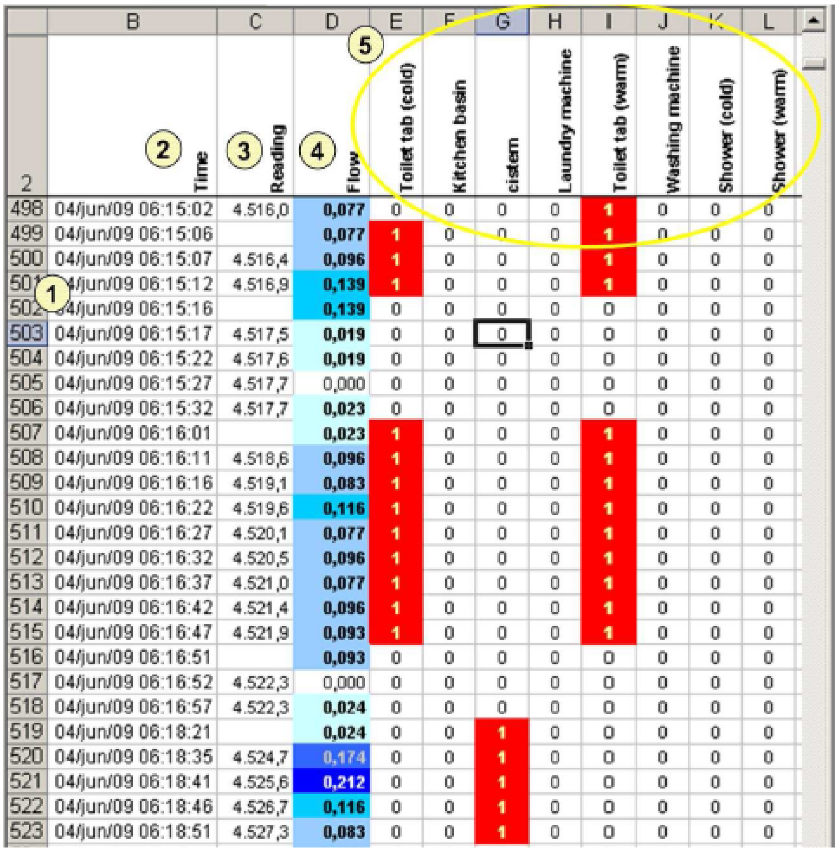
Figure 3.1: Household consumption data.

The system provided us with a data base containing the water use and meter data of 8 households over a period of 3 months; a huge amount of data. To automatically assemble, elaborate on and consolidate such information was a meticulous process. To illustrate this point, consider the spreadsheet snapshot below, which corresponds to a particular household in which the test was conducted (Fig. 3.2).

Number 1 indicates the rows that were sorted according to the time line evolution, with every row corresponding to a sample. The usual timing from sample to sample is 5 s (monitoring period). Number 2 (Column B) shows the time, with the day, month and hour when the sample was captured. Number 3 (Column C) corresponds to the meter's totalized reading at the time of the sample. Number 4 (Column D) is the flow in liters per second coming into the household. As a visual aid, the color code is proportional to the flow intensity. The group of columns represented by number 5 is the water use where each column corresponds to a unique, recognizable consumption point. The 0/1 sequence found in any water use column indicates whether or not the particular consumption point was used. The red and white backgrounds provide visual aid.

Despite the fact that two flow-switches were installed for hot and cold water, our interest was in detecting use activation without distinguishing between the two. For this reason, the sequences of 0/1 corresponding to the two types of water were combined.

Considering all the 8 households, a total of 12 different cisterns, 8 washing machines, 8 kitchen basins, 8 bath taps, 5 dishwashers and 10 showers are in the sample.



	B	C	D	E	F	G	H	I	J	K	L
2	Time	Reading	Flow	Toilet tab (cold)	Kitchen basin	cistern	Laundry machine	Toilet tab (warm)	Washing machine	Shower (cold)	Shower (warm)
498	04/jun/09 06:15:02	4.516,0	0,077	0	0	0	0	1	0	0	0
499	04/jun/09 06:15:06		0,077	1	0	0	0	1	0	0	0
500	04/jun/09 06:15:07	4.516,4	0,096	1	0	0	0	1	0	0	0
501	04/jun/09 06:15:12	4.516,9	0,139	1	0	0	0	1	0	0	0
502	04/jun/09 06:15:16		0,139	0	0	0	0	0	0	0	0
503	04/jun/09 06:15:17	4.517,5	0,019	0	0	0	0	0	0	0	0
504	04/jun/09 06:15:22	4.517,6	0,019	0	0	0	0	0	0	0	0
505	04/jun/09 06:15:27	4.517,7	0,000	0	0	0	0	0	0	0	0
506	04/jun/09 06:15:32	4.517,7	0,023	0	0	0	0	0	0	0	0
507	04/jun/09 06:16:01		0,023	1	0	0	0	1	0	0	0
508	04/jun/09 06:16:11	4.518,6	0,096	1	0	0	0	1	0	0	0
509	04/jun/09 06:16:16	4.519,1	0,083	1	0	0	0	1	0	0	0
510	04/jun/09 06:16:22	4.519,6	0,116	1	0	0	0	1	0	0	0
511	04/jun/09 06:16:27	4.520,1	0,077	1	0	0	0	1	0	0	0
512	04/jun/09 06:16:32	4.520,5	0,096	1	0	0	0	1	0	0	0
513	04/jun/09 06:16:37	4.521,0	0,077	1	0	0	0	1	0	0	0
514	04/jun/09 06:16:42	4.521,4	0,096	1	0	0	0	1	0	0	0
515	04/jun/09 06:16:47	4.521,9	0,093	1	0	0	0	1	0	0	0
516	04/jun/09 06:16:51		0,093	0	0	0	0	0	0	0	0
517	04/jun/09 06:16:52	4.522,3	0,000	0	0	0	0	0	0	0	0
518	04/jun/09 06:16:57	4.522,3	0,024	0	0	0	0	0	0	0	0
519	04/jun/09 06:18:21		0,024	0	0	1	0	0	0	0	0
520	04/jun/09 06:18:35	4.524,7	0,174	0	0	1	0	0	0	0	0
521	04/jun/09 06:18:41	4.525,6	0,212	0	0	1	0	0	0	0	0
522	04/jun/09 06:18:46	4.526,7	0,116	0	0	1	0	0	0	0	0
523	04/jun/09 06:18:51	4.527,3	0,083	0	0	1	0	0	0	0	0

Figure 3.2: Water use and meter data collection.

3.3 Data Analysis

In order to investigate and understand the variability in microcomponent data (data consumptions disaggregated at the devices level), the dataset was analyzed using a number of statistical methods. The main characteristics analyzed were the frequency of the water use (understood as the number of times each device is used in each property per day) and the volume per use.

Even after measures were taken to "guarantee" the data quality of the 8 monitored households and several corrections and repairs were performed on the loggers during the monitoring period, the database contained mistakes and inconsistencies. After thorough cleaning, the database was ready for analysis. It is worth mentioning that several of the problems were related with the loggers from 3 households that had dishwashers; thus, the information on the consumption patterns of this appliance is scarce. However, as we will see in the developed procedure, it is sufficient.

3.3.1 Definitions and General Characteristics

Before beginning the statistical analysis to characterize the consumption patterns of each device, it is necessary to establish some specific definitions.

Water use: continuous period of time in which the same device takes water. It is identified in the database by using a combination of flow and time criteria (Fig. 3.3).

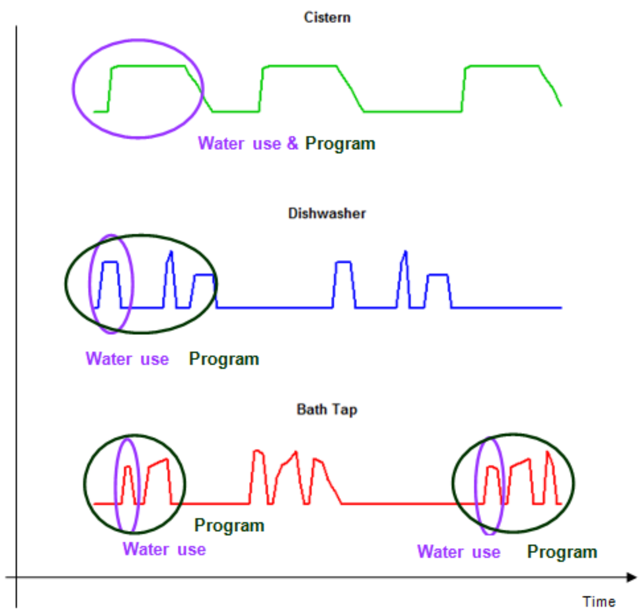


Figure 3.3: Definition of water use and programs.

Some devices, such as the cisterns, consume water in a continuous way. When they start, they begin consuming water and, when they end, consumption ends. However, dishwashers and washing machines take water in a discontinuous way. In general they take some water at the beginning, then stop the intake and perform other operations. A few minutes later, they take in water again. This pattern is repeated at different intervals until the device discontinues its use. In other words, their total water consumption is composed of different use groups. In these cases it will be necessary to define the concept of program.

Program: consecutive groups of water uses belonging to the same device that define the full use of the device. It can be composed of one or more water uses (Fig. 3.3).

The consumption of other devices, such as the shower, can be characterized by a program with a single water use or a program with several water uses. It depends on the shower habits of the individual.

The kitchen basin and bath tap are very complicated (in many cases impossible) for distinguishing whether different consecutive water uses belong to the same program or to a succession of single-use programs. Different households may have very different patterns of tap usage.

Given the need to correctly identify the programs, it was necessary to establish criteria for considering the limits between programs. The frontier between programs was based on the time period between two water uses by the same device; when this period exceeded a specified time –which was different for each type of device– the consumption was assigned to two different programs. Table 3.1 shows the maximum time period allowed between continuous water uses which were considered to be in the same program:

Device	Minutes
Cistern	0
Washing Machine	60
Dishwasher	60
Shower	15
Kitchen Basin	10
Bath Tap	10

Table 3.1: Maximum time (in min.) allowed.

Both programs and water uses were statistically analyzed in order to characterize the water consumption behavior of each device. For example, in the case of the washing machines, it would then be possible to answer questions like these: How long is a washing machine program? How many times does the washing machine take in water? For how long does the washing machine take in water?

3.3.2 Descriptive Analysis of Programs and Water Uses

One of the main objectives of the descriptive analysis is to understand the consumption profile of different devices.

In order to have a robust descriptive analysis –one that would give the best possible unbiased information on how each device consumes water– only programs that began and finished without any interference from other devices were taken into account (67% of the total number of programs). For the 8 monitored properties, Table 3.2 shows the total number of available programs and water uses.

Device	Number of water uses	Number of programs	Frequency of use (prog./property/day)	Consumption per prog. (liters/prog.)	Household Consumption (liters/property/day)	% of total Consumption
Cistern	3.077	3.077	6.88	5.56	30.10	32.11
Washing Machine	575	69	0.33	38.06	5.62	4.92
Dishwasher	538	88	***	***	***	***
Shower	1.261	591	1.31	36.37	35.47	40.31
Kitchen Basin	8.072	3.068	6.90	2.69	14.20	15.47
Bath Tap	3.362	2.183	6.80	1.35	4.75	5.52

Table 3.2: Statistics of program frequency and volume for each device.

The total daily household consumption is nearly 93 l per house. To interpret this number, it is necessary to keep in mind that not all devices in the households were monitored and that the results only include the devices when they are running without simultaneities. The shower represents 40% of this daily consumption and the cistern 32%. The kitchen basin and the bath tap constitute 21%, and a small percentage of about 6% is for the washing machine and the dishwasher.

Within each property, the distribution of the percentages for the total consumption of water was very similar. The higher percentages covered showers and cisterns. Bath and kitchen taps constitute the second block of consumption to be considered, followed by washing machines and dishwashers.

The percentage of total water consumption covered by devices does not differ from results reported by the Statistical Office of the European Commission (Eurostat [2007]) and several other published works (Mayer et al. [2003] and Richter and Stamminger [2012]).

3.3.3 Analysis of Possible Influential Factors in Domestic Water Consumption

Considering the main objective of distinguishing the consumption profiles of each of the devices, it would be useful to find variables that would help differentiate them. The pace of life today runs on timetables, therefore it is reasonable to assume that the time of day or day of the week (i.e. work days vs. weekends) can influence consumption habits.

Time of Day

Daily water consumption per hour, as averaged across all 8 households, is not useful in characterizing use by any of the devices. All the programs are more or less distributed throughout the day without any significant considerations. Cisterns and showers are distributed throughout the whole day, while kitchen basins register their maximum use at lunch time. Washing machines are distributed from 7.00 to 23.00, and dishwashers normally appear in the early afternoon (14.00-16.00). During the morning (6.00-10.00), cisterns and showers constitute the main use (Fig. 3.4).

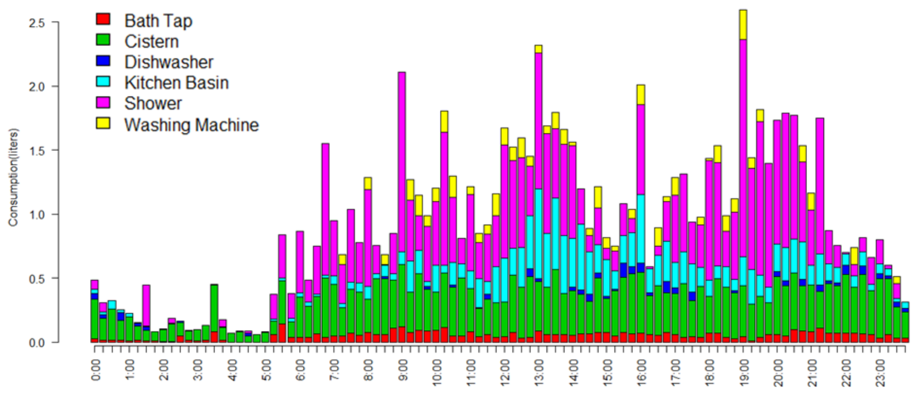


Figure 3.4: Average water consumption per day and hour.

Work Days and Weekend Water Consumption

People do not consume water differently on work days and weekends. Table 3.3 shows the distribution of the programs used during work days and weekends across the 8 properties (71.9% of programs are executed during work days and 28.1% during weekends). All weekday holidays were considered as labor days. The small differences are insignificant to the purpose of this study for detecting device consumption (Table 3.3). The distribution has also been tested within every property, and none of them show significant differences between the types of days considered.

Device	**Labor**	**Weekends**	**General**
Cistern	35.7%	33.2%	33.9%
Washing Machine	0.8%	0.8%	0.8%
Dishwasher	1.1%	0.9%	1.0%
Shower	6%	6.7%	6.5%
Kitchen Basin	33.2%	34%	33.8%
Bath Tap	23.3%	24.4%	24.1%

Table 3.3: Distribution of the program during weekdays, weekends and in general.

3.3.4 Descriptive Analysis of the Devices

In order to identify which device has consumed the water recorded by the meter, it is necessary to characterize (model) what kind of consumption each device produces. The level of detail in the microcomponent dataset used in this study reasonably represents the programs for characterizing and analyzing the devices. The descriptive analysis that follows will help develop the procedure for identifying microcomponents based on the information provided by the water meter.

The Cistern

Cisterns are the most used device in domestic habitats. Their total duration and the minimum and maximum levels they consume are the most significant results for characterizing their programs. Table 3.4 shows the descriptive statistics averaged across all the 8 properties (Min.: minimum, 1st Qu.: first quartile, 3rd Qu: third quartile, Max: maximum and Stdev: standard deviation).

According to the descriptive analysis, cisterns seem to take in nearly 5.5 l of water in 43 s. These are the combined results of both dual- and single-flush cisterns, and therefore the cause of such a high standard of deviation. In the sample, 40% of the cisterns were water efficient, dual-flush toilets with reduced flush and time consumption.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Stdev.**
Duration (sec.)	4.30	29.40	40.60	43.24	55.30	115.80	17.97
Consumption (l.)	1.00	3.49	4.62	5.56	7.68	21.60	2.86

Table 3.4: Descriptive statistics of the cisterns.

The Shower

Showers (and baths) are among the devices which consume the most water. Their consumption basically depends on flow rates and the duration in which they are turned on. Depending on individual habits and preferences, people may turn the water off periodically, decrease the water pressure at certain times, or they may have shorter showers with a constant flow of water.

The statistics used to characterize all of these habits are (per shower): duration, consumption, and the number of water uses (remember that this represents the number of times that the shower tap is turned on and off during a program, i.e., a single shower). The results are summarized in Table 3.5.

The Washing Machine

There are a wide range of washing machines on the market. They often have as many as ten programs for washing, rinsing and spinning. Moreover, energy-efficient washing machines use sensors to detect the size of a load and accordingly determine the appropriate amount of water (some even incorporate fuzzy logic strategies to vary the amount of water while running). All these factors increase the variability among programs and hinder their characterization.

The descriptive statistics are based on washing machine programs separated from any other device program. The useful variables for characterizing washing machines are: consumption, the number of uses during the programs, the duration of the program and the duration of time that the washing machine takes in water. Figure 3.5 shows the differences between the two last statistics.

The main summary statistics are presented in Table 3.6.

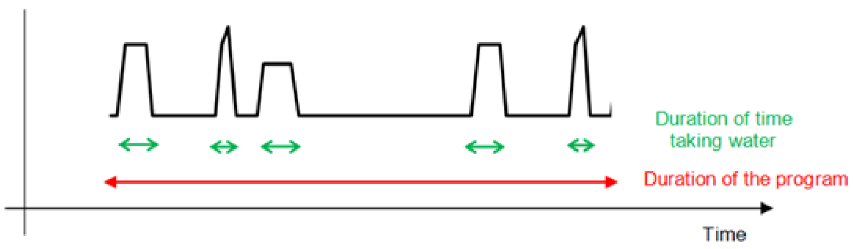


Figure 3.5: Differences between duration of program and duration of time taking in water.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Stdev.**
Duration (time)	0 min. 5 s.	1 min. 6 s.	4 min. 40 s.	5 min. 20 s.	7 min. 40 s.	21 min. 3 s.	4 min. 22 s.
Consumption (l.)	1.00	8.35	33.2	36.37	52.41	240.60	31.18
Water uses	1.00	1.00	2.00	2.13	3.00	14.00	4.2

Table 3.5: Descriptive statistics of the showers.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Stdev.**
Duration Program (time)	8 min. 15 s.	33 min. 55 s.	40 min. 45 s.	47 min. 33 s.	57 min. 50 s.	122 min. 35 s.	24 min. 33 s.
Duration Taking in water (time)	1 min. 25 s.	3 min. 50 s.	5 min. 30 s.	5 min. 10 s.	6 min. 2 s.	11 min. 28 s.	2 min. 15 s.
Consumption (l.)	1.15	30.05	42.55	38.06	47.34	61.69	12.75
Water uses	3.00	7.00	8.00	8.31	10.00	14.00	2.04

Table 3.6: Descriptive statistics of washing machines.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Stdev.**
Duration Program (time)	0 min. 10 s.	17 min. 12 s.	32 min. 55 s.	45 min. 56 s.	**84 min. 32 s.**	**135 min. 10 s.**	40 min. 02 s.
Duration Taking in water (time)	0 min. 5 s.	2 min. 10 s.	2 min. 44 s.	2 min. 32 s.	**3 min. 14 s.**	**4 min. 42 s.**	1 min. 13 s.
Consumption (l.)	0.3	5.27	10.79	10.16	**13.3**	**28.5**	7.03
Water uses	1.00	2.00	5.00	6.11	**9.00**	**14.00**	4.2

Table 3.7: Descriptive statistics of dishwasher programs.

The Dishwasher

Dishwashers consume water very similarly to washing machines. They have different programs and they are able to use the appropriate amount of water, depending on the load; so variability is very high. Therefore the variables considered are the same, and the summary statistics are presented in Table 3.7.

The numbers in Table 3.7 have to be considered very cautiously, because they are based on problematic data (loggers not functioning properly). However, and given that we had very few "clean" dishwasher observations, some facts could be derived, especially those corresponding to the columns in bold (numbers in the other columns have been represented in smaller typeface to consider their unreliability.

The Internal Taps

Tap use on a property is involved in various activities and housework. This is the reason behind the very high variability.

Table 3.8 shows the main statistics for all the programs associated to the kitchen basin:

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Stdev.**
Duration (sec.)	0.9	5.20	14.70	38.96	37.62	1087.0	72.44
Consumption (l.)	0.02	0.37	0.9	2.68	2.62	102.6	5.33
Water uses	1.0	1.0	1.0	2.6	3.0	31.0	3.02

Table 3.8: Descriptive statistics of the Kitchen Basin.

It is clear that there are extreme programs with an extreme maximum duration and consumption. Most of the programs have short duration and low consumption; but there are a negligible number of programs (again, for characterizing and identifying microcomponents) with very high consumption as well as very long duration.

The results of calculating the same statistics for the programs belonging to the microcomponents of the bath tap are presented in Table 3.9.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Stdev.**
Duration (sec.)	3.50	5.20	14.70	22.51	29.40	1753.4	44.59
Consumption (l.)	0.1	0.32	0.72	1.34	1.70	22.97	1.79
Water uses	1.0	1.0	1.0	1.54	2.0	15.0	1.17

Table 3.9: Descriptive statistics of the Bath Tap.

Once again eliminating the extreme programs, the distribution of duration and consumption is close to the distributions of the kitchen basin. This feature makes it difficult to differentiate the processes of bath taps and kitchen basin programs. Clearly it is very complicated, if not impossible, to find a criterion to separate and distinguish them.

Since consumption differences between kitchen basins and bath taps are not easy to identify, from now on the two devices will be treated under the unique use of internal taps.

3.3.5 Variability in the Database

A very interesting finding appeared during the analysis of device consumption and duration patterns inside each household. Our initial expectation was that variability would be much lower; that is, that the variability shown in the above analysis would vary mainly "between households". To our surprise, the independent data analysis of each household showed a similar degree of variability. In other words, the "within household" variability is at least as significant to total variability as that of the "between household" differences.

Figure 3.6 shows that the washing machine in household number 8 (Fig. 3.6a) exhibits different consumption (l.) and duration (sec.) profiles. It is very common that modern washing machines allow programming of various features such as temperature or rotation speed settings, among others. For this reason, the type of clothing in a household can cause great variation in how a washing machine uses water. In comparison to household number 2 (Fig. 3.6b), we can see that the same program is normally used for all loads; thus, there is low variability within the household. The use of a single program differs from household 8, and therefore there is greater variability between households. Household numbers 8 and 2 represent the respective maximum and minimum intra-variability among washing machine programs.

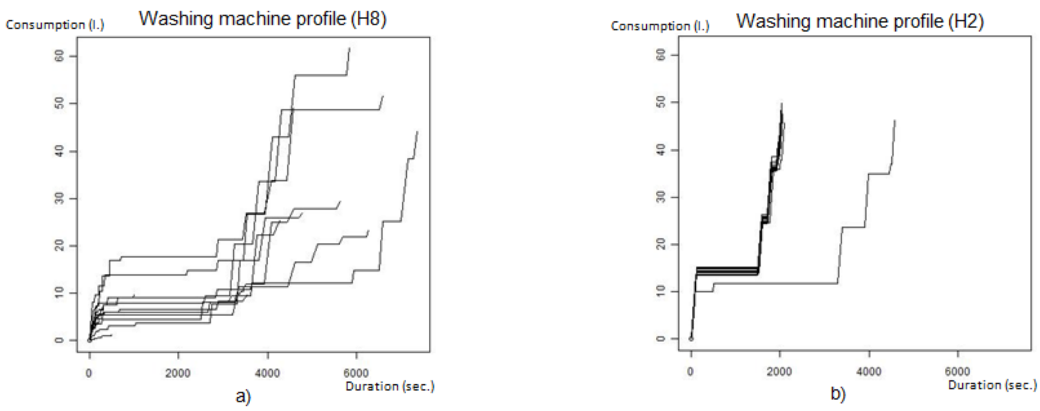


Figure 3.6: Example of washing machine variability between household 8 (H8) and household 2 (H2), and within them.

A similar effect occurs with the cistern profiles shown in Fig. 3.7. In this case, there is high variability within household number 8, where we can see (Fig. 3.7a) differences in cistern duration and consumption, probably because both a single and a dual flush cistern are present in the same household. Otherwise, household number 5 (Fig. 3.7b) is dominated by only one single-flush cistern, with a 10-liter mean consumption and 50-second duration. Household numbers 8 and 5 represent the respective maximum and minimum intravariability cistern profiles.

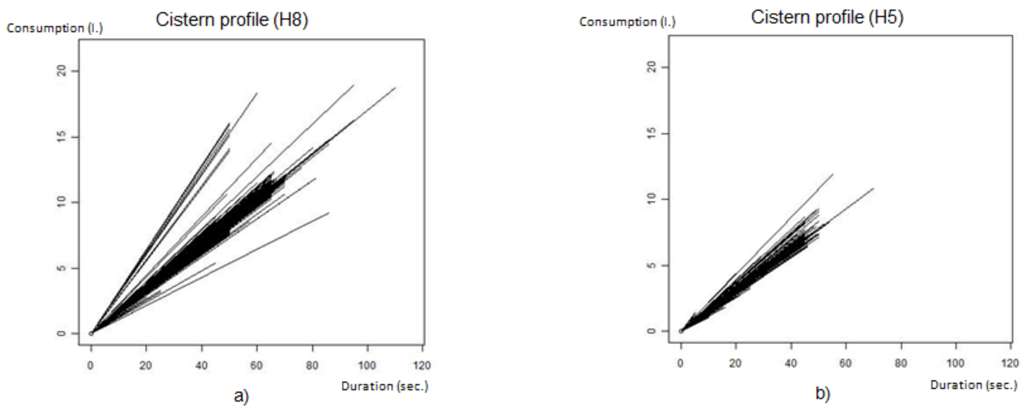


Figure 3.7: Example of cistern variability between household 8 (H8) and household 5 (H5), and within them.

3.4 Methodology

Statistically speaking, there was no previous, specific methodology for identifying which device consumes the water that passes through the water meter. Therefore, we faced the difficult task of devising a methodology. The proposed approach is a combination of a set of statistical techniques. The final procedure is the result of many trial and error efforts, and only a few of them are presented here.

3.4.1 First Approach: Profile Recognition

Given the project objectives, we first tried using pattern recognition techniques to identify which devices consumed the water indicated by the meter. Pattern recognition aims to classify different patterns based either on a priori knowledge or on statistical information extracted from the patterns. It seemed to be a good strategy, since each device has a pattern (or profile) based on the duration and the level of consumption. The first step tried to define the patterns to be classified. In this case the patterns to be classified were groups of water measurements that defined different device profiles. The schematic idea is to compare the pattern of a given device over the actual consumption profile (the measured values) and to do this over time; every time a god fit is identified this consumption is labeled as corresponding to this device and the corresponding consumption is extracted. The process is repeated for all device patterns. However, we tried several approaches, all of which were effective only for the cistern.

For all the remaining devices, the curve could not be estimated (or would be estimated with a very high variability, rendering it useless) because of the high variability from position and duration of water use inside each program. Moreover, the simultaneity of programs had a direct impact on flow volume, which distorted the metered value. Thus, the proposal of pattern recognition had to be discarded.

3.4.2 The Importance of Water Uses

At this point of the process, studying the water uses –i.e., the location and duration within the programs– would solve the variability problem.

Several attempts failed to fit a curve (estimated by several different procedures) to the water uses, though they took into account the program and considered them independent entities. The water uses also varied greatly, but they also indicated a good possibility of finding a statistical solution for summarizing the information acquired.

The proposal was to simplify the water use into its primary statistics. That is, using duration (in seconds), consumption (in liters) and flow (in liters/second) as working variables instead of the curve. This proposal simplified the information and allowed better treatment of meter variability (Fig. 3.8).

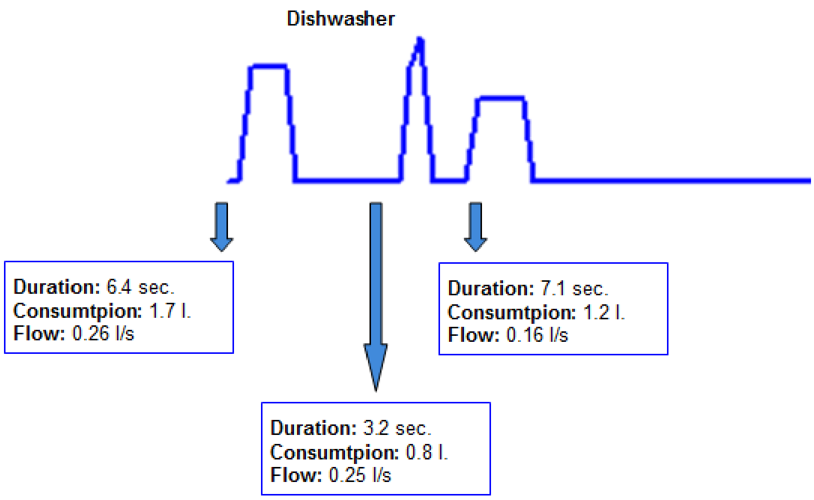


Figure 3.8: From curve characterization to water use characterization.

3.4.3 Combination of Water Uses

Once the water uses considered as independent entities (not part of programs) were characterized, the next problem was to associate them to a program based only on the information provided by the water meter. The idea was to find combinations of water uses that constituted a feasible program. Of course, on many occasions there were many different ways to combine water uses that produced feasible programs; thus a new problem arose: choosing among them (Fig. 3.9). The selection criteria was very important and it was necessary to keep in mind that most devices (washing machines, dishwashers, showers and internal taps) use water more than once during a program. The one exception is cisterns.

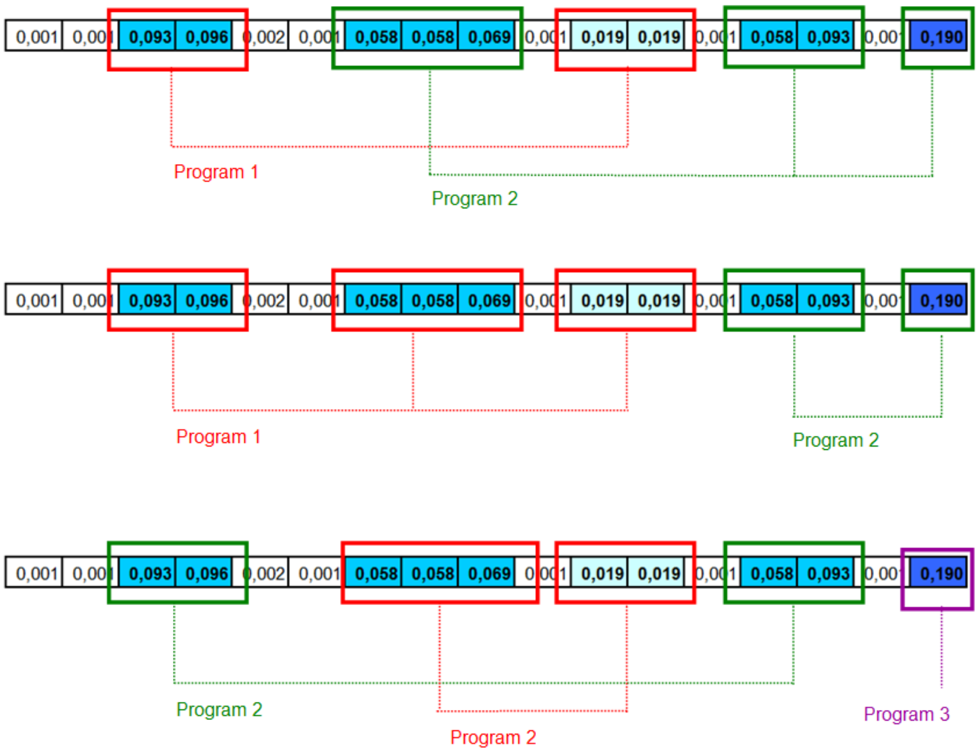


Figure 3.9: Multiple combination of water uses to create programs.

3.4.4 Evaluation of Variants in Identifying Water Uses

The combinatorial problem of grouping a random number of water uses associated to programs is very extensive. Computationally, not all combinations could be tested. Thus, it was necessary to create a methodology based on three steps:

1. Design a criterion for assigning the combinations.
2. Create a mechanism for producing variants that increase the probability of testing the correct combination.
3. Evaluate the programs and assign the use.

Designing a Criterion for Assigning the Combinations

Common sense indicates that a program is composed of a logical number of nearby water uses. We would consider that two consecutive water uses will belong to different programs when the time period between them exceeds 60 min for the washing machine and dishwasher, 15 min for the shower and 10 min for internal taps. Cisterns are considered to be generated by a singular water use (Fig. 3.10).

Because of the discontinuous water uses belonging to the same programs, especially the washing machines and dishwashers, it was necessary to identify some criteria for differentiating devices. In describing the different events, it was evident that the initial and final water uses of each device were quite particular, which helped classify the total program. This was the reason why the initial and final uses were used to evaluate the potential variants.

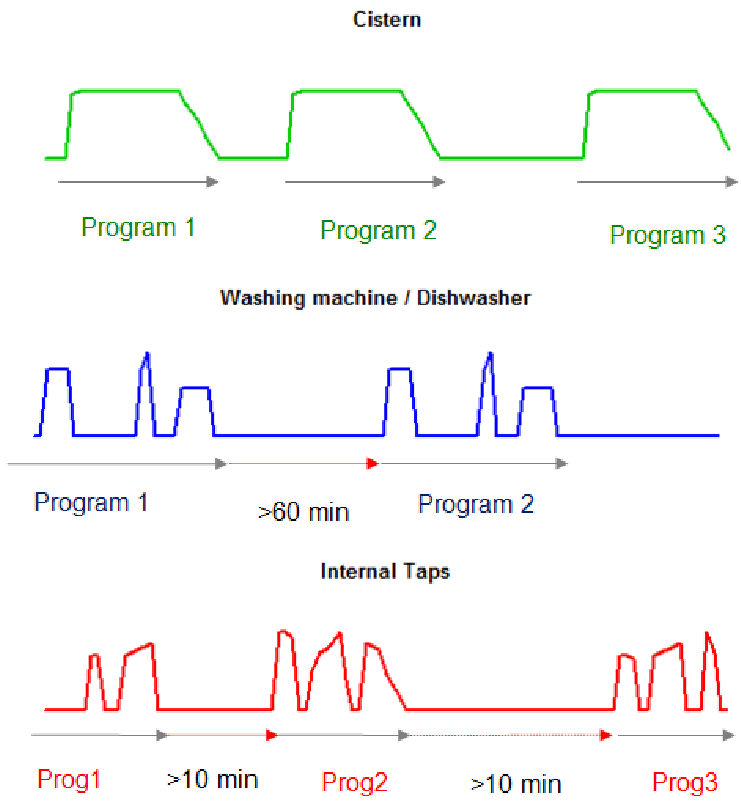


Figure 3.10: Example of maximum time allowed between water uses within the same program.

Generating the Variants

The first idea was to generate different possible programs (called "variants") by randomly grouping consecutive water uses. However, after some trials the procedure presented several complications, one of which was the improbability of calculating the correct program; so that procedure was discarded.

Then the investigation proceeded toward applying non-random heuristics. After designing and evaluating several heuristics, one was chosen which provided a very good relationship between computing time and accurate program identification. It began with a "compartment" composed of a consecutive number of water uses. This number had to be large enough to contain any possible program and small enough to maintain software execution time within feasible levels. The size of this compartment was treated as a random variable, so it was optimized to better identify the microcomponent within a reasonable time limit.

Once a compartment is established (keep in mind that it is moved along the water meter reading time series), the algorithm creates all possible variants obtained by removing one, two or three combinations of water uses inside it. Then each water use removed is evaluated to be classified as a possible cistern. All the variants generated in this way are tested as possible programs of any of the devices. In general, considering a compartment of m water uses, the number of variations of k water uses is:

$$\mathrm{Var}_{m,k} = \frac{m!}{(m-k)!k!}, \quad \text{where } m! = m(m-1)(m-2)\cdots 1$$

Evaluation of Variants for Identifying Programs

All the variants obtained with this heuristic are sequences of water uses which potentially belonged to any of the available microcomponents. To evaluate and classify each of the device programs, the heuristic calculated five variables for each sequence. These statistics were:

Referring to the program:

X_1 Total program duration (sec.): time between the beginning of the initial water use and the end of the final water use.

X_2 Total consumption (l.): total amount of water.

X_3 Time of service (sec.): total time water is flowing.

X_4 Mean Flow (l/sec.): total consumption divided by time of service.

X_5 Number of water uses.

Referring to the initial water use:

X_6 Total consumption (l.): total amount of water that flows during the initial water use.

X_7 Time of service (sec.): duration of initial water use.

X_8 First Delay (sec.): time between the end of the initial water use and the beginning of the second (by convention, 1 s if there is only one water use).

Referring to the final water use:

X_9 Total consumption (l.): total amount of water that flows during the final water use.

X_{10} Time of service (sec.): duration of the final water use.

Having the statistics, the program uses the likelihood function to apply a mathematical methodology related to the matching templates. In this problem the templates are the descriptive analysis of each device explained in section 3.3.4 (Descriptive analyses of the devices). The best approximation to the cistern model is the median values of duration, consumption, etc., shown in the statistical summary of the eight monitored properties. Thus, there are 5 available templates (or models), each one related to an device: cistern, shower, washing machine, dishwasher and internal taps. These templates are general and robust, since they summarize the behavior and variability of the different devices used by different users in different households.

The likelihood function is a function of the parameters of a statistical model. In this case, it depends on all the ten variables considered. In our case, the likelihood will help indicate the probability that a potential program belongs to any of the device models. For each of the obtained variants, the likelihood function is evaluated using a similarity measure. The device model with a maximum likelihood will be the device assigned to the potential program.

The empirical density shape of each water use time and consumption tend to be asymmetric –not surprising since none of these variables can take values lower than zero– and similar to an exponential decaying function. This distribution is very common in practice, and can be modeled as a Log-Normal parametric distribution. This probability family corresponds to situations where the logarithmic transformation of the data behaves as a Normal distribution; and thus the logarithmic transformation was applied to all time and consumption variables. It was also applied to variable X_5 (number of water uses) that reflects a counting process (a usual practice in counting variables; and to variable X_4 (mean flow) to enhance its symmetry and Normal distributed behavior.

Some of the transformed variables have bimodality. This indicates a mixed use pattern in the same household. But as a general approach, a Multivariate Gaussian model (Tong [1989]) is assumed for the vector of variables. For simplicity and assurance of a more robust model, correlations among variables are avoided. The model for the program description is marginally defined by its two first centered moments (mean and variance) in each component. The number of model parameters is 20 (two parameters for each component of the vector of descriptors).

General Model. By considering a common model for all the households, the Random Vector of Descriptors for use u, occurrence j has a multivariate Gaussian distribution:

$$(X_1, ..., X_{10})_{uj} \sim N((\mu_{u1}, ..., \mu_{u10}), \Sigma = \mathrm{diag}(\sigma^2_{u1}, ..., \sigma^2_{u10}))\,.$$

For a probabilistic model like this, a measure of the likelihood of one sample occurrence is the Normal density function applied to the values of the vector of statistics. If **??** $= (\mu_{u1}, ..., \mu_{u10}, \sigma_{u1}, ..., \sigma_{u10})$ represents the set of parameters describing use u for the general level, the likelihood for a specific program can be evaluated by this expression:

$$L(\Theta_u; \tilde{X}_{uj}) = f(\tilde{X}_{uj}; \Theta_u) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2_{ui}}} e^{-\frac{(x_{ui}-\mu_{ui})^2}{2\sigma^2_{ui}}}\,.$$

The parameters were estimated by Maximum Likelihood criterion (Casella and Berger [2001]) with the sample programs. For the General Model, all programs of each use are included. On the other hand, to estimate the Specific Level parameters, only the programs of each use for the specific household have been considered. The Maximum Likelihood estimators in this case correspond to the sample mean and variance for each component, calculated over the log-transformed data.

Although the Likelihood function has no units, it is possible to compare several parameters for a random vector in order to establish which of the different parameters it most likely corresponds to. So, for each potential program (artificial combination of water uses), the likelihood for all uses is calculated and the maximum value obtained indicates which use is more probable. The algorithm for assigning use to each program is based on this evaluation of all the artificial programs.

3.5 Results

The algorithm can be evaluated by comparing the microcompenents it identifies with the devices identified by the loggers.

The efficacy of the assignation process can be summarized in two different tables, where the identified water uses (tables reflect water uses instead of programs, because it is a better unit for making comparisons) are classified into two groups: correctly or wrongly identified. The two tables reflect two different ways of evaluating the results:

Sensitivity: Proportion of real water uses that the process has classified correctly.

Specificity: Proportion of the water uses that are well-classified.

The following tables provide the sensitivity and specificity results, using the specific model for each household:

Table 3.10 shows that the process correctly identifies 70% of the water uses that represent 67.8% of the monitored consumption. This percentage is the average of all devices. Internal taps and cisterns constitute the maximum rates with 76.9% and 62.8%, respectively. In spite of the fact that only 51.9% of showers are well classified, they consume 90.75% for this specific device. The minimum rates belong to devices with long programs: 50% of dishwashers and 30.78% of washing machines.

Table 3.11 shows that for all the programs identified as Internal taps, 91.37% are internal taps. This percentage is very high because the internal taps category works as an absorbent state, including more water uses than would be correct. 72.4% of all identified cisterns are correct and in the case of showers this percentage decreases to 49.03%. From this point of view, washing machine identification improves, with 69.41% of their water uses classified well. Dishwashers are problematic, since only 9.0% of the water uses associated to dishwashers are correct.

Real Water Uses		**CORRECT**	**WRONG**	**Total**
Internal Tap	Water Uses	8721	2611	11332
	% Water Uses	76.96%	23.04%	100.00%
	Consumption	4715.112	6449.516	11164.628
	% Consumption	42.23%	57.77%	100,00%
Cistern	Water Uses	1932	1145	3077
	% Water Uses	62.79%	37.21%	100,00%
	Consumption	10514.491	6598.271	17112.762
	% Consumption	61.44%	38.56%	100.00%
Shower	Water Uses	655	606	1261
	% Water Uses	51.94%	48.06%	100.00%
	Consumption	19507.844	1987.901	21495.745
	% Consumption	90.75%	9.25%	100.00%
Washing Machine	Water Uses	177	398	575
	% Water Uses	30.78%	69.22%	100.00%
	Consumption	968.555	1658.994	2627.549
	% Consumption	36.86%	63.14%	100.00%
Dishwasher	Water Uses	268	268	536
	% Water Uses	50.00%	50.00%	100.00%
	Consumption	504.891	388.697	893.588
	% Consumption	56.50%	43.50%	100.00%
Total of Programs		11753	5028	16781
Total of % Programs		70.04%	29.96%	100.00%
Total of Consumption		36210.892	17083.379	53294.271
Total of % consumption		67.95%	32.05%	100.00%

Table 3.10: Sensitivity table.

Programs identified as...		**CORRECT**	**WRONG**	**Total**
Internal Tap	Water Uses	8721	824	9545
	% Water Uses	91.37%	8.63%	100.00%
	Consumption	4715.112	1056.291	5771.403
	% Consumption	81.70%	18.30%	100.00%
Cistern	Water Uses	1932	736	2668
	% Water Uses	72.41%	27.59%	100.00%
	Consumption	10514.491	3299.119	13813.610
	% Consumption	76.12%	23.88%	100.00%
Shower	Water Uses	655	681	1336
	% Water Uses	49.03%	50.97%	100.00%
	Consumption	19507.844	6377.045	25884.889
	% Consumption	75.36%	24.64%	100.00%
Washing Machine	Water Uses	177	78	255
	% Water Uses	69.41%	30.59%	100.00%
	Consumption	968.555	325.986	1294.541
	% Consumption	74.82%	25.18%	100.00%
Dishwasher	Water Uses	268	2709	2977
	% Water Uses	9.00%	91.00%	100.00%
	Consumption	504.891	6024.937	6529.828
	% Consumption	7.73%	92.27%	100.00%
Total of Programs		11753	11753	5028
Total of % Programs		70.04%	70.04%	29.96%
Total of Consumption		36210.892	36210.892	17083.379
Total of % consumption		67.95%	67.95%	32.05%

Table 3.11: Specificity table.

3.6 Concluding Remarks

The problem of assigning microcomponents to devices based only on the water meter reading is very difficult because of the scarcity of information available: the water consumption time series and the variety of patterns in which people use different devices.

The method proposed is an algorithm that characterizes the water consumption of each device observed in Barcelona as initial parameters and correctly identifies programs and consumption about 70% of the time, which in these circumstances can be considered high. In our opinion, and given the variability already commented upon, it will be very difficult to significantly increase this percentage without increasing in one way or another information available for making the assignation.

The efficacy of the assignation process is summarized from two different points of view: the sensitivity and the specificity of the process. The sensitivity measures the proportion of devices that the process has classified correctly. The specificity refers to the proportion of the classified devices that are well classified.

The algorithm correctly identifies the devices between 70% and 80% of the time, or even higher, depending on how well the chosen parameters reflect the household consumption patterns. Considering the high variability of water consumption patterns –both between households and within each household (different persons)– and the fact that uses are characterized only on the basis of the aggregate consumption provided by the water meter, the results are quite satisfactory.

The algorithm has been implemented in a software program that has been optimized for processing time. The latest version takes between 5 and 10 min to process the consumption time series corresponding to 1 day of one household.

An additional feature of the devised method is that the water identified uses can be truncated according to the value of their likelihood. The likelihood can be interpreted as a measure of the probability that the assigned use is correct, so a minimum level of certainty can be defined. This means not classifying a certain percentage of water uses, let's say the 10% of those who have a lower likelihood (higher "probability" of being incorrect) in exchange for greater certainty of correct classification in the remaining 90%.

Once the Barcelona project was finished, the methodology was applied to ten houses in Murcia. The results were satisfactory obtaining a 71.8% of the programs well identified, representing the 60.8% of the total water consumption. The houses were selected using variability information. The experience from working in Barcelona informed the various changes introduced into the algorithm for improving the sensitivity and specificity of identification. The primary aims were to increase the general percentage of identification, improve the results for the devices with non-continuous water use, and to introduce a stage of learning procedure into the algorithm. The Murcia application was a validation stage that established a general methodology to be used in other municipalities.

In what follows, we briefly comment on possible ways to further improve the algorithm, which will hopefully result in small improvements in assignation processing and also in ways to increase the information available for making the assignment.

Possible algorithm improvements:

- Further investigation of variables that will help characterize long programs (washers and dishwashers) and thus, hopefully, have a higher correct identification rate.
- Devise ways to incorporate information regarding the number of persons living in the household.
- Although, in applying a general model to all households, we found no significant improvements in the assignment process by using device models specific to each household, a possible (however complicated and uncertain) improvement may be to introduce some type of learning procedure in the algorithm, so that it adapts itself to the household habits. This would be very difficult without human intervention.
- The group of internal taps presents very high variability; so it is not surprising that it absorbs water uses that belong to other devices, especially showers, dishwashers and washing machines. At this point, it is not clear how to mitigate this fact. It may not even be possible, but it certainly would be a nice improvement.

Acknowledgements. The authors are grateful to R+I Alliance for the financial support that made it possible to develop this project. The authors are also grateful to AQUAGEST SOLUTONS for their very useful comments and suggestions during the preparation of this manuscript.

Bibliography

Casella, G. and Berger, L. *Statistical inference.* Thomson Learning, United States, 2nd edition, 2001.

Eurostat. Consumers in europe – Facts and figures on services of general interest. *http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-DY-07-001/EN/KS-DY-07-001-EN.PDF*, page Accessed 16 April 2012, 2007.

Farinaccio, L. and Zmeureanu, R. Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses. *Energy and Buildings*, 30:245–259, 1999.

Fontdecaba, S., Grima, P., Marco, L., Rodero, L., Sánchez-Espigares, J., Solé, I., Tort-Martorell, X., Demessence, D., Martínez-De-Pablo, V., and Zubelzu, J. A methodology to model water demand based on the identification of homogenous client segments. Application to the city of Barcelona. *Water Resources Management*, 26:499–516, 2012.

Mayer, W. P., William, B., DeOreo, T. E., and Lewis, D. M. Residential indoor water conservation study: evaluation of high efficiency indoor plumbing fixture retrofits in single-family homes in the East BayMunicipal Utility District Service Area. Prepared to East BayMunicipal Utility District and The United States Environmental Protection Agency. 2003.

Richter, C. and Stamminger, R. Water consumption in the kitchen – a case study in four European countries. *Water Resources Management*, 26:1639–1649, 2012.

Tong, Y. L. *The multivariate normal distribution*. Springer Series in Statistics, United States, 1st edition, 1989.

Yamagami, S. and Nakamura, H. Non-intrusive submetering of residential gas appliances. *ACEEE Summer Study on Energy Efficiency in Buildings*, pages 265–273, 1996.

Chapter

4

Article 3

Proposal of a Single Critical Value for the Lenth Method

Sara Fontdecaba, Pere Grima and Xavier Tort-Martorell

Department of Statistics and Operational Research, Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, Spain

(*Received* September 2013, *accepted* February 2014)

Abstract. Different critical values deduced by simulation have been proposed that greatly improve Lenth's original proposal. However, these simulations assume that all effects are zero – something not realistic – producing bigger than desired critical values and thus significance levels lower that intended. This article, in accordance with Box [1992] well known idea that Experimental Design should be about learning and not about testing and based on studying how the presence of a realistic number and size of active effects affects critical values, proposes to use $t = 2$ for any number of runs equal or greater than 8. And it shows that this solution, in addition of being simpler, provides under reasonable realistic situations better results than those obtained by simulation.

Keywords: Industrial experimentation, Lenth method, pseudo standard error, significant effects, unreplicated factorial designs.

4.1 Introduction

Identifying which are the active factors is one of the steps required in analyzing the results of a factorial design. When replicates are available, the experimental error variance can be used to estimate the variance of the effects and judge whether they are active or not. In other cases it can be assumed that three or more factors interactions are zero and then their values can be used to estimate the effect's variance. Unfortunately, this estimate is made with very few degrees of freedom unless we are dealing with full factorial designs with five or more factors, or fractional designs with resolution greater than five, unusual situations in industrial environments where the number of runs is generally limited.

Daniel [1959] proposed a procedure based on the representation of the absolute value of the effects on a half-normal plot, since the non-active effects are distributed according to $N(0;\sigma_{ef})$ in these coordinates they are aligned approximately along a straight line that passes through the origin. A variant of this method, also used by Daniel [1959], is to represent the effects with their sign on a Normal Probability Plot (NPP). In this case, the non-significant effects are aligned according to a straight line which passes through the point $(0, 0.5)$.

Daniel's method, in the half-normal or NPP version, is the most used for analyzing the "significance" of effects in the absence of replicates; although, being based on a graphic interpretation, the conclusions that are drawn are subjective and may be debatable. What's more, their proper use requires a good deal of judgment and experience. If one simply applies the motto, "the active effects are those that are separated from the line," one can commit glaring errors, as described by De-León et al. [2011]. Another drawback from basing the observation on a graph is that it is not possible to automate the process of selecting active effects.

To solve these problems, Lenth [1989] proposed a method for estimating the standard deviation of the effects on the basis that if $X \sim N(0, \sigma)$, the median of $|X|$ is equal to 0.6475σ and therefore $1.5{\cdot}\text{median}\,|X| = 1.01\sigma \cong \sigma$. Considering that κ_i $(i = 1, ..., n)$ are the effects of interest and that their estimators c_i are distributed according to $N(\kappa_i, \sigma_{ef})$, he defines $s_0 = 1.5{\cdot}\text{median}\,|c_i|$ and uses this value to calculate a new median by excluding the estimates $|c_i| > 2.5s_0$ with the intention of excluding those which correspond to effects with $\kappa > 0$. In this way he obtains the so-called *pseudo standard error*:

$$PSE = 1.5 \cdot \underset{|c_i|<2.5s_0}{\text{median}}\,|c_i|$$

Lenth verifies that the PSE is a σ_{ef} estimator which is reasonably good when the significant effects are rare (sparse) and he proposes calculating a margin of error for c_i by means of $ME = t_{1-\alpha/2,d} \cdot PSE$ for a confidence level of $1 - \alpha$ and where t is distributed according to a Student's t distribution with $d = n/3$ degrees of freedom. On the other hand, since n significance tests are performed simultaneously, calculating a simultaneous margin of error (SME) is also proposed, which is obviously greater than the ME. Finally, he proposes performing a graphical analysis rather than formal significance tests. He suggests graphics that are similar to those used by Ott [1975] for the analysis of means. If $c_i < ME$ should be considered non-significant, if c_i is larger than ME but smaller than SME could be described as possibly active and if $c_i > SME$, then the effect is probably active.

Since its publication, the Lenth method has gained prominence. In fact, it has become the most employed method, being described in the books which are most prevalent in the field of industrial experimental design, such as those by Montgomery [2008] and Box et al. [2005]. It is also used by statistical software packages which are very common in the field of industrial statistics, such as MINITAB or SigmaXL that use the critical values originally proposed by Lenth or JMP which calculates critical values *ad hoc* for each case by simulation.

In our opinion the principal advantage of Lenth method is that it uses a simple and effective procedure (given the scarcity of the information available) to estimate the standard deviation of the effects. However, the critical values originally proposed are not the most appropriate and in many cases produce results that a representation in NPP show that are clearly incorrect. For example, Box et al. [2005] classical book provides a 2^{4-1} design example, page 237. The analysis is accompanied by the comment that it is reasonable to consider that factors A and B influence the response, which is very consistent with what is observed in the NPP representation of the effects. But, for example Minitab that uses Lenth's method with its critical values only presents A as significant (Figure 4.1).

Although the errors are less frequent when the number of runs increases, these also occur. The well-known example (Box [1992]) of optimizing the design of a paper helicopter using experimental design techniques can be useful to illustrate that. The article presents the results of a 2^{8-4}, 16 run design, saying that factors B and probably C are active. Minitab's representation of the effects on NPP and identifications of the significant ones using Lenth's method with its original critical values only identifies B as active (Figure 4.2).

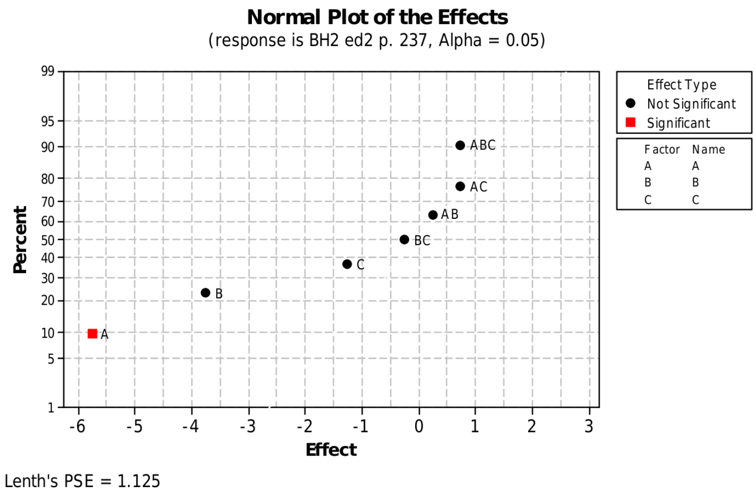


Figure 4.1: Effects of the Box, Hunter and Hunter example represented in NPP by Minitab. Significant ones identified by Lenth's method using its original critical values.

4.2 Alternatives to the Lenth method

Since its publication, many alternative methods to Lenth's proposal have appeared. Some, like Loughin [1998], show that a Student's t distribution with $d = n/3$ degrees

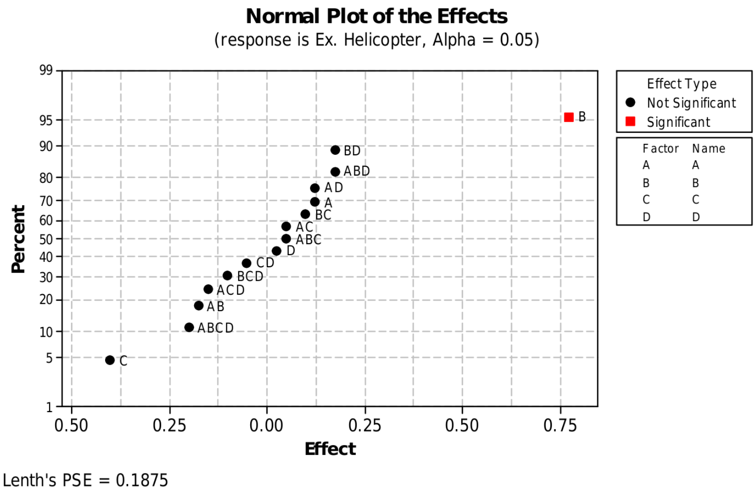


Figure 4.2: Effects of Box's helicopter paper represented by Minitab in NPP. Significant effects identified by Lenth's method using its original critical values.

of freedom is not a good reference distribution for Lenth's t-ratio, and he obtained by simulation that for $\alpha = 0.05$, it is necessary to have $t = 2.300$ and $t = 2.152$ for 8 and 16 runs, respectively. Along these lines, and also by simulation, Ye and Hamada [2000] determine the critical values for certain values of α by means of a computational method that is more efficient than Loughin. They also address three level and Plackett and Burman designs. In this case, the values obtained for $\alpha = 0.05$ and eight and sixteen-run designs are $t = 2.297$ and $t = 2.156$.

Other authors propose using different methods from Lenth but based on similar procedures, such as Dong [1993], who instead of using the effects' median to calculate the σ_{ef} estimator, uses the average of the squares and states that when the number of significant effects is low (less than 20%) and its value is not small (its tests are performed with a $\kappa \geq 5\sigma_{ef}$ value of significant effects), his method delivers better results than that of Lenth. In a similar vein, Juan and Peña [1992] deal with the problem by identifying outliers (significant effects) in a sample and, under certain circumstances, the Lenth method is a particular case of the one they propose. They claim that their method works better than Lenth when the number of active effects is large (greater than 20%), but it is more complicated.

Other strategies are those that may be called "step by step", like that of Venter and Steel [1998], which, through a simulation study for sixteen-run designs (with the same configurations that we have used), conclude that their σ_{ef} estimator is as good as Lenth's PSE when there are few significant effects, and that it is much better when there are many. Ye et al. [2001] also propose a method they call the "step-down version of the Lenth method". They begin by calculating the Lenth statistic (t-Lenth) for the biggest effect and, if it is considered significant, the PSE is recalculated with the remaining effects and a new t-Lenth is obtained for biggest of them. This is then compared with a critical value obtained by simulation in accordance with the number of effects considered. The procedure is repeated until the biggest effect remaining is not considered significant. After performing a simulation study, they conclude that their results are better than those obtained with the original Lenth method or that of Venter and Steel [1998]. Mee et al. [2011] also propose an iterative system, but in this case of stepwise regression assuming that at least a certain number of non-significant effects exist.

Edwards and Mee [2008] propose to calculate by simulation the p-value that corresponds to the t-ratios. To calculate the critical values they use a method similar to that of Ye and Hamada, but it can be applied to a wide variety of designs, particularly those which are not orthogonal. Bergquist et al. [2011] propose a Bayesian procedure. They begin with a study of published cases to establish the *a priori* probabilities based on the principles of sparsity, hierarchy and heredity, and then illustrate their method by analyzing cases in the literature.

A proposal that we find particularly pragmatic is that of Costa and Lopez-Pereira [2007], which takes advantage of current computational capabilities. Since most of the methods are simple to apply, they propose using various methods. If all of them give the same solution, it is clear that this it is a good one. On the contrary if there is a discrepancy, the solution can be decided by a majority, perhaps with careful consideration of the method or of the characteristics of the situation being analyzed. When there is a great disparity between some methods and others, the solution is certainly not clear; not for any failure of the methods but for lack of information, and the only way out is to perform new experiments.

Detailed studies and comparisons of existing methods have also been published, such as Haaland and O'Connell [1995]. After a simulation analysis of sixteen-run designs, they conclude that the Lenth method has good properties in a wide range of number and size of significant effects. Other alternatives, such as the previously mentioned Dong, are only recommended if the number of significant effects is known to be low. Another study, which is very complete and detailed, is that of Hamada and Balakrishnan [1998], where they summarize and discuss 17 methods and also propose another 5 that are new proposals or modifications of existing ones. They advise against the use of some methods and conclude that among those which perform reasonably well, there is no clear winner.

4.3 Lenth Method with Improved Critical Values

In our opinion, of all the alternatives considered, the ones that best combine simplicity and precision are those, in accordance with the proposals of Ye and Hamada [2000] or Edwards and Mee [2008], based on comparing the t-ratios with reference distributions obtained through simulation.

The problem with these t-ratios and the critical values proposed is that PSE values tend to be higher than those obtained by simulation under the assumption that all effects are zero. In practice, some non-null effects ($\kappa \neq 0$) may be incorporated into the PSE calculation producing an overestimation of σ_{ef} and therefore affecting the critical values. This drawback was already highlighted by Lenth and even improved versions as the widely acknowledged proposed by Ye and Hamada [2000] suffer from this problem. To further illustrate this phenomenon, figure 4.3(a) shows the distribution of 10,000 PSE

values obtained by simulation from a design with 8 runs, when all effects are values from a $N(0, 1)$ and Figure 4.3(b) gives the distribution obtained when there are five effects belonging to a distribution $N(0, 1)$ and two of a $N(2, 1)$. It can be clearly seen that the estimate of PSE tends to be higher in the latter case.

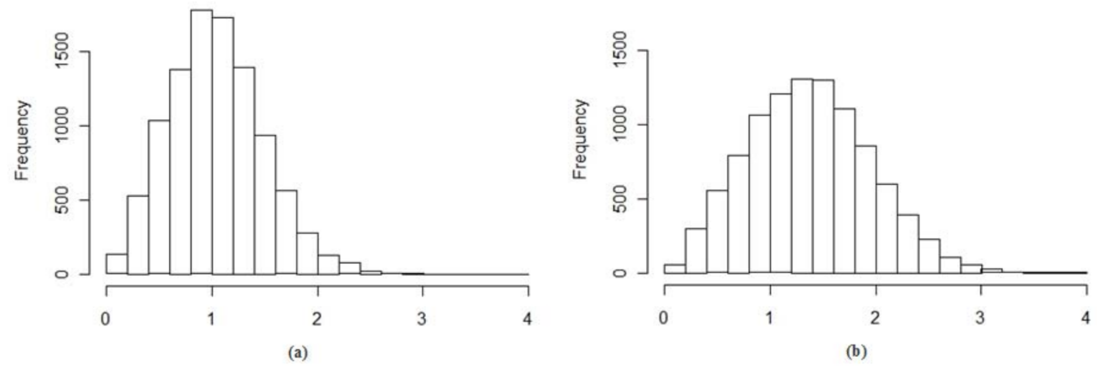


Figure 4.3: Distribution of PSE values in a design with 8 experiments. (a) when all 7 effects belong to a $N(0, 1)$ and (b) when five belong to the $N(0, 1)$ and two to $N(2, 1)$.

Depending on the number and magnitude of non-null effects ($\kappa \neq 0$) the overestimation of σ_{ef} can be higher or lower. To show this fact we have simulated a set of situations that reflect what one would expect in practice and calculated their PSE. For eight-run designs, we propose four configurations with up to three significant effects:

C1: $\kappa_1 = \cdots = \kappa_6 = 0,\ \kappa_7 = \Delta$

C2: $\kappa_1 = \cdots = \kappa_5 = 0,\ \kappa_6 = \kappa_7 = \Delta$

C3: $\kappa_1 = \cdots = \kappa_4 = 0,\ \kappa_5 = \kappa_6 = \kappa_7 = \Delta$

C4: $\kappa_1 = \cdots = \kappa_4 = 0,\ \kappa_5 = \Delta,\ \kappa_6 = 2\Delta,\ \kappa_7 = 3\Delta\,.$

For 16-runs designs we have used the same 6 configurations that both Ye et al. [2001] and Venter and Steel [1998] use in their proposals. They are:

C1: $\kappa_1 = \cdots = \kappa_{14} = 0,\ \kappa_{15} = \Delta$

C2: $\kappa_1 = \cdots = \kappa_{12} = 0,\ \kappa_{13} = \kappa_{14} = \kappa_{15} = \Delta$

C3: $\kappa_1 = \cdots = \kappa_{10} = 0,\ \kappa_{11} = \cdots = \kappa_{15} = \Delta$

C4: $\kappa_1 = \cdots = \kappa_8 = 0,\ \kappa_9 = \cdots = \kappa_{15} = \Delta$

C5: $\kappa_1 = \cdots = \kappa_{12} = 0,\ \kappa_{13} = \Delta,\ \kappa_{14} = 2\Delta,\ \kappa_{15} = 3\Delta$

C6: $\kappa_1 = \cdots = \kappa_{10} = 0,\ \kappa_{11} = \Delta,\ \kappa_{12} = 2\Delta,\ \kappa_{13} = 3\Delta,\ \kappa_{14} = 4\Delta,\ \kappa_{15} = 5\Delta\,.$

In all cases, the effects were generated as independent values of a Normal distribution with the specified mean and $\sigma_{ef} = 1$. As Ye et al. [2001], we call the Δ parameter *Spacing* and its value varies from 0.5 to 8 in steps of 0.5. For each of the configuration-spacing combinations, 10,000 situations were simulated using the R statistical package and for each one of them we have calculated the PSE. Figures 4.4 and 4.5 clearly confirm that, in eight as well as sixteen-run designs, the PSE overestimates the value of σ_{ef}.

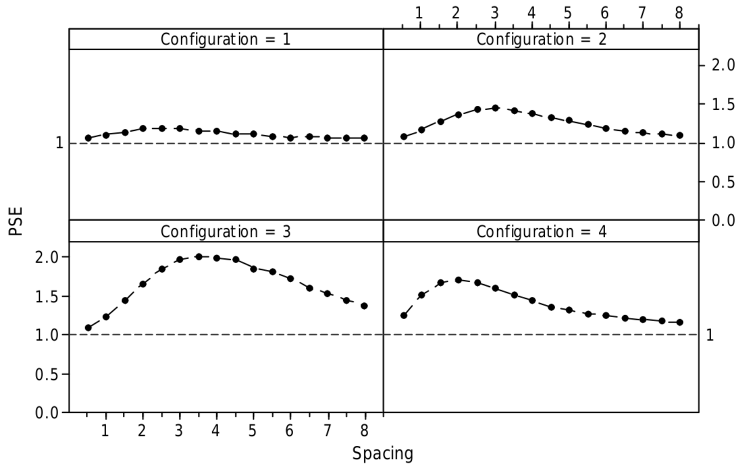


Figure 4.4: Eight run design. Average of 10,000 PSE estimates made for each configuration-spacing combination.

The fact that in practice PSE values tend to be higher than those obtained by simulation, suggest to use critical values smaller than the ones deduced by this procedure. At the same time it is not possible to propose a critical value that will always produce the desired significance level since it depends on the number and magnitude of significant effects, which is precisely what we want to know.

Therefore, using critical values that seem very accurate -with several decimal places- is misleading; it conveys the idea that the significance test has a meticulousness that it lacks. In accordance with Box's idea mentioned in abstract that aim of Design of Experiments should be to learn about the subject and not to test hypothesis, Box et al. [2005] do not perform significance tests for each effect, nor even when they have replicates. They place the effect values in a table together with their standard deviation in such a way that, in light of this information, the experimenter decides what factors should be considered to have an influence on the response. To go further than that when there are no replicates and, as such, less information is available, it doesn't make sense. Lenth himself doesn't even propose conducting formal significance tests, but rather an indicative graphical representation, as has been previously mentioned. Depending on which area of the graph the effects fall, they should be considered significant or not with higher or lower probability.

Paul Velleman, a student of John Tukey and a statistician at Cornell University, recounts that when he asked the reason for the 1.5 value when determining the area of anomalies in the boxplots, Tukey answered because 1 was too small and 2 was too large (De-Veaux et al. [2012], p. 91) and the number halfway between 1 and 2 is 1.5. With this same idea of simplicity, and in the absence of an exact value, our proposal is to always use a critical value of $t = 2$, independently of the number of runs conducted.

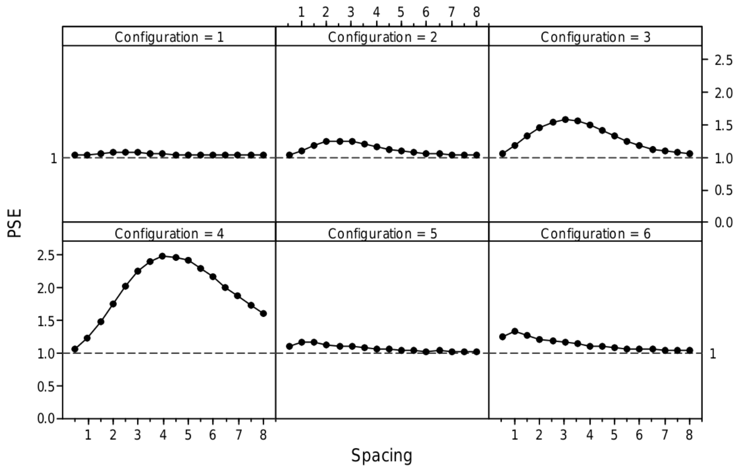


Figure 4.5: Sixteen run design. Average of 10,000 PSE estimates made for each configuration-spacing combination.

To show the usefulness and good properties of this value we compare the results produced by $t = 2$ with the ones with the t values proposed by Ye and Hamada [2000]. The comparison is made by computing the percentages of type I and type II errors that occur in each of the scenarios presented in the previous section. The significant effects identification error rates were calculated as the number of errors committed with respect to the total number that could have been committed. For example, for eight-run experiments in configuration 1 and with $\Delta = 0.5$ and considering the critical value proposed by Ye and Hamada ($t = 2.30$) 2759 type I errors have been obtained (effects which are considered significant when in reality they correspond to a distribution with $\kappa = 0$) and 9329 type II errors (effects with $\kappa \neq 0$ but which are considered not significant). As in configuration 1, there is only one effect with $\kappa \neq 0$. The error rates are:

$$\text{type I: } \frac{2759}{6 \times 10000} \times 100 = 4.60\% \quad \text{and type II: } \frac{9329}{1 \times 10000} \times 100 = 93.29\% \,.$$

Figures 4.6 and 4.7 show the type I error rate for each configuration-spacing combination. Except in one case (16 runs experiment, configuration 1, one significant effect) the type I error rate is closer to the intended 5% with $t = 2$ than with the values proposed by Ye and Hamada [2000].

Figures 4.8 and 4.9 show the type II error rate for each configuration-spacing combination. Although the differences seem small, in 8 run designs exceeds 10% in configurations 1, 2 and 3 and 8% in configuration 4. In designs with 16 runs the difference exceeds 6% in configurations 1, 2, 3 and 4 and 3% in 5 and 6.

For designs with more than 16 runs the difference between the critical value of $t = 2$ and those obtained through simulation by Ye and Hamada [2000] decrease. For 32, 64 and 128 runs designs the same authors propose as critical values: 2.064, 2.013 and 1.986; as can be seen the difference is irrelevant for practical purposes. For bigger number of runs the estimation of σ_{ef} is increasingly accurate and thus, the critical value tends to 1.96

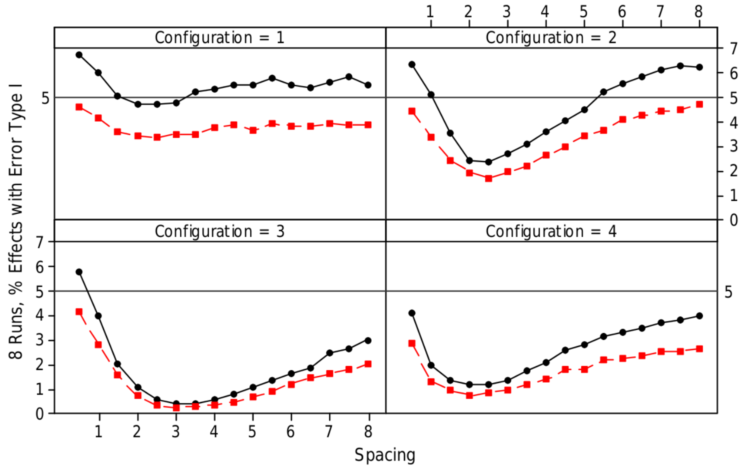


Figure 4.6: Eight-run designs. Percentage of Type I errors using $t = 2.3$ (square symbols) and $t = 2$ (round symbols).

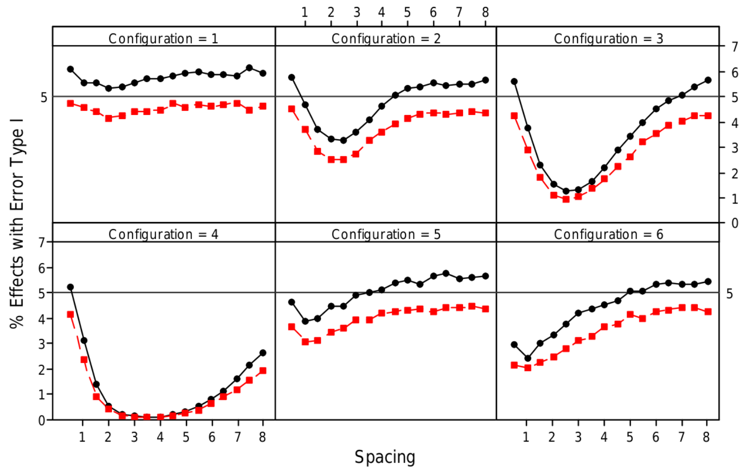


Figure 4.7: Sixteen-run designs. Percentage of Type I errors using $t = 2.156$ (square symbols) and $t = 2$ (round symbols).

($z_{0.025}$). Therefore, the differences with $t = 2$ keep being irrelevant.

Even with $t = 2$, the type I error rate is below 5% in many configuration-spacing combinations; naturally this implies a higher type II error risk, which is something unwanted in industrial contexts (De-León et al. [2006]). To mitigate this and in accordance with the already mentioned practice of Box et al. [2005] it seems adequate to define a doubtful zone for t-ratio values; a zone where it is not clear if the effects are significant or not but that the experimenter should be aware of this uncertainty. A reasonable choice for this zone could be the one corresponding to t-ratios between 1.5 and 2.

4.4 Conclusions

Among the many analytical procedures that have been proposed for identifying significant effects in not replicated two level factorial designs, the Lenth method constitutes

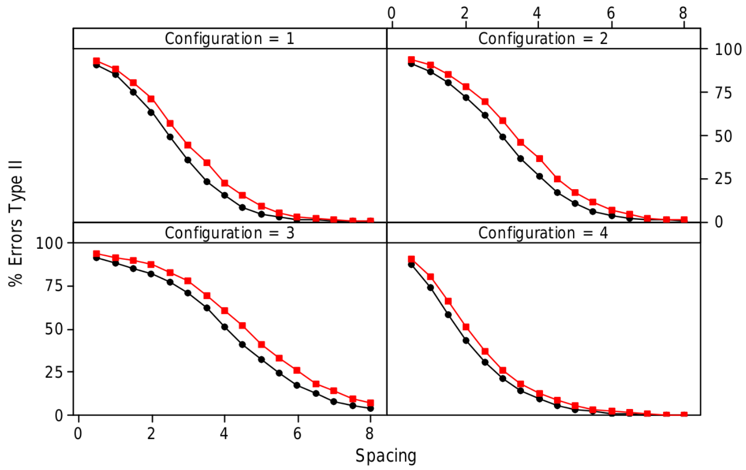


Figure 4.8: Eight-run designs. Percentage of type II errors using $t = 2.3$ (square symbols) and $t = 2$ (round symbols).

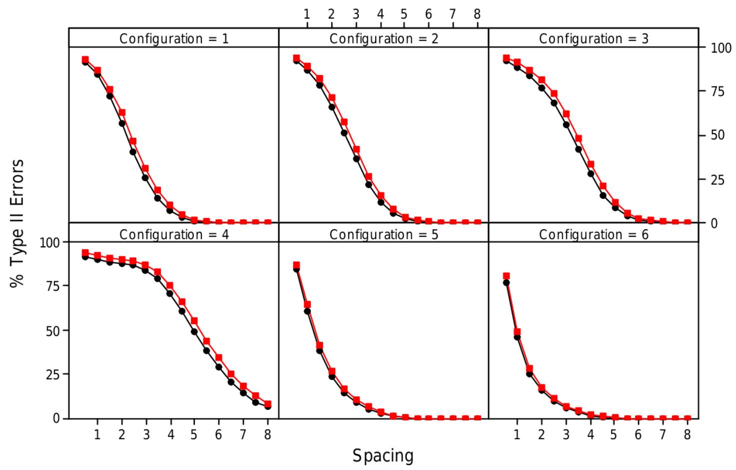


Figure 4.9: Sixteen-run designs. Percentage of type II errors using $t = 2.156$ (square symbols) and $t = 2$ (round symbols).

a central reference. It is included in well-known books on experimental design (Montgomery [2008]; Box et al. [2005]) and it is used by statistical software packages commonly employed in technical environments. And this is so, in spite that it is well known that the critical values employed lead to a type I error probability that is less than what is expected and, as an unwanted consequence, a greater probability of type II error. More accurate critical values derived by simulation produce the desired probability of a type I error when all factors included in the PSE calculation are inert. Unfortunately, the inclusion in the PSE calculation of one or more non-null factors, something difficult to avoid in practice, produces an overestimation of the σ_{ef} with the already mentioned undesired consequences.

The paper proposes a simple solution: to use a critical value of $t = 2$ for any 2^k or 2^{k-p} design. This proposal has several advantages:

- It conveys the idea that the procedure is approximate, it cannot be otherwise, and that therefore it requires a dose of good judgment by the experimenter. Completely in line with Box [1992] ideas of sequential experimentation and using DOE to learn.

- In a wide range of reasonable scenarios –number and size active effects– it produces type I errors closer to the desired and announced 5% than the values proposed by simulation methods.

- Finally, it is simple and easy and does not need simulations or tables.

Authors' Biographies:

Sara Fontdecaba has a Degree in Statistics from Universitat Politècnica de Catalunya (UPC) - BarcelonaTech and is a funded Ph.D. student at the Statistics Department of UPC. She is currently doing research in the topics of discrete time series analysis and design of experiments. She currently gives practical lectures at the Barcelona Industrial Engineering School. She has given presentations in different applied statistics seminars and she has collaborated in different consultancy projects.

Pere Grima is a Professor at the Universitat Politècnica de Catalunya - BarcelonaTech, where he also obtained his Ph.D.. One of the areas he specialises in is experimental design and he has more than ten years of experience in helping companies to implement statistical methods for quality control and improvement. He has been an advisor for the European Quality Award and has acted as a consultant to several multinational companies in Six Sigma projects.

Xavier Tort-Martorell has a Degree in Industrial Engineering, a Ph.D. in the same field and a Master's Degree in Statistics by the University of Wisconsin. He is currently a Professor at the Department of Statistics at UPC. He is the Director of UPC's Master's programme in TQM and the Six Sigma Black Belt training courses. Dr. Tort-Martorell has been a consultant in quality management and techniques to many firms and has taught seminars at many private and public organisations.

Bibliography

Bergquist, B., Vanhatalo, E., and Nordenvaad, M. L. A Bayesian analysis of unreplicated two-level factorials using effects sparsity, hierarchy, and heredity. *Quality Engineering*, 23:152–166, 2011.

Box, G. E. P. Teaching Engineers Experimental Design with a Paper Helicopter. *Quality Engineering*, 4(3):453–455, 1992.

Box, G. E. P., Hunter, J. S., and Hunter, W. G. *Statistics for Experiments: Design, Innovation and Discovery.* Wiley, Hoboken, 2nd edition, 2005.

Costa, N. and Lopez-Pereira, Z. Decision-making in the analysis of unreplicated factorial designs. *Quality Engineering*, 19:215–225, 2007.

Daniel, C. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4):311–341, 1959.

De-León, G., Grima, P., , and Tort-Martorell, X. Selecting significant effects in factorial designs taking type II errors into account. *Quality and Reliability Engineering International*, 22(7):803–810, 2006.

De-León, G., Grima, P., and Tort-Martorell, X. Comparison of normal probability plots and dot plots in judging the significance of effects in two level factorial designs. *Journal of Applied Statistics*, 38(1):161–174, 2011.

De-Veaux, R., Velleman, P. F., and Bock, D. E. *Intro Stats*. Addison-Wesley, Boston, 3rd edition, 2012.

Dong, F. On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica*, 3:209–217, 1993.

Edwards, D. J. and Mee, R. W. Empirically determined p-values for Lenth t-statistics. *Journal of Quality Technology*, 40(4):368–380, 2008.

Haaland, . P. D. and O'Connell, M. A. Inference for effect-saturated fractional factorials. *Technometrics*, 37(1):82–93, 1995.

Hamada, M. and Balakrishnan, N. Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, 8:1–41, 1998.

Juan, J. and Peña, D. A simple method to identify significant effects in unreplicated two-level factorial designs. *Communications in Statistics – Theory and Methods*, 21 (5):1383–1403, 1992.

Lenth, R. V. Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4): 469–473, 1989.

Loughin, T. M. Calibration of the length test for unreplicated factorial designs. *Journal of Quality Technology*, 30(2):171–175, 1998.

Mee, R. W., Ford-III, J. J., and Wu, S. S. Step-up test with a single cutoff for unreplicated orthogonal designs. *Journal of Statistical Planning and Inference*, 141(1): 325–334, 2011.

Montgomery, D. C. *Design and Analysis of Experiments*. John Wiley & Sons, Hoboken, New Jersey, 7th edition, 2008.

Ott, E. R. *Process Quality Control: Troubleshooting and Interpretation of Data*. McGraw-Hill, New York, 1975.

Venter, J. H. and Steel, S. J. Identifying active contrasts by stepwise testing. *Technometrics*, 40(4):304–313, 1998.

Ye, K. Q. and Hamada, M. Critical Values of the Lenth Method for Unreplicated Factorial Designs. *Journal of Quality Technology*, 32(1):57–66, 2000.

Ye, K. Q., Hamada, M., and Wu, C. F. J. A step-down lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology*, 33(2):140–152, 2001.

Chapter

5

Article 4

Analyzing DOE with Statistical Software Packages: Controversies and Proposals

Sara Fontdecaba, Pere Grima and Xavier Tort-Martorell

Department of Statistics and Operational Research, Universitat Politècnica de Catalunya -BarcelonaTech, Barcelona, 08028 Spain

Abstract. This article studies and evaluates how five well-known statistical packages: JMP, Minitab, SigmaXL, Statgraphics, and Statistica address the problem of analyzing the significance of effects in unreplicated factorial designs. All five use different methods and criteria that deliver different results, even for simple textbook examples. The article shows that some of the methods used are clearly incorrect and deliver incorrect results. Finally it raises the question of the impact that this may have in hindering the use of DOE by non-expert practitioners, and it provides suggestions for making this analysis more effective and easier to understand. Supplementary materials for this article are available online.

Keywords: Factorial design, significance of effects, statistical software, Lenth method, Industrial experimentation

5.1 Introduction

There is widespread agreement that Design of Experiments (DOE) is a very powerful technique, and that businesses should take more advantage of its potential to design and improve processes, products and services (see for example Tanco et al. [2009] and Tanco [2010]). Among the many reasons for its limited use, an important one is the lack of confidence that potential experimenters have in their DOE knowledge and skills. They fear that, after convincing skeptics and obtaining the necessary resources to conduct an experiment, the results –in terms of process learning and optimization– may be lower than expected, due to their inexperience or lack of knowledge.

In the industrial and business context, the designs most commonly used are two-level factorials. These designs are aimed at assessing the influence of a set of factors $(X_1, \ldots, X_k)$ on a response (Y). A design matrix is then defined by k orthogonal contrasts; for example, if $k = 3$, the design matrix is the one in Table 5.1.

$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$	**Y**
-1	-1	-1	y_1
$+1$	-1	-1	y_2
-1	$+1$	-1	y_3
$+1$	$+1$	-1	y_4
-1	-1	$+1$	y_5
$+1$	-1	$+1$	y_6
-1	$+1$	$+1$	y_7
$+1$	$+1$	$+1$	y_8

Table 5.1: Design matrix of a 2^3 design

The design matrix columns and their products in groups of two, three and so on define a total of 2^{k-1} orthogonal contrasts that allow estimating the same number of effects (so called because they measure the effect of the factors on the response). In the 2^3 case, the 7 contrasts allow the estimation of three main effects (X_1, X_2 and X_3), three two-factor interactions (X_1X_2, X_1X_3 and X_2X_3) and a three-factor interaction ($X_1X_2X_3$). Alternatively, the coefficients of a linear model are also used to measure the influence of the factors on the response. In the 2^3 example the model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \varepsilon$$

While the effects measure the change in the response when changing the factor from the low level (-1) to the high level $(+1)$, the coefficients measure the response change when the factor is increased by one unit; so effects are twice the size of coefficients. The clarification is important because, as will be seen, some software packages give effects while others use coefficients.

In principle, since there are as many parameters in the model as experimental conditions, the error standard deviation σ_ε can only be estimated if there are replicates; e.g., when more than one run has been conducted for each experimental condition. In such cases the standard error of the effects or the coefficients can be estimated and t-tests conducted to assess the significance of either the effects or the coefficients.

Unfortunately in industry, due to cost and time restrictions, one frequently cannot run replicates, and without them the identification of significant effects is a source of insecurity and confusion to non-expert experimenters. In the absence of replicates, one option is to estimate σ_ε by pooling the values of effects considered null; unfortunately, in the industrial context, this method often has the problem of providing estimates with few degrees of freedom.

Daniel [1976]proposed representing the effects in Normal or Half-Normal Plots to take advantage of the fact that non-significant effects would distribute roughly as $N(0;\sigma_\varepsilon)$ and would plot on these scales as a straight line; the effects lying outside this line are the ones considered significant. However, this analysis is not trivial, and mechanically applying the idea that "the significant effects are those that are separate from the straight line" frequently leads to errors, especially to non-specialists De-Leon et al. [2011]. Furthermore, the method requires human criteria and it is not suited to automatic implementation. Therefore, although most software packages present the Normal (or Half-Normal plot) of the effects marking in it the $N(0;\sigma_{ef})$ straight line and the significant effects, they have done this via an estimate of σ_{ef}, making the plot unnecessary.

Another possibility is to use Lenth's method (Lenth [1989]). The basic idea is to estimate σ_{ef}, on the basis that if $X \sim N(0,\sigma)$, the median of $|X|$ is equal to 0.6475σ and therefore $1.5 \cdot \text{median}\,|X| = 1.01\sigma \cong \sigma$. Considering that κ_i $(i = 1,\dots,n)$ are the effects of interest and that their estimators c_i are distributed according to $N(\kappa_i,\sigma_{ef})$, he defines $s_0 = 1.5{\cdot}\text{median}|c_i|$ and uses this value to calculate a new median by excluding the estimates $|c_i| > 2.5s_0$, with the intention of excluding those corresponding to effects with $|\kappa_i| > 0$. In this way he obtains the socalled pseudo *standard error*:

$$PSE = 1.5 \underset{|c_i|<2.5S_0}{\text{median}} |c_i|$$

Lenth verifies that the PSE is a reasonably good estimator of σ_{ef} when the significant effects are rare (sparse), and he proposes calculating a margin of error for c_i by means of $ME = t_{1-\alpha/2,d} \cdot PSE$ for a confidence level of $1-\alpha$, and where t is distributed according to a Student's t-distribution with $d = n/3$ degrees of freedom. However, it has been demonstrated (Ye and Hamada [2000]) that the critical values obtained from Lenth's reference distribution are not the most appropriate.

This paper reviews the methods employed by some of the statistical software packages most widely used in industrial settings, showing the variety of procedures applied and how they lead to different results. It also shows that some of these methods lead to clearly incorrect conclusions, even when applied to simple cases taken from very well-known textbooks. Finally, the paper presents a series of suggestions for making the topic of analyzing the significance of effects in unreplicated factorial designs more effective and easier to understand.

5.2 Methods used by some software packages

We have selected five of the most widely used industrial statistical packages. All of them deal in detail with DOE: they allow designing and analyzing experiments in a friendly manner (point-and-click approach, without programming or typing commands), and they automatically identify significant effects. The five packages are: JMP, Minitab, SigmaXL, Statgraphics, and Statistica. Two of them, JMP and SigmaXL, base the analysis and present the results using the model coefficients, while Minitab, Statgraphics and Statistica use the effects. In our comparisons and comments we disregard this fact, because the identification of significant contrasts depends on the method used and not on whether it is applied to effects or coefficients. The output of these programs is illustrated by using a simple example taken from Box et al. [2005] It is a 2^3 design used in the book to introduce factorial designs (p.177). Table 5.2 shows the design and the response.

X_1	X_2	X_3	**Y**
-1	-1	-1	60
$+1$	-1	-1	72
-1	$+1$	-1	54
$+1$	$+1$	-1	68
-1	-1	$+1$	52
$+1$	-1	$+1$	83
-1	$+1$	$+1$	45
$+1$	$+1$	$+1$	80

Table 5.2: 2^3 from Box, Hunter and Hunter -2005- used to illustrate the identification of signification effects performed by the selected packages

Other packages, such as the specialized *Design Expert* or its simplified version *Design Ease*, do not automatically identify significant effects but offer all available information for the experimenter to decide himself. This is, in our opinion, a good practice; but it impedes any comparison with the 5 selected packages. We did not evaluate other well-known statistical packages such as SPSS or STATA since they do not include factorial designs in a straightforward way. We believe that this makes them incomparable with the software that we are evaluating.

5.2.1 JMP (Version 11.0.0)

JMP provides two ways of analysis: 1) Estimating σ_{ef} by pooling effects considered to be null, and then using this estimate to assess significance via a t-test for each effect; JMP allows the user to select the effects to be used for estimating σ_{ef}; or 2) Using Lenth's *PSE* to estimate σ_{ef} without making any assumptions about the values of some effects, and then performing significance tests using a reference distribution of t-ratios obtained by simulation.

When using Lenth's PSE, the procedure for a design with n runs is as follows: first, JMP calculates the *PSE* and then for each effect e_i a t-ratio: $t_i = e_i/PSE$. Next it obtains a reference distribution for these t-ratios by simulating 10000 sets of $n-1$ random numbers from a $N(0;1)$, each set representing the $n-1$ effects from an n runs experiment with all factors inert (null hypothesis). The individual p-value is the interpolated fractional position among these t-ratios in descending order. The simultaneous p-value is the interpolation along the $max(|t_i|)$ of the $n-1$ values across the runs. This technique is similar to that in Ye and Hamada [2000].

JMP highlights the individual p-values smaller than 0.10 and, in addition, marks with an asterisk the ones smaller than 0.05. The Half Normal Plot (HNP) of the effects is also presented (this package does not generate Normal Probability Plots, NPP), identifying by name those that have been highlighted. Figure 5.1 shows the result obtained using Lenth's PSE.

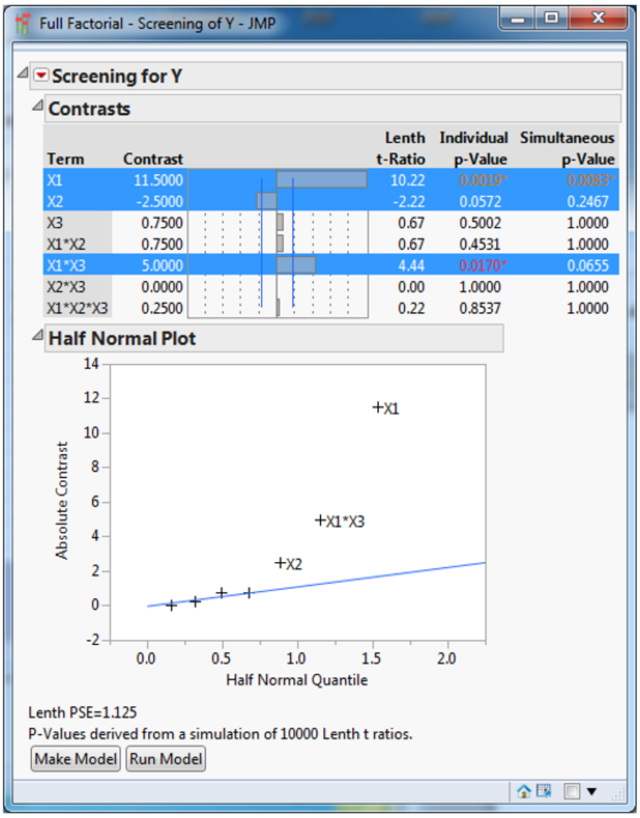


Figure 5.1: Analysis results with JMP for the 2^3 design by Box, Hunter and Hunter (2005), pg. 177

5.2.2 Minitab (Version 16.2.3)

Minitab also has two options for analysis: 1) Estimating σ_{ef} by pooling the effects that the user considers to be null, and using this estimation to perform a t-test for each effect; and 2) representing the effects graphically using an NPP, HNP or Pareto chart.

If the graph option is chosen, the representation of the effects in NPP or HNP includes the line corresponding to the distribution $N(0; PSE)$, and it marks as significant the effects that get a p-value< 0.05 when the individual tests are run using the Lenth method. Figure 5.2 shows the output of the estimated effects and coefficients and the Normal Plot of the effects. It is possible to change the α-level.

Frequently practitioners do not know Lenth's method and, thus, despite Minitab printing Lenth's PSE value on the PPN plot, a number of them (even some experienced professionals working in respected multinationals) believe that Minitab identifies significant effects by "judging" the plot as proposed by Daniel [1976].

Factorial Fit: BH2 p. 177 versus A, B, C

```
Estimated Effects and Coefficients for BH2 p. 177 (coded units)

Term        Effect    Coef
Constant            64.250
A           23.000  11.500
B           -5.000  -2.500
C            1.500   0.750
A*B          1.500   0.750
A*C         10.000   5.000
B*C         -0.000  -0.000
A*B*C        0.500   0.250

S = *   PRESS = *
```

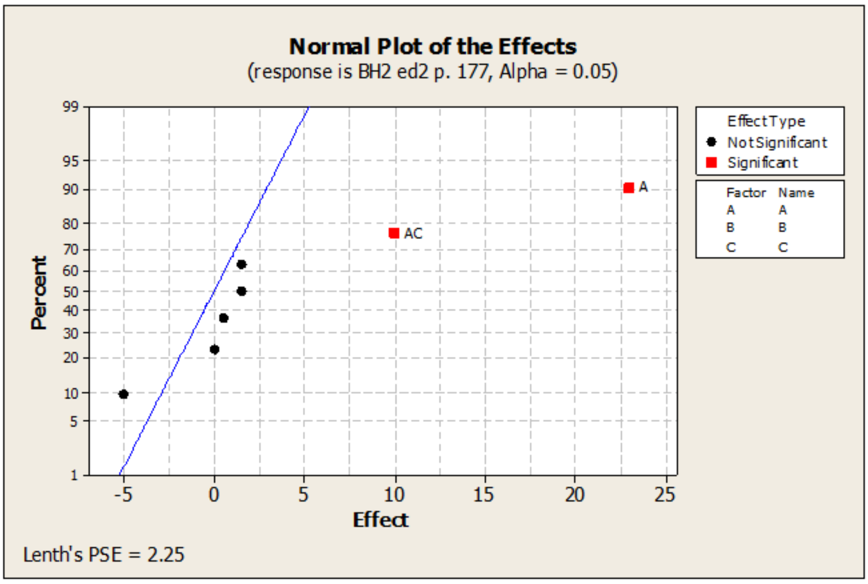


Figure 5.2: Output provided by Minitab (partial)

When the graph option is chosen after marking some effects as null, Minitab identifies the significant effects and marks them on the graph, according to the significance tests performed using the standard error estimated from effects that are considered null.

5.2.3 SigmaXL (Version 6.2)

This is a Microsoft Excel add-in. There is an option for non-experts, "Basic DOE Templates" that runs significance tests using the Lenth method and then displays a Pareto chart of the effects, including a critical value line for $\alpha = 0.05$ (5.3).
There is also another form of analysis that presents more possibilities: "2-Level Factorial/Screening". By default, this obtains the same results as the previous method; however, it also allows selecting effects that are considered null, pooling them to estimate the standard error, and using this to run significance tests. Neither of the two options allows for the effects to be represented on an NPP or HNP.

Calculation of Effects and Coefficients for Average (Y):

	Constant	A	B	C	AB	AC	BC	ABC
Avg(Avg(Y)) @ +1:		75.75	61.75	65	65	69.25	64.25	64.5
Avg(Avg(Y)) @ -1:		52.75	66.75	63.5	63.5	59.25	64.25	64
Effect (Delta):		23	-5	1.5	1.5	10	0	0.5
Coefficient (Delta/2):	64.25	11.5	-2.5	0.75	0.75	5	0	0.25
SE Coefficient:	1.125	1.125	1.125	1.125	1.125	1.125	1.125	1.125
T-value	57.111111	10.222	-2.222	0.6667	0.6667	4.4444	0	0.2222
P-value	0.0003064	0.0094	0.1564	0.5736	0.5736	0.0471	1	0.8448

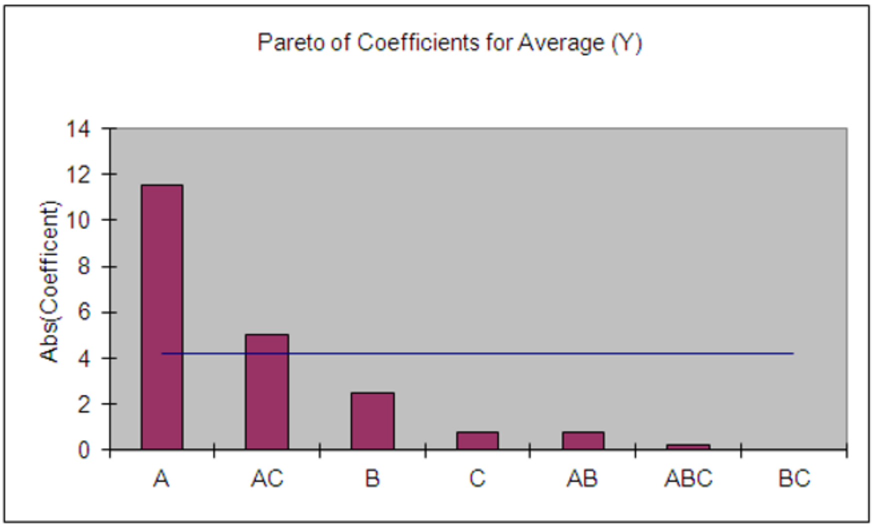


Figure 5.3: Output provided by the SigmaXL package (partial)

5.2.4 Statgraphics Centurion XVI (Version 16.2.04)

In this package there are also two ways to design and analyze an experimental plan: “Experimental Design Wizard”, that guides the user step by step, and “Legacy DOE Procedures” for more experienced users.
Using “Experimental Design Wizard”, the σ_{ef} is estimated by pooling the effects of three and more factor interactions that are treated as null by default. An ANOVA table is presented where p-values smaller than 0.05 are shown in red, along with a Pareto chart of the standardized effects, including a critical value line corresponding to $\alpha = 0.05$. Figure 5.4 displays the estimated effects and the ANOVA table. In this table, we have indicated with an asterisk the p-values that appear in red in Statgraphics’ original output.

These results agree with the ones presented by Minitab and JMP when instructed to use the three factor interaction to estimate σ_{ef}.

With “Legacy DOE Procedures”, the user can select up to what degree of interaction is of interest; from that degree onwards the interactions are considered null. Then a representation of the effects on an NPP can be obtained. This is illustrated in Figure 5.5 wherein only six effects are plotted because the interaction of three factors has been considered to be null, and null effects are not shown. The criterion for drawing the line is fitting a least squares regression line to the smallest, in absolute value, 50% of the effects. The Lenth method is not used under any circumstances. This graph does not

Estimated effects for BH2 p. 177

Effect	*Estimate*	*Stnd. Error*	*V.I.F.*
average	64.25	0.25	
A:A	23.0	0.5	1.0
B:B	-5.0	0.5	1.0
C:C	1.5	0.5	1.0
AB	1.5	0.5	1.0
AC	10.0	0.5	1.0
BC	0.0	0.5	1.0

Standard errors are based on total error with 1 d.f.

Analysis of Variance for BH2 p. 177

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
A:A	1058.0	1	1058.0	2116.00	0.0138*
B:B	50.0	1	50.0	100.00	0.0635
C:C	4.5	1	4.5	9.00	0.2048
AB	4.5	1	4.5	9.00	0.2048
AC	200.0	1	200.0	400.00	0.0318*
BC	0.0	1	0.0	0.00	1.0000
Total error	0.5	1	0.5		
Total (corr.)	1317.5	7			

Figure 5.4: Estimated effects and ANOVA Table with Statgraphics

indicate what effects should be considered significant.

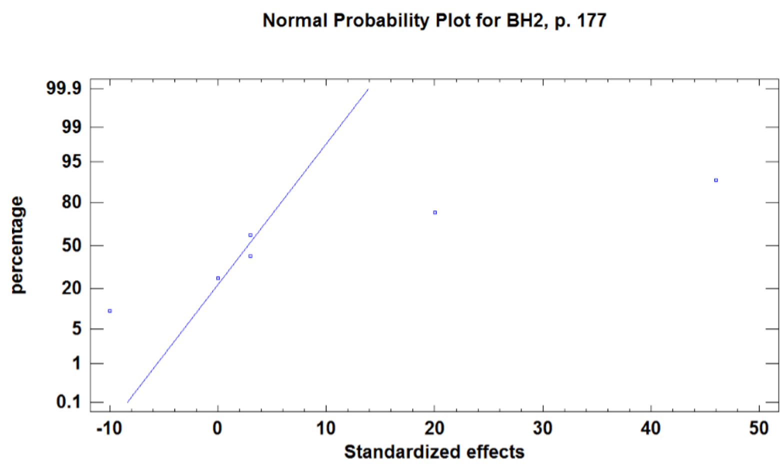


Figure 5.5: Representation of effects on an NPP with Statgraphics

5.2.5 Statistica

Results are analyzed in a window that offers the user several possibilities. By default (the "Quick" option), σ_{ef} is estimated by considering interactions of three or more factors to be null. The user can access a summary where the effects with a p-value < 0.05 appear in red. A Pareto chart of the effects, that includes a critical value line corresponding to $\alpha = 0.05$, can also be obtained (Figure 5.6).

There is also a "Model" option that allows the user to select specific interactions that are to be considered null. If none are considered null, it does not perform the significance analysis of the effects. In addition, the effects can be plotted on NPP or HNP, but only

Effect Estimates; Var.:BH2, p. 177; R-sqr=.99962; Adj:.99734 (Design: 2**(3-0) design (Spreadsheet1) in Workbook3)
2**(3-0) design; MS Residual=.5
DV: BH2, p. 177

Factor	Effect	Std.Err.	t(1)	p	-95.% Cnf.Limt	+95.% Cnf.Limt	Coeff.	Std.Err. Coeff.	-95.% Cnf.Limt	+95.% Cnf.Limt
Mean/Interc.	64.25000	0.250000	257.0000	0.002477	61.0734	67.42655	64.25000	0.250000	61.07345	67.42655
(1)A	23.00000	0.500000	46.0000	0.013837	16.6469	29.35310	11.50000	0.250000	8.32345	14.67655
(2)B	-5.00000	0.500000	-10.0000	0.063451	-11.3531	1.35310	-2.50000	0.250000	-5.67655	0.67655
(3)C	1.50000	0.500000	3.0000	0.204833	-4.8531	7.85310	0.75000	0.250000	-2.42655	3.92655
1 by 2	1.50000	0.500000	3.0000	0.204833	-4.8531	7.85310	0.75000	0.250000	-2.42655	3.92655
1 by 3	10.00000	0.500000	20.0000	0.031805	3.6469	16.35310	5.00000	0.250000	1.82345	8.17655
2 by 3	0.00000	0.500000	0.0000	1.000000	-6.3531	6.35310	0.00000	0.250000	-3.17655	3.17655

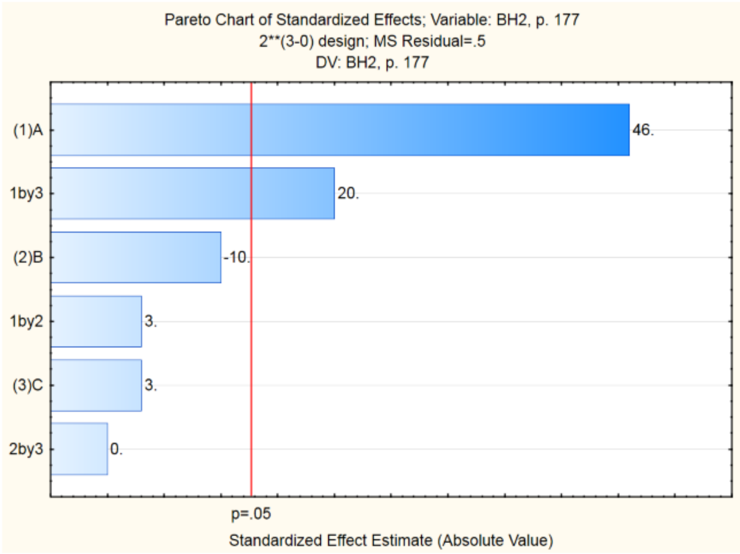


Figure 5.6: Output displayed by the Statistica package

those included in the model are shown. It does not include a line for null effects or highlight those considered to be significant. Like Statgraphics, it never uses the Lenth method.

5.3 Comparison

To be sure, the fact that different methods are used means that none of them are clearly better than the rest in all situations. However, we believe that some processes are not appropriate for being used by default.

5.3 summarizes the methods used by the 5 analyzed software packages to assess effects' significance via formal significance tests. All of the programs allow for identifying interactions assumed to be null, in order to estimate σ_{ef} using their values. But leaving aside the very rare case when the experimenter is sure that some interactions of two factors are null, this method is only useful when working with complete designs of 4 or more factors, or with fractional designs with 6 or more factors and resolution V or bigger. This is not a frequent situation. In 8 run designs, it can only be applied if the design is complete; and in this case the standard error of the effects will be estimated with only one degree of freedom –clearly not a very good idea, even though Statgraphics and Statistica do it by default.

Formal significance tests	**JMP**	**MTB**	**SXL**	**STG**	**STA**
Uses three and more factor interactions to estimate σ_{ef}				○	○
Manual selection of effects to be used to estimating σ_{ef}	○	○	○	○	○
Lenth method (individual contrasts)		○	○		
P-value derived from a simulation of Lenth t-ratios	○				

Table 5.3: Ways to run significance tests on the effects using the 5 analyzed software packages

Representation of effects on a Normal or Half Normal Plot	**JMP**	**MTB**	**SXL**	**STG**	**STA**
Does not include representation on NPP or HNP			○		
Does not draw any line					○
Draws a line adjusted by minimum squares at 50% of smaller effects				○	
Draws a line that represents distribution $N(0; PSE)$	○	○			
Identifies possible significant effects t-ratios	○	○			
Does not include effects not included in model		○		○	○

Table 5.4: Use of NPPs and HNPs in the packages studied

Moreover, it has been shown that the critical values proposed by Lenth are inadequate, yet they are used by SigmaXL and Minitab in their graphic analysis (perhaps we should say "pseudo-graphic analysis", since the conclusions are not derived from the graphs). Loughin [1998] and also Ye and Hamada [2000] show that a Student's t-distribution with $n/3$ degrees of freedom is not a good reference distribution for the Lenth t-ratio; they conclude through simulation that for $\alpha = 0.05$ more appropriate critical values are $t = 2.30$ and $t = 2.15$ for 8 and 16 run designs, while Lenth proposed 3.76 and 2.57. Therefore, both packages have a high tendency to ignore (mark as non-significant) effects that should have been considered significant at the selected α-level.

The use of NPP or HNP plots and methods associated therewith is summarized in Table 5.4. The proposed methods also vary. They range from not using this tool (SigmaXL), to using it without drawing the line representing the distribution of null effects (Statistica), or to drawing the line by different methods (JMP, Minitab, and Statgraphics). There are packages that, by default, identify possible significant effects on the graph (JMP and Minitab), and others do not (Statgraphics and Statistica), leaving this task to the analyst.
Table 5.5 shows 16 examples taken from Box et al. [2005] and Montgomery [2013]. In general, they are very simple, several of them are not real cases but examples especially prepared to be clear and unambiguous. Even in these simple cases, the significant effects proposed by the 5 packages do not agree; neither with each other nor with the book. In 16-run designs there is more agreement, possibly because in several of the cases some effects –the significant ones– are very, very large compared with σ_{ef}. Details of the output provided by each software package in each example can be found in the online supplemental material.

#	Book/Page	Design	Significant effects suggested by:					
			Book	JMP	Minitab	SigmaXL	Statgrapch	Statistica
1	BH2/177	2^3	A, B, AC	A, B*, AC	A, AC	A, AC	A, AC	A, AC
2	BH2/191	2^3	A, B, C	A, B, C	A, B, C	A, B, C	A	A
3	BH2/194 Y1	2^3	C, AB	C*	*None*	*None*	*None*	*None*
4	BH2/194 Y2	2^3	*None*	*None*	*None*	*None*	*None*	*None*
5	BH2/194 Y3	2^3	B, C	C	*None*	*None*	*None*	*None*
6	BH2/194 Y4	2^3	A, B, AB	A, B	B	B	B	B
7	BH2/207	2^3	A	A	*None*	*None*	*None*	*None*
8	BH2/237	2^{4-1}	A, B	A, B	A	A	*None*	*None*
9	BH2/199	2^4	A, B, D, BD	A, B, D, BD	A, B, D, BD	A, B, D, BD	A, B, D, BD	A, B, D, BD
10	BH2/265 y1	2^{8-4}	A, B	A, B	A, B	A, B	A, B	A, B
11	BH2/265 y2	2^{8-4}	A, B, F	A, B, D, F, G*	A, B, F	A, B, F	A, B	A, B
12	Mont/257	2^4	A, C, D, AC, AD	A, C, D, AC, AD	A, C, D, AC, AD	A, C, D, AC, AD	A, C, D, AC, AD	A, C, D, AC, AD
13	Mont/269	2^4	B, C, D, BC, BD	B, C, D, BC*, BD*	B, C, D	B, C, D	B, C, D, BC, BD	B, C, D, BC, BD
14	Mont/271	2^4	A, C	A, C	A, C	A, C	A, C	A, C
15	Mont/328	2^{5-1}	A, B, C, AB	A, B, C, AB	A, B, C, AB	A, B, C, AB	A, B, C, AB	A, B, C, AB
16	Mont/336	2^{6-2}	A, B, AB	A, B, AB, AC*, AD, BC*, ABF	A, B, AB, AD, ABF	A, B, AB, AD, ABF	A, B, AB	A, B, AB

Table 5.5: Results obtained for simple examples. "BH2" stands for Box, Hunter and Hunter (2005) and "Mont" stands for Montgomery (2013)

5.4 Proposals

The considered statistical packages provide several options for analyzing the significance of effects, allowing the analyst to use several or choose the one that seems most suitable for each case. In general, one of the options seems more immediate and also seems intended to make the task easier for non-expert users. We believe that the following proposals would improve the effectiveness and clarity, especially for non-experts.

With respect to formal significance tests:

1. Do not estimate the standard error of the effects with just one, or very few, degrees of freedom.

2. Do not calculate Lenth's PSE when there are only 4 runs. In this case it always equals 1.5 multiplied by the central value (in absolute value order). It is like estimating the standard error with one degree of freedom. Lenth, in his original article (Lenth [1989]), does not consider this possibility. However, the packages studied here that use this method do apply it with 4 runs.

3. Replace the critical values originally proposed by Lenth with those proposed by (Loughin [1998]) or Ye and Hamada [2000].

4. We believe that it is a good idea to highlight the effects that are significant at

$\alpha = 0.10$ in addition to the ones at $\alpha = 0.05$, as JMP does. In the industrial context, it is as risky to ignore variables that affect the output (Type II error) as it is to consider significant variables that do not affect it (Type I error). Thus, trying to balance the two types of error is good practice (De-Leon et al. [2006]).

With respect to graphical analysis:

1. Do not show effects on an NPP (or HNP) when only 4 runs have been conducted. The essence of the method is to distinguish the values that form a straight line (the null effects) from those that do not fall on that line. It is obvious that this method of analysis does not make sense with 3 points. However the 4 packages studied that produce NPP or HNP representations produce them when there are only 3 effects.

2. Do not include information on the NPP about effects that are significant and those that are not. Practitioners tend to believe that the software has made a good decision and do not dare to correct it, although sometimes a change is clearly justified.

3. Do not include by default a critical value line in the Pareto charts of effects. It should be clear that the Pareto chart is a graphical method that requires some judgment. If the decision is based on the critical value, the diagram is not required.

4. Always include the dot plot of the effects in the NPP representations by projecting the effects on the horizontal axis. Daniel [1976] and Box et al. [2005] always do this. This type of graphic representation is much easier to understand and can be as informative as the NPP representation of the effects. A study on this can be seen in De-Leon et al. [2011].

5. Include all the effects in the NPP representation. In some of the packages studied, the effects considered to be null and used to estimate σ_{ef} do not appear in the NPP representation. This is a bad idea. The NPP representation is an alternative form of analysis, and representing only part of the effects makes it more difficult to interpret the graph.

5.5 Concluding Remarks

Statistical software packages commonly used in industry for planning and analyzing factorial designs use different processes to analyze the significance of the effects, and it is not uncommon that they provide different results.

Non-expert users of DOE tend to trust the results that the software packages give them. As we have shown, these results are sometimes incorrect. Furthermore, it is an awkward situation when the instructor of a seminar says: “Do not trust the conclusions provided by your software, use the principles and criteria I’m telling you”. We wonder what the people attending the seminar may be thinking. In addition, it is unsettling for practitioners to see that different packages provide different answers.

The five packages that we evaluated can improve the way that the significance of effects in unreplicated factorial designs is analyzed. The amount of improvement that is needed is small in some cases and much larger in others. We have identified several errors, and also made suggestions that we hope will make the interpretation of results easier for practitioners

Acknowlegments: The authors are very grateful to the editor and the associate editor. Their many comments have been very useful to improve the paper in all aspects: technical, organizational, clarity and writing.

Bibliography

Box, G. E. P., Hunter, J. S., and Hunter, W. G. *Statistics for Experiments: Design, Innovation and Discovery.* Wiley, Hoboken, 2nd edition, 2005.

Daniel, C. *Applications of Statistics to Industrial Experimentation.* Wiley, New York, 1976.

De-Leon, G., Grima, P., and Tort-Martorell, X. Selecting Significant Effects in Factorial Designs Taking Type ii Errors into Account. *Quality and Reliability Engineering International*, 22(7):803–810, 2006.

De-Leon, G., Grima, P., and Tort-Martorell, X. Comparison of Normal Probability Plots and Dot Plots in Judging the Significance of Effects in Two Level Factorial Designs. *Journal of Applied Statistics*, 38(1):161–174, 2011.

Lenth, R. V. Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4): 469–473, 1989.

Loughin, T. M. Calibration of the length test for unreplicated factorial designs. *Journal of Quality Technology*, 30(2):171–175, 1998.

Montgomery, D. C. *Design and Analysis of Experiments.* Wiley, Hoboken, 8 edition, 2013.

Tanco, M. Why is not Design of Experiments Widely Used by Engineers in Europe? *Journal of Applied Statistics*, 37(12):1961–1976, 2010.

Tanco, M., Viles, E., Ilzarbe, L., and Alvarez, M. Barriers Faced by Engineers when Applying Design of Experiments. *The TQM Journal*, 21(6):565–575, 2009.

Ye, K. Q. and Hamada, M. Critical Values of the Lenth Method for Unreplicated Factorial Designs. *Journal of Quality Technology*, 32(1):57–66, 2000.

Part III
Conclusions and Future Directions

Chapter

6 Conclusions and Future Directions

In this chapter, we set out what has been accomplished in the thesis (Section 6.1) and the future lines of research it opens (Section 6.2).

6.1 Conclusions and Contributions

In what follows, the main contributions of the present thesis are set out. They include both theoretical aspects, and the methods and resources created, as well as, considerations regarding contributions to industrial statistics.

The list below summarizes the major contributions of water demand management:

1. A procedure that includes all the justifications and necessary steps to carry out the methodology to forecast future water demand (short, medium and long term), as well as to know more about the influence of socioeconomic factors on customer's consumption. This knowledge allows the detection of segments of clients with homogeneous consumption habits, which is also explained in detail.

2. The capability to create a unique data base by merging two different sources: household water consumption and sociological variables at the minimum disaggregation possible level (census tract).

3. The scope of the analysis with 23 municipalities of the Metropolitan Area of Barcelona, representing the total domestic water consumption of the area studied. A total of 27 variables were considered in the study, belonging to six major categories: territory, sex, nationality, household, income and population growth and 2.358 census tracts.

4. An algorithm that assigns water consumption to micro components end uses on the basis of the readings of water consumption. It evaluates the variants in identifying water uses in three steps: (*i*) define a criterion for assigning the combinations, (*ii*) create a mechanism for producing variants with higher probability of testing the correct combination (applying non-random heuristic methodologies) and (*iii*) evaluate the program and assign the device (using the likelihood function related to matching templates).

The main conclusions of the water demand management article are:

1. The use of two basic families of algorithms (hierarchical and partitional), leads to clustering the census tracts into 6 groups of homogenous socioeconomical information and water consumption habits. They have been described and represented on a spatial map, confirming the soundness of the clusters found, according to the structure of the Metropolitan Area of Barcelona.

2. The analysis using linear model regression and principal components analysis confirms the relationship between water consumption habits and socioeconomic characteristics. This relationship can be used to detect group of clients with similar consumption patterns.

3. The created groups, have been used to apply SARIMA models to short term prediction of water consumption, long term scenario planning and a better understanding of consumer habits.

4. The variability in micro component data can be characterized by analyzing the frequency of water uses and the volume of water per use. It leads to a total mean daily consumption of about 93 l. per house, distributed as follows: 40% shower, 32% cistern, 21% kitchen and bath taps, the remaining 6% corresponding to dishwasher and washing machine.

5. The "within household variability" is at least as significant a component of total variability as that of the "between household differences".

6. The sensitivity of the model is nearly 70%. This means that the real water uses that the process has classified correctly represents a 67.8% of the monitored consumption. The specificity analysis shows that internal taps category works as an absorbent state, with 91.37% of proportion uses well classified. Since, cisterns, showers and washing machine have acceptable levels of classifications, dishwasher are problematic with only 9% of the water uses well associated.

7. The methodology has also been applied to ten houses in the city of Murcia. The results are also satisfactory obtaining a 71.8% of the programs well identified, representing a 60.8% of the total water consumption.

Following the same structure as before, the main contributions of the determination of the significance effects in DoE are:

1. The proposal of using a simple critical value t = 2, for any 2^k or 2^{k-p} design, based on the idea, that Lenth's method is the central reference among the analytical procedures proposed in the literature for identifying significant effects. The

replacement of the simulations methods or tables for a single value conveys the idea that the procedure is approximate.

2. The simulation of a wide range of reasonable scenarios in order to compare the original Lenth method and other methods proposed in the literature. The results show that the previous proposal has type I errors closer to the established value of 5% and solves the common disadvantage of leading to type I error probabilities smaller than expected and, consequently, to a greater probability of type II error.

3. When comparing the results provided by analyzing simple cases with the five software packages they disagree among each other and with the results provided be the reference books. In 16-run designs, there is more agreement, possibly because in several of the cases some effects -the significant ones, are very large.

4. Referring to the formal significance tests, only two packages: *StatGraphics* and *Statistica* use, by default, three and more factor interactions to estimate σ_{ef}. The other three packages allow the user to manually select the effects to be used to estimate σ_{ef}. The more conservative packages, using the original Lenth procedure without any modification, are *Minitab* and *SigmaXL*; while *JMP*, the one doing a better job, uses p-values derived from a simulation of Lenth ratios.

5. Concerning graphical analysis, the proposed methods among the packages also differ. *SigmaXL* is the one which doesn't show to the user any representation and Stata uses the NPP or HNPP without drawing the line representing the distribution of null effects. *JMP*, *Minitab* and *Statgraphics* draw a line on the graph using different methods. By default, *JMP* and *Minitab* identify the possible significant effects on the graph.

Taking into account all the above comments and considering that in most cases the users of these software packages are practitioners and not statistical experts our conclusions and recommendations are:

7. With respect to formal significance tests, avoid estimating the standard error with very few degrees of freedom, as well as do not calculate PSE when there are only 4 runs. Software packages must replace Lenth's method original critical values by better alternatives proposed in the literature such as those proposed by Loughin or Ye and Hamada. Finally, highlight the effects that are significant at $\alpha = 0.10$ in addition to the ones at $\alpha = 0.05$.

8. With respect to graphical analysis, do not show effects on a NPP (or HNP) when only 4 runs have been conducted and do not include information on the NPP about significant and not significant effects; the same holds for Pareto graphs of the effects, do not include by default a critical value line, leave the task of deciding which effects are significant to the experimenter. Finally, represent all effects on the plots, even if some of them have been used to calculated σ_{ef}.

6.2 Future Directions

Naturally the possibilities of applying statistics in known or new ways to business and industry are endless. Therefore, we will concentrate in the research possibilities that this thesis opens with regard to the two topics covered: the application of the statistics in water demand management, and the identification of significant effects in unreplicated factorials.

With respect to water demand management:

Further work should be done in data collection; companies have little data about their own operations and are only beginning to realize the possibilities of gathering data form outside sources. The possibilities of Big Data in improving the management of water utilities is enormous and in fact the path opened in this thesis of mixing socioeconomic data with consumption data to obtain better forecasts is a first step in this direction. Clearly, the methodology proposed in article 1, can be used in other areas with only small adaptations. The only requirements are: having socioeconomic data (the proposed in the thesis is data on a census tract level, but it would be better on a disaggregate level), water consumption at the consumer (or other disaggregate) level and being able to situate the consumers in census tracts. Referring to the problem of assigning micro-components to appliances based only on the readings of water meters (article 2) is a very difficult one because of: the scarcity of the information (basically just one variable: the water consumption time series) and the variability of patterns in which people use the different appliances. In spite of that the devised method gives a reasonably high percentage of correct identification (58%)of consumption.

In our opinion, and given the variability already commented, it will be very difficult to increase in a significant way this percentage without increasing in one way or another the information available to make the assignation. What follows is our vision of the possible paths to follow to further improve the algorithm and also on ways to increase the information available to make the assignment:

1. Further investigation on variables that will help characterize long programs (washers and dishwashers) and thus, hopefully, have a higher correct identification rate of this devices.

2. Devise ways to incorporate information on the number of persons living in the household.

3. Even though we have not found significant improvements in the assignment process by using appliance models specific for each household with respect to using a general model for all the households, a possible (however complicated and uncertain) way to improve is to introduce some type of learning procedure in the algorithm so that it adapts itself to the household habits. This is very difficult without human intervention.

4. The group of internal taps presents a very large variability, so it is not surprising that it absorbs water uses that belong to other appliances, specially showers, dishwashers and washing machines. At this point it is not clear how, and it may just be not possible, to devise ways to mitigate this fact; but, certainly it would be a breakthrough improvement.

Possible improvements in the input information:

1. In some countries there are different water meters for hot and cold water. To have consumption from these two time series will be very helpful. The immediate consequence is that it will help to identify showers and thus eliminate the noise and problems generated by one of the "problematic" devices.

2. To incorporate a tactile screen on the household were people will input some information (even if partial) on the appliances being used.

3. To personally monitor the house during some time to gather firsthand information on uses (linked with the one provided by the meter through a synchronized time) and washer and dishwasher types programs.

With respect to the second block where the identification of significant effects in unreplicated DoE is analyzed, an area where a lot of work had been done that in general has led to proposals giving a false impression of accuracy in the α risks, our message is that any further work should be addressed towards finding simple solutions easy to interpret by non-expert practitioners and providing known and accurate α risks. Our proposal in this thesis is a clear step in this direction.

For the software designers of the five analyzed packages that we have evaluated we provided them with several improvement ideas and we are sure that they can from there undertake further work in the path of providing a clearer and better guidance to the non-experts users of DoE.

Bibliography

Arbués, F. and Villanua, I. Potential for pricing policies in water resource management: Estimation of urban residential water demand in Zaragoza, Spain. *Urban Studies*, 43: 2421–2442, 2006.

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

Babel, M. S. and Shinde, V. Identifying prominent explanatory variables for water demand prediction using artificial neural networks: A case study of Bangkok. *Water Resource Management*, 25(6):1653–1676, 2011.

Babel, M. S., Das-Gupta, A., and Pradhan, P. A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal. *Water Resource Management*, 21(3):573–589, 2007.

Baumann, D. D., Boland, J., and Hanemann, W. M. *Urban water demand management and planning*. McGraw-Hill, New York, 1998.

Beecher, J. Integrated resources planning for water utilities. *Water Resources Update*, 104, 1996.

Bergquist, B., Vanhatalo, E., and Nordenvaad, M. L. A Bayesian analysis of unreplicated two-level factorials using effects sparsity, hierarchy, and heredity. *Quality Engineering*, 23:152–166, 2011.

Box, G. E. P. Teaching Engineers Experimental Design with a Paper Helicopter. *Quality Engineering*, 4(3):453–455, 1992.

Box, G. E. P., Jenkins, M., and Reinsel, G. *Time series analysis: forecasting and control.* Prentice-Hall, 3rd edition, 1994.

Box, G. E. P., Hunter, J. S., and Hunter, W. G. *Statistics for Experiments: Design, Innovation and Discovery.* Wiley, Hoboken, 2nd edition, 2005.

Brooks, D. B. An operational definition of water demand management. *Water Resources Development*, 22:521–528, 2006.

Butler, D. and Memon, F. Water demand management. *International Water Association Publishing (IWAP)*, 2006.

Casella, G. and Berger, L. *Statistical inference.* Thomson Learning, United States, 2nd edition, 2001.

Corral-Verdugo, V., Frías-Armenta, M., Pérez-Urias, F., Orduña-Cabrera, V., and Espinoza-Gallego, N. Residential water consumption, motivation for conserving water and the continuing tragedy of the commons. *Environ. Manag.*, 30:527–535, 2002.

Costa, N. and Lopez-Pereira, Z. Decision-making in the analysis of unreplicated factorial designs. *Quality Engineering*, 19:215–225, 2007.

Daniel, C. Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4):311–341, 1959.

Daniel, C. *Applications of Statistics to Industrial Experimentation.* Wiley, New York, 1976.

De-León, G., Grima, P., , and Tort-Martorell, X. Selecting significant effects in factorial designs taking type II errors into account. *Quality and Reliability Engineering International*, 22(7):803–810, 2006.

De-Leon, G., Grima, P., and Tort-Martorell, X. Selecting Significant Effects in Factorial Designs Taking Type ii Errors into Account. *Quality and Reliability Engineering International*, 22(7):803–810, 2006.

De-León, G., Grima, P., and Tort-Martorell, X. Comparison of normal probability plots and dot plots in judging the significance of effects in two level factorial designs. *Journal of Applied Statistics*, 38(1):161–174, 2011.

De-Leon, G., Grima, P., and Tort-Martorell, X. Comparison of Normal Probability Plots and Dot Plots in Judging the Significance of Effects in Two Level Factorial Designs. *Journal of Applied Statistics*, 38(1):161–174, 2011.

De-Veaux, R., Velleman, P. F., and Bock, D. E. *Intro Stats.* Addison-Wesley, Boston, 3rd edition, 2012.

Dong, F. On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica*, 3:209–217, 1993.

Draper, N. and Smith, W. *Applied regression analysis.* Wiley, 3rd edition, 1998.

Duke, J. M., Ehemann, R. W., and Mackenzie, J. The distributional effects of water quantity management strategies: a spatial analysis. *Rev. Reg. Stud.*, 32(1):19–35, 2002.

Dziegielewski, B. Management of Water Demand: Unresolved Issues. *J. Water Resour. Update*, 114:1–7, 1993.

Edwards, D. J. and Mee, R. W. Empirically determined p-values for Lenth t-statistics. *Journal of Quality Technology*, 40(4):368–380, 2008.

European Commission. EU Water Framework Directive. Directive 2000/60/EC.

Eurostat. Consumers in europe – Facts and figures on services of general interest. *http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-DY-07-001/EN/KS-DY-07-001-EN.PDF*, page Accessed 16 April 2012, 2007.

Farinaccio, L. and Zmeureanu, R. Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses. *Energy and Buildings*, 30:245–259, 1999.

Fontdecaba, S., Grima, P., Marco, L., Rodero, L., Sánchez-Espigares, J., Solé, I., Tort-Martorell, X., Demessence, D., Martínez-De-Pablo, V., and Zubelzu, J. A methodology to model water demand based on the identification of homogenous client segments. Application to the city of Barcelona. *Water Resources Management*, 26:499–516, 2012.

Gleick, P. H. Water use. *Annu. Rev. Environ. Resour.*, 28:275–314, 2003.

Griffin, R. C. and Chang, C. Seasonality in community water demand. *West. J. Agric. Econ.*, 16(2):207–217, 1991.

Guy, S. Managing water stress: the logic of demand side infrastructure planning. *J. Environ. Plan. Manag.*, 39:123–130, 1996.

Haaland, . P. D. and O'Connell, M. A. Inference for effect-saturated fractional factorials. *Technometrics*, 37(1):82–93, 1995.

Hamada, M. and Balakrishnan, N. Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, 8:1–41, 1998.

Hamilton, L. Saving water: a causal model of household conservation. *Sociological Perspectives*, 26:355–374, 1983.

Hanke, S. and de Mare, L. Residential water demand: a pooled, time series, cross section study of Malmö, Sweden. *J. Am. Water Resour. As.*, 18(4):621–626, 1982.

Hasse, D. and Nuiss, H. Does urban sprawl drive changes in the water balance and policy? The case of Leipzig (Germany). *Landscape and Urban Planning*, 80:1–13, 2007.

Hellegers, P., Soppe, R., Perry, C., and Bastiaanssen, W. Remote sensing and economic indicators for supporting water resources management decisions. *Water Resour. Manag.*, 24(11):2419–2436, 2010.

ICWE (International Conference on Water and Environment). Dublin, 1992.

Johnson, R. A. and Wichern, D. W. *Applied multivariate statistical analysis*. Prentice Hall, 2002.

Juan, J. and Peña, D. A simple method to identify significant effects in unreplicated two-level factorial designs. *Communications in Statistics – Theory and Methods*, 21 (5):1383–1403, 1992.

Kahn, M. E. The Environmental Impact of Suburbanization. *J. Policy. Anal. Manage.*, 19:569–586, 2000.

Kanakoudis, V. K. Urban water use conservation measure. *Journal on Water Supply Research and Technology: AQUA*, 51(3):153–159, 2002.

Lavière, I. and Lafrance, G. Modelling the electricity consumption of cities: effect of urban density. *Energy Econ.*, 21:53–66, 1999.

Lebart, L., Morineau, A., and Piron, M. *Statistique exploratoire multidimensionnelle*. In Dunod, 4th edition, 2006.

Lenth, R. V. Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4): 469–473, 1989.

Liu, J., Daily, G. C., Ehrlich, P. C., and Luck, G. W. Effects of households dynamics on resource consumption and biodiversity. *Nature*, 421:530–533, 2003.

Loughin, T. M. Calibration of the length test for unreplicated factorial designs. *Journal of Quality Technology*, 30(2):171–175, 1998.

Mayer, W. P., William, B., DeOreo, T. E., and Lewis, D. M. Residential indoor water conservation study: evaluation of high efficiency indoor plumbing fixture retrofits in single-family homes in the East BayMunicipal Utility District Service Area. Prepared to East BayMunicipal Utility District and The United States Environmental Protection Agency. 2003.

Mee, R. W., Ford-III, J. J., and Wu, S. S. Step-up test with a single cutoff for unreplicated orthogonal designs. *Journal of Statistical Planning and Inference*, 141(1): 325–334, 2011.

Molino, B., Rasulo, G., and Tagliatela, L. Forecast model of water consumption for Naples. *Water Resour. Manag.*, 10(4):321–332, 1996.

Montgomery, D. C. *Design and Analysis of Experiments*. John Wiley & Sons, Hoboken, New Jersey, 7th edition, 2008.

Montgomery, D. C. *Design and Analysis of Experiments*. Wiley, Hoboken, 8 edition, 2013.

Montini, A. M. M. The determinants of residential water demand: empirical evidence for a panel of Italian municipalities. *Appl. Econ. Lett.*, 13:107–111, 2006.

Murdock, S. H., Albrecht, D. E., Hamm, R. R., and Backman, K. Role of sociodemographic characteristics in projections of water use. *J. Water Resour. Plann. Manag.*, 117:235–251, 1991.

Nauges, C. and Thomas, A. Long-run study of residential water consumption. *Environmental & Resource Economics, European Association of Environmental and Resource Economists*, 26(1):25–43, 2003.

Opaluch, J. J. Urban residential demand for water in the United Status: Further discussion. *Land. Econ.*, 58:225–227, 1982.

Ott, E. R. *Process Quality Control: Troubleshooting and Interpretation of Data.* McGraw-Hill, New York, 1975.

Peña, D. *Procesos de media móvil y ARMA.* Alianza. Análisis de series temporales, Madrid, 1st edition, 2005.

Peña, D. *El modelo general de regresión.* Alianza. Regresión y diseño de experimentos, Madrid, 2nd edition, 2010.

Postel, S. *The last oasis. Facing water scarcity.* London: W W Norton & Co Inc, 1992.

Renwick, M. E. and Green. Do residential water demand side management policies measure up? An analysis of eight California water agencies. *J. Environ. Econ. Manag.*, 40:37–55, 2000.

Renzetti, S. *The economics of water demand.* Kluwer Academic Publishers, Boston, 2002.

Richter, C. and Stamminger, R. Water consumption in the kitchen – a case study in four European countries. *Water Resources Management*, 26:1639–1649, 2012.

Smith, A. and Ali, M. Understanding the impact of cultural and religious water use. *Water. Environ. J.*, 20:203–209, 2006.

Stephenson, D. Demand management theory. *Water S.A*, 25:115–122, 1999.

Tanco, M. Why is not Design of Experiments Widely Used by Engineers in Europe? *Journal of Applied Statistics*, 37(12):1961–1976, 2010.

Tanco, M., Viles, E., Ilzarbe, L., and Alvarez, M. Barriers Faced by Engineers when Applying Design of Experiments. *The TQM Journal*, 21(6):565–575, 2009.

Tong, Y. L. *The multivariate normal distribution.* Springer Series in Statistics, United States, 1st edition, 1989.

United States Environmental Protection Agency (USEPA). *Water and energy savings from high efficiency fixtures and appliances in single family homes.* USEPA, Washington.

Venter, J. H. and Steel, S. J. Identifying active contrasts by stepwise testing. *Technometrics*, 40(4):304–313, 1998.

Yamagami, S. and Nakamura, H. Non-intrusive submetering of residential gas appliances. *ACEEE Summer Study on Energy Efficiency in Buildings*, pages 265–273, 1996.

Ye, K. Q. and Hamada, M. Critical Values of the Lenth Method for Unreplicated Factorial Designs. *Journal of Quality Technology*, 32(1):57–66, 2000.

Ye, K. Q., Hamada, M., and Wu, C. F. J. A step-down lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology*, 33(2):140–152, 2001.

Zhang, H. and Brown, D. Understanding urban residential water use in Beijing and Tianjin, China. *Habitat International*, 29:469–491, 2005.