Smart metering data for urban water demand modelling and Water Distribution Network management

Diana Fiorillo

Dipartimento di Ingegneria Civile, Edile e Ambientale

Università degli Studi di Napoli Federico II

Thesis submitted to the University of Naples Federico II for the degree of Doctor of Philosophy in Civil Systems Engineering in July 2021.

PhD Candidate: Diana Fiorillo

Supervisors: Prof. Maurizio Giugni, Prof. Francesco De Paola

Co-supervisor: Prof. Enrico Creaco (University of Pavia)

PhD Program Coordinator (XXXIII Cycle): Prof. Andrea Papola

Reading Committee: Prof. Chrysi Laspidou (University of Thessaly), Prof. Zoran Kapelan (Delft University of Technology)

Ai miei genitori,

con amore e gratitudine.

Abstract

The mitigation of the impacts of climate changes and anthropic pressures on water security calls for effective strategies for sustainable water distribution networks (WDNs) management. Innovative metering technologies, such as smart metering, able to provide high resolution water demand data, represent a useful tool still not fully exploited. The general objective of this thesis is to develop new water demand models to improve WDNs management, as well as optimize the use of smart meters. To this purpose, smart metering data from the District Metered Area (DMA) of Soccavo (Naples, Italy) were used.

A novel bottom-up methodology for the generation of demand time series of WDN users was developed. The methodology applies a copula-based re-sort to demand time series generated through a Beta or Gamma probability distribution. The methodology, applied to the literature case study of Milford (Ohio) and the Soccavo DMA, was able to reproduce the main statistics of measured demand time series and preserve their spatial and temporal cross-correlations. Thus, the methodology can effectively assist water utilities in design and management of WDNs by providing accurate estimates of water demand for numerical simulations of WDNs behaviour. Then, a comparative study of the performance of the proposed bottom-up methodology and a top-down one was carried out. The top-down methodology consists of a non-parametric disaggregation model based on the K-nearest neighbours approach. The comparison was performed by considering two case studies, both referred to the Soccavo DMA, with a different number of users. The bottom-up methodology performed better in reproducing cross-correlations between single users, and between single nodes. Whereas the top-down methodology was more effective in reproducing skewness and rank cross-correlations for spatially aggregated time series. For both methodologies, high levels of aggregation in nodes were found to be beneficial to preserve rank cross-correlations.

Successively, the focus was set on the reconstruction of the total temporal demand pattern of a DMA. Such aspect is important for water utilities for a variety of operational tasks, e.g. for the detection of anomalous events, such as unauthorized consumption and leakages. To address this issue, two procedures are proposed that can be applied when smart meters, originally installed at all locations, have to be replaced and when no smart meter is present, respectively. The first procedure uses the stepwise regression for the selection of the smart meters to be replaced for accurately reconstructing the total demand pattern. The second procedure consists of applying different criteria, based on easily available data (e.g. consumption on the annual bill and user typology), to identify the set of representative users to be provided with smart meters. Then, a novel linear model based on users billed annual consumption is used to estimate the total DMA demand. The procedures were applied

to the Soccavo DMA. In both cases the accuracy of the total demand pattern reconstruction was good already for low number of selected users. Identifying the users with the highest consumption, while distinguishing between the different categories of users, was the most effective strategy for the accurate reconstruction of the total demand. Overall, both procedures allow water utilities to optimize smart metering systems and reduce their costs.
Considering the increasing importance of water demand forecasting in light of the future climate changes, the prediction accuracy of models based on weather variables and a common machine learning technic, i.e. the Random Forests, was also investigated. The analysis was carried out for the Soccavo DMA by disaggregating water consumption based on the social characteristics of the users. The models were able to forecast the aggregated daily water demand, though their performances changed depending on the social characteristics of the users. The obtained results are useful for assessing future variations in water demand due to climate variability, thus reducing risks of supply and operational failures in WDNs.

In conclusion, the methodologies and the results presented in this thesis are expected to aid water utilities in developing more sustainable and cost-effective WDN management strategies.

Keywords: smart meters, water demand, water distribution network, district metered area, water demand model, water distribution network management.

Abstract

Per limitare gli impatti dei cambiamenti climatici e della pressione antropica sulle risorse idriche sono necessarie strategie efficaci per la gestione sostenibile delle reti di distribuzione idrica. A tal fine, tecnologie innovative per la telelettura dei consumi idrici, come gli smart meters, rappresentano un valido strumento le cui potenzialità non sono state ancora del tutto approfondite. Nel presente elaborato di tesi sono proposte nuove metodologie per la caratterizzazione della domanda idrica finalizzate a migliorare la gestione delle reti idriche e a ottimizzare l'utilizzo degli smart meters. Per lo sviluppo di tali metodologie si è usufruito dei dati registrati dal sistema di smart metering del distretto idrico di Soccavo (Napoli).

Nell'ambito del presente lavoro di tesi, è stata sviluppata una nuova metodologia bottom-up per la generazione di serie di dati sintetici di domanda idrica. Secondo tale metodologia, le serie sintetiche di domanda idrica, generate attraverso la funzione di probabilità Beta o Gamma, vengono riordinate tramite una funzione copula. La metodologia, applicata al caso studio di letteratura di Milford (Ohio) e al distretto idrico di Soccavo, è stata in grado di riprodurre le statistiche principali delle serie di domanda misurate e di preservare le correlazioni spaziali e temporali esistenti; può quindi contribuire in maniera efficace alla progettazione e alla gestione delle reti idriche, fornendo stime accurate della domanda idrica per le simulazioni idrauliche.

È stato poi condotto uno studio comparativo tra la metodologia bottom-up sviluppata e una metodologia top-down, che si fonda su un modello di disaggregazione non parametrico basato sull'algoritmo dei K-nearest neighbours. Nell'analisi comparativa si è fatto riferimento a due casi studio, entrambi relativi al distretto di Soccavo ma con un diverso numero di utenti. La metodologia bottom-up è risultata migliore nel preservare le correlazioni tra singoli utenti e tra singoli nodi. Invece, la metodologia top-down è risultata più efficace nel riprodurre il coefficiente di asimmetrica e le correlazioni delle domande aggregate. Per entrambe le metodologie è stata ottenuta una maggiore accuratezza nel riprodurre le correlazioni con elevati livelli di aggregazione spaziale nei nodi.

Il lavoro di tesi si è concentrato poi sull'importanza per diversi fini operativi, come l'identificazione di perdite idriche, di una ricostruzione accurata della domanda totale dei distretti idrici. A tal fine, sono state sviluppate due procedure innovative applicabili rispettivamente nel caso di un distretto in cui gli smart meters precedentemente installati debbano essere sostituiti e nel caso di un distretto in cui non sia presente un sistema di smart metering. La prima procedura si basa sull'applicazione di un modello di regressione stepwise per la selezione di un numero limitato di smart meters da sostituire per una ricostruzione accurata della domanda totale del distretto. Invece, la seconda procedura prevede l'applicazione di diversi criteri, basati sui consumi fatturati e la tipologia di utente, per

individuare gli utenti da monitorare per ricostruire la domanda totale. In questa procedura si applica un nuovo modello di regressione lineare per la stima della domanda totale basato sui consumi annuali fatturati. Dall'applicazione delle due procedure al distretto di Soccavo sono stati ottenuti buoni livelli di accuratezza già a partire da un numero limitato di utenti selezionati. Inoltre, selezionare gli utenti con i consumi maggiori in base alla loro tipologia è risultata la strategia di ottimizzazione più efficace. In definitiva, entrambe le procedure consentono l'ottimizzazione dei sistemi di smart metering, riducendone i costi.

In considerazione della crescente importanza della previsione della domanda idrica alla luce dei futuri cambiamenti climatici, è stata analizzata l'accuratezza di modelli di previsione basati sulle variabili meteorologiche e su un noto algoritmo di machine learning, il Random Forests. Le analisi sono state condotte per il distretto di Soccavo disaggregando la domanda idrica in base alle caratteristiche sociali degli utenti. I modelli sono stati in grado di prevedere la domanda giornaliera aggregata con diversi livelli di accuratezza a seconda delle caratteristiche sociali degli utenti. I risultati ottenuti possono essere utilizzati per determinare future variazioni della domanda idrica legate ai cambiamenti climatici, riducendo i rischi di carenze nell'approvvigionamento idrico.

In definitiva, le metodologie e i risultati descritti nel presente elaborato di tesi possono contribuire a una gestione delle reti idriche più sostenibile dal punto di vista ambientale e economico.

Acknowledgements

First I would like to thank my supervisor Prof. Maurizio Giugni for giving me the chance to join the PhD program and develop my research, for his assistance and advice over the last years. Next, I would like to thank my second supervisor Prof. Francesco De Paola, for always listening and helping me solve any issue and achieve my goals.

I would also like to express my great gratitude to my co-supervisor Prof. Enrico Creaco for his guidance and effort. He has steered me through my PhD and he has given me the invaluable chance to develop transferable skills. I wish, of course, to thank also Dr. Giacomo Galuppini for his ideas and his contribution to this thesis. Working with both of them has significantly enhanced my research skills. Also thank you to all the colleagues and professors that I met at the Department of Civil Engineering and Architecture of the University of Pavia for hosting me and making me so welcome each time.

I am very grateful to Prof. Zoran Kapelan for accepting me at the Department of Water Management of TU Delft, helping me improve this thesis with his comments and sharing his valuable expertise. I am also sincerely thankful to all the colleagues with whom I shared the office and this great experience at the TU Delft. Thank you all for being great friends, perfect colleagues and incredibly supportive. A big thank you also goes to Dr. Maria Xenochristou for her effort, help and advice during our work together.

Thank you also to Prof. Chrysi Laspidou for her effort in reviewing this thesis.

I would like to thank the Water Company ABC - Acqua Bene Comune Napoli for providing the data used in this thesis.

I wish to thank all my colleagues at the Department of Civil, Building and Environmental Engineering of the University of Naples Federico II with whom I shared most of my PhD, for their friendship, help and all the meaningful moments spent together. Special thank you to Maria Cristina Morani and Angela Romano for sharing with me all the difficulties and the goals achieved from the beginning to the end of our experience as PhD students.

A huge thank you goes to Dr. Gerardo Caroppi for his thoughtful and encouraging comments about my research and this thesis. No matter what time or how busy he was, he has always been there to listen to my worries, help me or engage with my work. For this and for supporting me every day over these years I am very thankful.

I am profoundly thankful to my friends and my family for always giving me their invaluable support, cheering me up and believing in my capabilities. I am most grateful to my parents for providing me with education, help and opportunities to achieve my goals.

Table of Contents

1 Introduction 1

1.1 Smart Water Distribution Network 2

1.1.1 Smart metering features 3

1.1.2 Soccavo District Metered Area 5

1.2 Water Demand Modelling 8

1.2.1 Investigating water demand patterns 9

1.2.2 Water demand time series generation 11

1.2.3 Water demand forecasting 12

1.3 Objectives 15

1.4 Thesis overview 16

1.5 Publications related to this thesis and other resources 18

References 21

2 Bottom-up generation of water demand time series 32

2.1 Methodology 33

2.1.1 Copula functions 34

2.1.2 Generation of demand time series 35

2.1.3 Imposition of spatial and temporal cross-correlation 36

2.2 Data 37

2.3 Results and discussion 38

2.3.1 Results – Case study 1 39

2.3.2 Results – Case study 2 42

2.4 Summary and conclusions 44

References 46

3 Water demand generation: comparison between bottom-up and top-down approaches 49

3.1 Methodology 50
3.1.1 Top-down methodology 50
3.1.2 Disaggregation model 51
3.2 Case studies 52
3.2.1 Applications 53
3.3 Results and discussion 54
3.3.1 Results – Case study 1 54
3.3.2 Results – Case study 2 59
3.4 Summary and conclusions 62
References 64
4 Identification of optimal smart meters locations for water demand pattern reconstruction 67
4.1 Methodology 68
4.1.1 Identification of the representative users based on measured demand 68
4.1.2 Identification of the representative users based on billed demand 71
4.2 Case study 75
4.3 Results and discussion 76
4.3.1 Procedure 1 - Results 76
4.3.2 Procedure 2 - Results 79
4.4 Summary and conclusions 83
References 86
5 Water demand forecasting by using weather data 89
5.1 Data 90
5.1.1 Consumption data and characteristics of the users 90
5.1.2 Weather data 91
5.2 Methodology 92

5.2.1 Random Forests....93

5.2.2 Weather-based predictive models....95

5.3 Results and discussion....97

5.3.1 RF models tuning....97

5.3.2 Prediction accuracy of RF weather-based models....99

5.4 Summary and conclusions....102

References....105

6 Conclusion....111

6.1 Thesis summary....111

6.2 Limitations and future directions....114

6.3 Thesis contributions....117

References....119

Appendix: Preliminary analysis – Chapter 5....122

A.1. Relationship between weather variables....122

A.2. Relationship between water demand and weather variables....123

References....129

Notation....131

List of Tables

Table 1 - Comparison of metering technologies (modified from Pericli and Jenkins, 2015). 3

Table 2 - Comparison of mean values (μ), standard deviation values (σ), skewness values (γ) and rank cross-correlations values (ρ) of measured and generated hourly demands, evaluated in terms of R^2 at both single user and aggregated scale, for the Case study 1. 40

Table 3 - Comparison of mean values (μ), standard deviation values (σ), skewness values (γ) and rank cross-correlations values (ρ) of measured and generated hourly demands, evaluated in terms of R^2 at both single user and aggregated scale, for the Case study 2. 43

Table 4 – Details of the performed applications. 54

Table 5 - Comparison of mean values μ, standard deviation values σ, skewness γ, rank cross-correlations ρ and $\rho - lag\ 0$ of measured and generated demands, evaluating the fit in terms of R^2 at both single and aggregated scales, for both applications to the first case study and for both the top-down and the bottom-up methodologies. 59

Table 6 - Comparison of mean values μ, standard deviation values σ, skewness γ, rank cross-correlations ρ and $\rho - lag\ 0$ of measured and generated demands, evaluating the fit in terms of R^2 at both single and aggregated scales, for each application of top-down methodology to Case study 2. 60

Table 7 - Comparison of mean values μ, standard deviation values σ, skewness γ, rank cross-correlations ρ and $\rho - lag\ 0$ of measured and generated demands, evaluating the fit in terms of R^2 at both single and aggregated scales, for each application of bottom-up methodology to Case study 2. 61

Table 8 - Criteria for the selection of the users to be monitored for a reliable total district demand characterization. The criteria refer to the cumulative frequency distributions of users' total (F_t), residential (F_r) and non-residential (F_{nr}) billed annual demand to select n, n_r and n_{nr} users among all, residential and non-residential users respectively. 73

Table 9 - Values of the coefficients α_i obtained by applying the stepwise regression using as stopping criterion the Fisher's test. 77

Table 10 - Results in terms of coefficient of determination (R^2) obtained by the stepwise regression. 78

Table 11 - Results in terms of coefficient of determination (R^2) obtained by the application of the stepwise regression and the Best Criterion for different numbers of users selected. 83

Table 12 - Classification of the households of the DMA according to their main social characteristics. 91

Table 13 - Explanatory (input) variables of each model. 96

Table 14 - *Description of household groups.* *96*

Table 15 - *Results in terms of root mean square error (RMSE) and coefficient of determination (R^2) obtained at aggregated scale for each model for the validation dataset.* *99*

Table 16 - *Results in terms of root mean square error (RMSE) and coefficient of determination (R^2) obtained at aggregated scale for the groups of households for each model.* *101*

Table 17 – Description of the weather variables. *122*

Table 18 – Spearman's coefficients between each pair of weather variables whit $p - value < 0.01$. *123*

Table 19 – Temporal segmentation and users groups. *124*

Table 20 - *Spearman's coefficients between each weather variable and daily demand with p-value < 0.01, for the entire DMA, Group 1, Group 2 and Group 3.* *125*

List of Figures

Figure 1 – Case study location. 5

Figure 2 - Boundaries of the Soccavo district and the DMA (red area). 6

Figure 3 - Wireless fixed data collection system architecture (modified from Bettin and Rogers, 2012). 7

Figure 4 – Overview of the thesis and the main contents of the chapters. 17

Figure 5 - Flow-chart of the methodology (modified from Creaco et al., 2020). 33

Figure 6 - Patterns of aggregated measured hourly demand in 31 days (each colour refers to one day) for the first (a) and the second (b) case study (Creaco et al., 2020). 38

Figure 7 - Comparison of mean values μ (a), standard deviation values σ (b), skewness γ (c) and rank cross-correlations ρ (d) of measured and generated single user demands for Case study 1 (Creaco et al., 2020). 39

Figure 8 - Comparison of mean values μ (a), standard deviation values σ (b), skewness γ (c) and rank cross-correlations ρ (d) of measured and generated demands at aggregated scale for Case study 1 (Creaco et al., 2020). 41

Figure 9 - Daily temporal pattern of first (blue), second (black) and third quantile (red) for measured (dots) and generated (lines) aggregated demand time series, for Case study 1 (modified from Creaco et al., 2020). 42

Figure 10 - Daily temporal pattern of first (blue), second (black) and third (red) quantile for measured (dots) and generated (lines) aggregated demand time series, for Case study 2 (modified from Creaco et al., 2020). 43

Figure 11 - Patterns of aggregated measured hourly demand for 31 days for (a) Case study 1 and (b) Case study 2 (Fiorillo et al., 2020). 53

Figure 12 - Case study 1 (Application 1.1) - Comparison between measured and generated single user demands: (a) mean values μ, (b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ (black) and $\rho - lag$ 0 (red), for the top-down methodology (Fiorillo et al., 2020). 55

Figure 13 - Case study 1 (Application 1.1) - Comparison between measured and generated aggregated demands: (a) mean values μ, (b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ, for the top-down methodology (Fiorillo et al., 2020). 56

Figure 14 - Case study 1 (Application 1.2) - Comparison between measured and generated single node demands: (a) mean values μ, *(b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ (black) and $\rho - lag$ 0 (red), for the top-down methodology (Fiorillo et al., 2020)....................................... 57*

Figure 15 - Case study 1 (Application 1.2) - Comparison between measured and generated aggregated demands: (a) mean values μ, *(b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ, for the top-down methodology (Fiorillo et al., 2020).. 58*

Figure 16 - Case study 2 (Application 2.2) - Daily temporal pattern of first (blue), second (black) and third (red) quantile for measured (dots) and generated (lines) aggregated demand time series: the generated demands were obtained applying the top-down (a) and the bottom-up (b) methodology respectively. 61

Figure 17 – Flow-chart of Procedure 1 (Fiorillo et al., 2020). 69

Figure 18 – Flow-chart of Procedure 2 (modified from Fiorillo et al., 2020)... 72

Figure 19 - Daily temporal patterns of total aggregated hourly demand and aggregated hourly demand of residential and non-residential users, respectively (Fiorillo et al., 2020). 75

Figure 20 - Cumulative frequency distributions (F) of users' total, residential, and non-residential mean daily demand (Fiorillo et al., 2020). 76

Figure 21 - Procedure 1 - Results in terms of coefficient of determination (R^2) obtained by the stepwise regression for the dataset of identification, validation and the whole dataset for different orders of the model (n) (Fiorillo et al., 2020). 78

Figure 22 - Procedure 2 - Comparison of the coefficient of determination (R^2) obtained using different criteria for each number of selected users (n) (Fiorillo et al., 2020)... 79

Figure 23 - Weekly temporal patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by the selection of 25 users through the application of Procedure 1 and the Best Criterion from Procedure 2 (Fiorillo et al., 2020).. 80

Figure 24 - Weekly temporal patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by the selection of 25, 50 and 100 users through the application of the Best Criterion from Procedure 2 (Fiorillo et al., 2020). 81

Figure 25 - Daily temporal patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by the selection of 25, 50 and 100 users through the application of Procedure 1 and Best Criterion from Procedure 2 (a) and residuals for aggregated hourly demands obtained by the selection of 25, 50 and 100 users through the application of the Best Criterion from Procedure 2 (modified from Fiorillo et al., 2020)... 82

Figure 26 - Boxplot of seasonal measured values of daily maximum temperature (a) and daily mean solar radiation (b).......... 92

Figure 27 - Training (a) and prediction (b) phase of Random Forest regression. 94

Figure 28 – Plot of RMSE values for the calibration dataset for various combinations of the hyperparameters for Model 1. 98

Figure 29 - Plot of RMSE values for the calibration dataset for various combinations of the hyperparameters for Model 2. 98

Figure 30 - Comparison between measured and forecasted aggregated daily demand for Models 1 (a) and Model 3 (b) (Fiorillo et al., 2021). 100

Figure 31 - Correlation between maximum daily temperature (°C) and daily total district demand during spring (a) and winter (b). 126

Figure 32 - Correlation between daily mean solar radiation (W/m^2) and daily total demand of Group 1 (a) and Group 2 (b) during working days. 127

1 Introduction

Water is essential for all living beings and the preservation of their natural environments. However, water resources are currently threatened by climatic changes, urbanisation, pollution and increasing water consumption due to the global population increase. The spatial and timely delivery of water also affects society and economy, since water is needed for public supply, agriculture, energy production and manufacturing. Even though the 71% of the Earth's surface is covered with water, only 2.5% of the total volume is freshwater. Furthermore, not all of the freshwater is available for consumption (Martyusheva, 2014). Given the risks to water security, it is important to well understand benefits and limitations of water management. In urban contexts, promoting adoption of water conservation practices is today considered an essential element of future water security (Arbués et al., 2003; De Loë et al., 2001).

Water demand can be defined as the water required by end-users for various usages (Merrett, 2004; Rinaudo, 2015). In this respect, water metering, namely the practice of measuring water use, has emerged as a key component of water demand management for water utilities. Water metering has been driven by the possibility to recover costs directly from the consumers and to apply the user-pays principle, allowing the users to pay for the actual amount of water used (Pericli and Jenkins, 2015). Traditional metering is characterized by the need of a meter-reading service, occurring few times per year, and the lack of a direct and easy access for users to their water consumption (Beal et al., 2013; Pericli and Jenkins, 2015). In response to these limitations, smart metering has emerged as fundamental tool in the evolution of demand management approaches, enabling remote meter reading and direct communication with the users.

The application of smart meter technologies is increasingly viewed as a useful tool to reduce water demand and facilitate more effective management of water distribution systems (Pericli and Jenkins, 2015). Smart metering data can be used to develop reliable water demand models aimed at optimising water distribution network (WDN) as well as effective strategies for water demand management. Additional data, including e.g. information on users and weather characteristics, can then be used to better understand and model water demand.

In this thesis smart demand metering data, as well as users' characteristics and weather data, were used to develop novel water demand models aimed at improving WDNs management. Furthermore, in order to optimise the use of smart meters, effective strategies for their allocation were developed. The present chapter provides the necessary background information. It starts describing the main characteristics of smart metering, then the key concepts of water demand modelling based on the literature review are provided. Lastly, the objectives of the thesis and the available resources, including publications and codes, are presented.

1.1 Smart Water Distribution Network

Smart water metering is part of an environmentally friendly policy aimed at minimizing water waste and energy consumption associated with the supply to end-users (Brentan et al., 2017a; Fox et al., 2009; Gato-Trinidad and Gan, 2011). Traditional metering can lead to several problems. Firstly, the meters can be hardly accessible, restricting users from being able to determine their water consumption. Secondly, metering data are usually provided long after the water usage, preventing responsive changes in behaviour from end-users. Furthermore, traditional meters generate recurrent financial costs for water utilities, since a meter-reading service is required to gain the consumption data (Beal et al., 2013). In view of these limitations, smart meters emerged as an essential tool to improve water demand management. The use of smart metering technologies mainly occurred in the energy sector, being smart gas/electricity meters more widely applied (Pericli and Jenkins, 2015). As the potential benefits of smart meters have been recognised, water utilities also started to implement smart metering in order to improve services and water distribution efficiency. Indeed, there is a growing trend for water utilities to create Smart Water Distribution Networks (SWDNs), by providing water networks with sensor nodes and data loggers which enable the periodical transmission of the network state to data centres (Kartakis et al., 2015).

Table 1 briefly explains how smart meters differ from traditional metering devices. Smart meters identify water consumption in greater detail by recording data at very frequent time intervals (even every second) via series of linked sub-meters. Smart meters are able to record extra information, such as water quality data and temperature, giving the end-users a clearer picture of water status (Beal et al., 2013). In addition, smart meters enable automated reading and communication. This is achieved via automated drive-by readings or data transfer using networks (Hunn, 2010). Unlike traditional metering devices, smart meters allow data to be used by consumers as well as water utility. Users may access to an internet portal or application and utility can use real-time data to monitor the water network. Essentially, in case of traditional meters the transfer of data relies on manual collection and processing. Thus, users are able to access these data only through a bill at specific interval, such as every three or six months (or even every year). In contrast, smart meters give users direct access to real-time (or almost real time) data, helping users and water utilities to better understand their water usage in a specific time slot.

Table 1 - Comparison of metering technologies (modified from Pericli and Jenkins, 2015).

Traditional metering	Smart metering
• Accumulation based measurement of consumption • Meters are read manually – often by staff of the utility provider, and over a longer time period • A single consumption value is given to the consumer	• Accumulation, pulse or time interval based measurement of consumption • Data logger and transponder • Remote meter reading • Multiple consumption values based on different time variables • Consumption data available through network connection

1.1.1 Smart metering features

Two main metering approaches exist, intrusive and non-intrusive metering (Cominola et al., 2015). The first consists of installing high-resolution sensors on each water-consuming appliance (e.g. toilet flush, washing machine, showerhead) and is generally costly and hardly accepted by end-users (Cordell et al., 2003; Kim et al., 2008). Non-intrusive metering is a more applicable alternative (Mayer et al., 1999) since it comprises water consumption measurements at household level.

One of the main reasons for using smart meters is to raise awareness of water consumption at individual user level. Previous studies showed that analysing water consumption at a fine temporal scale can generate virtuous behaviour in end-users, thus resulting in consumption reduction (Blokker, 2010; Britton et al., 2013; Makropoulos et al., 2014). Furthermore, smart metering data can provide a better understanding of consumption (Pericli and Jenkins, 2015) and the assessment of recurrent water use routines, allowing to improve water demand management strategies (Cominola et al., 2019). Indeed, measured demand time series can be used to improve water demand modelling. Numerous water demand models were developed thanks to the availability of metering data, as illustrated in detail in section 1.2. From an operating view point, an important feature of smart metering is the ability to apply more sophisticated tariffs. More flexible tariff structure can be applied to encourage water efficient behaviours (Zetland, 2011).

A further application of smart demand data is leak detection. On one hand, through real-time monitoring, water providers can quickly respond to water leaks, minimising operational costs, water waste and service issues for users (Pericli and Jenkins, 2015). On the other hand, analysing the revenue water – i.e. the users' consumption – water utility can identify the occurrence of anomalous events, such as water thefts and spills from pipe bursts. In recent years several technologies and strategies have been developed to allow water utilities to identify, measure and reduce leaks. Among them, a particularly cost-effective technique for quantifying the amount of water losses, is the network flows analysis. Through the districtualization (Galdiero et al., 2016; Savić et al., 2014; Stavenhagen et al., 2018; Vašak et al., 2014) large WDNs are usually subdivided into smaller and more easily manageable areas, called District Metered Areas (DMAs). Then, anomalous event detection in WDNs is typically performed by applying statistical methodologies to the total consumption pattern of a DMA (Alcocer-Yamanaka et al., 2012; Sophocleous, 2018; See Wong et al., 2010). Under these circumstances, the use of smart meters can improve the anomalous events identification by comparing the pattern of the total revenue water of a DMA with the corresponding inflow, usually monitored with flowmeters at DMA boundaries. If smart meters are available at all user locations, the pattern of the total revenue water can be easily calculated as the sum of the single readings at any temporal resolution. However, water utility managers may find the installation of smart meters at each user location exceedingly expensive to apply to WDNs.

Therefore, a suitable trade-off between the need for an accurate characterization of the total district demand and smart metering costs is needed. This issue can be addressed by identifying a limited number of users to be monitored for an accurate reconstruction of the temporal pattern of the DMA demand. The assessment of the minimum number of users to be monitored for reliable demand characterization can be required in two operating conditions. The first case consists of a DMA in which smart meters at the end of life, that is, 10–15 years after their installation, were previously installed at all user locations. In this case, the possibility of replacing a number of smart meters lower than the initial one represents a significant cost saving for the municipal water company. The second is a DMA where smart metering needs to be implemented for the first time. In this case, when the number of smart meters to be installed is fixed and a reliable estimate of the total district demand pattern is needed, the users to be supplied with a smart meter must be appropriately identified.

Overall, in order to ensure the widespread implementation of SWDN, the challenges relating to the high cost of smart metering technology should be overcome (Beal et al., 2013). The implementation phase should be managed appropriately to minimise costs and avoid technical problems, logistical issues and lack of preparation of water utilities. Therefore, effective strategies for SWDN management and optimization of smart meters allocation are needed. For developing new methodologies for the optimization of smart meters allocation, the availability of large amount of metering data from well monitored pilot areas is of fundamental importance. In the framework of this thesis, the data from the Soccavo DMA (Naples, Italy) were used. These data were collected and made available by the municipal Water Company ABC - Acqua Bene Comune Napoli. In the following, a brief description of the Soccavo DMA is provided.

1.1.2 Soccavo District Metered Area

The present study largely uses the smart metering data from Soccavo, a suburban area in the North-Western part of Naples, Italy (Figure 1). Soccavo district was chosen for a SWDN implementation, as part of a cooperation between the University of Naples Federico II and the municipal Water Company ABC - Acqua Bene Comune Napoli. Soccavo covers an overall area of 5.1 km², has a population of 45314 inhabitants (last survey, dated 2011) and is mainly a residential area. The topographic, urban and hydraulic characteristics of its water distribution network well represent common network conditions in Italy. The area is characterized by high variability in elevation and buildings height. The primary water network is made of different material (cast iron, ductile iron and steel), thus, the area is representative of both recent and old urban areas (such as historical centres). The water meters are located in different sites, such as: inside homes, in batteries at the basis of the buildings, and in underground structures. This allowed to test the data transmission under different working conditions.

Figure 1 – Case study location.

A DMA was identified on the basis of survey campaigns, including 5380 water meters. Then, 4989 traditional meters among all those present in the DMA were replaced by smart meters. The boundaries of the DMA are shown in Figure 2. The connections comprise residential flow meters (representing almost the 85% of the total number of meters) and non-residential flow meters (15%), mainly consisting of commercial activities.

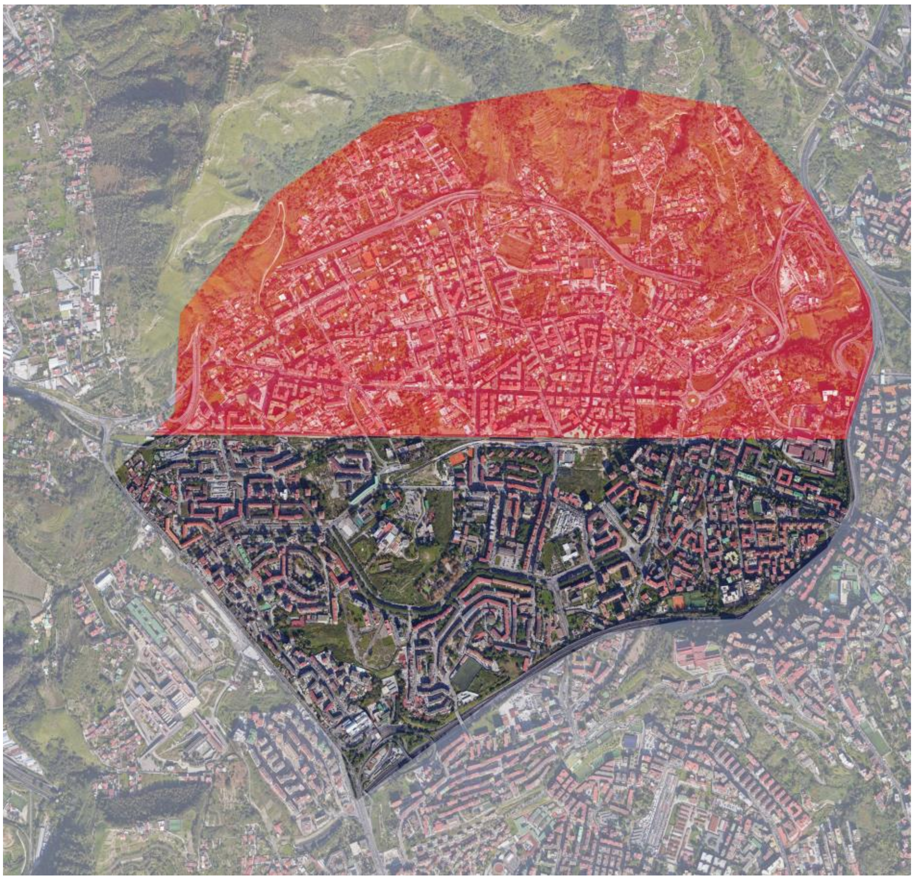

Figure 2 - Boundaries of the Soccavo district and the DMA (red area).

Specifically, the EverBlue wireless fixed data collection system provided by Itron was adopted. The system architecture is shown in Figure 3. Meter Interface Units (MIUs) are used to connect each meter to the wireless fixed network. The MIUs constantly monitor the meters and automatically communicate daily data to the utility server. MIUs transmit 24 hourly meter readings along with additional useful information (such as leakages, overflow, backflow, tampering, meter blocked or oversized/undersized), enabling early warnings to the end-users. The MIUs are connected to the collectors, which are dual-band radio routers. Once a day, each collector automatically receives data from a group of MIUs via Low-Power radio frequency (433 MHz/10 mW). Then, the stored data are daily transmitted to the Access Point (AP) by using High-Power radio frequency (868 MHz/200 mW). Six sub-districts were defined within the DMA, each served by its own AP. A preliminary site survey was conducted to assess likely signal transmission noises. In this way, each collector was associated to the AP for which the transmission was more stable. Each AP daily collects the hourly meter readings from the related group of collectors. The collected data, available for periods as long as one year, are transmitted to the utility server via GPRS communication at scheduled times (usually on a daily basis). Finally, the data from all APs are gathered by the FTP (File Transfer Protocol) server, becoming available for use by multiple operators. Daily data are automatically downloaded from the FTP server and can be easily exported using various export formats.

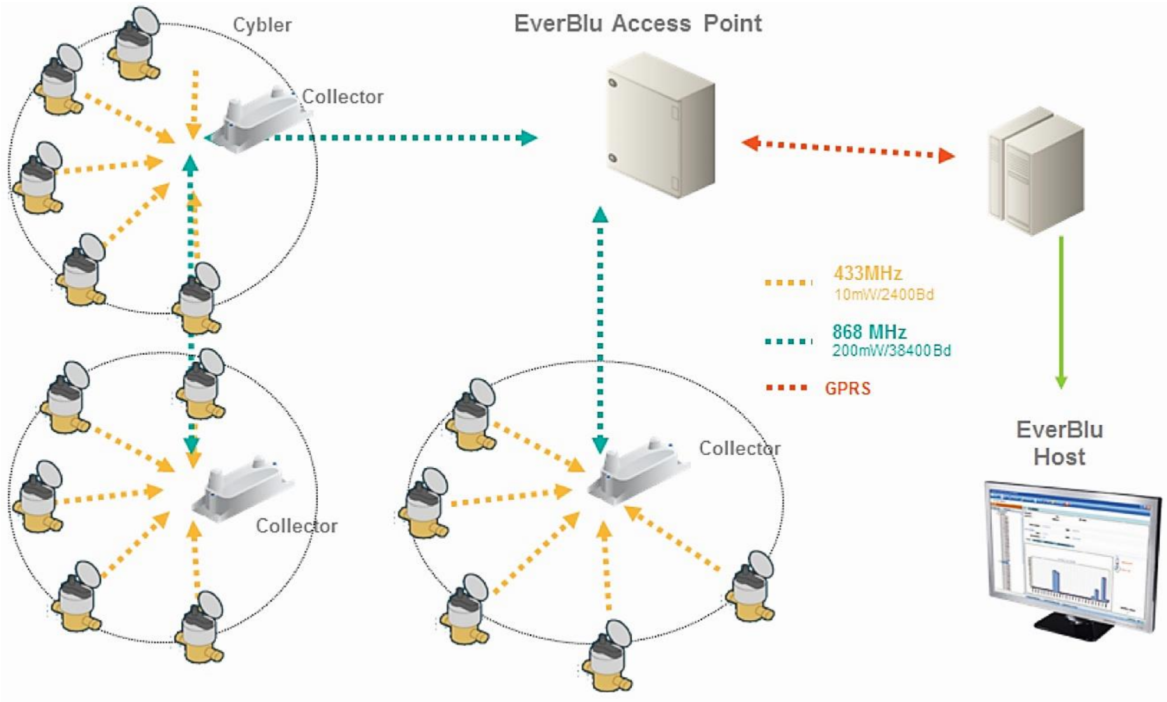

Figure 3 - Wireless fixed data collection system architecture (modified from Bettin and Rogers, 2012).

During the SWDN implementation, the Geographical Information System (GIS) of the DMA was built along with a database with all the information needed to identify the smart meters. The information available for each smart meter includes:

- the 9-digit code that identifies the user (i.e. the meter) and his position;
- the number of the contract signed by the user and the water utility;
- the name of the user that signed the contract for water supply and his address (street, building number and floor);
- the nominal diameter of the water meter;
- the 9-digit code of the MIU installed on the meter;
- the code that identifies the corresponding node in GIS;
- the number of the associated AP;
- type of user (residential or non-residential).

Such information is particularly useful for the correct interpretation of the recorded water consumption data.

The Soccavo SWDN is provided with the main smart metering features according to the European Standard EN14154 (Part 4 Additional Functionalities for Water Meters), i.e. alarm facilities for leakages/tampering, multiple readings at fixed intervals and improved communication with the end-users. Overall, the Soccavo DMA offered the possibility of collecting a large amount of water consumption data as well as useful information about the users (such as type and position). The extensive dataset was used to develop effective water demand models and SWDN management strategies, as shown in detail in the framework of this thesis. It is worth noting that, the Soccavo DMA represented a real case where effective strategies for smart meters allocation were needed. Indeed, some years after the SWDN implementation, the need to replace smart meters at the end of life emerged. On one hand, the Water Company ABC intended to replace only a limited number of smart meters to reduce smart metering costs. On the other hand, the Water Company needed to preserve the accurate reconstruction of the pattern of the total district demand. Therefore, the research described in this thesis started with the objective to provide an effective solution to this real operational problem, developing new strategies for the optimal allocation (or replacement) of smart meters. The strategies tailored to the Soccavo DMA can be easily extended to any other DMA.

1.2 Water Demand Modelling

Water demand modelling refers to a variety of numerical techniques used for investigating water demand under specific conditions. Specifically, water demand models can be developed for several purposes. One of the main objective of water demand modelling is to explore water demand patterns, aiming at improving the understanding of water uses. In water demand modelling several studies focused on investigating water demand patterns (section 1.2.1), by assessing the relationship between water demand and likely influencing factors (Beal and Stewart, 2014; Chang et al., 2010, 2014; Cole and Stewart, 2013; Goodchild, 2003; Hussien et al., 2016; Schleich and Hillenbrand, 2009; Willis et al., 2013; Xenochristou et al., 2020b), as well as identifying spatial trends (Chang et al., 2010; House-Peters et al., 2010) and users with similar consumption behaviours (Pullinger et al., 2013). A deeper understanding of water demand patterns and water uses can enable the development of effective water demand management strategies.

A further purpose of water demand models is to provide accurate estimates of nodal demands, through water demand time series generation models (section 1.2.2). Indeed, the assessment of nodal demands is needed for numerical simulation of WDNs behaviour. Synthetic water demand time series are used to obtain reliable results in terms of pipe discharges, nodal outflows and pressure-heads (Creaco et al., 2017b). More generally, water demand time series generation helps build more realistic hydraulic models (Brentan et al., 2018).

In addition, water demand modelling also includes forecasting purposes. Water demand projections, obtained by water demand forecasting models (section 1.2.3), are used for different objectives depending on the forecast horizon. Short-term forecasts (up to one month ahead) help to optimise operational and financial water systems management. Medium-term forecasts (from one to 10 years) are usually developed for planning improvements to water distribution and treatment systems, and adjusting water tariffs (Billings and Jones, 2008). In this case, the variability of water consumption due to weather and users behaviour changes is taken into account. Finally, long-term forecasts, covering time span ranging from 10 to 30 years, are useful in making long-term capital investments and implementing water conservation policies. Additional details on the models used for each purpose are provided in the following.

1.2.1 Investigating water demand patterns

Several approach have been proposed for analysing water demand, focusing on different aspects. In addition, it should be noted that, according to the resolution of the data available, the analysis can focus on identifying aggregated consumption patterns or on defining users' profiles (Cominola et al., 2015).

Descriptive statistics have been widely used to understand consumption data (Domene and Saurí, 2006; House-Peters et al., 2010; Pullinger et al., 2013). More specifically, measures, such as the mean and variance, have been used to determine the frequency of water demand occurrences. As regards the identification of similar water demand patterns, clustering methods have been proposed to gather users with similar consumption behaviours (Pullinger et al., 2013), while data visualization (Chang et al., 2010) has been used to gain additional information, such as the identification of spatial trends. GIS and spatial quantitative analysis techniques have become increasingly important in water demand analysis, making possible visualization and quantification of water use patterns across geographic areas (Chang et al., 2010; House-Peters et al., 2010; Lee and Wentz, 2008; Polebitski and Palmer, 2010; Shandas and Parandvash, 2010).

Several investigations have been focused on the identification of the most relevant drivers of water demand. In some studies, the influence of different factors on water demand was investigated by using statistical models. More specifically, these studies were based on piecewise and polynomial regression (Domene and Saurí, 2006; House-Peters et al., 2010; House-Peters and Chang, 2011;

Hussien et al., 2016) and log-log models (Schleich and Hillenbrand, 2009). The popularity of these models is due to their ease of use and interpretation. Another simple literature approach is represented by the use of correlation coefficients to estimate the relationship between water demand and likely influencing factors (Chang et al., 2010; Hussien et al., 2016; Xenochristou et al., 2020b).

In the existing literature many variables have been investigated as water demand influencing factors. Among them, the socio-economic characteristics of the users have been found important drivers of water demand (Ashoori et al., 2016; Balacco et al., 2018; Mamade et al., 2014; Manouseli et al., 2019; Parker, 2013; Romano et al., 2014; Villar-Navascués and Pérez-Morales, 2018; Wentz and Gober, 2007; Willis et al., 2013). Temporal characteristics, such as month, day of the week and season, have been also widely investigated. Indeed, different studies identified seasonal changes and, weekly and daily patterns in water demand (Cole and Stewart, 2013; Gato et al., 2007; House-Peters et al., 2010; Parker, 2013; Polebitski and Palmer, 2010; See Wong et al., 2010; Zhou et al., 2000). Finally, several studies have focused on the investigation of the effect of weather on water demand (Bakker et al., 2014; Balling et al., 2008; Beal and Stewart, 2014; Chang et al., 2014; Gato et al., 2007; Goodchild, 2003; Haque et al., 2017; Martinez-Espiñeira, 2002; Miaou, 1990; Mylopoulos et al., 2017; Slavíková et al., 2013; Statzu and Strazzera, 2009; Toth et al., 2018; Willis et al., 2013; Xenochristou et al., 2020b), accounting for a variety of weather variables.

From the literature review the following key knowledge gaps emerge:

- Even though a wide range of researchers has focused in achieving a better understanding of the complex spatial and temporal patterns of water demand (Lee and Wentz, 2008), few studies accounted for temporal and spatial variation of the effects of different variables on water demand (Xenochristou et al., 2020b). Therefore, temporal and spatial variability of the effects of water demand drivers should be more thoroughly investigated with reference to different case studies.
- Among the water demand influencing factors, users and temporal characteristics follow a more stable behaviour compared to weather fluctuations. Thus, the influence of weather on water demand remains one of the major uncertainties relating to water demand drivers.
- Climate changes are expected to affect water demand (Parandvash and Chang, 2016), leading to different variations in consumption according to geographic location and climatic conditions (Wang et al., 2014). Therefore, the weather impact on water demand should be investigated in different areas of the world to account for its variability. Despite the importance of such topic, few studies have investigated climate change effects on agricultural water demand and water supply in Italy (Bocchiola, 2013; Masia et al., 2018; Peres et al.,

2019). Moreover, to date no literature deals with the investigation of the impact of weather changes on urban water demand in Italy.

1.2.2 Water demand time series generation

Two different approaches are usually adopted for the assessment of nodal demands, the top-down and the bottom-up approach, respectively (Walski et al., 2003).

According to the first approach, the nodal demands are obtained by spatially disaggregating the total water demand time series of the whole WDN. In its common application, the water demand pattern is defined at high levels of spatial aggregation and then disaggregated into the individual nodes of the WDN in proportion to the average demand at each node. This deterministic approach does not consider the random character of water demand by assuming that all nodes are characterized by an identical demand pattern. Furthermore, it neglects spatial-temporal variability of water demand implicitly assuming the temporal patterns to be perfectly correlated in space (Alvisi et al., 2016). However, several studies (Blokker et al., 2010a, 2008; Filion et al., 2007, 2005) have highlighted the importance of taking into account the variability of water demand. In order to preserve the random nature of water demand, stochastic procedures for the disaggregation of time series have been developed. Among them, the models belonging to the hydrological field (Deidda et al., 2004; Koutsoyiannis and Manetas, 1996; Kumar et al., 2000) enable time series to be generated at different levels of aggregation while preserving the additivity property (i.e. the ability to generate time series whose sum is equal to the original aggregate time series), as well as the main distributional statistics (mean, variance and skewness) and spatial and temporal correlations of the disaggregated series (Alvisi et al., 2016). The disaggregation models available in the literature can be divided into two main approaches: parametric (Mejia and Rousselle, 1976; Santos and Salas, 1992; Todini, 1980) and non-parametric approach (Lee et al., 2010; Nowak et al., 2010; Tarboton et al., 1998). The parametric approach is based on a priori assumptions about the probability distributions of the modelled quantities, whereas the non-parametric models use observed nodal demands for their application.

The bottom-up approach, instead, reconstructs the WDN demand starting from the single nodes. In this case, the nodal demands are represented as stochastic variables. In order to obtain stochastic nodal demands, the results of pulse generation models can be temporal aggregated (Creaco et al., 2017a). Pulse models, such as the Poisson rectangular pulse process (Buchberger and Wu, 1995), operate at very small time steps (order of magnitude of one second) reproducing arrival time, intensity and duration of all pulses coming from individual appliances (such as dishwashers, washing machines, shower, etc.) in the generic household (Blokker et al., 2017; Buchberger et al., 2003). The Neyman–

Scott cluster has been also used for pulse demand generation (Alcocer-Yamanaka et al., 2012; Alvisi et al., 2003) providing more realistic bottom–up demand modelling. After adjustment of the Neyman–Scott parameters, a random distribution of water uses is combined with the elementary demands (such as those for washing machines and dishwashers) to generate single faucets demands. Due to the large computational burden of pulse models, different methods have been proposed when demand at larger time steps (from minutes to hours) is needed (Alvisi et al., 2014; Gargano et al., 2016). Gargano et al. (2016) proposed the overall pulse model based on the identification of the overall water demand of a single user at prefixed time steps. The random multinomial processes of Alvisi et al. (2014), instead, represent the demand of each user at the generic time step as a random variable and allow to reproduce the spatial and temporal cross-correlations. In this regard, previous studies (Blokker et al., 2008; Filion et al., 2007; Moughton et al., 2007) showed the importance of reproducing spatial and temporal cross-correlations, especially at large levels of temporal aggregation.

Concluding, the following key gaps in the knowledge on water demand time series generation models can be identified:

- Accurate synthetic demand time series are supposed to both preserve the main statistics (mean, standard deviation and skewness) of the measured time series and reproduce the existing cross-correlations between users and at all temporal lags. However, the generation of water demand time series which are consistent with the measured time series in terms of both statistics and correlations remains a challenging task, mainly when aggregated demand time series are considered. Further improvements are needed, since the large computational burden of existing water demand generation models is not always paid back by high levels of accuracy at both single user and aggregated scale (Alvisi et al., 2016; Creaco et al., 2017a).
- The assessment of the main differences in terms of accuracy and computational burden between the top-down and the bottom-up approaches based on real case studies could be beneficial to an improved understanding of water demand modelling. However, only few comparative studies between the top-down and the bottom-up approaches have been carried out so far (Blokker, 2010; Mamade et al., 2018; Sheng et al., 2017).

1.2.3 Water demand forecasting

Several water demand forecasting approaches have been proposed with the choice of the most suitable method depending on different factors. Besides the various horizons and planning levels, the forecast variable of interest and the determinants need to be assessed.

A survey by the American Water Works Association (AWWA), obtained from 662 North American water supply systems, portrayed the popularity among water utilities of various forecast variables in urban water demand forecasting. The major types of forecast variables and the percentage of utilities reporting each of them are reported below (Billings and Jones, 2008):

- peak day (73.9%);
- daily total system demand (65.9%);
- monthly total system demand (65.6%);
- annual per capita demand (65.4%);
- annual demand by customer class (58.0%);
- revenue forecasts linked with water demand forecasts (57.9%).

These results demonstrated that different variables with different resolution could be involved in water demand forecasting. Furthermore, in water industry many variables are considered factors influencing water demand (section 1.2.1), from socio-economic to weather variables (Donkor et al., 2014). Even if a good understanding of the influencing factors is needed to select effective predictors of water demand, due to the large number of these variables, choosing among them is often challenging. In addition, the forecasting models to be used also depend on the availability and choice of these independent variables.

When rudimentary forecasts are needed, qualitative methods, such as heuristics or rule-based methods, can be used for the sake of simplicity. These methods have been mostly used for unit water consumption forecasting, such as the consumption per unit of a customer category after disaggregating demand by customer segments (Brekke et al., 2002). Forecasting unit rates by rudimentary methods is the simplest approach used by most utilities (Billings and Jones, 2008). However, the reliability of these forecasts is questionable when simple rules of thumb are used.

A very common literature approach is using statistical models (Bakker et al., 2014; Downing et al., 2003; Fontanazza et al., 2014; Haque et al., 2014), since they integrate the effects of different factors, such as socio-economic, climatic and water policies-related factors. These models statistically estimate historical relationship between water demand and a variety of variables, providing water utilities with useful information about the factors affecting water demand.

Time series models (Brentan et al., 2017b; Chen and Boccelli, 2018; Kofinas et al., 2014) can be also used for water demand forecasting, by assuming that future trends can be predicted on the basis of past observations (Billings and Jones, 2008). This approach fails to take into account the effects of changes in demographic and economic variables and water demand management strategies. On the contrary, time series regression models account for the effect of exogenous variables by producing forecasts based on the relationship between water demand and its determinants (Polebitski and Palmer, 2010). However, regression models based on time series data are often characterized by the serial correlation of the error terms (Donkor et al., 2014). In this case, Autoregressive Moving Average models (ARMA) can be used to define the structure of the error terms for its integration with the model (Burney et al., 2001). In this context, Autoregressive Integrated Moving Average (ARIMA) models have been widely used (Kofinas et al., 2014; Quevedo et al., 2014) due to their ability to capture general and seasonal trends.

When there is a need to account for uncertainty in demand forecasts due to a limited number of combinations of the independent variables, scenario-based approaches are usually adopted. Scenario-based approaches basically determine the effect on water demand of various future scenarios of the determinants (Polebitski et al., 2011) and are generally used for long-term forecasts.

Several studies showed the effectiveness of machine learning algorithms in long-term water demand forecasting, as well as for short-term and medium-term forecasts (Bai et al., 2014; Bakker et al., 2014; Romano and Kapelan, 2014; Xenochristou et al., 2018; Xenochristou and Kapelan, 2020). Machine learning is a branch of artificial intelligence that allows computer systems to learn directly from data, enabling computers to perform specific tasks (Sharma and Kumar, 2017). Basically, the process of machine learning is to give training data to a learning algorithm that generates a new set of rules based on inferences from the data. This means generating a new algorithm formally referred to as the machine learning model. Machine learning techniques are usually divided in supervised learning, unsupervised learning, and reinforcement learning. Supervised learning consists in inferring a function from labelled training data (Sharma and Kumar, 2017). By analysing the training data, the algorithm produces an inferred function that can be utilized for mapping new data. Examples of supervised learning algorithm are Random Forest (RF) and Artificial Neural Networks (ANN). Unsupervised learning algorithms, such as Apriori algorithm and K-means, are used for identifying common features (i.e. inputs used for prediction or classification) and creating clusters of data points (Antunes et al., 2018), being the outcome unknown. Finally, in reinforcement learning the training data provide an indication as to whether an action is correct or not, while the task is to learn a control strategy in an unknown dynamical environment, with the aim to maximise a reward (Jordan and Mitchell, 2015). An example of reinforcement learning is Markov Decision Process. Although machine learning technics are becoming increasingly popular and more sophisticated, they are often considered black-box model. Therefore, the results obtained are difficult to interpret and directly use to develop water demand management strategies.

Finally, the approach based on hybrid models has found wide application in water demand forecasting (Alvisi et al., 2007; Anele et al., 2017; Bakker et al., 2014; Wang et al., 2009). Hybrid models combine different methods, improving positive and reducing negative capabilities of individual ones (Kofinas et al., 2014). These models usually involve the combination of forecasts from different models based on simple or weighted averages (Wang et al., 2009) or obtain forecasts at lower periodicities by adjusting higher periodicity forecasts (Alvisi et al., 2007). Also in this case, the interpretation of the results could be hard, since these models make prediction by combining the results of individual learners.

Overall, the limitations of existing water demand forecasting models relate to two main aspects:

- The most accurate existing models require the use of many variables or complex derivatives, becoming difficult to use for water utilities. In order to improve urban water demand management, forecasting models based on input variables that can be easily collected, monitored and used by utilities are needed. Therefore, the ability of utilities to acquire and monitor predictors should be given greater consideration in developing future predictive models (Donkor et al., 2014).
- In view of the future climate changes, the predictive models should be able to catch the likely variations in water demand due to climate variability. This could help to avoid capacity and operational problems for the existing water systems due to unexpected future variations in water demand (Colombo and Karney, 2003). However, the ability of predictive models to assess variations in urban water demand due to climate changes has largely been unexplored.

1.3 Objectives

As emerged from the literature review, important gaps exist between water demand models and their practical application as effective tool for WDN management. Reaching a suitable compromise between the need of accurate results and the search for models for practical application is a challenging task. In this respect, the use of smart metering data offers a significant benefit. However, since smart metering data are not always available and SWDNs implementation could be exceedingly expensive for water utilities it is important to explore new strategies.

The present thesis explores the topic of urban water demand with a view on SWDNs. The general objective of this research is to develop new methodologies to improve WDN management and optimize the use of smart meters.

One of the specific objectives is to improve the accuracy of existing water demand generation models, trying to overcome their limitations. Specifically, the thesis aims to develop a model that could preserve the main statistics (such as media, standard deviation, skewness and correlations) as well as spatial and temporal correlations of the measured time series, at both single and spatially aggregated scale.

The second specific objective of this thesis is to investigate benefits and limitations of the top-down and the bottom-up approach for water demand generation. To this purpose, in the framework of this thesis a comparative study between the two different approaches is carried out to highlight the differences in synthetic time series generation in terms of accuracy and computational burden.

A further specific objective is to develop innovative strategies for effective SWDN implementation and management. Specifically, the thesis is aimed at providing new procedures to identify the optimal location of a small subset of smart meters for the accurate reconstruction of the total temporal demand pattern of a DMA, as to find a suitable trade-off between the need for an accurate characterization of the total district demand and smart-metering costs.

Furthermore, considering that smart metering data are not always available for water utilities, another specific objective of this thesis is to explore new strategies for water demand forecasting in case of lack of past consumption data.

Finally, given the expected climate change impact on water demand and its likely variation according to geographic and climatic conditions, a final objective is to overcome the lack of knowledge of weather effect on urban water demand in Italy.

1.4 Thesis overview

The thesis consists of four chapters (from Chapters 2 to Chapter 5) that present new methodologies and the results of their application. Then, a final chapter (Chapter 6) summarizes the main contents of the thesis. Finally, the thesis includes an Appendix that provides supporting information for Chapter 5. Chapters 2-5 are structured in order to address the research objectives described in section 1.3 as schematised in Figure 4. They are based on journal publications, as explained in details in section 1.5. In each of these chapters a literature review, as well as a description of the data used, is provided.

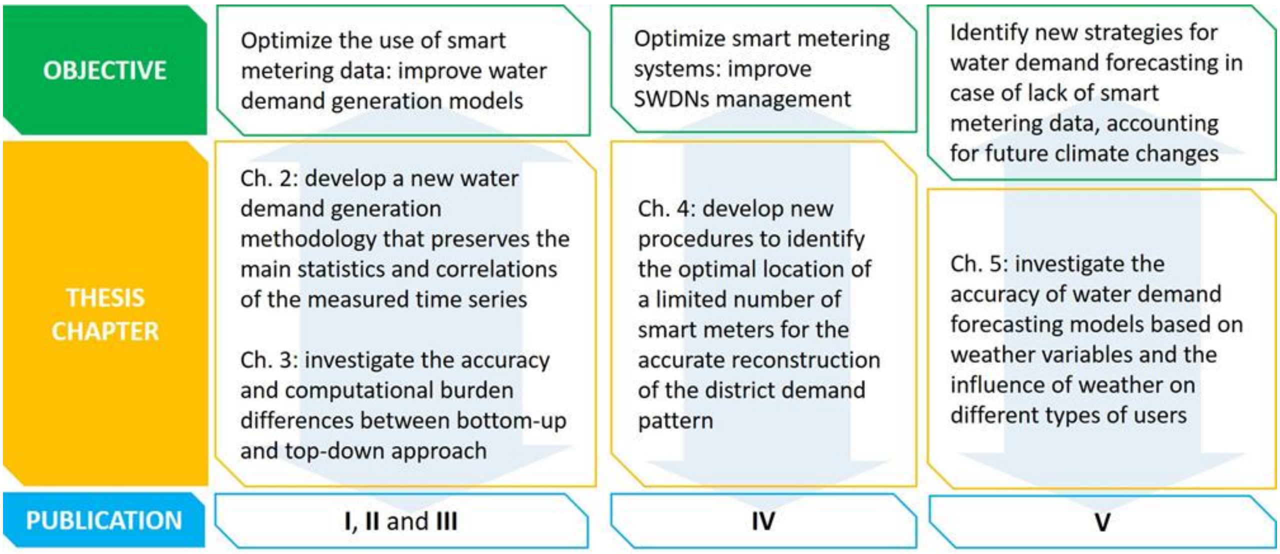

Figure 4 – Overview of the thesis and the main contents of the chapters.

In the following a brief summary of the chapters is reported.

In Chapter 2 a novel bottom-up methodology for the generation of demand time series is presented. The chapter is aimed at overcoming the limitations of existing water demand generation models. Thus, it provides a methodology that enables the preservation of the main statistics as well as spatial and temporal correlations of the measured time series, at both single and spatially aggregated scale. The methodology consists of a copula-based re-sort applied to demand time series generated through a Beta or Gamma probability distribution. In this chapter the results obtained applying the bottom-up methodology to two real case studies, namely the literature case study of Milford (Ohio) and the Soccavo DMA, are analysed. The accuracy of the methodology in reproducing the main statistics, as well as cross-correlations, of both single and spatially aggregated measured demand time series is thoroughly discussed.

Chapter 3 carries out a comparative study between the above bottom-up methodology and a top-down one. The top-down methodology relies on a non-parametric disaggregation model based on the K-nearest neighbours approach. The objective of the comparative study is to make a balance between the computational burden of the models and their level of accuracy in generating water demand time series. This chapter highlights the differences between the two methodologies in preserving the main statistics (i.e. media, standard deviation and skewness) and correlations of the measured time series. Indeed, the investigation is focused on the ability of the models of reproducing single user, nodal and aggregated demands. The chapter discusses the results of the two methodologies for two real case studies, both referred to the Soccavo DMA, differing for the number of users.

Chapter 4 addresses the objective of identifying new strategies for the optimization of SWDNs management. To this purpose, the chapter presents two procedures to identify the optimal allocation of a limited number of smart meters to obtain reliable estimates of the total district demand. The first procedure can be applied in a DMA in which smart meters at the end of life were previously installed at all user locations. This procedure uses the stepwise regression to assess the minimum number of smart meters to be replaced for a reliable demand characterization. The second procedure can be successfully applied during the initial phase of smart metering implementation to identify the users to be supplied with smart meters for an accurate reconstruction of the total district demand pattern. This procedure applies different criteria based on data available for any water utility, such as user typology and consumption on the annual bill. Then, it uses a novel linear model based on users billed annual consumption to estimate the total DMA demand. In this chapter the accuracy of the total demand pattern reconstruction resulting from both procedures is analysed for the Soccavo DMA, by showing the results obtained for different number of users selected. Finally, this chapter provides an effective strategy for water utilities to optimize smart metering systems while reducing their costs.

Chapter 5 is focused on the current need to obtain reliable water demand forecasts through models that can be easily applied by water utilities. Thus, the chapter explores new strategies for water demand forecasting based on information easily accessible to water utilities in case of lack of past consumption data. In light of the future climate changes, the accuracy of forecasting models based on weather variables and a common machine learning technic, the Random Forests model, is investigated for the Soccavo DMA. The investigation is carried out with the further objective of improving the knowledge of weather effect on urban water demand in Italy. Indeed, the analysis is conducted by disaggregating water consumption based on the social characteristics of the users to better determine the influence of weather on water demand. The chapter presents and discusses the results in terms of prediction accuracy obtained for the total district demand, as well as for the total demand of different groups of users.

Chapter 6 provides an overview of the thesis. This chapter summarizes the key findings of each chapter and highlights the limitations of the thesis, providing future research directions. Finally, the main contributions of thesis are presented.

1.5 Publications related to this thesis and other resources

The journal papers and proceedings related to this thesis are listed below:

I. Creaco, E., De Paola, F., Fiorillo, D., Giugni, M., 2020. Bottom-up generation of water demands to preserve main basic statistics and rank cross-correlations of measured time series. Journal of Water Resources Planning and Management, 146, 06019011. DOI: 10.1061/(ASCE)WR.1943-5452.0001142.

II. Fiorillo, D., Creaco, E., De Paola, F., Giugni, M., 2019. Generazione di domanda idrica a partire da serie storiche misurate/Generation of water demands starting from measured time series. L'Acqua, 5, 25-27.

III. Fiorillo, D., Creaco, E., De Paola, F., Giugni, M., 2020. Comparison of Bottom-Up and Top-Down Procedures for Water Demand Reconstruction. Water, 12, 922. DOI:10.3390/w12030922

IV. Fiorillo, D., Galuppini, G., Creaco, E., De Paola, F., Giugni, M., 2020. Identification of Influential User Locations for Smart Meter Installation to Reconstruct the Urban Demand Pattern. Journal of Water Resources Planning and Management, 146, 04020070. DOI: 10.1061/(ASCE)WR.1943-5452.0001269

V. Fiorillo, D., Kapelan, Z., Xenochristou, M., De Paola, F., Giugni, M., 2021. Assessing the Impact of Climate Change on Future Water Demand using Weather Data. Water Resources Management, 35**,** 1449–1462. DOI: 10.1007/s11269-021-02789-4

For publication I the author was mainly responsible for data collecting and quality control. The author also participated in the research design, interpretation of the results and writing the paper. Prof. Creaco was the main responsible for developing the codes used for the calculations and writing the paper. Prof. De Paola and Prof. Giugni participated in the research design and writing the paper. Publication II is based on the same codes used for publication I. As regards this publication, the author interpreted the results and wrote the paper, while Prof. Creaco participated in the research design and writing the paper, Prof. De Paola and Prof. Giugni provided comments on the paper.

Regarding publication III, the author designed the research, conducted the analysis, interpreted the results and wrote the paper. The author also partly developed the codes used for calculating the results along with Prof. Creaco, who also participated in designing the research and writing the paper. Prof. De Paola and Prof. Giugni participated in writing the paper.

For publication IV the author was mainly responsible for designing the research and developing the models and the codes. The author was also engaged in interpreting the results and writing the paper. Prof. Creaco and Dr. Galuppini participated in designing the research, developing the codes and writing the paper. Prof. De Paola and Prof. Giugni participated in writing the paper.

Concerning publication V, the author designed the research and developed the models and the codes. The author was also responsible for carrying the analysis, interpreting the results and writing the paper. Prof. Kapelan and Dr. Xenochristou participated in designing the research, developing the codes and writing the paper. Prof. De Paola and Prof. Giugni participated in designing the research and provided comments on the paper.

Chapter 2 is directly based on publications I and II, Chapter 3 on publication III, Chapter 4 on publication IV and Chapter V on publication V.

The data used in this study and in the related publications are not publicly available. The consumption and socio-economic characteristics data related to the Soccavo DMA were made available by Azienda Speciale Acqua Bene Comune Napoli (ABC) and are confidential in nature. The consumption data of the Milford households, used in Chapter 2, were made available by Prof. Buchberger, thus the author have restrictions on sharing them publicly.

The weather data, used in Chapter 5, were partly collected by the weather station of the University of Naples Federico II and are available upon request. Part of the weather data were recorded by the Italian Air Force weather station of Capodichino and are publicly available.

The codes used in Chapter 2 and 3 were developed in MATLAB and are available from the authors of the related publications. The codes used in Chapter 4 were developed in MATLAB and are available at the following github repository: https://github.com/DianaF92/Smart-water-networks. Finally, the codes used in Chapter 5 were developed in R and are available from the author.

References

Alcocer-Yamanaka, V.H., Tzatchkov, V.G., Arreguin-Cortes, F.I., 2012. Modeling of drinking water distribution networks using stochastic demand. Water Resour. Manag. 26, 1779–1792. https://doi.org/10.1007/s11269-012-9979-2

Alvisi, S., Ansaloni, N., Franchini, M., 2016. Comparison of parametric and nonparametric disaggregation models for the top-down generation of water demand time series. Civ. Eng. Environ. Syst. 33, 3–21. https://doi.org/10.1080/10286608.2015.1126823

Alvisi, S., Ansaloni, N., Franchini, M., 2014. Generation of synthetic water demand time series at different temporal and spatial aggregation levels. Urban Water J. 11, 297–310. https://doi.org/10.1080/1573062X.2013.801499

Alvisi, S., Franchini, M., Marinelli, A., 2007. A short-term, pattern-based model for water-demand forecasting. J. Hydroinformatics 9, 39–50. https://doi.org/10.2166/hydro.2006.016

Alvisi, S., Franchini, M., Marinelli, A., 2003. A stochastic model for representing drinking water demand at residential level. Water Resour. Manag. 17, 197–222. https://doi.org/10.1023/A:1024100518186

Anele, A.O., Hamam, Y., Abu-Mahfouz, A.M., Todini, E., 2017. Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction. Water 9, 887. https://doi.org/10.3390/w9110887

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A., Oliveira, M.S., 2018. Short-term water demand forecasting using machine learning techniques. J. Hydroinformatics. https://doi.org/10.2166/hydro.2018.163

Arbués, F., García-Valiñas, M.Á., Martínez-Espiñeira, R., 2003. Estimation of residential water demand: A state-of-the-art review. J. Socio. Econ. 32, 81–102. https://doi.org/10.1016/S1053-5357(03)00005-2

Ashoori, N., Dzombak, D.A., Small, M.J., 2016. Modeling the Effects of Conservation, Demographics, Price, and Climate on Urban Water Demand in Los Angeles, California. Water Resour. Manag. 30, 5247–5262. https://doi.org/10.1007/s11269-016-1483-7

Bai, Y., Wang, P., Li, C., Xie, J., Wang, Y., 2014. A multi-scale relevance vector regression approach for daily urban water demand forecasting. J. Hydrol. 517, 236–245. https://doi.org/10.1016/j.jhydrol.2014.05.033

Bakker, M., Van Duist, H., Van Schagen, K., Vreeburg, J., Rietveld, L., 2014. Improving the performance of water demand forecasting models by using weather input, in: Procedia Engineering. https://doi.org/10.1016/j.proeng.2014.02.012

Balacco, G., Carbonara, A., Gioia, A., Iacobellis, V., Piccinni, A.F., 2018. Investigation of Peak Water Consumption Variability at Local Scale in Puglia (Southern Italy). Proceedings 2, 674. https://doi.org/10.3390/proceedings2110674

Balling, R.C., Gober, P., Jones, N., 2008. Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona. Water Resour. Res. 44, 1–11. https://doi.org/10.1029/2007WR006722

Beal, C., Stewart, R., Giurco, D., Panuwatwanich, K., 2013. Intelligent Metering for Urban Water Planning and Management, in: Water Efficiency in Buildings: Theory and Practice. https://doi.org/10.1002/9781118456613.ch8

Beal, C.D., Stewart, R.A., 2014. Identifying residential water end uses underpinning peak day and peak hour demand. J. Water Resour. Plan. Manag. 140, 04014008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000357

Bettin, A., Rogers, D., 2012. Autoleak -Evaluating Leakage Economically And Environmentally, in: WaterLossEurope2012 IWA Specialist Conference. Ferrara.

Billings, R.B., Jones, C.V., 2008. Forecasting Urban Water Demand, Second. ed. American Water Works Association, Denver, CO.

Blokker, E.J.M., 2010. Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks. Water Management Academic Press, Delft. https://doi.org/10.1017/CBO9781107415324.004

Blokker, E.J.M., Vreeburg, J.H.G., Beverloo, H., Klein Arfman, M., Van Dijk, J.C., 2010. A bottom-up approach of stochastic demand allocation in water quality modelling. Drink. Water Eng. Sci. 3, 43–51. https://doi.org/10.5194/dwes-3-43-2010

Blokker, E.J.M., Vreeburg, J.H.G., Buchberger, S.G., van Dijk, J.C., 2008. Importance of demand modelling in network water quality models: a review. Drink. Water Eng. Sci. Discuss. 1, 27–38. https://doi.org/10.5194/dwesd-1-1-2008

Blokker, M., Agudelo-vera, C., Moerman, A., van Thienen, P., Pieterse-Quirijns, I., 2017. Review of applications of SIMDEUM, a stochastic drinking water demand model with small temporal and spatial scale. Drink. Water Eng. Sci. Discuss. 10, 1–15. https://doi.org/10.5194/dwes-2017-4

Bocchiola, D., 2013. Impact of potential climate change on crop yield and water footprint of rice in the Po valley of Italy. Agric. Syst. 116, 50–61. https://doi.org/10.1016/j.agsy.2015.07.009

Brekke, L., Larsen, M.D., Ausburn, M., Takaichi, L., 2002. Suburban water demand modeling using stepwise regression. Am. Water Work. Assoc. 94, 65–75. https://doi.org/10.1002/j.1551-8833.2002.tb09558.x

Brentan, B.M., Luvizotto, E., Herrera, M., Izquierdo, J., Pérez-García, R., 2017a. Hybrid regression model for near real-time urban water demand forecasting. J. Comput. Appl. Math. 309, 532–541. https://doi.org/10.1016/j.cam.2016.02.009

Brentan, B.M., Meirelles, G., Herrera, M., Luvizotto, E., Izquierdo, J., 2017b. Correlation Analysis of Water Demand and Predictive Variables for Short-Term Forecasting Models. Math. Probl. Eng. 2017, 1–10. https://doi.org/10.1155/2017/6343625

Brentan, B.M., Meirelles, G.L., Manzi, D., Luvizotto, E., 2018. Water demand time series generation for distribution network modeling and water demand forecasting. Urban Water J. 15, 150–158. https://doi.org/10.1080/1573062X.2018.1424211

Britton, T.C., Stewart, R.A., O'Halloran, K.R., 2013. Smart metering: Enabler for rapid and effective post meter leakage identification and water loss management. J. Clean. Prod. 54, 166–176. https://doi.org/10.1016/j.jclepro.2013.05.018

Buchberger, S.G., Carter, J., Lee, Y., Shade, T.., 2003. Random demands, travel times, and water quality in deadends, subject area: high-quality water. AWWA Research Foundation, Denver, CO.

Buchberger, S.G., Wu, L., 1995. Model for instantaneous residential water demands. J. Hydraul. Eng. 121, 232–246. https://doi.org/10.1061/(ASCE)0733-9429(1995)121:3(232)

Burney, N., Mukhopadhyay, A., Al-Mussallam, N., Akber, A., Al-Awadi, E., 2001. Forecasting of freshwater demand in Kuwait. Arab. J. Sci. Eng. 26, 99–113.

Chang, H., Parandvash, G.H., Shandas, V., 2010. Spatial variations of single-family residential water consumption in Portland, Oregon. Urban Geogr. 31, 953–972. https://doi.org/10.2747/0272-3638.31.7.953

Chang, H., Praskievicz, S., Parandvash, H., 2014a. Sensitivity of Urban Water Consumption to Weather and Climate Variability at Multiple Temporal Scales: The Case of Portland, Oregon. Int. J. Geospatial Environ. Res. 1, 1–7.

Chang, H., Praskievicz, S., Parandvash, H., 2014b. Sensitivity of Urban Water Consumption to Weather and Climate Variability at Multiple Temporal Scales: The Case of Portland, Oregon. Int. J. Geospatial Environ. Res. 1, 7.

Chen, J., Boccelli, D.L., 2018. Forecasting Hourly Water Demands With Seasonal Autoregressive Models for Real-Time Application. Water Resour. Res. 54, 879– 894. https://doi.org/10.1002/2017WR022007

Cole, G., Stewart, R.A., 2013. Smart meter enabled disaggregation of urban peak water demand: Precursor to effective urban water planning. Urban Water J. 10, 174–194. https://doi.org/10.1080/1573062X.2012.716446

Colombo, A.F., Karney, B.W., 2003. The labyrinth of water distribution systems: Demand, energy and climate change, in: Pumps, Electromechanical Devices and Systems Applied to Urban Water Management. Taylor & Francis, Leiden.

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., Rizzoli, A.E., 2015. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. Environ. Model. Softw. 72, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012

Cominola, A., Nguyen, K., Giuliani, M., Stewart, R.A., Maier, H.R., Castelletti, A., 2019. Data Mining to Uncover Heterogeneous Water Use Behaviors From Smart Meter Data. Water Resour. Res. 55, 9315–9333. https://doi.org/10.1029/2019WR024897

Cordell, D.J., Robinson, J.E., Loh, M.T.Y., 2003. Collecting residential end use data from primary sources: do's and dont's, in: Efficient 2003: Efficient Use and Management of Water for Urban Supply Conference. Tenerife, pp. 1–12.

Creaco, E., Blokker, M., Buchberger, S., 2017a. Models for generating household water demand pulses: Literature review and comparison. J. Water Resour. Plan. Manag. 143, 04017013. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000763

Creaco, E., Pezzinga, G., Savic, D., 2017b. On the choice of the demand and hydraulic modeling approach to WDN real-time simulation. Water Resour. Res. 53, 6159–6177. https://doi.org/10.1002/2016WR020104

De Loë, R., Kreutzwiser, R., Moraru, L., 2001. Adaptation options for the near term: Climate change and the Canadian water sector. Glob. Environ. Chang. 11, 231–245. https://doi.org/10.1016/S0959-3780(00)00053-4

Deidda, R., Badas, M.G., Piga, E., 2004. Space-time scaling in high-intensity Tropical Ocean Global Atmosphere Coupled Ocean-Atmosphere Response Experiment (TOGA-COARE) storms. Water Resour. Res. 40, W02506. https://doi.org/10.1029/2003WR002574

Domene, E., Saurí, D., 2006. Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona. Urban Stud. 43, 1605–1623. https://doi.org/10.1080/00420980600749969

Donkor, E.A., Mazzuchi, T.A., Soyer, R., Alan Roberson, J., 2014. Urban water demand forecasting: Review of methods and models. J. Water Resour. Plan. Manag. 140, 146–159. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000314

Downing, T.E., Butterfield, R.E., Edmonds, B., Knox, J.W., Moss, S., Piper, B.S., Weatherhead, E.K., 2003. Climate Change and Demand for Water, Research Report, Stockholm Environment Institute. Oxford (UK).

Filion, Y., Adams, B., Karney, B., 2007. Cross correlation of demands in water distribution network design. J. Water Resour. Plan. Manag. 133, 137–144. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:2(137)

Filion, Y.R., Karney, B.W., Adams, B.J., 2005. Stochasticity of demand and probabilistic performance of water networks, in: Impacts of Global Climate Change: Proceedings of World Water and Environmental Resources Congress 2005. ASCEE, Anchorage, Alaska, pp. 1–12. https://doi.org/10.1061/40792(173)49

Fontanazza, C.M., Notaro, V., Puleo, V., Freni, G., 2014. Multivariate statistical analysis for water demand modeling. Procedia Eng. 89, 901–908. https://doi.org/10.1016/j.proeng.2014.11.523

Fox, C., McIntosh, B.S., Jeffrey, P., 2009. Classifying households for water demand forecasting using physical property characteristics. Land use policy 26, 558–568. https://doi.org/10.1016/j.landusepol.2008.08.004

Galdiero, E., De Paola, F., Fontana, N., Giugni, M., Savic, D., 2016. Decision support system for the optimal design of district metered areas. J. Hydroinformatics 18, 49–61. https://doi.org/10.2166/hydro.2015.023

Gargano, R., Di Palma, F., de Marinis, G., Granata, F., Greco, R., 2016. A stochastic approach for the water demand of residential end users. Urban Water J. 13, 569–582. https://doi.org/10.1080/1573062X.2015.1011666

Gato-Trinidad, S., Gan, K., 2011. Characterizing maximum residential water demand. WIT Trans. Built Environ. 122, 15–24. https://doi.org/10.2495/UW120021

Gato, S., Jayasuriya, N., Roberts, P., 2007. Temperature and rainfall thresholds for base use urban water demand modelling. J. Hydrol. 337, 364–376. https://doi.org/10.1016/j.jhydrol.2007.02.014

Goodchild, C.W., 2003. Modelling the impact of climate change on domestic water demand. Water Environ. J. 17, 8–12. https://doi.org/10.1111/j.1747-6593.2003.tb00423.x

Haque, M.M., de Souza, A., Rahman, A., 2017. Water Demand Modelling Using Independent Component Regression Technique. Water Resour. Manag. 31, 299–312. https://doi.org/10.1007/s11269-016-1525-1

Haque, M.M., Hagare, D., Rahman, A., Kibria, G., 2014. Quantification of Water Savings due to Drought Restrictions in Water Demand Forecasting Models. J. Water Resour. Plan. Manag. 140, 04014035. https://doi.org/10.1061/(asce)wr.1943-5452.0000423

House-Peters, L.A., Chang, H., 2011. Urban water demand modeling: Review of concepts, methods, and organizing principles. Water Resour. Res. 47, W05401. https://doi.org/10.1029/2010WR009624

House-Peters, L.A., Pratt, B., Chang, H., 2010. Effects of Urban spatial structure, sociodemographics, and climate on residential water consumption in Hillsboro, Oregon. J. Am. Water Resour. Assoc. 46, 461–472. https://doi.org/10.1111/j.1752-1688.2009.00415.x

Hunn, N., 2010. Essentials of Short-Range Wireless, The Cambridge Wireless Essentials Series. Cambridge University Press. https://doi.org/10.1017/cbo9780511776991

Hussien, W.A., Memon, F.A., Savic, D.A., 2016. Assessing and Modelling the Influence of Household Characteristics on Per Capita Water Consumption. Water Resour. Manag. 30, 2931–2955. https://doi.org/10.1007/s11269-016-1314-x

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science (80-.). 349, 255–260. https://doi.org/10.1126/science.aaa8415

Kartakis, S., Abraham, E., McCann, J.A., 2015. WaterBox: A testbed for monitoring and controlling smart water networks, in: CySWater'15: Proceedings of the 1st ACM International Workshop on Cyber-Physical Systems for Smart Water Networks. Seattle, WA, USA, pp. 1–6. https://doi.org/10.1145/2738935.2738939

Kim, Y., Schmid, T., Charbiwala, Z.M., Friedman, J., Srivastava, M.B., 2008. NAWMS: Nonintrusive autonomous water monitoring system, in: SenSys'08 - Proceedings of the 6th ACM Conference on Embedded Networked Sensor Systems. https://doi.org/10.1145/1460412.1460443

Kofinas, D., Mellios, N., Papageorgiou, E., Laspidou, C., 2014. Urban water demand forecasting for the island of skiathos. Procedia Eng. 89, 1023–1030. https://doi.org/10.1016/j.proeng.2014.11.220

Koutsoyiannis, D., Manetas, A., 1996. Simple disaggregation by accurate adjusting procedures. Water Resour. Res. 32, 2105–2117. https://doi.org/10.1029/96WR00488

Kumar, D.N., Lall, U., Petersen, M.R., 2000. Multisite disaggregation of monthly to daily streamflow. Water Resour. Res. 36, 1823–1833. https://doi.org/10.1029/2000WR900049

Lee, S.J., Wentz, E.A., 2008. Applying Bayesian Maximum Entropy to extrapolating local-scale water consumption in Maricopa County, Arizona. Water Resour. Res. 44, W01401. https://doi.org/10.1029/2007WR006101

Lee, T., Salas, J.D., Prairie, J., 2010. An enhanced nonparametric streamflow disaggregation model with genetic algorithm. Water Resour. Res. 46, W08545. https://doi.org/10.1029/2009WR007761

Makropoulos, C., Kossieris, P., Kozanis, S., Katsiri, E., Vamvakeridou-Lyroudia, L., 2014. From smart meters to smart decisions: web-based support for the water efficient household, in: 11th International Conference on Hydroinformatics. CUNY Academic Works, New York, pp. 1–8.

Mamade, A., Loureiro, D., Alegre, H., Covas, D., 2018. Top-down and bottom-up approaches for water-energy balance in Portuguese supply systems. Water 10, 577. https://doi.org/10.3390/w10050577

Mamade, A., Loureiro, D., Covas, D., Coelho, S.T., Amado, C., 2014. Spatial and temporal forecasting of water consumption at the DMA level using extensive measurements. Procedia Eng. 70, 1063–1073. https://doi.org/10.1016/j.proeng.2014.02.118

Manouseli, D., Kayaga, S.M., Kalawsky, R., 2019. Evaluating the Effectiveness of Residential Water Efficiency Initiatives in England: Influencing Factors and Policy Implications. Water Resour. Manag. 33, 2219–2238. https://doi.org/10.1007/s11269-018-2176-1

Martinez-Espiñeira, R., 2002. Residential water demand in the Northwest of Spain. Environ. Resour. Econ. 21, 161–187. https://doi.org/10.1023/A:1014547616408

Martyusheva, O., 2014. Smart Water Grid. Plan B Technical Report. Fort Collins, Colorado.

Masia, S., Sušnik, J., Marras, S., Mereu, S., Spano, D., Trabucco, A., 2018. Assessment of irrigated agriculture vulnerability under climate change in Southern Italy. Water (Switzerland) 10, 1–19. https://doi.org/10.3390/w10020209

Mayer, P.W., Deoreo, W.B., Opitz, E.M., Kiefer, J.C., Davis, W.Y., Dziegielewski, B., Nelson, J.O., 1999. Residential End Uses of Water, Aquacraft, Inc. Water Engineering and Management. Denver, Colorado.

Mejia, J.M., Rousselle, J., 1976. Disaggregation models in hydrology revisited. Water Resour. Res. 12, 185– 186. https://doi.org/10.1029/WR012i002p00185

Merrett, S., 2004. The Demand for Water: Four Interpretations. Water Int. 29, 27–29. https://doi.org/10.1080/02508060408691745

Miaou, S. -P, 1990. A class of time series urban water demand models with nonlinear climatic effects. Water Resour. Res. 26, 169–178. https://doi.org/10.1029/WR026i002p00169

Moughton, L.J., Buchberger, S.G., Boccelli, D.L., Filion, Y.R., Karney, B.W., 2007. Effect of time step and data aggregation on cross correlation of residential demands, in: 8th Annual Water Distribution Systems Analysis Symposium 2006. ASCEE, Cincinnati, Ohio (USA), pp. 1–11. https://doi.org/10.1061/40941(247)42

Mylopoulos, N., Fafoutis, C., Sfyris, S., Alamanos, A., 2017. Impact of water pricing policy and climate change on future water demand in Volos, Greece. Eur. Water 58, 473–479.

Nowak, K., Prairie, J., Rajagopalan, B., Lall, U., 2010. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008530

Parandvash, G.H., Chang, H., 2016. Analysis of long-term climate change on per capita water demand in urban versus suburban areas in the Portland metropolitan area, USA. J. Hydrol. 538, 574–586. https://doi.org/10.1016/j.jhydrol.2016.04.035

Parker, J., 2013. Assessing the sensitivity of historic micro-component household water-use to climatic drivers. Loughborough University.

Peres, D.J., Modica, R., Cancelliere, A., 2019. Assessing future impacts of climate change on water supply system performance: Application to the Pozzillo Reservoir in Sicily, Italy. Water (Switzerland) 11. https://doi.org/10.3390/w11122531

Pericli, A., Jenkins, J.O., 2015. Smart Meters and Domestic Water usage. Foundation for Water Research, Marlow.

Polebitski, A.S., Palmer, R.N., 2010. Seasonal residential water demand forecasting for census tracts. J. Water Resour. Plan. Manag. 136, 27–36. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000003

Polebitski, A.S., Palmer, R.N., Waddell, P., 2011. Evaluating water demands under climate change and transitions in the urban environment. J. Water Resour. Plan. Manag. 137, 249–257. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000112

Pullinger, M., Browne, A., Anderson, B., Medd, W., 2013. Patterns of water: The water related practices of households in southern England, and their influence on water consumption and demand management. Lancaster UK. https://doi.org/10.1007/978-3-319-77748-1_3

Quevedo, J., Saludes, J., Puig, V., Blanch, J., 2014. Short-term demand forecasting for real-time operational control of the Barcelona water transport network, in: 22nd Mediterranean Conference on Control and Automation. IEEE, Palermo, pp. 990–995. https://doi.org/10.1109/MED.2014.6961503

Rinaudo, J.-D., 2015. Long-Term Water Demand Forecasting, in: Grafton, Q., Daniell, K., Nauges, C., Rinaudo, J.-D., Chan, N.W.W. (Eds.), Understanding and Managing Urban Water in Transition. Springer, Dordrecht, pp. 239–268. https://doi.org/10.1007/978-94-017-9801-3

Romano, G., Salvati, N., Guerrini, A., 2014. Estimating the determinants of residential water demand in Italy. Forests 5, 2929–2945. https://doi.org/10.3390/w6102929

Romano, M., Kapelan, Z., 2014. Adaptive water demand forecasting for near real-time management of smart water distribution systems. Environ. Model. Softw. 60, 265–276. https://doi.org/10.1016/j.envsoft.2014.06.016

Santos, E.G., Salas, J.D., 1992. Stepwise disaggregation scheme for synthetic hydrology. J. Hydraul. Eng. 118, 765–784. https://doi.org/10.1061/(ASCE)0733-9429(1992)118:5(765)

Savić, D., Vamvakeridou-Lyroudia, L., Kapelan, Z., 2014. Smart meters, smart water, smart societies: The iWIDGET project. Procedia Eng. 89, 1105–1112. https://doi.org/10.1016/j.proeng.2014.11.231

Schleich, J., Hillenbrand, T., 2009. Determinants of residential water demand in Germany. Ecol. Econ. 68, 1756–1769. https://doi.org/10.1016/j.ecolecon.2008.11.012

See Wong, J., Zhang, Q., Chen, Y.D., 2010. Statistical modeling of daily urban water consumption in Hong Kong: Trend, changing patterns, and forecast. Water Resour. Res. 46, W03506. https://doi.org/10.1029/2009WR008147

Shandas, V., Parandvash, G.H., 2010. Integrating urban form and demographics in water-demand management: An empirical case study of Portland, Oregon. Environ. Plan. B Plan. Des. 37, 112–128. https://doi.org/10.1068/b35036

Sharma, D., Kumar, N., 2017. A review on machine learning algorithms. Int. J. Adv. Reasearch Comput. Eng. Technol. 6, 495–507. https://doi.org/10.1007/978-981-15-6202-0_51

Sheng, Z., Jia, S., Michelsen, A., Abudu, S., 2017. Lessons learned from long term water resources plans : top town vs. bottom up, in: International Water Resources Association (IWRA), XVI World Water Congress “Bridging Science and Policy.” Cancun, Quintana Roo, Mexico, pp. 1–8.

Slavíková, L., Malý, V., Rost, M., Petružela, L., Vojáček, O., 2013. Impacts of Climate Variables on Residential Water Consumption in the Czech Republic. Water Resour. Manag. 27, 365–379. https://doi.org/10.1007/s11269-012-0191-1

Sophocleous, S., Savić, D.A., Kapelan, Z., 2018. Leak Detection and Localization Based on Search Space Reduction and Hydraulic Modelling, in: 1st International WDSA/CCWI Joint Conference. WDSA/CCWI, Kingston, Ontario, Canada.

Statzu, V., Strazzera, E., 2009. Water Demand for Residential Uses in a Mediterranean Region : Econometric Analysis and Policy Implications 1–31.

Stavenhagen, M., Buurman, J., Tortajada, C., 2018. Saving water in cities: Assessing policies for residential water demand management in four cities in Europe. Cities 79, 187–195. https://doi.org/10.1016/j.cities.2018.03.008

Tarboton, D.G., Sharma, A., Lall, U., 1998. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. Water Resour. Res. https://doi.org/10.1029/97WR02429

Todini, E., 1980. The preservation of skewness in linear disaggregation schemes. J. Hydrol. https://doi.org/10.1016/0022-1694(80)90093-1

Toth, E., Bragalli, C., Neri, M., 2018. Assessing the significance of tourism and climate on residential water demand: Panel-data analysis and non-linear modelling of monthly water consumptions. Environ. Model. Softw. 103, 52–61. https://doi.org/10.1016/j.envsoft.2018.01.011

Vašak, M., Banjac, G., Baotić, M., Matuško, J., 2014. Dynamic day-ahead water pricing based on smart metering and demand prediction. Procedia Eng. 89, 1031–1036. https://doi.org/10.1016/j.proeng.2014.11.221

Villar-Navascués, R.A., Pérez-Morales, A., 2018. Factors Affecting Domestic Water Consumption on the Spanish Mediterranean Coastline. Prof. Geogr. 70, 513–525. https://doi.org/10.1080/00330124.2017.1416302

Walski, T.M., Chase, D. V., Savic, D.A., Grayman, W., Koelle, E., 2003. Advanced Water Distribution Modeling and Management, First. ed, Methods. Haestad Methods, Waterbury, CT (USA).

Wang, X. jun, Zhang, J. yun, Shahid, S., Guan, E. hong, Wu, Y. xiang, Gao, J., He, R. min, 2014. Adaptation to climate change impacts on water demand. Mitig. Adapt. Strateg. Glob. Chang. 21, 81–99. https://doi.org/10.1007/s11027-014-9571-6

Wang, X., Sun, Y., Song, L., Mei, C., 2009. An eco-environmental water demand based model for optimising water resources using hybrid genetic simulated annealing algorithms. Part I. Model development. J. Environ. Manage. https://doi.org/10.1016/j.jenvman.2009.02.008

Wentz, E.A., Gober, P., 2007. Determinants of small-area water consumption for the City of Phoenix, Arizona. Water Resour. Manag. 21, 1849–1863. https://doi.org/10.1007/s11269-006-9133-0

Willis, R.M., Stewart, R.A., Giurco, D.P., Talebpour, M.R., Mousavinejad, A., 2013. End use water consumption in households: Impact of socio-demographic factors and efficient devices. J. Clean. Prod. https://doi.org/10.1016/j.jclepro.2011.08.006

Xenochristou, M., 2019. Water Demand Forecasting Using Machine Learning on Weather and Smart Metering Data. University of Exeter.

Xenochristou, M., Kapelan, Z., 2020. An ensemble stacked model with bias correction for improved water demand forecasting. Urban Water J. 17, 212–223. https://doi.org/10.1080/1573062X.2020.1758164

Xenochristou, M., Kapelan, Z., Hutton, C., 2020. Using Smart Demand-Metering Data and Customer Characteristics to Investigate Influence of Weather on Water Consumption in the UK. J. Water Resour. Plan. Manag. 146, 04019073. https://doi.org/10.1061/(asce)wr.1943-5452.0001148

Xenochristou, M., Kapelan, Z., Hutton, C., Hofman, J., 2018. Smart Water Demand Forecasting: Learning from the Data 3, 2351–2342. https://doi.org/10.29007/wkp4

Zetland, D., 2011. The End of Abundance: Economic Solutions to Water Scarcity. Aguanomics Press.

Zhou, S.L., McMahon, T.A., Walton, A., Lewis, J., 2000. Forecasting daily urban water demand: A case study of Melbourne. J. Hydrol. 236, 153–164. https://doi.org/10.1016/S0022-1694(00)00287-0

2 Bottom-up generation of water demand time series

Numerical simulation of WDNs behaviour enables to improve water system management. In order to obtain reliable simulations, accurate estimates of nodal demands are needed (Creaco et al., 2017c). In late 1990s the advent of smart meters improved the understanding of water demand pattern (Fox et al., 2009; Schleich and Hillenbrand, 2009). The use of smart meters has provided water consumption data at very high temporal (minutes or even seconds) and spatial resolution. Metering data allowed to model demand at household and micro-component level in order to maintain the heterogeneity of individual water uses (Cominola et al., 2015). Different approaches to generate nodal demand patterns at very small time steps (Blokker et al., 2017; Buchberger et al., 2003; Buchberger and Wu, 1995; Creaco et al., 2017a) as well as larger time steps, e.g. from minutes to hours (Alvisi et al., 2014; Gargano et al., 2016), have been developed. Furthermore, different studies (Blokker et al., 2008; Filion et al., 2007; Moughton et al., 2007) have shown the importance of accurately reproducing not only the main statistics of the measured time series but also their spatial and temporal correlations.

The present chapter proposes a novel bottom-up methodology for the generation of WDN demand time series. The methodology is based on two phases. First, it generates, for each user and for each time step, demand time series of the first attempt, by using a Beta probability distribution with tunable bounds or a Gamma distribution with shift parameter. The synthetic demand time series generated in the first phase are consistent with the measured time series in terms of mean, standard deviation and skewness. In the second phase, the existing cross-correlations between users and at all temporal lags are imposed on the generated demand time series through a single copula-based re-sort. Finally, the generated demand time series reproduce both the main statistics and the spatial and temporal correlations of the measured time series. The effectiveness of the methodology was proven in two case studies, the literature case study of Milford (Ohio) and the Soccavo DMA (section 1.1.2).

Compared to existing methodologies (Alvisi et al., 2014; Gargano et al., 2016), the novel methodology works on a single time step, which can be set according to modelling preferences. Moreover, compared to the methodology of Alvisi et al. (2014), based on two phases of data-reordering, a single data-reordering is performed to impose spatial and temporal correlations between users at all temporal lags. A further improvement of the methodology is that it enables to preserve the skewness of the measured demand time series at both user and aggregated scales.

This chapter was written by the author on the basis of publications I and II.

2.1 Methodology

The proposed methodology allows to generate demand time series for a fixed number of days (n_{days}), for each of the N_{user} users considered and for each of the $N_{\Delta t}$ time step Δt into which the day is divided. Figure 5 shows the flowchart of the methodology.

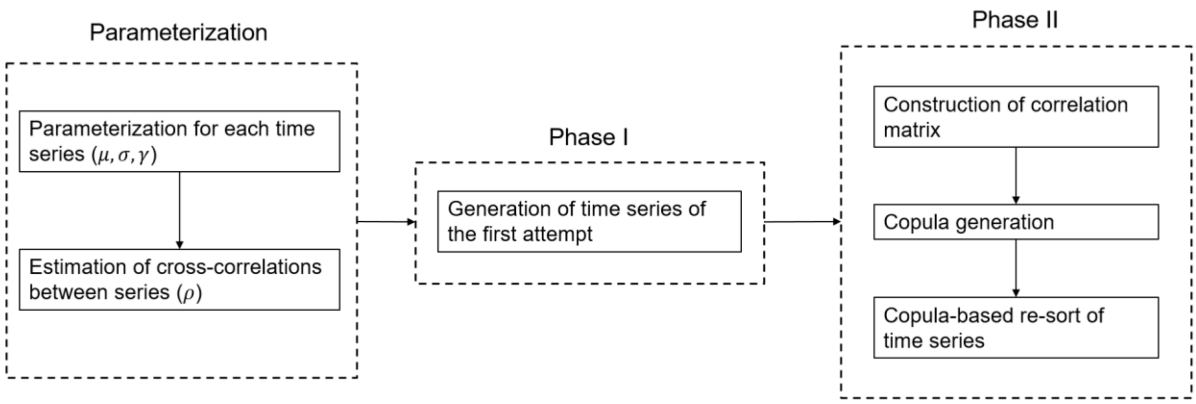

Figure 5 - Flow-chart of the methodology (modified from Creaco et al., 2020).

For each user and for each Δt, the parameterization requires the assessment of mean (μ), standard deviation (σ) and skewness (γ) of the measured demand time series. Thus, $3 \times N_{\Delta t} \times N_{user}$ parameters are required. As better explained in the following, in one of the two variants of the methodology the minimum values (a) of the measured demand time series is also needed (other $N_{\Delta t} \times N_{user}$ parameters should be assessed). Furthermore, rank cross-correlations between all users at all time lags are estimated based on the measured demand time series, to be used in the second phase of the methodology. Given that each time series is fully correlated with itself and the correlation between two time series is symmetrical, other $N_{\Delta t} \times N_{user} \times (N_{\Delta t} \times N_{user} - 1)/2$ must be assessed. The parameterization is followed by the first phase during which demand time series of the first attempt are generated. Finally, in the second phase the existing cross-correlation are imposed by using a copula (Nelsen, 1999).

It should be noted that the methodology can also be applied directly to demand generation at WDN nodes. This can be more useful than generating demands at single user scale when the overall behaviour of a WDN is of interest. In this case the measured demand time series of the generic node can be obtained summing the demand time series of the related users. Then, the measured nodal demand time series can be used for the parameterization of the methodology.

In this work, the generation of synthetic demand time series was carried out by subdividing the day into $N_{\Delta t} = 24$ time steps with Δt = 1 hr. For the sake of simplicity, all days from Monday to Sunday were considered identical. However, the methodology allows to account for the different days of the week, such as working days and weekends. To do this, the different days should be threatened separately starting from the parameterization phase. In order to attain representative results, the generation was performed for $n_{days} = 93$ days (i.e. a number of days three times greater than the measured time series, described in section 2.2) and was reiterated for 500 times. Finally, the average values over the 500 iterations were considered. The following sections give an insight about copulas and describe the two phases.

2.1.1 Copula functions

The study of the relationship between two or more random variables is an important statistical issue. Understanding the distribution of several random variables interacting together is a common need to different scientific contexts. Let us suppose that there are two important factors described by two random variables. Knowing marginal distributions of the individual variables and the linear correlation coefficient could be not enough to describe their joint behaviour. Marginal distributions and the linear correlation coefficient do not necessarily determine the joint distribution (Abozou, 2007). In this case, copula could be a convenient way to express the joint distribution, allowing to separate the joint distribution into the contribution from the marginal distributions and that from the interdependency of the probabilities.

The term copula comes from Latin and means “a link, tie, bond” (Nelsen, 1999). Copula can be defined as a function that "couple" multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copula can be defined as a multivariate distribution function whose one-dimensional margins are uniform on the interval $[0,1]$. To better understand copulas, Nelsen (1999) provided an effective and simple explanation. Consider two random variables X and Y, with distribution functions $F(x) = P[X \leq x]$ and $G(y) = P[Y \leq y]$, respectively. Let $H(x, y) = P[X \leq x, Y \leq y]$ be the joint distribution function. Note that for each pair of real numbers (x, y), the corresponding values of $F(x)$, $G(y)$ and $H(x, y)$ lie in the interval $[0,1]$. Each pair (x, y) can be associated to a point $(F(x), G(y))$ in the unit square $[0,1] \times [0,1]$, that in turn corresponds to a number $H(x, y)$ in $[0,1]$. This correspondence, which assigns the value of the joint distribution function to each ordered pair of values of the individual distribution functions, is copula. This can be easily extended to the multivariate case, thus when $\mathbb{R}^n$ random vector $\bar{X} = (X_1, \dots, X_n)$ with respective marginal distribution functions F_i is considered.

In this methodology a normal copula is used. The Gaussian or Normal copula is widely used in applications. It is derived from a multivariate Normal distribution function with mean zero by transforming the marginal distributions by the inverse of the standard normal distribution function (Mikosch, 2006).

Even if there may be statistical problems in handling copulas (Abozou, 2007), they are characterized by a flexible structure that can be applied in different conditions. Furthermore, they offer an enormous improvement in capturing the real correlation pattern. Overall, copulas can be used to simulate multivariate outcomes, providing an important tool when many variables need to be considered.

2.1.2 Generation of demand time series

The first phase is based on the assumption that the demand q_i^j of the $j-th$ user in the $i-th$ Δt follows a Beta probability distribution with tunable bounds (first variant of the methodology) or a Gamma distribution with shift parameter (second variant). In case of Beta probability distribution, the density function f of the random variable x is the following (Johnson et al., 1995):

$$f(x) = \frac{(x-a)^{(\alpha-1)}(b-x)^{(\beta-1)}}{B(\alpha,\beta)(b-a)^{(\alpha+\beta-1)}} \tag{1}$$

where $\alpha > 0$ and $\beta > 0$ are the shape parameters, B is the beta function, $a \geq 0$ and $b > a$ are the lower and upper bounds, respectively.

For the parameterization, the following equations can be used:

$$\alpha = \bar{\mu}\left[\frac{\bar{\mu}(1-\bar{\mu})}{\bar{\sigma}^2} - 1\right] \tag{2}$$

$$\beta = (1-\bar{\mu})\left[\frac{\bar{\mu}(1-\bar{\mu})}{\bar{\sigma}^2} - 1\right] \tag{3}$$

$$\bar{\mu} = \frac{(\mu - a)}{(b-a)} \tag{4}$$

$$\bar{\sigma} = \frac{\sigma}{(b-a)} \tag{5}$$

$$\gamma = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}} \tag{6}$$

The parameters α and β can be derived by using in equations 4 and 5 the mean and standard deviation of the measured time series for user and Δt under consideration, obtained through the method of the moments. The values of a and b can be determined to preserve the skewness of the measured time series. In this work, b was determined by minimizing the absolute value of the difference between the skewness expressed in equation 6 and the measured one, while a was set at the minimum value in the measured time series following preliminary analysis.

When a Gamma distribution with shift parameter is assumed, the density function f of the random variable x is (Johnson et al., 1995):

$$f(x) = \frac{(x - x_0)^{(k-1)} e^{-(x-x_0)/\vartheta}}{\vartheta^k \Gamma(k)} \tag{7}$$

where x_0 is the shift parameter, k and ϑ are the shape parameters, Γ is the gamma function.
The following equations can be used to determine the parameters of the distribution:

$$k = \frac{4}{\gamma^2} \tag{8}$$

$$\vartheta = \frac{\sigma\gamma}{2} \tag{9}$$

$$x_0 = \mu - \frac{2\sigma}{\gamma} \tag{10}$$

The gamma distribution can be parameterized by using the values of μ, σ and γ of the measured demand time series (method of the moments) in equations 8, 9 and 10. This parameterization may result in values of x_0 lower than 0. This condition should be avoided to prevent generation of negative values of water demand. When x_0 is lower than 0, it can be set at 0. In this case the other parameters can be determined by using the relationships of the gamma probability distribution with two parameters:

$$k = \frac{\mu^2}{\sigma^2} \tag{11}$$

$$\vartheta = \frac{\sigma^2}{\mu} \tag{12}$$

It follows from the above that the Gamma probability distribution parameterization is easier compared to the Beta probability distribution with tunable bounds, being fully based on analytical relationships. However, the Gamma probability distribution with two parameters failed to preserve skewness, as shown in the following. Therefore, the Beta probability distribution with tunable bounds should be used when numerous measured demand time series lead to negative values of x_0.

2.1.3 Imposition of spatial and temporal cross-correlation

Even though the demand time series generated in the first phase respect the basic statistics (mean, variance, and skewness), they fail to preserve the existing rank cross-correlations between users and at various temporal lags. Therefore, in this phase the generated demand time series are re-sorted through a copula to impose existing rank cross-correlations. These latter can be derived from the measured time series through the Spearman correlation index (Spearman, 1904):

$$\rho = \frac{\sum_{i=1}^{N_m}(r_{1,m} - \bar{r}_1)(r_{2,m} - \bar{r}_2)}{\sqrt{\sum_{i=1}^{N_m}(r_{1,m} - \bar{r}_1)^2}\sqrt{\sum_{i=1}^{N_m}(r_{2,m} - \bar{r}_2)^2}} \quad (13)$$

where $r_{1,m}$ and $r_{2,m}$ are the ranks of the $m-th$ element of the generic measured time series q_i^j and q_k^l respectively, $\bar{r}_1$ and $\bar{r}_2$ are the corresponding average ranks, N_m is the number of elements of both the measured time series.

The Spearman correlation index is derived from the Pearson index (Pearson, 1895) by converting the values into ranks before calculating the index. Compared to the Pearson index, the Spearman index is easier to impose on statistical series generated through any probability distributions.

Then, in order to impose the calculated correlations a multivariate normal probability distribution, with means and standard deviations equal to 0 and 1 respectively, is used as copula (Nelsen, 1999). The elements of its covariance matrix C are the measured cross-correlations. C is a square symmetric matrix ($N_{user} \times N_{\Delta t}$) with diagonal elements equal to 1. In fact, each time series is fully correlated with itself ($\rho = 1$) and the correlation between two time series is symmetrical. Then, n_{days} long time series of $N_{user} \times N_{\Delta t}$ random variables with $\mu = 0$ and $\sigma = 1$ are generated by using the multivariate normal probability distribution. These time series express the rank cross-correlations to be imposed on demand time series. Finally, in order to impose these rank cross-correlations, the copula-generated time series are used to re-sort according to their order the demand time series generated in the first phase.

2.2 Data

The novel methodology was applied to two case studies. The first case study (*Case study 1*) consisted of metering data from 19 households in Milford (Ohio) recorded from 11 May to 10 June 1997 (31 days). These data came from the experimental campaign carried out by (Buchberger et al., 2003) in 1997 by recording consumptions every second in 21 households. In order to use a more feasible scale for data acquisition through smart meters, in this work the demands were aggregated at hourly scale, i.e. for each hour all the recorded values were summed up. The second case study (*Case study 2*) is made up of the consumption data of 100 users from the Soccavo DMA (see Chapter 1). In this work the data for 31 days, from 1 January 2018 to 31 January 2018, were used. It is worth noting that in both case studies only one month was considered since the parameters of the methodology are characterized by monthly variations.

Figure 6 shows the daily patterns of aggregated measured hourly demand, i.e. the demand time series obtained by summing up the demand series of the single users, in the 31 days in both the first and second case studies. The measured demand is, on average, larger in the second case study due to the larger number of users considered. The shape of the patterns is different as well. Compared to the second case study, the morning, midday and evening peaks take place earlier in the first case study. Furthermore, in the first case study the peaks have a similar size while the aggregated demand of the second case study is characterized by different peaks sizes. For the second case study the morning peak is the largest, followed by the midday and evening peaks. These differences in the shape of the patterns are due the different habits of the users in the two sites. The differences between the two case studies allowed to test the effectiveness of the bottom-up model for different consumption behaviours and thus different water demand patterns.

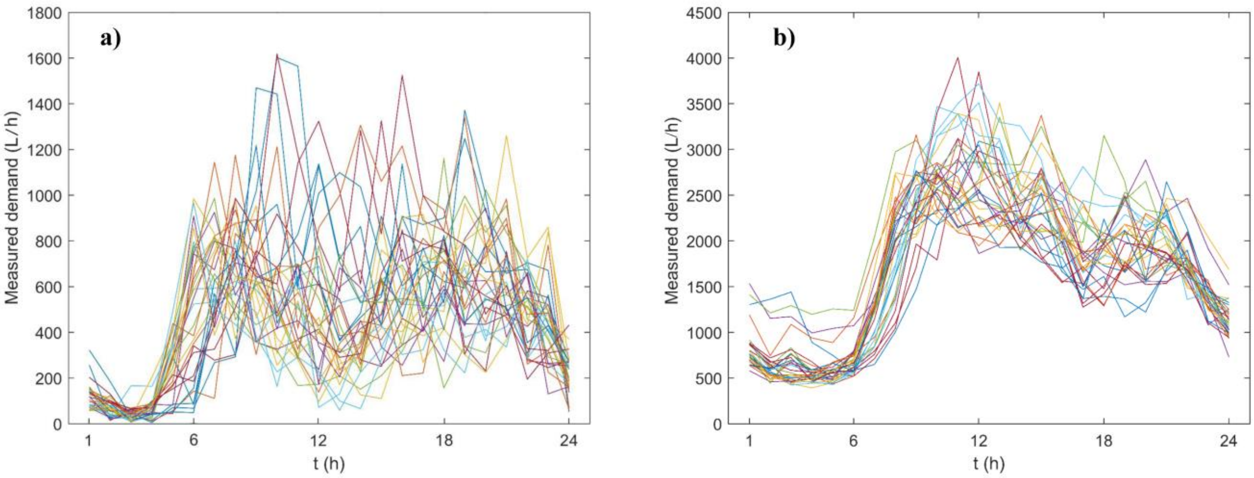

Figure 6 - Patterns of aggregated measured hourly demand in 31 days (each colour refers to one day) for the first (a) and the second (b) case study (Creaco et al., 2020).

2.3 Results and discussion

The following sections show and discuss the comparison between the basic statistics of measured and generated demand time series. The average values over 500 iterations were considered. First, the results obtained for the first case study by applying both Beta and Gamma probability distribution are discussed. In order to investigate the benefits of preserving rank cross-correlations, the results of the Beta probability distribution obtained by neglecting rank cross-correlations are also analysed. Then, the results for the second case study for both the distributions are discussed. The results of the generation of nodal demands by using the Beta probability distribution with tunable bounds are also presented. To this end, the users of the second case study were grouped in 20 nodes (close to the number of Milford households), by allocating five users to each node.

2.3.1 Results – Case study 1

Figure 7 shows the comparison in terms of mean (μ), standard deviation (σ), skewness (γ) and rank cross-correlations (ρ) between measured and generated hourly demands for the first case study. More specifically Figure 7a and Figure 7b report the comparison in terms of mean and standard deviation values, respectively. In both cases the dots of the graphs closely follow the graph bisectors, highlighting a high level of accuracy in reproducing both mean and standard deviation values. Figure 7c shows the fit of skewness values of measured demands against those of the generated demands. Also in this case the performance is good. However, the Beta distribution with tunable bounds was not flexible enough to reproduce all the measured values, underestimating the skewness values larger than 2. Finally, Figure 7d reports the comparison of spatial and temporal correlations of measured and generated hourly demands. The cases with $\rho = 1$ (i.e. rank cross-correlation between the generic demand time series and itself) were neglected and were not reported in Figure 7d. The fit is again excellent, demonstrating the effectiveness of the copula-based re-sort in the second phase of the methodology.

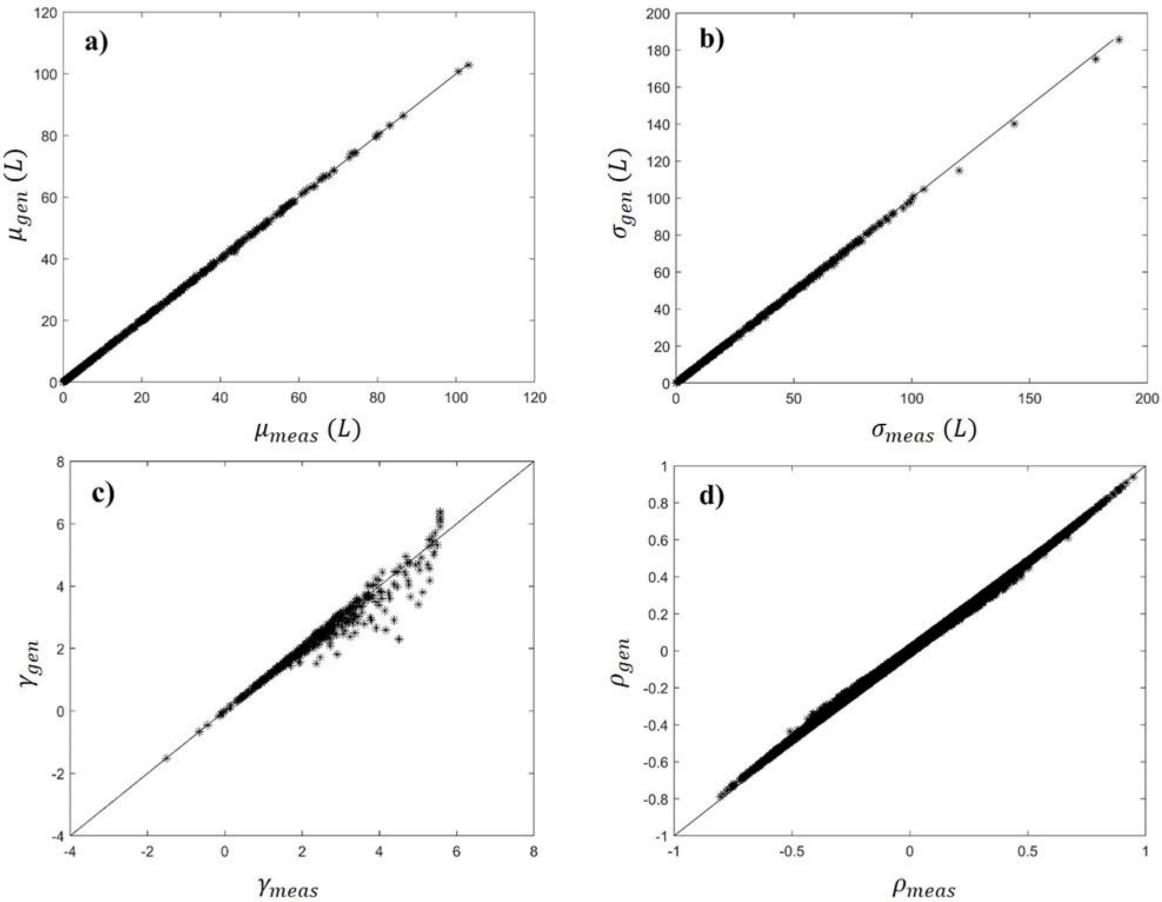

Figure 7 - Comparison of mean values μ (a), standard deviation values σ (b), skewness γ (c) and rank cross-correlations ρ (d) of measured and generated single user demands for Case study 1 (Creaco et al., 2020).

Table 2 shows the R^2 of the fit between the statistics values of measured and generated hourly demands at single user and aggregated scale for both the probability distributions. The table summarizes the results discussed above, showing high performances for μ, σ and ρ ($R^2 = 1$) in case of Beta probability distribution applied at single user scale. As already shown, the fit is excellent for γ as well. Overall, the effectiveness of the methodology in reproducing the main statistics and correlations at single user scale is due to the parameterization that was performed at the scale of the single user.

Table 2 - Comparison of mean values (μ), standard deviation values (σ), skewness values (γ) and rank cross-correlations values (ρ) of measured and generated hourly demands, evaluated in terms of R^2 at both single user and aggregated scale, for the Case study 1.

Distribution	**Single user's demand**				**Aggregated demand**			
	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\gamma}$	$\boldsymbol{\rho}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	$\boldsymbol{\gamma}$	$\boldsymbol{\rho}$
Beta	1.00	1.00	0.95	1.00	1.00	0.96	0.46	0.66
Beta no corr	1.00	1.00	0.95	0.00	1.00	0.77	0.43	0.00
Gamma	1.00	0.99	0.57	1.00	1.00	0.96	0.00	0.65

Figure 8 shows the results for the aggregated demand time series, obtained by summing up the demand time series generated for the single users, of the first case study. Figure 8a shows the comparison between mean values of measured and generated aggregated hourly demands. The level of accuracy of the methodology in reproducing mean values at aggregated scale was high ($R^2 = 1$). Figure 8b reports the comparison in terms of standard deviation values. The performance is again high ($R^2 = 0.96$). The fit of skewness values of measured aggregated demands against those of the generated aggregated demands is shown in Figure 8c. Compared to the other statistics and to the fit obtained at single user scale the accuracy of the methodology in reproducing the skewness values at aggregated scale was lower ($R^2 = 0.46$).

Finally, Figure 8d reports the comparison in terms of cross-correlations at all temporal lags, highlighting a lower performance compared to mean and standard deviation ($R^2 = 0.66$). The methodology overestimated the existing correlations when the measured correlations were smaller than 0. However, the fit is satisfactory. The lower performances in reproducing skewness and rank cross-correlation values at aggregated scale is due to both the parameterization performed at single user scale and the approximations in modelling. Using more complex probability distributions than those used in this work could lead to better results. On the other hand, this could result in the growth of the parameterization burden.

Overall, compared to the results of similar bottom up methodologies for demand generation (Alvisi et al., 2014) the performances at aggregated scale are good. The goodness of the fit at aggregated scale is due to the accuracy of the demand time series generated at single user scale. The aggregated demands were obtained by summing contributions that respected basic statistics and rank cross-correlations. As shown in Table 2, when the correlations were not imposed in the second phase (Beta no corr), worse results for σ ($R^2 = 0.77$), γ ($R^2 = 0.43$) and ρ ($R^2 = 0.00$) were obtained. Thus, accounting for spatial and temporal rank cross-correlations improved the results in terms of standard deviation, skewness and temporal correlations of the aggregated demand time series.

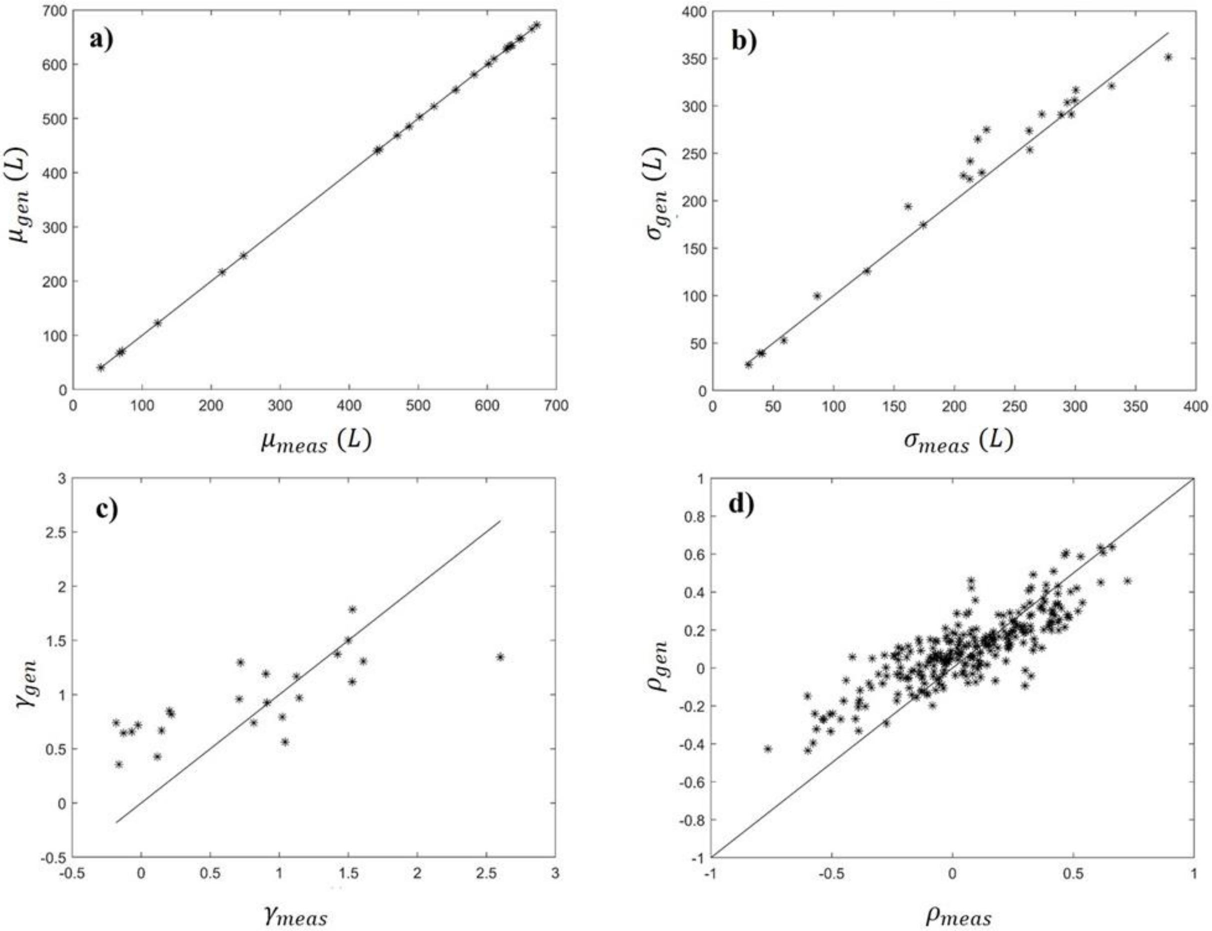

Figure 8 - Comparison of mean values μ (a), standard deviation values σ (b), skewness γ (c) and rank cross-correlations ρ (d) of measured and generated demands at aggregated scale for Case study 1 (Creaco et al., 2020).

Figure 9 shows the first, second and third quantiles for each time step for measured and generated aggregated demand time series. For the synthetic time series, the quantiles are the average over 500 generations of the quantiles of the aggregated demands generated in 93 days. The quantiles for the measured demand time series were obtained starting from the aggregated measured demands in 31 days. Overall, the graph shows the goodness of the fit at aggregated scale.

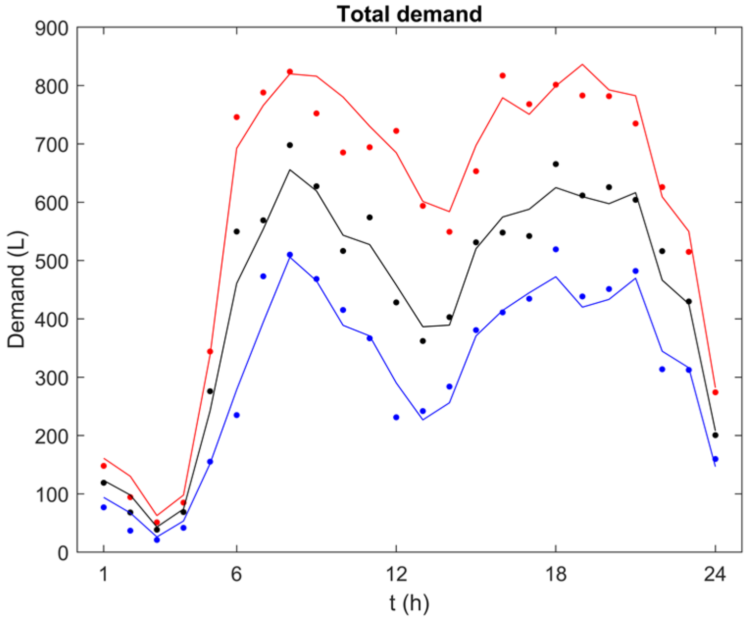

Figure 9 - Daily temporal pattern of first (blue), second (black) and third quantile (red) for measured (dots) and generated (lines) aggregated demand time series, for Case study 1 (modified from Creaco et al., 2020).

Finally, the results obtained applying the Gamma distribution are shown in Table 2. These results are similar to those of the Beta distribution except for the skewness. This is due to the use of the standard gamma distribution with two parameters that affected the reproduction of the skewness of the measured demand time series. In fact, due to the presence of numerous demand time series with $x_0 < 0$ it was necessary to use the standard gamma distribution with two parameters in many cases.

2.3.2 Results – Case study 2

Table 3 reports the results in terms of R^2 of the comparison of the main statistics of measured and generated hourly demands, for both the distributions, for the second case study. At single user scale, the fit of mean, standard deviation and rank cross-correlation values of measured demands to those of the generated demands is very good ($R^2 = 1$) for both the distribution. However, the performance in terms of skewness is better in case of Beta probability distribution with tunable bounds. At aggregated scale, the fit of mean values is again very good ($R^2 = 1$) for both the distributions. The results in terms of standard deviation are very satisfactory in both cases. However, the fit of skewness and rank cross-correlation values obtained with the Beta probability distribution at aggregated scale is better than the one obtained through the Gamma distribution with shift parameter.

Table 3 - Comparison of mean values (μ), standard deviation values (σ), skewness values (γ) and rank cross-correlations values (ρ) of measured and generated hourly demands, evaluated in terms of R^2 at both single user and aggregated scale, for the Case study 2.

Distribution	Single user/node 's demand				Aggregated demand			
	μ	σ	γ	ρ	μ	σ	γ	ρ
Beta	1.00	1.00	0.96	1.00	1.00	0.73	0.59	0.37
Gamma	1.00	1.00	0.80	1.00	1.00	0.75	0.37	0.26
Beta-20 nodes	1.00	1.00	0.79	1.00	1.00	0.92	0.72	0.69

Overall, the performances are better in case of application of the Beta distribution, as for the first case study. As regards the Beta distribution, the results of the generation of single user demand time series are similar to those of the first case study (see also Table 2), while some differences arise at aggregated scale. Compared to the first case study, the fit of skewness values of measured aggregated demands against those of the generated aggregated demands is better ($R^2 = 0.59$). On the other hand, the results in terms of σ and ρ at aggregated scale are worse. Given that the parameterization was carried out on the single user scale, the performances on the aggregated scale can be considered very satisfactory in both case studies. It is worth noting that, if the objective is the water demand generation at aggregated scale, the copula re-sort could be carried out to impose directly rank cross-correlations of the aggregated measured time series.

Figure 10 shows the first, second and third quantiles for each time step for measured and generated aggregated demand time series. This graph constitutes a further evidence of the goodness of the fit of aggregated demand time series.

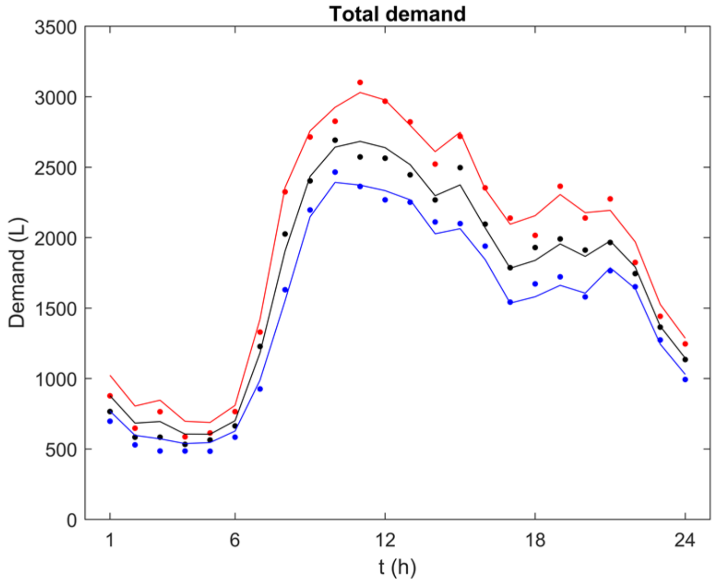

Figure 10 - Daily temporal pattern of first (blue), second (black) and third (red) quantile for measured (dots) and generated (lines) aggregated demand time series, for Case study 2 (modified from Creaco et al., 2020).

The results of the generation of demand time series at the 20 nodes are shown in Table 3. The performances are very good at both single node and aggregated scale. More specifically, the R^2 values for σ, γ and ρ at aggregated scale are higher than those obtained generating aggregated demands starting from single users. Thus, if the objective is to generate aggregated demand time series that preserve the main statistics and the spatial correlations, it is better to apply the methodology starting from the WDN nodes rather than from single users.

2.4 Summary and conclusions

The Chapter presents a novel bottom-up methodology for the generation of water demand time series that respect both the main statistics and the spatial and temporal correlations of the measured time series. According to the proposed methodology, first demand time series for each user and for each time step are generated applying a Beta probability distribution with tunable bounds or a Gamma distribution with shift parameter. Then, a copula based re-sort is applied to the demand time series generated to impose existing rank cross-correlations between users and at all temporal lags. The effectiveness of the methodology was proven in two case studies with different characteristics and number of users. The methodology was able to reproduce mean, standard deviation and above all skewness of measured demand time series, especially at single user scale. Furthermore, the methodology allowed to preserve rank cross-correlations at all time lags in both single user and aggregated time series. The methodology was also successfully applied to the generation of nodal demands. This application showed that starting from nodal demands can be beneficial for the generation of aggregated demands when the number of users increases.

Various methodologies for generating user water demand time series have been developed (Alcocer-Yamanaka et al., 2012; Alvisi et al., 2003; Blokker et al., 2010b; Buchberger and Wu, 1995; Guercio et al., 2001), which are able to represent synthetic water demand time series even with very small time steps (e.g. 1 second). However, such methodologies do not address the problem of how to preserve the accuracy from one level of spatial or temporal aggregation to another. Indeed, the procedures for temporal aggregation (e.g. to obtain hourly time series starting from series with 1 minute time step) and for spatial aggregation are not able to preserve the main statistics of the measured time series at the same level of spatial-temporal aggregation, especially variance and correlations at different time lags. However, it is important to accurately reproduce the demand time series of groups of users and at higher levels of temporal aggregation (e.g. 1 hour), because these are the time series typically used in hydraulic simulation models for designing and managing WDNs (Alvisi et al., 2014). At these higher levels of spatial and temporal aggregation it is also important that the main statistics and correlations of measured time series are preserved to attain a proper characterization of the system performances and avoid practical consequences (Blokker et al., 2008; Filion et al., 2007; Kapelan et al., 2005). Among the literature approaches, the methodology of Alvisi et al. (2014), based on two phases of data re-ordering through the Iman and Conover (1982) procedure, demonstrated to be capable of reproducing mean, variance and temporal correlations at large time steps at the level of individual users, as well as to preserve the same statistics at aggregated scale.

In comparison to the literature methodologies, the methodology presented in this chapter allows to accurately reproduce demand time series at large time steps (e.g. 1 hour time step) and satisfactorily preserve the main statistics and correlations of measured time series at high level of spatial aggregation, especially mean and standard deviation. Compared to the methodology of Alvisi et al. (2014), the presented methodology enables the preservation of not only mean, standard deviation and correlations, but also the skewness of demand time series at both user and aggregated scales. Overall, the presented methodology resulted in higher level of accuracy in reproducing the main statistics of aggregated measured time series. Moreover, the methodology is based on only one phase of data re-ordering to impose spatial and temporal correlations between users at all temporal lags. A further improvement compared to existing methodologies (Alvisi et al., 2014; Gargano et al., 2016) is that the methodology works with a single time step that can be set to any value.

Even though the methodology can work with any time step Δt, some remarks about the choice of the Δt should be made. The methodology is mainly based on the application of rank cross-correlations to demand time series. Therefore, its use is more suitable in case of Δt values in which these correlations are significant, i.e., starting from hourly time step (Moughton et al., 2007). Furthermore, the methodology neglects the pulsed nature of demand, that becomes predominant when Δt is small, i.e., of the order of minutes or seconds. Further remarks can be made about the parameterization of the methodology. The burden of the parameterization is quite high. However, the parameters needed for the methodology can be easily estimated when smart meter readings are available. Under this condition, the methodology enables the generation of consistent synthetic demand time series, thus repaying the effort for the parameterization.

The application to the two case studies showed that the effectiveness of the methodology is preserved even when the number of users grows. Overall, the methodology can be effectively used to generate water demands in WDNs, providing a useful tool for WDNs behaviour simulations.

References

Abozou, T.K., 2007. Copulas in Statistics. African Institute for Math. Sciences.

Alcocer-Yamanaka, V.H., Tzatchkov, V.G., Arreguin-Cortes, F.I., 2012. Modeling of drinking water distribution networks using stochastic demand. Water Resour. Manag. 26, 1779–1792. https://doi.org/10.1007/s11269-012-9979-2

Alvisi, S., Ansaloni, N., Franchini, M., 2014. Generation of synthetic water demand time series at different temporal and spatial aggregation levels. Urban Water J. 11, 297–310. https://doi.org/10.1080/1573062X.2013.801499

Alvisi, S., Franchini, M., Marinelli, A., 2003. A stochastic model for representing drinking water demand at residential level. Water Resour. Manag. 17, 197–222. https://doi.org/10.1023/A:1024100518186

Blokker, E.J.M., Vreeburg, J.H.G., Buchberger, S.G., van Dijk, J.C., 2008. Importance of demand modelling in network water quality models: a review. Drink. Water Eng. Sci. Discuss. 1, 27–38. https://doi.org/10.5194/dwesd-1-1-2008

Blokker, E.J.M., Vreeburg, J.H.G., van Dijk, J.C., 2010. Simulating Residential Water Demand with a Stochastic End-Use Model. J. Water Resour. Plan. Manag. 136, 375–382. https://doi.org/10.1061/(asce)wr.1943-5452.0000002

Blokker, M., Agudelo-vera, C., Moerman, A., van Thienen, P., Pieterse-Quirijns, I., 2017. Review of applications of SIMDEUM, a stochastic drinking water demand model with small temporal and spatial scale. Drink. Water Eng. Sci. Discuss. 10, 1–15. https://doi.org/10.5194/dwes-2017-4

Buchberger, S.G., Carter, J., Lee, Y., Shade, T.., 2003. Random demands, travel times, and water quality in deadends, subject area: high-quality water. AWWA Research Foundation, Denver, CO.

Buchberger, S.G., Wu, L., 1995. Model for instantaneous residential water demands. J. Hydraul. Eng. 121, 232–246. https://doi.org/10.1061/(ASCE)0733-9429(1995)121:3(232)

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., Rizzoli, A.E., 2015. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. Environ. Model. Softw. 72, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012

Creaco, E., Blokker, M., Buchberger, S., 2017a. Models for generating household water demand pulses: Literature review and comparison. J. Water Resour. Plan. Manag. 143, 04017013. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000763

Creaco, E., De Paola, F., Fiorillo, D., Giugni, M., 2020. Bottom-Up Generation of Water Demands to Preserve Basic Statistics and Rank Cross-Correlations of Measured Time Series. J. Water

Resour. Plan. Manag. 146, 06019011. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001142

Creaco, E., Pezzinga, G., Savic, D., 2017b. On the choice of the demand and hydraulic modeling approach to WDN real-time simulation. Water Resour. Res. 53, 6159–6177. https://doi.org/10.1002/2016WR020104

Filion, Y., Adams, B., Karney, B., 2007. Cross correlation of demands in water distribution network design. J. Water Resour. Plan. Manag. 133, 137–144. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:2(137)

Fox, C., McIntosh, B.S., Jeffrey, P., 2009. Classifying households for water demand forecasting using physical property characteristics. Land use policy 26, 558–568. https://doi.org/10.1016/j.landusepol.2008.08.004

Gargano, R., Di Palma, F., de Marinis, G., Granata, F., Greco, R., 2016. A stochastic approach for the water demand of residential end users. Urban Water J. 13, 569–582. https://doi.org/10.1080/1573062X.2015.1011666

Guercio, R., Magini, R., Pallavicini, I., 2001. Instantaneous residential water demand as stochastic point process. Prog. Water Resour. 129–138. https://doi.org/10.2495/WRM010121

Iman, R.L., Conover, W.J., 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. Commun. Stat. - Simul. Comput. 11, 311–334. https://doi.org/10.1080/03610918208812265

Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous univariate distributions. Volume 2 (2nd ed.), Wiley. ed.

Kapelan, Z.S., Savic, D.A., Walters, G.A., 2005. Multiobjective design of water distribution systems under uncertainty. Water Resour. Res. 41, W11407. https://doi.org/10.1029/2004WR003787

Mikosch, T., 2006. Copulas: Tales and facts. Extremes 9, 3–20. https://doi.org/10.1007/s10687-006-0015-x

Moughton, L.J., Buchberger, S.G., Boccelli, D.L., Filion, Y.R., Karney, B.W., 2007. Effect of time step and data aggregation on cross correlation of residential demands, in: 8th Annual Water Distribution Systems Analysis Symposium 2006. ASCEE, Cincinnati, Ohio (USA), pp. 1–11. https://doi.org/10.1061/40941(247)42

Nelsen, R.B., 1999. An Introduction to Copulas. Springer, New York. https://doi.org/10.1007/978-1-4757-3076-0

Pearson, K., 1895. Note on regression and inheritance in the case of two parents. Proc. R. Soc. London 58, 240–242. https://doi.org/10.1098/rspl.1895.0041

Schleich, J., Hillenbrand, T., 2009. Determinants of residential water demand in Germany. Ecol. Econ. 68, 1756–1769. https://doi.org/10.1016/j.ecolecon.2008.11.012

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. Am. J. Psychol. 15, 72–101.

3 Water demand generation: comparison between bottom-up and top-down approaches

The continuing growth in population of the cities is resulting in soaring water demands. At the same time, climate changes and extreme events, such as prolonged droughts, further increase stresses on regional water resources in many countries. More attention is being paid to water resources management and plans at different levels, such as regional, national and transboundary, to address water demands (Sheng et al., 2017).

The need of an effective management of the cities requires water supply utilities to reduce leakages and energy consumption. In this context, the use of new data in hydraulic models can improve the development of new design and management tools for distribution analysis (Brentan et al., 2018).

In common modelling practice, the water demand of the users is allocated to the nodes of the hydraulic model of the WDN. Two different approaches are usually used for allocating water demands (Walski et al., 2003): 'bottom-up' and 'top-down'. According to the top-down approach, first the water demand pattern is characterised at high levels of spatial aggregation and then the nodal patterns are defined based on the respective number of users. In the case of the bottom-up approach, in contrast, the pattern of water demand is characterised at low spatial aggregation levels, such as at single user level, and the nodal demand pattern is obtained by aggregating the water demands of the users allocated at the node considered (Alvisi et al., 2016).

As shown in Chapter 1, various applications of top-down (Deidda et al., 2004; Koutsoyiannis and Manetas, 1996; Kumar et al., 2000; Parsons et al., 2007) and bottom-up (Alcocer-Yamanaka et al., 2012; Alvisi et al., 2014; Blokker et al., 2017; Buchberger and Wu, 1995; Gargano et al., 2016) approaches exist in the scientific literature. However, few comparative works have been presented so far (Blokker, 2010), aimed at assessing the effectiveness of such approaches in both water resources planning (Sheng et al., 2017) and identifying the main energy inefficiencies in water systems (Mamade et al., 2018). However, the present chapter is aimed at presenting this comparison in a real case study, highlighting the differences in synthetic time series generation in terms of statistics, such as media, standard deviation, skewness and correlations. More specifically, the first methodology analysed is based on the top-down non-parametric disaggregation model developed by Nowak et al. (2010). The second one, is the bottom-up methodology presented in Chapter 2. The results of their

This chapter is directly based on publication III.

applications to two case studies, both consisting of hourly consumption data from the Soccavo DMA, are discussed in order to carry out a comparison between the two methodologies.

3.1 Methodology

In this work two methodologies for generating synthetic water demand time series were used. In order to generate demand time series for a fixed number of days (n_{days}) and for each of the N nodes (or users) considered, for both methodologies the generic day is subdivided into $N_{\Delta t}$ time steps Δt.

The first methodology is based on the top-down approach and consists of the non-parametric disaggregation model developed by Nowak et al. (2010). Starting from the temporal water demand pattern at a high level of spatial aggregation, the disaggregation model allows the generation of water demand time series at lower levels of spatial aggregation. The second one is the bottom-up methodology explained in Chapter 2. In this work it is assumed that the daily demand q_i^j of the generic $j-th$ user in the $i-th$ Δt follows the Beta probability distribution with tunable bounds.

For both methodologies, the generated demand time series were obtained by assuming the typical day of generation to be subdivided into $N_{\Delta t}$=24 time steps with Δt =1 hr. Furthermore, the generation of demand time series was performed for a number of days $n_{days} = 93$ and was reiterated for 500 times for both methodologies. Finally, the average values over the 500 iterations were considered.

The following sections present the top-down methodology and the non-parametric disaggregation model. The description of the bottom-up methodology is provided in Chapter 2.

3.1.1 Top-down methodology

The methodology allows to generate the water demand time series q_i^j of the generic $j-th$ node (or user) in the generic $i-th$ Δt, starting from the total amount of water Q_i supplied at the $i-th$ time step. The methodology is made up of two phases (Alvisi et al., 2016).

In the first phase a stochastic (Bras and Rodrìguez-Iturbe, 1993) or non-parametric algorithm (Lall and Sharma, 1996) is used to generate the total water demand time series of the area (i.e. one demand time series for each Δt of the day). In this work, the total demand at the generic time step of the day is sampled from a Beta probability distribution with tunable bounds. As shown in Chapter 2, this distribution enables preserving mean, variance and skewness of the total demand time series. Then, a copula re-sorting is applied on the generated time series to impose the existing temporal correlations on the total demand at all temporal lags. Specifically, a multivariate normal probability distribution, with means and standard deviations equal to 0 and 1 respectively, is used as copula (Nelsen, 1999). The rank cross-correlations are derived from the measured time series through the Spearman index (as shown in Chapter 2). The multivariate normal distribution is then used to generate time series expressing the rank cross-correlations to be imposed on demand time series at all temporal lags. Finally, in order to impose the rank cross-correlations, the demand time series generated are re-sorted following the order of the copula-generated time series.

For the parameterization of the methodology, mean μ, standard deviation σ and skewness γ for each total demand time series must be assessed. Therefore, $3 \times N_{\Delta t}$ parameters are required. In order to implement the Beta probability distribution with tunable bounds, the minimum value of the total demand time series is needed (see Chapter 2). Thus, further $N_{\Delta t}$ parameters must be assessed. Finally, temporal cross-correlations between total demand time series are needed. Assuming that each time series is fully correlated with itself and the correlation between two time series is symmetrical, $(N_{\Delta t} \times (N_{\Delta t} - 1))/2$ parameters are required. Concluding, the number of parameters to be assessed adds up to $4 \times N_{\Delta t} + (N_{\Delta t} \times (N_{\Delta t} - 1))/2$.

In the second phase, the water demand time series of each node are generated by spatially disaggregating the total demand time series.

3.1.2 Disaggregation model

In the present work the non-parametric disaggregation model proposed by Nowak et al. (2010) was used.

Let us assume an hourly time-step. According to the non-parametric disaggregation, the nodal demands are generated by random resampling from the conditional probability density function $f(q_h|Q_h)$, where Q_h and q_h are the random variables representing the aggregated demand in the $h-th$ hour and the disaggregated demand in the $h-th$ hour, respectively.

In the model of Nowak et al. (2010) the conditional density function is carried out using a K-nearest neighbours (K-NN) approach applied on the basis of the observed aggregated series. Specifically, let us assume the length of the generated and measured aggregated time series respectively equal to $n_{d,g} \times N_{\Delta t}$ and $n_{d,m} \times N_{\Delta t}$, where $n_{d,g}$ and $n_{d,m}$ are the numbers of days of generated and measured time series respectively. A number K of values (K-NN) near each generated value of the aggregated series ($Q_{n,h}^{gen}$, with $n = 1{:}\, n_{d,g}$) are identified from the measured aggregated demands related to the same hour h ($Q_{m,h}^{mea}$, with $m = 1{:}\, n_{d,m}$).

The number K of neighbours can be defined based on a heuristic approach. According to this approach, the optimal number K is equal to $\sqrt{n_{d,m}}$ (Lall and Sharma, 1996). The neighbours are selected based on the absolute value of the difference between the observed and generated aggregated values (Δ). The K values with the smallest Δ are selected. Then, after being reordered from the nearest to the farthest, the K-NN are assigned a weight W_l (with $l = 1{:}K$) according to their position l in the reordered vector:

$$W_l = \frac{1/l}{\sum_{j=1}^{K} 1/l} \tag{14}$$

Therefore each K-nearest neighbour has an extraction probability equal to W_l. Then, one of the K-NN is selected based on the weighted resampling. The corresponding proportions $p_{j,d,h}$ for each of the N nodes for the $h-th$ hour of the $d-th$ day are calculated on the basis of measured demands:

$$p_{j,d,h} = \frac{q_{j,d,h}^{mea}}{Q_{d,h}^{mea}} \tag{15}$$

where $q_{j,d,h}^{mea}$ is the generic disaggregated measured value of the j-th node for the $h-th$ hour of the $d-th$ day and $Q_{d,h}^{mea}$ is the associated aggregated value.

Finally, the obtained proportions are multiplied by the generated aggregated value $Q_{d,h}^{gen}$ to provide the generated disaggregated values $q_{j,d,h}^{gen}$:

$$q_{j,d,h}^{gen} = p_{j,d,h} Q_{d,h}^{gen} \tag{16}$$

3.2 Case studies

In this thesis two case studies were analysed. For both case studies the hourly consumption data from the Soccavo DMA were used. For the sake of comparison with the results presented in Chapter 2, in the first case study (*Case study 1*) the data of 100 users for 31 days, from 1 January 2018 to 31 January 2018, were considered (thus coinciding with the second case study of Chapter 2).

The second case study (*Case study 2*) is made up of 1000 users from the same smart water district, monitored from 1 October 2017 to 31 October 2017. It follows from the above that the case studies essentially differ because of the number of users.

Figure 11 shows the daily patterns of aggregated measured hourly demand in the 31 days considered in both case studies. The patterns highlight similar characteristics of water consumption.

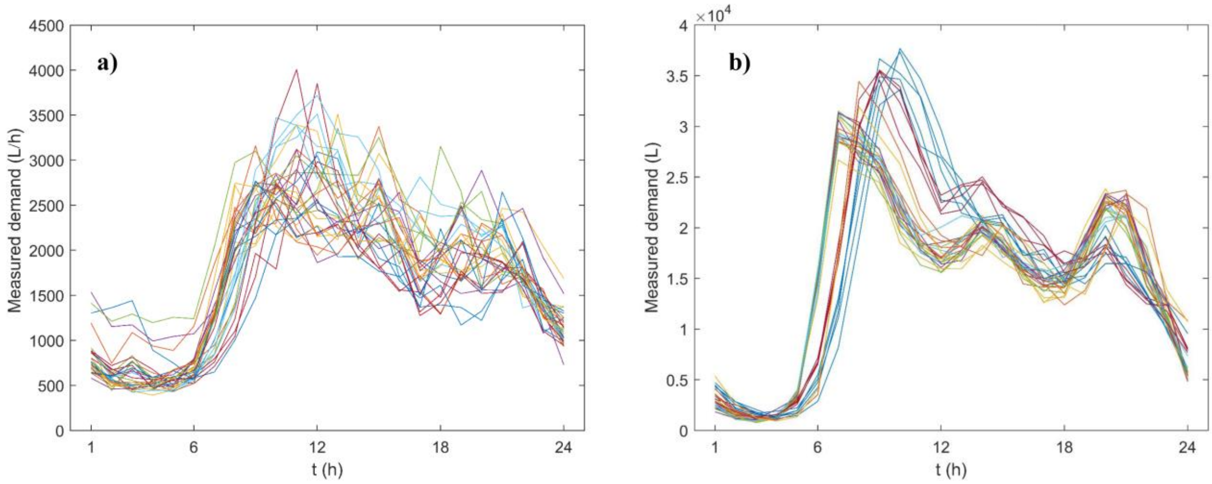

Figure 11 - Patterns of aggregated measured hourly demand for 31 days for (a) Case study 1 and (b) Case study 2 (Fiorillo et al., 2020).

3.2.1 Applications

The comparison between the top-down and the bottom-up methodologies was carried out by performing different applications for both case studies. The details of the performed applications are reported in Table 4.

For the sake of comparison with the results obtained for the bottom-up methodology shown in Chapter 2, the top – down methodology was applied to the *Case study 1* performing two applications. In the first application (*Application 1.1*) the top – down methodology was applied in order to generate single user and aggregated water demand time series, in the second application (*Application 1.2*) nodal demands were considered grouping the users in 20 nodes. In *Application 1.2* the measured demand time series of the generic node were estimated as the sum of the demand time series of the related users. Therefore, in *Application 1.1* the top-down methodology was parameterized based on the measured single user demand time series, while in *Application 1.2* nodal demand time series were used for the parameterization.

For the *Case study 2*, three applications were performed for both methodologies. In the first application (*Application 2.1*) the users were grouped in 10 nodes by allocating 100 users in each node. Similarly, in the second application (*Application 2.2*) in each node 20 users were allocated, resulting in 50 nodes. Finally, in the third application (*Application 2.3*) the users were grouped in 100 nodes with 10 users in each one. As for *Application 1.2* the methodologies were parameterized based on the measured nodal demand time series.

Table 4 – Details of the performed applications.

Application	Case study	N° of nodes	N° of users in each node
Application 1.1	*Case study 1*	100	1
Application 1.2	*Case study 1*	20	5
Application 2.1	*Case study 2*	10	100
Application 2.2	*Case study 2*	50	20
Application 2.3	*Case study 2*	100	10

3.3 Results and discussion

In the following sections the results obtained applying both the top-down and bottom-up methodologies to both case studies for each application are reported. The comparison between the methodologies is carried out by analyzing the accuracy in reproducing the main statistics and the correlations of the measured time series. In the comparison between basic statistics of measured and generated demand time series, the average values over 500 iterations were considered for generated demand time series for both methodologies.

3.3.1 Results – Case study 1

For the single user demands of *Application 1.1*, Figure 12 shows the comparison in terms of mean (μ), standard deviation (σ), skewness (γ) and rank cross-correlations (ρ) between measured and generated hourly demands for the top-down methodology. Figure 12a shows the comparison between measured and generated hourly demands in terms of mean values.

The level of accuracy of the methodology in reproducing mean values was high ($R^2 = 1$). Almost the same result was obtained for the standard deviation ($R^2 = 0.95$) and skewness values ($R^2 = 0.96$), as shown in Figure 12b and Figure 12c, respectively.

As regards the cross-correlations, in Figure 12d the dots representing the rank cross-correlations at lag 0, i.e. the spatial correlations between the N user demands in the same hour, are differentiated from the others. Indeed, the values of the rank cross-correlations at lag 0 show a good fit ($R^2 = 0.89$). However, the top – down methodology failed to preserve the rank cross-correlations between users and at various temporal lags ($R^2 = 0$). The non-parametric approach was unable to preserve the correlations between the demands associated with one hour and those of the previous hour. This is because, according to the non-parametric approach, for each hour the data were disaggregated independently of those related to other hours (Alvisi et al., 2016).

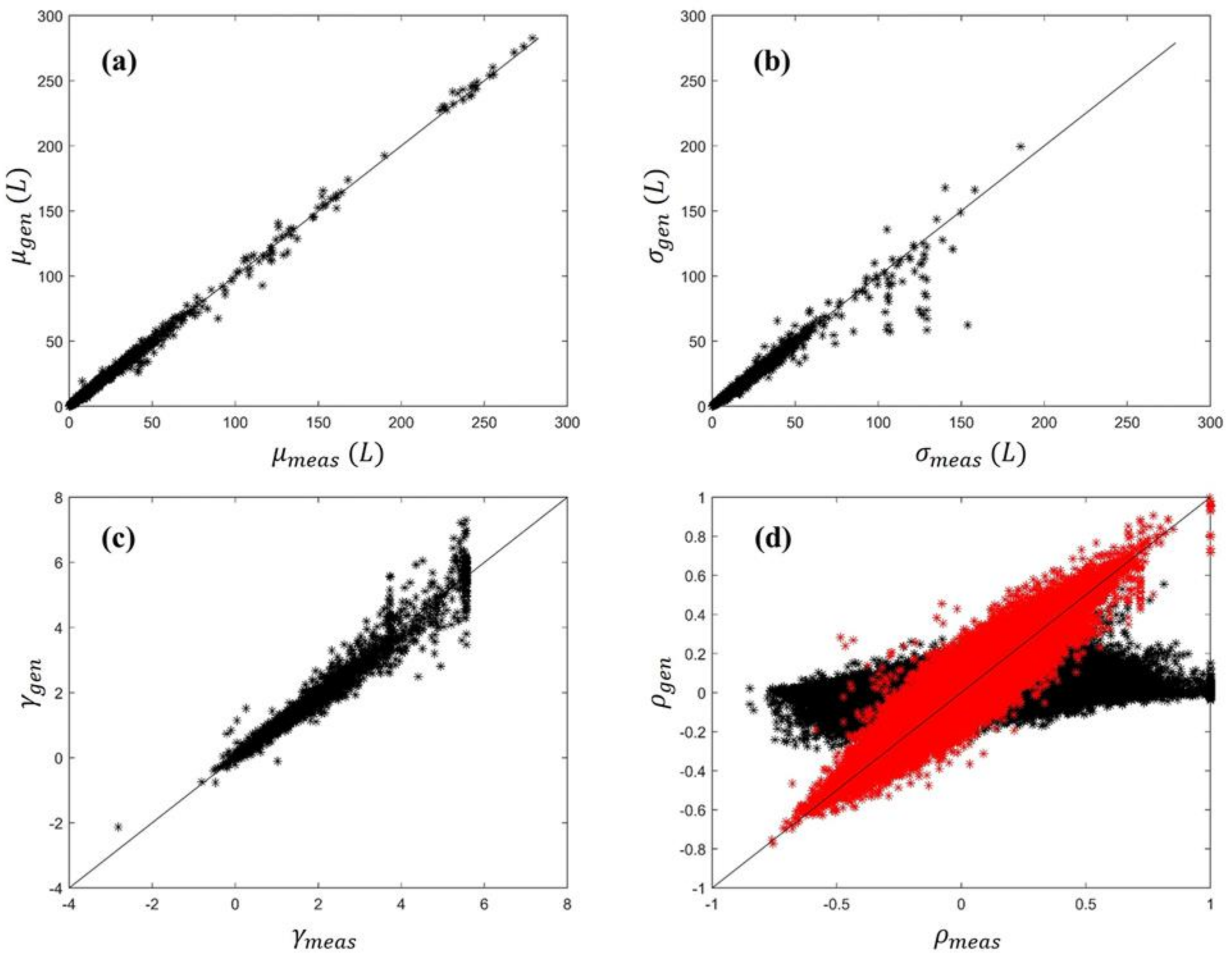

Figure 12 - Case study 1 (Application 1.1) - Comparison between measured and generated single user demands: (a) mean values μ*, (b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ (black) and $\rho - lag$ 0 (red), for the top-down methodology (Fiorillo et al., 2020).*

Figure 13 shows the results for the aggregated demand time series obtained applying the top-down methodology in *Application 1.1*.

As shown in Figure 13a, Figure 13b and Figure 13d, the comparison between measured and generated aggregated hourly demands in terms of mean, standard deviation and rank cross-correlations values highlights a very good agreement ($R^2 = 1$).

The good performance in reproducing the cross-correlations can be traced back to the parametrization performed on aggregated scale. Thus, a copula re-sorting is used to impose the existing temporal correlations to the generated aggregated demand time series, resulting in the preservation of the measured cross-correlations at aggregated scale.

As regards the skewness values (Figure 13c), the performance is good ($R^2 = 0.94$) as well.

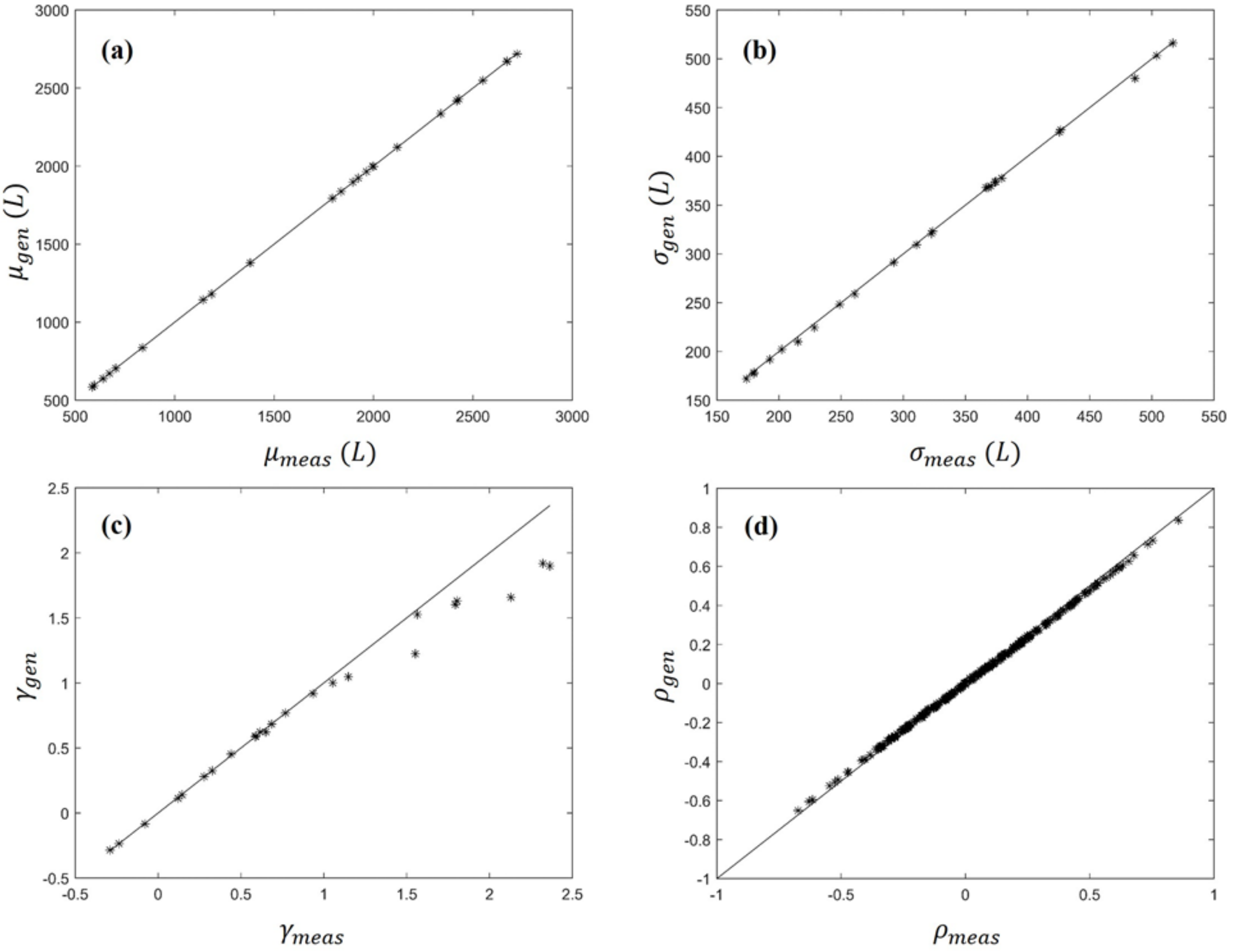

Figure 13 - Case study 1 (Application 1.1) - Comparison between measured and generated aggregated demands: (a) mean values μ, *(b) standard deviation values* σ, *(c) skewness* γ, *(d) rank cross-correlations* ρ, *for the top-down methodology (Fiorillo et al., 2020).*

The graphs in Figure 14 show the results obtained in *Application 1.2* for hourly demands at single node level generated through the top-down methodology, leading to similar considerations to *Application 1.1*.

Though respecting the basic statistics in terms of mean, variance and skewness the top-down approach was unable to reproduce the existing rank cross-correlations between nodes at various temporal lags.

As for every user in *Application 1.1.*, for every node the correlations between the demands in one hour and those associated with the previous hour were not preserved due to the non-parametric disaggregation.

According to the non-parametric approach, the nodal demands for each hour were obtained by disaggregating the total demand independently of the other hours. For this, the top-down methodology was not able to preserve the temporal correlations at node level.

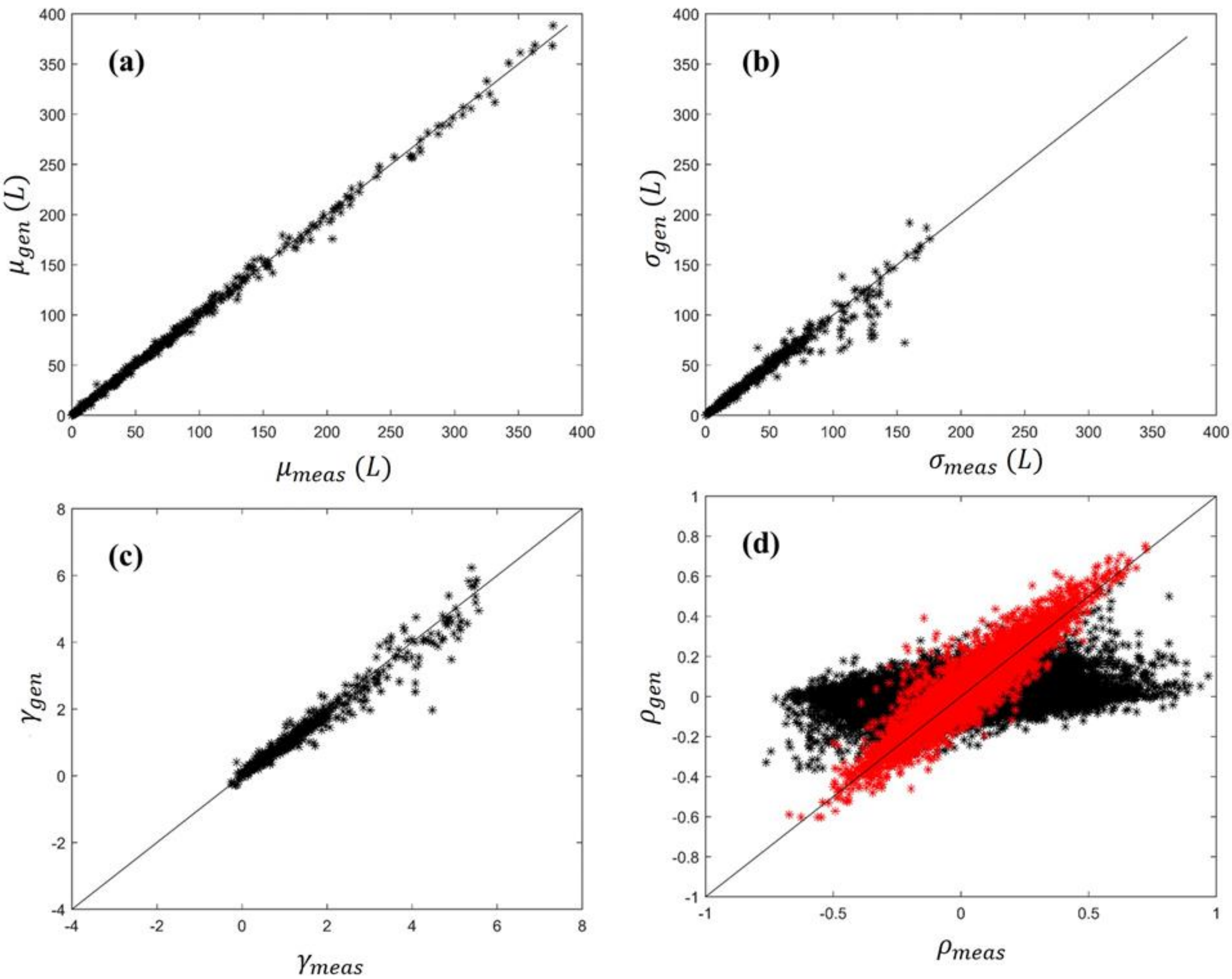

Figure 14 - Case study 1 (Application 1.2) - Comparison between measured and generated single node demands: (a) mean values μ, (b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ (black) and $\rho - lag\ 0$ (red), for the top-down methodology (Fiorillo et al., 2020).

Finally, Figure 15 reports the comparison of measured and generated aggregated hourly demands for the *Application 1.2*.

The performance on the aggregated scale is excellent as in the *Application 1.2*, demonstrating the effectiveness of the methodology in total district consumption reconstruction.

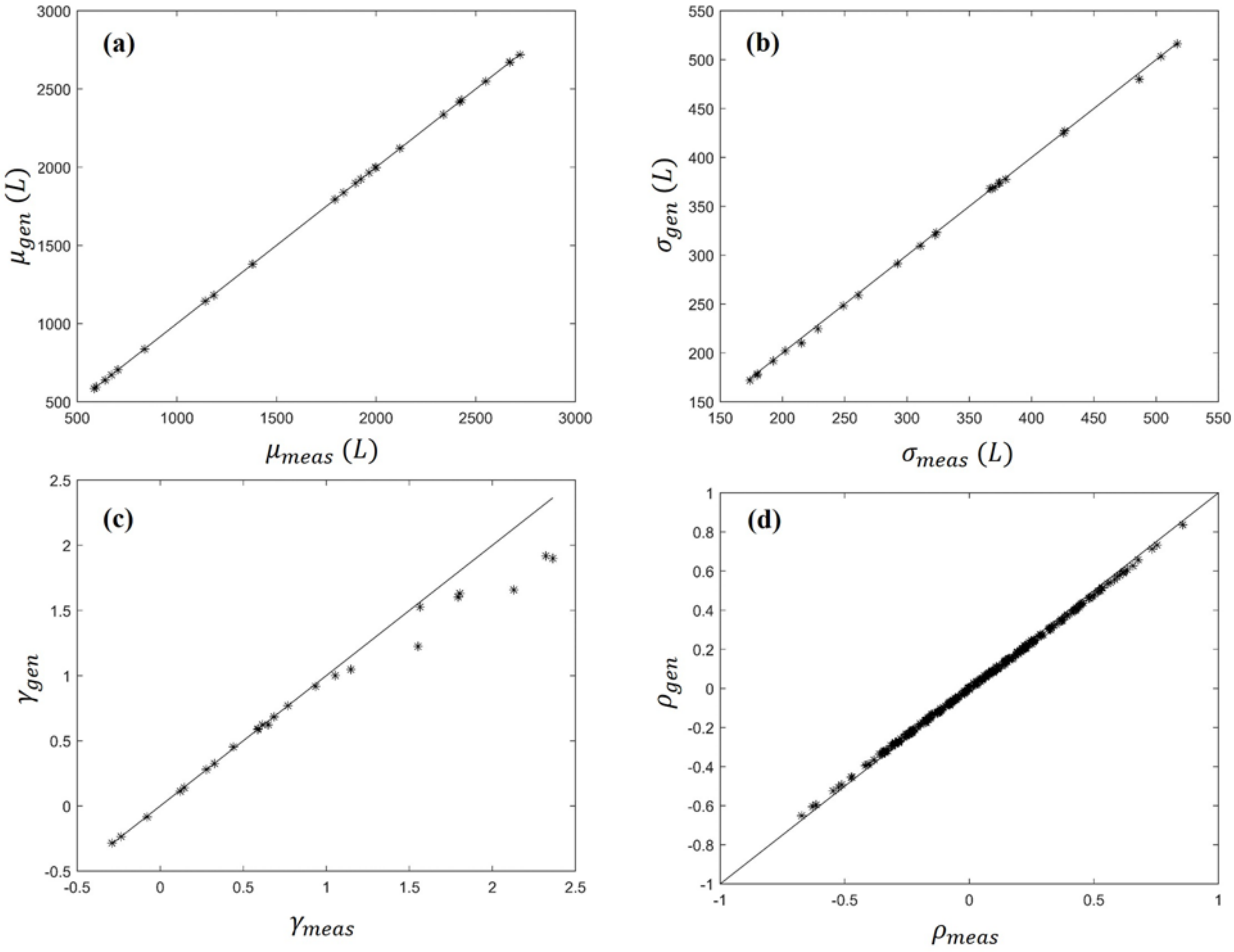

Figure 15 - Case study 1 (Application 1.2) - Comparison between measured and generated aggregated demands: (a) mean values μ, *(b) standard deviation values σ, (c) skewness γ, (d) rank cross-correlations ρ, for the top-down methodology (Fiorillo et al., 2020).*

Table 5 summarizes the results discussed above along with some of the results shown in Chapter 2. In order to facilitate the comparison between the top-down and bottom-up methodology, the results obtained applying the bottom-up methodology (in case of Beta probability distribution with tunable bounds) to the *Case study 1* starting from single user (*Application 1.1)* and nodal demands (*Application 1.2*) are reported again in Table 5. The results confirm the inability of the top-down methodology to reproduce the cross-correlations at various temporal lags at both single user and nodal scales. On the contrary, in both cases (single users and single nodes) the performance in terms of rank cross-correlation of the bottom-up methodology is excellent. The R^2 of the fit between measured and generated values is equal to 1 at both single user and nodal scales. This is due to the fact that the generated demand time series at user/node scale are re-sorted through a copula to impose existing rank cross-correlations between the users/nodes and at each time step.

When considering the aggregated demands, the results in terms of skewness and rank cross-correlations obtained with the top-down methodology are better than those of the bottom-up one. This is due to the parameterization. For the bottom-up methodology the parameterization performed on the single user scale led to the deterioration of the fit in terms of skewness and rank cross-correlation of the aggregated demand time series. Indeed, the level of spatial aggregation can affect the values of the main statistics and cross-correlations of water demands (Alvisi et al., 2014; Moughton et al., 2007). More specifically, cross-correlations tend to increase with increasing levels of spatial aggregation (Moughton et al., 2007). Therefore, the performances of the bottom-up methodology at aggregated scale were lower, since the generated aggregated demand time series were obtained by summing up the demand time series generated at single user/node scale and hence using the statistics and correlations of the measured user/node demand time series. On the contrary, the top-down methodology was parameterized based on the aggregated demand times series, leading to good results at aggregated scale.

Table 5 - Comparison of mean values μ, standard deviation values σ, skewness γ, rank cross-correlations ρ and $\rho - lag\ 0$ of measured and generated demands, evaluating the fit in terms of R^2 at both single and aggregated scales, for both applications to the first case study and for both the top-down and the bottom-up methodologies.

Application - Methodology	Single user/node 's demand					Aggregated demand			
	μ	σ	γ	ρ	$\rho - lag0$	μ	σ	γ	ρ
Application 1.1 – Top-down	1	0.95	0.96	0	0.89	1	1	0.94	1
Application 1.2 – Top-down	1	0.94	0.95	0	0.85	1	1	0.94	1
Application 1.1 – Bottom-up	1	1	0.96	1	-	1	1	0.59	0.37
Application 1.2 – Bottom-up	1	1	0.79	1	-	1	0.92	0.72	0.69

3.3.2 Results – Case study 2

The results in terms of R^2 of the fit obtained by applying the top-down methodology to *Case study 2* are shown in Table 6. For all the applications, the performance on mean, standard deviation, skewness and cross-correlations at lag 0 at single node level is excellent. However, the fit of rank cross-correlations at lag 0 seems to slightly improve with the increasing level of aggregation in nodes, i.e.

with the increasing number of users allocated in each node. The highest performance in reproducing rank cross-correlations at lag 0 ($R^2 = 0.92$) were obtained in *Application 2.1* (i.e. in case of 100 users in each node), while the *Application 2.3* (10 users in each node) resulted in the lowest performance ($R^2 = 0.88$).

Even though *Application 2.2* and *Application 2.3* confirmed the inability of the top-down methodology to reproduce the cross-correlations at various temporal lags, better results were obtained in case of 10 nodes (*Application 2.1*). For *Application 2.1* the R^2 reached a value of 0.54. Thus, it can be stated that a high level of aggregation in nodes can improve the performance in terms of cross-correlations for the top-down methodology.

In terms of aggregated demands, the fit is always excellent for all three applications, confirming the effectiveness of the methodology in total district consumption reconstruction.

Table 6 - Comparison of mean values μ, *standard deviation values* σ, *skewness* γ, *rank cross-correlations* ρ *and* $\rho - lag\,0$ *of measured and generated demands, evaluating the fit in terms of* R^2 *at both single and aggregated scales, for each application of top-down methodology to Case study 2.*

Application	**Single node 's demand**					**Aggregated demand**			
	μ	σ	γ	ρ	$\rho - lag0$	μ	σ	γ	ρ
Application 2.1	1	0.99	0.94	0.54	0.92	1	1	0.98	1
Application 2.2	1	0.98	0.92	0.00	0.90	1	1	0.98	1
Application 2.3	1	0.97	0.92	0.00	0.88	1	1	0.98	1

Table 7 shows the results obtained applying the bottom-up methodology to *Case study 2*. For all applications, the fit in terms of mean (μ), standard deviation (σ) and rank cross-correlations (ρ) between measured and generated nodal demands is excellent ($R^2 = 1$). The performances on skewness are excellent as well.

Overall, the bottom-up methodology performed better than the top-down methodology in terms of rank cross-correlations at single node level. The high performance confirmed the effectiveness of the copula-based re-sort during the second phase of the bottom-up methodology. At aggregated scale, for all applications the performances on mean, standard deviation and cross-correlations are excellent (Table 7). However, the performance in terms of skewness is better in case of application of the top-down approach. As already stated in Chapter 2, the deterioration of the fit in terms of skewness when aggregated demand is considered is due to both the parameterization, which was performed on the single user scale, and the approximations inherent in the modelling. However, the fit of the skewness values seems to improve with the increasing level of aggregation in nodes. The maximum value of R^2 was reached in case of *Application 2.1* ($R^2 = 0.64$), when users were grouped in 10 nodes by allocating 100 users in each node, while the *Application 2.3* (10 users in each node) led to the worst performance ($R^2 = 0.34$).

Table 7 - Comparison of mean values μ, standard deviation values σ, skewness γ, rank cross-correlations ρ and $\rho - lag\ 0$ of measured and generated demands, evaluating the fit in terms of R^2 at both single and aggregated scales, for each application of bottom-up methodology to Case study 2.

Application	Single user/node 's demand				Aggregated demand			
	μ	σ	γ	ρ	μ	σ	γ	ρ
Application 2.1	1	1	0.90	1	1	0.97	0.64	0.96
Application 2.2	1	1	0.88	1	1	0.98	0.40	0.94
Application 2.3	1	1	0.90	1	1	0.98	0.34	0.93

The graphs in Figure 16 report, for measured and generated aggregated demand time series, the first, the second and the third quantiles for each time step for both the methodologies in *Application 2.2*. The quantiles for the measured demand time series were obtained starting from the aggregated measured demands in 31 days, while those for the generated demand time series are the average over 500 generations of the quantiles of the aggregated demands generated in 93 days. These graphs constitute further evidence of the goodness of the fit at aggregated scale for both methodologies.

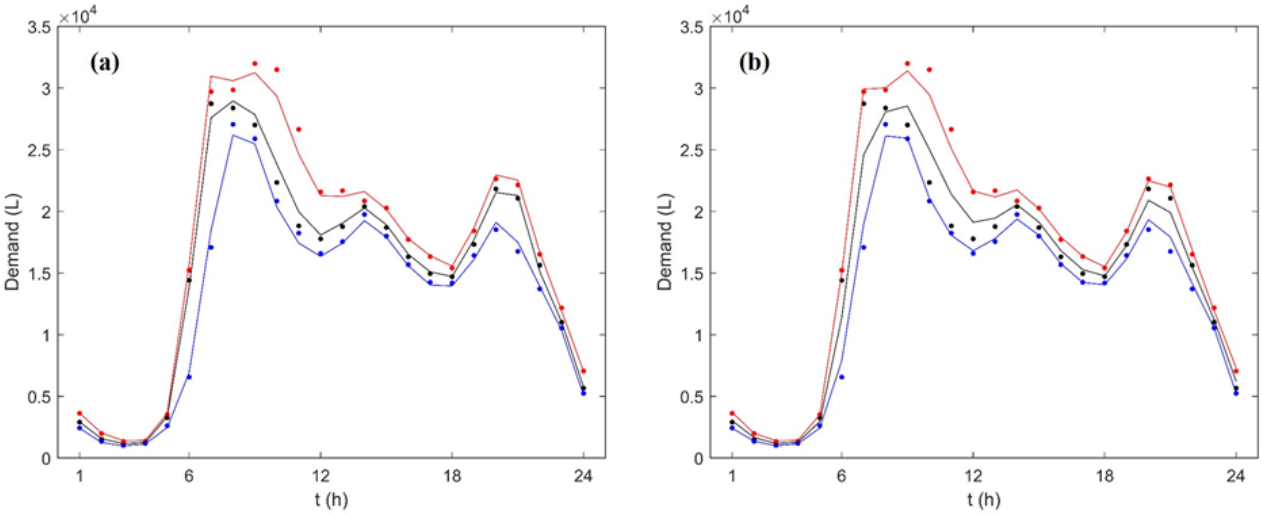

Figure 16 - Case study 2 (Application 2.2) - Daily temporal pattern of first (blue), second (black) and third (red) quantile for measured (dots) and generated (lines) aggregated demand time series: the generated demands were obtained applying the top-down (a) and the bottom-up (b) methodology respectively.

3.4 Summary and conclusions

This Chapter presents two methodologies for the generation of demand time series at both single user and nodal scale. The first methodology consists of a top-down approach based on the disaggregation model developed by Nowak et al. (2010). According to this methodology, once the aggregated water demand patterns have been defined, the disaggregation is applied to generate water demand time series at lower levels of spatial aggregation. The methodology uses a non-parametric disaggregation model based on the K-NN approach applied on the basis of the observed aggregated series. The second methodology is based on a bottom-up approach. Thus, it reconstructs WDN demand starting from the single nodes upward. A copula based re-sort is applied to demand time series of the first attempt, generated through a Beta probability distribution, to impose existing rank cross-correlations.

Both methodologies were applied to two case studies with different number of users, performing various application types. Then, the comparison between the methodologies was carried out by analysing their performances in reproducing the main statistics and cross-correlations of measured demand time series.

While the reproduction of mean and standard deviation of demand time series at single users, nodes and aggregated scale is satisfactory for both methodologies, differences arise for demand cross-correlations and skewness. The bottom-up methodology performed better than the top-down one at reproducing cross-correlations between single users (nodes). The top-down approach resulted to be poorly capable of reproducing the cross-correlations at various temporal lags, though results slightly improved in case of high levels of aggregation in nodes. The inability to accurately reproduce cross-correlations may become a not negligible problem (Alvisi et al., 2016; Blokker et al., 2011; Filion et al., 2007, 2005; Kapelan et al., 2005). Indeed, the value of auto- and cross-correlations can affect the system's resilience, namely the time necessary to restore normal operating conditions after a pipe break (Filion et al., 2005), the amount of time water remains standing in the network along with the simulated velocities (Blokker et al., 2011) and the mean and variance of pressure head values (Filion et al., 2007).

As regards the aggregated demands, the top-down methodology performed better in terms of skewness and rank cross-correlations. However, it was found that the application of the bottom-up methodology for the generation of nodal demands, rather than single users' demands, resulted in positive impacts in respecting rank cross-correlations of the aggregated time series. Overall, it can be successfully applied to the generation of nodal demands in WDNs.

It is worth noting that, even though an hourly time step was used in this work, both methodologies are expected to be effective also for different values of Δt. Anyway, when accounting for rank cross-correlations, large time step (e.g. hours) should be used (Moughton et al., 2007). Finally, the higher burden of parameterization of the bottom-up methodology should be taken into account when comparing the two methodologies. On one hand, the top-down methodology requires the definition of a lower number of parameters and a lower computational burden. On the other hand, the parameters required by the bottom-up methodology can be easily estimated when smart meter readings are widely available over the WDN.

References

Alcocer-Yamanaka, V.H., Tzatchkov, V.G., Arreguin-Cortes, F.I., 2012. Modeling of drinking water distribution networks using stochastic demand. Water Resour. Manag. 26, 1779–1792. https://doi.org/10.1007/s11269-012-9979-2

Alvisi, S., Ansaloni, N., Franchini, M., 2016. Comparison of parametric and nonparametric disaggregation models for the top-down generation of water demand time series. Civ. Eng. Environ. Syst. 33, 3–21. https://doi.org/10.1080/10286608.2015.1126823

Alvisi, S., Ansaloni, N., Franchini, M., 2014. Generation of synthetic water demand time series at different temporal and spatial aggregation levels. Urban Water J. 11, 297–310. https://doi.org/10.1080/1573062X.2013.801499

Blokker, E.J.M., 2010. Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks. Water Management Academic Press, Delft. https://doi.org/10.1017/CBO9781107415324.004

Blokker, E.J.M., Beverloo, H., Vogelaar, A.J., Vreeburg, J.H.G., Van Dijk, J.C., 2011. A bottom-up approach of stochastic demand allocation in a hydraulic network model: A sensitivity study of model parameters. J. Hydroinformatics 13, 714–728. https://doi.org/10.2166/hydro.2011.067

Blokker, M., Agudelo-vera, C., Moerman, A., van Thienen, P., Pieterse-Quirijns, I., 2017. Review of applications of SIMDEUM, a stochastic drinking water demand model with small temporal and spatial scale. Drink. Water Eng. Sci. Discuss. 10, 1–15. https://doi.org/10.5194/dwes-2017-4

Bras, R.L., Rodrìguez-Iturbe, I., 1993. Random functions and hydrology. Dover Publications, New York.

Brentan, B.M., Meirelles, G.L., Manzi, D., Luvizotto, E., 2018. Water demand time series generation for distribution network modeling and water demand forecasting. Urban Water J. 15, 150–158. https://doi.org/10.1080/1573062X.2018.1424211

Buchberger, S.G., Wu, L., 1995. Model for instantaneous residential water demands. J. Hydraul. Eng. 121, 232–246. https://doi.org/10.1061/(ASCE)0733-9429(1995)121:3(232)

Deidda, R., Badas, M.G., Piga, E., 2004. Space-time scaling in high-intensity Tropical Ocean Global Atmosphere Coupled Ocean-Atmosphere Response Experiment (TOGA-COARE) storms. Water Resour. Res. 40, W02506. https://doi.org/10.1029/2003WR002574

Filion, Y., Adams, B., Karney, B., 2007. Cross correlation of demands in water distribution network design. J. Water Resour. Plan. Manag. 133, 137–144. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:2(137)

Filion, Y.R., Karney, B.W., Adams, B.J., 2005. Stochasticity of demand and probabilistic performance of water networks, in: Impacts of Global Climate Change: Proceedings of World Water and Environmental Resources Congress 2005. ASCEE, Anchorage, Alaska, pp. 1–12. https://doi.org/10.1061/40792(173)49

Fiorillo, D., Creaco, E., Paola, F.D., Giugni, M., 2020. Comparison of bottom-up and top-down procedures for water demand reconstruction. Water 12, 922. https://doi.org/10.3390/w12030922

Gargano, R., Di Palma, F., de Marinis, G., Granata, F., Greco, R., 2016. A stochastic approach for the water demand of residential end users. Urban Water J. 13, 569–582. https://doi.org/10.1080/1573062X.2015.1011666

Kapelan, Z.S., Savic, D.A., Walters, G.A., 2005. Multiobjective design of water distribution systems under uncertainty. Water Resour. Res. 41, W11407. https://doi.org/10.1029/2004WR003787

Koutsoyiannis, D., Manetas, A., 1996. Simple disaggregation by accurate adjusting procedures. Water Resour. Res. 32, 2105–2117. https://doi.org/10.1029/96WR00488

Kumar, D.N., Lall, U., Petersen, M.R., 2000. Multisite disaggregation of monthly to daily streamflow. Water Resour. Res. 36, 1823–1833. https://doi.org/10.1029/2000WR900049

Lall, U., Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32, 679–693. https://doi.org/10.1029/95WR02966

Mamade, A., Loureiro, D., Alegre, H., Covas, D., 2018. Top-down and bottom-up approaches for water-energy balance in Portuguese supply systems. Water 10, 577. https://doi.org/10.3390/w10050577

Moughton, L.J., Buchberger, S.G., Boccelli, D.L., Filion, Y.R., Karney, B.W., 2007. Effect of time step and data aggregation on cross correlation of residential demands, in: 8th Annual Water Distribution Systems Analysis Symposium 2006. ASCEE, Cincinnati, Ohio (USA), pp. 1–11. https://doi.org/10.1061/40941(247)42

Nelsen, R.B., 1999. An Introduction to Copulas. Springer, New York. https://doi.org/10.1007/978-1-4757-3076-0

Nowak, K., Prairie, J., Rajagopalan, B., Lall, U., 2010. A nonparametric stochastic approach for multisite disaggregation of annual to daily streamflow. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008530

Parsons, J., Rees, P., Sim, P., Mcdonald, A., 2007. Working Paper 07 / 1 MACROWater : a Top-down , Policy-driven Model for Forecasting Domestic Water Demand. Leeds (UK).

Sheng, Z., Jia, S., Michelsen, A., Abudu, S., 2017. Lessons learned from long term water resources plans : top town vs. bottom up, in: International Water Resources Association (IWRA), XVI World Water Congress “Bridging Science and Policy.” Cancun, Quintana Roo, Mexico, 1-8.

Walski, T.M., Chase, D. V., Savic, D.A., Grayman, W., Koelle, E., 2003. Advanced Water

Distribution Modeling and Management, First. ed, Methods. Haestad Methods, Waterbury, CT (USA).

4 Identification of optimal smart meters locations for water demand pattern reconstruction

The WDNs management is typically and mainly oriented toward the minimization of water waste and energy consumption associated with the supply of water to the end-users (Brentan et al., 2017a; Fox et al., 2009; Gato-Trinidad and Gan, 2011). Consistently with this objective, smart metering can be used to improve WDNs management. In the past only few coarse measurements were available for a limited number of locations within the network. In recent years, smart meters are being widely installed in WDNs all over the world, providing automatized readings of users' consumption.

One of the objectives of smart metering is to raise awareness of water consumption at the individual user level. Different studies (Blokker, 2010; Britton et al., 2013; Kossieris et al., 2014; Makropoulos et al., 2014) showed that observing and analysing water consumption at a fine temporal scale can generate virtuous behaviours in end-users, thus resulting in consumption reductions and water resource savings.

Smart metering can also be used by water utilities to improve the management of WDNs through the districtualization. Analysing water consumption pattern in a generic DMA can help water utility managers in identifying the occurrence of anomalous events in non-revenue water, which includes authorized unbilled consumption (such as that used for firefighting) and water losses, and, in turn, help in planning suitable interventions. The water utility company can monitor inflows and outflows from the DMA using flowmeters at the boundaries to infer the total water consumption in the DMA, which is the sum of the revenue water (users' consumption) and the non-revenue water. Then, anomalous events can be identified by observing the non-revenue water, calculated as the difference between the total consumption of the DMA and the revenue water. In this framework, although the temporal pattern of the total consumption of the DMA is obtained by applying a water balance on inflows and outflows at the DMA boundaries, a methodology is required to reconstruct the temporal pattern of the revenue water. If all users are provided with smart meters, the pattern of the total revenue water (hereinafter indicated as the total demand of the DMA) can be easily calculated by summing the single readings at any temporal resolution. However, installing a smart meter at each user location may be exceedingly expensive for water utilities. Therefore, the question arises as to whether the number of monitored households can be limited to a small but still representative sample.

This chapter is directly based on publication IV.

Several authors provided statistical methods to identify the representative subset of users to ensure statistical reliability in their investigations of indoor and outdoor water uses (DeOreo, 2011; DeOreo et al., 2016; Mayer et al., 2009), as well as high consumption users aiming at reducing both water and energy consumption (Abdallah and Rosenberg, 2014). However, no literature deals with the identification of a small subset of users to be monitored for an accurate reconstruction of the temporal pattern of the DMA demand. In this chapter two procedures to face this issue are presented. The first procedure can be used to choose new meter installation sites in a DMA where smart meters at the end of life were previously installed at all locations. The second procedure can be applied starting from users' billed annual demand to a DMA in which no smart meters are present. Both procedures were tested against the Soccavo DMA. Finally, in the framework of this chapter, an effective strategy for the optimal allocation of smart meters for the accurate reconstruction of the total district demand pattern is provided.

4.1 Methodology

In this work two procedures aimed at optimizing smart meters allocation in two different contexts were developed.

The first procedure (*Procedure 1*) can be used in a DMA where smart meters at the end of life were previously installed at all user locations. It enables the assessment of the minimum number of smart meters to be replaced, namely the minimum number of users to be monitored, for reliable demand characterization starting from the available measured demand time series. The possibility of replacing a number of smart meters lower than the initial one results in an evident cost saving for municipal water companies.

The second procedure (*Procedure 2*) can be applied to a DMA where smart meters must be installed for the first time. It aims at identifying the users to be monitored starting from users' billed annual demand.

The following sections describe in detail the procedures.

4.1.1 Identification of the representative users based on measured demand

Procedure 1 is intended to identify a representative subset of users to be monitored for an accurate reconstruction of the district water demand pattern. The procedure is summarised in Figure 17.

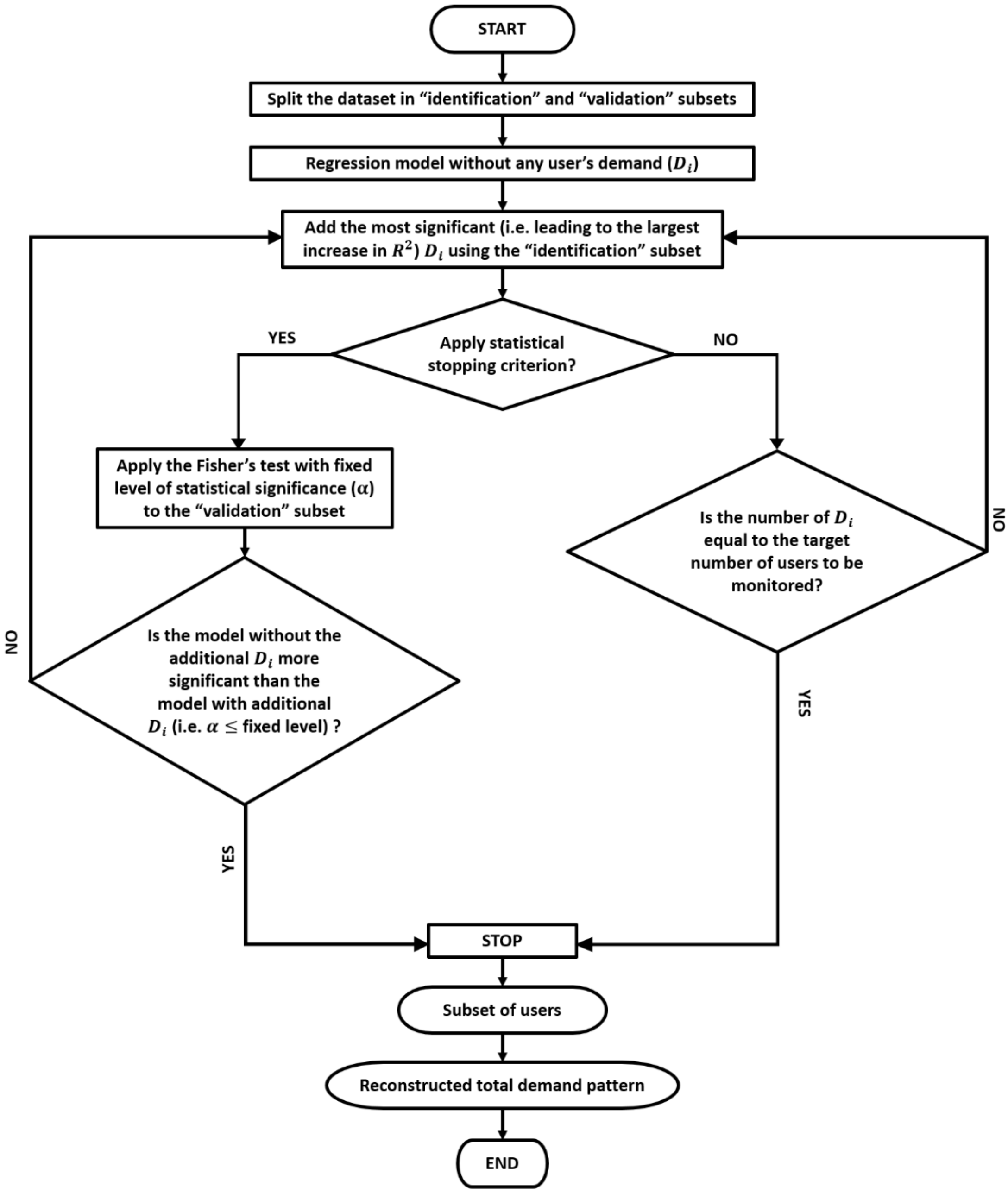

Figure 17 – Flow-chart of Procedure 1 (Fiorillo et al., 2020).

Determining a representative sample of users can be basically brought back to a variable selection problem that can be addressed by linear regression. Therefore, assuming N_{data} time steps available in the measured dataset, the following linear model structure was considered in *Procedure 1*:

$$\widehat{D}_{tot,k} = \alpha_1 D_{1,k} + \alpha_2 D_{2,k} + \ldots + \alpha_n D_{n,k} \tag{17}$$

where $\widehat{D}_{tot,k}$ is the total estimated demand of the DMA (dependent variable or output) at the generic time k ($k = 1,2, \ldots, N_{data}$), $D_{i,k}$ is the measured demand (regressor or input) of the generic $i - th$ of the n users included in the model at the generic time k, and $\alpha_i \geq 0$ is the regressor coefficient.

The coefficients $\alpha_i \geq 0$ must be estimated to maximize the fit of $\widehat{D}_{tot,k}$ to the measured total demand $D_{tot,k} = \sum_{i=1}^{N_{tot}} D_{i,k}$, being N_{tot} the total number of users in the DMA, by applying the Constrained Least Square (CLS) identification (James et al., 2013). The goodness of the fit can be estimated as a function of the sum of the square residuals $\sum_{k=1}^{N_{data}}\left(\widehat{D}_{tot,k} - D_{tot,k}\right)^2$(SSR). The lower this sum, the better the fit. This results in:

$$\widehat{\boldsymbol{\alpha}} = argmin\, SSR(\boldsymbol{\alpha}) \quad subject\ to\ \boldsymbol{\alpha} \geq 0 \tag{18}$$

where $\boldsymbol{\alpha}$ is the vector of the coefficients α_i.

Alternatively, the goodness of the fit can be assessed through the coefficient of determination R^2, expressed as follow:

$$R^2 = \frac{\sum_{k=1}^{N_{data}}\left(\widehat{D}_{tot,k} - \overline{D}_{tot}\right)^2}{\sum_{k=1}^{N_{data}}\left(D_{tot,k} - \overline{D}_{tot}\right)^2} = 1 - \frac{\sum_{k=1}^{N_{data}}\left(\widehat{D}_{tot,k} - D_{tot,k}\right)^2}{\sum_{k=1}^{N_{data}}\left(D_{tot,k} - \overline{D}_{tot}\right)^2} \tag{19}$$

where $\overline{D}_{tot}$ is the mean value of $D_{tot,k}$.

At this stage some remarks about the model structure, expressed by Equation 17, should be made. If $n = N_{tot}$, $\alpha_i = 1$ for all users, therefore $D_{tot,k}$ can be easily calculated through Equation 17, resulting in the perfect fit (R^2=1). When $n < N_{tot}$, each α_i , representing the influence of the $i-th$ user on the total DMA demand, must be estimated. In this work the forward stepwise regression (James et al., 2013) was used for determining the most influential users. The algorithm can be explained as follows. Let us indicate the model structure in Equation 17 as M_n because of the n present regressors. The algorithm starts at step 0 with no regressor in the model (M_0). Then, among the N_{tot} potentially available regressors the most beneficial regressor – i.e. leading to the largest increase in R^2 (or the largest decrease in SSR) – is inserted into the model, obtaining the model M_1. At step 1, M_1 is the starting point and, among the $N_{tot} - 1$ potentially available regressors, the most beneficial regressor is inserted into the model. At the generic step n, model M_n is considered as starting point and among the $N_{tot} - n$ potentially available regressors the most beneficial one is inserted in the model. The algorithm goes on until a stopping criterion is satisfied.

In this work, the total set of N_{data} data was split into two subsets, the identification subset, used for input identification and coefficient estimation, and the validation subset. The identification subset was made up of the odd days of the year (i.e. 1, 3, 5, …), while the validation subset was made up of the even days of the year (i.e. 2, 4, 6, 8, …). This approach for dividing the dataset has been chosen to calibrate (and hence validate) the model on the data from the whole year rather than from one season or few months. Indeed, as described in section 4.2, in this work the available dataset was related to only one year. In this case, dividing the dataset based on odd and even days of the year allows to easily and effectively include all months and each day of the week (from Monday to Sunday) in both the identification and validation subset. For the sake of comparison with the results of *Procedure 2*, the R^2 was also computed on the whole dataset.

Then, two different stopping criteria were used, statistical and economic criterion, respectively. The statistical criterion consists of the Fisher's test (Lehman, 1986) based on a fixed level of statistical significance α performed on the validation subset. In this work, $\alpha = 0.95$ was set. The economic criterion is based on the total cost allocated by the water utility to smart metering. The total cost for smart meters allocation impacts on the maximum number of installable smart meters. In other words, according to the economic criterion the maximum number of smart meters is set by the water utility based on the available economic budget. In this work, in order to explore the benefits in terms of goodness of the fit, in stepwise regression different model orders (i.e. number of users to be monitored) were considered. Based on economic considerations, $n = 100$ was chosen as the maximum number of smart meters, assuming that the replacement of 100 smart meters could be a worthwhile investment for the water company. Finally, the most suitable sample of users to characterize the total demand of the district is obtained.

It is worth noting that the water company can easily obtain any kind of information or statistic to improve the selection of the best number of smart meters to be installed according to the operating conditions of the DMA and its specific needs. Indeed, the procedure provides the synthetical pattern of the total district demand from which it can be easily calculated any useful statistic, such as the mean and maximum percentage errors, by referring to the measured demand. Calculating such statistics for different numbers of users selected can help water utility managers to choose the best solution.

4.1.2 Identification of the representative users based on billed demand

For a preassigned number of users to be monitored (n), *Procedure 2* allows to identify which user should be supplied with a smart meter to obtain a reliable estimate of the temporal pattern of the total district demand.

This procedure can be used in DMAs where smart metering needs to be implemented for the first time. Indeed, unlike *Procedure 1*, this procedure can be applied in the absence of metering data. The flow-chart in Figure 18 summarises the procedure.

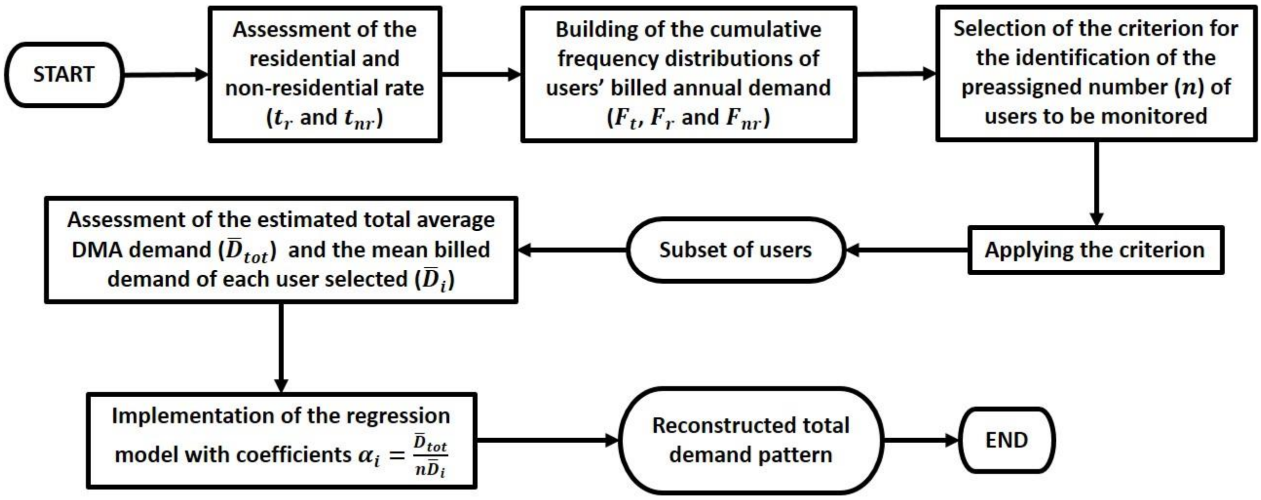

Figure 18 – Flow-chart of Procedure 2 (modified from Fiorillo et al., 2020).

First, in order to choose the n users to be supplied with a smart meter, *Procedure 2* uses criteria based on users' billed annual demand and consumption type (residential or non-residential). In what follows, it is assumed that in the analysed DMA there are a rate t_r and a number $N_{tot,r} = t_r N_{tot}$ of residential users, and a rate t_{nr} and a number $N_{tot,nr} = t_{nr} N_{tot}$ of non-residential users. The criteria are based on the cumulative frequency distributions of users' total (F_t), residential (F_r) and non-residential (F_{nr}) billed annual demand, that can be easily obtained by water utilities. Table 8 reports the criteria for the identification of n representative users. The list of criteria includes fully random selection (Criterion 1), criteria that consider only users' position in the cumulative distributions (Criteria 2, 3 and 4) and criteria that consider both users' type and position in the cumulative distributions (Criteria 5, 6 and 7). More specifically, Criteria 2 and 3 are probabilistic, in that they are based on random sampling from the cumulative frequency distributions, while privileging the users with a larger billed consumption. Criterion 4, instead, is deterministic, in that it selects the n users with the largest consumption. Criteria 5, 6 and 7 are the extension of criteria 2, 3 and 4, respectively, distinguishing users' type for the selection.

Regarding the probabilistic criteria, a preliminary analysis was carried out to test different criteria. Thus, different percentages of the n users to be selected in different intervals of the cumulative frequency distribution of the users' total demand F_t were investigated. On one hand, it was found that selecting the n users among those characterized by high values of F_t (i.e. $F_t > 0.5$) resulted in good performances. On the other hand, using several intervals for the selection of as many percentages of the n users was considered difficult to apply. Basically, the criteria could be reduced in two simpler criteria. The first is based on the selection of the n users among those with $F_t > 0.5$ (Criterion 3). The second consists of sampling the majority of the n users (e.g. the 80%) among those with the highest position in the cumulative distributions (e.g. $F_t > 0.75$), such as in Criterion 2. These observations were combined with considerations on the types of users to define Criteria 5 and 6.

Table 8 - Criteria for the selection of the users to be monitored for a reliable total district demand characterization. The criteria refer to the cumulative frequency distributions of users' total (F_t), residential (F_r) and non-residential (F_{nr}) billed annual demand to select n, n_r and n_{nr} users among all, residential and non-residential users respectively.

Criterion	**Description**
1	Random selection of n users
2	Selection of 20% of n among the users characterized by F_t ranging from 0.5 to 0.75 and 80% of n among users with $F_t > 0.75$
3	Selection of n among users with $F_t > 0.5$
4	Selection of the n users with the highest values of the mean daily demand among all users
5	Selection of 20% of n_r among the users characterized by F_r ranging from 0.5 to 0.75 and 80% of n_r among users with $F_r > 0.75$, selection of the remaining n_{nr} applying the same criterion referring to the F_{nr}
6	Selection of n_r among users with $F_r > 0.5$, selection of the remaining n_{nr} applying the same criterion referring to the F_{nr}
7	Selection of the n_r users with the highest values of the mean daily demand among the residential users, selection of the remaining n_{nr} applying the same criterion referring to the non-residential users

In this work, the criteria for users' selection were applied assuming different numbers n of users to be selected from 10 to 100. Excluding the deterministic criteria 4 and 7, the selection of the users was repeated several times for each probabilistic criterion in order to obtain reliable averaged results. More specifically, the number of times (n_{times}) to repeat the selection was assessed by using the following formula for the determination of the size of statistical samples valid in the case of large or unknown population (Cochran, 1953):

$$n_{times} = \frac{Z^2 pq}{e^2} \tag{20}$$

where Z^2 is the abscissa of the normal curve that cuts off an area α at the tails (1 - α equals the desired confidence level), e is the desired level of precision, p is the estimated proportion of an attribute that is present in the population, and q is $1-p$.

In the present study, confidence level equal to 0.95 ($Z = 1.96$), $\pm$ 5% precision ($e = 0.5$) and the maximum variability ($p = 0.5$) were considered, resulting in $n_{times} = 384$.

Then, the model structure of *Procedure 1* (Equation 17) was also adopted in *Procedure 2*. However, denoting with $\overline{D}_i$ the mean demand of the $i-th$ user as derived from billed consumption, in *Procedure 2* the coefficients α_i are evaluated by scaling down the estimated total average DMA demand $\overline{D}_{tot}$ to $\overline{D}_i$:

$$\alpha_i = \frac{\overline{D}_{tot}}{n\overline{D}_i} \geq 0 \tag{21}$$

By time averaging Equation 17 with coefficients evaluated with Equation 21, it can be shown that the total district demand is respected on average through *Procedure 2*. Therefore, in *Procedure 2*, Equation 17 can be written as follows:

$$\frac{\widehat{D}_{tot,k}}{\overline{D}_{tot}} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{D_{i,k}}{\overline{D}_i} \tag{22}$$

Thus, the reconstructed pattern results in the average of the regressors' patterns.

It should be noted that in Equation 22, as was mentioned above, $\overline{D}_{tot}$ and $\overline{D}_i$ can be derived from the billed annual demands, without the need of measured demand data. Therefore, the water company can easily calibrate the model by referring to each user's billed annual water consumption. In case of a DMA where some meters are not easily accessible for the replacement with smart meters, the water company should apply the procedure by neglecting such meters. In this case, the procedure would effectively select a subset of the most influential users for a reliable estimate of the total district demand among the available users. However, it is assumed that the water companies are interested in implementing smart metering systems in DMAs where the majority of the meters can be easily replaced by smart meters. Otherwise, economic and operating benefits would be very limited.

Furthermore, in the case of any malfunctioning smart meters, the model can be easily recalibrated, by excluding these smart meters from the model and by calculating coefficients α_i with Equation 21 on the remaining meters. In the same way an easy recalibration of the model can be made also in the event of variations in the number of users, i.e. if new customers are added to the DMA. However, regardless of the method, in the case of a new customer at least the estimated annual consumption should be available in order to apply this procedure.

4.2 Case study

Both procedures were applied to the Soccavo DMA. Hourly consumption data from both residential and non-residential meters, from 20 March 2017 to 19 March 2018, were used. Even though the pilot area comprises 4989 smart meters, only the demand time series from 1406 smart water meters, including 1067 residential flow meters and 339 non-residential flow meters, were considered in the calculations. The others were discarded because the measured demand patterns were considered unreliable, due to the presence of irregularities in the data possibly caused by malfunctioning smart meters.

The daily temporal patterns of the mean aggregated hourly demand are shown in Figure 19, while Figure 20 shows the cumulative distribution frequencies of the mean daily demand of the users. Summing up the results of both figures, most of the total district consumption consists of residential consumption. However, the non-residential users are characterized by the highest values of consumption.

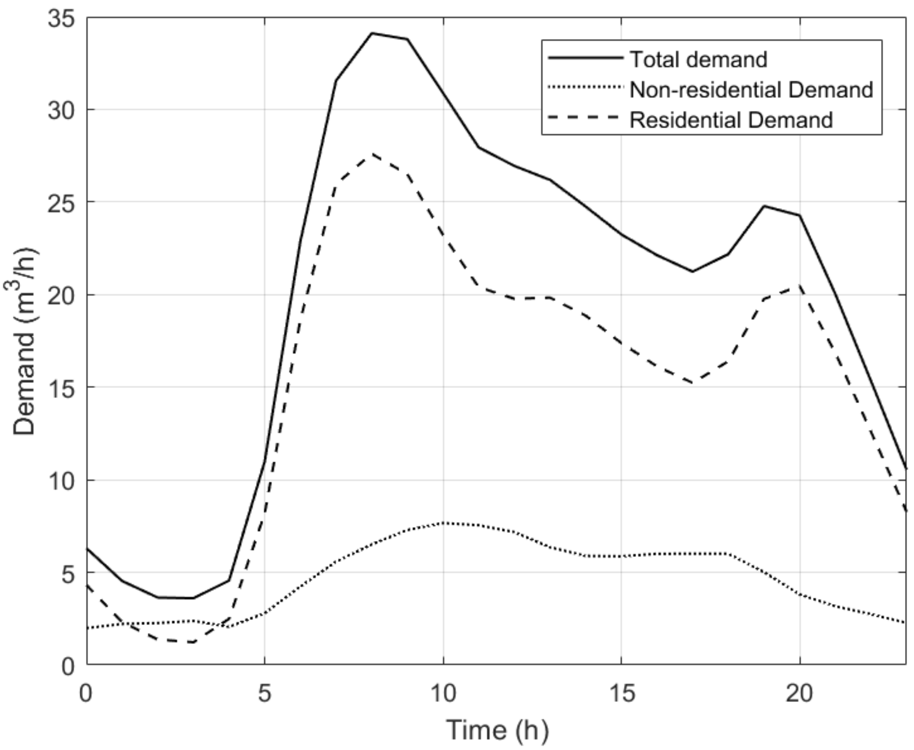

Figure 19 - Daily temporal patterns of total aggregated hourly demand and aggregated hourly demand of residential and non-residential users, respectively (Fiorillo et al., 2020).

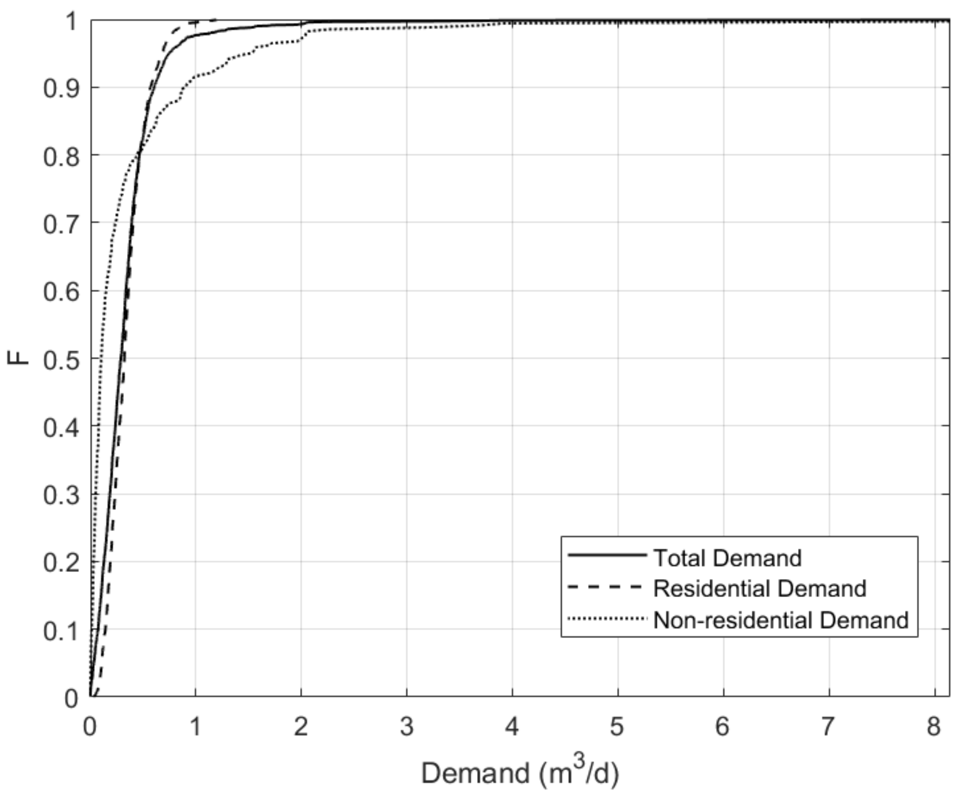

Figure 20 - Cumulative frequency distributions (F) of users' total, residential, and non-residential mean daily demand (Fiorillo et al., 2020).

4.3 Results and discussion

The following sections present and discuss the results obtained by applying both procedures to the case study. First the results of *Procedure 1* applied by using statistical and economic stopping criterion are discussed. Then, the results of each criterion of *Procedure 2* are presented. For each time and for each criterion, the comparison between the measured and estimated total district demand pattern is assessed in terms of R^2, referring to the mean of the $n_{times} = 384$ values obtained for each criterion.

4.3.1 Procedure 1 - Results

Applying *Procedure 1* to the case study by using as stopping rule the Fisher's test on the validation subset, the lowest number of users to be monitored for a reliable characterization of the total water demand of the DMA resulted to be 25. By way of example, Table 9 reports the values of the coefficients α_i obtained by applying the stepwise regression and the Fisher's test as stopping criterion. Each coefficient indicates how much the consumption of the generic $i - th$ selected user affects the total demand of the DMA in case of 25 users selected (i.e. in case of model order equal to 25).

Table 9 - Values of the coefficients α_i obtained by applying the stepwise regression using as stopping criterion the Fisher's test.

Model order	Regressor index	α_i
1	1405	9.71
2	176	24.46
3	1404	7.35
4	1063	28.90
5	1335	22.13
6	571	28.65
7	1047	22.65
8	279	21.30
9	726	26.72
10	41	23.39
11	1349	31.60
12	1403	10.36
13	336	45.60
14	842	16.38
15	694	19.22
16	1120	19.81
17	878	20.42
18	848	39.56
19	387	21.40
20	605	29.10
21	407	21.29
22	337	25.37
23	845	30.74
24	668	39.55
25	78	28.66

Figure 21 reports the results in terms of R^2 for different orders of the model, referring to the identification, validation and test datasets. Even for low model order, i.e. 20-40 users selected, the fit is good for identification, validation and test datasets. Then, the fit becomes excellent in case of high number of users selected, such as 100. Therefore, in the event that the water company wants to install less than 100 smart meters (i.e. the maximum number of smart meters set in this work), *Procedure 1* can be effectively used by setting a lower value for the maximum number of installable smart meters. The procedure provides a list of as many locations as the maximum number of installable smart meters set by the water company. Furthermore, since at each step of the forward stepwise regression the most beneficial regressor among all those available is inserted in the model (i.e. in the list of locations where the smart meter should be replaced), the users selected are automatically ranked from the most to the list influential user. Therefore, for example, each model order shown in Table 9 coincides with the rank of the user inserted in the model at the corresponding step.

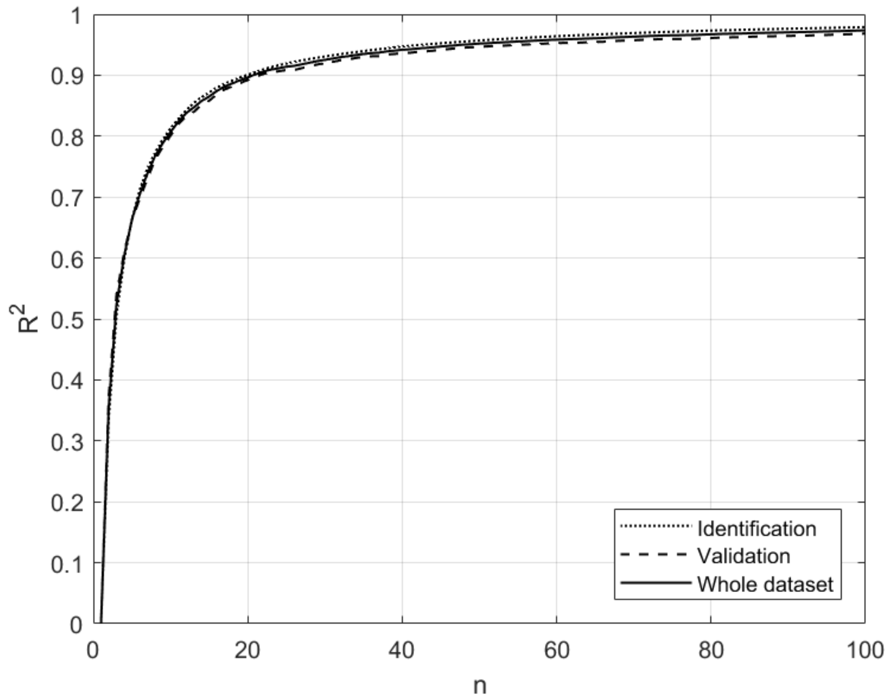

Figure 21 - Procedure 1 - Results in terms of coefficient of determination (R^2) obtained by the stepwise regression for the dataset of identification, validation and the whole dataset for different orders of the model (n) (Fiorillo et al., 2020).

In Table 10 the results obtained for model order respectively equal to 25, 50, 75 and 100 are reported. The values of R^2 do not significantly differ between identification and validation datasets, attesting to the robustness of *Procedure 1*. It is worth noting that for all model orders considered most of the selected users were non-residential and characterised by the highest values of consumption. This demonstrated that results are mostly affected by the highest consumptions.

Table 10 - Results in terms of coefficient of determination (R^2) obtained by the stepwise regression.

Model order	**R^2**		
	Identification	**Validation**	**Whole dataset**
25	0.92	0.91	0.91
50	0.96	0.95	0.95
75	0.97	0.96	0.96
100	0.98	0.97	0.97

4.3.2 Procedure 2 - Results

Figure 22 shows the comparison between the criteria of *Procedure 2* in terms of R^2 for each number of users selected (n). Criterion 7 was found to be the *Best Criterion*, closely followed by Criterion 4 in the case of high numbers of selected users, i.e. 90 – 100 users. It is worth noting that Criterion 4 and 7 are the only two deterministic criteria among all those of *Procedure 2*. Then, it can be stated that, in the case of the installation of a suitable number of smart meters (from 90 to 100), choosing the users with the highest values of the mean daily demand is a suitable trade-off between the need of an accurate characterization of the total district demand and smart-metering costs. However, the additional information about the type of users, included in Criterion 7, resulted in better results.

At this stage, it must be remarked that for the probabilistic criteria only the minimal performance can be ensured, i.e. the minimum value of R^2 obtained for each criterion, which is smaller than the mean value shown in Figure 22. Instead, the use of a deterministic criterion, such as the *Best Criterion*, does not cause such uncertainty since the users' selection is univocal.

Except for Criterion 1, which did not consider the mean daily demand of the users in the phase of users' selection, all the criteria resulted in a good fit even if the number of the selected users was not high (i.e. 50 users). Therefore, privileging the selection of users with high values of consumption seems to be beneficial in terms of goodness of the fit.

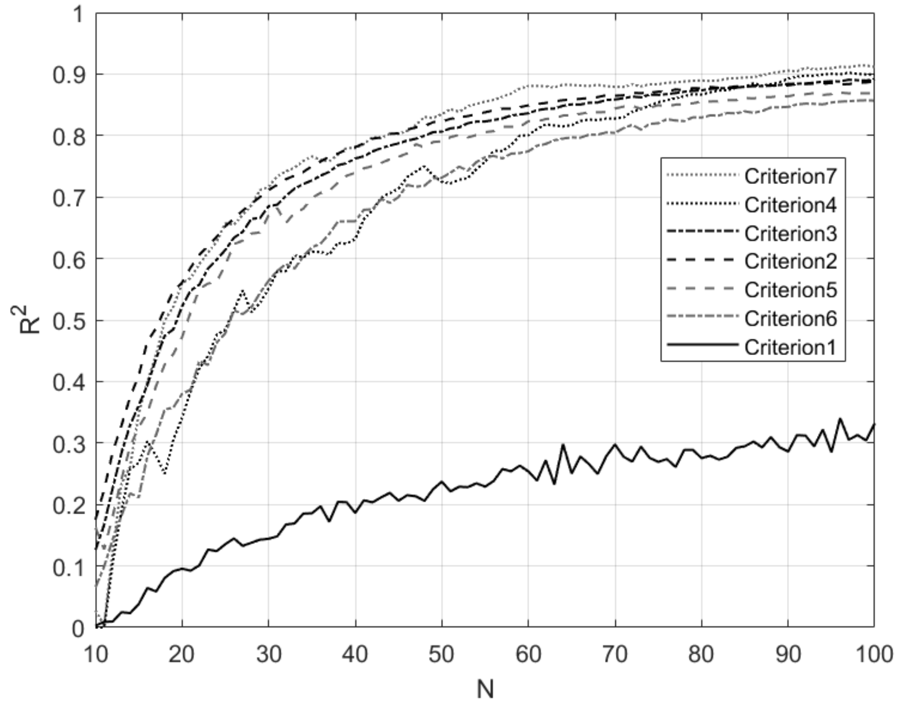

Figure 22 - Procedure 2 - Comparison of the coefficient of determination (R^2) obtained using different criteria for each number of selected users (n) (Fiorillo et al., 2020).

Figure 23 shows the weekly temporal patterns of measured aggregated hourly demands and synthetical aggregated hourly demands obtained through *Procedure 1* and the *Best Criterion* from *Procedure 2* in case of 25 selected users. *Procedure 1* is more accurate than *Procedure 2-Best Criterion*, since this latter does not appear to be able to reproduce demand peaks.

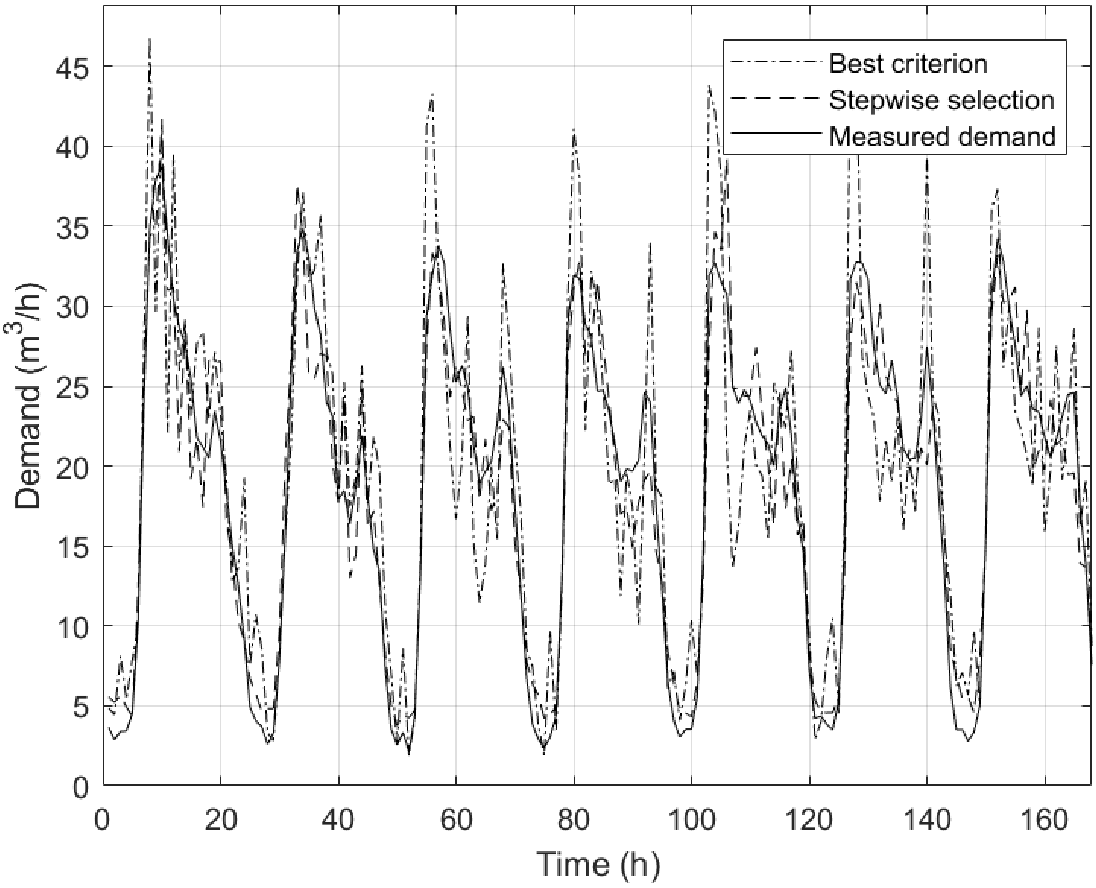

Figure 23 - Weekly temporal patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by the selection of 25 users through the application of Procedure 1 and the Best Criterion from Procedure 2 (Fiorillo et al., 2020).

Figure 24 reports the weekly patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by using the *Best Criterion* for the selection of 25, 50 and 100 users.

The goodness of the fit obtained applying *Procedure 2-Best Criterion* improves with the increasing number of users. Indeed, in the case of 100 selected users, *Procedure 2-Best Criterion* reaches a high level of accuracy. This is particularly remarkable, since *Procedure 2* is parameterized based on users' billed annual consumption, without using any measured demand time series.

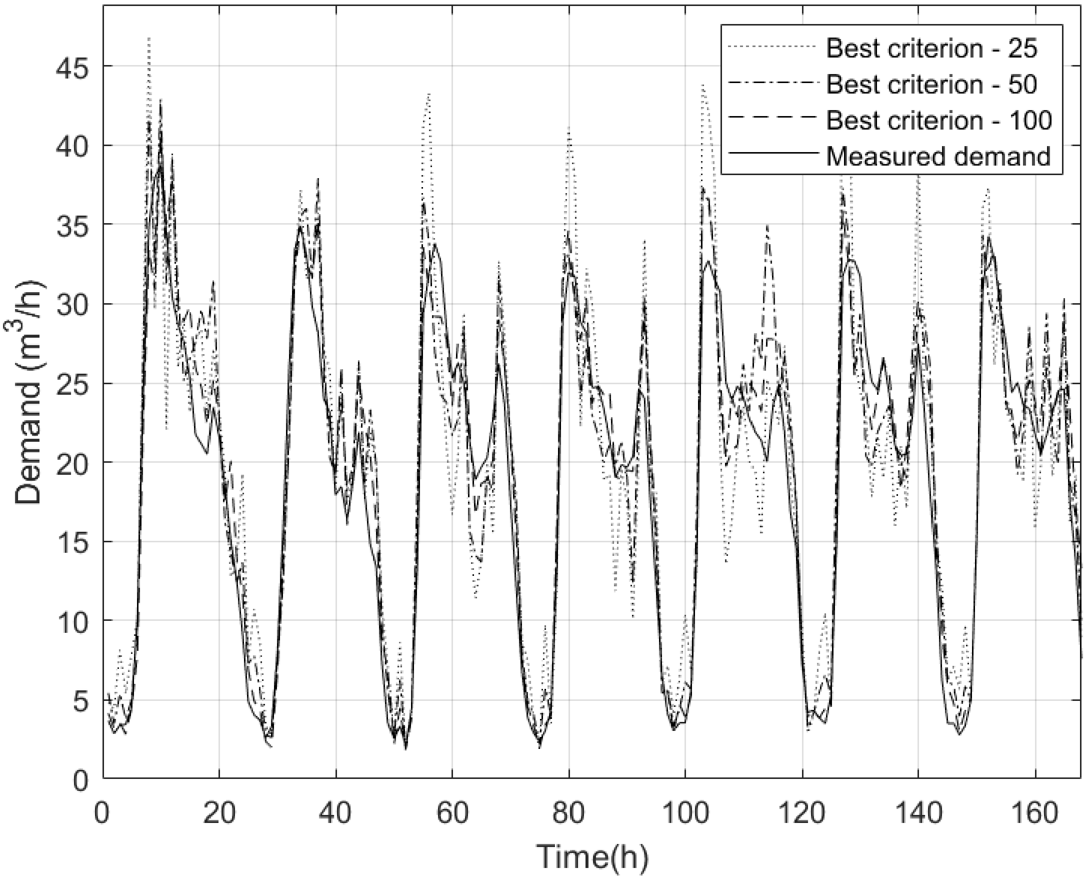

Figure 24 - Weekly temporal patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by the selection of 25, 50 and 100 users through the application of the Best Criterion from Procedure 2 (Fiorillo et al., 2020).

In order to better visualize the differences for the maximum and the minimum consumption hours, Figure 25a shows the daily temporal patterns of measured aggregated hourly demands and synthetical aggregated hourly demands obtained through *Procedure 1* and the *Best Criterion* for different number of users selected. Both procedures tend to overestimate the peak demand. However, *Procedure 1* shows slightly better performances in reproducing peak demand. In both cases the performances for the minimum consumption hours are good, especially in case of higher number of users selected.

Overall, considering the low numbers of users selected compared to the total number of users (1406), the performances of both the procedures are satisfactory. Figure 25b reports the distribution of the residuals for aggregated hourly demands, obtained by applying the *Best Criterion* for the selection of 25, 50 and 100 users. The highest residuals in absolute value are shown in case of 25 users selected. Overall, the residuals are randomly distributed indicating that the linear model provided a good fit to the data. This demonstrates the effectiveness of the regression model used in *Procedure 2*.

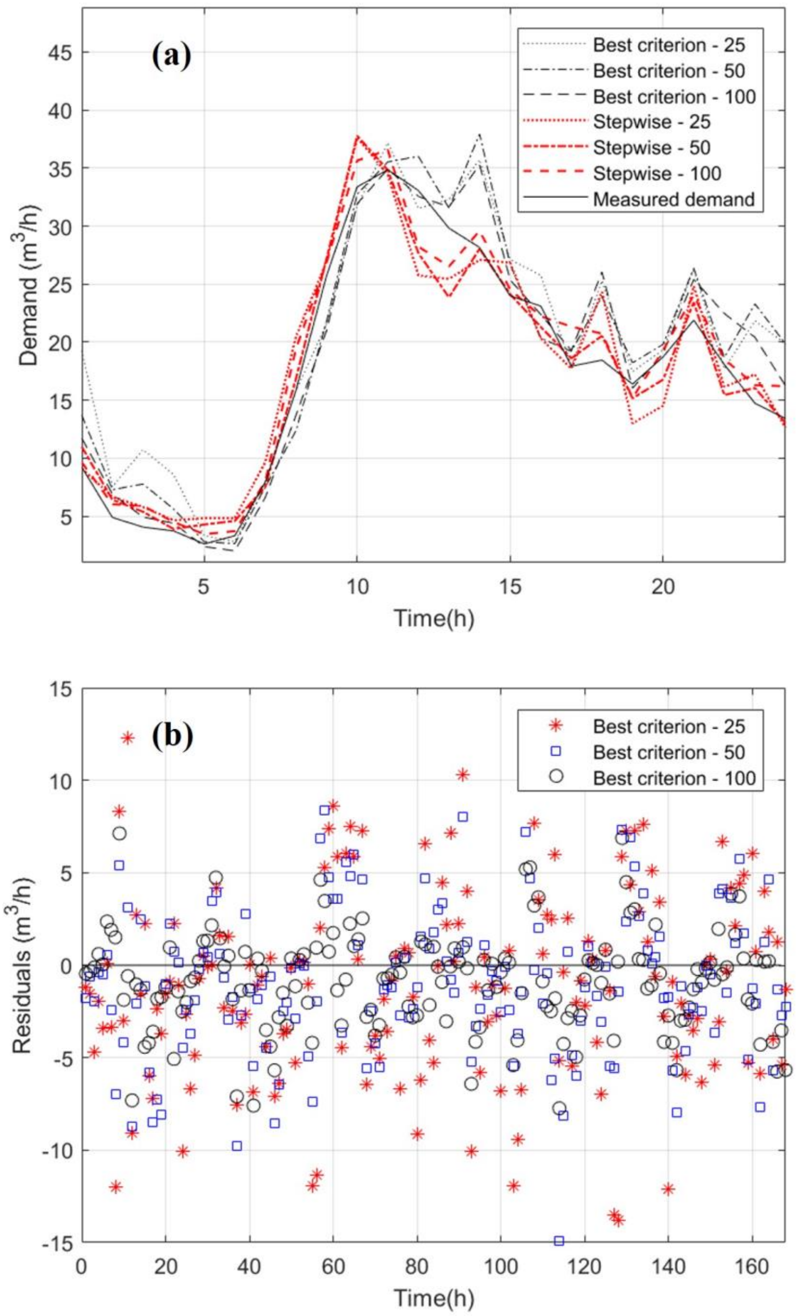

Figure 25 - Daily temporal patterns for measured aggregated hourly demands and for aggregated hourly demands obtained by the selection of 25, 50 and 100 users through the application of Procedure 1 and Best Criterion from Procedure 2 (a) and residuals for aggregated hourly demands obtained by the selection of 25, 50 and 100 users through the application of the Best Criterion from Procedure 2 (modified from Fiorillo et al., 2020).

Finally, Table 11 reports the results in terms of R^2 obtained by applying *Procedure 1* and *Procedure 2 – Best Criterion* for different numbers of users selected. For high numbers of selected users, i.e. 100, *Procedure 2-Best Criterion* enables achieving almost the same results as *Procedure 1*, which is calibrated using a significantly larger amount of data (i.e. the measured demand time series). Furthermore, it can be stated that the R^2 grows quickly as the number n of selected users increases.

Table 11 - Results in terms of coefficient of determination (R^2) obtained by the application of the stepwise regression and the Best Criterion for different numbers of users selected.

Number of users selected	**R^2**	
	Stepwise Regression	**Best Criterion**
25	0.91	0.66
50	0.95	0.83
100	0.97	0.91

4.4 Summary and conclusions

In this chapter two procedures for identifying suitable sites for smart meters installation in a DMA are presented. By using historical measured data, the first procedure (*Procedure 1*) enables the identification of the suitable sites to obtain a reliable prediction of the temporal pattern of the total district demand. Therefore, *Procedure 1* can be applied in the case of a DMA where smart meters present at all user locations have reached the end of their service life, i.e. 10-15 years after their installation. Its use helps in avoiding re-installation of smart meters at all sites. *Procedure 1* consists of the application of the stepwise regression for the identification of the users to be monitored, while using a statistical approach, based on Fisher's test, or an economic approach as stopping criterion. The second procedure proposed in this chapter (*Procedure 2*) allows the identification of the users to be monitored starting from the billed annual consumption of each user in the DMA. Therefore, *Procedure 2* can be successfully applied during the initial phase of smart metering implementation. *Procedure 2* is based on the application of seven different criteria for users' selection, which consider both billed annual demand and type of users (residential or non-residential). Then, *Procedure 2* uses a novel linear model to estimate the temporal pattern of the total demand in the DMA. This model is parameterized based on the billed annual demands of the users. Thus, it can be easily calibrated by the water utility.

The application to the Soccavo DMA proved the effectiveness of both procedures as the number of selected users to be monitored increases. When this number reached the 7% of the total number of users in the DMA, the fit was very accurate ($R^2 = 0.97$ for *Procedure 1*). As for *Procedure 2*, the best criterion for users' selection resulted the deterministic one, which identified the users with the highest values of billed consumption while distinguishing between residential and non-residential users.

Both the presented procedures can be successfully applied to WDNs, allowing the water company to find an adequate trade-off between the need of an accurate characterization of the total demand in a DMA and smart metering costs.

Overall, the applications revealed that an effective strategy for the allocation of smart meters for an accurate reconstruction of the total district demand pattern consists of selecting the users with the highest annual consumptions while distinguishing them based on their typology (i.e. residential and non-residential users).

It is important to note that even if this study provides general criteria for the allocation of smart meters, the results could change in terms of both type and number of users selected according to the typology of the DMA. However, in case of residential areas, the percentage of users selected on the total number of users in the DMA (i.e. about 7% for the highest accuracy) should remain almost unchanged.

This work provides strategies for the allocation of smart meters in order to obtain an accurate characterization of the total demand with the lowest number of meters, resulting in an innovative tool in water management. Therefore, this study can contribute to expand previous researches based on readings acquired in few locations of water systems, aimed at forecasting water consumption (Kozlowski et al., 2018; Zhou et al., 2000) reproducing water demand time series (Alvisi et al., 2014; Creaco et al., 2017a; Di Nardo et al., 2018; Gargano et al., 2016; Kossieris et al., 2019) and characterising peak water demands (Balacco et al., 2019; Bougadis et al., 2005; Gargano et al., 2017; Tricarico et al., 2007). Furthermore, it must be noted that such models as that used in this work (Equation 17) can be used for detection of anomalous events (i.e. unexpected consumption, pipe bursts and so forth), whether used on real time. Indeed, the anomalous event detection can be performed through application of statistical methodologies on the non-revenue water, estimated as the difference between the total consumption (obtained by the flowmeter typically present at the DMA entrance) and the revenue water (estimated by the models).

Though the main objective of this study is the reconstruction of the total demand of a DMA to facilitate anomalous event detection, it also has other potential applications. Specifically, after the total demand $\widehat{D}_{tot,k}$ has been reconstructed, the demand at unmetered users, hence even in case of meters locations that are not easily accessible for the replacement with smart meters, can be defined as follows. First, its total value can be calculated as $\widehat{D}_{tot,k} - \left(D_{1,k} + D_{2,k} + \ldots + D_{n,k}\right)$, i.e. by subtracting from $\widehat{D}_{tot,k}$ the total metered demand. Then, this quantity can be allocated to the unmetered users applying any criterion, e.g. proportionally to the average nodal demand. As a result, a realistic set of nodal demands can be constructed as input to WDN models for the real time modelling of WDNs. Indeed, the use of realistic input is a prerequisite for obtaining reliable results from real time WDN modelling (Creaco et al., 2017b). Furthermore, a reliable and realistic WDN model is very beneficial for the real time regulation of the devices (De Paola et al., 2017a, 2017b, 2016; Galuppini et al., 2019) and it facilitates assessing the stability and robustness of control algorithms (Galuppini et al., 2020).

References

Abdallah, A.M., Rosenberg, D.E., 2014. Heterogeneous residential water and energy linkages and implications for conservation and management. J. Water Resour. Plan. Manag. 140, 288–297. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000340

Alvisi, S., Ansaloni, N., Franchini, M., 2014. Generation of synthetic water demand time series at different temporal and spatial aggregation levels. Urban Water J. 11, 297–310. https://doi.org/10.1080/1573062X.2013.801499

Balacco, G., Gioia, A., Iacobellis, V., Piccinni, A.F., 2019. At-site assessment of a regional design criterium for water-demand peak factor evaluation. Water 11. https://doi.org/10.3390/w11010024

Blokker, E.J.M., 2010. Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks. Water Management Academic Press, Delft. https://doi.org/10.1017/CBO9781107415324.004

Bougadis, J., Adamowski, K., Diduch, R., 2005. Short-term municipal water demand forecasting. Hydrol. Process. https://doi.org/10.1002/hyp.5763

Brentan, B.M., Luvizotto, E., Herrera, M., Izquierdo, J., Pérez-García, R., 2017. Hybrid regression model for near real-time urban water demand forecasting. J. Comput. Appl. Math. 309, 532–541. https://doi.org/10.1016/j.cam.2016.02.009

Britton, T.C., Stewart, R.A., O'Halloran, K.R., 2013. Smart metering: Enabler for rapid and effective post meter leakage identification and water loss management. J. Clean. Prod. 54, 166–176. https://doi.org/10.1016/j.jclepro.2013.05.018

Cochran, W.G., 1953. Samling Techniques. John Wiley & SoNs, Inc., New York.

Creaco, E., Blokker, M., Buchberger, S., 2017a. Models for generating household water demand pulses: Literature review and comparison. J. Water Resour. Plan. Manag. 143, 04017013. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000763

Creaco, E., Pezzinga, G., Savic, D., 2017b. On the choice of the demand and hydraulic modeling approach to WDN real-time simulation. Water Resour. Res. 53, 6159–6177. https://doi.org/10.1002/2016WR020104

De Paola, F., Galdiero, E., Giugni, M., 2017a. Location and setting of valves in water distribution networks using a harmony search approach. J. Water Resour. Plan. Manag. 143, 04017015. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000760

De Paola, F., Galdiero, E., Giugni, M., 2016. A jazz-based approach for optimal setting of pressure reducing valves in water distribution networks. Eng. Optim. 48, 727–739.

https://doi.org/10.1080/0305215X.2015.1042476

De Paola, F., Giugni, M., Portolano, D., 2017b. Pressure Management Through Optimal Location and Setting of Valves in Water Distribution Networks Using a Music-Inspired Approach. Water Resour. Manag. 31, 1517–1533. https://doi.org/10.1007/s11269-017-1592-y

DeOreo, W., 2011. Analysis of water use in new single family homes. Aquacraft Water Engineering and Management, Boulder, Colorado (USA).

DeOreo, W.B., Mayer, P., Kiefer, J., 2016. Residential End Uses of Water, Version 2: Executive Report., Water Research Foundation.

Di Nardo, A., Di Natale, M., Gargano, R., Giudicianni, C., Greco, R., Santonastaso, G.F., 2018. Performance of partitioned water distribution networks under spatial-temporal variability of water demand. Environ. Model. Softw. 101, Pages 128-136. https://doi.org/10.1016/j.envsoft.2017.12.020

Fiorillo, D., Galuppini, G., Creaco, E., De Paola, F., Giugni, M., 2020. Identification of Influential User Locations for Smart Meter Installation to Reconstruct the Urban Demand Pattern. J. Water Resour. Plan. Manag. 146, 1–9. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001269

Fox, C., McIntosh, B.S., Jeffrey, P., 2009. Classifying households for water demand forecasting using physical property characteristics. Land use policy 26, 558–568. https://doi.org/10.1016/j.landusepol.2008.08.004

Galuppini, G., Creaco, E., Toffanin, C., Magni, L., 2019. Service pressure regulation in water distribution networks. Control Eng. Pract. 86, 70–84. https://doi.org/10.1016/j.conengprac.2019.03.007

Galuppini, G., Magni, L., Creaco, E., 2020. Stability and Robustness of Real-Time Pressure Control in Water Distribution Systems. J. Hydraul. Eng. 146, 04020023. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001722

Gargano, R., Di Palma, F., de Marinis, G., Granata, F., Greco, R., 2016. A stochastic approach for the water demand of residential end users. Urban Water J. 13, 569–582. https://doi.org/10.1080/1573062X.2015.1011666

Gargano, R., Tricarico, C., Granata, F., Santopietro, S., de Marinis, G., 2017. Probabilistic models for the peak residential water demand. Water 9, 417. https://doi.org/10.3390/w9060417

Gato-Trinidad, S., Gan, K., 2011. Characterizing maximum residential water demand. WIT Trans. Built Environ. 122, 15–24. https://doi.org/10.2495/UW120021

James, G., Witten, D., Hastie, T., Tibishirani, R., 2013. An Introduction to Statistical Learning with Applications in R (older version), Springer Texts in Statistics.

Kossieris, P., Kozanis, S., Hashmi, A., Katsiri, E., Vamvakeridou-Lyroudia, L.S., Farmani, R., Makropoulos, C., Savic, D., 2014. A web-based platform for water efficient households. Procedia Eng. 89, 1128–1135. https://doi.org/10.1016/j.proeng.2014.11.234

Kossieris, P., Tsoukalas, I., Makropoulos, C., Savic, D., 2019. Simulating marginal and dependence behaviour of water demand processes at any fine time scale. Water 11, 855. https://doi.org/10.3390/w11050885

Kozlowski, E., Kowalska, B., Kowalski, D., Mazurkiewicz, D., 2018. Kozłowski_2018.pdf. Arch. Civ. Mech. Engeneering 18, 140–148. https://doi.org/http://dx.doi.org/10.1016/j.acme.2017.05.006

Lehman, E.L., 1986. Testing Statistical Hypotheses, Second. ed. Springer, New York. https://doi.org/10.1007 /978-1-4757-1923-9

Makropoulos, C., Kossieris, P., Kozanis, S., Katsiri, E., Vamvakeridou-Lyroudia, L., 2014. From smart meters to smart decisions: web-based support for the water efficient household, in: 11th International Conference on Hydroinformatics. CUNY Academic Works, New York, pp. 1–8.

Mayer, P., DeOreo, W., Hayden, M., Davis, R., Caldwell, E., Miller, T., Bickel, P.J., 2009. Evaluation of california weather-based “smart” irrigation controller programs. Aquacraft, Water Engineering and Management.

Tricarico, C., De Marinis, G., Gargano, R., Leopardi, A., 2007. Peak residential water demand. Proc. Inst. Civ. Eng. Water Manag. 160, 115–121. https://doi.org/10.1680/wama.2007.160.2.115

Zhou, S.L., McMahon, T.A., Walton, A., Lewis, J., 2000. Forecasting daily urban water demand: A case study of Melbourne. J. Hydrol. 236, 153–164. https://doi.org/10.1016/S0022-1694(00)00287-0

5 Water demand forecasting by using weather data

The impact of the climate on water supply is a major concern in Italy. The climate influence on water demand can cause problems in terms of water shortages and energy waste (Colombo and Karney, 2003).

In the context of water demand modelling, several studies have used weather variables to explain demand variability (Adamowski, 2008; Ashoori et al., 2016; Beal and Stewart, 2014; Cole and Stewart, 2013; Dos Santos and Pereira Filho, 2014; Gato et al., 2007; Goodchild, 2003; Haque et al., 2017; Manouseli et al., 2019; Miaou, 1990; Papageorgiou et al., 2016; Slavíková et al., 2013; Toth et al., 2018; Willis et al., 2013; Xenochristou et al., 2018, 2020a, 2020b; Xenochristou and Kapelan, 2020; Zhou et al., 2000). Furthermore, different studies showed the importance of accounting for additional explanatory factors, such as socio-demographic variables and types of users, in the characterization of water demand patterns (Laspidou et al., 2015; Mamade et al., 2014; Parker, 2013; Xenochristou et al., 2020a, 2020b).

Considering that water utilities do not always have access to water consumption data with high temporal and spatial resolution, it is important to develop alternative strategies to investigate water demand by using different information, such as weather data and socio-economic characteristics of the users (Xenochristou, 2019). Furthermore, since daily and weekly fluctuations in weather variables are highly associated with changes in water demand (Adamowski, 2008; Chang et al., 2014; Zhou et al., 2000), climate changes are also expected to influence water demand (Parandvash and Chang, 2016).

The studies focusing on climate change impacts on water demand (Babel et al., 2014; Downing et al., 2003; Jampanil et al., 2012; Kanakoudis et al., 2017; Neale et al., 2007; Parker, 2013; Polebitski et al., 2011; Rasifaghihi et al., 2020; Zachariadis, 2010; Zubaidi et al., 2020) showed that the likely variations in water demand due to climate change would vary widely with geographic location and climatic conditions (Wang et al., 2014). Changes in water demand could affect the existing water systems in terms of capacity and operation. Specifically, increases in water demand can cause imbalance in water resources and problems in storage capacity, worsening the situation of water shortages that Mediterranean countries, such as Italy, are already experiencing (La Jeunesse et al., 2016). Despite the clear benefits, few studies have investigated the impacts of weather changes on

This chapter is directly based on publication V.

water demand in Italy, focusing on climate change effects on agricultural water demand (Bocchiola, 2013; Masia et al., 2018) and water supply (Peres et al., 2019).

This study investigated the possibility of using weather variables to forecast water demand and its likely variations due to climate change for the case study of Soccavo, with the aim of improving the understanding of weather effects on urban water demand in Italy. The water demand forecasting accuracy of Random Forests models (RFs) based on weather variables was analysed. The analysis was carried out by disaggregating water consumption based on the social characteristics of the users. The models based on weather variables were able to forecast aggregated water demand, allowing to catch the variations in water demand due to climate variability. However, the effectiveness of the models changed depending on the social characteristics of the users, highlighting the relevance of disaggregating consumption to both determine the influence of weather on water demand and improve forecasting models.

5.1 Data

In this work an extensive dataset consisting of water consumption and social characteristics belonging to the users of the Soccavo DMA, as well as weather data, was used.

In the following subsections, first an overview of the consumption data and social characteristics of the users is presented, followed by a description of the weather data.

5.1.1 Consumption data and characteristics of the users

In this work, 1067 residential flow meters from the Soccavo DMA were considered. More specifically, the hourly consumption data running from 20 March 2017 to 19 March 2018 were used. In order to analyse daily water demand for each user, the hourly data were aggregated at the daily scale.

In addition, the available information about the social characteristics of the users of the DMA were used. According to the data provided by Istat (Italian National Institute of Statistics) for each census section included in the DMA the following characteristics were assessed:

- the average level of employment of the inhabitants;
- the average educational level of the inhabitants.

Then, the social characteristics of each household of the DMA were determined based on the related census section. Finally, each household was classified on the basis of state of employment and educational level as shown in Table 12.

Table 12 - Classification of the households of the DMA according to their main social characteristics.

State of Employment	Educational level
Employed	*Primary/Secondary school degree*
Unemployed/stay-at-home	*High school/university degree*

5.1.2 Weather data

In this study six weather variables were initially considered: daily maximum air temperature, rainfall amount, rainfall rate, solar radiation, wind speed and air humidity. Recorded weather data over the same period as for the water consumption data (20 March 2017 - 19 March 2018) were used. The humidity data were recorded at daily intervals by the Italian Air Force weather station of Capodichino. The other variables were collected at 30 minutes intervals by the weather station of the University of Naples Federico II. These data were also aggregated at the daily scale. Both the weather stations were chosen due to their proximity to the DMA.

In order to choose suitable predictors of water demand, a preliminary analysis was carried out, reported in the Appendix. Specifically, the correlation between weather variables and water demand was investigated, as explained in the Appendix. The highest degrees of correlation with water demand were observed for the daily maximum air temperature and the daily mean solar radiation. For the other variables only weak relationships with water demand were found (see the Appendix). Therefore, among the weather variables only daily maximum temperature and daily mean solar radiation were considered in the forecasting models.

As shown in Figure 26, the highest values of daily maximum temperature occurred in summer, with an average of almost 30°C. With regards to the daily mean solar radiation, the highest values were recorded in both spring and summer, with an average of approximately 270 W/m^2.

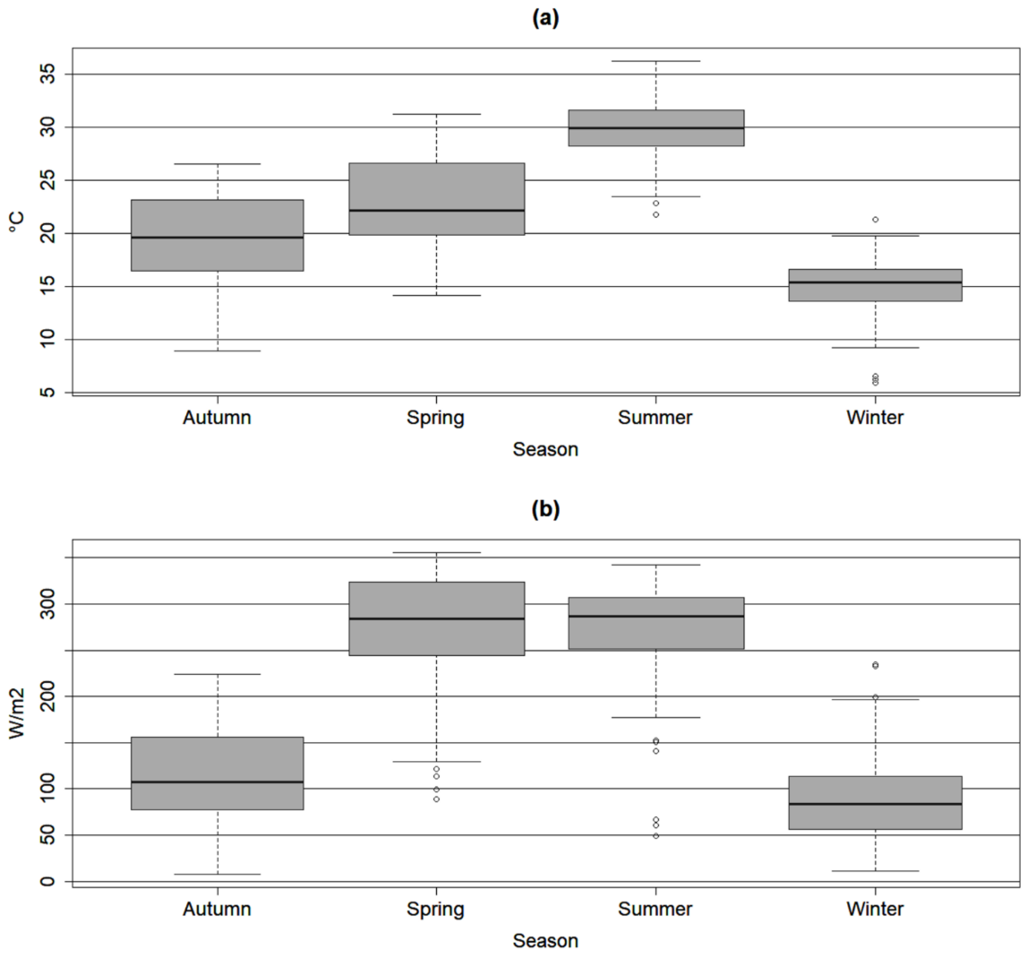

Figure 26 - Boxplot of seasonal measured values of daily maximum temperature (a) and daily mean solar radiation (b).

5.2 Methodology

This section presents a new methodology to characterize water demand. The methodology is based on 3 different configurations of RFs, aiming at assessing the forecasting accuracy of different weather variables, as well as their combined effect.

RFs were chosen since they have proven to perform better than other machine learning techniques (Chen et al., 2017). In recent years, ensemble methods, such as RFs, have been found the most successful models among machine learning techniques (see Chapter 1). Basically, they combine individual weak learners to create a strong learner. Furthermore, RFs can be easily trained since they have a limited number of parameters to be tuned.

The RFs and the methodology are illustrated in detail in the following.

5.2.1 Random Forests

In this study, regression RFs were implemented. RFs are data driven models consisting of an ensemble of decision trees. Such models can be used for both classification and regression, for categorical and continuous response variable, respectively (Cutler et al., 2012).

According to the RF regression, from the training data $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where $x_i = (x_{i,1}, \dots, x_{i,p})^T$represents the p predictors and y_i denotes the response, for the generic tree j ($j = 1, 2, \dots, n_t$) a bootstrap sample D_j of size N is taken from D (Breiman, 2001). Then, the tree is fitted by using D_j as training data and applying the binary recursive partitioning (Cutler et al., 2012). Specifically, starting with all observations in a single node, for each un-split node, m predictors among the p available predictors are randomly selected. The node is then split into two descendant nodes using the best binary split among all binary splits on the m predictors. In the regression context, the mean squared residual at the node is usually used as a splitting criterion. The algorithm goes on until a stopping criterion is satisfied, i.e. when the tree has reached the maximum allowed depth. All the resulting trees are finally combined by averaging their responses.

Therefore, the prediction at the new point x is made as follows:

$$\hat{y}(x) = \frac{1}{n_t} \sum_{j=1}^{n_t} \widehat{h_j}(x) \tag{23}$$

where $\widehat{h_j}(x)$ is the prediction of the response variable at x using the $j-th$ tree.

In the RF models, the tree are the base learners of the ensemble predictor $\hat{y}(x)$. Basically, RFs allow merging together the predictions of multiple decision trees to get a prediction more accurate and stable than the one provided by individual decision trees. The strength of the model lies principally in implementing the randomness in the modelling process, as the variable for splitting each node is selected among a random sample of independent variables (Herrera et al., 2010).

The performance of RFs depends on three hyperparameters, i.e. the parameters whose values are fixed before the learning process:

- the number of predictors randomly selected at each node (m);
- the number of trees (n_t);
- the minimum size of terminal nodes (n_d).

The hyperparameters influence the structure of machine learning models, determining how closely the model fit on the training data. In case of underfitting, namely fitting too loosely, the model does not learn how to represent the patterns in the data. On the other hand, in case of overfitting (i.e. fitting to closely) the model learns from the noise in the training dataset, resulting in a poor prediction on the test dataset. For example, in RFs small values of m can increase the randomness of the trees, creating trees that are less similar to each other, while values of m equal to the total number of input variable (maximum value of m) may optimize the split of the node along all possible directions (Scornet, 2017). Overall, there is no theoretical framework for choosing the optimum values of the hyperparameters (Scornet, 2017). Therefore, RFs must be fine-tuned to select the optimum set of hyperparameters (m, n_d, n_t) that minimizes errors while avoiding overfitting.

The architecture of the RF regression is shown in Figure 27.

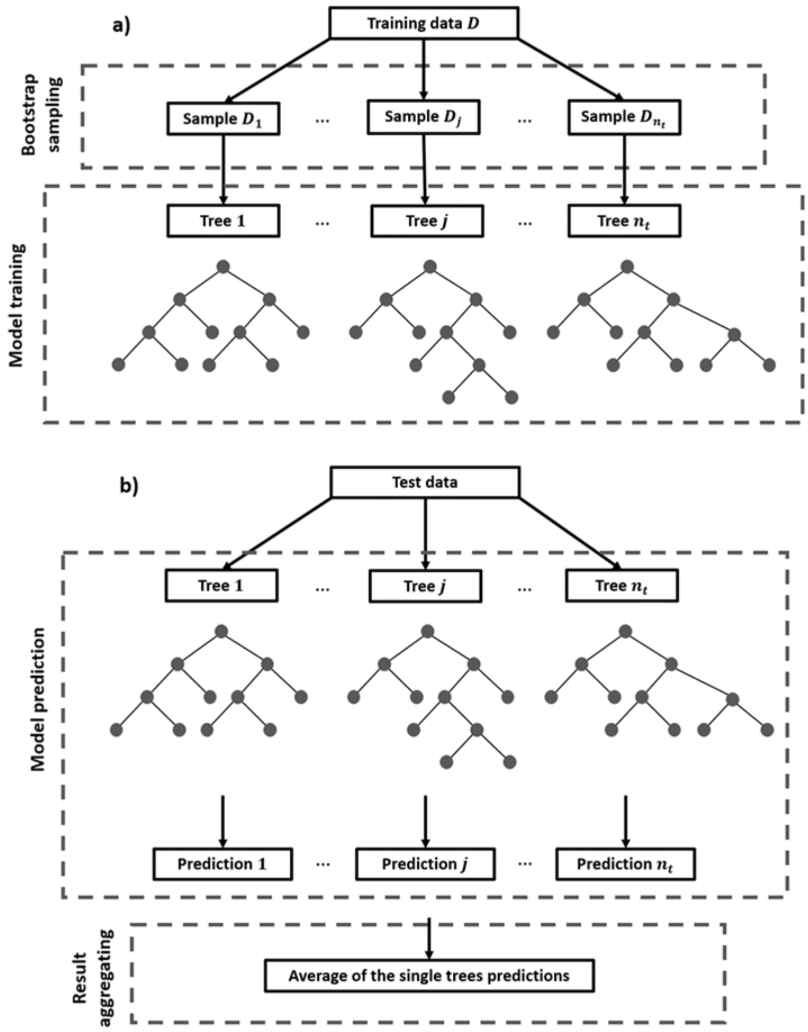

Figure 27 - Training (a) and prediction (b) phase of Random Forest regression.

5.2.2 Weather-based predictive models

The RF model was used to predict daily water demand at aggregated scale, i.e. the demand obtained by summing the daily demand of each user of the DMA. The available dataset was split in calibration subset, made up of the odd lines, and validation subset, made up of the even lines. The model was trained using the calibration dataset, whereas the validation dataset was used to evaluate the model performance on unseen data.

The accuracy was assessed using the Root Mean Square Error (RMSE) and the coefficient of determination (R^2):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{24}$$

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{25}$$

where y_i and $\hat{y}_i$ are the observed and forecasted values respectively, $\bar{y}$ is the mean value of the observed values y_i and N is the number of observations.

These metrics were chosen because they provide different information. The RMSE is sensitive to outliers and large errors, which are particularly undesirable in water demand forecasting. On the other hand, R^2 indicates the amount of variance explained by the model.

Since past consumption is not always available for water utilities it is important to explore an alternative strategy for water demand forecasting. Furthermore, past consumptions can masque the effect of other predictors (that are characteristic of the user or the days the consumptions belong to), as they can carry the same information (Xenochristou, 2019). For these reasons, 3 configurations of the RF model were developed to investigate the performances of weather variables as predictors, on the basis of the results of the preliminary analysis showed in the Appendix. The configurations are shown in Table 13. The first configuration of the model (Model 1) accounts for the combined effect of temperature and solar radiation, whereas Models 2 and 3 investigate the influence of temperature and solar radiation respectively. Specifically, Models 2 and 3 were developed to reveal the influence of each weather variable without being concealed by overlapping information (see the Appendix). Temporal characteristics – i.e. type of day (working day or holiday), season, month and weekday - were considered in all configurations since they are always reliable and easily accessible to network operators.

Table 13 - Explanatory (input) variables of each model.

Models\Variables	Temporal characteristics	Daily maximum temperature	Daily mean solar radiation
Model 1	x	x	x
Model 2	x	x	
Model 3	x		x

In order to investigate the influence of weather on different user types, the models were applied to forecast the aggregated demand of three groups of households that were identified according to the available information about the social characteristics of the users (i.e. state of employment and educational level). The description of the groups is reported in Table 14. The groups differ in employment state and educational level of the residents. *Group 1* consists of households where members are on average employed with a high average educational level (high school/university degree). Group 2 has members that are on average unemployed with primary/secondary school degree. *Group 3* is made up of households where members are on average employed with primary/secondary school degree.

It is worth noting that, according to the available information about the social characteristics of the users (i.e. state of employment and educational level), all the possible groups were identified. Further groups can be obtained by grouping the households based on one classification rather than on both state of employment and educational level (e.g. grouping together all households with employed members). However, these groups would be very heterogeneous, reducing the differences between each group and, thus, the benefit of disaggregating water consumption.

Table 14 - Description of household groups.

Group	Description	Number of households
Group 1	Employed with High school/University degree	622
Group 2	Unemployed with Primary/Secondary school degree	125
Group 3	Employed with Primary/Secondary school degree	320

In order to take into account the effect of the group size on the forecasting accuracy (Xenochristou et al., 2020a), groups with the same number of households were required. Given that Group 2 was the smallest group, samples with the same size of Group 2 (i.e. 125 households) were considered for Group 1 and Group 3. In order to both limit the calculation time and obtain representative results for each group, the number of samples was chosen proportionally to each group size. In order to select a number of samples that was proportional to the size of Group 1, 5 samples were randomly selected for this group, since it was almost 5 times greater than the samples' size. Then, the models were applied to each sample and the results in terms of RMSE and R^2 were averaged among all samples. The average results, and hence the forecasting accuracy, are expected to remain almost unchanged even for higher number of samples, since the samples selected are enough to be truly representative of all the users of the group. Similarly, the values of RMSE and R^2 for Group 3 were determined by averaging the results obtained for 3 random samples.

5.3 Results and discussion

The following sections 5.3.1 and 5.3.2 present and discuss the results obtained applying the methodology proposed in this chapter to the case study of Soccavo. First the results obtained tuning the model for the optimum set of hyperparameters are presented. Then, the results obtained applying the RF models to forecast the daily aggregated water demand of the Soccavo DMA are discussed. The accuracy of the models is assessed by referring to two metrics, the RMSE and R^2. More specifically, the RMSE values are expressed in L. However, it is worth noting that the reference time unit of these values is the day, since the daily water demand is analysed.

5.3.1 RF models tuning

The RF models were tuned over a three dimensional grid search space including multiple values of m, n_d and n_t by using the calibration dataset. The grid search space was built using five values of m for Model 1 (2, 3, 4, 5, 6) and four values of m for Model 2 and 3 (2, 3, 4, 5). Values from the default value $m = \text{total number of predictors}/3$ (~ 2 for all models) to the maximum value $m = \text{total number of predictors}$ (equal to 6 for Model 1 and 5 for Model 2 and 3) were selected for each model. For computation time reasons, six values of n_d (5, 10, 25, 50, 100, 200) and five values of n_t (50, 100, 200, 300, 500) were used for all models.

Figure 28 and Figure 29 show the values of RMSE for the calibration dataset for various combinations of the hyperparameters, for Model 1 and Model 2, respectively. At the same values of n_t, the lowest values of RMSE were obtained for low values of n_d and high values of m. Similar results were obtained for Model 3 as well. The optimum value of RMSE (i.e. the lowest value) corresponded to $m = 6$, $n_d = 5$ and $n_t = 100$ for Model 1, and $m = 5$, $n_d = 5$ and $n_t = 300$ for Model 2. However, these combinations of hyperparameters led to a too large difference in accuracy between the calibration and validation datasets, being the values of RMSE in validation almost double those obtained for the calibration dataset. This led to the conclusion that the model was overfitted. Therefore, taking into account the difference in accuracy between the calibration and validation datasets, the following combinations were selected: $m = 6$, $n_d = 10$ and $n_t = 50$ for Model 1, and $m = 3$, $n_d = 10$ and $n_t = 200$ for Model 2, $m = 3$, $n_d = 10$ and $n_t = 100$ for Model 3. As shown in Figure 28 and Figure 29, these combinations of hyperparameters resulted in good level of accuracy, i.e. low values of RMSE.

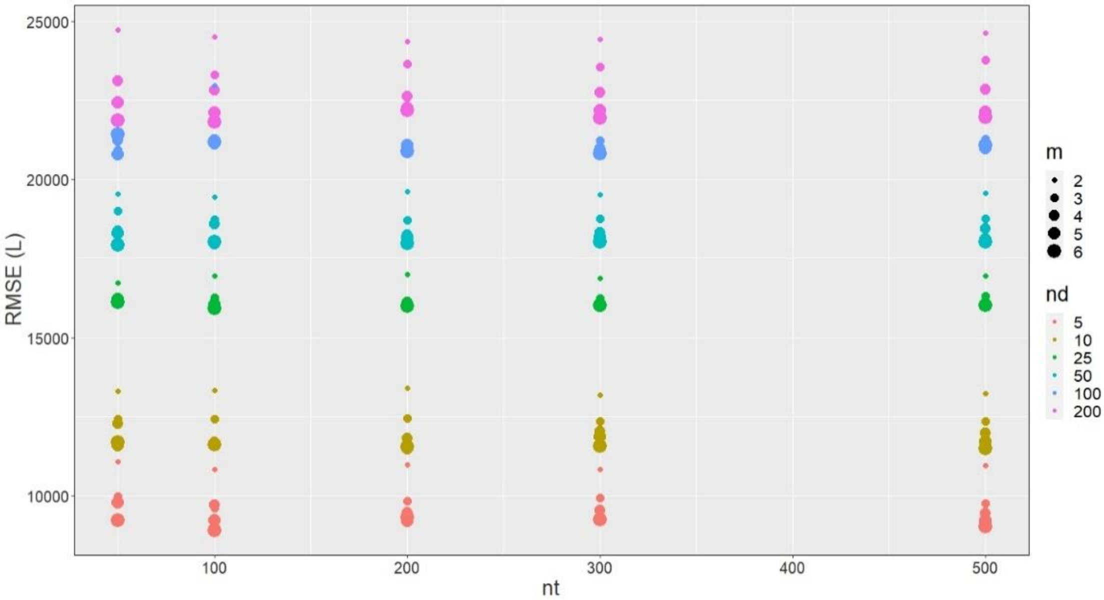

Figure 28 – Plot of RMSE values for the calibration dataset for various combinations of the hyperparameters for Model 1.

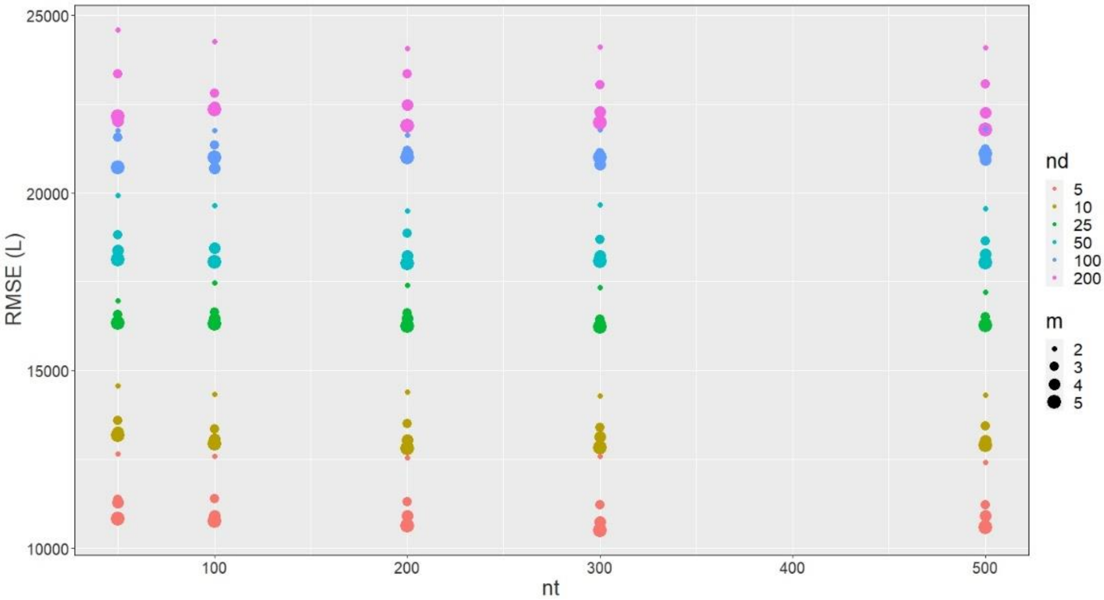

Figure 29 - Plot of RMSE values for the calibration dataset for various combinations of the hyperparameters for Model 2.

5.3.2 Prediction accuracy of RF weather-based models

First, the results obtained applying the RF models to forecast the daily aggregated water demand of the DMA are presented.

Table 15 shows the results of each model in terms of RMSE and R^2 obtained for the validation dataset. All models led to good results, showing good performances in terms of R^2. Model 1 (based on temporal characteristics, temperature and solar radiation input) resulted in the best performances, leading to the highest value of R^2 (0.67) and the lowest value of RMSE (16448 L). This demonstrated the benefit of including both the weather variables (temperature and solar radiation) as predictors. Model 3, that besides temporal characteristics included solar radiation as input data, led to slightly better performances compared to Model 2.

Table 15 - Results in terms of root mean square error (RMSE) and coefficient of determination (R^2) obtained at aggregated scale for each model for the validation dataset.

Model	**RMSE (L)**	**R^2**
Model 1	16448	0.67
Model 2	17539	0.64
Model 3	17031	0.65

For example, Figure 30 reports the comparison between measured and forecasted aggregated daily demands for the application of Models 1 and 3. Each point represents one day. The most of the points follow the bisectors of the graphs highlighting a good agreement between measured and forecasted aggregated daily demands. However, both models seem to overestimate the lowest daily demands and underestimate the highest ones.

This result can be traced back to the structure of RFs which is based on averaging among different predictions. This could lead to underpredict the highest demands and overpredict the lowest demands. Furthermore, many forecasting models usually struggle to predict outliers (Xenochristou and Kapelan, 2020). In this case, bias correction methods can be used to improve the forecasting accuracy of the peak days. Similar results were obtained for Model 2 as well.

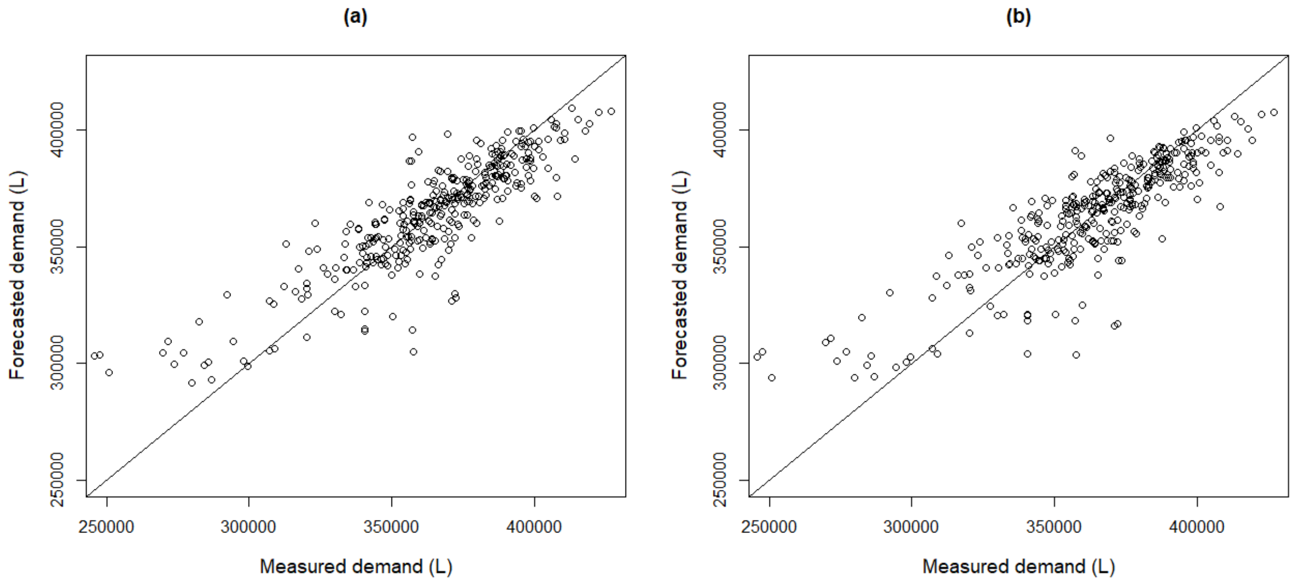

Figure 30 - Comparison between measured and forecasted aggregated daily demand for Models 1 (a) and Model 3 (b) (Fiorillo et al., 2021).

The good performances obtained for the developed models showed that weather variables can be effectively used to forecast water demands without the need of recorded consumptions, in agreement with previous studies that demonstrated the benefits of adding weather variables in forecasting models (Xenochristou, 2019; Xenochristou et al., 2018). Therefore, in the event no smart metering data are available for water utilities, weather variables could offer a reasonable alternative to forecast the total water demand of the districts, being weather data easily accessible to network operators. In addition, these models may be effectively used to estimate future demand changes based on climate change scenarios, since they were able to catch the variations due to weather.

The results obtained using the RF models to forecast the aggregated demand of each group of households are reported in Table 16. For *Group 1*, all models led to similar results. More specifically, Model 1 including both temperature and solar radiation as predictors showed slightly better performances in terms of RMSE (i.e. the lowest value of RMSE = 2937 L). Model 3, which was based only on solar radiation and temporal characteristics, also showed good level of accuracy (R^2 = 0.66 and RMSE = 2948 L). Slightly worse performances (R^2 = 0.65 and RMSE = 2985 L) were obtained using temperature and temporal characteristics as predictors (Model 2), although the level of accuracy is good. With regards to *Group 2*, the models showed the lowest performances, resulting in low values of R^2. The worst performances (R^2= 0.51 and RMSE = 3216) were obtained using only temporal characteristics and solar radiation as predictors (Model 3). For *Group 3*, the best results (RMSE = 2630 L and R^2 = 0.62) were obtained for Model 1 based on both temperature and solar radiation.

Models 2 and 3, including only temperature and solar radiation input respectively, led to reasonable levels of accuracy. Specifically, using only solar radiation input beside temporal characteristics (Model 3) slightly increased the prediction accuracy compared to Model 2 (based on temperature).

Table 16 - Results in terms of root mean square error (RMSE) and coefficient of determination (R^2) obtained at aggregated scale for the groups of households for each model.

	Group 1		**Group 2**		**Group 3**	
Model	**RMSE (L)**	**R^2**	**RMSE (L)**	**R^2**	**RMSE (L)**	**R^2**
Model 1	2937	0.66	3177	0.53	2630	0.62
Model 2	2985	0.65	3096	0.55	2713	0.59
Model 3	2948	0.66	3216	0.51	2701	0.60

Overall, the best performances were observed for *Group 1*, although good levels of prediction accuracy were obtained for *Group 3* using both solar radiation and temperature input. The models showed the lowest performances for *Group 2*. These results prove a stronger relationship between weather variables and water demand for *Group 1* and *Group 3*, meaning that employed users appear to be on average more affected by weather than the unemployed ones. Indeed, employed users spend more time outside and have more scheduled habits that can be easily affected by weather. Notably, better results were obtained for *Group 1* (consisting of users with high school/university degree), thus suggesting the educational level as a further discriminating factor in investigating not only the water uses (Hurd, 2006; Makki et al., 2013) but also the effects of weather variables on water consumptions.

The performed analysis showed that there are types of users that are more affected by weather variables, highlighting the relevance of disaggregating consumption to both determine the influence of weather on water demand and improve forecasting models. Therefore, the presented results are in agreement with previous studies that proved the importance of including the socio-economic status of the users when investigating the effects of weather on water demands (Chang et al., 2010; Domene and Saurí, 2006; Xenochristou et al., 2020a, 2020b, 2018).

In addition, these results can be used in investigating the effect of climate changes on water demand. The types of users mostly affected by weather will likely be primarily responsible of potential future variations in water demand related to climate change impacts. At the same time, the types of users less sensitive to weather will probably have less of an impact. Knowing the number of users belonging to each type will enable the water utility to easily assess if the total district water demand is expected to rise and thus avoid possible failures in water system capacity.

5.4 Summary and conclusions

In this study the use of weather variables in water demand forecasting models was investigated. The Random Forests model was used to predict the daily aggregated water demand of the Soccavo DMA. For this reason, 3 model configurations were developed to investigate the prediction accuracy of weather variables. The models based on weather variables resulted in good performance in forecasting aggregated demands. Thus demonstrating that, in the absence of past consumption data, weather variables can be effectively used to forecast the total water demand of the DMA, being weather data easily accessible to water utilities.

In addition, the analysis performed on different types of users showed that the effectiveness of the models depended on the social characteristics of the users. Thus, the results highlighted the relevance of disaggregating water consumption based on the social characteristics of the users to both determine the influence of weather on water demand and improve forecasting models. It is worth noting that the methodology was applied to a residential neighbourhood. Therefore, the results could vary in case of different district typology, such as commercial or industrial district.

Overall, the RF models based on weather variables can be effectively used to estimate future variations in residential water demand due to climate change, since they were able to catch the variations due to weather. Indeed, the weather-based RF models can be applied to predict the daily aggregated water demands by using reliable climate change scenarios. Furthermore, the investigation can be optimized by taking into account the social characteristics of the users. Indeed, the variations in water demand can be different for users with different social characteristics, since they resulted to be affected by weather to varying degrees.

It is worth noting that future variations in water demand due to climate changes could affect the existing water systems in terms of capacity and operation. Owing to likely increases in water demand, water utilities could have significant problems securing sufficient storage capacity. Therefore, it is essential to ensure optimum distribution of water supply based on the calculation of present and future demand (Patel and Katiyar, 2014; Shamsi, 2005). In addition, the sizing of treatment and pumping facilities is obviously linked to the magnitude and pattern of demand. Thus, the effectiveness of such facilities can be easily affected by unexpected variations in water demand. A reliable prediction of future water demand is beneficial from an operational viewpoint as well. For example, the effectiveness of the pump scheduling (procedure needed to minimize the energy cost of water distribution systems) depends on water demand forecasting accuracy. Therefore, knowing the future likely changes in water demand allows to improve the optimization of the pumps. A further operational issue regards the water quality. Water treatment procedures and infrastructures are sized according to water demand pattern. Therefore, if higher peak demands due to climate changes are ignored, shortfall in treatment capacity could occur.

Concluding, accounting for future variations in water demand due to climate changes is needed in order to avoid risks of supply and operational failures in water distribution systems. This need is particularly remarkable for those countries, like Italy and most of the Mediterranean countries, that are already coping with water shortages (La Jeunesse et al., 2016).

In previous studies for the Mediterranean area (Collet et al., 2015; Kanakoudis et al., 2017; Zachariadis, 2010) climate change scenarios were used for assessing future vulnerability of water resources. However, in these studies demographic statistics and past consumption trends were used to determine future water demand variations. Instead, the models developed in the framework of this study can be used to directly determine future variations in water demand due to climate change.

Few studies (Edwards and Martin, 1995; Parandvash and Chang, 2016; Parker and Wilby, 2013) have investigated spatial and temporal variations in water demand, focusing on the determination of the relationships between a set of explanatory variables and water demand. However, in the study presented in this chapter the temporal and spatial variation are taken into account in water demand forecasting.

Furthermore, compared with most literature models which use property characteristics, weather and economic data to predict residential water demand at monthly scale (Babel et al., 2014; Duerr et al., 2018; Toth et al., 2018; Williamson et al., 2002), the presented methodology provides predictions at daily scale. Thus, the methodology allows to account for temporal variability by forecasting the daily water demand. Xenochristou et al. (2021) also attempted to predict residential daily water demand by using household characteristics, temporal variables (e.g. season, month and day of the week) and weather data as predictors. In all the predictive models developed by Xenochristou et al. (2021) weather input are coupled with household characteristics. In the present study the water demand forecasting models are instead completely based on weather variables, excluding temporal characteristics, enabling a better assessment of the effectiveness of such variables in water demand forecasting. Moreover, in the study of Xenochristou et al. (2021) the weather input did not improve the accuracy of the modelling results. On the contrary, in this study the forecasting models based on weather variables attained good level of prediction accuracy. Such difference in results could also relate to geographical aspects. Indeed, Xenochristou et al. (2021) investigated a case study in UK, characterized by a mild climate and lacks of seasonal extremes. However, in this study the forecasting models were applied to a case study located in Southern Italy along the Tyrrhenian coast. The Mediterranean climate of Southern Italy is characterized by mild and rainy winters and hot and dry summers. However, the Tyrrhenian coast has a warm Mediterranean climate with long, hot and very dry summers. Such climatic characteristics can have a greater impact on water demand compared to the UK climate, especially due to the seasonal extremes, such as the peak temperatures that can be reached during late spring and summer.

Overall, the good results obtained using air temperature input in water demand forecasting models are in agreement with previous studies (Al-Zahrani and Abo-Monasar, 2015; Ashoori et al., 2016; Chang et al., 2014). In such studies the air temperature has proved an important predictor in water demand forecasting.

References

Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. J. Water Resour. Plan. Manag. 134, 119–128. https://doi.org/10.1061/(asce)0733-9496(2008)134:2(119)

Al-Zahrani, M.A., Abo-Monasar, A., 2015. Urban residential water demand prediction based on artificial neural networks and time series models. Water Resour. Manag. 29, 3651–3662. https://doi.org/10.1007/s11269-015-1021-z

Ashoori, N., Dzombak, D.A., Small, M.J., 2016. Modeling the Effects of Conservation, Demographics, Price, and Climate on Urban Water Demand in Los Angeles, California. Water Resour. Manag. 30, 5247–5262. https://doi.org/10.1007/s11269-016-1483-7

Babel, M.S., Maporn, N., Shinde, V.R., 2014. Incorporating Future Climatic and Socioeconomic Variables in Water Demand Forecasting: A Case Study in Bangkok. Water Resour. Manag. 28, 2049–2062. https://doi.org/10.1007/s11269-014-0598-y

Beal, C.D., Stewart, R.A., 2014. Identifying residential water end uses underpinning peak day and peak hour demand. J. Water Resour. Plan. Manag. 140, 04014008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000357

Bocchiola, D., 2013. Impact of potential climate change on crop yield and water footprint of rice in the Po valley of Italy. Agric. Syst. 116, 50–61. https://doi.org/10.1016/j.agsy.2015.07.009

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chang, H., Parandvash, G.H., Shandas, V., 2010. Spatial variations of single-family residential water consumption in Portland, Oregon. Urban Geogr. 31, 953–972. https://doi.org/10.2747/0272-3638.31.7.953

Chang, H., Praskievicz, S., Parandvash, H., 2014. Sensitivity of Urban Water Consumption to Weather and Climate Variability at Multiple Temporal Scales: The Case of Portland, Oregon. Int. J. Geospatial Environ. Res. 1, 1–7.

Chen, G., Long, T., Xiong, J., Bai, Y., 2017. Multiple Random Forests Modelling for Urban Water Consumption Forecasting. Water Resour. Manag. 31, 4715–4729. https://doi.org/10.1007/s11269-017-1774-7

Cole, G., Stewart, R.A., 2013. Smart meter enabled disaggregation of urban peak water demand: Precursor to effective urban water planning. Urban Water J. 10, 174–194. https://doi.org/10.1080/1573062X.2012.716446

Collet, L., Ruelland, D., Estupina, V.B., Dezetter, A., Servat, E., 2015. Water supply sustainability and adaptation strategies under anthropogenic and climatic changes of a meso-scale Mediterranean catchment. Sci. Total Environ. 536, 589–602. https://doi.org/10.1016/j.scitotenv.2015.07.093

Colombo, A.F., Karney, B.W., 2003. The labyrinth of water distribution systems: Demand, energy and climate change, in: Pumps, Electromechanical Devices and Systems Applied to Urban Water Management. Taylor & Francis, Leiden.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random Forests, in: Zhang, C., Ma, Y. (Eds.), Ensemble Machine Learning. Springer, Boston, MA, pp. 157–175. https://doi.org/10.1007/978-1-4419-9326-7

Domene, E., Saurí, D., 2006. Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona. Urban Stud. 43, 1605–1623. https://doi.org/10.1080/00420980600749969

Dos Santos, C.C., Pereira Filho, A.J., 2014. Water Demand Forecasting Model for the Metropolitan Area of São Paulo, Brazil. Water Resour. Manag. 28, 4401–4414. https://doi.org/10.1007/s11269-014-0743-7

Downing, T.E., Butterfield, R.E., Edmonds, B., Knox, J.W., Moss, S., Piper, B.S., Weatherhead, E.K., 2003. Climate Change and Demand for Water, Research Report, Stockholm Environment Institute. Oxford (UK).

Duerr, I., Merrill, H.R., Wang, C., Bai, R., Boyer, M., Dukes, M.D., Bliznyuk, N., 2018. Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study. Environ. Model. Softw. 102, 29–38. https://doi.org/10.1016/j.envsoft.2018.01.002

Edwards, K., Martin, L., 1995. A Methodology for Surveying Domestic Water Consumption. Water Environ. J. 9, 477–488. https://doi.org/10.1111/j.1747-6593.1995.tb01486.x

Fiorillo, D., Kapelan, Z., Xenochristou, M., De Paola, F., Giugni, M., 2021. Assessing the Impact of Climate Change on Future Water Demand by using Weather Data. Water Resources Management (Accepted).

Gato, S., Jayasuriya, N., Roberts, P., 2007. Temperature and rainfall thresholds for base use urban water demand modelling. J. Hydrol. 337, 364–376. https://doi.org/10.1016/j.jhydrol.2007.02.014

Goodchild, C.W., 2003. Modelling the impact of climate change on domestic water demand. Water Environ. J. 17, 8–12. https://doi.org/10.1111/j.1747-6593.2003.tb00423.x

Haque, M.M., de Souza, A., Rahman, A., 2017. Water Demand Modelling Using Independent Component Regression Technique. Water Resour. Manag. 31, 299–312.

https://doi.org/10.1007/s11269-016-1525-1

Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. J. Hydrol. 387, 141–150. https://doi.org/10.1016/j.jhydrol.2010.04.005

Hurd, B.H., 2006. Water conservation and residential landscapes: Household preferences, household choices. J. Agric. Resour. Econ. 31, 173–192. https://doi.org/10.2307/40987314

Jampanil, D., Suttinon, P., Seigo, N., Koontanakulvong, S., 2012. Application of Input-Output Table for future water resources management under policy and climate change in Thailand: Rayong Province Case study, in: PAWEES 2012 International Conference Challenges of Water & Environmental Management in Monsoon Asia. Pakkred, Thailand.

Kanakoudis, V., Tsitsifli, S., Papadopoulou, A., Cencur Curk, B., Karleusa, B., 2017. Water resources vulnerability assessment in the Adriatic Sea region: the case of Corfu Island. Environ. Sci. Pollut. Res. 24, 20173–20186. https://doi.org/10.1007/s11356-017-9732-8

La Jeunesse, I., Cirelli, C., Aubin, D., Larrue, C., Sellami, H., Afifi, S., Bellin, A., Benabdallah, S., Bird, D.N., Deidda, R., Dettori, M., Engin, G., Herrmann, F., Ludwig, R., Mabrouk, B., Majone, B., Paniconi, C., Soddu, A., 2016. Is climate change a threat for water uses in the Mediterranean region? Results from a survey at local scale. Sci. Total Environ. 543, 981–996. https://doi.org/10.1016/j.scitotenv.2015.04.062

Laspidou, C., Papageorgiou, E., Kokkinos, K., Sahu, S., Gupta, A., Tassiulas, L., 2015. Exploring patterns in water consumption by clustering. Procedia Eng. 119, 1439–1446. https://doi.org/10.1016/j.proeng.2015.08.1004

Makki, A.A., Stewart, R.A., Panuwatwanich, K., Beal, C., 2013. Revealing the determinants of shower water end use consumption: Enabling better targeted urban water conservation strategies. J. Clean. Prod. 60, 129–146. https://doi.org/10.1016/j.jclepro.2011.08.007

Mamade, A., Loureiro, D., Covas, D., Coelho, S.T., Amado, C., 2014. Spatial and temporal forecasting of water consumption at the DMA level using extensive measurements. Procedia Eng. 70, 1063–1073. https://doi.org/10.1016/j.proeng.2014.02.118

Manouseli, D., Kayaga, S.M., Kalawsky, R., 2019. Evaluating the Effectiveness of Residential Water Efficiency Initiatives in England: Influencing Factors and Policy Implications. Water Resour. Manag. 33, 2219–2238. https://doi.org/10.1007/s11269-018-2176-1

Masia, S., Sušnik, J., Marras, S., Mereu, S., Spano, D., Trabucco, A., 2018. Assessment of irrigated agriculture vulnerability under climate change in Southern Italy. Water (Switzerland) 10, 1–19. https://doi.org/10.3390/w10020209

Miaou, S. -P, 1990. A class of time series urban water demand models with nonlinear climatic effects. Water Resour. Res. 26, 169–178. https://doi.org/10.1029/WR026i002p00169

Neale, T., Carmichael, J., Cohen, S., 2007. Urban Water Futures: A Multivariate Analysis of Population Growth and Climate Change Impacts on Urban Water Demand in the Okanagan Basin, BC. Can. Water Resour. J. 32, 315–330. https://doi.org/10.4296/cwrj3204315

Papageorgiou, E.I., Poczęta, K., Laspidou, C., 2016. Hybrid model for water demand prediction based on fuzzy cognitive maps and artificial neural networks, in: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Vancouver, BC, pp. 1523–1530. https://doi.org/10.1109/FUZZ-IEEE.2016.7737871

Parandvash, G.H., Chang, H., 2016. Analysis of long-term climate change on per capita water demand in urban versus suburban areas in the Portland metropolitan area, USA. J. Hydrol. 538, 574–586. https://doi.org/10.1016/j.jhydrol.2016.04.035

Parker, J., 2013. Assessing the sensitivity of historic micro-component household water-use to climatic drivers. Loughborough University.

Parker, J., Wilby, R.L., 2013. Quantifying Household Water Demand: A Review of Theory and Practice in the UK. Water Resour. Manag. 27, 981–1011. https://doi.org/10.1007/s11269-012-0190-2

Patel, A., Katiyar, S.K., 2014. Prediction of water demand and water storage capacity of municipal system by using geospatial techniques. J. Geomatics 8, 78–85.

Peres, D.J., Modica, R., Cancelliere, A., 2019. Assessing future impacts of climate change on water supply system performance: Application to the Pozzillo Reservoir in Sicily, Italy. Water (Switzerland) 11. https://doi.org/10.3390/w11122531

Polebitski, A.S., Palmer, R.N., Waddell, P., 2011. Evaluating water demands under climate change and transitions in the urban environment. J. Water Resour. Plan. Manag. 137, 249–257. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000112

Rasifaghihi, N., Li, S.S., Haghighat, F., 2020. Forecast of urban water consumption under the impact of climate change. Sustain. Cities Soc. 52, 1011848. https://doi.org/10.1016/j.scs.2019.101848

Scornet, E., 2017. Tuning parameters in random forests. ESAIM Proc. Surv. 60, 144–162. https://doi.org/10.1051/proc/201760144

Shamsi, U.M., 2005. GIS Applications for Water, Wastewater, and Stormwater Systems. Taylor & Francis, Boca Raton (Florida). https://doi.org/10.1201/9781420039252

Slavíková, L., Malý, V., Rost, M., Petružela, L., Vojáček, O., 2013. Impacts of Climate Variables on Residential Water Consumption in the Czech Republic. Water Resour. Manag. 27, 365–379. https://doi.org/10.1007/s11269-012-0191-1

Toth, E., Bragalli, C., Neri, M., 2018. Assessing the significance of tourism and climate on residential water demand: Panel-data analysis and non-linear modelling of monthly water consumptions.

Environ. Model. Softw. 103, 52–61. https://doi.org/10.1016/j.envsoft.2018.01.011

Wang, X. jun, Zhang, J. yun, Shahid, S., Guan, E. hong, Wu, Y. xiang, Gao, J., He, R. min, 2014. Adaptation to climate change impacts on water demand. Mitig. Adapt. Strateg. Glob. Chang. 21, 81–99. https://doi.org/10.1007/s11027-014-9571-6

Williamson, P., Mitchell, G., McDonald, A.T., 2002. Domestic water demand forecasting: A static microsimulation approach. Water Environ. J. 16, 243–248. https://doi.org/10.1111/j.1747-6593.2002.tb00410.x

Willis, R.M., Stewart, R.A., Giurco, D.P., Talebpour, M.R., Mousavinejad, A., 2013. End use water consumption in households: Impact of socio-demographic factors and efficient devices. J. Clean. Prod. https://doi.org/10.1016/j.jclepro.2011.08.006

Xenochristou, M., 2019. Water Demand Forecasting Using Machine Learning on Weather and Smart Metering Data. University of Exeter.

Xenochristou, M., Hutton, C., Hofman, J., Kapelan, Z., 2021. Short-Term Forecasting of Household Water Demand in the UK Using an Interpretable Machine Learning Approach. J. Water Resour. Plan. Manag. 147, 04021004. https://doi.org/10.1061/(asce)wr.1943-5452.0001325

Xenochristou, M., Hutton, C., Hofman, J., Kapelan, Z., 2020a. Water Demand Forecasting Accuracy and Influencing Factors at Different Spatial Scales Using a Gradient Boosting Machine. Water Resour. Res. 56, e2019WR026304. https://doi.org/10.1029/2019WR026304

Xenochristou, M., Kapelan, Z., 2020. An ensemble stacked model with bias correction for improved water demand forecasting. Urban Water J. 17, 212–223. https://doi.org/10.1080/1573062X.2020.1758164

Xenochristou, M., Kapelan, Z., Hutton, C., 2020b. Using Smart Demand-Metering Data and Customer Characteristics to Investigate Influence of Weather on Water Consumption in the UK. J. Water Resour. Plan. Manag. 146, 04019073. https://doi.org/10.1061/(asce)wr.1943-5452.0001148

Xenochristou, M., Kapelan, Z., Hutton, C., Hofman, J., 2018. Smart Water Demand Forecasting: Learning from the Data 3, 2351–2342. https://doi.org/10.29007/wkp4

Zachariadis, T., 2010. Residential water scarcity in Cyprus: Impact of climate change and policy options. Water 2, 788–814. https://doi.org/10.3390/w2040788

Zhou, S.L., McMahon, T.A., Walton, A., Lewis, J., 2000. Forecasting daily urban water demand: A case study of Melbourne. J. Hydrol. 236, 153–164. https://doi.org/10.1016/S0022-1694(00)00287-0

Zubaidi, S.L., Ortega-Martorell, S., Kot, P., Alkhaddar, R.M., Abdellatif, M., Gharghan, S.K., Ahmed, M.S., Hashim, K., 2020. A Method for Predicting Long-Term Municipal Water

Demands Under Climate Change. Water Resour. Manag. 34, 1265–1279. https://doi.org/10.1007/s11269-020-02500-z

6 Conclusion

In the framework of this thesis the topic of urban water demand modelling is investigated as to provide a contribution in developing effective strategies for WDN management. Given the significant benefits provided by smart metering, the thesis also aimed at contributing to the optimization of SWDN implementation. In order to achieve these objectives, smart metering data from a DMA in Soccavo, a suburban area in the North-Western part of Naples (Italy), were extensively used.

In this chapter, a brief summary of the contents of the thesis is presented. Conclusively, recommendations for future research and final remarks are provided.

6.1 Thesis summary

The following is a summary of the objectives, the developed approaches, and the main conclusions of the topics covered by the different chapters. In Chapter 1 the general background for the research described within this thesis is provided.

The second chapter focused on the development of a novel bottom-up methodology for the generation of water demand time series. The objective of the study was to improve the accuracy of existing water demand generation models by developing a new methodology to enable the preservation of the main statistics as well as spatial and temporal correlations of the measured time series, at both single and aggregated scale. The proposed methodology is characterized by two phases. In the first phase, the model generates, for each user and for each time step of the day, first attempt demand time series by applying a Beta probability distribution with tunable bounds or a Gamma distribution with shift parameter. The demand time series generated during the first phase are consistent in terms of mean, standard deviation and skewness with the measured time series. In the second phase, rank cross-correlations between users and at all temporal lags are imposed to the generated demand time series through a single copula-based re-sort. The effectiveness of the methodology was tested against two real case studies: the literature case study of Milford (Ohio), and the Soccavo DMA. The application to the two case studies showed the effectiveness of the methodology to be preserved even when considering larger number of users.

The bottom-up methodology was able to adequately reproduce mean, standard deviation and skewness of measured demand time series. Furthermore, the methodology allowed to preserve rank cross-correlations at all time lags in both single user and spatially aggregated time series. The methodology was also successfully applied to the generation of nodal demands, leading to good results at both single node and aggregated scales. It was found that starting from nodal demands, rather than from single users, can be beneficial for the generation of aggregated demand time series, improving the preservation of temporal cross-correlations.

Successively, a comparative analysis between the proposed bottom-up methodology and a top-down one, based on a literature disaggregation model, was carried out in order to investigate benefits and limitations of the two different approaches. The comparison, described and discussed in Chapter 3, was aimed at highlighting the differences in synthetic time series generation in terms of statistics (mean, standard deviation and skewness) and spatial and temporal correlations, as well as in the computational burden of the models. For the top-down approach, once the aggregated water demand patterns are defined, the disaggregation is used to generate water demand time series at lower levels of spatial aggregation. More specifically, in the first phase of the top-down methodology, the Beta probability distribution with tunable bounds is used to generate the total water demand time series of the considered area. Then a copula resorting is used to impose to the generated time series the existing temporal correlations at all temporal lags. During the second phase, a non-parametric disaggregation model based on the K-nearest neighbours approach is used to generate water demand time series at user or nodal level. In order to establish a consistent comparison, two case studies with different number of users were considered, both referred to the Soccavo DMA. Furthermore, different applications for both case studies were performed by grouping the users in different number of nodes.

The results of the comparative analysis showed that the bottom-up methodology performed significantly better than the top-down one in reproducing rank cross-correlations between single users, as well as between single nodes. On the other hand, the top-down model showed a better performance in terms of skewness and rank cross-correlation when the spatially aggregated demands were considered. Finally, the level of aggregation in nodes was found to positively affect the performance of both the considered methodologies. Overall, even though the top-down approach required the definition of a lower number of parameters and, therefore, a lower computational burden, the parameters of the bottom-up methodology were easily estimated, showing that the model can be easily parameterized when smart meter readings are available.

Chapter 4 focused on operational aspects for the effective SWDN implementation and management. The analysis of the total water demand pattern in a DMA can help water utilities identify anomalous events (i.e. pipe bursts, unauthorized consumption and leakages) and, as a result, plan suitable interventions in the WDN. Indeed, by comparing the total demand pattern with the inflow into the DMA, which is usually monitored by flowmeters at DMA boundaries, water utilities can easily identify anomalous events. Installing smart meters at all user locations represents the optimum solution for water demand pattern reconstruction. However, the installation of smart meters at each user location may be exceedingly expensive for water utilities. Driven by the interest in finding a satisfactory solution to this issue, the objective of the work was to define innovative strategies for the optimal selection of metering points. Therefore, two procedures to identify the optimal allocation of a limited number of smart meters for an accurate reconstruction of the temporal pattern of DMA demand were developed, as to find a suitable trade-off between the need for an accurate characterization of the total district demand and smart-metering costs. The two proposed procedures cover two possible different DMAs scenarios: the first scenario consists of a DMA in which smart meters at the end of life were previously installed at all user locations; the second is a DMA where smart metering needs to be implemented for the first time. The first procedure is based on the application of the stepwise regression for the selection of user locations, while using Fisher's test or economic considerations as stopping criterion. The procedure leads to the assessment of the minimum number of smart meters to be replaced for a reliable demand characterization, allowing to replace a number of smart meters lower than the initial one. The second procedure can be applied during the initial phase of smart metering implementation, enabling the identification of the users to be monitored starting from the billed annual consumption of each user in the DMA. More specifically, it is based on the application of different criteria for users' selection based on billed annual demand and type of users (residential or non-residential), while using a novel linear model to estimate the temporal pattern of the total daily demand of the DMA. This model can be easily calibrated by the water utility since it is parameterized based on the billed annual demands of the users. Both procedures were applied to the Soccavo DMA by using hourly consumption data from both residential and non-residential users.

<u>The results revealed that for both the proposed procedures for the optimal smart meters allocation the accuracy of the total demand pattern reconstruction was good even for low number of selected users (almost the 2% of the total number of users), becoming particularly high as the number of selected users reached the 7% of the total number of users. Concluding, an effective strategy for the allocation of smart meters for the accurate reconstruction of the total district demand pattern was provided. Thus, the deterministic criterion for users' selection resulted to be the best one, identifying the users with the highest values of billed consumption while distinguishing between residential and non-residential users.</u>

Finally, Chapter 5 was devoted to the current need for water demand forecasting models that can be easily applied by water utilities. Indeed, considering that SWDNs implementation could be too costly for water utilities and the increasing importance of water demand forecasts due to climate and demographic changes, the study was aimed at exploring new strategies for water demand forecasting. Therefore, one of the objectives was to obtain reliable water demand forecasts for DMAs lacking smart demand metering data by using information easily accessible to network operators. A further objective was to improve the knowledge of weather effect on urban water demand in Italy, in view of the expected climate change impact on water demand and its likely variation according to geographic and climatic conditions. For these purposes, the prediction accuracy of Random Forests models based on weather variables was investigated for the Soccavo case study. Indeed, an extensive dataset including water consumptions and social characteristics belonging to the Soccavo DMA, as well as weather data, was used. Then, the water consumption was disaggregated based on the social characteristics of the users.

The water demand forecasting models based on weather variables were able to predict aggregated daily water demand, allowing to catch the variations in water demand due to weather and climate changes. However, the effectiveness of the models changed depending on the social characteristics of the users, highlighting the relevance of disaggregating consumption to both determine the influence of weather on water demand and improve forecasting models in both current and future climate change scenarios.

6.2 Limitations and future directions

The results gathered in this thesis provide a contribution in urban water demand modelling and SWDN management. This thesis also highlights future research opportunities, as illustrated in the following.

The use of realistic input plays a major role in the numerical simulation of WDN behaviour. Only by using accurate estimates of water demand, numerical simulation can yield consistent results in determining hydraulic variables, such as nodal outflows and pressure-heads (Creaco et al., 2017a, 2017b). In the framework of this thesis, this issue was addressed by developing a bottom-up approach for water demand time series generation based on the use of the Beta probability distribution with tunable bounds or the Gamma distribution with shift parameter. Among the two probability distributions, the Gamma distribution is easier to use because its parameterization is completely based on analytical relationships. However, in the application to the considered case study, in most cases it was not possible to use the Gamma distribution with shift parameter due to the generation of infeasible negative demand values (see Chapter 2). Therefore, the Gamma probability distribution with two parameters was used, resulting in failure to preserve the skewness of measured demand time series. Furthermore, some limitations in preserving the basic statistics at aggregated scale due to the approximations in modelling were also observed when the Beta probability distribution with tunable bounds was used.

Improved results could be achieved by using more complex probability distributions than those applied in this work. On the other hand, this could cause the growth of the parameterization burden. Therefore, future works should be dedicated to an in depth study of such issue. More specifically, the use of different probability distributions to generate water demand time series should be investigated in order to assess if a larger computational burden is effectively paid back by higher levels of accuracy at both single user and aggregated scale. In this investigation, the application of the Gamma probability distribution with shift parameter to new case studies can be beneficial as well. This can help to assess the performance of the Gamma distribution when the shift parameter is non-zero.

Furthermore, even though the parameterization of the presented bottom-up model is generally easy, the methodology is inapplicable for water demand time series generation in the absence of smart meter readings. This limitation can be overcome by using the top-down model presented in this work because it is parameterized based on the measured total demand of the DMA (Chapter 3). However, the top-down model resulted poorly capable of reproducing the cross-correlations between single users (or nodes) at various temporal lags.

Considering the importance of accurately reproducing both the main statistics of the measured time series and their spatial and temporal correlations (Alvisi et al., 2016; Blokker et al., 2011, 2008; Filion et al., 2007, 2005; Moughton et al., 2007), future works should focus on developing new models with a lower burden of parameterization, that are able to achieve high levels of accuracy as that obtained for the bottom-up methodology presented in this thesis. Developing an accurate water demand time series generation model that does not require smart meter readings widely distributed over the WDN represents a challenging task for future researches, as well as a significant opportunity to improve WDNs behaviour simulations.

As regards the WDNs management, the reconstruction of the total demand patterns of a DMA is a topic of high interest for water utilities, allowing to improve a variety of different tasks, including e.g. the detection of anomalous events. To this purpose, in the framework of this thesis, two procedures to identify a small subset of smart meter locations for an accurate reconstruction of the temporal pattern of the total DMA demand were developed (Chapter 4). Both procedures are based on a linear regression model. However, different models, such as neural networks (Haykin, 1999) and evolutionary polynomial regression (Giustolisi and Savic, 2006), could be used to improve the fit of the pattern for a preassigned installation of smart meters. Then, future efforts should be dedicated to explore this issue.

Furthermore, the procedures proposed are expected to be effective also at node level. Therefore, future research could focus on expanding the procedures through their use at node level to identify the most influential users of the aggregated demand to provide accurate estimates of nodal demand for numerical simulations and pressure management optimization models.

Further improvements can be obtained by applying the presented procedures to new case studies. Even though this study provided a general effective strategy for the allocation of smart meters, the results could change based on the typology of the DMA. Indeed, for residential areas, such the one considered in this study, the percentage of users selected out of the total number of users in the DMA is expected to remain almost unchanged. However, for DMAs with different characteristics (commercial or industrial districts) the results could change in terms of both type and number of users selected. Therefore, future developments could be dedicated to an in-depth study of such issues by applying the procedures presented to other case studies.

A further essential task for water utilities is water demand forecasting, with increasing importance in light of the future climate changes. This thesis addressed this topic by investigating the accuracy of forecasting models based on weather variables, disaggregating water consumption based on the social characteristics of the users (Chapter 5). However, additional socio-economic information about the users could lead to an improved disaggregation of the consumption, improving in turn the water demand forecasting models.

Furthermore, since the case study included only residential users, future studies should focus on assessing the effects of including industrial or commercial users. In addition, the investigation presented in the framework of this thesis was conducted for a case study in southern Italy. In different areas, with different climate, under different socio-economic conditions, the results could be very different. Therefore, future research should be dedicated to perform similar investigations in different areas of Italy and in un-investigated areas of the world. This is even more important considering that the expected climate change impact on water demand would vary widely with geographic and climatic areas (Babel et al., 2014; Collet et al., 2015; Downing et al., 2003; Jampanil et al., 2012; Kanakoudis et al., 2017; Karamouz et al., 2011; Neale et al., 2007; Parker, 2013; Polebitski et al., 2011; Rasifaghihi et al., 2020; Wang et al., 2014; Zachariadis, 2010; Zubaidi et al., 2020).

In this context, future research should focus on using the outcomes of this thesis to analyse climate change impacts on water demand of different types of users under different climate scenarios.

Overall, in order to improve the consistency and reliability of the results attained in the framework of this thesis, future developments should be dedicated to the application of the proposed methodologies to new case studies.

6.3 Thesis contributions

In conclusion, this research provided different contributions to existing knowledge of urban water demand, WDNs management and SWDN implementation. Specifically, this thesis offered the following main contributions:

- The first contribution is an accurate water demand time series generation model which can lead to improvements in design and management of WDNs. Indeed, the bottom-up model presented in this thesis allows to appropriately characterize the inputs of hydraulic simulation models for designing, verifying and managing WDNs. Furthermore, the ability of the proposed bottom-up methodology to reproduce all the main statistics and existing correlations of the measured time series at both single users and aggregated scale is beneficial from an operational point of view, enabling a proper characterization of the system performances.
- The second contribution is an improved understanding of the benefits and limitations of the main approaches in water demand modelling, i.e. the bottom-up and the top-down approach. The comparative study between the bottom-up and top-down methodology presented in this thesis increased the existing knowledge about water demand modelling, providing indications for choice of the most suitable approach to be used according to the specific needs.
- In addition, the research provided effective tools for SWDN management and implementation. The procedures to identify the optimal allocation of a limited number of smart meters presented in this thesis allow water utilities to overcome operational problems, such as the replacement of smart meters at the end of life. Using the proposed procedures, water utilities can improve a cost-effective management of the SWDNs. This can lead to costs and water resources saving. Furthermore, the procedures enable the implementation of an effective smart metering system even in case of a relative low initial budget.
- A further contribution of this thesis is a novel linear model to estimate the temporal pattern of the total daily demand of a DMA based on the billed annual demands of the users. This model represents a useful tool for water utilities managing DMA in absence of smart demand metering data, being completely based on information generally accessible by water utilities.

- Finally, the thesis contributed to the search of reliable water demand explanatory factors to be used when consumption data are limited, showing that weather variables can be effectively used to forecast the total water demand of a DMA. The research also confirmed the importance of social and temporal factors for the assessment of the influence of weather on water demand and the improvement of forecasting models. Furthermore, the performed analysis partially contributed to an improved understanding of weather effect on urban water demand in Italy, especially as regards the south of the country. The obtained results can assist with addressing future requirements related to climate change mitigation.

Overall, the results attained in the framework of this thesis are expected to contribute to the improvement of WDNs management and to provide useful tools for optimizing smart metering systems.

References

Alvisi, S., Ansaloni, N., Franchini, M., 2016. Comparison of parametric and nonparametric disaggregation models for the top-down generation of water demand time series. Civ. Eng. Environ. Syst. 33, 3–21. https://doi.org/10.1080/10286608.2015.1126823

Babel, M.S., Maporn, N., Shinde, V.R., 2014. Incorporating Future Climatic and Socioeconomic Variables in Water Demand Forecasting: A Case Study in Bangkok. Water Resour. Manag. 28, 2049–2062. https://doi.org/10.1007/s11269-014-0598-y

Blokker, E.J.M., Beverloo, H., Vogelaar, A.J., Vreeburg, J.H.G., Van Dijk, J.C., 2011. A bottom-up approach of stochastic demand allocation in a hydraulic network model: A sensitivity study of model parameters. J. Hydroinformatics 13, 714–728. https://doi.org/10.2166/hydro.2011.067

Blokker, E.J.M., Vreeburg, J.H.G., Buchberger, S.G., van Dijk, J.C., 2008. Importance of demand modelling in network water quality models: a review. Drink. Water Eng. Sci. Discuss. 1, 27–38. https://doi.org/10.5194/dwesd-1-1-2008

Collet, L., Ruelland, D., Estupina, V.B., Dezetter, A., Servat, E., 2015. Water supply sustainability and adaptation strategies under anthropogenic and climatic changes of a meso-scale Mediterranean catchment. Sci. Total Environ. 536, 589–602. https://doi.org/10.1016/j.scitotenv.2015.07.093

Creaco, E., Blokker, M., Buchberger, S., 2017a. Models for generating household water demand pulses: Literature review and comparison. J. Water Resour. Plan. Manag. 143, 04017013. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000763

Creaco, E., Pezzinga, G., Savic, D., 2017b. On the choice of the demand and hydraulic modeling approach to WDN real-time simulation. Water Resour. Res. 53, 6159–6177. https://doi.org/10.1002/2016WR020104

Downing, T.E., Butterfield, R.E., Edmonds, B., Knox, J.W., Moss, S., Piper, B.S., Weatherhead, E.K., 2003. Climate Change and Demand for Water, Research Report, Stockholm Environment Institute. Oxford (UK).

Filion, Y., Adams, B., Karney, B., 2007. Cross correlation of demands in water distribution network design. J. Water Resour. Plan. Manag. 133, 137–144. https://doi.org/10.1061/(ASCE)0733-9496(2007)133:2(137)

Filion, Y.R., Karney, B.W., Adams, B.J., 2005. Stochasticity of demand and probabilistic performance of water networks, in: Impacts of Global Climate Change: Proceedings of World Water and Environmental Resources Congress 2005. ASCEE, Anchorage, Alaska, pp. 1–12. https://doi.org/10.1061/40792(173)49

Giustolisi, O., Savic, D.A., 2006. A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinformatics 8, 207–222. https://doi.org/10.2166/hydro.2006.020b

Haykin, S., 1999. Neural networks: A comprehensive foundation, Second edi. ed, Prentice Hall International, Inc. Prentice Hall International, Inc., Upper Saddle River (New Jersey).

Jampanil, D., Suttinon, P., Seigo, N., Koontanakulvong, S., 2012. Application of Input-Output Table for future water resources management under policy and climate change in Thailand: Rayong Province Case study, in: PAWEES 2012 International Conference Challenges of Water & Environmental Management in Monsoon Asia. Pakkred, Thailand.

Kanakoudis, V., Tsitsifli, S., Papadopoulou, A., Cencur Curk, B., Karleusa, B., 2017. Water resources vulnerability assessment in the Adriatic Sea region: the case of Corfu Island. Environ. Sci. Pollut. Res. 24, 20173–20186. https://doi.org/10.1007/s11356-017-9732-8

Karamouz, M., Zahmatkesh, Z., Nazif, S., 2011. Selecting a domestic water demand prediction model for climate change studies, in: World Environmental and Water Resources Congress 2011: Bearing Knowledge for Sustainability. ASCE, Palm Springs, California (USA), pp. 1338–1346. https://doi.org/10.1061/41173(414)139

Moughton, L.J., Buchberger, S.G., Boccelli, D.L., Filion, Y.R., Karney, B.W., 2007. Effect of time step and data aggregation on cross correlation of residential demands, in: 8th Annual Water Distribution Systems Analysis Symposium 2006. ASCEE, Cincinnati, Ohio (USA), pp. 1–11. https://doi.org/10.1061/40941(247)42

Neale, T., Carmichael, J., Cohen, S., 2007. Urban Water Futures: A Multivariate Analysis of Population Growth and Climate Change Impacts on Urban Water Demand in the Okanagan Basin, BC. Can. Water Resour. J. 32, 315–330. https://doi.org/10.4296/cwrj3204315

Parker, J., 2013. Assessing the sensitivity of historic micro-component household water-use to climatic drivers. Loughborough University.

Polebitski, A.S., Palmer, R.N., Waddell, P., 2011. Evaluating water demands under climate change and transitions in the urban environment. J. Water Resour. Plan. Manag. 137, 249–257. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000112

Rasifaghihi, N., Li, S.S., Haghighat, F., 2020. Forecast of urban water consumption under the impact of climate change. Sustain. Cities Soc. 52, 1011848. https://doi.org/10.1016/j.scs.2019.101848

Wang, X. jun, Zhang, J. yun, Shahid, S., Guan, E. hong, Wu, Y. xiang, Gao, J., He, R. min, 2014. Adaptation to climate change impacts on water demand. Mitig. Adapt. Strateg. Glob. Chang. 21, 81–99. https://doi.org/10.1007/s11027-014-9571-6

Zachariadis, T., 2010. Residential water scarcity in Cyprus: Impact of climate change and policy options. Water 2, 788–814. https://doi.org/10.3390/w2040788

Zubaidi, S.L., Ortega-Martorell, S., Kot, P., Alkhaddar, R.M., Abdellatif, M., Gharghan, S.K., Ahmed, M.S., Hashim, K., 2020. A Method for Predicting Long-Term Municipal Water Demands Under Climate Change. Water Resour. Manag. 34, 1265–1279. https://doi.org/10.1007/s11269-020-02500-z

Appendix: Preliminary analysis – Chapter 5

The predictors in water demand forecasting models were chosen based on the results of a preliminary analysis, aimed at assessing reliable predictors of water demand among weather variables. First, the relationship between each pair of weather variables was investigated. Then, the degree of relationship between each weather variable and water consumption was assessed. The study investigated the varying effect that weather changes can have across time, accounting for users characteristics.

A.1. Relationship between weather variables

In this study weather data over the period from 20 March 2017 to 19 March 2018 were used. Six weather variables, air temperature, air humidity, rainfall amount, rainfall rate, solar radiation and wind speed, were initially considered in the analysis. The information from the weather station of the University of Naples Federico II and the Italian Air Force at Capodichino was used to calculate one daily value for each weather variable. Table 17 summarizes the weather variables that were used in this study.

Table 17 – Description of the weather variables.

Weather variable	Description	Units	Weather station
Air temperature	Maximum daily temperature	°C	University of Naples Federico II
Humidity	Daily mean humidity	%	Italian Air Force - Capodichino
Rainfall amount	Total daily rainfall	mm	University of Naples Federico II
Rainfall rate	Daily mean rainfall rate	mm/h	University of Naples Federico II
Solar radiation	Daily mean solar radiation	W/m^2	University of Naples Federico II
Wind speed	Mean daily wind speed	Km/h	University of Naples Federico II

The Spearman's rank correlation coefficient ρ (Spearman, 1904) was used to assess both the strength and the direction of association between each pair of weather variables. The Spearman's correlation coefficient was chosen as it is a non-parametric test allowing the identification of non-linear relationships. Then, the $p - value$ of the correlation was used to assess the statistical significance of the relationship. Therefore, correlation with $p - value > 0.01$ were considered statistically insignificant.

Table 18 reports the Spearman's ρ correlation coefficient for each pair of ranked variables. The strongest correlation was observed between rainfall amount and rate ($\rho = 0.93$), followed by solar radiation and air temperature ($\rho = 0.77$). Solar radiation was also found inversely correlated with rainfall amount ($\rho = -0.60$), rainfall rate ($\rho = -0.55$) and humidity ($\rho = -0.57$). Finally, a moderate relationship was observed between rainfall amount and humidity ($\rho = 0.51$). The other correlations between weather variables were not statistically significant ($\rho < |\pm 0.5|$).

Table 18 – Spearman's coefficients between each pair of weather variables whit $p - value < 0.01$.

Weather variable	Rainfall amount	Rainfall rate	Wind speed	Solar radiation	Humidity	Air temperature
Rainfall amount	1.00	0.93	0.35	-0.60	0.51	-0.46
Rainfall rate	0.93	1.00	0.33	-0.55	0.49	-0.41
Wind speed	0.35	0.33	1.00	-	-0.20	-0.13
Solar radiation	-0.60	-0.55	-	1.00	-0.57	0.77
Humidity	0.51	0.49	-0.20	-0.57	1.00	-0.42
Air temperature	-0.46	-0.41	-	0.77	-0.42	1.00

These results are in agreement with a previous study referred to the UK (Xenochristou et al., 2020b) that showed a strong correlation between solar radiation and air temperature, and an equally strong but inverse correlation between solar radiation and humidity.

Overall, these results constitute a good basis for further analysis aimed at developing improved water demand forecasting models. Indeed, forecasting models can be affected by variable interactions. In the case of variable interaction, the predictors can provide overlapping information to the model. Thus, the influence of each predictor can be concealed by overlapping information, affecting the forecasting model. Therefore, the interactions between weather variables should be taken into account when choosing the model predictors.

A.2. Relationship between water demand and weather variables

In order to assess the impact of weather on water demand at different times, the consumption was segmented based on the season (into summer, spring, autumn, and winter) and the type of day (into working day and weekends/holidays), as shown in Table 19. Furthermore, in order to take into account the social characteristics of the users, this segmentation was applied to the daily total district demand and the daily total demand of the groups of users that were selected on the basis of employment state and educational level of the residents (Table 19).

Table 19 – Temporal segmentation and users groups.

Temporal segmentation	**Household groups**
Whole year	All
Working days	*Group 1*: Employed with High school/University degree
Holidays	*Group 2*: Unemployed with Primary/Secondary school degree
Spring	*Group 3*: Employed with Primary/Secondary school degree
Summer	
Autumn	
Winter	

Then, the Spearman's rank correlation coefficient ρ (Spearman, 1904) was used as an indicator of the degree of relationship between each weather variable and water consumption, considering correlations with $p - value > 0.01$ statistically insignificant.

Table 20 reports the Spearman's ρ correlation coefficient for each temporal segmentation and each group of households. The strongest correlations were observed for air temperature during spring, with the total district demand ($\rho = 0.65$) and the total demand of *Group 1* ($\rho = 0.65$). The temperature also showed a moderate correlation with the total demand of *Group 3* during autumn ($\rho = 0.57$). Moderate correlations were also found for solar radiation during spring ($\rho = 0.51$ for the total district and $\rho = 0.50$ for *Group 1*), summer ($\rho = -0.50$ for *Group 2*), autumn ($\rho = 0.58$ for *Group 3*) and holidays ($\rho = -0.50$ for *Group 1*). The other variables, rain, wind speed and humidity, showed only weak relationships with water demand ($\rho < |\pm 0.5|$).

Similar results were obtained by previous studies (Adamowski, 2008; Balling et al., 2008; Chang et al., 2014; Cole and Stewart, 2013; Downing et al., 2003; Goodchild, 2003; Slavíková et al., 2013; Statzu and Strazzera, 2009; Toth et al., 2018; Xenochristou, 2019; Xenochristou et al., 2020b). Water demand was found strongly correlated with air temperature (Adamowski, 2008; Balling et al., 2008; Chang et al., 2014; Cole and Stewart, 2013; Downing et al., 2003; Goodchild, 2003; Slavíková et al., 2013; Statzu and Strazzera, 2009; Toth et al., 2018; Xenochristou, 2019; Xenochristou et al., 2020b) and sunshine hours (Goodchild, 2003; Xenochristou, 2019), whereas a limited to no effect was observed for rainfall (Chang et al., 2014; Cole and Stewart, 2013; Downing et al., 2003; Goodchild, 2003; Slavíková et al., 2013; Toth et al., 2018; Xenochristou, 2019; Xenochristou et al., 2020b).

Table 20 - Spearman's coefficients between each weather variable and daily demand with p-value < 0.01, for the entire DMA, Group 1, Group 2 and Group 3.

Temporal segment	Household group	Rain height	Rain rate	Wind speed	Solar radiation	Humidity	Air temperature
Whole year	All	-	-	-	-	-	-
	Group 1	-	0.15	-	-0.37	0.18	-0.29
	Group 2	-0.18	-0.17	-	0.28	-	0.20
	Group 3	-0.32	-0.30	-0.14	0.45	-0.16	0.36
Working days	All	-	-	-	-	-	-
	Group 1	-	-	-	-0.28	0.18	-0.24
	Group 2	-0.21	-0.19	-	0.26	-	0.20
	Group 3	-0.35	-0.33	-0.20	0.49	-	0.38
Holidays	All	-	-	-	-	-	-
	Group 1	-	-	-	**-0.50**	-	-0.39
	Group 2	-	-	-	0.32	-	-
	Group 3	-0.28	-0.25	-	0.44	-	0.34
Spring	All	-	-	-	**0.51**	-	**0.65**
	Group 1	-	-	-	**0.50**	-	**0.64**
	Group 2	-	-	-	-	-	-
	Group 3	-	-	-	0.39	-	0.36
Summer	All	0.32	0.30	-	-	0.42	-0.47
	Group 1	0.28	0.27	-	-	0.37	-0.46
	Group 2	0.35	0.31	-	**-0.50**	0.31	-0.37
	Group 3	-	-	-	-	0.39	-
Autumn	All	-	-	-	0.44	-	0.40
	Group 1	-	-	-	0.31	-	-
	Group 2	-	-	-	0.37	-	0.40
	Group 3	-0.33	-	-	**0.58**	-	**0.57**
Winter	All	-0.44	-0.42	-	-	-	0.28
	Group 1	-0.40	-0.36	-	-	-	-
	Group 2	-0.32	-0.33	-	-	-	-
	Group 3	-0.35	-0.34	-	-	-	-

Figure 31 and Figure 32 show the correlation between consumption and weather variables across different days and type of users. Each point represents one day and the trend line indicates the linear regression model that best fits the data, showing the degree of the weather effect on water demand.

Figure 31 shows the simultaneous effect of maximum daily temperature and mean daily solar radiation on the total district demand during spring (Figure 31a) and winter months (Figure 31b). The temperature is represented as the independent variable, while for the solar radiation the colour ranges are used. During spring days there is less uncertainty and a clearer trend that shows a steady increase for an increase in temperature. On the other hand, during winter months the trend is less clear, showing a lower impact of temperature on water demand. Furthermore, for the same temperature higher solar radiation (lighter blue) corresponds to higher values of consumption during spring.

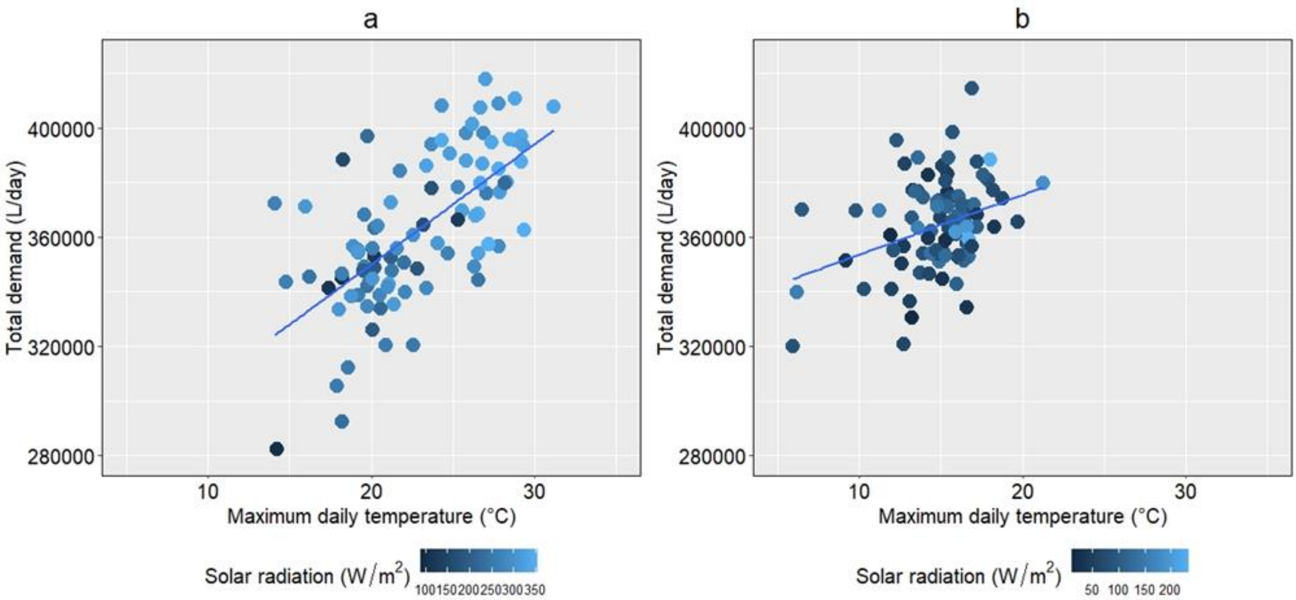

Figure 31 - Correlation between maximum daily temperature (°C) and daily total district demand during spring (a) and winter (b).

Figure 32 shows the simultaneous effect of solar radiation and temperature during working days for *Group 1* (Figure 32a) and *Group 3* (Figure 32b). The consumption of *Group 1* is characterized by a higher variability compared to *Group 3*. Although the trend is less clear, the consumption seems to decrease with increasing solar radiation. On the contrary, the total demand of *Group 3* moderately increases as solar radiation increases. In addition, the highest temperatures correspond to the lowest values of the total demand for *Group 1,* while for *Group 3* high values of total demand correlate with high values of temperature.

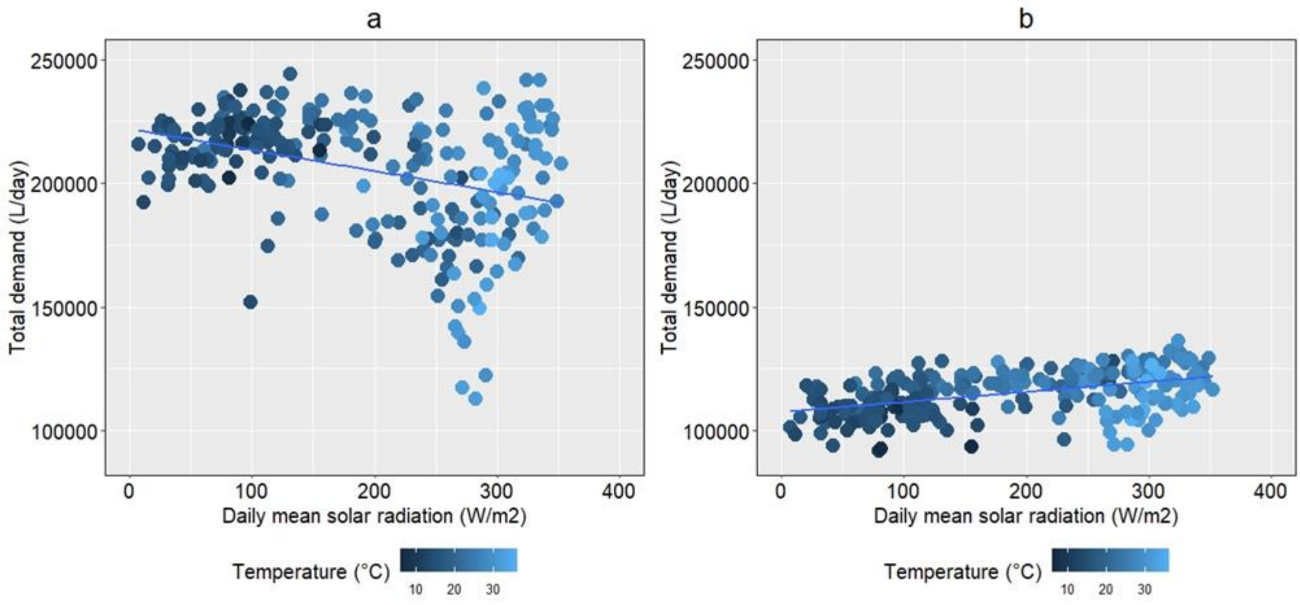

Figure 32 - Correlation between daily mean solar radiation (W/m^2) and daily total demand of Group 1 (a) and Group 2 (b) during working days.

The above results showed that the effect of weather on water demand can change according to season and, type of day (working day or weekend) and user. Certain type of users during certain time of the year or the week can be more sensitive to weather than others (Ashoori et al., 2016; Chang et al., 2010; Cole and Stewart, 2013; Domene and Saurí, 2006; Downing et al., 2003; Parker, 2013; Xenochristou et al., 2020b). Indeed, there are times and users types for which primary water uses, such as outdoor use and use for personal hygiene, are more likely to occur (Chang et al., 2010; Cole and Stewart, 2013; Domene and Saurí, 2006; Downing et al., 2003; Parker, 2013). Water is more likely to be used for outdoor activities during summer (Cole and Stewart, 2013; Downing et al., 2003; Parker, 2013; Statzu and Strazzera, 2009; Toth et al., 2018) and for users with higher socio-economic status (Balling et al., 2008; Chang et al., 2010; Domene and Saurí, 2006; Villar-Navascués and Pérez-Morales, 2018; Wentz and Gober, 2007), while weekly pattern indicated that people are more likely to wash clothes over the weekend (Parker, 2013). Therefore, it is expected that for these times and users types the consumption is more affected by weather.

Overall, the performed analysis was needed for better understanding water demand pattern and developing effective water demand forecasting models. Indeed, since water demand showed significant correlations ($\rho > |\pm 0.5|$) with the daily maximum air temperature and the daily mean solar radiation, these weather variables should be considered in the forecasting models.

Furthermore, the analysis showed that accounting for the social characteristics of the users contributes in explaining sensitivity to weather. Different types of users can have different habits that in turn can be affected by weather to varying degrees. On one hand, given the importance of understanding habits related to water use for the characterization of water demand patterns (Allon and Sofoulis, 2006), accounting for the social characteristics of the users can contribute in improving water demand forecasting models. On the other hand, investigating the accuracy of water forecasting models across different types of users can enable a deeper understanding of climate and weather changes impact on water demand, extending and strengthening the results of preliminary analysis such that performed in this study.

References

Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. J. Water Resour. Plan. Manag. 134, 119–128. https://doi.org/10.1061/(asce)0733-9496(2008)134:2(119)

Allon, F., Sofoulis, Z., 2006. Everyday Water: Cultures in transition. Aust. Geogr. 37, 45–55. https://doi.org/10.1080/00049180500511962

Ashoori, N., Dzombak, D.A., Small, M.J., 2016. Modeling the Effects of Conservation, Demographics, Price, and Climate on Urban Water Demand in Los Angeles, California. Water Resour. Manag. 30, 5247–5262. https://doi.org/10.1007/s11269-016-1483-7

Balling, R.C., Gober, P., Jones, N., 2008. Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona. Water Resour. Res. 44, 1–11. https://doi.org/10.1029/2007WR006722

Chang, H., Parandvash, G.H., Shandas, V., 2010. Spatial variations of single-family residential water consumption in Portland, Oregon. Urban Geogr. 31, 953–972. https://doi.org/10.2747/0272-3638.31.7.953

Chang, H., Praskievicz, S., Parandvash, H., 2014. Sensitivity of Urban Water Consumption to Weather and Climate Variability at Multiple Temporal Scales: The Case of Portland, Oregon. Int. J. Geospatial Environ. Res. 1, 1–7.

Cole, G., Stewart, R.A., 2013. Smart meter enabled disaggregation of urban peak water demand: Precursor to effective urban water planning. Urban Water J. 10, 174–194. https://doi.org/10.1080/1573062X.2012.716446

Domene, E., Saurí, D., 2006. Urbanisation and water consumption: Influencing factors in the metropolitan region of Barcelona. Urban Stud. 43, 1605–1623. https://doi.org/10.1080/00420980600749969

Downing, T.E., Butterfield, R.E., Edmonds, B., Knox, J.W., Moss, S., Piper, B.S., Weatherhead, E.K., 2003. Climate Change and Demand for Water, Research Report, Stockholm Environment Institute. Oxford (UK).

Goodchild, C.W., 2003. Modelling the impact of climate change on domestic water demand. Water Environ. J. 17, 8–12. https://doi.org/10.1111/j.1747-6593.2003.tb00423.x

Parker, J., 2013. Assessing the sensitivity of historic micro-component household water-use to climatic drivers. Loughborough University.

Slavíková, L., Malý, V., Rost, M., Petružela, L., Vojáček, O., 2013. Impacts of Climate Variables on Residential Water Consumption in the Czech Republic. Water Resour. Manag. 27, 365–379.

https://doi.org/10.1007/s11269-012-0191-1

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. Am. J. Psychol. 15, 72–101.

Statzu, V., Strazzera, E., 2009. Water Demand for Residential Uses in a Mediterranean Region : Econometric Analysis and Policy Implications 1–31.

Toth, E., Bragalli, C., Neri, M., 2018. Assessing the significance of tourism and climate on residential water demand: Panel-data analysis and non-linear modelling of monthly water consumptions. Environ. Model. Softw. 103, 52–61. https://doi.org/10.1016/j.envsoft.2018.01.011

Villar-Navascués, R.A., Pérez-Morales, A., 2018. Factors Affecting Domestic Water Consumption on the Spanish Mediterranean Coastline. Prof. Geogr. 70, 513–525. https://doi.org/10.1080/00330124.2017.1416302

Wentz, E.A., Gober, P., 2007. Determinants of small-area water consumption for the City of Phoenix, Arizona. Water Resour. Manag. 21, 1849–1863. https://doi.org/10.1007/s11269-006-9133-0

Xenochristou, M., 2019. Water Demand Forecasting Using Machine Learning on Weather and Smart Metering Data. University of Exeter.

Xenochristou, M., Kapelan, Z., Hutton, C., 2020. Using Smart Demand-Metering Data and Customer Characteristics to Investigate Influence of Weather on Water Consumption in the UK. J. Water Resour. Plan. Manag. 146, 04019073. https://doi.org/10.1061/(asce)wr.1943-5452.0001148

Notation

Abbreviations

ANN	Artificial Neural Network
AP	Access Point
CLS	Constrained Least Square
DMA	District Metered Area
FTP	File Transfer Protocol
GIS	Geographical Information System
GPRS	General Packet Radio Service
K-NN	K-nearest neighbours
MIU	Meter Interface Unit
RF	Random Forest
RMSE	Root Mean Square Error
SSR	Sum of the Square Residuals
SWDN	Smart Water Distribution Network
WDN	Water Distribution Network

Symbols

$\boldsymbol{\alpha}$	Vector of the regressors coefficients
α_i	Regressor coefficient
α, β	Shape parameters of the Beta distribution
a, b	Lower and upper bounds of the Beta distribution
B	Beta function
γ	Skewness
Γ	Gamma function
Δt	Time step

$\overline{D}_i$	User's mean water demand
$\widehat{D}_{tot,k}$	Estimated total water demand of the DMA
$D_{tot,k}$	Measured total water demand of the DMA
$F_t, F_r, F_{n,r}$	Cumulative frequency distribution
$F(x), G(y)$	Distribution functions
$f(x)$	Probability density distribution
$H(x,y)$	Joint distribution function
k, ϑ	Shape parameters of the Gamma distribution
μ	Mean
m, n_d, n_t	Hyperparameters of the Random Forest model
n_{days}	Number of days
n_{times}	Number of times to repeat users selection
$N_{\Delta t}, N_{data}$	Number of time steps
N_{user}, N_{tot}	Number of users
q_i^j, q_h	User/node water demand
Q^i, Q_h	Aggregated water demand
R^2	Coefficient of determination
ρ	Spearman coefficient
σ	Standard deviation
t_r, t_{nr}	Rates by type of users
x_0	Shift parameter of the Gamma distribution
$\hat{y}(x)$	Prediction at the generic point x