Environmental Modelling & Software 100 (2018) 48–66

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A methodology for synthetic household water consumption data generation

Dimitris T. Kofinas, Alexandra Spyropoulou, Chrysi S. Laspidou*

Department of Civil Engineering, University of Thessaly, Pedion Areos, Volos 38334, Greece

ARTICLE INFO

Article history:
Received 1 December 2016
Received in revised form
7 November 2017
Accepted 9 November 2017

Keywords:
Synthetic water consumption data generation
Pulse models
Missing data
Water consumption patterns

ABSTRACT

In the smart cities context, real-time knowledge of residential water consumption has become increasingly important, especially given the fast evolution of sensors, ICT and the production of big, high-resolution data coming from the urban environment. A variety of reasons often leads to the creation of continuity gaps in these data series, thus making the need for a methodology that produces reliable and realistic synthetic data urgent. In this article, we present a methodology that generates synthetic household water consumption data; we showcase it in two case studies, Skiathos, Greece and Sosnowiec, Poland, which exhibit significant differences in water consumption patterns. The methodology captures the stochasticity of daily residential water use. Algorithm validation is implemented through the comparison of various metrics for actual and generated data; this way, we show that the suggested approach is capable of adequately simulating water consumption in both micro- and macro-time scale.

Data availability

The data used in this article are water consumption data collected by a total of 16 households for a period of 13 months—starting from February 1st, 2015—in two locations: Skiathos, Greece (10 sampling points—faucets—each one corresponding to a different household) and Sosnowiec, Poland (9 sampling points—faucets and appliances—in 6 households). The water consumption monitoring system was installed in a diverse group of households that were specifically chosen in order to provide needed data to help comprehend human behavior and water consumption patterns by different users in a household in various socio-economic settings. The criterion for the selection of the households was the availability and promptness of the housekeepers. Additionally, the households were chosen so that they were diverse, regarding their location in the network and number of occupants. Wireless sensors were installed in various sampling points in the households, i.e. faucets, washers and showers. 30-second step records were transmitted to a remote central server in real time. Technical details on the water consumption monitoring system are provided in Chen et al. (2015). The data used are available online, at validation.issewatus.eu/data-re-use/. This work was undertaken for the EC FP7 project ISS-EWATUS (2016).

1. Introduction

Key to a smart city concept is the idea of measurement, of instrumenting the urban landscape and associated activity and monitoring their state and behavior in a way that leads to technological, governmental and societal advances. It has been said that "you can't manage what you can't measure", which greatly applies to a city of the future, in which near real-time measurements enable stakeholder awareness, engagement and quick response to new conditions, thus leading to a new model of civic behavior and involvement. This new paradigm is based on almost individualized planning on one hand, and near real-time information on another (Lim et al., 2010). A recent study (Cominola et al., 2015) reviews water smart metering projects taking place in the last decades worldwide. According to this work, these projects, which focus on real-time water use monitoring at high spatial and temporal granularity, stimulate modeling approaches and behavior adaptive urban water management strategies. Consumer awareness campaigns have been documented in the literature in the last decade (Russel and Fielding, 2010; Novak et al., 2016; Perren and Yang, 2015; Shan et al., 2015), while latest advances include the development of gaming platforms (Wang and Capiluppi, 2015) for water management and the involvement of social media for citizen engagement in water saving practices. The European Commission

* Corresponding author.
E-mail address: laspidou@uth.gr (C.S. Laspidou).

https://doi.org/10.1016/j.envsoft.2017.11.021
1364-8152/

has funded a series of research projects that developed a series of diverse case studies that all showed how building consumer awareness could limit water consumption (all these projects are grouped under the ICT4WATER cluster—http://ict4water.eu). A number of water utilities increasingly attempt to influence the behavior of consumers towards improving water consumption, by using communication tools to give information back to users and display their consumption or customized feedbacks or water-saving tips. At the same time, various companies have been established lately that specialize solely on transforming the way customers think about their household water consumption, as well as the way utilities engage with their customers. Such companies combine Machine Learning (ML) and other data science tools with cloud computing and behavioral science to develop Software-as-a-Service solution to customer engagement and efficiency issues faced by utilities.

Subsequently, the need for the collection and management of large quantities of temporal and spatial high-resolution data emerges as the core of urban planning, while at the same time, the radical evolution in the technological sector of sensors, Information and Communication Technologies (ICTs), social network data analysis and Data Mining (DM) techniques reveals new potentials for more efficient planning (Laspidou, 2014; Laspidou et al., 2015; Yang et al., 2017). In the urban water domain, due to fast urbanization, increasing demands, climate change and high pressure on water resources, research activity increasingly focuses on monitoring, understanding and better managing urban water activities. Detailed monitoring of household water consumption can reveal useful information about citizen behavioral patterns, not only related to their water use *per se*, but also concerning a range of socio-economic factors, directly or indirectly related to water, such as circadian rhythms, working hours, daily habits, house amenities, familial structure and profile, etc. Furthermore, the spatiotemporal analysis of household water use can help make water consumption a key indicator of human behavior, thus helping authorities and relevant stakeholders identify changes in city-living conditions, such as local development, migration, epidemics, or it can disclose population shifts due to events, such as terrorist attacks, natural disasters, large-scale organized meetings or tournaments, etc. Besides the wealth of information potentially extracted by monitoring water consumption, channeling this data back to the consumers will contribute to an increased awareness that will lead to a smaller household water footprint (Lanzarone and Zanzi, 2010; Perren and Yang, 2015; Al-Hoqani and Yang, 2015). The effectiveness of similar schemes regarding energy consumption through energy metering, billing and direct display methodologies has already been documented (Darby, 2006), concluding that feedback to consumers is an important element of an energy savings scheme for consumers. Numerous relative examples are reported in Ehrhardt-Martinez et al. (2010) and Fischer (2008) works. Indicatively, in the Staats et al. (2004) study an energy savings increase of approximately 3% in 16 months in Netherland is reported and in Wilhite and Ling (1995) study an increase of 2.4% in energy saving in one year for a Norway case study is presented.

In fields such as DM, ML and Knowledge Discovery from Databases (KDD), a commonly emerging issue, which is the main focus of this paper, is that of missing values or missing data. Numerous reasons can lead to such a problem: Equipment malfunctions, refusal of respondents to fill in questionnaires and gathering of erroneous data, etc. (Schafer and Graham, 2002; Batista and Monard, 2002). Demand management initiatives rely on good comprehension of water usage practices, as well as of factors influencing water demand (White et al., 2003). The emerging Data-Driven Demand Management has been supported by cloud-based data platforms and represents a new, critical element to improve decision-making in today's water industry. Utility managers can achieve the sustainability and affordability objectives they desire through the practical application of data analytics (Fielding et al., 2012). In this context, the implication of data gaps is really important, since the decision-making process relies on continuous data sets. Such continuous data sets improve the resilience of new decision-making schemes.

Based on the reason why a gap is created, missing data is categorized into three classes, depending on the level of randomness of the incident: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR) are commonly used classes that imply that the incident either does not depend on the missing value, or depends on a related to the value attribute, or directly depends on the value, respectively (Little and Rubin, 1987; Little, 1988). An example for a MCAR would be the interruption of functioning of a sensor. That would create a gap no matter what the measurements would be. An example for MAR would be the absence of answer in a questionnaire about an attribute that is indirectly related to the gender of the respondent. An example for NMAR would be the case of a sensor not recording a value, because it lies outside its measuring capacity range. Thus, a missing or erroneous value would imply that it is out of this range. The level of randomness is conclusive of the method that the missing data are treated. Depending on the class that the data gap belongs to, a different methodology for treating the missing data is selected.

Another criterion for choosing the method to treat an incident of missing value is the nature of the attribute. Specifically, if the attribute were a time-series, the treatment would involve analysis of components, such as trend and seasonality. Moreover, if the missing attribute value were correlated to another known attribute, then the method of treatment would be selected based on this correlation, which would imply the implementation of multivariate analysis, as opposed to univariate. Lastly, a criterion is the “length” of the missing part—this can vary from a single missing value to a larger gap of data. The aforementioned criteria are decisive of the treatment of an incident: variable methods are applied for this purpose. Some commonly applied tactics include ignoring and discarding the incident, case substitution mean or mode imputation, hot deck and cold deck method, applying a predictive model and others (Lakshminarayan et al., 1999; Grzymala-Busse and Hu, 2001; .Batista and Monard, 2003). The imputation of a missing value is generally classified into deterministic or stochastic (Rao, 1996).

Other than filling missing data gaps, the production of data that mimic the properties of a data set (synthetic data) can be essential in situations in which available real data are limited and longer data sets are required for evaluation, validation and/or testing of models, platforms, algorithms, or Decision Support Systems (DSSs). Barse et al. (2003) define synthetic data as ‘generated data by simulated users in a simulated system, performing simulated actions’. A typical example of need for synthetic data is the case when privacy constraints block the direct use of original sets. In other words, water utilities may not agree to provide actual water consumption data, being concerned about violating the privacy of their customers, even if data is anonymized. Synthetically generated data overcome problems related to data privacy (Cominola et al., 2016). In such cases, the use of a tool that provides synthetic water consumption data will serve well the needs of water utilities, including decision-making platforms used in data-driven demand management schemes. Another example is the training and adapting of a Fraud Detection System (FDS) on a synthetic data set, testing its properties by injecting synthetic frauds or comparing the performance of different FDSs (Barse et al., 2003).

Past research works have focused on investigating whether urban water consumption time series can be simulated in multiple