CSCE Annual Conference

Growing with youth – Croître avec les jeunes

Laval (Greater Montreal)

June 12 - 15, 2019

STOCHASTIC METHOD FOR PREDICTING LONG-TERM URBAN WATER CONSUMPTION

RasiFaghihi, N.[1], Li, S.S.[2], Haghighat, F.[3]

[1] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada, n_rasifa@encs.concordia.ca

[2] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada sam.li@concordia.ca

[3] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada, fariborz.haghighat@concordia.ca

Abstract: The purpose of this paper is to improve our understanding of how urban water consumption depends on a set of influencing variables. The traditional methods for predicting urban water consumption are typically based on historical data and assumed that the data are linear and stationary over time. A significant gap exists in the future changes and associated uncertainties of the influencing variables that have not been dealt with adequately. Consequently, it would be difficult to propose reliable planning and management strategies for water consumption sustainability. This paper extends the previous research of urban water consumption by treating consumption prediction as a stochastic process. The predictions explore the uncertainties associated with each influencing variables and a combination of some of these variables. To demonstrate its relevance, the proposed method is applied to analyse urban water consumption of the City of Brossard, Quebec, Canada. Daily records of urban water consumption are divided into: 1) base water consumption, which reflects winter consumption, and 2) seasonal use, which depends on seasonal and climatic variables. Various climate and socio-economic variables are investigated as the major influencing variables of urban water consumption. The analysis uses probabilistic data mining techniques and produces quantitative results of the correlations among the variables as well as their influences on urban water consumption.

Keywords: Urban water consumption; prediction; data mining; uncertainty analysis; the City of Brossard

1 INTRODUCTION

It is challenging to achieve sustainable management of urban water consumption (UWC). The challenges are partly due to an increase in urban population, rapid urbanization, and limited freshwater resources. This is particularly problematic for water-scarce countries. Even countries with abundant freshwater resources like Canada are facing the same issues. Canada has freshwater resources in different forms of water bodies; however, the sustainable management is a major issue in major cities and municipalities. Thus, it is important to study sustainable strategies for UWC.

This paper chooses the City of Brossard in the metropolitan area of Montreal, Quebec, Canada, as a study site. The main land-use types in Brossard are residential and commercial with many parks scattered through the city (Eslamian et al. 2016). The previous two censuses, conducted in 2011 and 2016, show that the city had a population of 85,721 in 2016, representing a percentage growth of 8.13% from 2011.

According to the city by-law (Brossard 2019a), watering with sprinklers not equipped with a timer is permitted for even-numbered addresses on Wednesdays, Fridays and Sundays, while for odd numbered addresses permitted days are Tuesdays, Thursdays and Saturdays (Brossard 2019b). This example points to the need to manage both peak and daily-averaged water demands. There is also a need to plan future consumption under a changing climate.

Previously, researchers have made good efforts to identify variables that influence UWC. Some researchers (Eslamian et al. 2016; Wong et al. 2010) decomposed complex UWC into components and associated each component with less complex factors. The decomposition gives base water consumption (BWC), seasonal water consumption (SWC), and calendrical use. Other researchers (Gato et al. 2007; Gato et al. 2005) decomposed UWC only into BWC and SWC. Parandvash and Chang (2016) proposed a regression model for estimates of per capita daily water demand. They formulated the demand as a function of seasonal variables; weather variables; indicators or dummy variables that reflect weekends, holidays and other data anomalies; unemployment rate; and long-term trend variables for the time period 1983-2012. Previous studies typically list temperature, precipitation, population, and income as key variables of UWC. Romano et al. (2014) expanded the list to include the altitude of the location in question; annual expenditure of residential households (tariff); utility ownership, including public utilities or non-public utilities; and the geographic location of the chief towns of Italian provinces, distinguishing between the northern, central, and southern regions of Italy. Romano et al. (2014) covered the time period of 2007-2009. From a different perspective, Kenney et al. (2008) suggested that water demand is a function of variables, which are within the control of water utilities. Such control includes water price; non-price strategies such as public education, technological improvement and water restriction; and other factors not related to weather and population.

In majority of existing studies of urban water demand, the prediction techniques used are not capable of addressing uncertainties associated with the future changes in key variables. The existing studies have used regression models (Eslamian et al. 2016; Parandvash and Chang 2016; Stoker and Rothfeder 2014; Chang et al. 2014; Wong et al. 2010; House-Peters et al. 2010; Adamowski and Karapataki 2010; Polebitski and Palmer 2009; Praskievicz and Chang 2009; Kenney et al. 2008; Ruth et al. 2008; Gato et al. 2005), linear mixed effect models (Romano et al. 2014), and other methods. These methods include factor analysis, wavelet transform, and support vector machine. Some of the key variables are a random variable. The existing studies have not adequately addressed the issue of uncertainties associated with random variables that influence UWC. As a result, it has been a big challenge to develop reliable planning and management strategies for sustainable water consumption.

UWC and its drivers contain abundant valuable information and data of climatic and socio-economic variables. Data mining is a powerful technique to extract valuable information from the data. Data mining is the process of extracting interesting patterns or knowledge from a huge amount of data, preferably in an efficient, scalable, and practical manner (Han and Kamber 2011). Previously, researchers adopted various data mining approaches and coupled them to create more advanced and capable tools for data mining. Yu et al. (2011) proposed clustering, decision tree, and association rule mining for studying of occupants' behavior in residential buildings. Singh and Yassine (2018) used clustering analysis, association rule mining, and Bayesian network for the analysis and forecast of energy consumption time series. They extracted various temporal energy consumption patterns. Adamowski et al. (2012) proposed a method based on coupling discrete model wavelet-neural for short-term forecast of water consumption. They suggested using the bootstrap method in order to account for uncertainty, which has not been considered in the coupling discrete model. Tiwari and Adamowski (2013) developed a coupled model (wavelet-bootstrap-neural network) for short-term forecasting.

A further improvement of data mining techniques for application to UWC prediction is needed in order to obtain results that are more reliable. The purpose of this paper is to improve our understanding of how UWC depends on a set of key variables. Some of them display random characteristics. We propose a stochastic method for predicting UWC under changes in climatic variables. The method uses data mining techniques. In the following sections, we discuss the methodology (Section 2) and the results (Section 3), and finally conclusions (Section 4).

2 METHODOLOGY

2.1 Data of daily water consumption

In this paper, we use records of daily water consumption of the City of Brossard over the time period of January 2011 to October 2015. For this time period, we obtained climatic and socio-economic data. The climatic data include daily minimum temperature (T_{min}), daily maximum temperature (T_{max}), daily mean temperature (T_{mean}) and total precipitation. The climatic data were measurements made at Pierre Elliott Trudeau International Airport (45° 28′11.06″ N, 73°44′41.71″ W). The data source is Environment Canada. The socio-economic variables are population, age, water price, and household income, acquired from Statistics Canada. Following previous studies, as an initial trial, we divided the records of daily water consumption into: 1) BWC, which reflects winter consumption, and depends on such socio-economic variables as population, age and water price, and 2) SWC, which depends on seasonal and climatic variables. Using correlation methods, we assess a possible two-way linear association between two continuous variables. Since SWC and BWC are two separate drivers, we perform a correlation analysis for each.

2.2 Preprocessing data

It is important to pre-process data, to find outliers, replace missing values and transform the values. In this paper, the applicable preprocessing step in accordance with the available data is detecting outliers. Outliers are values, which behave differently from expectation (Han and Kamber 2011). Several approaches to outlier detection exist, including clustering-based methods and statistical methods. The statistical methods assume a normal distribution of data; therefore, values in a low probability region are considered as outliers. The clustering-based methods accept that outliers might belong to a small or sparse cluster or might be far from the clusters to which they are closest.

We use both the clustering-based and statistical methods in this paper. By grouping UWC data, we identify potentially unexpected behaviours and reveal their hidden patterns on the basis of variations in air temperature. Subsequently, we apply the concept of maximum likelihood in statistical methods to those clusters that contain values with unusual behaviours. Specifically, UWC values outside the range of $\mu \pm 2\sigma$ are labeled as the outlier, where μ is the mean value, and σ is the standard deviation. Note that $\mu \pm 2\sigma$ contains 95% of data under the assumption of normal distribution.

2.3 Cluster analysis

The cluster analysis is a process of dividing the observed records into classes or clusters so that objects in the same cluster have a high similarity, while objects in a different cluster have a low similarity. k-means approach was used in this study. Clustering can be regarded as a form of classification, which creates labeling of objects with cluster labels derived from data. Hence, this methodology might be referred to as unsupervised classification. We ignore the effect of socio-economic variables and focus on climatic variables. Grouping the data by weather conditions leads to the identification temporal water consumption. Note that by applying clustering analysis, potential outliers are also detected.

2.4 Bayesian network

Bayesian networks are a graphical model based on Bayes' theorem. They are capable of modeling probabilistic relationships among a set of variables (Heckerman, 1997). This methodology is considered as classification approach and it can potentially be used for prediction (Hawarah et al. 2010). One of the most significant characteristics of this approach is their ability to account for uncertainties associated with inaccurate and incomplete databases. The idea is to reveal the interdependency of variables in the form of probability distribution. In the Bayesian networks, variables of interest are represented as child nodes and parent nodes, and the links between them indicate informational or causal dependencies among them (Ismail et al. 2011).

We construct a Bayesian network in two steps. The first step is structure learning, which produces a graphical structure of dependencies between nodes. Even though a number of machine-learning algorithms are available for the determination of the number of parents based on the strength of the relationship between each pair of variables, this paper develops the structure on the basis of the meaningful interdependency of predictors and predictant. The second step is parameter learning, which determines a conditional probability distribution among nodes. Initially, the number of intervals should be defined on the basis of user's choice. In this study, the variables are continuous. As a result, we can discretize data in either equal distance or equal frequency. The latter is less sensitive to outliers and provides better accuracy, it is chosen in this paper.

3 RESULT

Clustering analysis is unsupervised learning due to the fact that the class label is unknown. It is necessary to determine a clustering algorithm and the number of desired clusters. In k-means algorithm, we evaluate the sufficient number of clusters and the records in each cluster, using the elbow method.

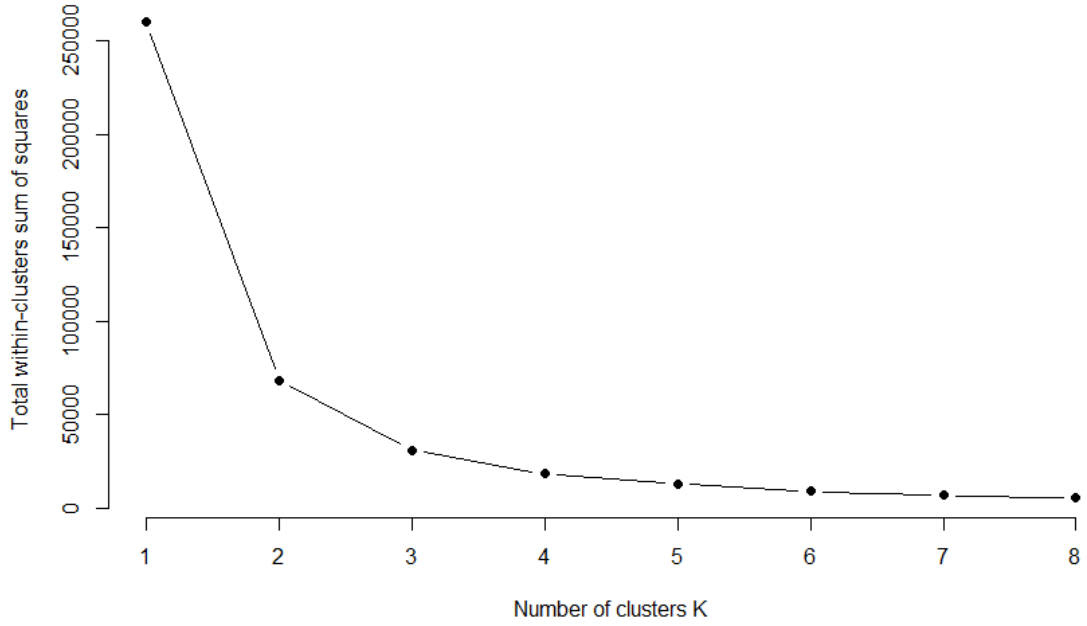

Figure 1: Determination of the number of clusters using elbow method

In Figure 1, it is clear that three clusters will be enough to create the clusters. Generally speaking, air temperature in the region of Montreal varies considerably during the day. Therefore, clustering analysis includes SWC, T_{min}, T_{max} and even T_{mean} as variables.

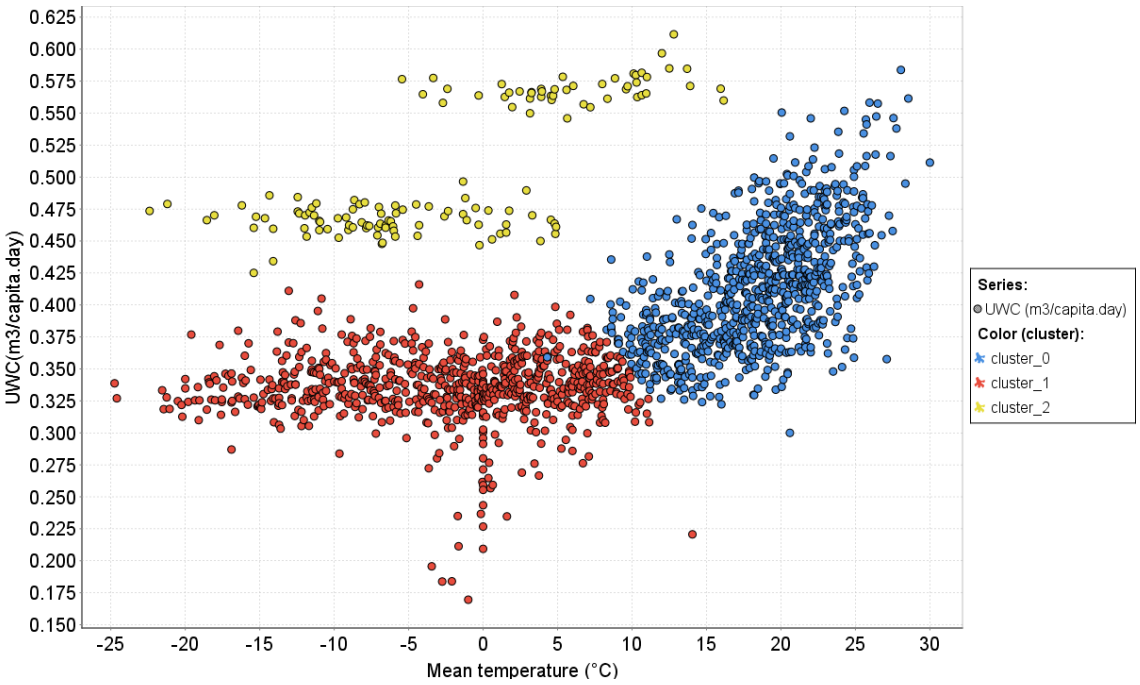

Figure 2: UWC vs. T_{mean}, showing the clusters (red, blue and yellow data points)

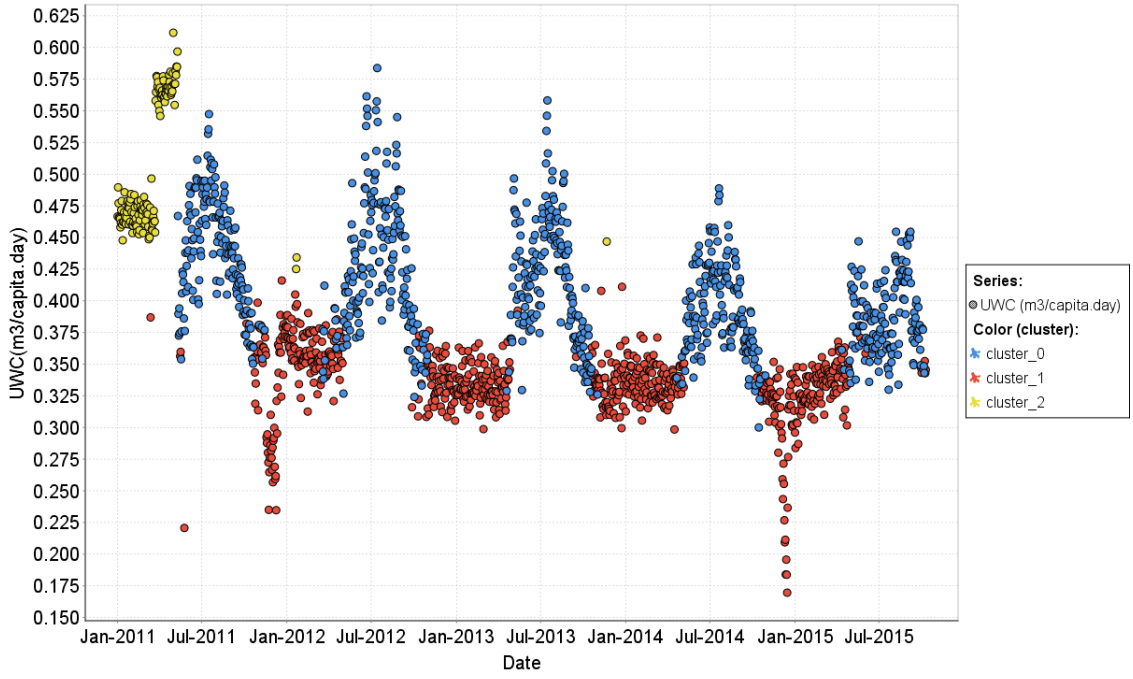

Figure 3: Time series of UWC, showing the clusters (red, blue and yellow data points)

According to Figures 2 and 3, UWC has meaningful patterns: cluster 1 (red data points) represents BWC, which is mostly the water consumption in cold season (between November and April); cluster 0 (blue data points) is SWC, which includes warm season (between May and October). Cluster 0 confirms the positive correlation between SWC and T_{mean}. On the contrary, the data points for the period of January – May 2011

do not follow the dominant patterns of UWC for the other years. These data points, along with three other data points for 2012 and 2013, belong to cluster 2 (yellow data points) and are considered as outliers. Moreover, in Figure 3, a number of BWC data points display unexpected behaviours. They are removed using the maximum likelihood approach and are excluded from subsequent analysis.

We present the results from correlation analysis for both SWC and BWC, along with their key influencing variables, in Tables 1 and 2. In Table 1, the population is organised into three age groups: child (0 – 14 year old), young adult (15 – 29 year old), and senior adult (30 years or older).

Table 1: Correlation coefficient for BWC and its key influencing variables

Attributes	Child	Young adult	Senior adult	Payment (CAD/ household)	Water price (CAD)	Income (CAD/ household)	BWC (m^3/ capita.day)
Child	1	1	1	0.935	0.983	0.996	-0.314
Young adult	1	1	1	-0.935	-0.983	-0.996	0.314
Senior adult	1	1	1	0.935	0.983	0.996	-0.314
Payment (CAD/household)	0.935	-0.935	0.935	1	0.971	0.942	-0.288
Water price (CAD)	0.983	-0.983	0.983	0.971	1	0.984	-0.295
Income (CAD/household)	0.996	-0.996	0.996	0.942	0.984	1	-0.292
BWC (m^3/capita·day)	-0.314	0.314	-0.314	-0.288	-0.295	-0.292	1

Table 2: Correlation coefficient for SWC and key influencing variables

Attributes	T_{max} (°C)	T_{min} (°C)	T_{mean} (°C)	Total precipitation (mm)	SWC (m^3/capita·day)
T_{max} (°C)	1	0.758	0.940	-0.081	0.615
T_{min} (°C)	0.758	1	0.928	0.124	0.0497
T_{mean} (°C)	0.940	0.928	1	0.021	0.591
Total precipitation (mm)	-0.081	0.124	0.021	1	-0.130
SWC (m^3/capita·day)	0.615	0.497	0.591	-0.130	1

As expected, the correlations among different groups of people's age are strong as well as the correlations among water price, household income and payment (Table 1). The important information in Table 1 is the direction of the correlation. Among the BWC key influencing variables, the child and senior adult groups of population age have a negative correlation with BWC, whereas the young adult group has a positive correlation with BWC. This means that raising young population leads to an increase in BWC. Besides, BWC has a negative correlation with water price, household payment and household income.

Table 2 demonstrates that SWC has a moderate, positive correlation with daily maximum temperature, daily mean temperature, and daily minimum temperature. In addition, SWC has almost no correlation with total precipitation. This is due to considerable variations in the total precipitation and the presence of nearly zero values of precipitation.

The results from the clustering analysis as well as correlation analysis validate the hypothesis that the seasonal variations in air temperature during the year have a strong influence on UWC. Therefore, we develop a Bayesian network based on SWC. The network could be use for predicting UWC probability distribution in response to climate change. Even though total precipitation appears to have almost no correlation with UWC, we include it in the development of Bayesian network.

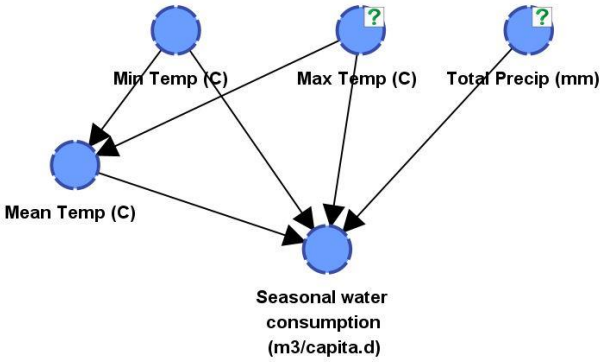

Figure 4: Bayesian network demonstrating SWC and key influencing variables

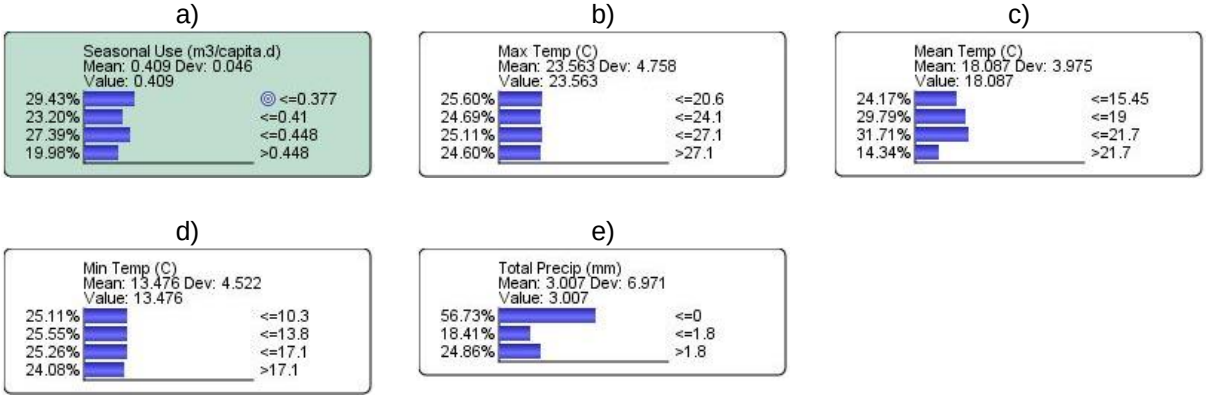

Figure 5: Probability distributions of variables: a) SWC; b) T_{max}; c) T_{mean}; d) T_{min}; e) Total precipitation

Figure 4 shows SWC and the key influencing variables. Arrows demonstrate the effect of key influencing variables on SWC as well as the influence of T_{max} and T_{min} on T_{mean}. Those variables, which have missing data, are labeled with a question mark. In order to validate the accuracy of the developed model, we split the dataset into 20% as test data and 80% as training data. The results show that the p-value is 2.47%, which proofs the high confidence of the developed Bayesian network. The network in validation mode reveals the probability distribution of all variables and provides the chance to manipulate the probabilities of different states to observe the outcomes on other variables (Figure 5). In fact, if T_{max} is higher than 27.1°C, the most probable interval of SWC will change to "more than 0.448 (m^3/capita·day)" by 17%.

With the Bayesian network, we are able to not only manipulate the probabilities of intervals of child nodes and demonstrates the new probability distribution of parent node, but also obtain the variation of the probability distribution of child nodes based on the absolute interval of the parent node. We demonstrate these useful features by showing the new probability distributions of variables in the condition that T_{mean} is in the range of 19 – 21.7°C with 100% probability (Figure 6), and the probability distribution of T_{max}, T_{min}, T_{mean} and total precipitation in the case that the probability of occurrences of SWC > 0.448 (m^3/capita·day) is 100% (Figure 7).

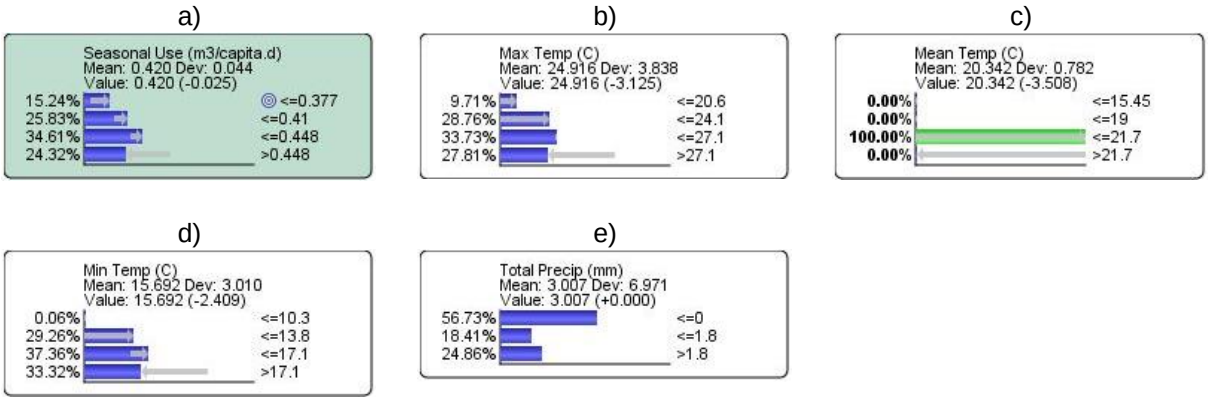

Figure 6: New probability distributions of variables when T_{mean} is in the range of 19 – 21.7°C; a) SWC; b) T_{max}; c) T_{mean}; d) T_{min}; e) Total precipitation

Figure 6 shows that when T_{mean} is in the range of 19 – 21.7°C, the relationships between T_{min}, T_{max} and T_{mean} give the most probable interval of T_{min} and T_{max}, being 13.8 – 17.1°C and 24.1 – 27.1°C, respectively. In addition, the probability distribution for a target variable will change to the condition where SWC is most probable to be 0.410 – 0.448 (m^3/capita·day).

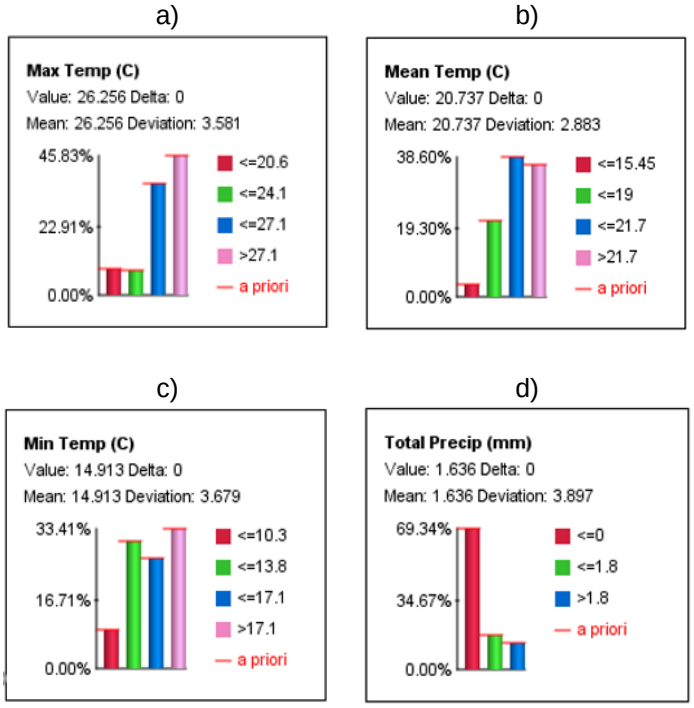

Figure 7: Conditional probability of key influencing variables when SWC > 0.448 (m^3/capita·day); a) T_{max}; b) T_{mean}; c) T_{min}; d) Total precipitation

Figure 7 shows that the occurrence probability of SWC > 0.448 (m^3/capita·day) is 19.98%. Consider the situation that this probability is 100%. The associated probability distribution of key influencing variables will change. For instance, the associated probability of T_{mean} occurring between 19 – 21.7°C and above 21.7°C changes from 31.71% to 38.60% and from 14.34% to 36.74%, respectively.

4 CONCLUSION

This paper describes the development of a procedure to improve our understanding of UWC based on data mining. Daily data of UWC decomposes into BWC and SWC. Clustering analysis of the data leads to the detection of a significant number of outlier values in the data. The detection makes use of the maximum likelihood method. We conclude that the young adult group has a positive, moderate correlation with UWC. BWC has a negative correlation with water price, household payment and household income. SWC has a moderate, positive correlation with daily maximum temperature, mean temperature, and minimum temperature, but has almost no correlation with total precipitation. We develop a Bayesian network for the analysis of SWC to in order to reveal the dependency of water on climatic variable. Under the condition that the mean temperature ranges from 19 to 21.7°C. SWC is most probable to be 0.410 to 0.448 m^3/capita·day. The occurrence probability of SWC > 0.448 m^3/capita·day being 100% is associated with the probability of occurrence T_{mean} between 19–21.7°C rising to 38.60% from 31.71% and T_{mean} above 21.7°C raising to 36.74% from 14.34%. The methods discussed in this paper are useful for sustainable water consumption management.

References

Adamowski, J. and Karapataki, C. 2010. Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN learning algorithms. *Journal of Hydrologic Engineering*, **15**(10), 729-743.

Adamowski, J. Fung Chan, H. Prasher, S. O. Ozga-Zielinski, B. & Sliusarieva, A. 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research*, **48**(1).

Adamowski, J. Adamowski, K. & Prokoph, A. 2013. A spectral analysis based methodology to detect climatological influences on daily urban water demand. *Mathematical Geosciences*, **45**(1), 49-68.

Brossard, 2019a. Municipal services, By-Law http://www.ville.brossard.qc.ca/services-citoyens/Reglements/Reglements.aspx?lang=en-ca, Montreal, Canada.

Brossard, 2019b. Municipal services, Watering, http://www.ville.brossard.qc.ca/services-citoyens/eau/Eau/Arrosage.aspx?lang=en-ca, Montreal, Canada.

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. *International Journal of Geospatial and Environmental Research*, **1**(1), 7.

Eslamian, S. A. Li, S. S. & Haghighat, F. 2016. A new multiple regression model for predictions of urban water use. *Sustainable Cities and Society*, 27, 419-429.

Gato, S. Jayasuriya, N. Hadgraft, R. & Roberts, P. 2005. A simple time series approach to modelling urban water demand. *Australasian Journal of Water Resources*, **8**(2), 153-164.

Gato, S. Jayasuriya, N. & Roberts, P. 2007. Temperature and rainfall thresholds for base use urban water demand modelling. *Journal of hydrology*, **337**(3-4), 364-376.

Han, J. Pei, J. & Kamber, M. 2011. *Data mining: concepts and techniques*. 3rd ed. Elsevier., Waltham, MA, USA.

Heckerman, D. 1997. Bayesian networks for data mining. *Data mining and knowledge discovery*, **1**(1), 79-119.

House-Peters, L. Pratt, B. & Chang, H. 2010. Effects of Urban Spatial Structure, Sociodemographics, and Climate on Residential Water Consumption in Hillsboro, Oregon 1. *Journal of the American Water Resources Association*, **46**(3), 461-472.

Kenney, D. S. Goemans, C. Klein, R. Lowrey, J. & Reidy, K. 2008. Residential water demand management: lessons from Aurora, Colorado1. *Journal of the American Water Resources Association*, **44**(1), 192-207.

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. *Journal of Hydrology*, **538**, 574-586.

Polebitski, A. S. & Palmer, R. N. 2009. Seasonal residential water demand forecasting for census tracts. *Journal of Water Resources Planning and Management*, **136**(1), 27-36.

Praskievicz, S. & Chang, H. 2009. Identifying the relationships between urban water consumption and weather variables in Seoul, Korea. *Physical Geography*, **30**(4), 324-337.

Romano, G. Salvati, N. & Guerrini, A. 2014. Estimating the determinants of residential water demand in Italy. *Water*, **6**(10), 2929-2945.

Ruth, M. Bernier, C. Jollands, N. & Golubiewski, N. 2007. Adaptation of urban water supply infrastructure to impacts from climate and socio-economic changes: the case of Hamilton, New Zealand. *Water Resources Management*, **21**(6), 1031-1045.

Singh, S. & Yassine, A. 2018. Big data mining of energy time series for behavioral analytics and energy consumption forecasting. *Energies*, **11**(2), 452.

Stoker, P. & Rothfeder, R. 2014. Drivers of urban water use. *Sustainable Cities and Society*, **12**, 1-8.

Wong, J. S. Zhang, Q. & Chen, Y. D. 2010. Statistical modeling of daily urban water consumption in Hong Kong: Trend, changing patterns, and forecast. *Water resources research*, **46**(3).

Yu, Z., Fung, B. C. Haghighat, F. Yoshino, H. & Morofsky, E. 2011. A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, **43**(6), 1409-1417.