Environmental Modelling & Software 102 (2018) 29–38

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study

Isaac Duerr [a, 1], Hunter R. Merrill [a, 1], Chuan Wang [b], Ray Bai [b], Mackenzie Boyer [a], Michael D. Dukes [c], Nikolay Bliznyuk [d, *]

[a] Graduate Assistants, Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611, United States
[b] Graduate Assistants, Department of Statistics, University of Florida, Gainesville, FL 32611, United States
[c] Professor, Department of Agricultural and Biological Engineering, Director, Center for Landscape Conservation and Ecology, University of Florida, Gainesville, FL 32611, United States
[d] Assistant Professor of Statistics, Departments of Agricultural and Biological Engineering, Biostatistics and Statistics, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:
Received 14 May 2017
Received in revised form
27 December 2017
Accepted 5 January 2018

Keywords:
Predictive modeling
Spatial modeling
Time series
Tree-based methods
Uncertainty quantification
Urban water use

ABSTRACT

Forecasts of water use are crucial to efficiently manage water utilities to meet growing demand in urban areas. Improved household-level forecasts may be useful to water managers in order to accurately identify, and potentially target for management and conservation, low-efficiency homes and relative high-demand customers. Advanced machine learning (ML) techniques are available for feature-based predictions, but many of these methods ignore multiscale spatiotemporal associations that may improve prediction accuracy. We use a large dataset collected by Tampa Bay Water, a regional water wholesaler in southwest Florida, to evaluate an array of spatiotemporal statistical models and ML algorithms using out-of-sample prediction accuracy and uncertainty quantification to find the best tools for forecasting household-level monthly water demand. Time series models appear to provide the best short-term forecasts, indicating that the temporal dynamics of water use are more important for prediction than any exogenous features.

1. Introduction

Rapidly increasing urban populations have put new stress on water utilities as they try to meet growing demands for water. Effective management of water utilities requires anticipating future water use both in the short term, in order to efficiently meet demand, and over longer temporal ranges, so that they may best allocate scarce investment resources. Accurate household-level forecasts may be used for correctly anticipating future demand and identifying households that are predicted to use large quantities of water (relative to lot and/or household size) in the future as potential targets for efficiency monitoring. However, creating these accurate predictive models for utilities data can be challenging due to the vast amounts of data involved.

Donkor et al. (2014) summarized water demand forecasting methodologies used by utilities. Some of these methods rely on simple per capita usage averages, which have considerable deviation around the mean. The public-supply gross per capita use in Florida was 134 gallons per person per day in 2010, but ranged from less than 100 to more than 200 by county (Marella, 2014). Others use simulations and scenario-based decision systems to model future water demand (Williamson et al., 2002; Billings and Jones, 2008; Polebitski et al., 2011). More advanced forecasting techniques have been proposed using time series models including linear regression, autoregressive integrated moving average (ARIMA) models, stochastic process models, and artificial neural networks (ANN). The success of ANN models in particular (Billings and Jones, 2008; Bennett et al., 2013; Tiwari and Adamowski, 2015) suggests that additional nonparametric machine learning (ML) techniques could produce similar or superior forecasts. However, very few studies have assessed the quality or uncertainty of water

* Corresponding author.
E-mail address: nbliznyuk@ufl.edu (N. Bliznyuk).
[1] Duerr and Merrill are first co-authors.

https://doi.org/10.1016/j.envsoft.2018.01.002
1364-8152/

demand forecasts made by machine learning methods. This paper seeks to address this gap in the literature. Other studies have modeled water use on finer timescales (e.g., daily and hourly) with the data recorded by smart meters (Herrera et al., 2010; Cominola et al., 2015). However, not all water providers collect data at the daily or hourly resolution.

Specialized approaches have been developed that focus on household demand and in particular that of single family homes. DeOreo and Mayer (2012) provide a thorough end use analysis of fixtures and various uses of water in the household. This detailed analysis results in a precise per capita use estimate for households. Presumably this information could be used to forecast future water demand if one knew the relative growth and composition of future household construction. However, even knowing details on end uses requires knowledge of population growth and prediction of new construction trends. DeOreo and Mayer (2012) point out that indoor use declined from 187 gpd/household on homes built in the early 1990s to 162 gpd/household for homes built around 2007 and as low as 107 gpd/household for new high-efficiency homes. This decrease of indoor demand was primarily due to high efficiency toilets and clothes washers mandated and used in newer construction. Similarly, Abdallah and Rosenberg (2014) attributed changes in indoor water use over time to efficiency gains in technology and not to conservative behavior. However, DeOreo and Mayer (2012) point out that outdoor use did not change between the older and newer homes. This observation is important since outdoor use (e.g., irrigation) in southern states like Florida has been shown to be as high as 74% of total household water use in typical landscapes (Haley et al., 2007).

The purpose of this paper is to evaluate alternatives for water demand forecasting and to compare several advanced machine learning techniques and spatio-temporal models that are available for feature-based predictions. Our goal is to determine if these alternative models can improve prediction and uncertainty quantification of the forecasts, as judged by root mean squared error (RMSE) and prediction interval-based metrics (such as the width and empirical coverage rate), respectively. The ML algorithms considered in this paper are Random Forest (RF), Bayesian additive regression trees (BART), and gradient boosting algorithms (GBM) (see, e.g., Friedman et al., 2001, for an overview and algorithmic details of these models). These methods make use of exogenous features but do not explicitly take into account multiscale spatial and temporal associations that may improve predictive quality of the models in high spatio-temporal resolution (e.g., monthly household water demand). For comparison, we also evaluate the prediction accuracy and uncertainty quantification of an array of statistical models which explicitly include parametric dependence and semiparametric trend structures to account for spatio-temporal dynamics and compare these models with our nonparametric ML algorithms in order to identify the best tools for forecasting water demand. The models developed here do not make any assumptions about rates of urban growth, changes in construction methods, improvements in water efficiency, or other global drivers of water use. If the true relationship between the features used to create the models and water usage is altered by changes in exogenous factors, the models need only be refit using updated data. The only constraint on the models adapting to exogenous changes is the rate at which new data becomes available which includes these altered dynamics.

The data set used for this study contains monthly total water usage by single family homes from Tampa Bay Water (TBW), a water authority operating within the Southwest Florida Water Management District (SWFWMD), for three member governments located in Hillsborough, Pasco, and Pinellas counties recorded approximately from 1998 to 2010, with some variation by region (Boyer et al., 2014). The data includes monthly water usage collected for each household (parcel) based on billing records. Parcel and billing records were augmented with environmental covariates including precipitation, evapotranspiration, and temperature. This case study allows us access to a large amount of high-quality data which has been screened for accuracy and completeness and hence presents an excellent opportunity to evaluate ML and spatio-temporal models for predicting water usage within a municipal utility.

This manuscript is organized as follows. Section 2 outlines the available data and the covariates used in our study. Section 3 describes the models used in our analysis and the criteria by which they are evaluated. Section 5 presents our model comparisons and Section 6 provides further discussion of these results and directions for possible future study. As per Editor's request, we provide magnified versions of all figures in the Supplements at the end of the manuscript.

2. Data

2.1. Water use data

The dataset considered here consists of monthly potable water billing records water usage data from TBW for parts of three counties in east-central Florida (Hillsborough, Pasco, and Pinellas) and recorded approximately from 1998 to 2010, with some variation by region (Boyer et al., 2014). The data includes between 6 months and 12 years of monthly records from over one million unique customers, with water usage collected in monthly increments for each unique parcel-based on billing records. For this study, we use only parcels which contain exactly one single-family residence to focus on household-level water demand. The data is both spatial and temporal in nature with observations throughout the Tampa Bay region. This paper focuses on forecasting the total water usage per month in each parcel. To increase the number of statistical and ML methods for comparison and to mitigate computational burdens in comparing the methods, our analysis focused on a representative subset of over 100,000 observations (household-month records) within the City of Tampa, shown in Fig. 1 (c). To allow for the full scope of spatio-temporal models to be used for forecasting, we focused on 973 households without missing data over 137 months (see Section 4.2.2 for details on this constraint).

The data exhibits significant spatio-temporal heterogeneity. Fig. 1 shows the location of the study within Florida, and the spatial distribution of the households and water usage included in this study.

Fig. 2 shows the total monthly water use for each household as multiple time series. While there are aggregate trends, the patterns within individual households are much less regular. The highlighted lines indicate the monthly water usage of two typical individual households which shows strong autocorrelation and possible seasonal variation as well as less structured, household-specific temporal variation. The average across all households is also shown, and the figure indicates that the variation of household water use around the mean is quite large.

As an exploratory step, we view the spatial heterogeneity and temporal structure of the data aggregated to census blocks in Fig. 3(a), which shows the total water use for each census block averaged over time and indicates that some spatial dependence may be present in the data, but with significant between-household variation observed within even very small spatial areas. The totals are highly skewed and many billing records are missing from the full data set. A simple exploratory time series model was fit to each census block, and the estimated one-month