Environmental Modelling & Software 105 (2018) 257–269

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Automated detection of unusual soil moisture probe response patterns with association rule learning

Ziwen Yu [a,*], Alex Bedig [b], Franco Montalto [a], Marcus Quigley [b]

[a] Civil, Architectural, and Environmental Engineering, Drexel University, 3141 Chestnut Street, Philadelphia, PA, USA
[b] OptiRTC, Inc., 356 Boylston St, Boston, MA 02116, USA

ARTICLE INFO

Article history:
Received 16 January 2018
Received in revised form
26 March 2018
Accepted 3 April 2018

Keywords:
QAQC
Association rule learning
Green infrastructure
Anomalous pattern detection
Dynamic time warping

ABSTRACT

In-situ field monitoring networks generate vast quantities of continuous data can help to improve the design, management, operation and maintenance of Green Infrastructure (GI) systems. However, such actions require efficient and reliable quality assurance quality control (QAQC). In this paper, we develop a rule-based learning algorithm involving Dynamic Time Warping (DTW) to investigate the feasibility of detecting anomalous responses from soil moisture probes using data collected from a GI site in Milwaukee, WI. As an enhancement to traditional QAQC methods which rely on individual time steps, this method converts the continuous time series into event sequences from which response patterns can be detected. Association rules are developed on both environmental features and event features. The results suggest that this method could be used to identify abnormal change patterns as compared to intra-site historical observations. Though developed for soil moisture, this method could easily be extended to apply on other continuous environmental datasets.

1. Introduction

In-situ monitoring with sensors is increasingly being used to improve the design and management of Green Infrastructure (GI) systems in urban areas, also creating new opportunities for scientific discovery (Porter et al., 2009) at both private and public sites. For example, DiGiovanni et al. (2010) characterized the hydrologic performance of a small scale green roof with data collected from a custom designed lysimeter, a rain gage and soil moisture sensors; Carson et al. (2013) performed a similar study using an Onset Hobo U30 weather station with a tipping bucket rain gauge and a custom designed weir; Berretta et al. (2014) investigated the ET characteristics of a green roof by tracking the soil moisture content and temperature using Campbell Scientific CS616 Water Content Reflectometer.

The potential for deriving new insights from these efforts work is, however, contingent upon good quality data, a goal not always easy to achieve given the ease with which data streams can be corrupted (Hill et al., 2009) in the harsh environmental conditions found in GI systems. Erroneous data frequently occurs in hydrometeorological monitoring (Sciuto et al. (2009), and poor quality data may even significantly hinder analysis, inhibit modeling, and lead to poor decision making (Madsen, 1989). New approaches to automate Quality Assurance and Quality Control (QAQC) are thus necessary to service the expanding volume of data collected by growing sensors networks (National Science Foundation, 2005). Overall, the purpose of QAQC is to detect the anomalies by both time steps (e.g. an unsensational value) and events (e.g. odd response series with strong autocorrelation).

Traditional QAQC protocols emphasize detection of anomalies and outliers for individual time points only, and can broadly classified into three categories. The first is the **record/limit check method,** which compares individual sensor values to a pair of upper and lower limits. For example, Durre et al. (2010) used global temperature, precipitation, snowfall, and other extremes to flag sensor data that was physically impossible because it fell outside the range of values observed anywhere on the Earth.

The second includes a family of **statistical methods** that define outliers as sensor values that fall in the far head or tail of the frequency distribution at a specific location and time. A z-score, derived from the mean and standard deviation of the normalized data values, can be used, for example, to estimate how likely a sensor value is to be an outlier. By setting a specific confidence threshold, z-score exceedances can be used to flag outliers

* Corresponding author.
E-mail address: zy32@drexel.edu (Z. Yu).

https://doi.org/10.1016/j.envsoft.2018.04.001

(Hubbard and You, 2005; Kunkel et al., 2005). Such methods are extensions of traditional statistical process control techniques and assume that the variability in the underlying phenomena is constant over time (e.g. stationary) such that deviations from historical distributions indicate anomalous behaviors. Such assumption do not always hold in GI systems where natural processes such as soil erosion, or human-included perturbation such as irrigation can lead to gradual and abrupt changes in the field.

The third, and most recent category of approaches are **machine learning methods.** Unsupervised machine learning methods such as k-means have become very popular for outlier detection (Bolton and Hand, 2001). Bayesian methods such as Dynamic Bayesian networks (DBNs) identify anomalies by dynamically tracking changes in the historical data (Hill et al., 2009). Supervised learning methods such as neural networks (Kozma et al., 1994; López-Lineros et al., 2014), support vector machines (SVM) (Buluta et al., 2005) and decision trees (John, 1995) can classify sensor values as either valid or invalid based on the properties of a training dataset. To date, application of these methods has principally focused on each time stamped data point, frequently underestimating the autocorrelation inherent to the time series data.

As mentioned previously, QAQC methods should scrutinize both individual points and point series in an event. One short come of traditional machine learning methods is lack the ability to identify autocorrelation in patterns of change within, and between, events. To enhance the traditional methods to complete QAQC, similarity comparisons on patterns or shapes need to be applied across the time series events generated by sensors. Using GI monitoring as an example, the QAQC process should align with our physical understanding of a GI site's rainfall/runoff response, infiltration characteristics as well as the surrounding microclimatic limits of evapotranspiration. A first step would confirm whether the raw observations are responding consistently to historical change patterns. Few, if any, existing QAQC methods in hydrology monitoring address this need.

This paper aims to improve data quality in terms of detecting unusual change series patterns in situations where historical observations are of inadequate quality. The approach includes a rule-based machine learning algorithm to investigate the feasibility of detecting anomalous response patterns from soil moisture probes within a single site and across multiple sites with similar configurations. The data used to demonstrate the technique are collected at a GI site in Milwaukee, WI. To enhance traditional QAQC methods, which focus on outlier detection on a single time step without considering the autocorrelation across multiple time points, this new method splits the whole data set into chunks based on precipitation events and employs Dynamic Time Warping (DTW) to align and compare events at different lengths. The paired-event distances calculated by DTW are generalized to categorize the similarity between events. Other, more stable observed site phenomena are also included in the model to reflect the physical conditions that underlie the association rules.

2. Methodology

2.1. Data and site

The monitored GI site is a green roof located in UWM Golda Meir library in Milwaukee, WI. The site is about 1350 m^2 and has a growth media depth of 0.1524 m. It is monitored by OptiRTC (optirtc.com) using soil moisture probes installed at the bottom of the growth media (Decagon EC-5 Soil moisture sensor with hobo link data logger). Three soil moisture probes are installed at different locations at the site (see Fig. 1). Instant precipitation and running median temperature are also recorded by a climate station located in the same site. These data were collected at 5-min

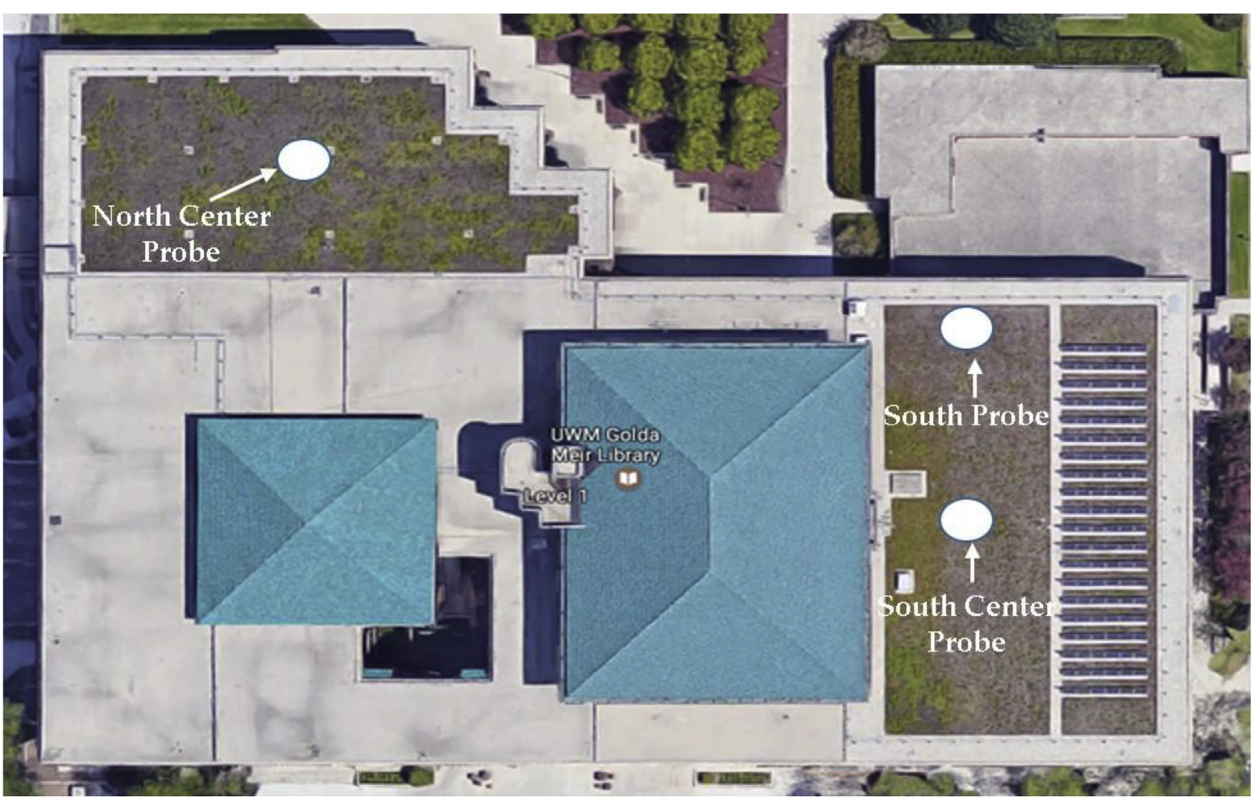

Fig. 1. Overview of the monitoring site.