Journal of Hydrology 519 (2014) 909–917

Contents lists available at ScienceDirect

Journal of Hydrology

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

A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations

Andrea Bichler [a], Arnold Neumaier [b], Thilo Hofmann [a,*]

[a] University of Vienna, Department of Environmental Geosciences, Althanstrasse 14, UZA2, 1090 Vienna, Austria
[b] University of Vienna, Department of Mathematics, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

ARTICLE INFO

Article history:
Received 15 November 2013
Received in revised form 8 July 2014
Accepted 4 August 2014
Available online 11 August 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Niko Verhoest, Associate Editor

Keywords:
Groundwater quality
Drinking water
Faecal indicator bacteria
Total coliforms
Classification tree
CHAID

SUMMARY

Microbial contamination of groundwater used for drinking water can affect public health and is of major concern to local water authorities and water suppliers. Potential hazards need to be identified in order to protect raw water resources. We propose a non-parametric data mining technique for exploring the presence of total coliforms (TC) in a groundwater abstraction well and its relationship to readily available, continuous time series of hydrometric monitoring parameters (seven year records of precipitation, river water levels, and groundwater heads). The original monitoring parameters were used to create an extensive generic dataset of explanatory variables by considering different accumulation or averaging periods, as well as temporal offsets of the explanatory variables. A classification tree based on the Chi-Squared Automatic Interaction Detection (CHAID) recursive partitioning algorithm revealed statistically significant relationships between precipitation and the presence of TC in both a production well and a nearby monitoring well. Different secondary explanatory variables were identified for the two wells. Elevated water levels and short-term water table fluctuations in the nearby river were found to be associated with TC in the observation well. The presence of TC in the production well was found to relate to elevated groundwater heads and fluctuations in groundwater levels. The generic variables created proved useful for increasing significance levels. The tree-based model was used to predict the occurrence of TC on the basis of hydrometric variables.

1. Introduction

Safe drinking water is essential to good health and a basic human right; it is considered by the WHO (2011) to be a component of effective policy for health protection. Although drinking water is subject to strict quality controls (EC, 1998; USEPA, 1996) and a great amount of effort is put into protecting water resources, a certain level of risk cannot be avoided. Apart from chemical hazards, microbial hazards are of particular concern for water suppliers. This is due to the fact that source water quality can vary rapidly with respect to microbial parameters, while chemical properties are generally subject to long term variations (Dechesne and Soyeux, 2007). Waterborne diseases can also result from very limited exposure to contaminated water (Macler and Merkle, 2000). Many regulations and guidelines promote a preventive approach to mitigate the risks to drinking-water quality: this strategy emphasises the identification of possible sources of contamination (USEPA, 1996). A multi-barrier approach has been proposed by the WHO (2011) as a first step towards microbial safety, focusing on the reduction of pathogen entry into water sources and raw water. The identification of potential hazards and hazardous events is a basic requirement for developing effective mitigation measures to secure water quality and reduce the purification treatment required. Since microbial contamination cannot always be prevented, the prediction of microbial pollution is of great interest in water management.

A useful first step towards identifying possible contamination sources is to analyse any existing data that might be available. Since water suppliers are often subject to strict regulation (EC, 2000, 2006; USEPA, 1996), during both approval and operational processes, they are likely to possess large sets of investigative and operational monitoring data on hydro-meteorological, chemical, and microbial parameters. In addition to data from water suppliers, other valuable datasets from local authorities may also be available for analysis. Such datasets may never have been

* Corresponding author. Tel.: +43 1 4277 53320.
E-mail addresses: andrea.bichler@univie.ac.at (A. Bichler), arnold.neumaier@univie.ac.at (A. Neumaier), thilo.hofmann@univie.ac.at (T. Hofmann).

http://dx.doi.org/10.1016/j.jhydrol.2014.08.013
0022-1694/

statistically analysed and may represent a hidden treasure for predicting drinking water quality.

Statistical analysis methods have been widely used to investigate large data sets: various studies have presented statistical techniques for relating microbial indicators to catchment processes and for identifying possible sources of pollution (Cinque and Jayasuriya, 2010; Cruz et al., 2012; Nnane et al., 2011). In these approaches data exploration techniques such as factor analysis, principal component analysis, or discriminant analysis, are fitted to the observations to reveal relationships and/or predict future scenarios. Although these methods can be powerful tools, they rely on assumptions about the data distribution, linearity, independence, etc., that are not always valid in environmental data. In contrast, algorithmic models use only the input variables to explore relationships without making any assumptions about the distribution of the data (Breiman, 2001). For water resource data, which are commonly characterised by skewness and outliers, non-parametric tests can be more powerful than parametric tests (Helsel and Hirsch, 2002). Tree-based models are a type of algorithmic model that is already widely used as a data-mining technique in medical, social, and economic sciences (Murthy, 1998). In the context of water quality management, however, very few studies have employed this robust and versatile data analysis method. Litaor et al. (2010) used a binary tree-model to classify spring water samples according to their hydrochemical constituents, in order to identify factors affecting water quality. Parkhurst et al. (2005) and Jones et al. (2013) explored readily available monitoring datasets using data-driven tree regression to predict the concentration of faecal indicator bacteria in bathing water. These applications show that non-parametric models have the power to (1) explore large datasets and reduce them to a smaller number of significant variables, (2) assign a probability to these explanatory variables, and (3) predict future scenarios. They can yield results that are similar to, or even better than, parametric models without having made any assumptions about the data distribution (Álvarez-Álvarez et al., 2011; Litaor et al., 2010).

In this study we explore the relationship between readily available monitoring data such as precipitation and river water level (continuous explanatory variables) from a large water supplier and the presence of total coliforms (categorical response variable) in a groundwater well used for drinking water production. The association of the explanatory variables with total coliforms (TC) was assessed by recursive partitioning using the Chi-Squared Automatic Interaction Detection (CHAID) algorithm (Kass, 1980).

The main aims of this study were to identify monitoring parameters associated with the occurrence of TC in a production well and a nearby observation well and to rank these variables according to their significance. The question of whether hydrometric data can be used to predict the presence of TC in the production well was also addressed.

2. Materials and methods

2.1. Site description

The area investigated lies within a mesoscale alpine headwater catchment. Land use within the catchment area is dominated by pastures with patches of forest and scattered urban areas, mainly comprising small villages and farms (EEA, 2006). The glaciofluvial aquifer consists predominantly of coarse carbonate gravel with a mean hydraulic permeability (k_f) of 1.5×10^{-2} ms^{-1} and average linear flow velocities (v_a) of 60–70 md^{-1}. Water abstraction has led to the removal of fine sediments in the immediate vicinity of the horizontal wells, resulting in increased hydraulic permeability (k_f up to 5.5×10^{-2} ms^{-1}) and flow velocities (v_a up to 90–100 md^{-1}). Four piezometers (A–D) are located on a profile along the main groundwater flow direction between the production well and the R1 river (Fig. 1).

The catchment is drained by two rivers and water from one of these (R1) infiltrates into the aquifer. Water from this river is also diverted for hydropower generation a few kilometres upstream of the investigation site. At the river stretch under consideration a minimum residual flow of 1.1 m^{3}s^{-1} remains in the river bed over approximately 75% of the year, with higher discharge rates occurring at times.

The groundwater level has fallen several metres since the extraction of groundwater commenced. The depth to the groundwater table (which extends beneath the river) is at present approximately 5–6 m in the wellhead area, leading to constant influent flow conditions: river water infiltration recharges the aquifer under all flow conditions and has created a zone within the aquifer that is constantly under the influence of bank filtrate.

Groundwater is abstracted with a horizontal drainage well of 600 m length (the production well, PW) and used directly for drinking water supply, generally without any further treatment or disinfection. The production well is situated in the central part of the aquifer, at a depth of 9 m; water is drained from the aquifer by gravitational forces only, without being pumped. Well discharge (Q_{PW}) is mainly driven by the response of the water works operations to the hydraulic flow conditions in the aquifer, but it can also be regulated by operating a valve and adjusted to match consumption. Well discharge ranges from 1.1–2.6 m^{3}s^{-1}, with a mean of 1.5 m^{3}s^{-1}. A second horizontal well that has now been abandoned is used for monitoring purposes only (the observation well, OW). It is located between the R1 river and the production well and is operated with a continuous discharge of only 0.5×10^{-3} m^{3}s^{-1}. During average flow conditions neither the production well nor the observation well receive any bank filtrate. The wellheads are sealed and only accessible through a well house to prevent any direct contamination. The area above the horizontal wells is covered with grassland on shallow soils of rendzina and brown earth. Agriculture is prohibited in the wellhead area but it is accessible to the public for recreational activities. The entire area is protected by levees against flooding from the adjacent rivers.

2.2. Dataset and pre-processing

A seven year record (from 01 January 2006 to 31 January 2013) of microbial, hydrologic, and hydraulic data provided by the local waterworks served as a data base for this research. The time series

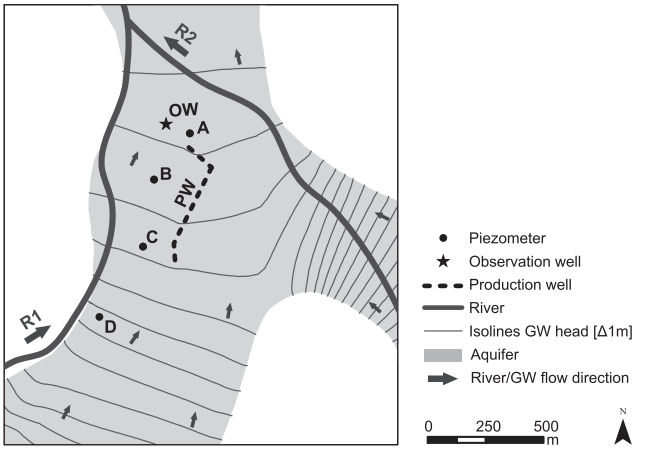

Fig. 1. Investigation site with production well (PW) and observation well (OW) during mean groundwater flow conditions.