Journal of Hydrology xxx (2014) xxx–xxx

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 xxxx
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 statistically

* 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/

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 to 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

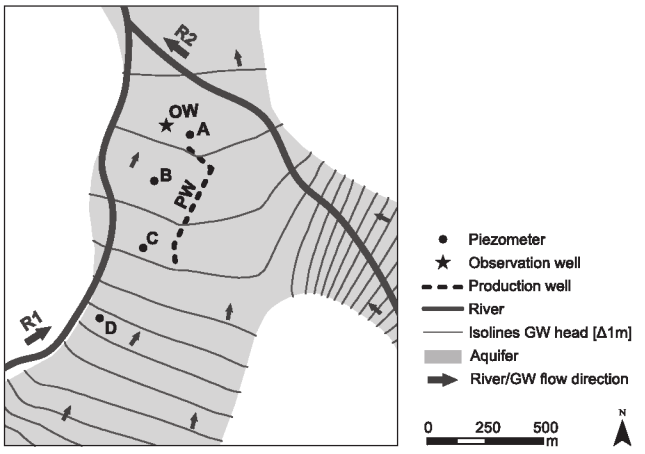

Fig. 1. Investigation site with production well (PW) and observation well (OW) during mean groundwater flow conditions.

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013

were checked for outliers and inconsistencies and corrected, where possible. Data points were only used if microbial data was also available, resulting in a total sample size covering 1746 days.

Total coliforms (TC) were sampled once a day (Monday–Friday) in the production well and the observation well, but not in any other monitoring wells. We therefore based our analysis on these two observation points. TC were monitored using the Colilert®-18/Quanti-Tray method with a detection limit of 1 cfu in 100 ml. For the statistical analysis TC were only reported in a binary code signifying either presence or absence. The categorical response variables were set to 1 for positive test results and to 0 for negative results.

The precipitation (P), the water level (stage) of the river (WL), the groundwater head (GWH), and the discharge from the production well (QW_{PW}) were all recorded at hourly intervals. The hourly values for P were converted to daily totals and the hourly values for WL, GWL, and QW_{PW} into daily means. In addition to these daily values, sums (s) and means (m) were calculated for periods of up to five days. In order to take into account the dynamic behaviour of the system, variations in the individual variables were integrated in the analysis: differences (d) between sequential observations for the river level, groundwater level, and well discharge over periods of between one and five days were calculated (e.g. the change in groundwater head from piezometer A over one day). In addition, time lags of up to five days were applied to all variables in order to analyse how earlier events might have influenced the occurrence of TC. Fifteen types of variables were thus derived (i.e. sums, means, and differences) with thirty variations for each type (see for example, Fig. 2 for variations to the discharge from the production well). The variables were labelled according to their type, time range, and time lag: e.g. GWH D m_3 t_{-5} refers to the mean groundwater head (GWH) from piezometer D over three days (m_3), five days prior (t_{-5}) to the occurrence of TC$_t$.

In total, 402 variables were generated for inclusion in the CHAID analysis (Table 1).

2.3. Classification tree – CHAID algorithm

2.3.1. Theoretical background

The relationships between hydro-meteorological and hydraulic variables (explanatory variables) and the occurrence of TC (response variable) were assessed using a tree-based model, i.e. the Chi-squared Automatic Interaction Detector (CHAID) developed by Kass (1980). Tree-based models operate in a sequential procedure where each explanatory variable is split into more homogenous subgroups with respect to the response variable. The most significant variable with the highest X^2 value is then chosen and each of its subgroups is re-analysed independently to create further subgroups. Instead of expressing relationships between variables by linear combinations (e.g. as in linear regression analysis) this type of data model proceeds stepwise, thus being more adept at handling interactions between variables (Clark and Pregibon, 1992). Depending on the measurement scale of the response variable, classification trees (categorical) can then be distinguished from regression trees (continuous) (Hill and Lewicki, 2007).

The technique proposed by Kass (1980) allows large data sets of categorical data to be described, with one variable being the response variable (e.g. TC) and the remainders being the explanatory variables (e.g. water level of river, groundwater heads, production well discharges). If the explanatory variables are of a continuous nature, categories with approximately equal numbers of observations need to be created. In the presented study the response variable is categorical (positive/negative TC test result) while the explanatory variables (e.g. precipitation) are continuous. In order to be able to include this type of data the explanatory variables are banded into discrete categories prior to analysis. In a stepwise procedure the CHAID algorithm identifies for each explanatory variable in turn the pair of categories that is least significantly different with respect to the response variable. If the significant difference is below a critical significance level (p-value), the two categories are considered to be homogenous and are merged into a single category. This step is repeated until the differences between the categories are significant with respect to the response variable, thus allowing the formation of multi-way splits. A split criterion can be defined in order to ensure that the best partition is found for each explanatory variable. The X^2 value for each explanatory variable is then calculated, the one with the highest score (X^2 value) is chosen and the analysis is run again on each subgroup (node). If the significance for splitting a variable it is below the predefined significance level (p-value) the algorithm stops, thus creating a 'terminal leaf' of the classification tree. A detailed description and evaluation of the computational procedure can be found in Kass (1980) and van Diepen and Franses (2006).

2.3.2. Application

All calculations were run using SPSS 19 software (IBM, 2010). The continuous explanatory variables were grouped into 10 categories with equal numbers of observations. The results were evaluated using cross-validation; the significance level for splitting nodes and merging categories was set to a p-value of 0.01. The CHAID algorithm was applied in three different ways: (1) All variables except for the response variable were used as explanatory variables. This was repeated twice: once for TC in the production well, and once for TC in the observation well ($TC_{PW/OW}$ t_{-0}), thus creating two classification trees. (2) The algorithm was run again but including only one variable type at a time (e.g. WLR d_{1-5} $t0$–5) and with the tree depth limited to one level. This step identified the most significant explanatory variable of each type. Variables that are only slightly less significant than the most significant variable may be revealed by this step. (3) To evaluate the significance of individual variables the algorithm was used for recursive partitioning only. In this case only one variable at a time was included in the analysis. The selected variable was split into subcategories and assigned an X^2 value. In addition to the classification trees produced, this allowed different variable types to be compared. Again, all analyses were performed with respect to TC in the production and observation wells.

2.3.3. Prediction

Classification trees can also be used to predict the response variable and allow weightings to be defined for prediction errors. These weightings can represent either the actual costs of prediction errors, or the level of risk of incorrect prediction that is considered acceptable. The acceptable risk can be expressed in monetary terms but qualitative values can also be used. If weightings were set to equal values (ratio 1:1), a false negative error and a false positive error would be expected to have similar consequences. Using this ratio the overall predictive performance is at its

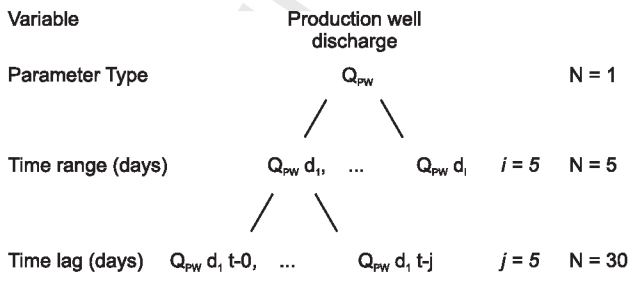

Fig. 2. Schematic view of variable definition, taking into account time range (i) and time lag (j) in an example of changes in discharge from the production well (Q_{PW} d).

Table 1
Data set, variations in monitoring variables.

Variable	Total coliforms PW/OW	Precipitation	Water level river R1	Groundwater head piezometers A–D	Discharge PW	*N* variables
ID	$TC_{PW/OW}$	P	WLR	GWH_{A-D}	Q_{PW}	9
Type	Presence	Sum (s)	Mean (m) difference (d)	Mean (m) difference (d)	Mean (m) difference (d)	15
Unit	(yes/no)	(mm)	(m)	(m a.s.l.)	(m^3/s)	
Time range *i* (days)	–	1–5	1–5	1–5	1–5	67
Time lag *j* (days)	0–5	0–5	0–5	0–5	0–5	
N variables	12	30	60	240	60	402

maximum. Depending on the objective of a study, it might be desirable to improve the correct prediction rate for one particular outcome (e.g. correct prediction of TC positives). Changing the initial weightings (1:1 ratio) can result in a better prediction of the chosen category (TC positives) but will also lead to a worse prediction of the other category (TC negatives), thus reducing the overall predictive performance (predicting both positives and negatives correctly). The outcomes predicted will therefore be influenced by the analyst's preferences, the problem definition, and the structure of the tree-model itself, i.e. the homogeneity of the subgroups. This study has not taken into account monetary costs. We considered a false negative error (no TC predicted even though test result is positive) to be associated with a potential health risk and therefore to have a greater impact than a false positive error. Assuming that a possible threat to human health is of critical importance we adjusted the weights stepwise until all TC positives (>99%) could be predicted correctly, and also evaluated the impact that changing the weightings had on the predictive outcome.

3. Results

3.1. Temporal variability of monitoring variables

Flow conditions are subject to seasonal variations but can also vary from year to year (Fig. 3). A seasonal pattern can be observed for 2010, between a dry winter period from mid-November to mid-May and a wetter summer period (Fig. 3a). Large rainfall events during the summer months resulted in higher water levels in the river and groundwater levels also responded to these high flows, exhibiting a damped version of the river hydrograph signal, indicating connectivity between the river and the aquifer system. The well discharge is closely related to the groundwater level and consequently yielded higher discharge rates during the summer season. Positive TC results were mainly clustered around precipitation and high flow events but also occurred sporadically in between, and after, such events (upper row of Fig. 3a).

In contrast, the year 2012 was characterised by average to low flow conditions; rainfall events were of lower intensity and river water levels lower throughout the year than in 2010. Groundwater levels during 2012 remained nearly constant and production well discharge was below average. TC occurred less frequently than in 2010 but with individual incidents recorded throughout the year (Fig. 3b).

3.2. Correlation of TC between production well and observation well

The production well was affected by TC on 52 of the 1751 sampling days (3.0%) and the observation well on 142 days (8.1%). The contamination was generally only sporadic. The production well was mostly affected by TC on solitary days only, while the observation well was often affected on consecutive days (Fig. 3).

The algorithm was first used to examine the correlation between the microbial parameters for both wells. TC_{PW} and TC_{OW} with increasing time lags were used as explanatory variables and

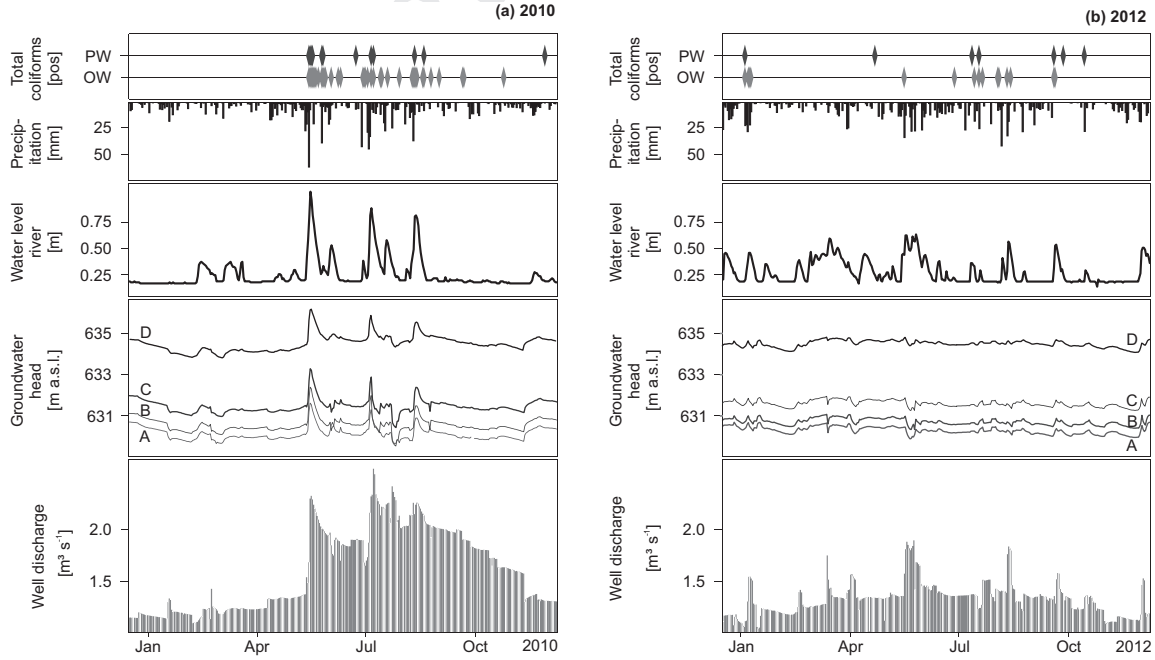

Fig. 3. Overview of monitoring variables for selected years: 2010 (high flow conditions) and 2012 (average/low flow conditions).

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013

response variables. Both wells were affected simultaneously by TC on 25 days, representing 48.1% of all TC positives in the production well and 17.6% in the observation well. No further correlation was found when using TC_{PW} as an explanatory variable, while using TC_{OW} as an explanatory variable was found to be also significant with longer time lags. If TC were detected in the observation well, there was an 8.3% probability of the production well being affected on the following day (solid line, Fig. 4).

TC records from the observation well showed a strong auto-correlation, with a positive result on one day increasing the probability of a positive result on the next day by 38%. This relationship became weaker as the time lag increased, but even after five days the probability of TC occurrence remained at 29% (dashed line, Fig. 4).

3.3. Classification tree for hydrometric variables

In order to explore the relationships between hydrometric variables and the occurrence of microbial indicators, classification trees were grown using all generic explanatory variables with respect to TC in the production and observation wells.

Cumulative precipitation over two days (P s_2 t_- 0) was found to be the most significant variable affecting the production well. At the first tree level this variable was split into three subgroups (Fig. 5), with the highest probability (13.7%) of TC being found in the subset when rainfall exceeded 22 mm in 48 h. This node represented a leaf (terminal node), while the remaining two subsets of this variable were divided further. The splits were realised by variables derived from groundwater heads, based either on groundwater levels from piezometer D (left branch), or fluctuations in the near-by piezometers A and B (centre branch). Elevated TC probabilities were found for groundwater heads from piezometer D in excess of 634.45 m a.s.l. (GWH D m_1 t_- 0), and also for groundwater levels from piezometer B that rose more than 0.16 m over three days (GWH B d_3 t_- 2). The subsets of the third level (GWH A d_2 t_- 5) did not show any clear pattern.

In the observation well precipitation was again found to be the most significant variable with respect to TC (as was the case for the production well, described above): the highest probability (32.6%) of TC occurrence in the observation well was found to be when cumulative precipitation exceeded 46 mm over 5 days (P s_5 t_- 0). This subset was split further by groundwater heads from piezometer C, where 44.6% of the cases with values above 631.42 m a.s.l. were associated with TC in the observation well (Fig. 6). Variables generated from river water levels were identified as significant explanatory variables (WL m_1 t_- 0, WL m_1 t_- 4) if cumulative precipitation over 5 days was below 46 mm. Different levels of precipitation resulted in further splits in the centre branch. The combination of high river water levels and cumulative rainfall exceeding 45 mm over five days, together with a time lag of five days (P s_5 t_- 5) accounted for 16 microbial results, corresponding to a 30.8% probability of TC occurring within this subset. In the left branch an elevated probability of TC can be related to high river water levels prior to a microbial incident. At water levels ranging between 0.18 and 0.2 m the probability of TC increased to 14.7% if water levels had already exceeded 0.17 m four days earlier.

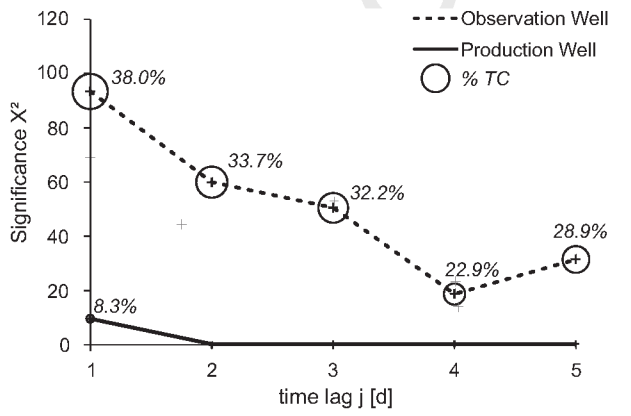

Fig. 4. Correlation and auto-correlation of microbial variables: TC_{OW} with different time lags (j) used as explanatory variable for TC in the production well (solid line) and observation well (dashed line).

3.4. Predictors of microbial water quality – individual analysis of hydrometric variables

A possible shortcoming of the algorithm is that only the most significant variable is entered in the first level of the classification tree and variables that are only slightly less important may be ignored. Actual dominant variables that are overshadowed by slightly more dominant secondary variables may also be overlooked completely. In addition to the tree-model therefore, each variable type was also analysed individually and ranked according to its X^2 value.

Fig. 7 shows the X^2 values of all variables; the three most significant variables are labelled. All eleven variables were statistically significant with respect to the observation well, while for the production well only ten out of thirteen variables were eligible for the analysis. As already demonstrated in Section 3.3, precipitation showed the strongest relationship to the response variable in both wells. Mean values and differences in river water levels were significant for the observation well, while for the production well groundwater heads from piezometer D and fluctuations in groundwater levels from piezometer A had the strongest relationship to TC. Table 2 lists the subcategories of these explanatory variables and includes information on absolute and relative TC and median values for each of the variables.

Groundwater levels from piezometer D were significantly related to TC in the production well (GWH D m_3 t_- 5, X^2: 26.9): TC were most likely to occur when groundwater levels were between 634.37 and 634.45 m a.s.l. These levels represent moderate to low flow conditions and are below the median value of 634.54 m a.s.l. Short term changes in groundwater head recorded by piezometer A (GWH A d_1 t_- 0, X^2: 22.1) also showed a significant relationship to the presence of microbial indicators in the production well. In the subset with rising groundwater levels, TC were detected in 8% of the cases where levels were rising at more than 0.01 m per day.

The presence of TC in the observation well showed a strong correlation with flow conditions in the river. The highest frequency of TC in the observation well was observed when daily mean water levels in the river exceeded 0.43 m (WL m_1 t_- 0, X^2: 96.0). River water levels higher than the median of 0.18 m still resulted in an increased probability of TC (12.2%). Apart from the means of river water levels, fluctuations in these water levels over two day periods (WL d_2 t_- 2, X^2: 76.6) also showed a significant relationship to TC. The dominant flow condition can be considered to be stable over the two day period with only minor variations, ranging between −0.08 and +0.07 m and accounting for only 5% of the TC counts. Fluctuations in the river water level increased this frequency to 15.9% for falling water levels and to 24.4% for rising water levels.

3.5. Effect of generic variables on significance

All monitoring parameters were used to create an extended data set of generic variables by varying the time period for accumulation, averaging, and calculating differences (time range j)

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013

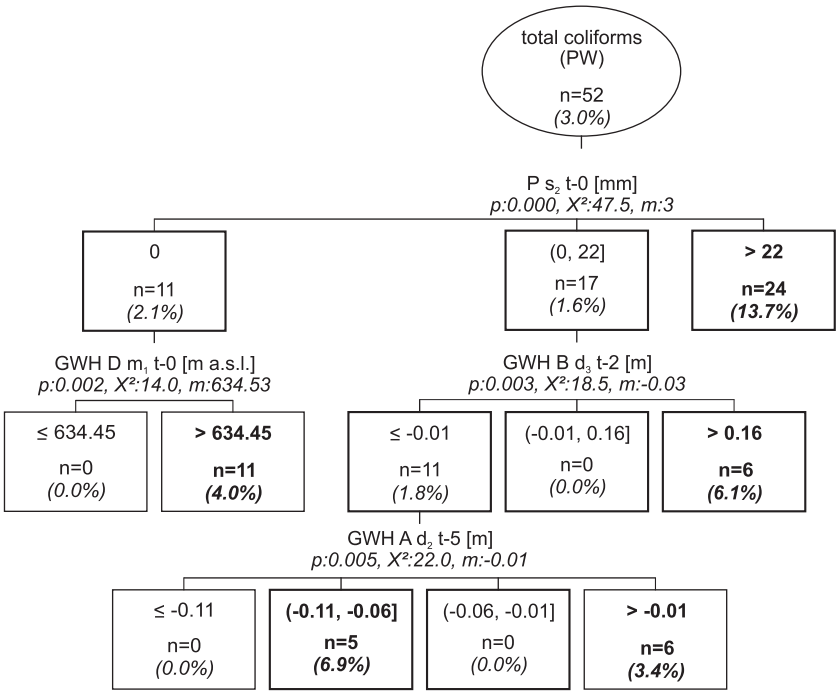

Fig. 5. Classification tree of hydrometric variables for the production well, with *p*-value, significance (X^2), and median (*m*). Each box (node) represents a subclass of an explanatory variable, with the variable value on top and the number of TC positives (*n*) in the centre; values in brackets denote percentages within the subset. Nodes are marked in bold font where the initial probability exceeds 3.0%.

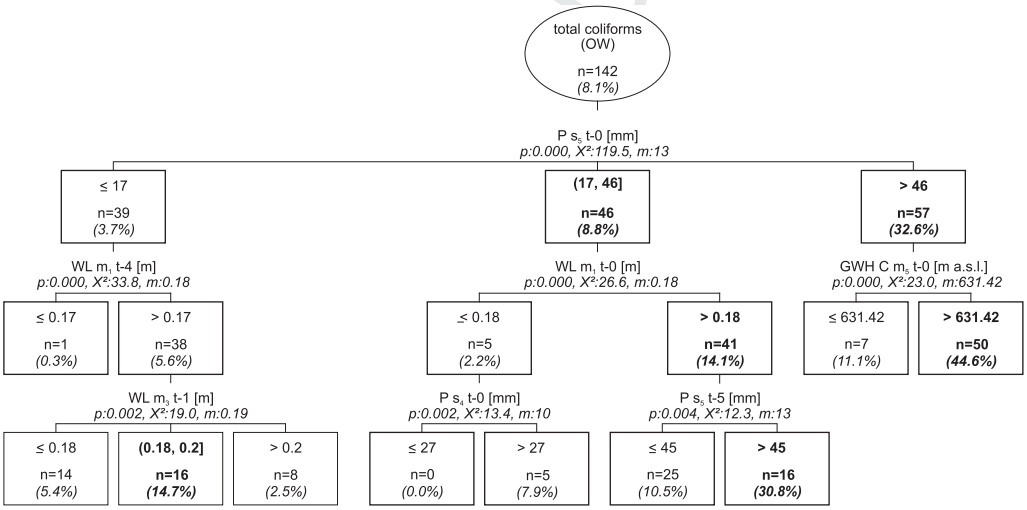

Fig. 6. Classification tree of hydrometric variables for the observation well, with *p*-value, significance (X^2), and median (*m*). Each box (node) represents a subclass of an explanatory variable, with the variable value on top and the number of TC positives (*n*) in the centre; values in brackets denote percentages within the subset. Nodes are marked in bold font where the initial probability exceeds 8.1%.

and imposing a time lag (*i*). Fig. 8 illustrates how these two time variables affect the significance of two selected monitoring variables. Changing the time range for the accumulation of daily precipitation values had only a minor influence on significance in the production well (Fig. 8a). In contrast, the significance increased substantially in the observation well when the time range was extended (Fig. 8b). Introducing a time lag (*j*) made no improvement to the significance of this variable in either well.

Groundwater heads from piezometer D were rather invariant when considering variations in *i* and *j* for the production well (Fig. 8c), with only a slight increase in significance for a time lag of more than three days. For the observation well (Fig. 8d), variations in the time range (*i*) had a limited effect on the significance, while a time lag (*j*) had a negative effect.

3.6. Prediction of TC using the CHAID classification tree

Classification trees can also be used for prediction of a response variable. The frequency of positive TC was below 50% in all subcategories of the predictors ('leaves' in Figs. 5 and 6). At the initial 1:1 ratio of false negative to false positive errors all leaves predicted the absence of TC; the overall correct prediction rate (positive and negative TC results) was above 90% for both the production and the observation well. Gradually increasing the weighting for false negatives resulted in the number of leaves predicting TC (and consequently the correct prediction rate for TC positives) also increasing gradually, while the overall prediction rate decreased (Fig. 9). The correct prediction rate for TC positives in the PW increased rather uniformly as the weightings were increased until 100%

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013

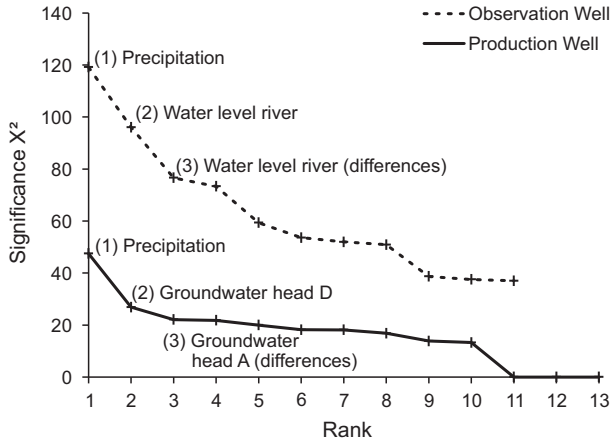

Fig. 7. Significance of explanatory variables.

correct prediction was achieved at a weighting of 30. To obtain this result the model predicted TC occurrence on 798 (46%) of the total number of days, of which 52 were positives and 746 were false positives. The prediction of TC positives in the observation well achieved an 80% success rate a weighting of 10; a weighting of 40 was required to achieve a success rate greater than 90%. In the latter case, TC occurrence was predicted at 1213 (69%) of the total number of days, of which 141 were positives and 1072 false positives. At a correct prediction rate of greater than 99% for TC positives the overall correct prediction rate dropped to 57% for the production well and 39% for the observation well (Table 3).

4. Discussion

4.1. Microbial water quality and correlation of indicator parameters

The water quality differed between the observation well and the production well, with TC being detected less frequently in the latter. Almost 50% of all TC positives in the production well coincided with TC positives in the observation well, suggesting that a common source of contamination is likely since they are located in close proximity to each other. However, the observation well suffered more often from microbial water deterioration, probably as a result of being located closer to the river and the aquifer zone being under the constant influence of bank filtrate, which is thought to be a possible source of contamination. Another reason for the increased contamination may be the operational mode of the observation

Table 2
Classification of the three explanatory variables with respect to TC in the production and observation wells, together with their median values (m). Numbers shown in bold font are for subsets in which the initial frequency of TC (PW: 3.0%, OW: 8.1%) is exceeded.

	Production well			Observation well		
	(1) Precipitation $P\ s_2\ t-0$, m: 3 mm			(1) Precipitation $P\ s_5\ t-0$, m: 13 mm		
Class	0	(0–22]	**>22**	⩽17	**(17–46]**	**>46**
TC	11 *(2.1%)*	17 *(1.6%)*	**24 *(13.7%)***	39 *(3.7%)*	**46 *(8.8%)***	**57 *(32.6%)***
	(2) Groundwater head D *GWH D* $m_3\ t-5$, m: 634.54 m a.s.l.			(2) Water level river *WL* $m_1\ t-0$, m: 0.18 m		
Class	⩽634.37	**(634.37–.45]**	>634.45	⩽0.18	**(0.18–0.43]**	**>0.43**
TC	3 *(0.6%)*	**14 *(8.0%)***	34 *(3.3%)*	25 *(2.7%)*	**78 *(12.2%)***	**39 *(22.7%)***
	(3) Groundwater head A (changes) *GWH A* $d_1\ t-0$, m: −0.01 m			(3) Water level river (changes) *WL* $d_2\ t-2$, m: 0.00 m		
Class	⩽−0.06	(−0.06 to 0.01]	**>0.01**	⩽**−0.08**	(−0.08 to 0.07]	**>0.07**
TC	1 *(0.3%)*	36 *(2.9%)*	**14 *(8.0%)***	**29 *(15.9%)***	69 *(5.0%)*	**44 *(24.4%)***

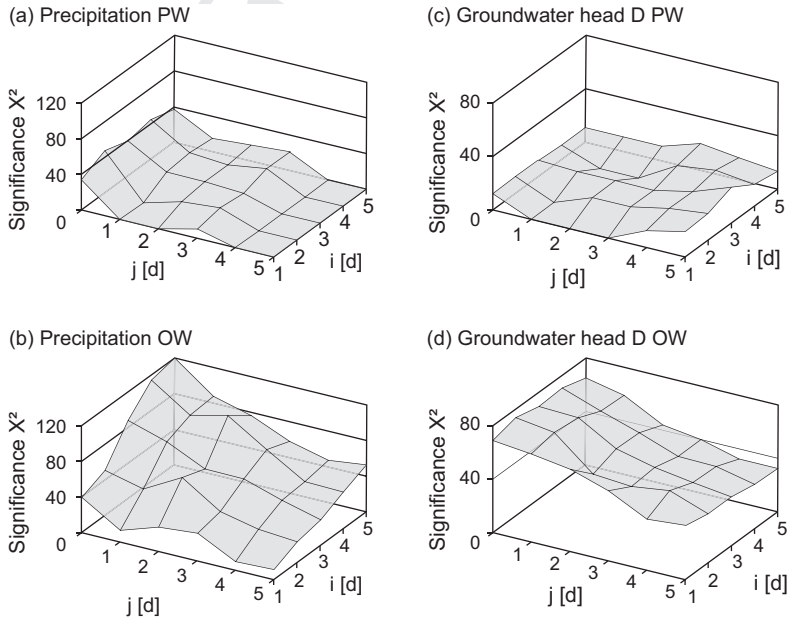

Fig. 8. Effect of time lag (j) and time range (i) on significance (X^2). Changes in significance with variations in precipitation, with respect to TC in the production well (a) and observation well (b), and with variations in groundwater head (from piezometer D), with respect to TC in the PW (c) and OW (d).

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013

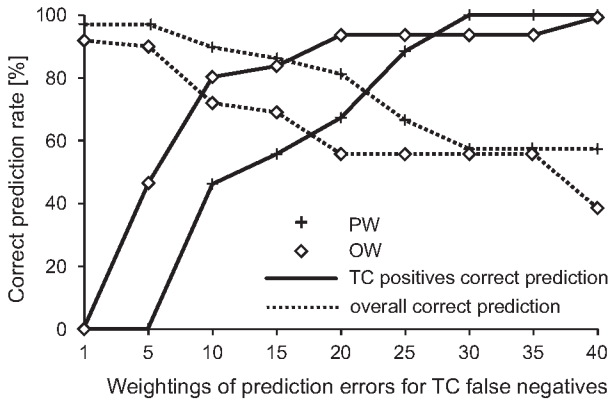

Fig. 9. Effect of weights for prediction errors on the correct prediction rate for TC positives (solid line) and the overall prediction rate (dashed line) in the production well (crosses) and observation well (squares).

well, which is run continuously with a very low rate of flow (Q_{OW} 0.5×10^{-3} m^3 s^{-1}); the higher flow rate in the production well (Q_{PW} mean: 1.5 m^3 s^{-1}) may lead to dilution below the coliform detection limit. Since TC frequently occurred in the observation well on the same day as in the production well, the use of this well as a 'sentinel' site is of only limited value. The high auto-correlation of TC_{OW} reflected the fact that this site was frequently affected on consecutive days; this could again be related to the different operational modes of the two wells as microbial contamination may be more rapidly removed from the production well.

4.2. Significant hydrometric variables

Relationships between explanatory variables and TC were generally of greater significance in the observation well than in the production well. This may be due to TC often occurring on consecutive days in the observation well, making it easier to identify homogenous subgroups. Different explanatory variables were selected for each well, even though both wells are located in close proximity to each other and in similar hydrogeological settings. The most significant variable for both wells was the precipitation. Variables related to groundwater flux were only significant with respect to the production well, while variables describing river water dynamics were found to be only significant in the observation well. For both, the production and observation wells there was a marked difference in significance between the most significant variable (precipitation) and the second and third most important variable (groundwater heads, water level river. This suggests that, these variables may represent different catchment processes, although they are still likely to be interrelated to some extent.

The presence of TC showed a strong correlation with heavy rainfall in both wells, with cumulative rainfall over two days being highly significant for the production well, and over five days for the observation well.

The outbreak of waterborne diseases is known to often be triggered by extreme weather events involving heavy rainfall and flooding (Cann et al., 2013). During extreme rainfall events microorganisms can be mobilized from non-point sources such as agricultural surfaces (Collins et al., 2005), consequently increasing the availability of bacterial loads in a catchment by some orders of magnitude compared to dry conditions (Dechesne and Soyeux, 2007; Kistemann et al., 2002). Our analysis of TC in the groundwater wells has revealed a very rapid response to precipitation. This suggests high transport velocities for the bacterial indicators through both the vadose zone and the aquifer. Bacteria may reach the aquifer much more rapidly than suggested by the average pore water flow velocity due to preferential flow paths (Beven and Germann, 2013; Taylor et al., 2004). Moreover, size exclusion effects on microorganisms, in combination with macropore flow, impose a considerable risk on shallow aquifers with shallow soils (Unc and Goss, 2003).

River water levels were only significant with respect to TC in the observation well, where the occurrence of TC was always more likely during high flow conditions. These results are in agreement with the findings of Dechesne and Soyeux (2007), who associated microbial risk with rapidly rising water levels in a river. Infiltration of river water into the aquifer is a highly dynamic process and influenced by the discharge rates, with enhanced infiltration occurring during periods of high flow (Schubert, 2002). Flooding may remove the clogging layer of the riverbed, thus increasing the hydraulic conductivity and water fluxes into the aquifer (Hiscock and Grischek, 2002; Mutiti and Levy, 2010). The statistical significance of rising water levels indicates enhanced infiltration during short-term fluctuations, as shown by Derx et al. (2010). Within the area investigated the travel times between the river and the observation well ranged between five and seven days. In view of the detection of TC very soon after storm runoff (within <2 days), direct contamination by freshly infiltrated river water is considerd to be unlikely. The relationship between TC and elevated river water levels may instead be explained by a different process: increased infiltration can result in an expansion of the aquifer zone that is under the constant influence of bank filtrate towards the central part of the aquifer and into the catchment area of the production well. This change in the groundwater flow field, as well as the subsequent return to normal flow conditions, can occur within a few hours of the runoff and infiltration peak and could explain the occurrence of TC in the observation well.

Parameters related to groundwater flux were found to be of major importance for the production well. During moderate rainfall fluctuating groundwater levels in piezometers A and B, which were located in the immediate vicinity of the production well, became important explanatory variables. Both variables were most significant with time lags of two days and five days; changes may have been caused by infiltration of surface water from the river, or by rapid groundwater recharge from the unsaturated zone. Rising groundwater heads may also mobilize mircoorganisms from the interphase between the phreatic and the vadose zones (Pang, 2009; Unc and Goss, 2003).

Not all TC occurrences could be explained solely on the basis of hydrometric variables and it is likely that other contamination sources remained unidentified. However, our analysis did reveal certain situations under which TC occurrence is more likely and also allowed the identification of processes that could possibly lead to microbial contamination: in general TC occurred more frequently during wet conditions (heavy precipitation, storm runoff) and during variations in the hydraulic system (fluctuating water levels in the river and in the groundwater).

Table 3
Predictive performance of the classification tree at a correct prediction rate for TC positives of >99%. Comparison of predicted (*P*) and observed (*O*) TC for the production and observation well.

	Production well			Observation well		
	P neg	*P* pos	%corr	*P* neg	*P* pos	%corr
O neg	948	746	56	532	1072	33
O pos	0	52	100	1	141	99
% overall	*54*	*46*	*57*	*31*	*69*	*39*

4.3. Generic variables

Generic variables (time range and lag time) had different impacts on significance. Varying the time range had a marked effect on precipitation, but much less effect on groundwater heads.

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013

Longer periods for accumulation or averaging had a greater impact on variables that were highly variable and had a low level of autocorrelation. A longer accumulation period for the rainfall variable yielded an improved reflection of pathogen occurrence, as had also been found by Wilkes et al. (2009). This was of particular importance when considering the observation well, which was often affected for a number of consecutive days. Increasing the temporal offset generally resulted in a decrease in significance, implying that the most recent information is the most important. However, the influence of time range and time lag was shown to depend on both the original variable and the response variable, and is difficult to estimate *à priori*.

4.4. Prediction

Our objective was the correct prediction of all TC occurrences (100%) in the production well. The model predicted TC positives on nearly half of all monitoring days (46%) and although this led to many false negative errors and a reduced overall correct prediction rate, it also indicated that protection measures such as water disinfection on approximately half of all operating days is likely to be sufficient to prevent any risks from microbial contamination. In order to achieve this result the weightings for a false negative prediction of TC had to be increased to 30. With regard to monetary costs, this appears to suggest that the poor overall correct prediction rate is acceptable provided that the cost of a false negative error (e.g. medical treatment following an outbreak of water-borne disease) is more than 30 times greater than the cost of a false positive error (e.g. water disinfection when not actually necessary).

5. Conclusions

The CHAID algorithm was able to identify hydrometric parameters that were significantly related to the occurrence of TC in a drinking water production facility. Exploring statistical relationships between hydrometric variables and microbial indicators provided valuable indications of the likely sources of the pathogens. The significant variables could be used as proxy indicators for critical conditions, which would be of particular interest as they can be predicted using independent models (from, for example, precipitation, or a river's stage).

Moreover, the proposed algorithm proved to be a useful tool with which to reduce a large data set to a much smaller number of significant variables. This is particularly valuable as enlarging the readily available monitoring dataset by creating generic variables was clearly shown to improve significance levels.

Additional insights were also gained by separately analysing the individual variable types. It was possible to identify explanatory variables other than the common explanatory variables dominating both systems, in which the X^2 value provided valuable information concerning the strength and importance of a relationship.

A major advantage of the tree-like structure is the intuitive interpretation of the results. Complex relationships can be displayed in a clear and comprehensible way, providing easily understandable information to researchers and water managers. Moreover, tree-based models can provide valuable support to decision makers in evaluating the consequences of false negative errors (e.g. the cost of medical treatment following an outbreak of water-borne disease) and false positive errors (e.g. the cost of water disinfection when not actually necessary).

References

Álvarez-Álvarez, P., Khouri, E.A., Cámara-Obregón, A., Castedo-Dorado, F., Barrio-Anta, M., 2011. Effects of foliar nutrients and environmental factors on site productivity in Pinus pinaster Ait. stands in Asturias (NW Spain). Ann. For. Sci. 68 (3), 497–509.
Beven, K., Germann, P., 2013. Macropores and water flow in soils revisited. Water Resour. Res. 49 (6), 3071–3092.
Breiman, L., 2001. Statistical modeling: the two cultures. Stat. Sci. 16 (3), 199–231.
Cann, K.F., Thomas, D.R., Salmon, R.L., Wyn-Jones, A.P., Kay, D., 2013. Extreme water-related weather events and waterborne disease. Epidemiol. Infect. 141 (4), 671–686.
Cinque, K., Jayasuriya, N., 2010. Catchment process affecting drinking water quality, including the significance of rainfall events, using factor analysis and event mean concentrations. J. Water Health 8 (4), 751–763.
Clark, L.A., Pregibon, D., 1992. Tree based models. In: Hastie, T.J. (Ed.), Statistical Models. S. Chapman & Hall, London, pp. 337–420.
Collins, R., Elliott, S., Adams, R., 2005. Overland flow delivery of faecal bacteria to a headwater pastoral stream. J. Appl. Microbiol. 99 (1), 126–132.
Cruz, M.C. et al., 2012. The impact of point source pollution on shallow groundwater used for human consumption in a threshold country. J. Environ. Monit. 14 (9), 2338–2349.
Dechesne, M., Soyeux, E., 2007. Assessment of source water pathogen contamination. J. Water Health 5 (SUPPL. 1), 39–50.
Derx, J., Blaschke, A.P., Blöschl, G., 2010. Three-dimensional flow patterns at the river–aquifer interface – a case study at the Danube. Adv. Water Resour. 33 (11), 1375–1387.
EC, 1998. Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption. Official Journal of the European Communities L333, pp. 32–54.
EC, 2000. Directive 2000/60/EC of the European Parliament and of the Councils of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327, pp. 1–73.
EC, 2006. Directive 2006/118/EC of the European Parliament and of the Council of 12 December 2006 on the protection of groundwater against pollution and deterioration. Official Journal of the European Communities L 372, pp. 19–31.
EEA, 2006. Corine Land Cover 2006 raster data. European Environmental Agency, Copenhagen.
Helsel, D.R., Hirsch, R.M., 2002. Statistical methods in water resources. In: U.S.G.S. (Ed.), Techniques of Water-Resources Investigations, p. 522.
Hill, T., Lewicki, P., 2007. Statistics: Methods and Applications. StatSoft, Tulsa, OK.
Hiscock, K.M., Grischek, T., 2002. Attenuation of groundwater pollution by bank filtration. J. Hydrol. 266 (3–4), 139–144.
IBM, C., 2010. IBM SPSS Statistics for Windows, Version 19.0.0.1. IBM Corp., Armonk, NY.
Jones, R.M., Liu, L., Dorevitch, S., 2013. Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. Environ. Monit. Assess. 185 (3), 2355–2366.
Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. Appl. Stat. 29 (2), 119–127.
Kistemann, T. et al., 2002. Microbial load of drinking water reservoir tributaries during extreme rainfall and runoff. Appl. Environ. Microbiol. 68 (5), 2188–2197.
Litaor, M.I., Brielmann, H., Reichmann, O., Shenker, M., 2010. Hydrochemical analysis of groundwater using a tree-based model. J. Hydrol. 387 (3–4), 273–282.
Macler, B.A., Merkle, J.C., 2000. Current knowledge on groundwater microbial pathogens and their control. Hydrogeol. J. 8 (1), 29–40.
Murthy, S.K., 1998. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min. Knowl. Discov. 2 (4), 345–389.
Mutiti, S., Levy, J., 2010. Using temperature modeling to investigate the temporal variability of riverbed hydraulic conductivity during storm events. J. Hydrol. 388 (3–4), 321–334.
Nnane, D.E., Ebdon, J.E., Taylor, H.D., 2011. Integrated analysis of water quality parameters for cost-effective faecal pollution management in river catchments. Water Res. 45 (6), 2235–2246.
Pang, L., 2009. Microbial removal rates in subsurface media estimated from published studies of field experiments and large intact soil cores. J. Environ. Qual. 38 (4), 1531–1559.
Parkhurst, D.F., Brenner, K.P., Dufour, A.P., Wymer, L.J., 2005. Indicator bacteria at five swimming beaches—analysis using random forests. Water Res. 39 (7), 1354–1360.
Schubert, J., 2002. Hydraulic aspects of riverbank filtration – field studies. J. Hydrol. 266 (3–4), 145–161.
Taylor, R., Cronin, A., Pedley, S., Barker, J., Atkinson, T., 2004. The implications of groundwater velocity variations on microbial transport and wellhead protection – review of field evidence. FEMS Microbiol. Ecol. 49 (1), 17–26.
Unc, A., Goss, M.J., 2003. Movement of faecal bacteria through the vadose zone. Water, Air, Soil Pollut. 149 (1), 327–337.
USEPA, 1996. Safe Drinking Water Act US Environmental Protection Agency.
van Diepen, M., Franses, P.H., 2006. Evaluating chi-squared automatic interaction detection. Inf. Syst. 31 (8), 814–831.
WHO, 2011. Guidelines for Drinking-Water Quality, fourth ed. World Health Organisation, Geneva.
Wilkes, G. et al., 2009. Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape. Water Res. 43 (8), 2209–2223.

Please cite this article in press as: Bichler, A., et al. A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. J. Hydrol. (2014), http://dx.doi.org/10.1016/j.jhydrol.2014.08.013