Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems

S. R. Mounce, E. J. M. Blokker, S. P. Husband, W. R. Furnass, P. G. Schaap and J. B. Boxall

ABSTRACT

Particulate material accumulates over time as cohesive layers on internal pipeline surfaces in water distribution systems (WDS). When mobilised, this material can cause discolouration. This paper explores factors expected to be involved in this accumulation process. Two complementary machine learning methodologies are applied to significant amounts of real world field data from both a qualitative and a quantitative perspective. First, Kohonen self-organising maps were used for integrative and interpretative multivariate data mining of potential factors affecting accumulation. Second, evolutionary polynomial regression (EPR), a hybrid data-driven technique, was applied that combines genetic algorithms with numerical regression for developing easily interpretable mathematical model expressions. EPR was used to explore producing novel simple expressions to highlight important accumulation factors. Three case studies are presented: UK national and two Dutch local studies. The results highlight bulk water iron concentration, pipe material and looped network areas as key descriptive parameters for the UK study. At the local level, a significantly increased third data set allowed K-fold cross validation. The mean cross validation coefficient of determination was 0.945 for training data and 0.930 for testing data for an equation utilising amount of material mobilised and soil temperature for estimating daily regeneration rate. The approach shows promise for developing transferable expressions usable for pro-active WDS management.

Key words | discolouration, evolutionary polynomial regression, material accumulation, operation and maintenance strategies, self-organising maps, turbidity

S. R. Mounce (corresponding author)
S. P. Husband
W. R. Furnass
J. B. Boxall
Pennine Water Group,
Department of Civil and Structural Engineering,
University of Sheffield,
Sheffield S1 3JD,
UK
E-mail: *s.r.mounce@sheffield.ac.uk*

E. J. M. Blokker
KWR Watercycle Research Institute,
Groningenhaven 7, Postbus 1072,
3430 BB Nieuwegein,
The Netherlands

P. G. Schaap
PWN Water Supply Company North-Holland,
PO Box 2113,
1990 AC Velserbroek,
The Netherlands

INTRODUCTION

Discoloured water is generally viewed as an aesthetic problem, however the possibility of a high content of metals, organic/inorganic compounds and micro-organisms could potentially pose a health risk. It has been shown that particulate material accumulates over time as cohesive layers on pipeline surfaces in drinking water distribution systems (WDS) as a ubiquitous process. When subsequently mobilised, this material can be responsible for causing discolouration and other water quality issues, such as exceeding iron and manganese prescribed concentration values. Previous work (Husband & Boxall 2010, 2011) has demonstrated the cohesive nature and variable shear strength properties of these material layers, and how the layers are conditioned by the daily hydraulic regime and their causal relationship to discolouration.

The factors influencing this accumulation rate (also referred to here as regeneration) might include localised asset properties such as pipe age, material or diameter, while hydraulic conditions and bulk water quality (particularly iron concentration and water treatment type) are

doi: 10.2166/hydro.2015.140

likely to be important. Several studies have explored how temperature influences discolouration material accumulation rates. Sharpe (2013) studied the impact of temperature (comparing 8 and 16 °C) and prevailing shear stress on accumulation rates in a realistic-scale HDPE pipe rig over 28 days and found that accumulation was most greatly influenced by temperature. Schaap & Blokker (2013) found a strong correlation between temperature and accumulation rates in district metered areas (DMAs). Figure 1 captures some of the possible factors and potential interrelated complexities the literature suggests are important. The thick lines represent factors directly effecting accumulation rate, all others having secondary or more complex associations.

Fieldwork results (Cook & Boxall 2011; Blokker *et al.* 2011) suggest that accumulation rates are a linear function of time, with the magnitude dominated by the supplied water quality (with pipe material also being an important factor). There is limited work in the literature on predicting discolouration material accumulation and identifying the most important factors for estimating this rate. Models that can provide site-specific predictions of regeneration rates do not yet exist but a basic bi-variate categorical breakdown of discolouration rates are presented by Husband & Boxall (2011). They showed that the development of material layers is a reproducible and repetitive process. Given the complex, interrelated nature of the physical, chemical and biological reactions that are considered to contribute to discolouration material regeneration it could be expected that predictive models of regeneration rates will only be engendered through the application of multivariate, regressive

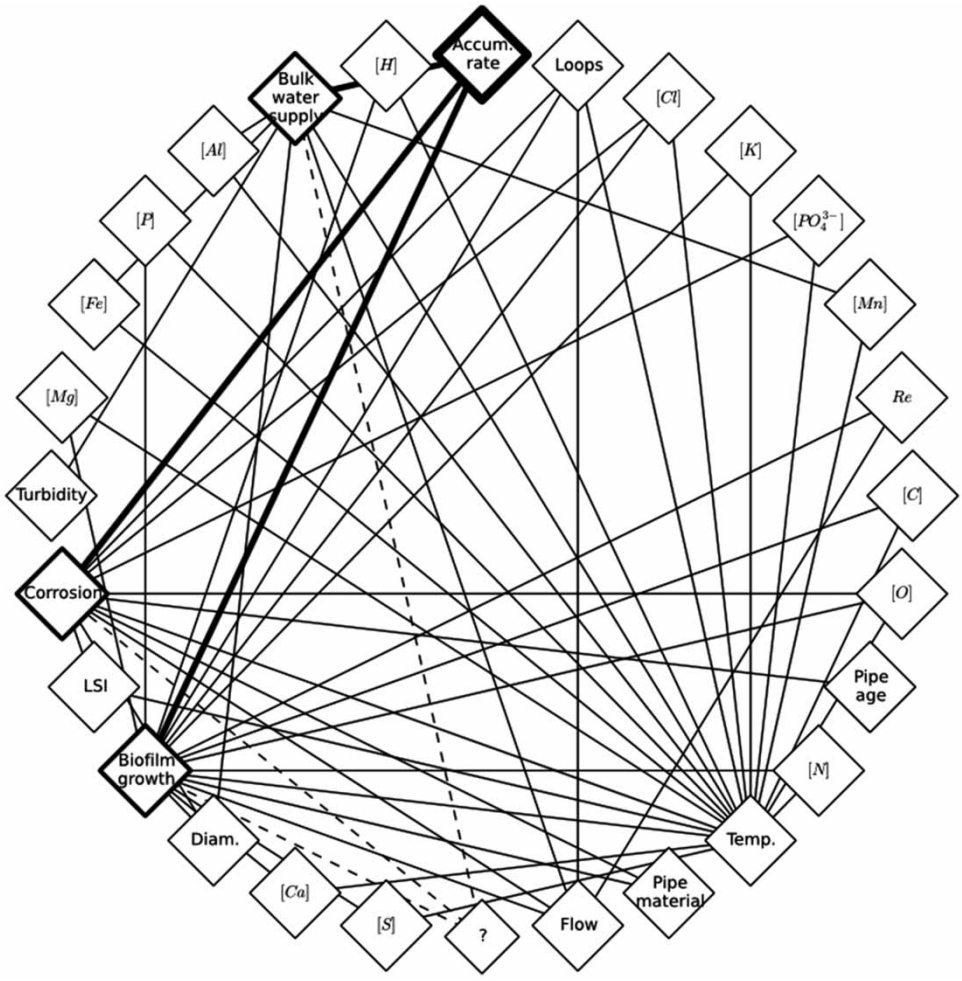

Figure 1 | Potential factors determining material accumulation rate (Temp. is an abbreviation for temperature).

methods to sufficient volumes of representative data. McClymont *et al.* (2013) presented an approach for the multi-objective optimisation of WDS using a new hyper-heuristic called the Markov-chain hyper-heuristic, for which one of the objectives was discolouration risk. They specifically sought to examine the impact of pipe diameter on discolouration risk. Trading off various considerations including optimising network design and rehabilitation costs along with discolouration risk is possible, with constraints such as pipe velocities and node heads. However, specific models of material regeneration are simplistic and are not sufficiently representative of reality to allow for prediction. This paper investigates how accumulation (regeneration) rate can be correlated with other system information, such as source water quality and pipe material.

Data mining for water resources knowledge discovery

Data-driven techniques from the field of machine learning are capable of identifying complex nonlinear relationships between inputs (factors potentially affecting accumulation) and output (the accumulation rate). Models capturing such relationships from historical training data can then be used for prediction for new input data. Some examples of this approach are present in the literature for similar applications. Opher & Ostfeld (2011) used genetic algorithms (GAs) to optimise model-tree regression methods for learning pipeline biofouling rates (focussed on biofilm measures as the output) from a large number of predictor variables in pilot studies. Giustolisi & Savic (2009) reported using the evolutionary polynomial regression (EPR) technique that utilises a multi-objective GA and applied it to a case study relating groundwater level predictions to total monthly rainfall. Savic *et al.* (2009) demonstrated that EPR offers a way to model multi-utility data of asset deterioration in order to render model structures transportable across physical systems. A polynomial expression for burst rate occurrence was derived using only asset data – the equation contained pipe length, diameter and age. EPR was also used to explore the relationships between climate data (such as temperature and precipitation-related covariates) and pipe bursts during a 24-year period in Ontario, Canada (Laucelli *et al.* 2014) with the models for the cold seasons showing best accuracy. Artificial neural networks (ANNs) have been used for modelling water quality variables for different aspects of drinking water systems and a comprehensive review is contained in Wu *et al.* (2014). Bhattacharya & Solomatine (2006) used MLP ANNs and M5 model trees to predict sedimentation in a harbour basin. One of their findings highlighted the importance of bringing a considerable amount of domain knowledge and expertise into the process of machine learning and this research follows this general precept.

A data-driven modelling approach is adopted for this paper, whereby two machine learning methods, Kohonen self-organising maps (SOMs) and EPR, make use of several sets of real world data for multivariate data mining based on the observed phenomena from both a qualitative and a quantitative perspective. The initial focus was on knowledge discovery of correlation across factors affecting accumulation rate for the national scale case study, with further investigation into actual EPR estimation accuracy and validation of the model in the second detailed local scale study. Figure 2 provides a flow chart of the methodology.

METHODOLOGY

SOMs

Initially, Kohonen SOMs were used for integrative and multivariate data mining of the potential factors affecting regeneration. SOMs are a type of ANN which draw inspiration from biological processes and resemble brain maps in the way they spatially order their responses by modelling the self-organising and adaptive learning features of the brain. The map evolves localised response patterns to input vectors. SOMs can be used for clustering and visual data mining and exploration. In unsupervised learning (also referred to as self-organisation) the inputs are presented to an ANN which forms its own classifications of the training data. The SOM is one of the most well-known ANNs employing unsupervised learning, first proposed by Kohonen (1990), and has the properties of both vector quantisation and vector projection algorithms. The prototype vectors are positioned on a regular

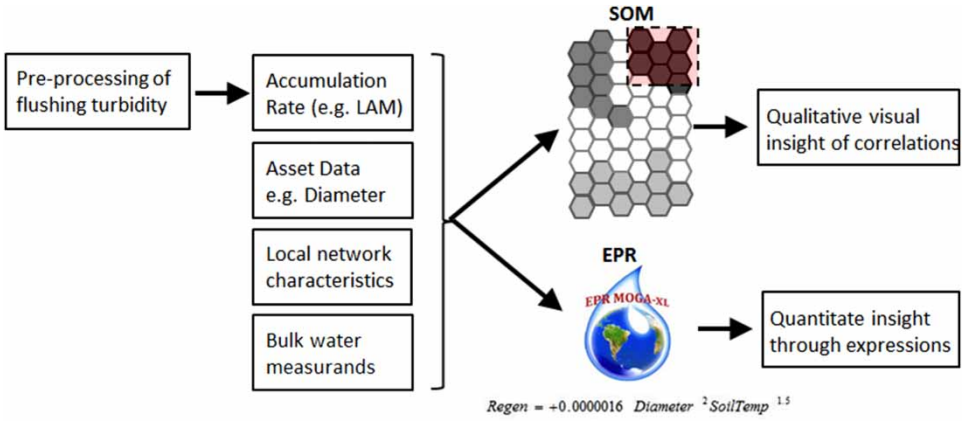

Figure 2 | Flow chart of methodology.

low-dimensional grid in a spatially ordered fashion allowing improved visualisation.

SOMs are commonly used for visual data mining/exploration and as pattern classifiers (such as for speech recognition) but also have potential in such areas as process control. SOMs have been used for analysis and modelling of water resources, including applications such as river flow and rainfall–runoff and surface water quality, as reviewed in Kalteh *et al.* (2008). Mounce *et al.* (2012) proposed their use in data mining microbiological and physico-chemical data for laboratory pipe rig data and for knowledge discovery from large corporate water company databases, through linking water quality, asset and modelled data (Mounce *et al.* 2014). SOM analysis does not make any assumptions about the distribution of the input variables or their relationship to one another. It allows higher dimensional data to be given a simpler visual representation in a smaller n-dimensional space determined by the investigator.

A SOM has two layers, an input layer (with the same number of nodes n as input variables) and an output layer. The output neurons are arranged into a one, two (usually) or possibly more dimensional lattice (often rectangular or hexagonal). Each output neuron is connected to the inputs by a vector of weights and also to its neighbours in the array.

Let $x = [x_0, x_1, \ldots, x_{n-1}] \in \Re$ be the input vector, where n is the number of input nodes. Let $w_j = [w_{0j}, w_{1j}, \ldots, w_{n-1j}] \in \Re$ be the weight vector of output neuron j. An outline of the basis of the algorithm follows.

Kohonen self-organising map algorithm

(1) Initialise network: Define $w_{ij}(t)(0 \le i \le n-1)$ to be the weight from input i to output node j at time t. Initialise these weights to small random values. Let the initial radius of the neighbourhood around node j, $N_j(0)$ be large.

(2) Present input: Present input $x = x_0(t), x_1(t), \ldots, x_{n-1}(t)$, where $x_i(t)$ is the input to node i at time t. Normalise the input vector.

(3) Calculate distances: Compute the distance d_j between the input and each of the weights of each output node j, given by:

$$d_j = \sum_{i=0}^{n-1} (x_i(t) - w_{ij}(t))^2$$

(4) Select minimum distance: Designate the output node with minimum distance d_j as c.

(5) Update weights: Update weights for node c and its neighbours, as defined by the neighbourhood $N_c(t)$. New weights are given by

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)) \quad \text{for} \quad 0 \le i \le n-1$$

for j in $N_c(t)$

where $\eta(t)$ is a learning coefficient which decreases with time, as does the neighbourhood size $N_c(t)$.

Normalise each weight vector that is updated.

(6) Repeat by going to 2.

This process repeats over a number of iterations, resulting in clusters of winning nodes that correspond to clusters within the input data thus evolving localised response patterns to input vectors. If input vectors are similar then they evoke a topologically close response. The algorithm is robust when presented with input vectors containing missing values: the dimensions corresponding to missing input vector values are simply ignored when finding the best matching output neuron and then updating the output neuron weights.

The SOM for the training vectors was generated using the program MATLAB (Version 7.14.0.739; The Mathworks Inc.) using the SOM toolbox (Version 2.0beta) developed at the Helsinki University of Technology (available online at http://www.cis.hut.fi/projects/somtoolbox). For this work, the input layer consisted of a number of neurons corresponding to the number of variables used and the output layer consisted of a hexagonal Kohonen map whose size was optimally selected by the SOM toolbox. A batch training method was used with a Gaussian neighbourhood. The initial learning rate of 0.5 was used for the first rough phase of training corresponding to the creation of a 'coarse' mapping which is when the global order is imposed on the map. Later the learning rate is reduced to 0.05 for the second phase, in which the fine structure is added to the map while preserving the global order. Kohonen (2001) reported that system parameters are not 'brittle' as is the case for other types of network algorithm and that the self-ordering phenomenon occurs for quite diverse values of the parameters.

The SOM is a useful tool in visual correlation discovery (the primary use in this application), that is in inspecting the possible correlations in the input data – the component plane representation allows visualisation of the relative component distributions, and these planes are effectively slices of the SOM (with each slice a dimension). Each plane represents the value of one component in each node in the SOM (typically using a colour range) and by comparing these planes even partial correlations may be found. By comparing component planes we can see if two or more components (dimensions) correlate. By picking the same neuron in each plane (in the same location), we could assemble the relative values of a 'codebook' vector of the network.

EPR

EPR (Giustolisi & Savic 2006), a hybrid data-driven technique, was applied, which combines GA with numerical regression for developing simple and easily interpretable mathematical model expressions. Polynomial models are generated combining the independent variables together with the user-defined function as in Equation (1):

$$\hat{Y} = \sum_{i=1}^{m} F(\boldsymbol{X}, f(x), a_i) + a_0 \quad (1)$$

where $\hat{Y}$ is the EPR estimated dependent variable, $F(.)$ the polynomial function constructed by EPR, $\mathbf{X}$ is the independent variable matrix, $f(.)$ is a user defined function, a_i the coefficient of the ith term in the polynomial (with a_0 the bias) and m is the total number of polynomial terms. The multi-objective genetic algorithm (MOGA) (Giustolisi & Savic 2009) allows the development of multiple models by simultaneously optimising fitness to training data and parsimony of resulting mathematical expressions (in terms of numbers of terms and equation complexity). The principle of parsimony states that for a set of otherwise equivalent models of a given phenomenon one should choose the simplest one to explain a given data set (Savic *et al.* 2009). These models have a capability to select a subset of the most relevant inputs and the relationship type relevant for model predictions, i.e., identify the most relevant input covariates. This is in contrast to some other data-driven models such as ANNs which are usually focussed on goodness of fit only, and may be prone to over-fitting. EPR consists of a two-stage process: a GA identifies the model structures and a numerical least-squares regression estimates the coefficients in the selected expressions. Usually, a pseudo-polynomial expression is used, where each term comprises a combination of the candidate inputs and each covariate has its own power (exponent) value. Each polynomial term is multiplied by a constant coefficient(s) which is determined during the search, and can include user-selected functions. In the GA search, the candidate power values are selected from a user-defined set of values, which generally includes zero (any candidate raised to the power zero is excluded from that model).

MOGA ranks the candidate model based on three criteria: goodness of fit, parsimony of covariate variables (number of inputs) and parsimony of mathematical equations (number of polynomial terms). The result is a set of models returned as formulae. Their symbolic nature allows their inspection in light of the physical and domain knowledge of the phenomena. Full details of the methodology are contained in Giustolisi & Savic (2006).

The EPR MOGA – XL tool version 1.0 was used for the static regression modelling of the regeneration rate. This software uses a MOGA optimisation strategy based on the Pareto dominance criteria (Giustolisi & Savic 2009). Data were not scaled in accordance with general procedure for this type of static regression application, and proportionality factors for generations were default as suggested in the literature. In this research, possible power values were [−2, −1.5, −1, −0.5, 0, 0.5, 1, 1.5, 2] thus considering most well-known relationships, e.g., linear, quadratic, inverse linear, square root, etc. The regression method for parameter estimation was non-negative least squares (i.e., $a_j > 0$) and the bias term was assumed equal to zero. The reader is referred to the user manual for the details of the EPR toolbox and the various different components of its graphical user interface (Laucelli *et al.* 2005). EPR was used to explore producing novel simple expressions to capture and highlight the important factors in the accumulation of discolouration material, based on the case studies, and to explore estimation of this rate.

CASE STUDIES

Case study 1: UK national

Description

An extensive nationwide data set (presented in Husband & Boxall 2011) was collated comprising field data collected during uni-directional flushing operations within live WDS in partnership with nine collaborating UK water service providers who serve over 40 million customers. These water companies provided access to a total of 36 sites and were selected to cover a range of factors suspected to have an influence on material accumulation rates, including pipe material, diameter, volume, source water and bulk water quality factors, such as the presence of upstream unlined cast iron pipes and water treatment processes such as coagulation and hydraulic conditions. Site-specific details could then be correlated to identify influencing factors. The study used 15 different network locations from across the UK with 67 monitored pipe sections. Site selection included, wherever possible, sites that had not previously experienced hydraulic disturbances (as indicated by water company records), such that the initial flushing of each pipe mobilised a large amount of material from its wall. Ideally, sites should also then have no large hydraulic disturbance between flushes. For each site, the two operations (initial visit and repeat) were planned to be completed under identical conditions (same time of day, flushing flow rates and duration) but a year apart, thereby producing two sets of comparable turbidity data. These were used to calculate an annual regeneration rate of erodible discolouration material, irrespective of operational date or location. The annual regeneration rate is a percentage figure relating the amount of material accumulated and subsequently mobilised 12 months after the initial site visit.

The full fieldwork methodology employed is presented in Husband & Boxall (2011) along with further site details. These comprehensive site data were collected by a single researcher who was directly involved on site with all operations, in collaboration with water company personnel, hence although of limited size the data set has been assembled into very high quality information. By collating data sources (i.e., asset and water quality data) from all sites it is then possible to explore the relative influence of the factors identified as possibly influencing material accumulation rate.

Data sets and preparation

The observed temporal turbidity traces for each pipe section from the initial and repeat operations can be plotted together with the measured flushing flow rate. Figure 3 shows an example of a turbidity trace for a 75 mm CI pipe. In Husband & Boxall (2011), in order to determine a useful regeneration index, three methods were trialled to obtain a score from these plots allowing simple comparisons between the initial trial and the repeat trial as an indicator of the regeneration rate. These measures were peak turbidity, average turbidity (mean of all data within measured time frame) and finally a

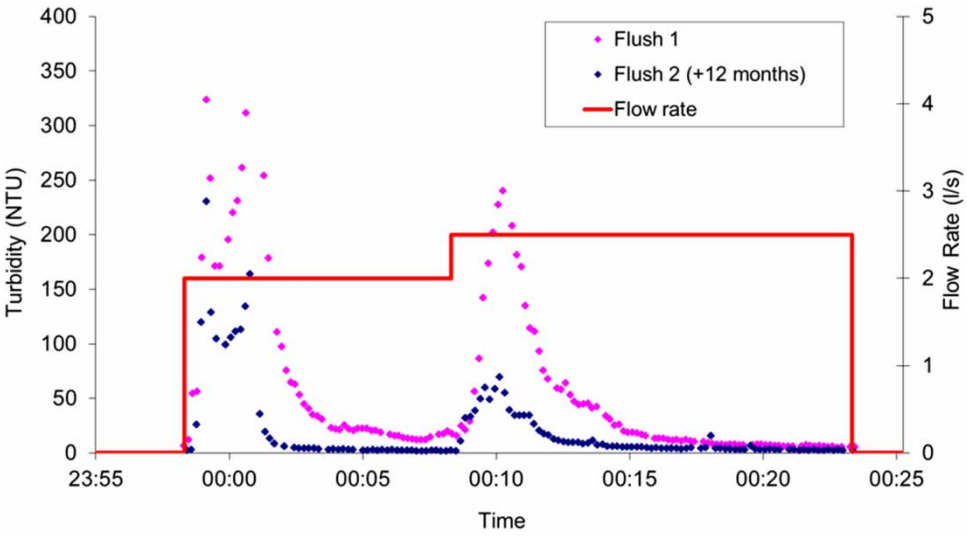

Figure 3 | Turbidity trace for initial and return flush, 75 mm, 92 m cast iron pipe (with flushing steps of 2 L/s (0.45 m/s) and 2.5 L/s (0.55 m/s).

metric based on integration of the time turbidity plots: effectively a step in calculating an amount of material (Boxall *et al.* 2003a). It was concluded for this data set that the material mobilised from sites at this national scale is not consistent due to differences in water quality so multiplication by a common conversion factor to suspended solids, a mass of material, was not undertaken. Further, it was observed that the average and integration method for calculating regeneration percentages return comparable results. Consequently, a similar approach of using the average turbidity was utilised here with the appropriate score determined for each operation, being a percentage annual regeneration value, indicating the rate material returns to a fully developed and maximum discolouration risk.

The results from the fieldwork were compiled and preprocessed to include variables considered to be possible influencing factors in the regeneration of discolouration material including bulk water iron concentration, source water, coagulation treatment type, presence of upstream unlined cast iron pipes, pipe material, pipe diameter, daily hydraulic conditions and configuration. The type of the site refers to a subjective classification based on the flow route – 'main' refers to pipe lengths that do not terminate in a dead end and are not part of a loop, 'loop' referring to an area likely to be affected by flow reversals or so-called 'tidal points' due to multiple potential flow paths, and dead end being self-explanatory. Some of the variables were binary or ordinal/nominally encoded and these are detailed in Table 1. Most materials were either cast iron (all unlined) or a plastic.

Continuous variables included Fe concentration, diameter (mm), volume (m^3) and modelled shear stress values – both daily and (maximum) flushing values (Pa). In general, the higher the scoring the higher the risk of material accumulation due to perceived increased material source or accumulation mechanism, e.g., an unlined CI pipe has a score of 3 in the pipe category as it is regarded as a possible source of corrosion products (informed by Husband & Boxall 2011). Upstream iron is based on information supplied by the consortium water companies and incorporates a

Table 1 | Encoding for discrete discolouration factors

Value	Supply	Treatment	Upstream iron	Pipe material	Area	Dead end	Loop	Main
0						Not DE	Not loop	Not main
1	Ground	None	None	PE/PVC	Urban	DE	Loop	Main
2	Blend	Al Coag	Minor	AC	Rural			
3	Surface	Fe Coag	Significant	CI				

subjective element. A number of water quality measures have been considered as possibly influencing discolouration material accumulation processes. These measures could be considered as independent sources of material contributing to the generation of discolouration material. However, analysis of flushing samples has shown iron to be the dominant constituent of discolouration material, independent of site conditions (Seth *et al.* 2003).

Case study 2: Dutch local scale long-term monitoring

Description

An area in Purmerend, a town in the Netherlands, has been flushed four times in 5 years. The area has the same water supply and treatment and a total of 12.3 km of mainly AC and PVC pipes, with 2,310 home connections. Note that the length variable is the length of pipe affected by the flushing operation. Typical source water values for Fe and Mn were 1.7 μg/L and 0.2 μg/L, respectively. The flushing programme is described in detail in Blokker (2010).

Data sets and preparation

For each pipe the turbidity (FTU) that was measured during flushing at the pipe location was recorded. From this the locally accumulated material (LAM) was computed according to Equation (2). In 2013, the flushing programme was slightly changed and higher flushing velocities were used, hence in order to obtain comparable quantities the turbidity was multiplied with the Q_{flush} (in litres) as shown in Equation (2):

$$\text{LAM}_{\text{year}} = \text{Turbidity} \cdot Q_{\text{flush}} \cdot \Delta t \cdot \frac{Q_{\text{flush}} \sum_{t_i \geq t_{\min}}^{t_i \leq t_{\min}} |\text{Turbidity}(t_i)}{\text{year\_between\_flushes}} \left[\frac{\text{FTU} \cdot L}{\text{year}}\right] \quad (2)$$

Figure 4 shows an example of a turbidity trace for a 1,000 m length pipe, 200 mm AC and some pieces of 150 mm PVC (internal diameter of 141 mm).

Case study 3: Dutch local scale highly repeated flushing

Description

An area in Volendam, a town in the Netherlands, has had an extensive and intensive flushing programme for several years, where every 3–6 weeks up to 21 pipe lengths in an urban distribution area have been flushed as part of a continuing monitoring programme. Results have been used from January 2012 to August 2014. The site characteristics are as follows: surface water supplied, UV and H_2O_2 treatment and pipes mainly of PVC and AC material (with no upstream iron).

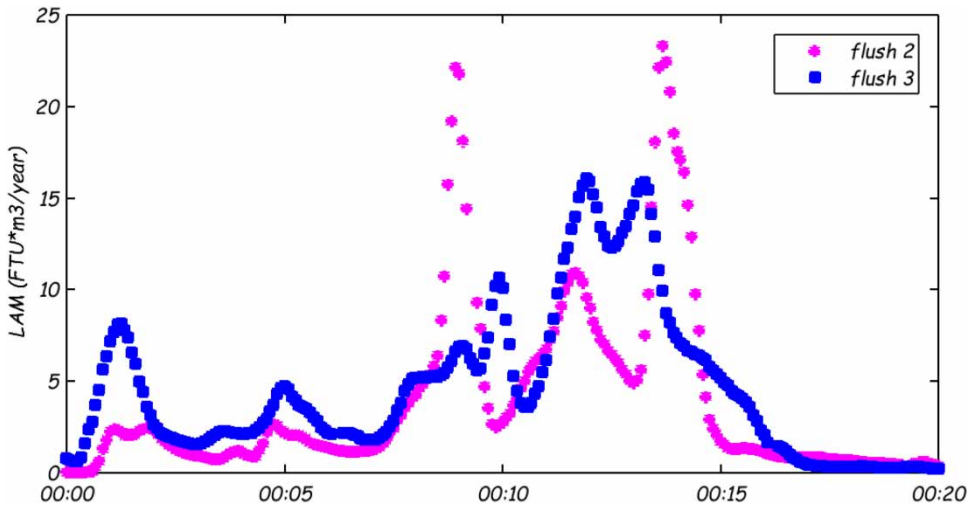

Figure 4 | LAM per year for first return flush in March 2010 (17 months after initial flush) and second return flush in October 2010 (24 months after initial flush), flushing action 2 on mainly a 200 mm, 1,000 m AC pipe, with flushing of 42 L/s (1.42 m/s) and 50 L/s (1.71 m/s).

Data sets and preparation

The data set contains some asset information, soil temperature, flushing shear stress, the average turbidity (using an integrative method of calculation) and the accumulation rate (per day). A similar process was used as in case study 2 to calculate LAM as shown in Equation (3):

$$\mathrm{LAM}_{\mathrm{day}} = \mathrm{Turbidity} \cdot Q_{\mathrm{flush}} \cdot \Delta t \cdot \frac{Q_{\mathrm{flush}} \sum_{t_i \geq t_{\min}}^{t_i \leq t_{\min}} |\mathrm{Turbidity}(t_i)}{\mathrm{days\_between\_flushes}} \left[\frac{\mathrm{FTU} \cdot L}{\mathrm{day}}\right] \quad (3)$$

Figure 5 shows an example of a turbidity trace for 710 m pipe of PVC 300 (270 mm inside pipe diameter).

RESULTS

Case study 1: UK national

SOM results

A top level data set was assembled for case study 1 and this was used as input to a SOM, along with the return flush NTU (Flush2) and the average annual percentage regeneration obtained by dividing the repeat by the initial results (Flush2/Flush1). Figure 6 provides the component planes of the SOM allowing visual inspection of how variables change relative to each other and in comparison to the regeneration index (REGEN). Note that SOMs are able to interpolate missing values, unlike EPR.

The resulting Kohonen map comprises colour-coded or greyscale shaded hexagons that summarise all of the component planes that represent individual variables. In Figure 6, there are two separate parts of the SOM display. These include the summary U-matrix and then the 15 component planes for individual variables. The U-matrix shows the distances between the reference vectors of adjacent cells. Ridges in the U-matrix therefore delineate clusters in the trained SOM.

Each hexagonal cell represents multiple neurons, which are the mathematical linkages between the input and output layers. In the component planes for individual variables, the colouring or shading corresponds to actual numerical values for the input variables that are referenced in the scale bars adjacent to each plot. Blue (dark) shades show low values and red (white) corresponds to high values. This allows visual comparison of their clustering relationships with other variables by comparing regions of the map across component planes.

The SOM confirms that the sites with the greatest regeneration rates (bottom right component plane, lower right area) are surface source water sites, non-plastic pipes, and high iron concentration, with iron coagulation treatment and combined with unlined upstream cast iron pipes.

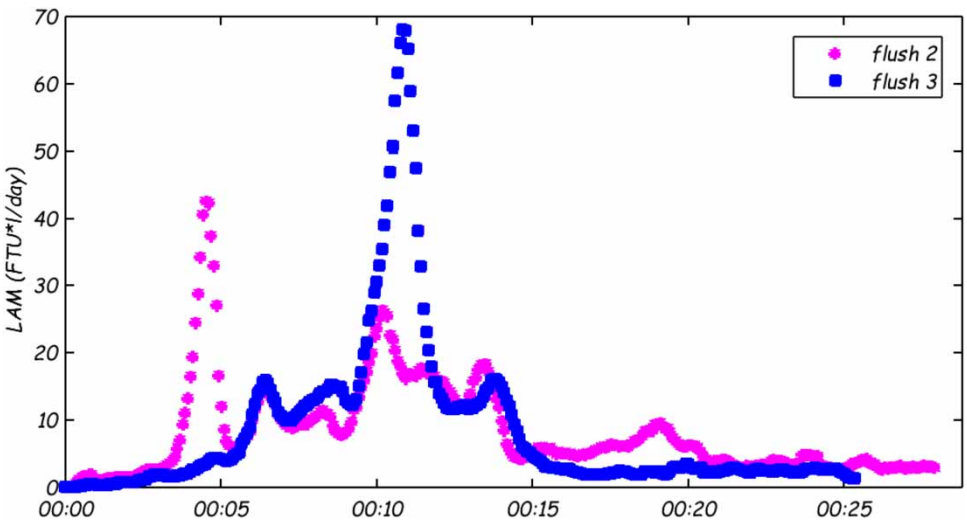

Figure 5 | LAM per day for first return flush on 23 February 2012 (4 weeks after initial flush) and second return flush on 29 March 2012 (9 weeks after initial flush), flushing action 3 on mainly a 270 mm, 710 m PVC pipe, with flushing of 44 L/s (0.77 m/s) and 42 L/s (0.73 m/s).

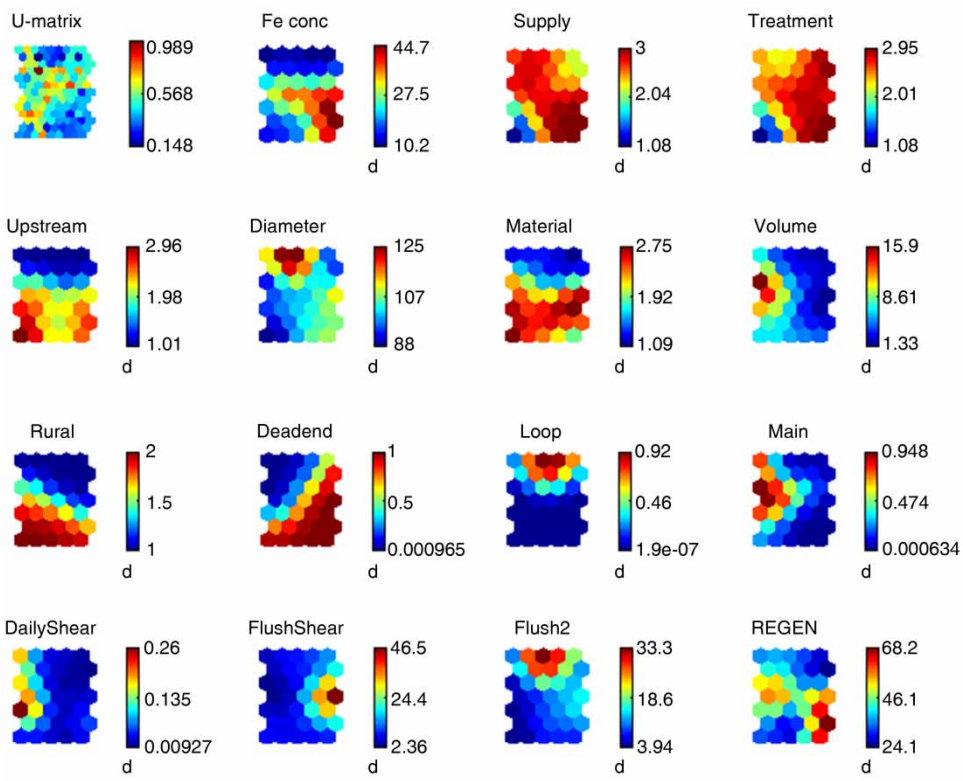

Figure 6 | SOM for regeneration fieldwork results with UK site determinants. Please refer to the online version of this paper to see this figure in colour.

Dead ends would also appear to be a factor for increased regeneration rate. In contrast, trunk mains and urban locations would appear to result in a lower regeneration rate. Combinations of factors are evident, such as medium diameter pipes in rural areas with upstream unlined cast iron, are related to higher rates. These findings support the hypothesis that certain key factors will, in general, determine the rate of material accumulation on pipe walls.

EPR results

EPR is used here to attempt to derive an expression for an annual regeneration rate and not the risk or magnitude of any potential discolouration event occurring. All the variables in Figure 6 (apart from Flush2 which is used in the calculation of the annual regeneration rate) were initially used, i.e., 13 candidate inputs in total. Input and output test (or cross validation) data were not used, due to the limited data points available for this case study and the principal goal being knowledge discovery. The seven model structures described in Giustolisi & Savic (2006) were applied, although the use of an inner function (such as logarithm, exponential, etc.) did not provide any additional accuracy and merely led to longer model run times. Results are presented here for standard polynomial structured equations produced by EPR. The MOGA process

Table 2 | Selected Pareto optimal regeneration rate estimation models identified by EPR

Model structure	CoD
$\text{Regen} = +9.5843\,\text{Fe Conc}^{0.5}$	0.40
$\text{Regen} = +6.6894\,\text{Fe Conc}^{0.5}\text{Material}^{0.5}$	0.49
$\text{Regen} = +9.7883\,\text{Material}^{2}\,\text{Loop}^{1.5} + 1.4608\,\text{Fe Conc}$	0.58
$\text{Regen} = +0.2990\,\text{Loop Flush Shear}^{1.5} + 6.382\,\text{Fe Conc}^{0.5}\,\text{Material}^{0.5}$	0.67

ran for 7,020 generations. Table 2 provides five models produced along with the coefficient of determination (CoD) which is based on the sum of squared errors and reflects model accuracy.

Table 2 reveals that in the simplest models the iron concentration and pipe material are key factors. Models with reduced parsimony, such as those with greater than seven terms could provide CoD greater than 0.8; however, due to the data set size for this case study overfitting is inevitable. By incorporating Flush2 as an additional input, the best single variable equation was still the first listed equation in Table 2, but the second equation then incorporated Flush2, along with FeConc with a CoD of 0.65. The multi-case strategy (MCS) variant of EPR utilises splitting data into subsets according to, for example, failure history, and it has been used to develop distinct models for different subsets of pipes (Giustolisi & Berardi 2009). An additional EPR study was conducted on cast iron material only for the case study. Table 3 provides two of these models (the simplest). Note that it was possible to get CoD greater than 0.9 for only five terms in this case, however caution over the subset data size needs to be emphasised since the data set was reduced to 44% of the original size. However, CoD for the equation with only the iron concentration term is improved over that in Table 2. Bulk water iron concentration could potentially be a single measure capturing the dominant influence of a number of other water quality factors: source water, coagulation treatment processes and quantity of unlined upstream iron.

Case study 2: Dutch local scale long-term monitoring

In Figure 7, the discolouration response due to flushing is provided as measured at each visit to each pipe in the network. In the figure, pipes are coloured or shaded according to percentiles of turbidity response for each visit's distribution of turbidity response values: yellow (thin black line): no value (not flushed or not recorded); black (very light grey): lower 50 percentile; cyan (light grey): 50–80 percentile; blue (grey): 80–90 percentile; purple (dark grey): 90–95 percentile; red (thick black line): upper 5 percentile. From the figure it can be seen that the degree and location of accumulation changes across the network with each visit. It should be noted that the flushing operations in this area were from routine operations, not rigorously managed for scientific investigation. The time of day, flushing rate, etc. were not rigorously repeated between operations. Additionally, it should be noted that the durations between the sequence of images in Figure 7 are not consistent.

Table 3 | Selected Pareto optimal regeneration rate estimation models (cast iron material) identified by EPR

Model structure	CoD
$\text{Regen} = +10.8244\,\text{Fe Conc}^2$	0.46
$\text{Regen} = +68.5251\,\text{Loop} + 10.1151\,\text{Fe Conc}^{0.5}$	0.79

Total accumulation rate

To explore beyond the pipe level variability evident in Figure 7, it was decided to calculate total network behaviour. A total accumulated material (TAM) value was calculated from the summation of all measured turbidities during each visit and the total pipe length (Equation (4)).

$$\text{TAM} = \frac{\sum_{i=1}^{F}\sum_{t=1}^{\infty}\text{Turbidity}(t)\cdot Q_{\text{flush},i}\cdot\Delta t}{\sum_{i=1}^{F} L_i}\,[\text{FTU}\cdot l/m] \tag{4}$$

with Δt the measurement time step in s, i counting all flush actions, L is length per flush action, Q_{flush} is the flushing flow.

Figure 8 presents the total network accumulation behaviour as a function of time, from which it can be seen that the overall rate of accumulation was highly consistent, evident in the similar gradients. Thus while the individual pipe accumulation behaviour may be suspect, the overall system behaved in a repeatable manner.

While having a similar number of data points as case study 1, the flushing exercises for case study 2 did not have the high degree of control of the former and were confined to a specific area. A SOM is provided in Figure 9 by way of example, and it is difficult to draw any strong conclusions. This led to EPR analysis being unfeasible for providing any generic results. Encodings are as set out in Table 1, and note that most material is AC.

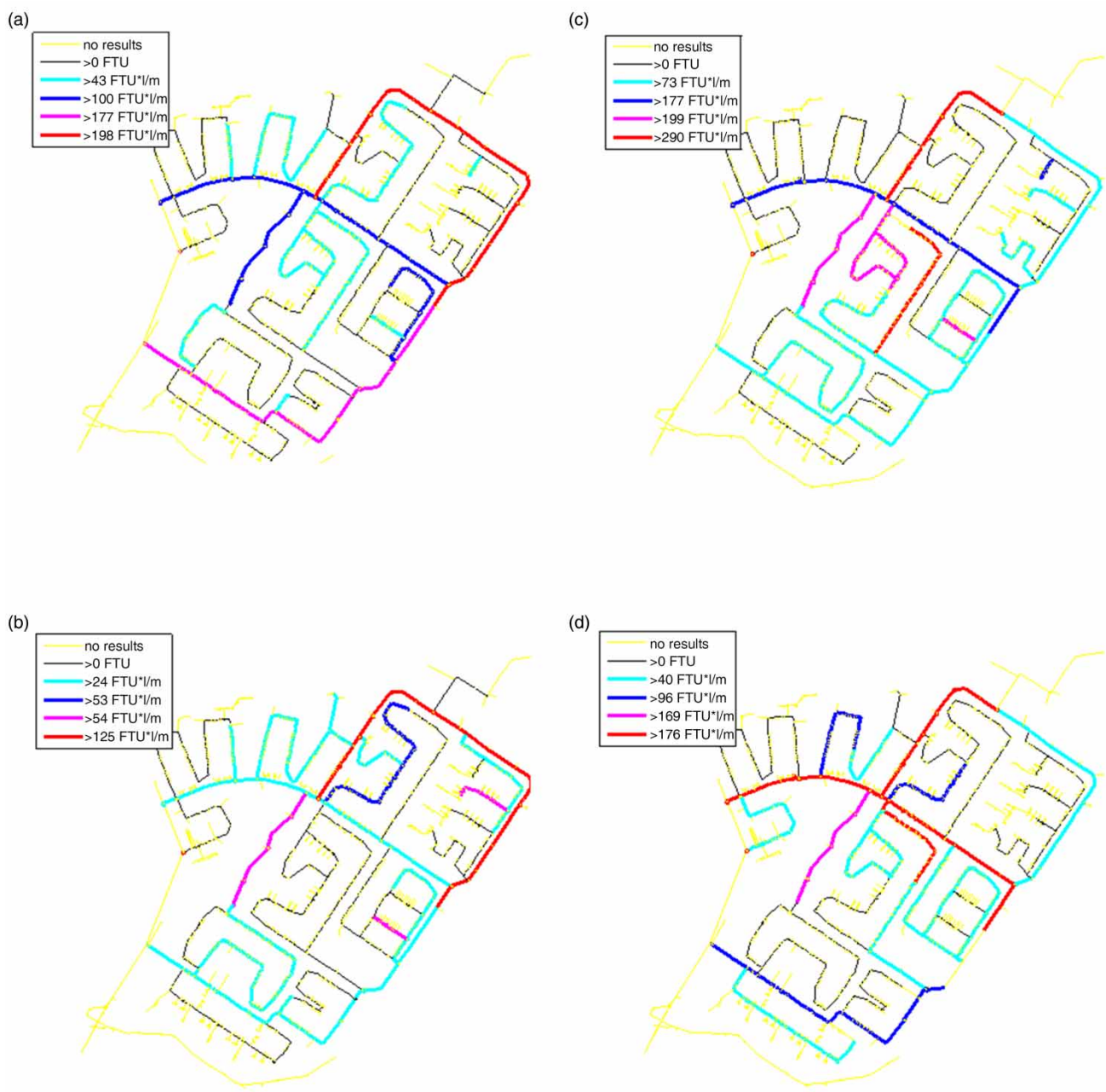

Figure 7 | Accumulation rate per flushing action: (a) March 2010; (b) October 2010; (c) August 2013; (d) October 2014. Please refer to the online version of this paper to see this figure in colour.

Case study 3: Dutch local scale highly repeated flushing site

SOM results

Figure 10 provides the component planes of the SOM for case study 3. This allows visual inspection of how variables vary relative to each other and in comparison to the daily regeneration index (LAM). The amount of material mobilised is particularly clearly correlated with regeneration rate.

In Figure 11, all the accumulation rates for all the flushing actions over time are provided, normalised to a maximum rate at each site. The black dotted line is the temperature of the soil at 1 m depth. The correlation is apparent

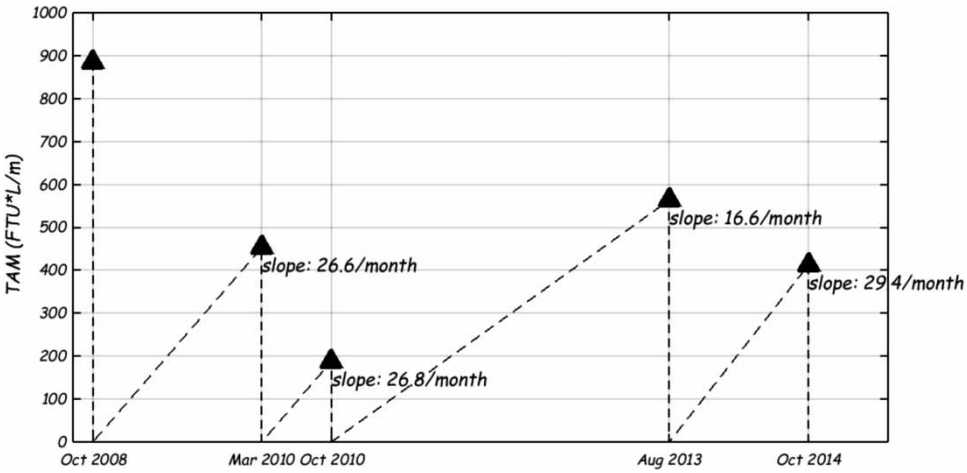

Figure 8 | Total accumulated material in Purmerend area over time.

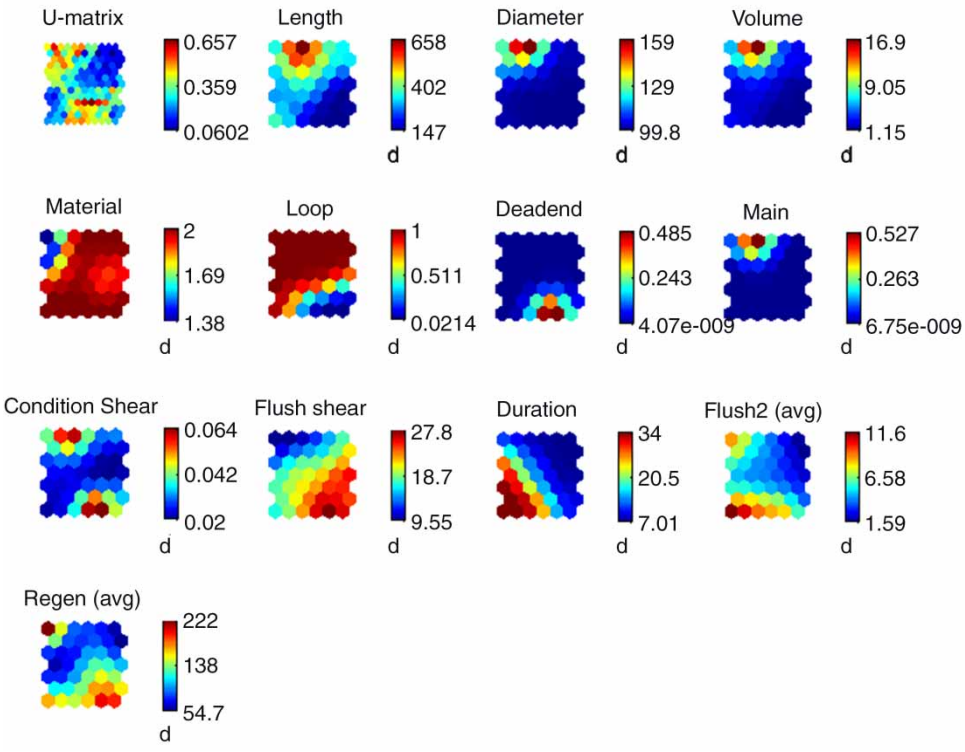

Figure 9 | SOM for flushing operations in case study 2. Please refer to the online version of this paper to see this figure in colour.

and hence the importance of temperature, which is also partially evident in Figure 10, for larger diameter pipes. Up to 21 sites were flushed for up to 33 times over the studied time period; in total this provided 495 data points. Some data were removed, because: (1) during the summer vacation there was a lot less flow into the system which affects the accumulation rate, but this variable was not used in the analyses (50 out of 495 were removed); and (2) visual inspection on turbidity measurements indicated that some measurements were not reliable (24 out of 495 were removed). This means 425 data points remained. The removed data explain the gaps in Figure 11.

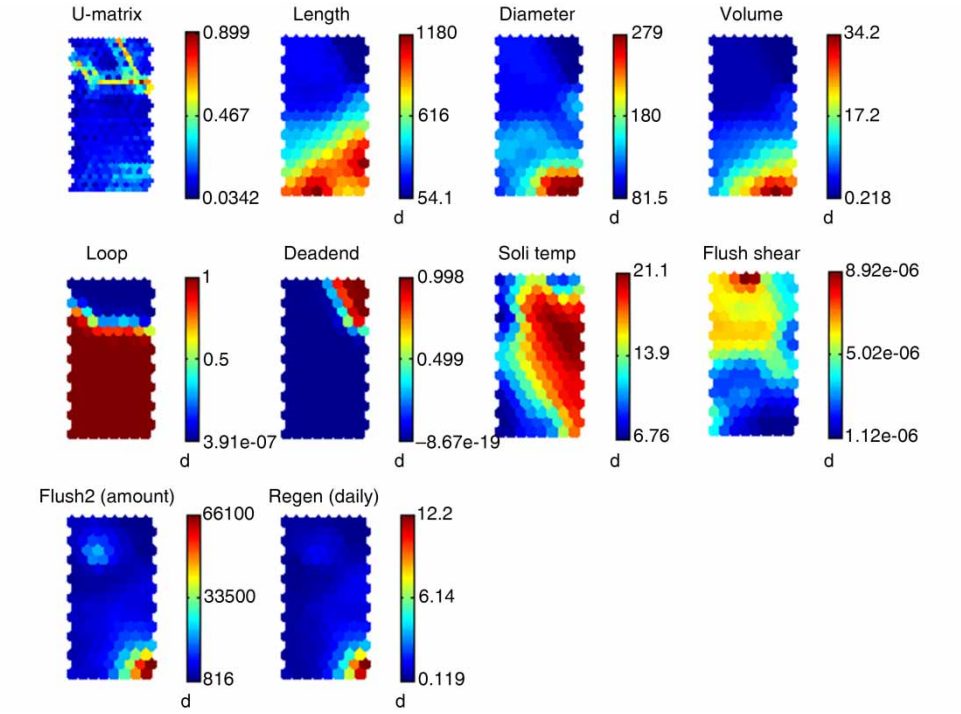

Figure 10 | SOM for multiple flushing operations in case study 3. Please refer to the online version of this paper to see this figure in colour.

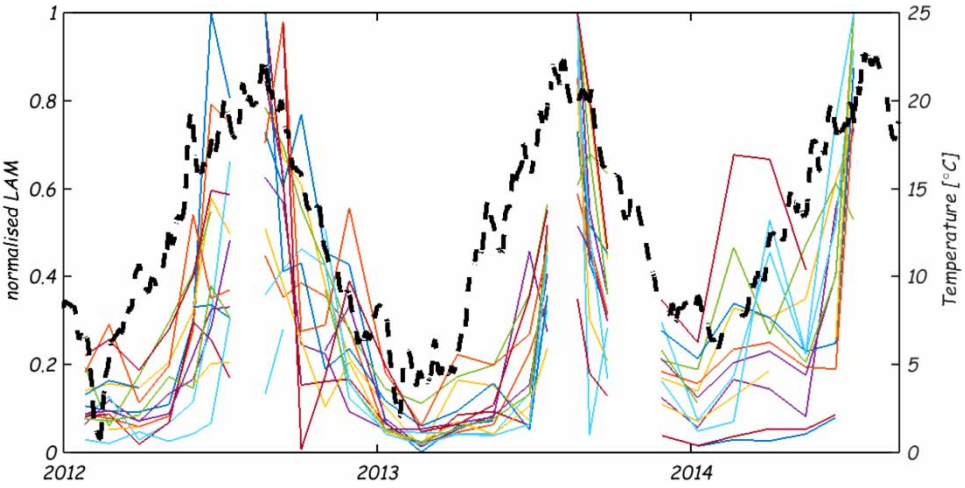

Figure 11 | Dotted black line = calculated soil temperature at −1 m (validated with drinking water temperature samples). The other lines are the accumulation rates (per day) over time for all the flushing action locations, normalised to a maximum of 1.

EPR results

EPR was used in a similar manner to previously to attempt to derive an expression for a daily regeneration rate based on all flushing actions and their repeats. An extensive data set was available with 624 individual flushing actions available in theory (before issues of missing data had to be dealt with) with accumulation rates calculable for each successive pair of flushing operations at each site. All the variables in Figure 10 were used in the EPR, i.e., eight candidate inputs in total. In

this case study, the amount of material in the second flush was used as calculated by the integrative method (recall that in the UK national case study 1 the average turbidity had not been the dominant predictive factor, however it had proved to be so for the Dutch local scale study). Results are presented here for standard polynomial structured equations produced by EPR. The MOGA process ran for 4,320 generations. Table 4 provides three models produced along with the CoD when utilising all data. Note the relatively high prediction from just using the Flush2 amount, but that the addition of soil temperature further improves the forecast.

Table 4 reveals that in the simplest models the amount of material mobilised in the second flush dominated the contribution to the models (somewhat in contrast to case study 1) with only the soil temperature feature further adding to the CoD excellent accuracy. With very complex models with far reduced parsimony, including other asset parameters, a CoD of 0.96 could not be exceeded. The relatively large number of data points and repetitive nature of the flushing operations in time and space allowed for independent training and validation. A K-fold cross validation approach was applied rather than hold-out validation (Kohavi 1995). The data are broken into K-blocks (five were used here, resulting in an 80/20 split). Then, for K=1 to X, the K*th* block becomes the validation (or test) block with the remaining data becoming the training data. EPR training and testing is conducted and K then updated. Recall that 21 specific locations were flushed 33 times each, within a period of between 3 and 6 weeks. The K-folds were randomly selected from all flushing actions. Table 5 provides the results – the best equation with material amount and soil temperature is used. Mean cross validation CoD was 0.945 for training data and 0.930 for testing data. Average CoD for the single variable equation (structure from Table 4) was 0.883 for unseen data.

Figure 12 provides scatter plots of the regeneration rates from the observed data and the EPR model-predicted data

Table 4 | Top three Pareto optimal daily regeneration rate estimation models identified by EPR

Model structure	CoD
$\text{Regen} = +0.16156\,\text{Flush 2 Amount}$	0.897
$\text{Regen} = +0.0097246\,\text{Soil Temp Flush 2 Amount}$	0.947
$\text{Regen} = +0.072493\,\text{Flush 2 Amount} + 0.00031\,\text{Soil Temp}^2\text{Flush 2 Amount}$	0.954

Table 5 | K-fold validation results for optimal equation for case study 3

F-fold	Equation	Training CoD	Testing CoD
1	$\text{Regen} = +0.010069\,\text{Soil Temp Flush 2 Amount}$	0.954	0.903
2	$\text{Regen} = +0.0098479\,\text{Soil Temp Flush 2 Amount}$	0.942	0.956
3	$\text{Regen} = +0.038656\,\text{Soil Temp}^{0.5}\,\text{Flush 2 Amount}$	0.948	0.897
4	$\text{Regen} = +0.009523\,\text{Soil Temp Flush 2 Amount}$	0.938	0.955
5	$\text{Regen} = +0.0099064\,\text{Soil Temp Flush 2 Amount}$	0.944	0.937

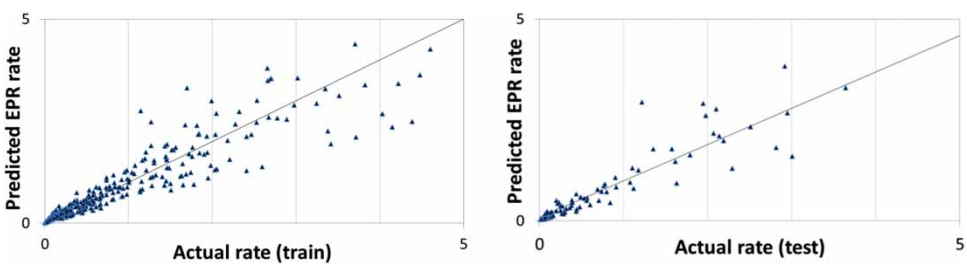

Figure 12 | Scatter plots of the observed data and the EPR model-predicted regeneration rate (train and test for F-fold 5).

for K-fold 5, and with a small percentage (< 5%) of outliers out of scale.

If EPR is run without the amount of material from the second flush, the best equation (Equation (5)) involves soil temperature and diameter with CoD of 0.624 on training data – comparable to similar equations in Table 2 for case study 1. Duration between flushes was utilised as an additional input but did not contribute (the temporal element was already accounted for in the regeneration rate):

$$\text{Regen} = +0.0000016\,\text{Diameter}^2\,\text{Soil Temp}^{1.5} \quad (5)$$

DISCUSSION

There exists a significant contrast between case study 1 (the national UK data set) and the two local Dutch case studies. The large scale UK data set was collected in very carefully controlled conditions, over many varying areas with different materials, source waters, treatment types, etc. (Husband & Boxall 2011). In this case regeneration rates were available from only a single pair of flushing operations a year apart. This case study shows that discolouration processes were dominated by iron, although manganese, which has been shown to be the other dominant metal in UK discolouration samples (Seth *et al.* 2003), was not available. It was not possible to explore seasonal effects for this data set as flushing was at identical times of the year.

In the Dutch systems, local areas were used with identical source water, treatment type and exclusively plastic or AC material pipelines. For case study 2, flushing operations were deemed less repeatable in nature, evidently an impact of the variability in the flushing shear stress between repeat visits. Schematics showing accumulation per pipe did not show repeatable patterns (Figure 7). However, when summed over the area, generally similar regeneration rates were observed except for a slight difference during the third period (Figure 8). Only four repeats were conducted (and extrapolated to an annual regeneration rate), but with different durations between flushing operations such that it was not possible to explore seasonal or temperature-dependent effects. Difficulties in obtaining pipe level understanding and expressions from case study 2 highlight the need for trial accuracy, replication (since the number of data points available here was very small) and control.

Case study 3 was a very extensive and intensive flushing programme concentrating on regular and repeated flushing of 21 specific locations (still within one section of the supply system). These were flushed 33 times each, over multiple years, with a return period of between 3 and 6 weeks. Pairs of flushes from this data set allowed calculation of daily regeneration rates at high spatial and temporal resolution with results revealing seasonality, or more specifically, temperature dependent, effects (Figure 11 and Table 4). Temperature has an obvious link to biological reactions (Sharpe 2013) and suggests the importance of biofilm processes in material accumulation within pipes. At the scale of 3–6 weeks, linearity of regeneration rate is suggested (Table 4); this is in agreement with field data and analysis reported by Cook & Boxall (2011).

It might be queried whether the use of SOM and EPR can be complementary? In case study 3, we can see that when EPR was applied, this was dominated by Flush2Amount and soil temperature (see Table 4). However, by inspecting the SOM (Figure 10) we can see evidence for the combination of high soil temperature and high diameter being linked to higher daily regeneration rate (bottom right area of component plane). It is only when Flush2Amount is specifically excluded from the EPR input list that the best performing equation which uses these two variables emerges (Equation (5)). Hence, the SOM can suggest and inform combinations of variables for EPR particularly if equations may be required using variables of certain types.

The equations that have been derived here provide the opportunity to make estimates of regeneration rates/discolouration material accumulation rates. Such estimates are vital for planning proactive management, in particular the return frequency between operations to maintain a desired level of system cleanliness or level of discolouration risk. The Dutch data sets suggest that pipe-specific estimates can be made with a high degree of accuracy from previous flushing results for a given pipe. However, such data are often unavailable, and the UK data set suggests that where only general asset and bulk water quantity is available estimates have a high degree of uncertainty. In this case, the ability to quantify uncertainty would be desirable to allow practitioners to calculate a range of scenarios and hence inform risk-based management decisions.

It should be noted that UK and Dutch systems have some fundamental differences, affecting both in-pipe hydraulic conditions and bulk water qualities. For example, Dutch systems have reduced diameters and generally plastic materials whereas UK systems generally have older, larger pipes and much more diverse materials. There is also the issue of chlorinated versus non-chlorinated systems. While these differences will have significant effects on discolouration processes, important and relevant findings were revealed here irrespective of these. It is suggested that future work in each country should address the other factors that seem to have dominated the results presented here. Future UK flushing programmes could usefully explore the effect of seasonality by conducting repeat trials at less than annual return intervals. Such a study was reported in Boxall *et al.* (2003b) but for one groundwater supplied site only that experienced very little change in water temperature at the pipeline studied, hence showing little temperature or seasonal dependence. From this and the case study 3 results reported here, it is interesting to speculate if it is variation in source water temperature or change in water temperature due to soil/ground conditions that is more important or a combination of the two. In the Netherlands, future work could explore a wider range of systems, perhaps using a similar strategy to case study 1 to investigate the effects of bulk water, as bulk water iron concentration was found to be the dominant factor in the UK data set reported here.

CONCLUSIONS

Improved comprehension of discolouration material accumulation processes and prediction of accumulation rates in WDS are essential for proactively managing mains rehabilitation and subsequently assessing the effectiveness of any interventions made. Discolouration risk and material accumulation rate are not the same. Discolouration risk consequence may be considered primarily a function of the location of an event within a network and the population exposed. Discolouration risk probability may be considered as a cross product of the duration since last disturbance or cleaning operation, the accumulation rate and the specifics of a given hydraulic mobilisation event. Of the three components, accumulation or regeneration rate is currently the most uncertain, but potentially controllable through treatment works or network interventions.

This paper presents some findings of applying data mining techniques to investigate material accumulation (regeneration) rate and the dependence on factors believed to influence the process. Three case studies were examined, having differing data quantity and quality resolutions: (1) a high quality and representative data set compiled across the whole of the UK (including varying conditions and source waters); (2) a Dutch local scale area flushed four times in 5 years; and (3) a Dutch local scale area very extensively flushed 32 times (periods of between 3 and 6 weeks) in 21 locations over multiple years. Case studies 2 and 3 had identical source waters and treatment. Key findings include the following:

- SOMs are a very useful tool for visual correlation discovery, that is, in inspecting the possible correlations in the input data across multiple dimensions, with each component plane being effectively a slice of the SOM. By comparing component planes we can see if two or more components (dimensions) correlate. This ability to synthesise and present multi-dimensional data (which might otherwise be impossible for humans to interpret) in a higher-fidelity representation is particularly useful for qualitative and intuitive communication with practising engineers. While not as definitive as equations derived from EPR, this data-driven approach still provides a level of knowledge discovery and evidence/audit trail beyond 'engineering judgement'. For the nationwide UK case study, the SOM helped confirm that high bulk water iron concentration, surface source water sites, non-plastic pipes, iron coagulation treatment and the presence of unlined upstream cast iron pipe are all factors contributing to a higher material accumulation rate.
- The EPR modelling paradigm implements a multi-objective genetic search algorithm, where the objective functions are accuracy (measured using CoD) and parsimony (number of covariates and equation complexity). EPR was applied to the application of predicting material accumulation (regeneration) rate for two of the case studies. In case study 1, the simplest equations involved bulk iron concentration, pipe material and looped

network areas. In case study 3, the significantly increased data set allowed K-fold cross validation. The optimal equations utilised the amount of material mobilised and soil temperature. Mean cross validation CoD was 0.945 for training data and 0.930 for testing data for an equation with both terms. When not using the material from the second flush as one of the inputs, an equation was derived with only diameter and soil temperature which illustrates the potential for ultimately developing transferable expressions for WDS.

- The type of EPR-derived equations that appear in Table 1 relate to bulk water iron concentrations and treatment and more definitive versions would allow network scale adjustments in these parameters to be assessed. The type of equations in Table 4 would be useful for stable networks and planning operations and their frequency.

There is expected to be much more potential for MCS experimentation for predicting accumulation rates based on pipe material and other cohort sub-divisions (such as source water or treatment type) as increased data sets become available as part of the PODDS programme of work (www.podds.co.uk), and with further international partners. Manganese would be a very useful input variable for future work, and other parameters identified through improved understanding of biofilm physiology due to the emerging importance of biofilm processes in discolouration (Douterelo *et al.* 2014a, b). More definitive EPR generated expressions should follow. Such predictive models could be very valuable for discolouration risk models (e.g., McClymont *et al.* 2013; Furnass *et al.* 2014).

ACKNOWLEDGEMENTS

This work was supported by the Pennine Water Group – Urban Water Systems for a Changing World Platform Grant (EP/I029346/1) and by the Pipe Dreams project (EP/G029946/1), both funded by the UK Science and Engineering Research Council. The authors would also like to thank the PODDS consortium of seven UK water companies for supporting the ongoing programme of discolouration research at the University of Sheffield. The PWN Water Supply Company is acknowledged for data provision. In addition, the hydroinformatics.it team at Politecnico di Bari are kindly thanked for provision of EPR-MOGA-XL for research usage.

REFERENCES

Bhattacharya, B. & Solomatine, D. P. 2006 Machine learning in sedimentation modelling. *Neural Networks* **19**, 208–214.

Blokker, E. J. M. 2010 Stochastic water demand modelling for a better understanding of hydraulics in water distribution networks. PhD thesis, Delft University of Technology, p. 212.

Blokker, E. J. M., Schaap, P. G. & Vreeburg, J. H. G. 2011 Comparing the fouling rate of a drinking water distribution system in two different configurations. In: *Urban Water Management: Challenges and Opportunities* (D. A. Savic, Z. Kapelan & D. Butler, eds). Centre for Water Systems, University of Exeter, Exeter, UK.

Boxall, J. B., Skipworth, P. J. & Saul, A. J. 2003a Aggressive flushing for discolouration event mitigation in water distribution networks. *Water Sci. Technol. Water Supply* **3** (1/2), 179–186.

Boxall, J. B., Saul, A. J., Gunstead, J. D. & Dewis, N. 2003b Regeneration of discolouration in distribution systems. In: *Proceedings of ASCE, EWRI, World Water and Environmental Resources Conference*, 23–26 June, Philadelphia, PA, USA.

Cook, D. & Boxall, J. B. 2011 Discolouration material accumulation in water distribution systems. *J. Pipeline Syst. Eng. Pract.* **2**, 113–122.

Douterelo, I., Sharpe, R. & Boxall, J. 2014a Bacterial community dynamics during the early stages of biofilm formation in a chlorinated experimental drinking water distribution system: implications for drinking water discolouration. *J. Appl. Microbiol.* **117** (1), 286–301.

Douterelo, I., Husband, S. & Boxall, J. B. 2014b The bacteriological composition of biomass recovered by flushing an operational drinking water distribution system. *Water Res.* **54**, 100–114.

Furnass, W. F., Collins, R. P., Husband, P. S., Sharpe, R. L., Mounce, S. R. & Boxall, J. B. 2014 Modelling both the continual erosion and regeneration of discolouration material in drinking water distribution systems. *Water Sci. Technol. Water Supply* **14**, 81–90.

Giustolisi, O. & Berardi, L. 2009 Prioritizing pipe replacement: from multiobjective genetic algorithms to operational decision support. *J. Water Resour. Plann. Manage.* **135**, 484–492.

Giustolisi, O. & Savic, D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinform.* **8**, 207–222.

Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modelling using EPRMOGA. *J. Hydroinform.* **11**, 225–236.

Husband, P. S. & Boxall, J. B. 2010 Field studies of discolouration in water distribution systems: model verification and practical implications. *J. Environ. Eng.* **136**, 86–94.

Husband, P. S. & Boxall, J. B. 2011 Asset deterioration and discolouration in water distribution systems. *Water Res.* **45**, 113–124.

Kalteh, A. M., Hjorth, P. & Berndsson, R. 2008 Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environ. Modell. Softw.* **23**, 835–845.

Kohavi, R. 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 20–25 August, Montreal, Quebec, Canada, 2, pp. 1137–1143.

Kohonen, T. 1990 The self-organizing map. *Proceedings of the IEEE* **78**, 1464–1480.

Kohonen, T. 2001 *Self-Organising Maps*, 3rd edn. Springer, Berlin, Germany.

Laucelli, D., Berardi, L. & Doglioni, A. 2005 *Evolutionary Polynomial Regression Toolbox: version 1. SA*. Department of Civil and Environmental Engineering, Technical University of Bari, Bari, Italy. http://www.hydroinformatics.it (accessed December 2014).

Laucelli, D., Rajani, B., Kleiner, Y. & Giustolisi, O. 2014 Study on relationships between climate-related covariates and pipe bursts using evolutionary-based modelling. *J. Hydroinform.* **16** (4), 743–757.

McClymont, K., Keedwell, E., Savic, D. & Randall-Smith, M. 2013 A general multi-objective hyper-heuristic for water distribution network design with discolouration risk. *J. Hydroinform.* **15**, 700–716.

Mounce, S. R., Douterelo, I., Sharpe, R. & Boxall, J. B. 2012 A bio-hydroinformatics application of self-organizing map neural networks for assessing microbial and physico-chemical water quality in distribution systems. In: *Proceedings of 10th International Conference on Hydroinformatics*, 14–18 July, Hamburg, Germany.

Mounce, S. R., Sharpe, R., Speight, V., Holden, B. & Boxall, J. B. 2014 Knowledge discovery from large disparate corporate databases using self-organising maps to help ensure supply of high quality potable water. In: *Proceedings of 11th International Conference on Hydroinformatics*, New York, USA.

Opher, T. & Ostfeld, A. 2011 A coupled model tree (MT) genetic algorithm (GA) scheme for biofouling assessment in pipelines. *Water Res.* **45**, 6277–6288.

Savic, D. A. O., Giustolisi, O. & Laucelli, D. 2009 Asset deterioration analysis using multi-utility data and multi-objective data mining. *J. Hydroinform.* **11**, 211–224.

Schaap, P. & Blokker, E. J. M. 2013 Zooming in on network fouling locations. In: *ASCE World Environmental and Water Resources Congress 2013: Showcasing the Future*, Cincinnati, OH, USA, pp. 1033–1043.

Seth, A., Bachmann, R., Boxall, J. B., Saul, A. J. & Edyvean, R. 2003 Characterisation of materials causing discolouration in potable water systems. *Water Sci. Technol.* **49** (2), 27–32.

Sharpe, R. 2013 Laboratory investigations into processes causing discoloured potable water. PhD thesis, University of Sheffield, UK.

Wu, W., Dandy, G. C. & Maier, H. R. 2014 Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environ. Modell. Softw.* **54**, 108–127.

First received 15 December 2014; accepted in revised form 13 May 2015. Available online 9 July 2015