environments

MDPI

Article

Relating Water Quality and Age in Drinking Water Distribution Systems Using Self-Organising Maps

E.J. Mirjam Blokker [1,*], William R. Furnass [2], John Machell [2], Stephen R. Mounce [2], Peter G. Schaap [3] and Joby B. Boxall [2]

1 KWR Watercycle Research Institute, Groningenhaven 7, 3433 PE Nieuwegein, The Netherlands
2 Department of Civil and Structural Engineering, University of Sheffield, Mappin street, Sheffield s1 3jd, South Yorkshire, UK; wrfurnass1@sheffield.ac.uk (W.R.F.); j.machell@sheffield.ac.uk (J.M.); s.r.mounce@sheffield.ac.uk (S.R.M.); j.b.boxall@sheffield.ac.uk (J.B.B.)
3 PWN Water Supply Company North Holland, Postbus 2113, 1990 AC Velserbroek, The Netherlands; peter.schaap@pwn.nl
* Correspondence: mirjam.blokker@kwrwater.nl; Tel.: +31-6-1586-1099

Academic Editors: Luigi Berardi and Daniele Laucelli
Received: 29 January 2016; Accepted: 13 April 2016; Published: 20 April 2016

Abstract: Understanding and managing water quality in drinking water distribution system is essential for public health and wellbeing, but is challenging due to the number and complexity of interacting physical, chemical and biological processes occurring within vast, deteriorating pipe networks. In this paper we explore the application of Self Organising Map techniques to derive such understanding from international data sets, demonstrating how multivariate, non-linear techniques can be used to identify relationships that are not discernible using univariate and/or linear analysis methods for drinking water quality. The paper reports on how various microbial parameters correlated with modelled water ages and were influenced by water temperatures in three drinking water distribution systems.

Keywords: drinking water distribution; water quality; water age; SOM

1. Introduction

Water quality deteriorates as it travels through a drinking water distribution system (DWDS). Parameters and reactions that are considered of interest include disinfectant residual and disinfectant by-product formation, nitrification, bacterial regrowth, corrosion, sedimentation, temperature, and taste and odour [1]. Many of these are kinetic in nature and hence residence time within a system (or water age) may be an indicator of such deterioration [1]. It has also been shown that higher water temperatures may enhance water quality deterioration [2–6]. Many of the chemical changes that influence water quality are driven by reaction kinetics which are temperature dependent, and temperature also influences microbial populations [7]. Temperature is particularly important because it influences water chemistry, bacteriology and even physical parameters such as water density. Chemical reaction rates and numbers of microorganisms generally increase with increasing temperature, and higher water temperature can dissolve more material thereby increasing conductivity. Conversely, disinfectant residual and dissolved oxygen decrease at higher water temperatures.

It is common practice in many countries to add a disinfection residual as water leaves the water treatment works to protect against regrowth and contamination within the DWDS. This together with relatively cheap, simple, rapid and on-site measurement has often led to the use of free and total chlorine residuals as a general indication of water quality within a distribution network. Changes in free and total chlorine residuals provide a reliable indication of pollution by certain chemical species, such as ammonia, and other polluting species, including microbial contamination. However, not

all water quality parameters are influenced by chlorine, nor is it practicable to maintain a chlorine residual in the entire DWDS without expending significant time and cost, and potential taste and odour complaints. This is especially true in the outskirts of the network [4] where water age, and hence the period for chlorine to react and decay, can be as long as 10 days or more. There are also networks where no chlorine residual is maintained at all, e.g., in the Netherlands, Denmark and Switzerland [8,9].

Water age cannot be measured by physical means; it is an artefact that is an output from mathematical modelling tools. As such, it can be determined without the need for the complexity and uncertainty of reaction kinetics and associated rate coefficients, and therefore has potential pragmatic value as a useful indicator of system specific degradation of water quality. Every DWDS will have unique characteristics influenced by source water quality, treatment process history, the mix of pipe materials, and even the operational regimes that affect system flows, flow directions and storage times. Each system will also have a unique pattern of bulk water volumes, each with a different water age, with the potential to influence water quality in a different manner. There will therefore be no specific "use by date", *i.e.*, no upper water age target that should be applied to all distribution systems.

Common hydraulic network simulations calculate average water age (1) assuming water age at junctions is determined as the flow-weighted mean water age in the joining pipes; and (2) using averaged, repeating demand patterns to find the average water age over the day. The first issue can be resolved by adding water-age bin propagation functionality to a hydraulic network simulator [10]. It was shown that the older (maximum) water age fractions in such simulations can be more important than average water age when trying to interpret bacteriological parameter information [10]. The second issue can be addressed by using a network model with a higher temporal and spatial resolution, e.g., a hydraulic time step of 5 min, and an all-pipes and all-nodes network model with specific, rather than average, water demand patterns [11].

As microbial water quality is not only potentially affected by water age, but also by parameters such as the concentration of biodegradable compounds in the water and pipe material, network layout, hydraulic regime and temperature amongst others, a technique was required to explore the relative influence of several of these factors simultaneously, by means of multivariate analysis. Self-Organising Maps (SOMs) offer such an approach, having been previously applied for determining the parameters that influence the regeneration of discolouration material [2], and the influence of asset and hydraulic modelled parameters, including water age on the physical and chemical quality of drinking water [12].

Data sets from Dutch and UK water companies were enriched with information such as water age and pipe material taken from hydraulic models of the associated networks. The very large Dutch data set represented ten years of regulatory water quality sample results. Average water age for this data set was calculated using a hydraulic model with typical 2012 daily demand patterns. The UK dataset was much smaller, but a more accurate hydraulic model was used and intensive water quality sampling was targeted at areas of the network containing specific water age volumes [7,13].

In this paper we report work undertaken to determine how microbial parameters in a supply zone are influenced by water temperatures or correlate with modelled water ages. The research explores if multivariate, non-linear techniques, such as SOMs, can be used to identify relationships that are not discernible using univariate or linear analysis methods.

2. Methods and Materials

2.1. Hypotheses on Factors Influencing Microbial Water Quality Data

Both the Dutch and UK datasets contain several microbial parameters. Based on the literature, some hypotheses were formed on correlation between these parameters and characteristics of the DWDSs.

Despite known limitations (primarily the selective measure of organisms culturable at relatively high temperatures on agar media, representing less than 0.1% of the community possible within DWDS), heterotrophic plate counts (HPC) remain a widely used measure in practice. According to the

World Health Organisation, the broadest application of HPC within the water industry is when analysing samples taken in response to illness complaints, and complaints of taste and odour. HPC bacteria are not directly related to ill health, but unexpected changes in HPC populations can be indicative of the expanding of biofilms, or the result of an unplanned event within the distribution network [14]. These in turn can be the source of bacteriologically generated organoleptic compounds [15]. Plate counts are also used as a water quality indicator for the commissioning of new or refurbished water mains prior to being put into, or returned to, supply. The 22 °C HPC has been used as a general indicator since 1885. The 37 °C HPC was introduced as an indicator of potential faecal contamination. This has now been dropped from EU Directives, but is often used for day to day operational management. Likewise, coliform bacteria have been discarded as a faecal contamination indicator, but are still put to use as indicators of general bacteriological water quality and are regularly used especially for monitoring water treatment [16]. HPCs are not related to any specific effects on health; they are only indicative of the general water quality in distribution networks. However, 2, 3, 5 and 7-day HPCs can be used as an indicator of unexpected "change" in general water quality, that may trigger more detailed investigations [17]. According to Sartory [18]:

> *"The number of bacteria enumerated at 20–22 °C provides some indication of (1) the amount of food substance available for bacterial nutrition and (2) the amount of soil, dust and other extraneous material that had gained access to the water. The HPC at 37 °C affords more information relating to potentially dangerous pollution, as the organisms developing at this temperature are chiefly those of soil, sewage, or intestinal origin. Their number, therefore, may be used as an index of the degree of purity of the water" and "The ratio of the HPC at 22 °C to that at 37 °C is helpful in explaining sudden fluctuations, high ratios being associated with bacteria of clean soil or water saprophyte origin and, therefore, of 'small significance'".*

Some of the factors influencing HPCs include temperature, residence time within and through the network (water age), disinfection regime and residual disinfectant, organic molecule food sources, flow velocity and patterns, and sediments [19–22]. HPC only determines a very low percentage (<1% to 10%) of the total bacteria. Since this can change between samples, it has no relation with total bacterial cells or total bacterial biomass in DWDS. From studies in the USA, it was concluded that coliform regrowth was significantly reduced in chlorinated supplies at AOC (Assimilable Organic Carbon) values below 50–100 µg carbon/litre [23,24], indicating that AOC is much more influential than water age or temperature [25]. Some correlation was found between some types of HPC numbers in the DWDS and temperature and distance from the water treatment works (WTW) [26]. Even though the meaning of HPC is not very clear, it is widely used and much data is available.

According to the Drinking Water Decree [27] samples from the DWDS have to be tested on various microbial parameters. Next to HPC, these are spores of sulphite reducing *Clostridia* and *Aeromonas* and *E. coli.* The spores of sulphite reducing *Clostridia* are an indicator of effectiveness of the treatment process. A spore is the rest phase of bacteria and thus they do not grow; they are very resistant to environmental influences such as heat and disinfection. *Aeromonas* bacteria are able to replicate and thrive in the presence of easily degradable compounds [28]. Significant correlations were found between the occurrence of *Aeromonas* and the amount of AOC in the drinking water. It is assumed that *Aeromonas* multiplies in the sediment and the biofilm on the pipe wall. It is expected that increased numbers of *Aeromonas* in the distribution system correlate with high water age [5] and high temperature [29]. *E. coli* are indicators of faecal contamination and do not grow in the DWDS, so should not correlate with water age. Table 1 lists the key microbial parameters of interest available in the study data sets, and how they may be expected to be influenced by temperature or correlated with water age.

Table 1. Microbial parameters and influences.

Indicator	Description	Expected Correlation with Water Age	Expected Influence of Temperature	Dataset
Aeromonas (at 30 °C)	Grows in the sediments and biofilm	No direct correlation with water age, but may be indirectly through exchange with sediments and biofilm	Positive influence of temperature	Dutch
Spores of sulphite reducing *Clostridia*	Used as a process indicator for the treatment works in removal of pathogens No growth	No correlation with water age	No influence from temperature	Dutch
E. coli	Faecal indicator, does not grow in the DWDS	No correlation with water age	No influence from temperature	Dutch, UK
HPC at 37 °C (2 days)	Indicator of bacteria capable to grow at human body temperature	No growth in the DWDS, no correlation with water age	Optimum growth at high temperature, limited influence of temperature in DWDS	UK
HPC at 22 °C (3 days)	Indicator for bacteria in water that grow at 22 °C	Fast grower, possible correlation with water age	Could grow in the water, possible influence of temperature	Dutch, UK
HPC at 30 °C (5 days)		No growth in the DWDS, no correlation with water age	Optimum growth at high temperature, limited influence of temperature in DWDS	UK
HPC at 22 °C (7 days)	Indicator for bacteria in water that grow at 22 °C	Slow grower, limited correlation with water age	Could grow in the water, possible influence of temperature	UK

2.2. Dutch Dataset

Water company PWN provided more than ten years of analytical results from regulatory samples. Depending on the water quality parameter, this means 1000 to 25,000 entries (Table 2). This data was filtered to remove spurious examples such as those samples taken from locations that (1) were on pipes with an installation date newer than the date the sample was taken; and (2) had address and postal code that did not match. Something to consider is that, over time, methods of lab analyses or the recording of analytical results may have changed. For example, the threshold value moved, or the way of reporting values below the threshold altered ("0" or "<threshold"). The data was coupled by measurement location and measurement day. In this dataset not all data points have values for all parameters. Temperature was not captured in the database between 2004 and 2010. The ^{2}log was used on all microbial parameters to account for microbial growth typically being an exponential function of time.

Table 2. Dutch water quality data.

Title	Unit	Number of Data Points	First Date	Last Date
Temperature	°C	25,098	29 January 1997	2 March 2012
Turbidity	FTU	14,587	5 January 2004	30 September 2011
Aeromonas	#cfu/100 mL	14,577	3 July 1997	2 March 2012
E. coli	#cfu/100 mL	32,185	5 January 2004	30 September 2011
HPC	#cfu/mL	15,067	5 January 2004	30 September 2011
Fe	mg/L	15,440	6 April 2004	26 March 2013
Spores of *Clostridia*	#cfu/100 mL	6593	6 April 2004	30 September 2011

PWN provided the average modelled water age at each measurement location in the northern part of their network. The network is supplied from two WTWs: WTW Bergen supplies the western part and WTW Andijk supplies the eastern part, both supply surface water but different water, and

with different treatment schemes. The network contains 61,254 nodes for which the water age was calculated; 4886 nodes corresponded to locations where samples were taken. The sampling locations were matched to node names in the hydraulic model using postal codes. A Synergi Water® hydraulic model (DNV GL AS: Mechanicsburg, PA, USA) was used, with the network and demand settings of an average day in 2012. It was run for 696 h (29 days); the average water age of the last 24 h (simulation hours 673–696) was then determined. Any water age above 672 was treated as suspicious, since apparently the modelled water age was still being initialised. The percentage of cast iron pipes around the sampling location (postal code area) was determined from information in the hydraulic model files.

The enriched dataset thus contained less data points (Table 3). Water age and the percentage of cast iron water main were determined for a little over 23,000 data points, but these specific parameters do not change over time. Thus, there are only 4886 unique values (per measurement location). As *E. coli* and spores of Clostridia do not grow in the DWDS (Table 1), and there are only very few (<10%) strictly positive, non-null data points in the dataset (Table 3), these parameters were not used in the SOM analysis. The turbidity and iron data are very skewed (Figure 1), so these may not be very discriminating in the data analysis. There are no location plus time combinations where all parameters are available: Each record contains at least one null (not a number, NaN) value.

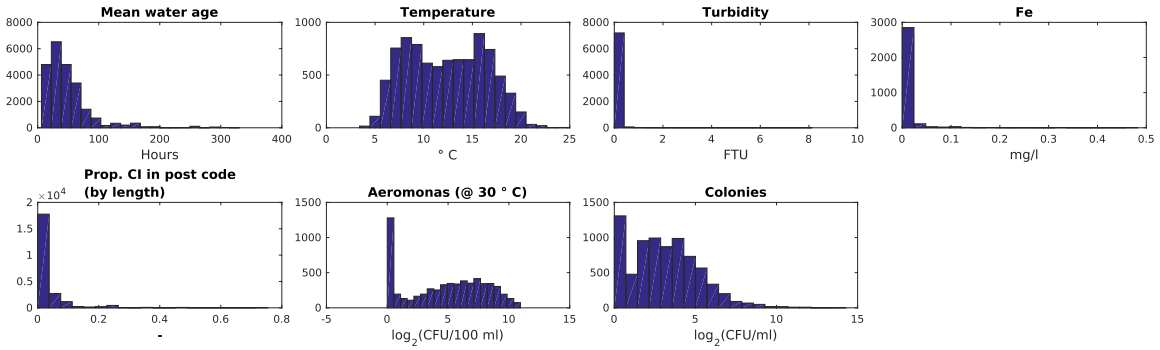

Figure 1. Data distribution of Dutch dataset.

Table 3. Dutch dataset, after adding water age and percentage with Cast Iron pipes for PWN North.

Title	Unit	Not NaN		Non-NaN > 0		Min	Max
Temperature	°C	8763	(37.9%)	8763	(100.0%)	3.4	24.8
Turbidity	FTU	7286	(31.6%)	6592	(90.5%)	0.0	8.1
Aeromonas	^{2}Log (#cfu)/100 mL	6156	(26.7%)	4877	(79.2%)	0.0	11.0
E. coli	^{2}LOG (#CFU)/100 mL	12,855	(55.7%)	176	(1.4%)	0.0	9.7
HPC	^{2}Log(#cfu)/mL	7689	(33.3%)	6381	(83.0%)	0.0	14.3
Fe	mg/L	3076	(13.3%)	2769	(90.0%)	0.0	0.5
Spores of *Clostridia*	#cfu/100 mL	3856	(16.7%)	25	(0.6%)	0.0	2.6
Water age	h	23,092	(100.0%)	23,092	(100%)	6.1	330.5
Percentage of Cast Iron	%	22,955	(99.4%)	17,044	(74.2%)	0.0	75.0

When plotted as bi-variate scatter diagrams, there are no obvious correlations between any of the parameters. When the data is sorted, e.g., by temperature, and divided into groups of equal size, some parameters show a different frequency distribution for each group. Figure 2 shows that at higher temperatures there tend to be more *Aeromonas* and a higher HPC; but above 15–17 °C this relation ceases to exist. Additionally, it shows that in the groups with the lowest water ages (roughly below 24 h), the lowest *Aeromonas* and HPC are found. It seems to suggest that where there are higher percentages of cast iron in the network, less *Aeromonas* are found. There also seems to be a positive correlation between *Aeromonas* and HPC. Figure 2 shows the combination of parameters where the groups exhibit different frequency distributions; the other combinations had coinciding frequency distributions. This also means that water age does not strongly correlate with the water temperature

and the two parameters can be treated as independent. When taking into account the water treatment works (WTW Andijk or Bergen), some of the correlations were even stronger in one area, and almost none existent in another. This was especially true for ATP *versus* temperature and ATP *versus* water age where there was a clear correlation for the water from WTW Andijk and none for the water from WTW Bergen. While these observations can be tentatively made or extrapolated from the bi-variate analysis they are not well evidenced. To explore how water age, temperature and percentage cast iron may in combination correlate with *Aeromonas* and/or HPC the SOM analysis was undertaken.

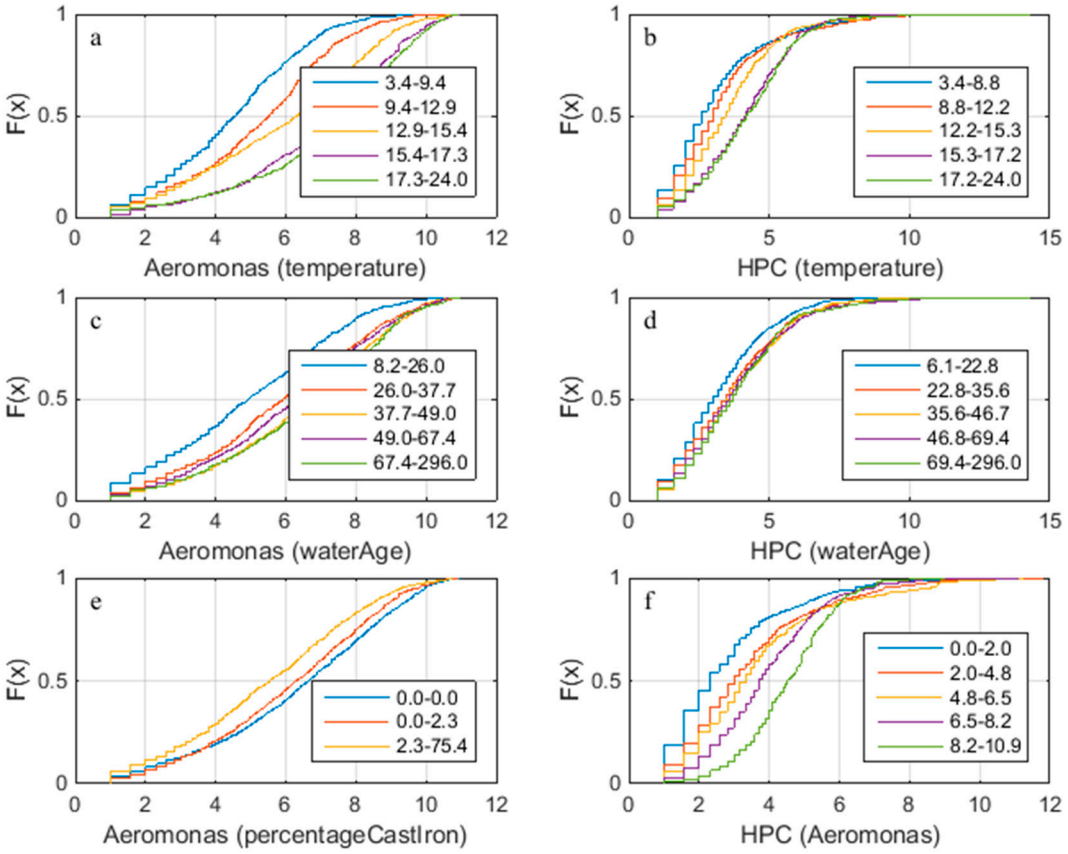

Figure 2. Cumulative frequency distribution for sorted groups (**a**) ^{2}log(*Aeromonas*) sorted by temperature (5 groups, n = 978); (**b**) ^{2}log(*Aeromonas*) sorted by water age (5 groups, n = 976); (**c**) ^{2}log(*Aeromonas*) sorted by percentage Cast Iron (3 groups, n = 1616); (**d**) ^{2}log(HPC) sorted by temperature (5 groups, n = 1279); (**e**) ^{2}log(HPC) sorted by water age (5 groups, n = 1277); (**f**) ^{2}log(HPC) sorted by *Aeromonas* (5 groups, n = 585).

2.3. UK Datasets

The UK "customer taps" study area (13 sampling locations: 11 at customers' taps, 2 at service reservoirs) is described in Machell and Boxall [13]. The UK "small looped network" study area (5 sampling locations in a compact geographical area) is described in Machell and Boxall [7] and Sekar *et al.* [30]. With earlier approaches it was difficult to find clear relations between water quality data and mean or maximum water age [13]. For the SOM analysis, some data processing steps were taken. Non-detect values were replaced with half the detection threshold value, the ^{2}log was used on all microbial parameters, and outliers were not removed. Parameters that had only zero values were removed (Faecal *streptococci*, *Clostridium perfringens*, Total *coliforms*, *Escherichia coli*). When the SOMs showed microbial parameters where less than 10% of values were non-zero, these parameters were discarded. Not all parameters were expected to hold information as, while intensive sampling was undertaken, it was within a small geographical area, and over a short time

period when pH, conductivity, turbidity, temperature and chlorine levels show limited variability (Table 4, Figures 3 and 4).

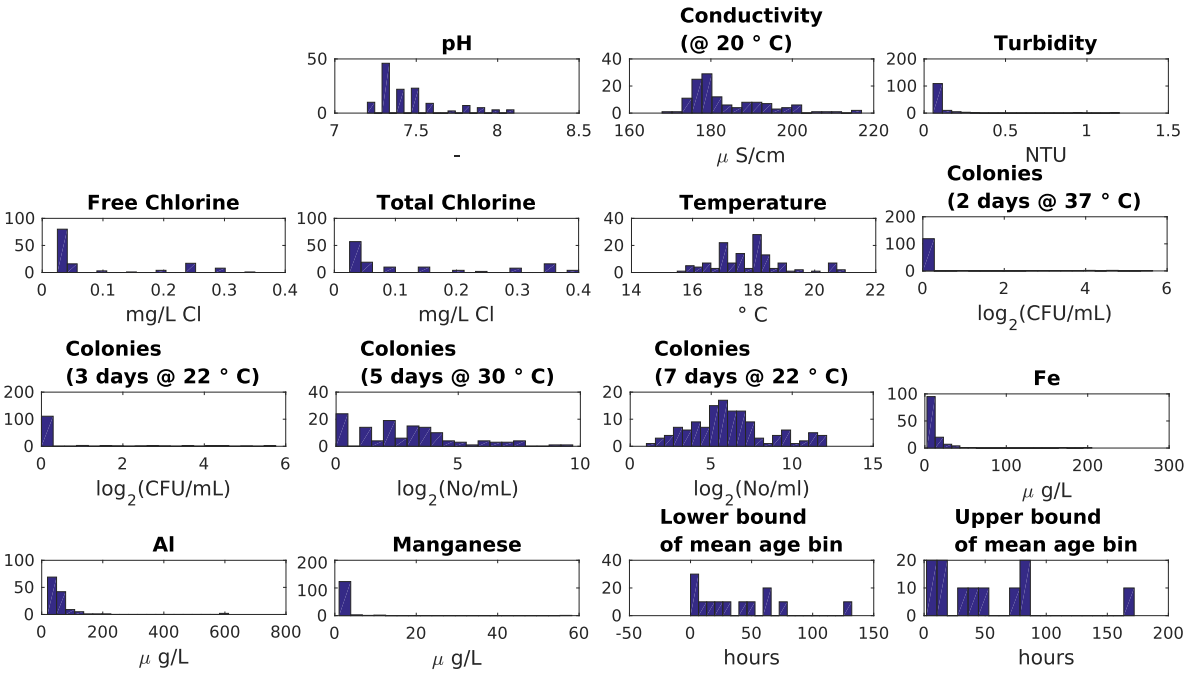

Figure 3. UK customer taps dataset data distribution.

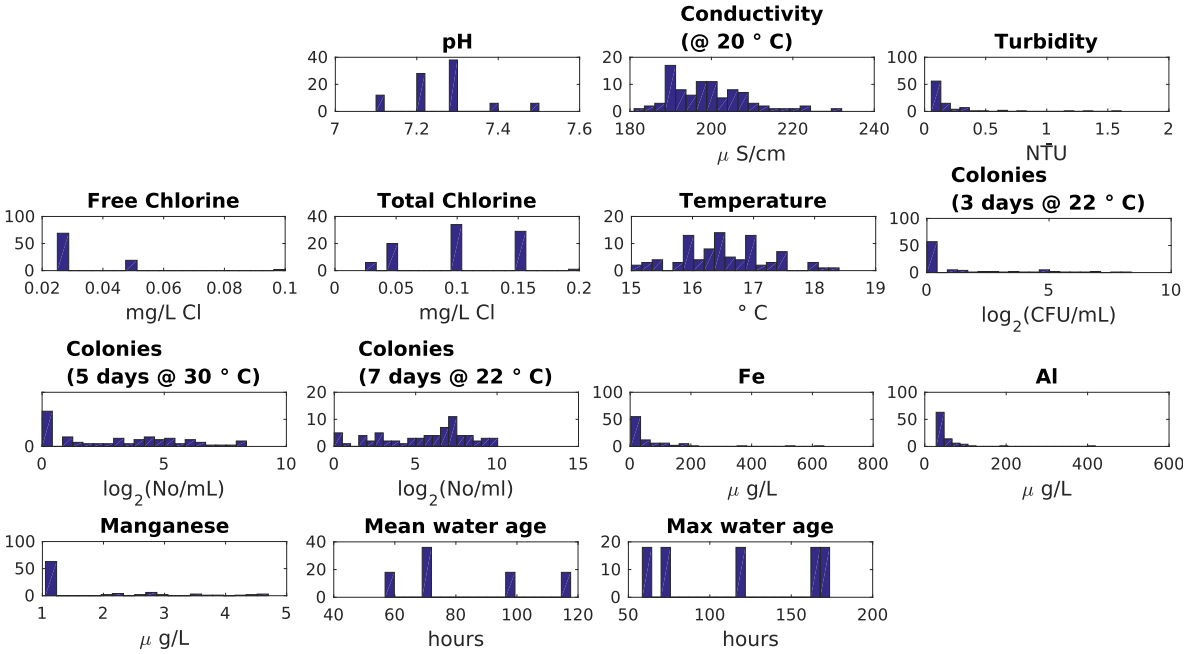

Figure 4. UK small looped network dataset data distribution.

Table 4. Summary for the UK datasets.

Parameter	Customer Taps	Small Looped Network	Of Interest? (After Preliminary Analysis)
pH	7.2–8.1	7.1–7.5	No, small variation
Conductivity	170–215	180–220	No, small variation (<10%)
Turbidity	<0.4	<1.5	Not likely, values are very low
free and total chlorine	0.05–0.4	<0.15	Not for "small looped network": values are too low to have any meaning
Temperature	16–21	15–18	No, small variation (for biological processes)
HPC at 37 °C for 2 days			Yes, if positive samples available
HPC at 22 °C for 3 days			Yes, if positive samples available
HPC at 30 °C for 5 days			Yes, if positive samples available
HPC at 22 °C for 7 days			Yes, if positive samples available
Fe	<50	<200	Yes
Al	<200	<100	Yes
Mn	<10	<5	Yes
water age lower and upper bound	0–170 h	60–170 h	Yes

2.4. SOM Analysis

Clustering aims to discover structure in complex data and is useful when natural groupings based on common properties are suspected. The technique used for this data mining and clustering analysis was the Kohonen Self-Organising Map (SOM), a class of unsupervised Artificial Neural Network (ANN) that performs a dimensionality reduction of the feature space to yield topologically-ordered maps. All variants of the SOM training process generate a many-to-one mapping between n input data records and k *map units*, thus they *discretize/cluster* the input data. The map units are arranged in a 2D lattice and each is associated with a *reference vector* (a weight vector). The properties and values of the resulting map can be best appreciated with some knowledge of the (unsupervised) training process.

Firstly, an appropriate number and arrangement of map units (lattice size/shape) is chosen. Each map unit has an associated reference vector (set of weights) that has the same cardinality, d, as each of the input records used to train/construct the SOM. These weights are then initialised using one of several methods (e.g., randomly or using the Principle Component Analysis (PCA) of the (normalised) training data). The classical SOM algorithm then processes each (normalised) input record in turn. A scalar measure of distance between the record and each reference vector is calculated (typically Euclidean distance), then (a) the most similar reference vector (Best Matching Unit (BMU)) is made more similar to the input record, as are (b) the reference vectors of neighbouring map units in the lattice.

An attractive property of SOMs is that the presence of missing (null) values in the input data do not preclude the training process. Classical SOM training ignores vector dimensions corresponding to null input values when calculating all distance metrics. Alternatively, when updating reference vectors during the iterative process, null values can be substituted for expectations by utilising an Imputation SOM (used herein) as described in Vatanen *et al.* [31].

The result of training is a lattice of map units, each of which is associated with a reference vector and zero or more input records. A particular advantage of the approach is the ability to present the data in a human-readable way that provides for easy visual synthesis and interpretation of multi-dimensional and complicated data sets. The SOM can be visualised as a set of *component planes*: a lattice of tessellating shapes (typically hexagons) is presented for each input dimension, with the colours of hexagons in the i^{th} lattice indicating the magnitude of dimension i of the corresponding map units' reference vectors. Training results in reference vectors being more similar to their neighbours than more distant map units within the lattice (*i.e.*, provides multi-dimensional clustering). Therefore, by comparing component planes for different input variables one can identify general (potentially non-linear) correlations, or correlations that are specific to just a portion of the input data. SOMs can also be interpreted using a *U-matrix*, which is a visualisation of the dissimilarity between the reference

vectors of all map units in the lattice. Ridges of high values in the U-Matrix indicate boundaries between clusters in the input data. The map of the number of hits shows the number of input vectors associated with each hexagon in the component planes (after training) and thus shows where the map unit values (shading) are most supported by data and where values are the result of the influence of neighbouring units. If all the input records are clustered around the edges of the lattice then this might suggest that the SOM is not a good representation of the data (and there are possibly too many input variables or the input variables do not explain the variance in the data). Each input vector can contain NaNs (and these missing values are infilled or ignored during SOM training) but cannot consist solely of NaN values. Some of the hexagons are empty, these only contain information based on interpolation between surrounding hexagons.

A review of the use of SOMs in the water resources domain is presented in Kalteh *et al.* [32] and a comparison against similar approaches appears in Astel *et al.* [33]. Mounce *et al.* [34] proposed their use in data mining microbiological and physio-chemical analytical results data from a laboratory scale pipe rig. SOM analysis of the growth phase revealed that strong clustering of PCA reduced T-RFLP profiles existed in terms of similarity of microorganisms over time. An initial study for their use in clustering water quality, hydraulic model and asset data appears in Mounce *et al.* [12]. A key finding from the analysis is that the risk for water quality stagnation appears to be associated with high water age. SOMs have also been used to explore the factors contributing to higher material accumulation rates in water distribution pipeline systems [2].

All SOMs were generated using the Imputation SOM algorithm [31] and the SOM Toolbox v2.1 for MATLAB [35], used with MATLAB 2015b (Mathworks: Natick, MA, USA). Info on the heuristics the MATLAB SOM Toolbox uses to determine the number of map units and aspect ratio of the map are provided by the authors of the toolbox [35].

3. Results

Figure 5 shows the number of data per hexagon in the component planes for all data sets.

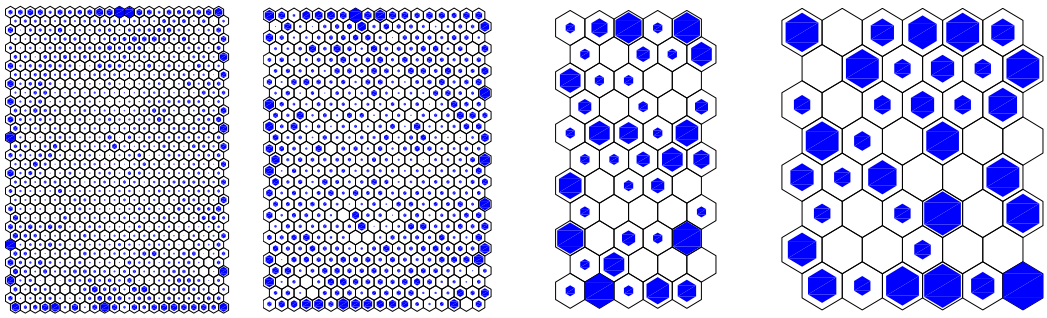

Figure 5. Map of number of hits (**a**) Dutch full dataset; (**b**) Dutch WTW Andijk dataset; (**c**) UK customer taps dataset; (**d**) UK small looped network dataset.

3.1. Dutch Dataset

Figure 6 shows the component planes from the SOM analysis, Figure 5a shows the number of data per hexagon in the component planes. The component planes demonstrate the following:

- Water age and temperature are not correlated; this is as expected [3]
- Total iron and turbidity partly correlate; the highest turbidity values coincide with raised iron levels, presumably particulate iron, whereas the highest iron levels do not correlate with raised turbidity levels and this may indicate dissolved iron. Note that the values for iron and turbidity are all quite low. There seems to be a moderate correlation between higher iron levels and higher local cast iron percentage in the network.

- HPC and *Aeromonas* correlate with each other and with temperature. There does not seem to be a correlation between *Aeromonas* and water age. Repeating the analysis segregated for the two main water sources (WTW Bergen and WTW Andijk) we see similar results, including a weak correlation between water age and *Aeromonas* for WTW Andijk (Figures 7 and 5b).

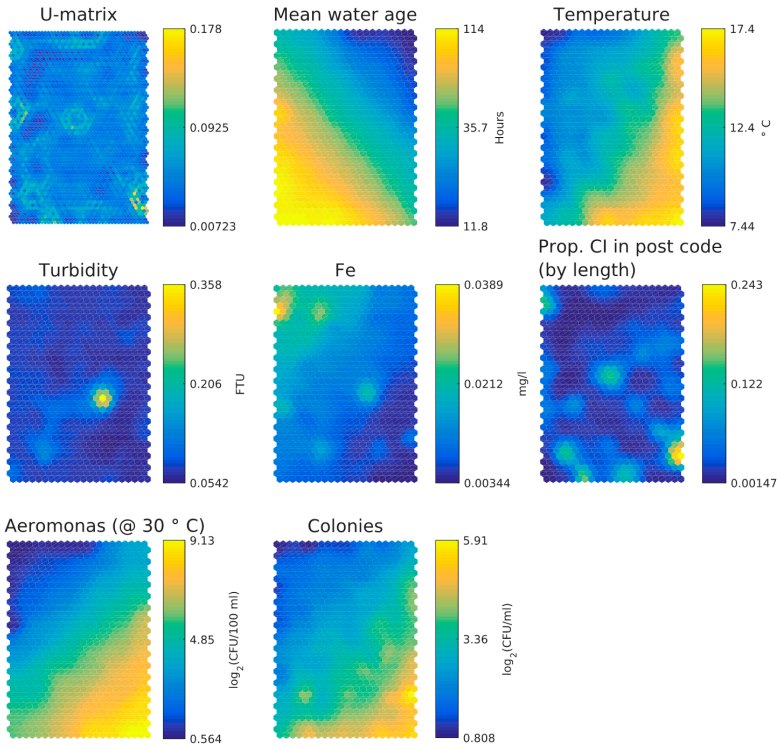

Figure 6. SOM component planes for the Dutch dataset.

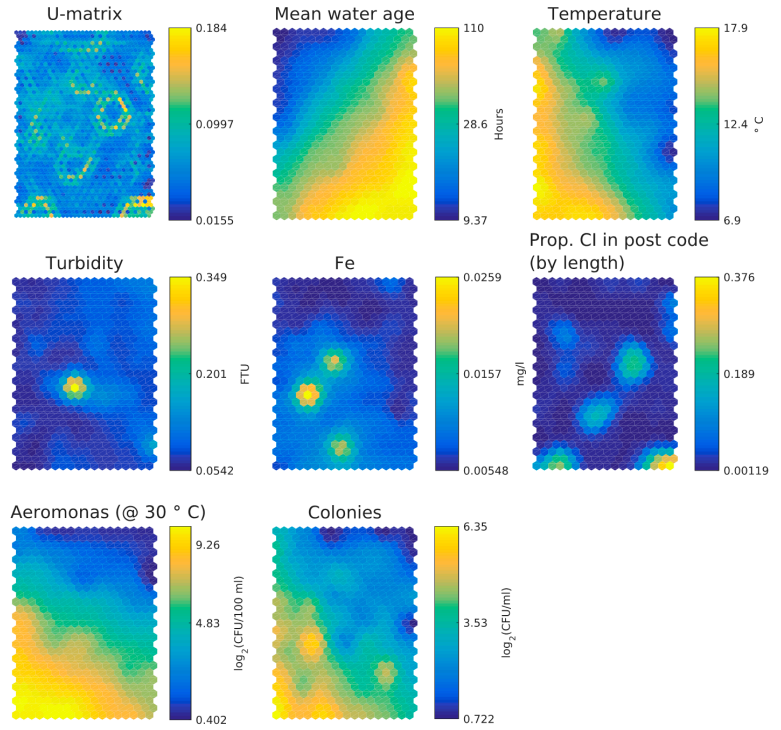

Figure 7. SOM component planes for only WTW Andijk in the Dutch dataset.

3.2. UK Customer Taps Dataset

Figure 8 shows the component planes from the SOM analysis, Figure 5c shows the number of data per hexagon in the component planes. The component planes suggest the following:

- There is little difference between lower and upper bound of mean water age bin in SOM component plane plots.
- Inverse relationship between HPC and chlorine concentrations, strongest for the lower day HPC counts.
- Inverse relationship between chlorine and temperature/age.
- Weak relationship between age and temperature.
- Weak relationship between two days and three days HPC and age/temperature (HPC two days at 37 °C and three days at 22 °C strongly clustered).
- Weak relationship between five days and seven days HPC and age/temperature (HPC five days at 30 °C and seven days at 22 °C strongly clustered).
- Higher values of aluminium in the upper right-hand corner correlate with higher chlorine values and smaller water age values. Higher values of iron, aluminium and turbidity correlate in the lower right-hand corner but higher values of manganese, aluminium and turbidity correlate in the lower left-hand corner. Aluminium coagulant was used historically at the WTW and some of it entered the DWDS and some is found together with iron and manganese, in which case also a higher turbidity is found.

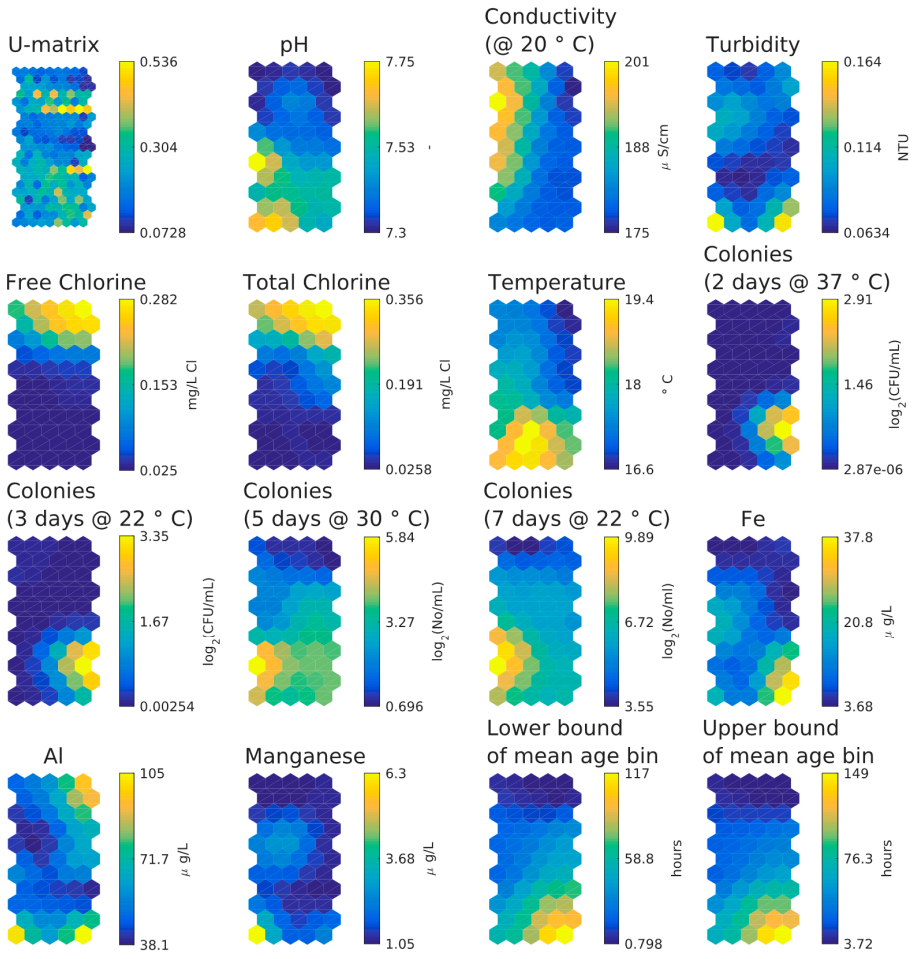

Figure 8. SOM component planes for the customer taps dataset.

3.3. UK Small Looped Network Dataset

Figure 9 shows the component planes from the SOM analysis, Figure 5d shows the number of data per hexagon in the component planes. The component planes suggest the following:

- Little difference between mean and max age bins in SOM component planes.
- Inverse relationship between chlorine concentration and three days HPC and weak inverse relationship between chlorine concentration and five days or seven days HPC.
- Temperature and water age appear to be independent.
- Five-day HPC features same higher values as three-day HPC plus some extra distinct higher values. This second cluster is possibly more representative of biofilm-associated microbial activity than bulk water activity.
- Five-day HPC and seven-day HPC are different.
- Higher values of iron, manganese and turbidity correlate in lower left-hand corner but higher values of manganese and aluminium correlate in upper left-hand corner (aluminium coagulant was used at the WTW and was historically suspected to be entering the DWDS as a result of filter breakthrough).

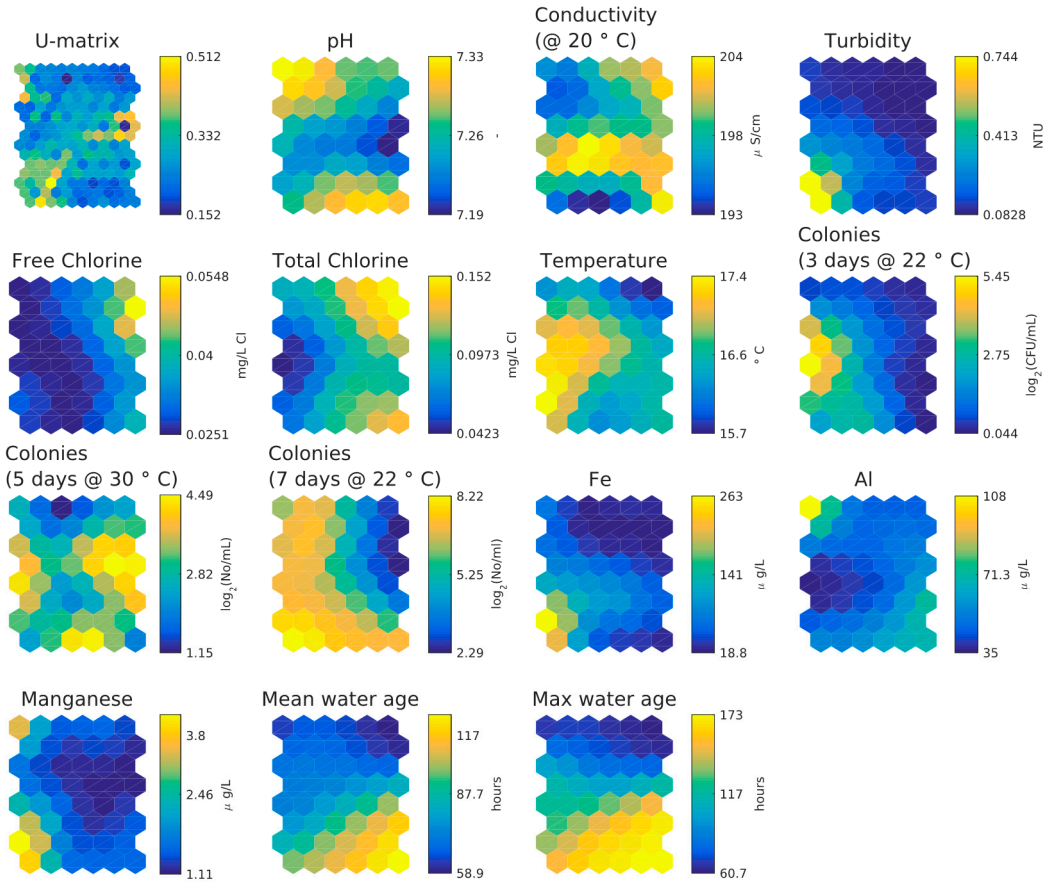

Figure 9. SOM component planes of the small looped network dataset.

4. Discussion

All three datasets showed that water age and temperature could be treated as independent parameters when considering their correlation or influence on the microbial parameters.

The temperature range in the Dutch system was greater that the UK system. Due to the long-term nature of the Dutch data it included both seasonal temperature effects and changes occurring within

the pipe system. Temperature changes in the UK system were likely to be the result of in-system processes only, due to the intensive period of sampling. The low temperature range in UK systems limited the detection of any effects with temperature, indicating that in system temperature changes of this magnitude (16–21 °C and 15–18 °C) were not significant for bacterial water quality. The Dutch study showed that three-day HPC at 22 °C and *Aeromonas* were influenced by temperature, as was expected (Table 5, Havelaar *et al.*, 1990), clearly evidencing that seasonal temperature effects influence these parameters. This influence of temperature on colony counts (*Aeromonas* or HPC) is more likely explained by (more) growth in the DWDS than by more colonies entering the DWDS at higher temperatures. The effect of temperature on colony counts at the WTW is unclear, as usually the micro-organisms do not exceed the detection limit. This suggests that this temperature range (3–25 °C) was significant for in pipe bacterial growth mechanisms.

Table 5. Summary of correlations found between water age, temperature and microbial parameters for Dutch and UK datasets.

Indicator	Expected Correlation with Water Age	Expected Influence from Temperature	Relation with Water Age Found			Relation with Temperature Found	
			Dutch Dataset	Customer Taps	Small Looped Network	Dutch Dataset	UK Datasets
Aeromonas	No direct correlation with water age	Positive influence of temperature	Yes, weak	N.A.	N.A.	Yes	N.A.
HPC at 37 °C (2 days)	No growth in the water, no correlation with water age	Optimum growth at high temperature, limited influence of temperature in DWDS	N.A.	Yes	N.A.	N.A.	N.A.
HPC at 22 °C (3 days)	Fast grower, correlation with water age	Could grow in the water, possible influence of temperature	Yes, weak	Yes	No	Yes	?
HPC at 30 °C (5 days)	No growth in the water, no correlation with water age	Optimum growth at high temperature, limited influence of temperature in DWDS	N.A.	No	No	N.A.	?
HPC at 22 °C (7 days)	Slow grower, limited correlation with water age	Could grow in the water, possible influence of temperature	N.A.	No	No	N.A.	?

"Yes" if there is a correlation, "No" if there is not, "?" if analysis was inconclusive, mainly due to the relatively small temperature range in the UK dataset, and "N.A." if the parameter was not included in the study.

Aeromonas was expected to be correlated with water age; and the Dutch dataset suggested that there was a positive correlation for one of the two WTW. The two-day HPC at 37 °C was not expected to be correlated with water age; but the customer taps dataset suggests that there was a positive correlation between the two. The UK small looped network dataset held too few positive data points so this parameter was not studied. The three-day HPC at 22 °C was expected to be correlated with water age; the customer taps dataset suggest that there was such a positive correlation, but the small looped network dataset and the Dutch dataset suggest that there was not. The five-day HPC at 30 °C and the seven-day HPC at 22 °C were not expected to be directly correlated with water age; the UK datasets suggest that indeed there was no such correlation. *Aeromonas* and the two-day HPC at 37 °C were not expected to directly correlate with water age, as they do not grow in the water phase. However, they may grow in the biofilm and, as the interaction time with the biofilm was influenced by water age, there still may be a correlation with water age on these parameters. Instead of an exponential growth

over time, there may be more of a linear increase. Additionally, as the correlation with water age was not found in all datasets, there may be an influence of other water quality parameters, such as AOC.

The weak correlation between colony counts (*Aeromonas* and HPC) and water age suggests that there is limited potential for growth in the studied DWDS. However, there are some remarks that can be made regarding the mean water ages used in this study. The Dutch dataset was enriched with flow-weighted mean water age, based on an average water demand day in 2012. This meant that the *actual* network demands at the time of taking the water quality samples were not considered. Water age could be quite different depending on the network configuration on the sampling days (which may be different due to e.g., mains repairs), and the time of day. A modelling approach with stochastic demands in the hydraulic network model and an exponentially growing micro-organism in the DWDS showed that the variation in micro-organisms at a node could vary by a factor of ten depending on the time of day, or as a result of variable demands leading to different travel times and potentially different routes through the network [36]. Hence for micro-organisms that grow in the DWDS, it would seem inappropriate to expect correlation of colony counts with the modelled mean water age. Let us assume two equal flows in a hydraulic model, flow A with a water age of 12 h and flow B with a water age of 48 h, the sum being flow C. Under the assumption of exponential growth with a doubling time of 12 h, stream B would have 8 times more micro-organisms than stream A; stream C would contain 4.5 times more micro-organisms than stream A. Based on the average water age stream C would have a water age of 30 h that would as a result contain three times more micro-organisms than stream A, thus underestimating the number of micro-organisms. This would suggest that maximum water age could have a better correlation with the colony counts. However, the SOMs of the UK datasets did not show that maximum water age had a better correlation with microbial parameters than mean water age when considering planktonic phase micro-organisms in bulk water flow. This is contradictory to what was found earlier [10]. In the customer taps dataset the upper bound of mean water age was interpreted as maximum water age, which is not the same as the maximum water age according to the water age bin propagation method. For the small looped network there was significant mixing and Machell and Boxall [7] reported an improved correlation with water quality parameter when the sample points where ordered by maximum rather than mean age, using bi-variate analysis. This is not observed here from the multivariate analysis, perhaps as the parameter is considered as a numerical value here rather than a categorical variable. Another possibility is that the micro-organisms grow in the biofilm and could under certain shear stresses be detached and enter the water. This would more likely be a linear process with contact time (and thus water age). However, the limited correlation between water age and ^{2}log of colony counts in the Dutch dataset does not support this theory. In a DWDS the detachment of biofilm is different in each area of the network, where different water ages are found, depending on local shear stresses, the correlation between water age and colony counts is then unpredictable. Before a DWDS specific upper desired water age can be determined, more understanding is needed on the effect of water age and the process behind this. This could possibly be tested on a trunk main where flows and thus water age in the system could be controlled, provided that the system where long enough to generate suitable extreme ages. The resulting desired upper water age would be system specific as in this test setup incoming water quality, pipe material and diameter, *etc.* would be fixed. Surface area to volume effects of larger *versus* smaller pipes would also have to be carefully considered, along with hydraulic regime driving transport and mixing between the bulk water and boundary layer.

As the Dutch dataset was sparse and the values of the microbial parameters tended to be very variable, it is hence not expected that a more quantitative algorithm will lead to more useful results. Conversely the UK datasets were intensive, and conditions better defined hence qualitative algorithms may be informative but would require extensive effort and further data collection, such as rate coefficients. The SOM analysis presented here provides an efficient method to determine insight into correlations between many parameters and hence gain understanding of the water quality degradation occurring within DWDS and hence how to manage this.

5. Conclusions

Self-organising maps (SOMs) are a useful tool for visual correlation discovery. That is, in inspecting the possible correlations in input data across multiple dimensions, with each component plane being effectively a slice of the multidimensional space. By comparing component planes we can see whether two or more components (dimensions) correlate. This ability to synthesise and present multi-dimensional data (which might otherwise be too complex for humans to interpret) in a higher-fidelity representation is particularly useful for qualitative and intuitive communication with practising engineers. This data-driven approach provides a level of knowledge discovery and evidence/audit trail beyond "engineering judgement".

For the multi-year Dutch dataset and the detailed UK case studies, the SOM showed that water age and temperature may be treated as independent parameters. It also showed that that there is a clear influence of temperature on *Aeromonas* and HPC at 22 °C. The correlation with water age seems less apparent, and there is little added value in considering maximum rather than average modelled water age. Water age as an artefact from mathematical modelling tools is considered an indicator for catch all system specific degradation of water quality, but seems to be of little value as an indicator for specific microbial water quality. This is likely due to a need to improve hydraulic models beyond current best practice of assuming a repeating ideal 24 h average demand pattern and more accurately represent the complex demand driven system dynamics. A water quality model that considers microbial growth over time under specific DWDS circumstances such as local shear stresses, temperature and substrate may also be a much better method for understanding vulnerable locations in DWDS.

Acknowledgments: The authors would like to gratefully thank PWN and Yorkshire Water Services for data provision and permission to publish the details included herein. The research reported in this paper was in part supported by EPSRC platform grant EP/I029346/1.

Author Contributions: E.J. Mirjam Blokker and Peter G. Schaap collected the Dutch data, John Machell collected the UK data, they together interpreted the results, William R. Furnass did the SOM analyses, and E.J. Mirjam Blokker and John Machell wrote the manuscript. Joby B. Boxall and Stephen R. Mounce were responsible for supervising the study, providing key input for the methods and materials used in this research, and revising the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. US Environmental Protection Agency. *Effects of Water Age on Distribution System Water Quality*; EPA: Washington, DC, USA, 2002.
2. Mounce, S.R.; Husband, S.P.; Furnass, W.R.; Boxall, J.B. Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems. *IWA J. Hydroinform.* **2016**, *18*, 96–114. [CrossRef]
3. Blokker, E.J.M.; Pieterse-Quirijns, E.J. Modeling temperature in the drinking water distribution system. *J. Am. Water Works Assoc.* **2013**, *105*, E19–E29. [CrossRef]
4. Blokker, E.J.M.; Vreeburg, J.; Speight, V. Residual chlorine in the extremities of the drinking water distribution system: The influence of stochastic water demands. In Proceedings of the 12th International Conference on Computing and Control for the Water Industry, Perugia, Italy, 20 February 2014.
5. Havelaar, A.; Versteegh, J.; During, M. The presence of Aeromonas in drinking water supplies in the Netherlands. *Int. J. Hyg. Environ. Med.* **1990**, *190*, 236–256.
6. Van der Kooij, D. Legionella in drinking-water suppliies. In *Microbial Growth in Drinking-Water Supplies. Problems, Causes, Control and Research Needs*; van der Kooij, D., van der Wielen, P.W.J.J., Eds.; IWA: London, UK, 2013; pp. 127–175.
7. Machell, J.; Boxall, J. Field studies and modeling exploring mean and maximum water age association to water quality in a drinking water distribution network. *J. Water Res. Plan. Manag.* **2012**, *138*, 624–638. [CrossRef]

8. Medema, G.J.; Smeets, P.W.M.H.; Van Blokker, E.J.M.; Lieverloo, J.H.M. Safe distribution without a disinfectant residual. In *Microbial Growth in Drinking-Water Supplies. Problems, Causes, Control and Research Needs*; van der Kooij, D., van der Wielen, P.W.J.J., Eds.; IWA: London, UK, 2013; pp. 95–125.
9. Smeets, P.; Medema, G.; Van Dijk, J. The Dutch secret: How to provide safe drinking water without chlorine in the Netherlands. *Drink. Water Eng. Sci.* **2009**, *2*, 1–14. [CrossRef]
10. Machell, J.; Smeets, P.W.M.H.; Van Blokker, E.J.M.; Lieverloo, J.H.M. Improved Representation of Water Age in Distribution Networks to Inform. Water Quality. *J. Water Res. Plan. Manag.* **2009**, *135*, 382–391. [CrossRef]
11. Blokker, E.J.M.; Beverloo, H.; Vogelaar, A.J.; Vreeburg, J.H.G.; Van Dijk, J.C. A bottom-up approach of stochastic demand allocation in a hydraulic network model: A sensitivity study of model parameters. *J. Hydroinform.* **2011**, *13*, 714–728. [CrossRef]
12. Mounce, S.R.; Sharpe, R.; Speight, V.; Holden, B.; Boxall, J. Knowledge discovery from Large disparate corporate databases using Self-Organizing Maps to help ensure supply of high quality potable water. In Proceedings of the 11th International Confernece on Hydroinformatics, New York, NY, USA, 17–21 August 2014.
13. Machell, J.; Boxall, J. Modeling and Field Work to Investigate the Relationship between Age and Quality of Tap Water. *J. Water Res. Plan. Manag.* **2014**, *140*, 431–439.
14. World Health Organization. *Guidelines for Drinking-Water Quality: Recommendations*; World Health Organization: Geneva, Switzerland, 2004; Volume 1.
15. Standing Committee of Analysts. The assessment of taste, odour and related aesthetic problems in drinking waters. In *Methods for the Examination of Waters and Associated Materials*; Environmental Agency: London, UK, 1998.
16. Standing Committee of Analysts. The microbiology of drinking water 2002—Part. I—The enumeration of heterotrophic bacteria by pour and spread plate techniques. In *Methods for the Examination of Waters and Associated Materials*; Environmanetal Agency: London, UK, 2002.
17. Bartram, J.; Bartram, J.; Cotruvo, J.; Exner, M.; Fricker, C.; Glasmacher, A.; Bartram, J.; Cotruvo, J.; Exner, M.; Fricker, C.; *et al. Heterotrophic Plate Counts and Drinking-Water Safety: The Significance of HPCs for Water Quality and Human Health*; IWA Publishing: London, UK, 2003.
18. Sartory, D.P. Heterotrophic plate count monitoring of treated drinking water in the UK: A useful operational tool. *Int. J. Food Microbiol.* **2004**, *92*, 297–306. [CrossRef] [PubMed]
19. Liu, G.; Van der Mark, E.J.; Verberk, J.Q.J.C.; Van Dijk, J.C. Flow cytometry total cell counts: A field study assessing microbiological water quality and growth in unchlorinated drinking water distribution systems. *BioMed Res. Int.* **2013**, *2013*. [CrossRef] [PubMed]
20. Geldreich, E.E. *Microbial Quality of Water Supply in Distribution Systems*; CRC Press: Boca Raton, Fl, USA, 1996.
21. LeChevallier, M.W. Biofilms in drinking water distribution systems: Significance and control. In *Identifying Future Drinking Water Contaminants*; National Research Council: Washington, DC, USA, 1999; p. 206.
22. Lehtola, M.J.; Michaela, L.; Miettinen, I.T.; Arja, H.; Terttu, V.; Martikainen, P.J. The effects of changing water flow velocity on the formation of biofilms and water quality in pilot distribution system consisting of copper or polyethylene pipes. *Water Res.* **2006**, *40*, 2151–2160. [CrossRef] [PubMed]
23. LeChevallier, M.W.; Schulz, W.; Lee, R.G. Bacterial nutrients in drinking water. *Appl. Environ. Microbiol.* **1991**, *57*, 857–862. [PubMed]
24. LeChevallier, M.W.; Welch, N.J.; Smith, D.B. Full-scale studies of factors related to coliform regrowth in drinking water. *Appl. Environ. Microbiol.* **1996**, *62*, 2201–2211. [PubMed]
25. Van der Kooij, D. Assimilable organic carbon as an indicator of bacterial regrowth. *J. Am. Water Works Assoc.* **1992**, *84*, 57–65.
26. Power, K.N.; Nagy, L.A. Relationship between bacterial regrowth and some physical and chemical parameters within Sydney's drinking water distribution system. *Water Res.* **1999**, *33*, 741–750. [CrossRef]
27. Drinking Water Decree. 2015. Available online: http://wetten.overheid.nl/BWBR0030111/geldigheidsdatum_27-11-2015 (accessed on 27 November 2015).
28. Van der Kooij, D.; Visser, A.; Hijnen, W. Growth of Aeromonas hydrophila at low concentrations of substrates added to tap water. *Appl. Environ. Microbiol.* **1980**, *39*, 1198–1204. [PubMed]
29. Rouf, M.; Rigney, M.M. Growth temperatures and temperature characteristics of Aeromonas. *Appl. Microbiol.* **1971**, *22*, 503–506. [PubMed]

30. Sekar, R.; Deines, P.; Machell, J.; Osborn, A.M.; Biggs, C.A.; Boxall, J.B. Bacterial water quality and network hydraulic characteristics: A field study of a small, looped water distribution system using culture-independent molecular methods. *J. Appl. Microbiol.* **2012**, *112*, 1220–1234. [CrossRef] [PubMed]
31. Vatanen, T.; Osmala, M.; Raiko, T.; Lagus, K.; Sysi-Aho, M.; Orešič, M.; Honkela, T.; Lähdesmäki, H. Self-organization and missing values in SOM and GTM. *Neurocomputing* **2015**, *147*, 60–70. [CrossRef]
32. Kalteh, A.M.; Hjorth, P.; Berndtsson, R. Review of the Self-Organizing Map (SOM) approach in water resources: Analysis, modelling and application. *Environ. Model. Softw.* **2008**, *23*, 835–845. [CrossRef]
33. Astel, A.; Tsakovski, S.; Barbieri, P.; Simeonov, V. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Res.* **2007**, *41*, 4566–4578. [CrossRef] [PubMed]
34. Mounce, S.; Douterelo, I.; Sharpe, R.; Boxall, J. A bio-hydroinformatics application of self-organizing map neural networks for assessing microbial and physico-chemical water quality in distribution systems. In Proceedings of the 10th International Conference on Hydroinformatics, Hamburg, Germany, 14–18 July 2012.
35. Helsinki University of Technology. SOM Toolbox (for MATLAB) 2014. Available online: https://github.com/ilarinieminen/SOM-Toolbox/ (acccessed on 14 April 2016).
36. Blokker, E.J.M.; Pieterse-Quirijns, E.J.; Vogelaar, A.; Sperber, V. *Bacterial Growth Model in the Drinking Water Distribution System—An Early Warning System*; KWR Watercycle Research Institute: Nieuwegein, The Netherlands, 2014; p. 31.