10th International Conference on Hydroinformatics
HIC 2012, Hamburg, GERMANY

A BIO- HYDROINFORMATICS APPLICATION OF SELF-ORGANIZING MAP NEURAL NETWORKS FOR ASSESSING MICROBIAL AND PHYSICO-CHEMICAL WATER QUALITY IN DISTRIBUTION SYSTEMS

STEPHEN MOUNCE, ISABEL DOUTERELO, REBECCA SHARPE AND JOBY BOXALL
Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, Sheffield S1 3JD, UK.

Water quality in a water distribution system (WDS) is determined by a variety of complex processes, affected directly or indirectly by numerous factors including changing water demand, infrastructure condition and environmental variation. Whilst most of the hydraulic and physico-chemical variables that are relevant to water quality are quite well understood, measures of microbiological processes are less developed and have so far been difficult to use in water quality decision support tools. DNA-based molecular techniques are now being used to analyse environmental samples. Bio- and HydroInformatics can be defined as disciplines that generate computational methods and tools, databases, and methods to support DNA-based and hydraulic related research. This paper demonstrates how Kohonen self-organizing maps (SOM) can be used for integrative data mining of disparate hydraulic, physico-chemical and microbiological data sources from a unique experimental pipe test facility. Results are reported from a four week test period to examine the impact of three separate flow profiles on the accumulation and mobilization of particles. Genetic signatures acquired by terminal restriction fragment length polymorphisms (T-RFLPs) were obtained from samples, and analysed using principal component analysis (PCA). A range of single parameter hydraulic and chemical variables were logged. These datasets were then analysed by SOM networks. Results show that the visual output of the SOM analysis provides a useful tool for identifying novel microbiological relationships.

INTRODUCTION

Customers regard a safe supply of water as one of the most important aspects of the water supply service. Water quality in a WDS is a very complex system, affected directly or indirectly by numerous factors including changes in the source and final water quality, as well as changing demand and climatic variation. Changing hydraulic operations at the utility (e.g. the operation of tanks, pumps, and valves etc.) and failure of infrastructure can all cause a change in water quality.

High quality water leaving treatment facilities generally deteriorates as it travels through extensive, often convoluted, distribution networks, via a number of mechanisms associated with distribution network materials, hydraulic conditions, chemical and biological reactions, or ingress of polluting materials. The presence of biofilms attached to the inner pipe surface is a major concern but not yet well understood. The increase of microorganisms in distribution networks generates a number of problems such as loss of residual chlorine, discolouration, negative changes in water taste and odour, pipe corrosion etc. Traditionally, microbial assessment of drinking water has been based on the study of plankton (microorganisms inhabiting the bulk water). However recent research has observed that the majority of microorganism in the network is actually attached to the inner pipe surface as biofilms. Bacterial communities within biofilm can be analysed to determine their abundance, diversity and to compare communities separated in space and/or time. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in IT have combined to produce large datasets in the area of molecular biology. Bioinformatics is the name given to mathematical and computing approaches used to glean understanding of biological processes and it covers the creation and development of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from analysis of this data.

The Kohonen self-organizing feature map, generally referred to as the Self Organising Map (SOM), is a neural network model which draws inspiration from biological processes. In this paper SOMs are used for examining the relationships between water physico-chemical and microbiological characteristics. The data mining approach was applied to the results from experiments conducted in a test loop facility. The research presented here was a part of the Pipe Dreams Project (http://www.sheffield.ac.uk/pipedreams).

BACKGROUND

T-RFLP profiling

Terminal restriction fragment length polymorphism (T-RFLP) is a fast way of screening and analysing complex microbial communities. T-RFLP is based on the amplification of the 16S rRNA gene with a fluorescent label attached to the end of one or both primers followed by digestion of the PCR product with restriction enzymes [1]. The sizes of the resulting terminal restriction fragments (TRFs) containing the fluorescent label are subsequently determined using an automated fragment length analysis system. The number of peaks and peak area in a T-RFLP profile immediately give insight into the richness and evenness of the population providing a “fingerprint” of the bacterial community.

A recurring problem in T-RFLP analysis is the comparison of profiles. In the past, comparison of multiple T-RFLP profiles to identify shared and unique components of microbial communities has been a manual process which is both time consuming and liable to the introduction of errors. The use of a web based tool, T-Align [2] allows rapid comparison of numerous T-RFLP profiles. T-Align has been applied to T-RFLP profiles obtained in response to discolouration events in an experimental WDS [3]. The comparison matrix produced by T-Align is of uniform size, containing all the consensus profiles compared with all others, with each point containing either a zero (in the absence of a TRF) or the relative percentage fluorescence when the TRF was present. These uniform profiles facilitate the comparison of samples and can be subsequently further analysed using ordination statistics or transformed for further comparison with binary matching. Techniques from the field of Artificial Intelligence, such as Cellular Automata, Genetic

Algorithms and Artificial Neural Networks (ANNs), are now being explored to provide sophisticated data mining in the bioinformatics domain [4].

SOMs and water resources

There are some applications for which the ‘correct’ outputs are unknown. 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 thus allowing it to derive information from data when it is suspected that distinct classes exist in a collection of samples. The SOM is one of the most well-known ANNs employing unsupervised learning having properties of both vector quantization and vector projection algorithms [5]. The prototype vectors are positioned on a regular low-dimensional grid in a spatially ordered fashion hence facilitating improved visualisation.

SOMs have been used for analysis and modelling of water resources as reviewed in [6]. Carstea et al. [7] presented results of real-time fluorescence excitation-emission matrices (EEM) spectroscopy using an in-situ fibre-optic probe installed in a small urban river. SOMs were used to cluster fluorescence EEMs of different character, demonstrating seven distinct clusters. Chang et al. [8] used SOMs for expressing water quality comprehensive evaluation in a water network by high-dimensional water quality indicator projection to a low dimensional topology grid for higher level interpretation. Mustonen et al. [9] generated pressure shocks in a pilot-scale drinking water distribution system to explore water quality changes, with a particular emphasis on particle size. The data collected was analysed with SOM and Sammon’s mapping. The pressure shocks led to detachment of biofilms and soft deposits, and this was observed to increase electrical conductivity, turbidity and the number of particles in drinking water.

MATERIALS AND METHODS

Test loop facility and experimental work

Following the findings of and development after Husband et al. [10] a test loop facility, housed in a temperature controlled room, has been constructed at the University of Sheffield (Figure 1). It consists of three connecting loops which can be individually controlled to represent three different hydraulic regimes. Individual loops consist of nine-and-a-half 21.4m long coils of HPPE pipe, with a total length of ~200m and a combined height of 4m. Unlike bench scale experiments the full scale pipe surface area of the test loop facility enables fully realistic exchange processes and interactions between the bulk fluid and the pipe wall to occur, replicating realistic conditions in a typical WDS including boundary layer hydraulics. This facility has been used to conduct growth and flushing experiments to determine the processes which lead to discoloured potable water [11]. Investigating the microbiological component of the pipe wall material was achieved by fitting the test loop facility with PWG coupons, as described in Deines et al. [12]. Coupons were fitted along the length of each pipe loop to facilitate microbiological studies of the accumulation and mobilisation of material on the pipe wall.

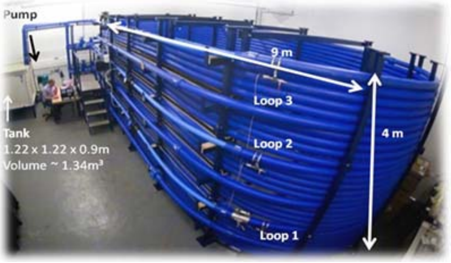

Figure 1. Test loop facility [11]

A 16°C study was conducted to assess microbiological and physico-chemical parameters over a 28 day experimental growth phase, followed by a flushing phase. During the growth phase mains water was circulated through the three loops, each set with a different flow condition. The three loops compared a low varied flow, ranging from 0.2 to 0.8 l/s (loop1), a steady state 0.4 l/s flow (loop 2) and a high varied flow, ranging from 0.2 to 1.2 l/s (loop 3). The varied flow profiles were based on the daily pattern usually observed in WDS [10]. To maintain consistent water quality and hydraulic condition, water was re-circulated throughout the test loops, but with a system trickle drain and refill rate set to ensure a system hydraulic residence time of 24 hours. In the flushing phase, the flow is incrementally increased in the pipes up to 4.5 l/s, allowing for three turnovers of the system for each flush so that the water becomes well mixed. PWG coupons were removed from multiple positions on the pipe wall during the accumulation phase (after 3, 7, 14, 21 and 28 days) and at the beginning and end of flushing. Water chemistry spot samples (chlorine, ORP, temperature, pH, manganese, iron and turbidity) were taken for each loop after 1 turnover of the system (where the peak was expected) at the start and end of the flushing.

Microbiology

Biofilms were removed from PWG coupons as described in [12], filtered through 0.22 μm nitrocellulose membrane filters (Millipore, Corp) and DNA extracted using a standard phenol:chloroform method. To study the planktonic communities within the pipes, 1L of bulk water was filtered as above and DNA extracted from the filter. Bacterial DNA was amplified by PCR with the primer set FAM-63F and 518R, followed by individual digest with the restriction enzyme AluI (Roche Diagnostic). T-RFLPs were separated by capillary electrophoresis using an automater sequencer 3730 (Applied Biosystems). Differences in abundance and length of T-RFLPs were determined by comparison with a known size internal standard (ROX$^{®}$500 size standards, Applied Biosystems) and the actual sizes were estimated by interpolation using a Local Southern algorithm with the software GeneMapper 3.7 (Applied Biosystems). Subsequently, T-Align was applied with confidence interval 0.5.

Data mining with SOM and PCA

In this work T-RFLP profiles were utilised both in isolation and in conjunction with physico-chemical data from test loop experiments. The SOM was generated using the program MATLAB (Version 7.2.0.635; The Mathworks Inc.) using the SOM toolbox developed at the Helsinki University of technology (available on line at http://www.cis.hut.fi/projects/somtoolbox) [13]. T-RFLP T-Aligned profiles can be reduced in dimensionality by Principal Components Analysis (PCA). Microbiological data were normalised as part of the T-Align processing and linear scaling was also conducted on the PCA variables so that the variance of each was one. The network parameters were selected on the basis of trial runs and default suggested values in the SOM toolbox. The input layer consisted of a number of neurons corresponding to PCA components (and, where relevant, physico-chemical parameters) and the output layer consisted of a hexagonal Kohonen map whose size was optimally selected by the SOM toolbox. Figure 2 shows an example SOM for the case study for the flushing phase.

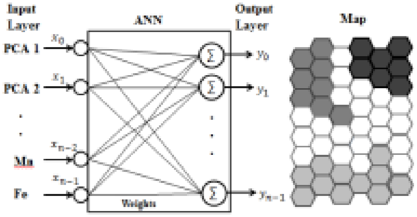

Figure 2. Self Organising Map structure for test loop flushing data

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 – 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. A trained network can be labelled in a manner described in [5].

RESULTS

Firstly, the T-RFLP data was examined independently from other parameters in the accumulation (material growth) phase to assess microbiological similarity over time only. After T-Align was applied, the complete input data set for the growth phase consisted of 51 profiles with 329 potential peaks representing the relative abundance of each type of bacteria. PCA was applied to this full T-RFLP data setusing the *princomp* function in MATLAB. The first five principal components account for 68.1% of the total variability of the data set (with the first two components describing only 46.6 %).

This dataset was then presented to a SOM in order to examine temporal variation in the growth phase as shown in figure 3a. The U-matrix allows examination of the overall cluster patterns in the input data set after the model has been trained. In the component planes for individual variables, the colouring corresponds to actual numerical values for the input variables that are referenced in the scale bars adjacent to each plot. The first two strongest principal components are shown along with a labelled map with sampling days as categories. The map shows, in general, clustering based on the date of the sample i.e. the microbial similarity over the growth phase. Once a SOM map is trained it can also be used for classification via the Best Matching Unit (BMU). This feature is not generally available in conventional clustering analysis. In order to test the labelled maps, a leave-one-out bootstrap approach (used in ANN testing when a dataset size is limited) was carried out so that all testing was conducted unseen. For two principal components, the accuracy was 41.2% and to within the nearest sampling date 74.5%.

In the final flushing phase, microbiology samples were taken as previously described from both coupons (i.e. pipe wall biofilm) and flushed water. The T-RFLP data was examined independently to explore micro-organism difference in these two mediums. After T-Align was applied, the complete input data set consisted of 36 profiles with 393 peaks. The first two principal components account for 67.1% of the total variability of the data set. Experimentation revealed that only two PCAs were needed to describe the data clustering around sample type successfully. Figure 3b shows the SOM with two PCA component planes and a labelled map with type of sample as category. Clearly, there is significant microbiological diversion between the two categories. Using the bootstrap testing approach, the SOM had 100% accuracy classifying an unseen profile as either the biofilm or water.

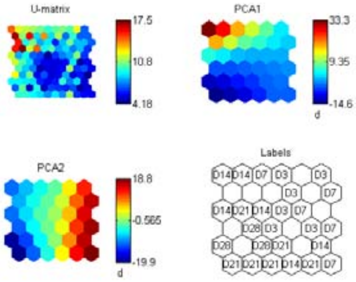

Figure 3a. SOM for growth phase data

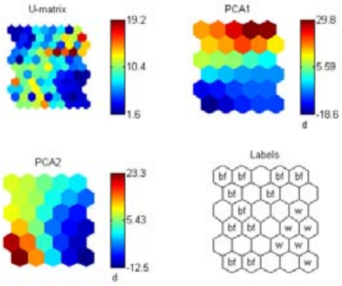

Figure 3b. SOM for flushing phase (type)

Finally, the start (0.4 l/s) and post flushing along with individual loop conditions were analysed by bringing together the microbiological data and measured physico-chemical parameters. Five variables describing T-RFLP patterns were used to represent community structure (largest PCA components) leading to 12 component planes as shown in Figure 4.

DISCUSSION

The value of the SOM analysis is in observing interrelationships that exist between the various variables and potentially providing a basis for generating hypotheses that can be subsequently examined experimentally. In Figure 4 it can be observed that by integrating the data sources the labeled maps show distinctive regions of clusters when comparing variables from pre/post flushing and between loops. For example, the temperature is clearly lower in loop 1, chlorine is lower in loop 3 and a low value of PCA3 seems to relate to pre-flushing conditions in loop 1. Some common patterns exist between several variables; for example, turbidity, Fe and Mn all have higher values after flushing (as should be expected) hence corroborating the key role of these metals in the process of water discolouration [10].

The SOM does not replace existing statistical tools, particularly those in the bioinformatics domain, but enables the visual examination of relationships between disparate types of variables. Future work is planned to explore the development of this approach further, first by conducting additional test loop studies with more significant water source variation between loops; for example, a higher iron content in one of the loops. Secondly, field trials in a real distribution system enabling microbiology and physico-chemical sampling of multiple sites will allow further evaluation and data mining.

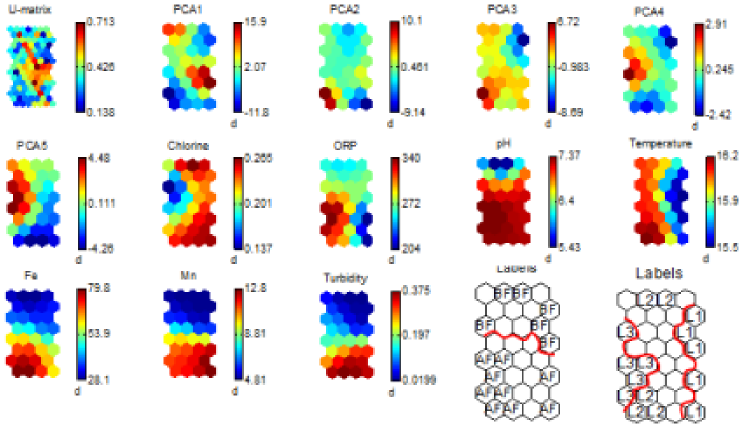

Figure 4. SOM for integrated data with labelling for before/after flushing and loop number

CONCLUSIONS

In this paper the bacterial communities grown in a test loop facility were analysed for a 28 day accumulation (growth) period and a final day flushing phase. 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 (day of sample taken). T-RFLPs of samples from the coupon wall (biofilm) versus from the water revealed divergent typical profiles, and using the SOM as a classifier on unseen data achieved 100% accuracy. Finally, by combining the T-RFLPs with physico-chemical sampled data in the flushing phase, the interrelationships between variables for differing conditions was explored.

The results show that the visual output of the SOM provides a rapid and intuitive means of examining covariance between variables and exploring hypotheses for increased understanding. A particular advantage of the approach is the ability to present data in a visual way that provides easy interpretation of multi-dimensional and complicated data sets.

ACKNOWLEDGEMENT

This work is part of the Pipe Dreams project supported by the U.K. Engineering and Physical Sciences Research Council (Challenging Engineering: EP/G029946/1).

REFERENCES

[1] Avaniss-Aghajani, E., Jones, K., Chapman, D. and Brunk, C., “A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences”, *Biotechniques* Vol. 17, (1994), pp 144-149.

[2] Smith, C. J., Danilowicz, B. S., Clear, A. K., Costello, F. J., Wilson, B. and Meijer, W. G., “T-Align, a web-based tool for comparison of multiple terminal restriction fragment length polymorphism profiles”, *FEMS Microbiology Ecology*, Vol. 54, (2005), pp 375-380.

[3] Smith, C.J., Sharpe, R.L., Boxall, J.B., ”Investigating bacterial community response during discolouration events in an experimental water distribution system”, *Proceedings of the 13th International Symposium for Microbial Ecology*, Seattle, USA, August, (2010).

[4] Keedwell E. and Narayanan, A., *“Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems”*, John Wiley, (2005).

[5] Kohonen, T., “The Self-Organizing Map”, *Proceedings of the IEEE*, Vol. 78, No. 9, (1990), pp 1464-1480.

[6] Kalteh, A. M., Hjorth, P. and Berndsson, R., “Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application”, *Environmental Modelling and Software*, Vol. 23, (2008), pp 835-845.

[7] Carstea, E. M., Baker, A., Bieroza, M. and Reynolds, D., “Continuous fluorescence excitation-emission matrix monitoring of river organic matter”, *Water Research*, Vol. 44, (2010), pp 5356-5366.

[8] Chang, K., Gao, J. L., Wu, W. and Yuan, Y. X., “Water quality comprehensive evaluation method for large water distribution network based on clustering analysis”, *Journal of HydroInformatics*, Vol. 13, No. 3, (2011), pp 390-400.

[9] Mustonen S. M., Tissari S., Huikko L., Kolehmainen M., Lehtola M. J. and Hirvonen A., “Evaluating online data of water quality changes in a pilot drinking water distribution system with multivariate data exploration methods”, *Water Research*, Vol. 42, No. 10-11, (2008), pp 2421-2430.

[10] Husband, S., Boxall, J. B., Saul, A. J., “Laboratory studies investigating the processes leading to discolouration in water distribution networks”, *Water Research*, Vol. 42, (2008), pp 4309-4318.

[11] Sharpe, R. L. Smith, C. J., Biggs, C. A. and Boxall, J. B., “Pilot scale laboratory investigations into the impact of steady state conditioning flow on potable water discolouration”, *Proceedings of WDSA*, Tucson, USA, (2010), ASCE online.

[12] Deines, P., Sekar, R., Husband, P. S., Boxall, J. B., Osborn, A. M., and Biggs, C.A., “A new coupon design for simultaneous analysis of in situ microbial biofilm formation and community structure in drinking water distribution systems”, *Applied Microbiology and Biotechnology*. DOI: 10.1007/s00253-010-2510-x, (2010).

[13] Vesanto, J., Alhoniemi, E., Himberg, J., Kiviluoto, K. and Parviainen, J., “Self-Organising Map for Data Mining in Matlab: the SOM Toolbox”, *Simulation News Europe*, Vol. 25, No. 54, (1999).