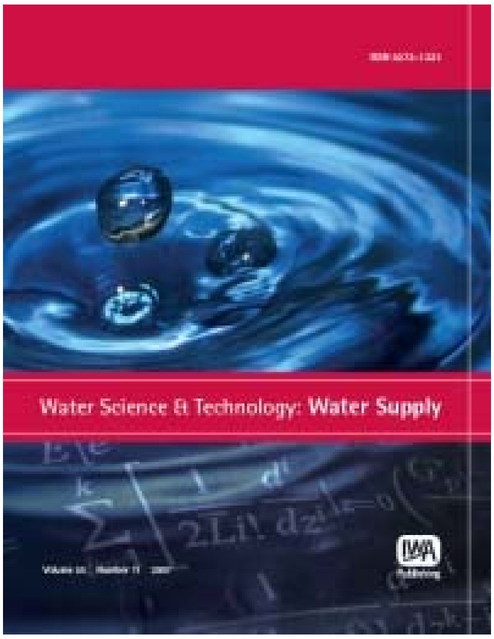
Water Science & Technology: Water Supply
IWA Publishing

# Relating organic matter character to trihalomethanes formation potential: a data mining approach

J. Roe, A. Baker and J. Bridgeman

## ABSTRACT

The removal of natural organic matter (NOM) during water treatment is becoming more important for all water utilities in the UK, as a result of tightened regulatory standards for trihalomethanes (THM), disinfection by-products (DBP) formed when residual organics react with chlorine. This paper considers the spatial and temporal variability of raw and clarified water arising from 16 surface water treatment works in the Midlands region of the UK. A wide range of investigation techniques are applied in order to develop effective strategies for the treatment of NOM-rich water. For the first time, rigorous data mining techniques are applied to a major dataset in order to examine potential inter-relationships between a wide range of quality parameters including, *inter alia*, total organic carbon (TOC), $UV_{254}$, coagulation pH, resin fractionation (hydrophilic acids (HPIA), hydrophobic acids (HPOA), hydrophilic non-acids (HPINA)) and total THM formation potential (TTHMFP). This paper focuses on the use of principal component analysis (PCA) to develop robust algorithms for the prediction of TOC removal and hence THM formation. Results show that raw water characteristics can be categorised into three main types, according to their HPOA content and specific ultraviolet absorbance (SUVA.). PCA identified possible THMFP precursors, according to raw water type verified by strong statistical relationships.

**Key words** | discriminant analysis, natural organic matter (NOM), principal component analysis (PCA), Trihamomethane formation potential (THMFP)

**J. Roe**
**J. Bridgeman**
School of Civil Engineering,
University of Birmingham,
Edgbaston,
Birmingham B15 2TT,
UK
E-mail: *jlr553@bham.ac.uk;*
*j.bridgeman@bham.ac.uk*

**A. Baker**
School of Geography,
Earth and Environmental Sciences,
University of Birmingham,
Edgbaston,
Birmingham B15 2TT,
UK

## INTRODUCTION

Natural organic matter (NOM) is found in abundance in nearly all natural waters around the world due to interactions between the hydrological cycle and both the biosphere and geosphere. It is available from a number of different sources, both pedogenic (e.g. soil and terrestrial vegetation) and aquagenic (e.g. biota in a water body), and is known to be variable both spatially and temporally (Kitis *et al.* 2002; Scott *et al.* 2002). NOM exhibits a complex mixture of organic compounds such as carbohydrates, lipids, protein biopolymers and humic macromolecules: the latter varying in terms of molecular weight and charge density (Edzwald 1993). The variability of NOM is due to a combination of variations in seasonal production and transport of pedogenic and aquagenic NOM, variations in their lability due to differences in chemical structure, together with variability in their in-stream microbial processing (Battin *et al.* 2008). Organic loading varies seasonally with increased levels occurring in the late summer and early autumn periods in the UK (Hurst *et al.* 2004). This has been attributed to increased microbial degradation of organic matter in the warmer summer months, with limited rainfall and enhanced evaporation, followed by a release of potential DOM in early autumn rainfall (Scott *et al.* 2002). Over the past decade, UK and US utilities have reported operational difficulties coinciding with rapid influxes of organic material at certain times of year, usually after periods of high and intense rainfall (Sharp *et al.* 2006).

doi: 10.2166/ws.2008.150

Conventional and well established treatment processes for NOM removal include coagulation using trivalent metal salts, flocculation and filtration. However, organics can cause severe problems during coagulation by coating particles and dominating the properties of inorganic colloids (Wilkinson *et al.* 1997) as NOM has a much higher surface area and negative charge than turbidity-causing matter (Kim & Yu 2005). NOM composition is dependent on source; however hydrophobic (HPO) and hydrophilic acid (HPI) fractions make up the largest percentage of NOM composition (Bache 2004). HPO material consists of humic and fulvic material. Humic acids are heterogeneous polyfunctional polymers formed by the breakdown of plant and animal tissue by chemical and biological processes, and so are more prevalent in surface waters and, because of their complex properties, are still among the least understood and characterised components in the environment (McDonald *et al.* 2004). HPO material is larger and more amenable to traditional removal methods. Unlike HPO, HPI material is much more difficult to remove through conventional treatment, being smaller, colourless, and with little or no charge density.

When residual NOM molecules react with chlorine in the final stage of treatment, potentially carcinogenic disinfectant by-products (DBPs) are formed. Any increased loading of NOM in untreated water systems is therefore an escalating risk to water treatment works (WTW) and customers alike. The two most abundant DBPs commonly found in disinfected waters are trihalomethanes (THM) and haloacetic acids (HAA) (Kitis *et al.* 2001). Investigations into THMs and HAAs have highlighted potential reproductive, carcinogenic and mutagenic effects (Milliarou *et al.* 2005; Babi *et al.* 2007; Wang *et al.* 2007). Over 500 DBPs have been identified to date, but only THMs are routinely measured in the UK (Milliarou *et al.* 2005). The current consent in the UK for total THMs is $100\,\mu g.L^{-1}$, with no standard for HAAs. The US THM and HAA standards are $80\,\mu g.L^{-1}$ and $60\,\mu g.L^{-1}$ respectively. It is possible that the UK could adopt similar standards in the future (Bose & Reckhow 2007).

This paper aims to identify key trends and potentially overlooked relationships between NOM character and trihalomethane formation potential (THMFP). A large data set has been developed using data from 16 WTWs over a two year period, to assess the changes in NOM character and composition, and relate this to THMFP. This paper applies data mining methods (i.e. discriminant analysis and principal component analysis (PCA)) to identify relationships ultimately relating key THMFP precursors to NOM character.

## MATERIALS AND METHODS

### Water source

PCA was undertaken on 16 WTWs located within the Severn and the Trent catchments, and owned and operated by Severn Trent Water. Samples were collected on a quarterly basis between March 2006 and February 2008, and a series of bench scale tests were undertaken in order to monitor source water characteristics.

### NOM Characterisation

#### DOC

Samples were analysed using a PPM Labtoc Analyser, with a range of $0.18–10\,mg.L^{-1}$ C. Samples were then filtered though a $0.45\,\mu m$ membrane prior to analysis. Samples were firstly mixed with persulphate, and inorganic carbon was purged off as $CO_2$. Samples were then swept by $N_2$ carrier to an infra-red detector to determine $CO_2$ at a wavelength of $4.4\,\mu m$, which was then related to the concentration of total carbon in sample.

#### $UV_{254}$

$UV_{254}$ absorbance analysis was performed using a Biochem Libra S12 Spectrophotometer, at a wavelength of 254 nm. Monthly and quarterly bench scale samples were filtered through a $0.45\,\mu m$ membrane to remove turbidity, and analysed with a 1 cm quartz cell which was rinsed with deionised (DI) water prior to each sample.

#### Turbidity

Turbidity measurements were performed using a HACH 2100N IS Turbidimeter. 30 ml of unfiltered sample was placed in a vial, which was pre-rinsed with DI water before each sample.

### Fractionation

Quarterly raw water samples were fractionated using XAD-7HP and XAD-4 resins. The resins were cleaned before each sample using 200 ml of 0.1 molar NaOH, 200 ml of 0.1 molar HCl and then with 200 ml of DI water. Samples were filtered though a Pall 0.45 μm membrane filter, pre-washed with 50 ml of DI water. Samples were then adjusted to pH 2 using 3 molar HCl. The water was pumped though the XAD-7HP resin, with the HPO fraction being absorbed. Samples were then passed through the XAD-4 resin, absorbing the HPIA fraction, leaving only the HPINA fraction.

### THM formation potential

Low pH jar tests were conducted to obtain THM formation potential (THMFP) under optimum pH conditions. Coagulant dose was set to current plant conditions whilst the pH was reduced to 4.5 using 0.1 m or 3 m NaOH and HCl. Jar tests were performed using a Phipps and Bird six paddle jar test apparatus. Ferric sulphate was added at the start of a 1.5 minute rapid mix stage at 200 rpm. A 15 minute slow mix stage at 30 rpm followed, after which jars were allowed to settle for 20 minutes. THMFP analysis was performed using a HP6890 Gas Chromatograph, fitted with electron capture detector (ECD). Samples were firstly filtered through a 0.45 μm membrane, and then buffered to pH 7 and spiked to approximately 5 mg.L$^{-1}$ free $Cl_2$, prior to storage for 7 days. On day 3, extra free $Cl_2$ was added if necessary. After the required time, portions of the sample were transferred to a septum vial. After equilibrium with headspace at 80°C, a sample vapour was injected by autosampler onto a capillary column GC fitted with ECD to determine quantitatively THMs present.

Statistical techniques are increasingly being used in data analysis as they allow the rapid analysis of large and multidimensional data sets. Discriminant and principal component analysis were performed using SPSS Inc. Version 16.0. Discriminant analysis is a tool to determine an optimum combination of variables to provide the optimum discrimination between sites (Spencer *et al.* 2007). Data are split into functions, function 1 providing the most variation. Functions are orthogonal to one another and so their contributions to the discrimination do not overlap (Spencer *et al.* 2007). PCA is a way of reducing and simplifying data sets by means of linear transformations, detecting significant patterns in the data.

## RESULTS AND DISCUSSION

Raw water characteristics from all 16 surface water sites can be seen in Table 1.

### Discriminant analysis of raw water NOM character

Discriminant analysis was performed only on the fractionation and SUVA data to detect potential differences between sites. The main component of Discriminant Function 1 is SUVA, a prime indicator of NOM character, obtained by dividing the UV absorbance of a given sample at a wavelength of 254 nm, by the DOC concentration in mg.L$^{-1}$. The main component in Discriminant Function 2 is the total HPO fraction in mg.L$^{-1}$. These two functions are able to achieve a broad spread of the sites, implying these variables are important for site discrimination (Figure 1). On the basis of the discriminant analysis, the sites can be split into three main types.

Type 1 consists of sites 1, 7 and 13, which are typically moorland source waters, characterised by higher total fractions of high molecular weight HPO material, NOM consisting of aquatic humic and fulvic acids. Type 1 water usually reacts favourably to coagulation and flocculation processes, with large amounts of total DOC removed during treatment, but the large scatter visible in Figure 1 demonstrates that this water type is the most susceptible to seasonal variation. Type 2 waters include sites 5, 6, 8, 9, 10, 12, 14, 15 and 16. Type 2 waters contain a mixture of molecular weights, and hydrophilic and hydrophobic NOM. Due to the HPI fraction in the water, sites are usually less amenable to standard treatment processes, but good removal can still be achieved with optimised processes. Type 2 waters are intermediate water types between types 1 and 3. Finally, Type 3 waters include sites 2, 3, 4 and 11. These sites exhibit high levels of HPI material in raw source waters and removal of total DOC is generally below 25%. Type 3 sites are typically situated in lowland, more

**Table 1** | Average raw water characteristics and plant removal

| Site | pH | $UV_{254}$ (abs.$m^{-1}$) | Turbidity (NTU) | DOC (mg.$l^{-1}$) | SUVA* ($m^{-1}$.l.$mg^{-1}$) | HPIA (mg.$l^{-1}$) | HPINA (mg.$l^{-1}$) | Average % Plant Removal | Standard Deviation |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.17 | 32.55 | 1.61 | 6.12 | 5.12 | 0.95 | 1.22 | 78.63 | 4.10 |
| 2 | 7.59 | 14.74 | 3.73 | 5.72 | 2.45 | 1.44 | 1.59 | 23.51 | 10.06 |
| 3 | 7.58 | 12.91 | 7.64 | 5.30 | 2.32 | 1.37 | 1.83 | 21.38 | 12.12 |
| 4 | 7.88 | 12.15 | 1.06 | 5.96 | 2.25 | 1.61 | 1.95 | 18.50 | 6.72 |
| 5 | 7.42 | 13.19 | 7.16 | 4.36 | 2.93 | 0.88 | 1.34 | 35.01 | 13.40 |
| 6 | 7.39 | 14.79 | 3.06 | 4.49 | 3.14 | 1.00 | 1.25 | 48.60 | 7.08 |
| 7 | 6.69 | 11.04 | 1.13 | 2.68 | 3.92 | 0.44 | 0.74 | 60.17 | 7.20 |
| 8 | 7.43 | 8.67 | 3.52 | 2.81 | 2.86 | 0.73 | 0.97 | 48.38 | 17.23 |
| 9 | 7.16 | 13.24 | 5.65 | 3.98 | 3.12 | 1.15 | 0.98 | 48.39 | 9.15 |
| 10 | 7.55 | 13.14 | 1.63 | 4.68 | 2.70 | 1.06 | 1.50 | 36.41 | 9.00 |
| 11 | 7.57 | 19.41 | 4.01 | 7.15 | 2.61 | 2.16 | 2.00 | 46.49 | 6.11 |
| 12 | 7.51 | 8.85 | 3.92 | 2.87 | 2.83 | 0.66 | 1.14 | 35.98 | 7.91 |
| 13 | 6.84 | 24.28 | 3.49 | 6.71 | 3.47 | 1.66 | 1.5 | 67.16 | 4.84 |
| 14 | 7.60 | 8.85 | 1.42 | 3.08 | 2.72 | 0.79 | 0.90 | 25.38 | 15.14 |
| 15 | 7.37 | 14.61 | 6.91 | 4.29 | 3.18 | 1.11 | 1.20 | 36.67 | 9.33 |
| 16 | 7.63 | 11.79 | 1.18 | 4.32 | 4.53 | 1.14 | 1.52 | 18.88 | 7.57 |

∗ SUVA—Specific UV Absorbance ($UV_{254}$ [$m^{-1}$]/DOC [mg.$l^{-1}$])

urbanised catchments and are less influenced by seasonal variability.

## Principal component analysis & stepwise regression

PCA was performed in order to determine any potential link between source water characteristics and THMFPs present after conventional treatment processes.

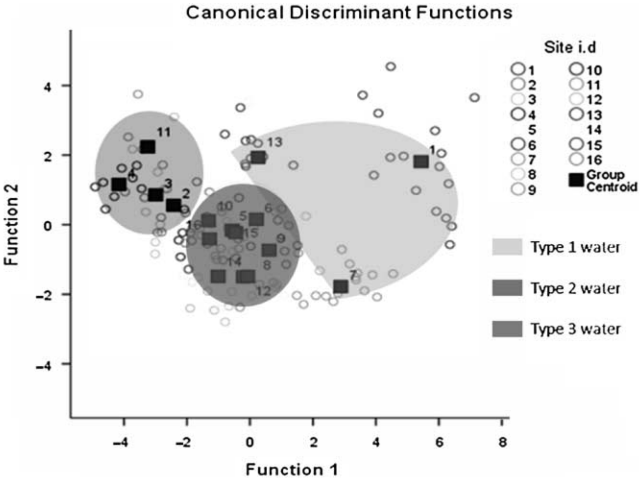

**Figure 1** | Discriminant analysis.

Due to the large number of sites, for the purpose of this paper, only sites 1, 4 and 10 are considered in detail. Sites 1 and 4 are at opposing ends of the SUVA scale and are typical examples of Types 1 and 3 waters (Figure 1). Site 10 is closest to the average SUVA for all sites and is an example of Type 2 waters.

## Type 1

For Type 1 waters (typical of moorland catchment raw waters) component 1 is characterised by high UV, HPO and DOC, with chloroform and bromodichloride forming after disinfection (Figure 2). Component 2 is characterised by high NTU and HPIA, and are low in remaining THMFP per $mgL^{-1}$ of DOC. As shown in Figure 2, samples are identified by month number, with a clear distinction occurring seasonally even with some overlap between quarters. Months April (4) and October (10) are typically identified as having lower turbidity and less HPINA, whereas months July (7) and January (1) are the opposite. October waters are consistently higher in HPO material and total DOC. Stepwise regression was found to show a positive relationship between THM chloroform and HPO material.

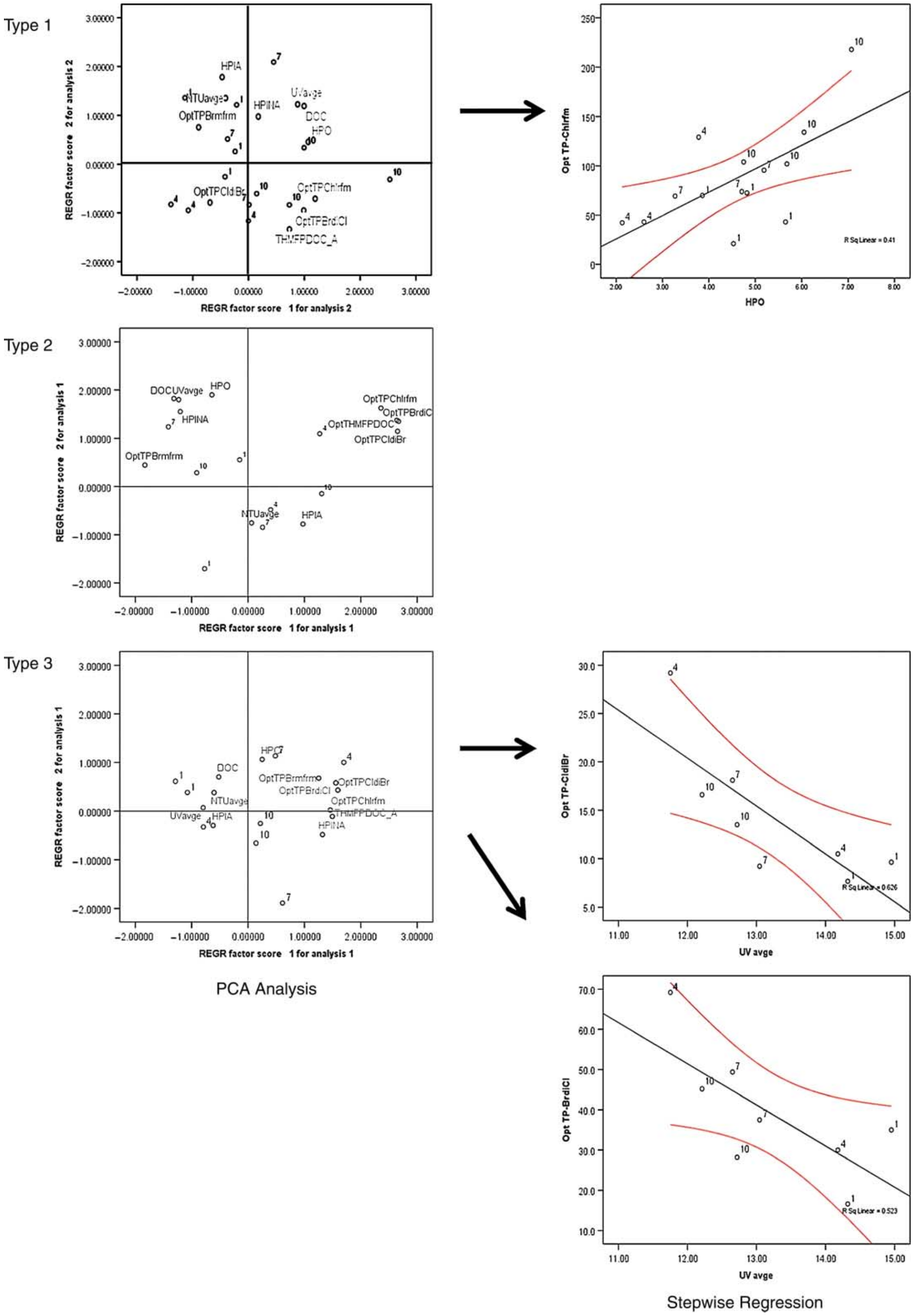

**Figure 2** | PCA and Stepwise Regression. Samples are labelled by months, e.g. January (1), April (4), with tinted text indicating spread of variables in plotted components. Stepwise regression graphs show trend lines with (coloured) mean confidence bands. In the stepwise regression, scatter is evident, occurring as a result of natural variation. Subscribers to the online version of *Water Science and Technology: Water Supply* can access the colour version of this figure from http://www.iwaponline.com/ws

**Table 2** | Stepwise regression relationship in raw waters. Table identifies the variables with which stepwise regression relationships occurred, and the statistical significance. The statistical significance levels for the sites remain high for the number of data points, with site 1 as 0.01, $n = 14$; 0.05 for site 4, $n = 8$

| | Site | Chloroform | Chlorodibromide | Bromoform | Bromodichloride |
|---|---|---|---|---|---|
| Type 1 | 1 | HPO (0.64) | | | |
| | | HPO + HPIA (0.77) | | | |
| | 7 | HPO (0.65) | | | |
| | 13 | | | | |
| Type 2 | 5 | | | | |
| | 8 | DOC (0.93) | HPO (0.76) | HPINA (0.68) | |
| | 9 | | | | |
| | 10 | | | HPINA (0.51) | |
| | | | | HPINA + NTU (0.85) | |
| | 14 | | | | |
| | 16 | | | | |
| | 15 | | NTU (0.62) | | |
| Type 3 | 3 | | HPO (0.81) | HPINA (0.84) | |
| | 2 | | UV (0.63) | HPIA (0.78) | UV (0.52) |
| | | | UV + NTU (0.85) | | |

### Type 2

Type 2 waters typically consist of mixed molecular weight and HPI/HPO content (Figure 2). Component 1 is high in chloroform, chlorodibromide and bromodichloride, with a high residual THMFP μg.l$^{-1}$ per DOC mgL$^{-1}$. Also, there is a negative relationship with UV, DOC and bromoform. Component 2 is characterised as high in UV, HPO and DOC, with a negative relationship with HPIA and turbidity. April (4) is consistently high in residual THMs, and January (1) THMFPs after treatment occur less, however there is no clear trend occurring throughout the diurnal profile over the 2 year sampling period.

### Type 3

Type 3 waters are typical of lowland sources, high in total HPI material and consistently difficult to remove total DOC. In this case, component 1 is high in HPINA and all THMFPs, with a high remaining THMFP μg.l$^{-1}$ per DOC mgL$^{-1}$ (Figure 2). Component 2 waters are characterised by being high in HPO, DOC and bromoform. October (10) and January (1) remain consistent over the two year sampling period, but April (4) and July (7) do not. July does remain high in HPINA and THMFPs after treatment; however there are notable differences in the component 2 characteristics. In the stepwise regression, Type 3 waters were found to have strong relationships occurring between chlorodibromide, bromodichloride and UV.

The stepwise regression relationships obtained on each of the four THMs and the statistical significance of the $r^2$ relationship are shown in Table 2. Type 1 waters show only relationships occurring with chloroform, with HPO material being a common precursor. Where there is a dominance of HPI material in the raw waters, particularly HPINA, relationships with the remaining THMs are also found.

## CONCLUSIONS

- Using discriminant analysis, it is possible to split Severn Trent Water source water into three distinct types based on the raw water SUVA and the HPO fraction.
- Type 1 waters are typically moorland source waters, with a dominance of HPO material. Type 1 waters show distinct seasonal variations, with late summer and autumn periods experiencing notably higher total DOC concentrations and HPO content.

- Stepwise regression for Type 1 waters indicates a relationship between chloroform and HPO. Months with increased levels of HPO were found to have higher chloroform levels occurring after treatment, indicating HPO is typically a precursor for chloroform at these sites.
- Type 2 waters contain a mixture of both HPO and HPI material. Seasonal trends were not observed at the representative site, although THM precursor identification was observed at some sites. The only notable trend in the stepwise regression analysis occurred with bromoform and HPINA within the Type 2 source water group.
- Type 3 waters characteristically consist of low molecular weight, HPI NOM. This is known to hinder the removal of NOM in typical water treatment processes, leaving higher amounts of NOM remaining to potentially react with disinfectants and form undesirable THMs. In the PCA, seasonal trends were apparent in the autumn and winter months, however this trend was not carried on into the remaining quarters. No relationships were encountered with the occurrence of chloroform, possibly due to the absence of HPO material. However strong relationships between potential precursors and the remaining three THMs were observed.
- Statistical analysis can be hugely beneficial with large datasets, and is useful for identifying the key components for sites. Techniques can confirm existing data interpretations and detect new relationships between variables which would have previously been overlooked.

## ACKNOWLEDGEMENTS

The authors would like to thank Severn Trent Water Ltd and the University of Birmingham for funding and technical support.

## REFERENCES

Babi, K., Koumenides, K., Nikolaou, A. *et al.* 2007 Pilot study if the removal of THMs, HAAs and DOC from drinking water by GAC adsorption. *Desalination* **210**(1–3), 215–224.

Bache, D. 2004 Floc rupture and turbulence: a framework for analysis. *Chem. Eng. Sci.* **59**, 2521–2534.

Battin, T., Kaplan, L., Findlay, S., Hopkinson, C., Marti, E. *et al.* 2008 Biophysical controls on organic carbon fluxes in fluvial networks. *Nat. Geosci.* **1**, 95–100.

Bose, P. & Reckhow, D. 2007 The effect of ozonation on natural organic matter removal by alum coagulation. *Water Res.* **41**, 1516–1524.

Edzwald, J. 1993 Coagulation in drinking water treatment: particles, organics and coagulants. *Water Sci. Technol.* **27**(11), 21–35.

Hurst, A., Edwards, M., Chipps, M., Jefferson, B. & Parsons, S. 2004 The impact of rainstorm events on coagulation and clarifier performance in potable water treatment. *Sci. Total Environ.* **321**, 219–230.

Kim, H. & Yu, M. 2005 Characterisation of natural organic matter in conventional water treatment processes for selection of treatment processes on DBPs control. *Water Res.* **39**, 4779–4789.

Kitis, M., Kranafil, T., Kilduff, J. *et al.* 2001 The reactivity of natural organic matter to disinfection by-products formation and its relation to specific ultraviolet absorbance. *Water Sci. Technol.* **43**(2), 9–16.

Kitis, M., Karanfil, T., Wigton, A. & Kildiff, J. 2002 Probing reactivity of dissolved organic matter for disinfection by-product formation using XAD-8 resin adsorption and ultrafiltration fractionation. *Water Res.* **36**, 3834–3848.

Milliarou, E., Collins, C., Graham, N. & Nieuwenhuijsen, M. 2005 Haloacetic acids in drinking water in the United Kingdom. *Water Res.* **39**, 2722–2730.

McDonald, S., Bishop, A., Prenzler, P. *et al.* 2004 Analytical chemistry of freshwater humic substances. *Anal. Chim. ATAC* **527**(2), 105–124.

Spencer, R. G. M., Baker, A., Uher, G. & Goddard, R. 2007 Discriminatory classification of natural and anthropogenic waters in two U.K. estuaries. *Sci. Total Environ.* **373**, 305–323.

Scott, M., Jones, M., Woof, C., Simon, B. & Tipping, E. 2002 The molecular properties of humic substances isolated from a UK upland peat system. A temporal investigation. *Environ. Int.* **27**, 449–462.

Sharp, E. L., Parsons, S. A. & Jefferson, B. 2006 The impact of seasonal variations in DOC arising from a moorland peat catchment on coagulation with iron and aluminium salts. *Environ. Pollut.* **140**, 436–443.

Wang, G., Deng, Y. & Lin, T. 2007 Cancer risk assessment from trihalomethanes in drinking water. *Sci. Total Environ.* **387**, 86–95.

Wilkinson, K., Negre, J. & Buffle, J. 1997 Coagulation of colloidal material in surface waters: the role of natural organic matter. *J. Contam. Hydrol.* **26**, 229–243.