Co-occurrences of EDCs/PPCPs in Surface Water Using Chemometrics

MARK HAGEMANN,1 MINJI PARK,1 VARUN SRINIVASAN,1 DAVID A. RECKHOW,1 MICHAEL LAVINE,2 ERIK ROSENFELDT,3 BENJAMIN D. STANFORD,4 AND MI-HYUN PARK1

1Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, Mass.
2Department of Mathematics and Statistics, University of Massachusetts, Amherst, Mass.
3Hazen and Sawyer P.C., Richmond, Va.
4Hazen and Sawyer P.C., Raleigh, N.C.

This study investigated co-occurrences of endocrine-disrupting compounds (EDCs), and pharmaceuticals and personal care products (PPCPs) in order to develop effective monitoring strategies. EDCs/PPCPs were clustered on the basis of similarities in their occurrence in surface waters to reduce analytical complexity. Chemometric approaches were applied to three water bodies with different water systems and climate conditions: Lake Mead in Nevada, the Assabet River in Massachusetts, and the Santa Ana River in California. The results show that site-specific co-occurrences among EDCs/PPCPs exist, though these co-occurrences do not translate well between sites. Therefore, the usefulness of this study is in demonstrating the approach of selecting indicator/surrogate compounds using chemometrics rather than providing a single list of recommended compounds for long-term monitoring. This study offers a systematic and practical approach to selecting a suite of analytes when implementing a monitoring program for EDCs/PPCPs in surface waters.

Keywords: chemometrics, endocrine-disrupting compounds, indicators, occurrences, patterns, pharmaceuticals and personal care products, surrogate

Public concerns about endocrine-disrupting compounds (EDCs) and pharmaceuticals and personal care products (PPCPs) in aquatic environments have been increasing since these compounds have been linked to adverse effects on ecosystem and/or human health even at low concentrations (Kumar & Xagoraraki 2010, Bercu et al. 2008, Webb et al. 2003). Although EDCs/PPCPs are not currently regulated in the United States, many of these compounds are frequently found in surface water, wastewater treatment plant (WWTP) effluents, and finished drinking water worldwide (Meyer et al. 2011, Eaton et al. 2010, Lee & von Gunten 2010, Benotti et al. 2009, Dickenson et al. 2009), leading to increasing pressures to monitor their levels in human-affected waters. This is an issue for water utilities especially in the case of direct or indirect water reuse. The combined trends of increasing concern for such compounds and increasing popularity of water reuse (Angelakis & Gikas 2014) suggest the importance of designing effective monitoring campaigns for EDCs/PPCPs in drinking water sources.

Assessing EDCs/PPCPs is expensive, time-consuming, and complex, requiring the analyses of large numbers of compounds, possibly at multiple locations (Kolpin et al. 2002), with some degree of a priori knowledge of what compounds may be of concern for a community and why they should be monitored. Therefore, the selection of analytes should be carefully designed in order to effectively and efficiently assess EDCs/PPCPs in surface waters. It may be possible to reduce the amount of redundant information in a suite of analytes by forgoing the measurement of compounds whose presence is indicated by either other co-occurring compounds or related conventional water quality constituents. Such information will be useful in developing site-specific monitoring strategies.

Previous studies have proposed various tracers of wastewater that could indicate the presence of EDCs/PPCPs because wastewater effluent is a major source of EDCs/PPCPs. For example, boron has been used as an indicator of wastewater in surface waters in which ambient levels are suitably low (Schreiber & Mitch 2006). Caffeine and 35 other commonly detected chemicals including PPCPs have been identified as indicator candidates for wastewater (Glassmeyer et al. 2005). Caffeine is a good indicator of anthropogenic effects (Buerge et al. 2003) but is not a conservative tracer because of its biodegradation in water (Bradley et al. 2007). Primidone and carbamazepine have also been suggested as indicators of wastewater (Guo & Krasner 2009, Drewes et al. 2008). Recent studies have demonstrated the use of sucralose (Mawhinney et al. 2011, Oppenheimer et al. 2010) and gadolinium...
(Lawrence & Bariel 2010, Verplanck et al. 2010) to indicate the presence of domestic and hospital wastewater, respectively.

Common chemometric approaches to finding patterns among conventional water quality parameters include cluster analysis (CA) and principal component analysis (PCA). CA divides a set of observations into groups that are based on their similarity. PCA is commonly used to identify underlying patterns in the data and to find combinations of variables explaining the largest amount of a data set's total variance (Ouyang 2005). In water quality monitoring applications, CA has been used for grouping water monitoring data according to their similarity in time and space. PCA has been used to identify combinations of water quality parameters that describe the most variation within a data set (Razmkhah et al. 2010, Varol & Sen 2009, Bouza-Deano et al. 2008). Astel et al. (2006) applied CA and analysis of variance to a complex data set in order to evaluate temporal and spatial variations in drinking water quality. CA, PCA, and discriminant analysis have been used to select groups of water quality monitoring locations with similar pollution characteristics (Osman et al. 2012, Varol et al. 2012, Al-Odaini et al. 2011, Mustapha & Aris 2011, Shrestha & Kazama 2007, Kowalkowski et al. 2006). CA, PCA, and a self-organizing map were compared to assess river water quality and find natural clusters of monitoring stations within similar water quality type (Astel et al. 2007, Brodnjak-Voncina et al. 2002). However, chemometrics has not yet been extensively applied for EDCs/PPCPs to find relationships among compounds.

The objectives of this study were to characterize the patterns of EDCs/PPCPs and/or conventional water quality constituents and to demonstrate systematic and practical approaches to selecting indicator/surrogate compounds for long-term monitoring. The main purposes are to find co-occurrences of compounds and indicator/surrogate compounds, and then to assess how accurately these indicator/surrogate compounds represent the presence and concentration of EDCs/PPCPs in surface water. This methodology could be applied as an initial dimensionality reduction step in the design of a long-term monitoring campaign, especially one that deals with a large number of difficult-to-analyze chemicals.

METHODS

Study areas. Three water bodies in the continental United States were used for this study: Lake Mead in Nevada, the Assabet River in Massachusetts, and the Santa Ana River in California (Figure 1). These waters were selected to represent different water systems and climates.

Lake Mead, the largest reservoir in the United States, receives approximately 97% of its inflow from the Colorado River and 1.5% from the wastewater-dominated Las Vegas Wash (LVW). Treated wastewater effluent provides approximately 90% of LVW base flow, which averages around 160 mgd (Shanahan & Zhou 2011). Annual precipitation in the region is 9 in., most of which falls in winter. Frequent high-flow events result from stormwater discharge from metropolitan Las Vegas, although it is attenuated via detention basins. Lake Mead supplies water for 30 million residents in the Las Vegas Valley through the Alfred Merritt Smith (AMS) Water Treatment Facility and River Mountain Water Treatment Facility, which convey water from intakes in the

Lake. Raw water at AMS includes water from LVW but only after substantial transit time and dilution with Colorado River water. The 31-mi-long Assabet River flows north from its headwaters in Westborough, Mass., to its confluence with the Concord River. The river is used as recreation and public water supply for seven towns and receives effluent from four municipal WWTPs (approximately 16 mgd) and runoff from the watershed. Annual precipitation is 48 in. with high flow during spring snowmelt conditions and low flow during summer.

The Santa Ana River, the longest river in Southern California (96 mi), provides water to approximately two million residents in more than 20 cities and contributes groundwater recharge to the Orange County Basin. The river receives tertiary treated wastewater (>50% of base flow) as well as stormwater runoff and groundwater (Guo et al. 2010). Wastewater discharge contributes the majority of stream flow during dry-weather periods in summer, whereas stormwater runoff contributes most of the stream flow during wet-weather periods in winter. Mean annual precipitation is 10 in.

Sampling data acquisition and preprocessing. A 10-year record of water quality data from Lake Mead was obtained from the Southern Nevada Water Authority. The data include monthly samples of 20 EDCs/PPCPs as well as conventional water quality constituents. Detailed information on sample collection and analytical methods is available in Snyder et al. (2008). Individual data sets collected at LVW and raw water at AMS taken from intakes were analyzed separately because of the different hydrological regimes represented by these two sampling stations. Conventional water quality data including total dissolved solids (TDS), turbidity, and temperature (obtained from Benotti et al. 2010) that coincided with the EDC/PPCP sampling campaign were used for analysis. These data were available only at LVW and AMS between 2006 and 2010, overlapping with fewer than half of the EDC/PPCP samples.

Assabet River grab samples were collected and analyzed for 27 compounds and conventional water quality constituents including pH, temperature, and conductivity. Monthly sampling was conducted at 10 sampling stations along the river between August 2011 and March 2013 including stations upstream and downstream of four WWTPs and two tributaries. Detailed information of sampling and analytical methods is available in Srinivasan (2012).

Santa Ana River sampling data for 49 EDC/PPCP compounds were obtained from Guo et al. (2010). The data set contains data from quarterly sampling at 13 sites between May 2008 and February 2009. No coincident conventional water quality data were available for the sampling dates and locations.

Certain characteristics of the different data sets complicated the analysis and preprocessing was required. Because of the disparity in analytes across systems, patterns among compounds were characterized separately for each system. Most compounds included values reported as below a method detection limit (MDL), requiring imputation of these values. Instead of using half the MDL (Antweiler & Taylor 2008), data values below MDLs were stochastically imputed by sampling from a maximum-likelihood distribution truncated to the interval (0, MDL) (Delignette-Muller & Dutang 2015). In most cases, data were right-skewed and a lognormal distribution was used for imputation, but in certain
cases in which data were approximately symmetrical, a normal distribution was used instead. Those variables with right-skewed distributions were then transformed using either logarithmic or square-root transformation in order to approximate normality (Esmen & Hammad 1977). To facilitate cluster analysis and to compare relationships across stations in the Assabet River and Santa Ana River data, the (potentially transformed) data from each site were scaled to have zero mean and unit variance prior to analysis (Cao et al. 1999). Compounds from stations with fewer than two values above the MDL were excluded from the analysis. The distribution of each variable was then assessed using boxplots, quantile-quantile plots (Wilk & Gnanadesikan 1968), and the Shapiro–Wilk normality test (Shapiro & Wilk 1965) to evaluate their normality.

Chemometric analyses. Chemometric methods were used to explore the correlation structure within each suite of compounds, with the objective of identifying groups of co-varying compounds such that one member of each group could be used as an indicator of the other members. Specifically, hierarchical CA was employed to establish a hierarchy of similarity among compounds, while PCA with varimax rotation, a modification of PCA that seeks to find components largely composed of relatively few variables, was used to identify groups of compounds that explained the largest amount of total variance in the data matrix. The R statistical computing language (version 3.1)\(^1\) was used for statistical analysis.

As an exploratory tool, correlation matrixes were constructed using the Pearson correlation, which measures the linear dependence between two variables. Pairwise scatter plots of the transformed compounds were used to visually examine relationships between them.

To cluster compounds on the basis of similarity of occurrence, hierarchical CA was used on the basis of squared Euclidean distance, \(D(x,y) = \sum (x_i - y_i)^2\), using Ward’s minimum variance method (Ward 1963). Note that this is mathematically similar to the Pearson correlation, \(r(x,y) = 1/n \sum x_i y_i\), when variables have zero mean and unit variance; i.e., pairs of variables with high correlation have low squared Euclidean distance. The result is graphically represented as a dendrogram that links groups of

---

**FIGURE 1** Location and sampling stations of study areas\(^a\)

\(^a\)Lake Mead, Nev.; Assabet River, Mass.; and Santa Ana River, Calif.
variables into progressively less similar clusters (Simeonov et al. 2010). Using the resulting hierarchy, it is straightforward to assign groups with an arbitrary degree of internal similarity; this study used one-third the maximum squared Euclidean distance (Hothorn & Everitt 2009).

PCA was used to find patterns of co-occurrence among compounds and to identify which linear combinations of compounds explain the most variance in the data set. PCA transforms the original data to a set of new orthogonal variables called principal components (PCs) that reduce the dimensionality of the original data while preserving much of the data’s original information (Mas et al. 2010). The optimal number of PCs was quantified using the combination of eigenvalue-one criterion and Scree test (Vandeginste et al. 1998). Varimax rotation was applied to the PCs before interpreting the PCA results (Abdi & Williams 2010).

RESULTS

Because the suite of analytes was different for each monitoring campaign, the consistency of occurrence can be compared only...
for certain compounds (Figure 2). For example, sulfamethoxazole was frequently detected (>73%) in all water systems, whereas estradiol had low detection frequency (<6%) in all water systems in which it was measured. Other compounds, such as diclofenac, varied widely in their detection frequency across water systems. However, such differences may be a function of analytical precision as well as concentration of compounds.

Although the log-transformation ameliorated the right-skewed distributions of the raw concentration data, the transformed data were not always normally distributed (36 out of 107 cases were significantly non-normal by the Shapiro–Wilk test). Non-normality violates the assumptions of CA and PCA based on Pearson correlation and can lead to spurious correlations in which the departure from normality is large. In order to establish the effect of non-normality in the data, a separate data set was obtained by coercing the values to be normal by mapping the data quantiles to the cumulative distribution function (cdf) of a standard normal distribution

\[ x_{i}^{*} = \Phi^{-1}\left(\frac{\text{rank}(x_{i}) - 0.5}{n}\right) \]

where \( \Phi^{-1}(\cdot) \) is the inverse of the standard normal cdf. The deviation between the Pearson correlation coefficients in the normal-coerced data and the original (log-transformed) data was <0.15 for all pairs of EDCs/PPCPs in LVW, AMS, and Assabet River, suggesting that the non-normal distributions did not substantially bias these analyses. Deviations in Santa Ana River were up to 0.19, but the majority of cases were <0.1, and no apparent bias was observed toward higher or lower coefficients.

Las Vegas Wash. At LVW, 13 compounds (65% of total analytes) were detected with >85% frequency, with pharmaceuticals making up the most frequently detected analytes (Figure 2, part A). The detection frequencies of many of these compounds were greater than those previously found in US waters (Eaton et al. 2010, Snyder 2010, Barnes et al. 2002, Kolpin et al. 2002) as LVW is dominated by wastewater discharge, which is the primary source of PPCPs. Conversely, hormones (i.e., estradiol, ethynylestradiol, progesterone, testosterone) were detected at low frequency (<10%) possibly because of their high removal efficiencies through wastewater treatment processes (Johnson & Sumpter 2001).

CA revealed eight clusters of EDCs/PPCPs and conventional water quality constituents (Figure 3, part A), with varying degrees of internal similarity. Clusters 1–3 and three components of cluster 4, mainly composed of pharmaceuticals, had relatively high internal similarity, with total squared Euclidean distance well below the separating threshold of one-third of the maximum. Clusters 6–8, in contrast, had relatively low internal similarity, and were composed mainly of compounds with low detection frequency and conventional water quality constituents with relatively few sample points \( n < 32 \). These groups composed a separate branch from clusters 1–3, indicating low similarity with the pharmaceuticals.

PCA results (Figure 4) for LVW distinguished between many of the groups from CA. PC1 distinguished cluster 2 (Dilantin, tris(2-carboxyethyl)phosphine [TCEP], carbamazepine, and meprobamate) from the others, confirming both the mutual similarity of its compounds and the relatively large distance between them and the other compounds. PCs 2 and 3 distinguished clusters 1 (diazepam, diclofenac, and sulfamethoxazole), and 4 (naproxen, gemfibrozil, and trimethoprim) excluding fluoxetine, respectively.

Scatterplots of clustered compounds (Figure 5) show that certain compounds may be used as indicators for one another, while others might not, despite belonging to the same cluster. The similarity of compounds, reflected in the squared Euclidean distance of clusters in CA and similarity of loadings in PCA, indicated their co-occurrences, as revealed in the scatterplots. For example, gemfibrozil and trimethoprim (Figure 5, part A) were sufficiently collinear \( r = 0.78 \); therefore, either of these compounds can be used as an indicator of the other. Fluoxetine and trimethoprim (Figure 5, part C) were less collinear \( r = 0.56 \) and might not be useful as indicators of each other. Nitrate-N, a conventional water quality parameter, was not likely to be useful as a surrogate of triclosan \( r = 0.31 \) despite belonging to the same cluster with moderate internal similarity (Figure 5, part D).

Alfred Merritt Smith treatment plant inflow. In AMS raw water from Lake Mead, five compounds (25% of total analytes) were detected above 80% frequency, with the remaining compounds detected less than 25% of the time (Figure 2, part B). Most of the five frequently detected compounds were pharmaceuticals that were also frequently detected at LVW except atrazine, which was infrequently detected at LVW. Several compounds detected with >85% frequency at LVW (i.e., N,N-Diethyl-meta-toluamide [DEET], trimethoprim, triclosan, naproxen, gemfibrozil) were less frequently detected in AMS. Even compounds detected with >60% frequency at LVW (i.e., TCEP, fluoxetine, diclofenac, musk ketone, diazepam) were detected in only a single sample or not at all in AMS inflow.

CA grouped the compounds into six clusters (Figure 3, part B), with cluster 1 composed of the four most frequently detected compounds (i.e., sulfamethoxazole, Dilantin, carbamazepine, meprobamate) and the remaining compounds composing a separate branch with low similarity to cluster 1. The most similar pharmaceuticals (i.e., carbamazepine, meprobamate and Dilantin in cluster 1) were also highly similar at LVW. Naproxen, gemfibrozil, and trimethoprim, which were similar in AMS inflow, also formed a cluster at LVW, although their low detection frequency relative to those at LVW resulted in lower internal similarity.

PCA results (Figure 6) similarly show the co-occurrence of the pharmaceuticals in cluster 1, with PC1 distinguishing them from other compounds. Cluster 2 (turbidity, triclosan, atrazine, and TDS) was distinguished in PC2. Other clusters were not well delineated in PCA, reflecting both their low internal similarity and relatively low between-cluster distance.

Pairwise scatter plots show strong positive relationships \( r > 0.6 \) between compounds in cluster 1 (Figure 7, parts A, B, C), suggesting any of them could be used as an indicator for the others. Other clusters were less well suited to indicate the presence of
other compounds because of their lower internal similarity. For example, TDS did not appear reliable for use as a surrogate for atrazine from the scatter plot (Figure 7, part D) despite belonging to the same cluster (Figure 3, part B).

**Assabet River.** In the Assabet River, eight compounds (25% of total analytes) were detected with >50% frequency—most of these pharmaceuticals (Figure 2, part C). The most frequently detected compounds were DEET and sulfamethoxazole (>70%), which were previously reported to be frequently present in US waters (Snyder 2010, Kolpin et al. 2002). Hormones (i.e., estrone, estradiol) were detected at low frequency (<10%).

The dendrogram from CA revealed eight clusters (Figure 3, part C). The clusters in the dendrogram show larger distance among clusters than those in Lake Mead data, indicating lower internal

**FIGURE 3** Dendrograms from cluster analysis of compounds

![Dendrograms from cluster analysis of compounds](image-url)

DEET—N,N-Diethyl-meta-toluamide, NO₃−N—nitrate as nitrogen, TCEP—tris(2-carboxyethyl)phosphine, TDS—total dissolved solids

Dashed lines separate groups as using one-third maximum squared Euclidean distance.
similarity among compounds in the clusters. The pairs of carbamazepine and Dilantin, and sulfamethoxazole and diclofenac were highly similar despite their low detection frequencies (Figure 2, part C). Conventional water quality constituents (i.e., conductivity, pH, temperature) did not show similarity to any EDC/PPCP compounds.

PCA results for Assabet River data (Figure 8) were less coherent than those for Lake Mead data, in agreement with the relatively small squared-Euclidean distance between CA clusters. PC3 distinguished cluster 1 compounds; other groups were not well delineated by the PCs.

Scatterplots of many compounds in CA clusters (Figure 9) do not show significant collinearity, suggesting less potential for their use as indicators. This is likely due to the low detection frequencies of the compounds, resulting in noise from the imputation of the data values below MDLs. For example, diclofenac and sulfamethoxazole show only fair correlation ($r = 0.50$; Figure 9, part A) despite having among the highest similarity from CA, as 75% of diclofenac and 27% of sulfamethoxazole data were randomly imputed. Although conventional water quality constituents, caffeine, and sucralose were not useful to indicate the presence of EDCs/PPCPs (Figure 9, parts C and D), gadolinium appeared to be a potential indicator of TCEP (Figure 9, part B).

**Santa Ana River.** In the Santa Ana River, eight compounds were detected with >85% frequency (50% of total analytes), with only three detected in fewer than half of the samples (Figure 2, part D). The most frequently detected compounds above MDLs were caffeine, TCEP, diuron, DEET, sulfamethoxazole, primidone, carbamazepine, and Dilantin (≥75%). TCEP and caffeine, the most frequently detected compounds in SAR, were previously identified as frequently detected compounds in US waters (>30%)

---

**FIGURE 4** Loading plots from principal component analysis of compounds in Las Vegas Wash

![Diagram showing loading plots from principal component analysis with scatterplots for PC1, PC2, PC3, PC4, and PC5. Plotting symbols are the group number of individual analytes from cluster analysis. Axes are correlation loadings for the corresponding principal component.]
(Eaton et al. 2010, Barnes et al. 2002, Kolpin et al. 2002). Diuron, an herbicide frequently detected in US surface waters and one of the most-used herbicides in California (Bergin et al. 2008, Eaton et al. 2010), showed the highest maximum concentration of compounds in the Santa Ana River.

The dendrogram of CA resulted in six clusters (Figure 3, part D) with relatively large distances between them. Cluster 2, composed of gemfibrozil, caffeine, and ibuprofen, and cluster 5, composed of carbamazepine, primidone, and Dilantin had the highest internal similarity. Caffeine and ibuprofen are frequently found in US waters (>20%) (Eaton et al. 2010) and are often active ingredients of over-the-counter pain relievers (e.g., Anacin, Motrin, Advil). Cluster 4 consisted of the herbicides atrazine and atrazine-desisopropyl with high similarity, together with the insect repellent, DEET. These compounds exhibit seasonal patterns, with high values in late spring and early summer and low

---

**FIGURE 5** Scatter plots of clustered compounds in Las Vegas Wash

**A**

![Scatter plot A](image1)

**B**

![Scatter plot B](image2)

**C**

![Scatter plot C](image3)

**D**

![Scatter plot D](image4)

**NO₃-N**—nitrate as nitrogen

Units are standard deviations from the mean of the scaled, transformed concentration.
values in autumn and early spring. Cluster 1 comprised the herbicides Diuron and simazine. Cluster 5 was composed of the anticonvulsants carbamazepine, Dilantin, and primidone. Clusters 3 and 6 contained compounds of different uses; cluster 3 paired the pain reliever acetaminophen with the antifungal agent triclosan, while cluster 6 contained antibiotics sulfamethoxazole and azithromycin as well as the biochemical reducing agent TCEP.

PCA results (Figure 10) reflected many of the clusters from CA results. PCs 1, 2, 3, 4, and 5 distinguished clusters 5, 2, 6, 4, and 3, respectively. Clusters 1 and 6 were on opposite sides of PC3, in agreement with the negative correlations between compounds in these groups.

Pairwise scatter plots of compounds in Santa Ana River clusters generally show collinearity, although not necessarily strongly enough to be used reliably as indicators (Figure 11). Compounds in cluster 2 were sufficiently collinear ($r \geq 0.60$) to be potentially used as indicators of each other (Figure 11, parts A and B), as were members of cluster 5 ($r \geq 0.59$; Figure 11, parts C and D), although the relationship between carbamazepine and Dilantin (Figure 11, part D) was weaker in Santa Ana River ($r = 0.59$) than in AMS ($r = 0.73$).

DISCUSSION

The patterns among EDCs/PPCPs revealed through chemometrics could reflect various factors influenced by their source, transport, and degradation. PPCPs are mainly introduced from point sources and tend to vary widely in their removal from conventional WWTPs. Nonpoint sources tend to be directly introduced to receiving waters without treatment and may...
introduce chemicals such as pesticides and herbicides. Hormones are introduced from both point sources and nonpoint sources. The fate of EDCs/PPCPs varies because of their diverse physical and chemical properties. Some EDCs/PPCPs are degraded through photolysis, biodegradation, and/or sorption, while other compounds are persistent in waters. The rates of degradation are dependent on characteristics of the analyte and environment, leading to the different behaviors across sites and compounds.

Site-specific conditions such as source distribution and hydrologic/climatic conditions resulted in differences in groupings across water systems. For example, the presence and the concentrations of sulfamethoxazole and carbamazepine were more similar in AMS than in either LVW or Santa Ana River. Atrazine and DEET showed co-occurrences in Santa Ana River but not in Lake Mead (LVW and AMS). Gemfibrozil and trimethoprim had co-occurrences in Lake Mead but not in Assabet River. Thus, the approach of clustering compounds by similarity should be made on a site-by-site basis.

Functionally similar compounds had high similarity in their occurrences and concentrations in certain cases. In particular,
anticonvulsants including carbamazepine, propranolol, and Dilantin were highly similar in all water systems in which they were measured. Certain herbicides were paired with high similarity, but not all formed clusters. Hormones were not clustered together in any water system and very infrequently detected at all sampling stations (<40% of samples), likely influencing their low similarity.

Because chemical analyses of EDCs/PPCPs are expensive, a well-designed preliminary monitoring program should include a suite of more easily measurable parameters as potential indicators. Previously suggested wastewater tracers had limited utility as surrogates for EDCs/PPCPs, and this utility varied by different water systems. TDS and conductivity occasionally clustered with certain compounds (in Lake Mead), but nowhere was this relationship strong enough to be considered predictive. It is unlikely that TDS or conductivity would be a suitable surrogate where ambient levels are high, for example, in brackish water or when road runoff containing deicing agents is introduced in winter. Other indicators of wastewater such as nitrate were available only in limited numbers for LVW, and were not available in the other water systems in this study. Among more easily measured water quality parameters, only gadolinium was found to be useful as a surrogate of EDCs/PPCPs in Assabet River. However, the applicability of such surrogates will vary by each water system.

The stochastic imputation method for data below MDL used in this study resulted in conservative estimates of similarity, with infrequently detected compounds tending to receive lower similarity scores than they would if no imputation took place. While the stochastic imputation offers the benefit of incorporating data tuples

---

**FIGURE 8** Loading plots from principal component analysis of compounds in the Assabet River

Plotting symbols are the group number of individual analytes from cluster analysis. Axes are correlation loadings for the corresponding principal component.
containing values below MDLs (as opposed to discarding the data points) and avoids potential biases arising from deterministic imputation methods, it can fail to identify potentially useful relationships if a large portion of data is imputed. The amount of information available based on relationships between compounds is therefore a function of their detection frequency and will improve as analytical methods continue to be refined.

This study was based on the best available data in the United States; further investigation with more data collection in other water systems will complement our findings to discover patterns among other compounds and especially other conventional water quality constituents. Chemometrics and further analysis of time series data rely on complete data sets and are degraded by the presence of missing values. There is no generally accepted rule for minimum data size, but the number of complete data sets without missing values should be greater than the number of compounds from sampling campaigns. It is therefore recommended that sampling campaigns should be conducted at a

FIGURE 9  Scatter plots of clustered compounds in the Assabet River

A  B

Caffeine  Naproxen

Sulfamethoxazole  Diclofenac

Estrone  Sucralose

TCEP—tris(2-carboxyethyl)phosphine

Units are standard deviations from the mean of the scaled, transformed concentration.
consistent set of stations at a regular time interval. In order to investigate the patterns of compounds in different weather conditions or different hydrological years, the sampling campaigns should be designed to accommodate such requirements to collect sufficient data size.

CONCLUSION

This study illustrates an approach to designing efficient monitoring campaigns for EDCs/PPCPs in surface water systems that are used for drinking water. Chemometric techniques including principal component analysis and cluster analysis revealed underlying co-occurrence patterns in high-dimensional water quality data sets, guiding the selection of surrogate compounds whose measurement can be used as a proxy for other compounds. This can reduce the analytical burden of a monitoring campaign by reducing the number of necessary analyses. Relationships revealed by chemometrics were evaluated using pairwise scatterplots in order to gauge their usefulness as surrogates. While certain co-occurrences aligned with chemical similarities between compounds, the majority did not, providing new information not available to an analysis based solely on chemical properties. The application of this approach to data sets from three different surface water systems revealed differences in co-occurrence behavior between the different data sets, implying that investigations of this type must be undertaken at each individual water system. However, continued research may discover universally applicable insights as more data of this type become available. Such insights could be further used to develop models of the fate and transport of EDCs/PPCPs in surface waters and manage these compounds to protect ecosystem and public health.
ACKNOWLEDGMENT

This study is funded by the Water Research Foundation (#4260). The authors would like to thank the Southern Nevada Water Authority for providing monitoring data of Lake Mead, and Larry Kramer, Soon-mi Kim, and Woo-Jung Choi at the University of Massachusetts for assisting sampling, chemical analysis, and data analysis for the Assabet River.

ENDNOTE

1 R Foundation for Statistical Computing, Wien, Austria

ABOUT THE AUTHORS

Mark Hagemann is a fourth-year PhD candidate in the Department of Civil Environmental Engineering at the University of Massachusetts, Amherst. He received his BA degree in geology from Carleton College, Northfield, Minn. Hagemann is a recipient of the Raymond A. Noga and Karen Murphy Noga Fellowship from the University of Massachusetts and the Duncan Stewart Fellowship from Carleton.
College. Minji Park is a researcher at the Han River Environment Research Center, Yangpyeong-gun, Gyeonggi-do, Korea. Varun Srinivasan is a PhD candidate, David A. Reckhow is a professor in the Department of Civil and Environmental Engineering, and Michael Lavine is a professor in the Department of Mathematics and Statistics, all at the University of Massachusetts, Amherst. Erik Rosenfeldt is the director of drinking water process technology at Hazen and Sawyer, Richmond, Va. Benjamin D. Stanford is the director of applied research at Hazen & Sawyer, Raleigh, N.C. Mi-Hyun Park (to whom correspondence may be addressed) is an assistant professor at the University of Massachusetts, 130 Natural Resources Rd., Amherst, MA 01003 USA; mtpark@ecs.umass.edu.

PEER REVIEW

Date of submission: 12/19/2014
Date of acceptance: 10/26/2015

REFERENCES


