Cluster Analysis as a Tool for Characterizing Water Reclamation Plant Influent

Junjie Zhu[a]; Javier Segovia[b]; Paul R. Anderson[a]
[a] *Department of Civil, Architectural and Environmental Engineering, Illinois Institute of Technology, Chicago, IL, USA*
[b] *School of Computer Science, Universidad Politecnica de Madrid, Boadilla del Monte, 28660, Madrid, Spain*

Abstract: Water reclamation plants (WRP) face the challenging task of simultaneously managing fluctuating influent conditions and satisfying effluent discharge requirements. To better prepare WRP operators for this task, we combined *k*-means cluster analysis with cross-tabulation analysis to develop an influent classification system for the Calumet WRP in Chicago, IL. We considered weather and influent composition characteristics to identify 25 clusters, nine of which were significant (99% confidence level). For example, dry weather with mid-range temperature characteristics are common after wet weather days, and these conditions typically present low influent concentrations. In addition, compared to cold-weather characteristics, warm-weather flows are more likely to have large precipitation events and more variation in influent quality. The duration of storm events is also important for planning. Large storms during warm weather feature relatively low influent concentrations and have a high probability of lasting only one day, whereas warm and dry-weather conditions that bring relatively high influent concentrations have a high probability of lasting more than one day. We believe the approach used in this study can be replicated and will provide useful risk management information at other WRPs.

Keywords: *k*-means cluster analysis, cross-tabulation analysis, water reclamation plant, influent scenario.

1. Introduction

This study is part of a collaborative project between the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC) and the Illinois Institute of Technology (IIT) on applications of cyber-physical systems (CPS) to the Chicago Area Waterway System (CAWS). Ultimately we hope to develop an agent-based intelligent sensor network system that integrates data from on-line sensors with predictions developed from historical data. The overall objective is to minimize energy demands and improve control of nutrient loading (CPS 2010). In this study, cluster analysis and cross-tabulation analysis were applied to data from the MWRDGC Calumet WRP to characterize its historical influent scenarios. Results from this study are part of the information required for intelligent real-time process control.

© 2014 *by a Junjie Zhu, Javier Segovia, Paul R. Anderson Printed in USA.*

2. Background Information

2.1. THE CALUMET WRP

The Calumet WRP, which began operations in 1922, now serves more than one million people in the Chicago area. Wastewater influent to the plant comes from municipal sources and tunnel and reservoir plan (TARP) flow (TARP is the system for managing combined sewer overflows in Chicago). Based on historical data from 2002-2011 (MWRDGC, 2013), average influent conditions include 262 MGD flow, 11 mg NH_3-N/L, 74 mg $CBOD_5$/L, and 136 mg SS /L. Precipitation events and patterns of industrial water use can lead to influent characteristics that are substantially different from these average conditions. For example, the standard deviation for the influent flowrate is 90 MGD; the ammonia concentration can be as large as 26 mg/L, and the $CBOD_5$ can be lower than 15 mg/L. The purpose of the cluster analysis is to better understand the types of perturbations that affect the influent.

2.1. COMMON CLUSTERING METHODS

The most common clustering methods are hierarchical and *k*-means (Mooi and Sarstedt 2011). Hierarchical clustering typically applies algorithms to measure the Euclidean distances among different observations, *k*-means methods use a centroid-based approach to minimize within-cluster variation, and each method has advantages and disadvantages. Hierarchical methods provide higher quality on likelihood classification with small datasets, whereas the *k*-means method is less sensitive to outliers and typically more efficient at processing large (> 500) sample sizes (Abbas 2008; Mooi and Sarstedt 2011).

The literature suggests that the hierarchical method has been used more frequently to classify water quality data. For example, Astel et al. (2007) applied the hierarchical method to classify 14 common chemical indicators at 24 on the Struma River in Bulgaria, and identified three groups of indicators and four clusters of sites. Kamble and Vijay (2011) used the hierarchical method to classify 17 coastal locations near Mumbai, India, and identified three clusters based on water quality. Mukhopadhay et al. (2011) also employed the hierarchical method to investigate groundwater contamination in 29 wells, and identified five clusters based on microbiological and chemical parameters. Similarly, Nnane et al. (2011) used the hierarchical method to classify 14 sites on the Ouse River, England, and identified six clusters based on microbial water quality.

Examples where *k*-means methods were used typically involve larger sample sizes. For example, Albazzaz et al. (2005) used the *k*-means method as one of several statistical tools to identify unusual operations at a WRP, and identified 17 abnormal cases out of 527 operating days. Brena et al. (2005) analyzed 23 groundwater wells in the Aquifer Punta Espinillo (Uruguay) based on either α-naphthol concentrations (18 wells) or hydrogeological data (22 wells) and identified three clusters. Akbar et al. (2011) combined principal component analysis and the *k*-means method to analyze water quality for 18 lakes in Alberta based on 11 years of historical data. De la Vega et al. (2012) applied a *k*-means method to divide historical data (sample size = 2000) on wet- or dry-weather conditions, to investigate oxidation–reduction potential (ORP) and D.O. profiles along aeration basins in a biological treatment process. They reported that nitrification and denitrification processes were inhibited during wet weather due to lower ammonia and higher nitrate and D.O. concentrations.

The historical data for our study involved a relatively large sample size (> 3000), so we used the *k*-means method to identify influent scenarios. One challenge of *k*-means cluster analysis is that the number of

clusters has to be specified in advance. Typically, several different values of *k* are considered and the domain expert selects the partition that appears most meaningful (Jain 2010). Because we wanted to identify important clusters from a process control perspective, we adopted two criteria:

- Each cluster should represent distinctive characteristics.

- Data included in a cluster should cover more than 5% of the records. (Low population (< 5%) clusters represent exceptional scenarios that will be the subject of future work.)

3. Research Methods

3.1. *K*-MEANS CLUSTER ANALYSIS

We defined scenarios based on weather and influent composition. The four factors that make up the weather scenarios are the raw influent flowrate, the flowrate associated with TARP, the amount of precipitation, and the water temperature. The *composition scenarios* were defined based on five water quality parameters: SS, $CBOD_5$, TKN, VSS/SS, and NH_3-N/TKN. These parameters were selected based on the available data for representing influent characteristics, and because these parameters are important in the WRP simulation model.

Before clustering, we screened the data to identify missing elements, remove outliers, and normalize the values. Where part of a data element was missing, the entire day's data was removed from the assessment. Extreme outliers were removed by following a simplified distance-based outlier detection method described by Angiulli et al. (2006). In that approach, the steps are to calculate parameter mean value, arrange data in an ordered list from lowest to highest values, and calculate differences between adjacent values in the list. If a difference value exceeds the mean value, values from that point to the end of the list are outliers and excluded from further study. This simple heuristic detects a gap in the tails of the continuum of values, which could be due to error or the gap could identify extreme values that we consider as out of scope of this study.

The final step before clustering was value normalization. Influent parameters were converted to *Z*-scores () so that different parameters could be treated equally for cluster analysis. Specifically, the average value ($\bar{x}$) and standard deviation (*s*) of influent variable were calculated, then the original data (*x*) were normalized using $\bar{x}$ and *s* as shown in .

$$Z - score(x) = \frac{x-\bar{x}}{s} \text{Z-score(x)} = \frac{\text{x-}\bar{\text{x}}}{\text{s}} \quad (1)$$

3.2. CROSS-TABULATION ANALYSIS

Clusters were subjected to cross-tabulation analysis to identify associations between weather and composition conditions using independence and homogeneity tests. For example, if there are "*I*" clusters classified from weather scenarios and "*J*" clusters classified from composition scenarios, there will be a

total of "*I*×*J*" possible combinations. The Pearson chi-square score (χ^2) was used as the criterion to test the independence at the 99% confidence level ($\alpha = 0.01$) with *(I-1)×(J-1)* degrees of freedom.

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

$$E_{ij} = NP_wP_c \quad (3)$$

$$P_w = \frac{number\ of\ days\ with\ W_i}{N} \quad (4)$$

$$P_c = \frac{number\ of\ days\ with\ C_j}{N} \quad (5)$$

The test statistic depends on observed days (O_{ij}) and expected days (E_{ij}) for a joint cluster (a day simultaneously classified as weather cluster W_i and composition cluster C_j identified here as W_iC_j, where $1 \le i \le I$ and $1 \le j \le J$). Specifically, O_{ij} is the number of observed days in a joint cluster; E_d is the number of expected days in this joint cluster considering that W_i and C_j are independent ($E_{ij} = NP_wP_c$). In that case the number of expected days for a joint cluster is the number of records, *N*, multiplied by P_w the probability of weather cluster W_i occurring ($P_w = \frac{number\ of\ days\ with\ W_i}{N}$), and by P_c the probability of composition cluster C_j occurring ($P_c = \frac{number\ of\ days\ with\ C_j}{N}$).

The test of homogeneity identifies homogeneous distributions for weather and composition clusters. Significant scenarios were identified from possible combinations based on the criterion of adjusted standardized residual (ASR)

$$ASR = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij} \times (1 - P_w) \times (1 - P_c)}} \quad (6)$$

The ASR value depends on observed days (O_{ij}), expected days (E_{ij}), row proportion (P_w), and column proportion (P_c) (Haberman 1973; Agresti 2007). ASR measures the association between variables. For example, at the 99% confidence level ($\alpha = 0.01$) the critical value is $Z_{\alpha/2} = \pm 2.58$, and the level of association between weather and composition clusters can be described as follows:

- ASR ≥ 2.58 identifies a statistically significant association between weather and composition clusters. If the ASR for influent scenario W_iC_j falls in this range, we reject the hypothesis that clusters W_i and C_j are independent. They occur together more frequently than by chance, the observed days are significantly more than the expected days, and therefore W_i and C_j are associated. The higher the value of ASR, the stronger the association.

- -2.58 < ASR < 2.58 means that we cannot reject the hypothesis that the influent scenario happens by chance.

– ASR $\leq$ -2.58 identifies a significant negative association between W_i and C_j; it is unusual for clusters W_i and C_j to occur simultaneously.

In this study, significant scenarios (ASR $\geq$ 2.58) were emphasized because they represented strong patterns of influent scenarios at the Calumet WRP.

4. Quality Assurance and Quality Control (QA/QC)

Historical data for the Calumet WRP (MWRDGC 2013) were provided by MWRDGC. Because these data must satisfy the QA/QC requirements of the USEPA, there was no additional QA/QC assessment before the analysis in this study.

5. Results

5.1. PRELIMINARY TREATMENT

Twenty-three outliers (0.7% of the data) were identified; no outliers were detected in the temperature, flow, or NH_3-N/TKN data; the precipitation data had the largest number of outliers (Table 1). There were 453 days with missing data and 23 days with outliers, leaving 3176 days for the cluster analysis.

Table I. Average Values of influent conditions, number and critical values of outliers.

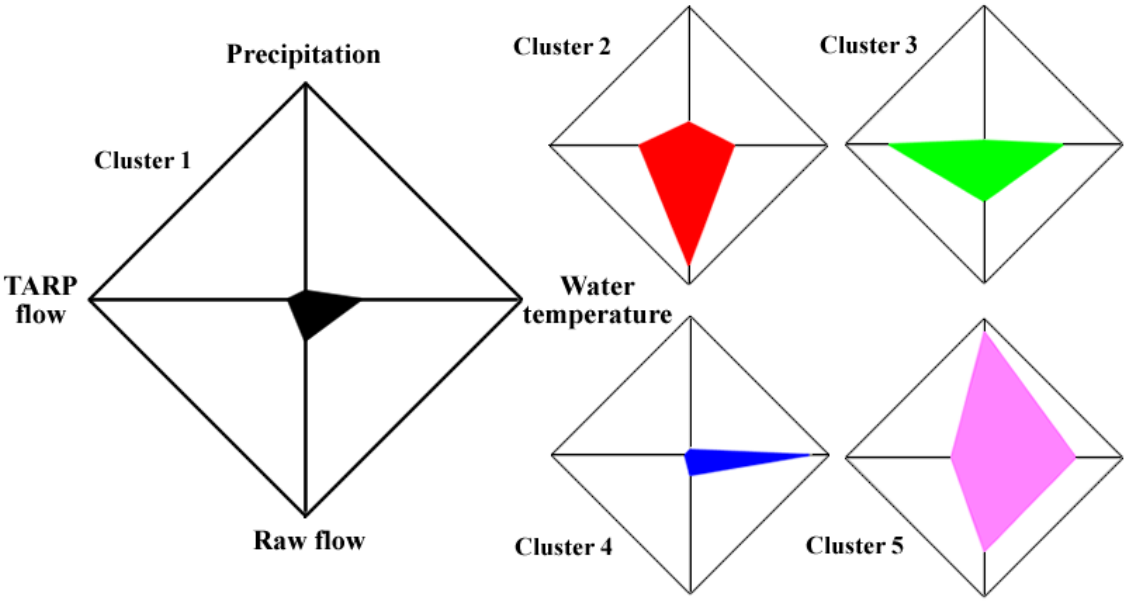
		Treated data		Outliers	
	Unit	Mean	Standard deviation	Number	Critical value
Precipitation	in	0.10	0.24	12	> 2
Water temperature	F^o	58.1	10.5	0	
Raw flow	MGD	206.8	56.8	0	
TARP flow	MGD	54.9	51.6	0	
SS	mg/L	135.6	92.5	4	> 1200
TKN	mg/L	20.1	7.2	2	> 320
$CBOD_5$	mg/L	74.3	30.5	2	> 390
VSS/SS	-	0.69	0.09	3	> 1
NH_3-N/TKN	-	0.55	0.09	0	

5.2. HISTORICAL CLUSTERS

5.2.1. *Weather clusters*

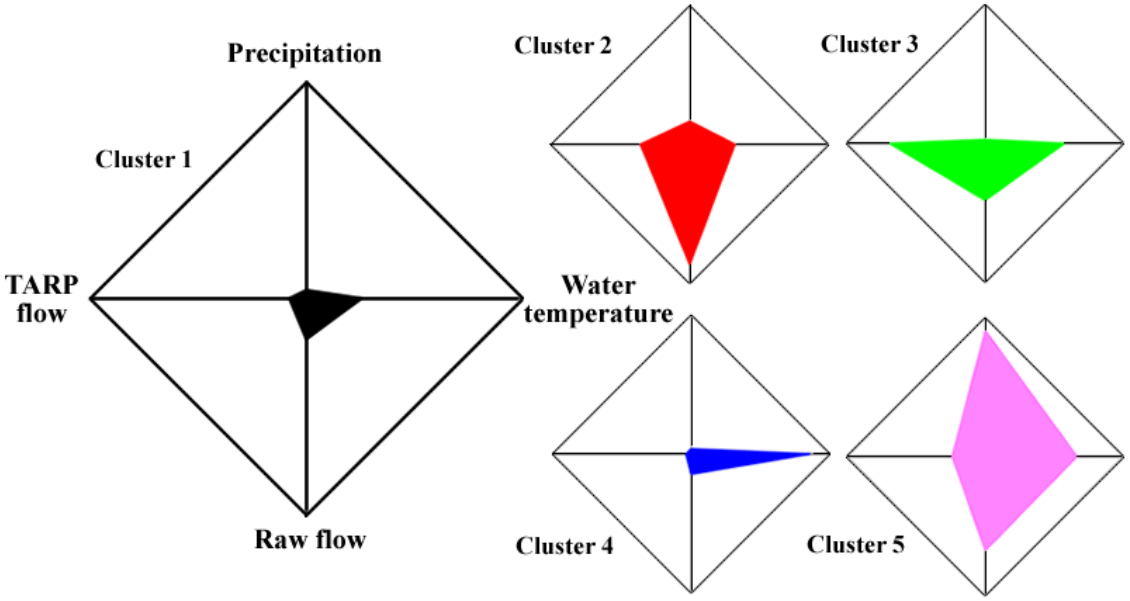
We explored *k*-means models using a range of three to seven clusters. The model with three clusters included two clusters representing relatively dry weather (average precipitation $\approx$ 0) and one very large

storm cluster (average ≈ 0.83 in). The major difference between these two dry-weather clusters was TARP flow; one had relatively high flows (≈ 121 MGD) and the other had low flows (≈ 27 MGD). The model with four clusters resulted from the division of dry weather days with low TARP flows into two clusters: relatively low temperatures (≈ 48 ^{o}F) and relatively high temperatures (≈ 66 ^{o}F). The model with five clusters provided one more cluster, representative of small storms (≈ 0.16 in). This additional cluster was valuable because it provided information for common storm events between the two extreme conditions (dry weather and very large storms). The models with six or more clusters were rejected because the additional clusters either had similar characteristics or accounted for less than 4% of the data.



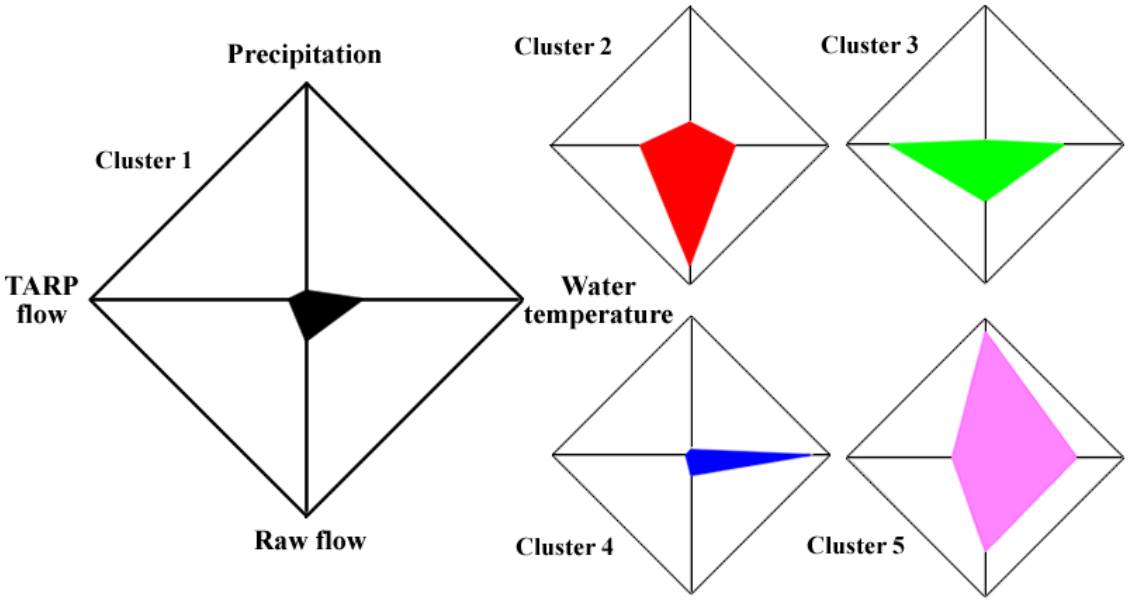
Parameter	Precipitation (in)	Water temperature (^{o}F)	Raw flow (MGD)	TARP flow (MGD)
Range of the axis	0 ~ 1	40 ~ 70	150 ~ 350	20 ~ 200

The final classification for this study is based on five weather clusters



Parameter	Precipitation (in)	Water temperature (^{o}F)	Raw flow (MGD)	TARP flow (MGD)
Range of the axis	0 ~ 1	40 ~ 70	150 ~ 350	20 ~ 200

Figure 1 and Table 2). Cluster 1 includes relatively low values for all the weather parameters. Cluster 2 includes days with little precipitation, high raw and TARP flows, and relatively low temperatures. Cluster 3 represents dry weather in the median range of temperatures, but the TARP flow is very high, indicating that most days in cluster 3 are dry-weather days following wet-weather days. Cluster 4 includes the largest sample size, representing dry weather and relatively high temperatures. Cluster 5 is the smallest cluster; it includes days with very large storms and high temperatures.



Parameter	Precipitation (in)	Water temperature (°F)	Raw flow (MGD)	TARP flow (MGD)
Range of the axis	0 ~ 1	40 ~ 70	150 ~ 350	20 ~ 200

Figure 1. Radar charts for five weather clusters (average values); the table shows scales of four weather parameters in the radar charts. Cluster 1 is presented larger so the axes can be labeled.

Table II. Average conditions in five weather clusters and their observed counts.

Cluster	Precipitation	Water temperature	Raw flow	TARP flow	Observed counts
	in	°F	MGD	MGD	day
1	0.04	48	188	34	822
2	0.16	50	322	84	227
3	0.03	57	232	145	538
4	0.04	66	179	27	1398
5	0.90	60	284	63	191

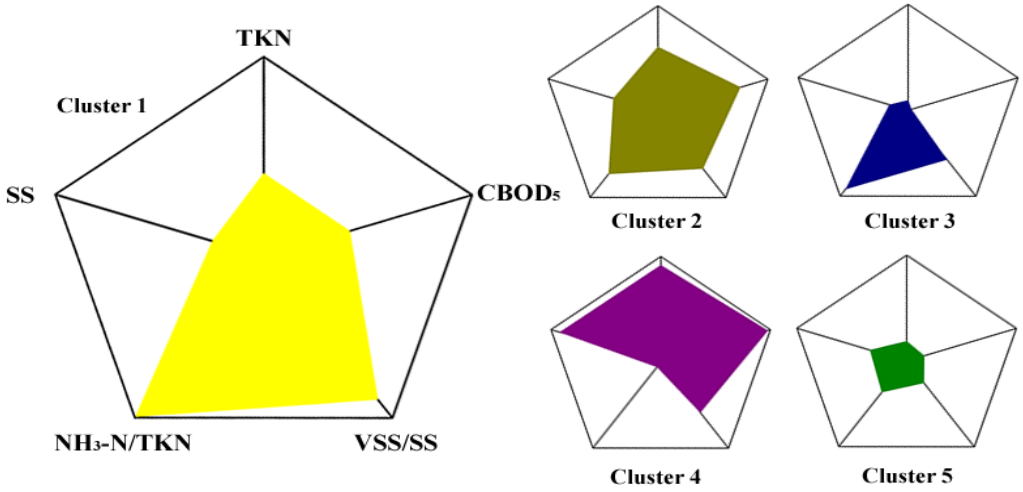
5.2.2. *Composition clusters*

A similar trail-and-error approach was used to define composition clusters. Three or four clusters may not provide enough detail and six or more clusters begin to yield clusters with very small sample sizes (cluster 6 included 1.5% of the data). Therefore, five clusters were classified.

Among these five composition clusters, cluster 1, which includes the most cases, represents relatively medium-to-low influent concentrations but high concentration ratios (

Figure 2 and Table 3). Cluster 2 includes days with average influent quality. Most days in cluster 3 have very low influent concentrations and medium-to-high concentration ratios. Clusters 4 and 5 include days

with extreme conditions, but they are different types of extremes. The smallest cluster is number 4, which has very high concentrations (high concentrations can result in average (VSS/SS) or low (NH_3-N/TKN) ratios). The extreme conditions in cluster 5 represent relatively low concentrations.



Parameter	SS (mg/L)	TKN(mg/L)	CBOD5 (mg/L)	VSS/SS	NH3-N/TKN
Range of the axis	0 ~ 400	10 ~ 35	40 ~ 120	0.5 ~ 0.8	0.4 ~ 0.6

Figure 2. Radar charts of five composition clusters (average values). The table shows scales of five composition parameters in the radar charts.

Table III. Average conditions in five composition clusters and their observed counts.

Cluster	SS	TKN	$CBOD_5$	VSS/SS	NH_3-N/TKN	Observed counts
	mg/L	mg N/L	mg O_2/L	-	-	day
1	98	20	73	0.77	0.60	919
2	162	25	100	0.70	0.54	851
3	67	12	42	0.67	0.58	626
4	368	33	118	0.68	0.41	241
5	133	15	53	0.58	0.47	539

5.3. SIGNIFICANT INFLUENT SCENARIOS

The Pearson chi-square test was used to verify independence between weather clusters and composition clusters. A calculated test statistic, χ^2 = 1787 with a *P*-value substantially smaller than the critical value 0.01, indicates that the null hypothesis can be rejected and we can conclude that there is a significant association between weather and composition scenarios.

With five weather clusters and five composition clusters, there are 25 possible scenarios. In that group there is significant positive association with nine influent scenarios (red rectangles), three scenarios that

lack significant association (orange rectangles), and 13 scenarios (dark blue rectangles) that exhibit a significant negative association (Figure 3). The nine significant influent scenarios cover 2403 days (about 76%), the three scenarios without significant association cover 295 days (9%), and the 13 scenarios with significant negative association cover 478 days (15%). Among these last 13 scenarios, the combination between weather cluster 4 and composition cluster 3 (W4C3) has the lowest ASR value (-18.3). For dry weather days with relatively high temperature (weather cluster 4), it is very unlikely that the Calumet WRP will experience low influent concentrations (composition cluster 3).

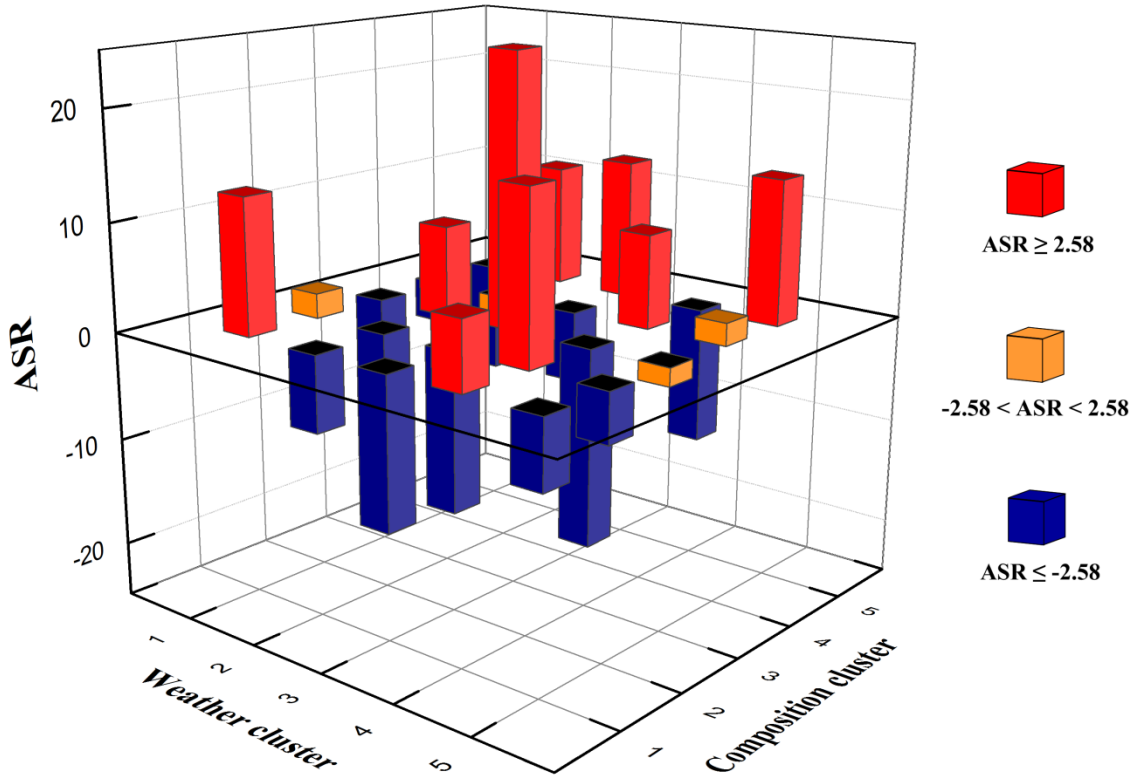


Figure 3. Adjusted standardized residual (ASR) of 25 possible influent scenarios, which combine five weather clusters and five composition clusters, based on the chi-square test.

Among the nine significant influent scenarios, scenario W3C3 has the highest ASR value (25.0), indicating that influent concentrations are usually low (composition cluster 3) for dry-weather days with median temperature and just after (a) wet-weather day(s) (weather cluster 3). Scenario W4C1 has the smallest ASR value; the observed number of days (487) is slightly more than expected (407). Scenario W4C2 includes the most observed days (566), which represents dry, warm weather with relatively medium influent quality. The smallest number of expected days is in scenario W5C5 (33), which applies to large storms in warm weather with low concentrations.

Influent conditions are composed of nine variables, so we used a parallel coordinates plot to exhibit characteristics of significant scenarios. An advantage of such a plot is that the distribution of each parameter and the relationships between two parameters can be read clearly (Inselberg 1997). Three

scenarios (W1C1, W2C3, and W4C4) were selected to show distributions of daily data for all influent parameters (Figure 4). The W1C1 scenario, which includes days in dry and cold weather with low concentrations and high concentration ratios, has relatively small coefficients of variation for all parameters. For example, water temperature ranges from 41 to 57 ^{o}F with the smallest SD of 4 ^{o}F in all nine scenarios; SS concentration ranges from about 36 to 219 mg/L with a SD of 28 mg/L. When most of the lines between the axes are parallel, it's an indication of a strong linear relationship between two parameters (Inselberg 1997). With the exception of water temperature, neighbor parameters in the W1C1 scenario follow a roughly linear relationship. The W2C3 scenario represents days with small storms and cold weather with very low concentrations and medium-to-high concentration ratios, but water temperature and TARP flow have large variations in this scenario. For example, although on most days (72%) the water temperature is below average (48 ^{o}F), the maximum value could reach 70 ^{o}F with a SD of 8 ^{o}F. Moreover, the TARP flow ranges from 17 to 170 MGD with a SD of 33 MGD. The W4C4 scenario, which represents dry, warm weather with high concentrations, exhibits clear, linear relationships among neighbor parameters, especially water temperature, precipitation, and raw flow. Compared to other scenarios, influent concentrations in this scenario cover wider ranges. For example, the SS concentration ranges from 126 to 863 mg/L with the biggest SD (115 mg/L) among the nine scenarios, the TKN concentration ranges from 19 to 65 mg/L also with the biggest SD (7 mg/L), and the $CBOD_5$ concentration also has the biggest SD (26 mg/L). The other six scenarios (not shown in*Figure*) all have distinct characteristics. For example, raw (SD ≈ 59 MGD) and TARP (≈ 39 MGD) flows both have the biggest variations in scenario W2C5; precipitation (≈ 0.34 in) and water temperature (≈ 11 ^{o}F) have the biggest deviation in scenario W5C5.

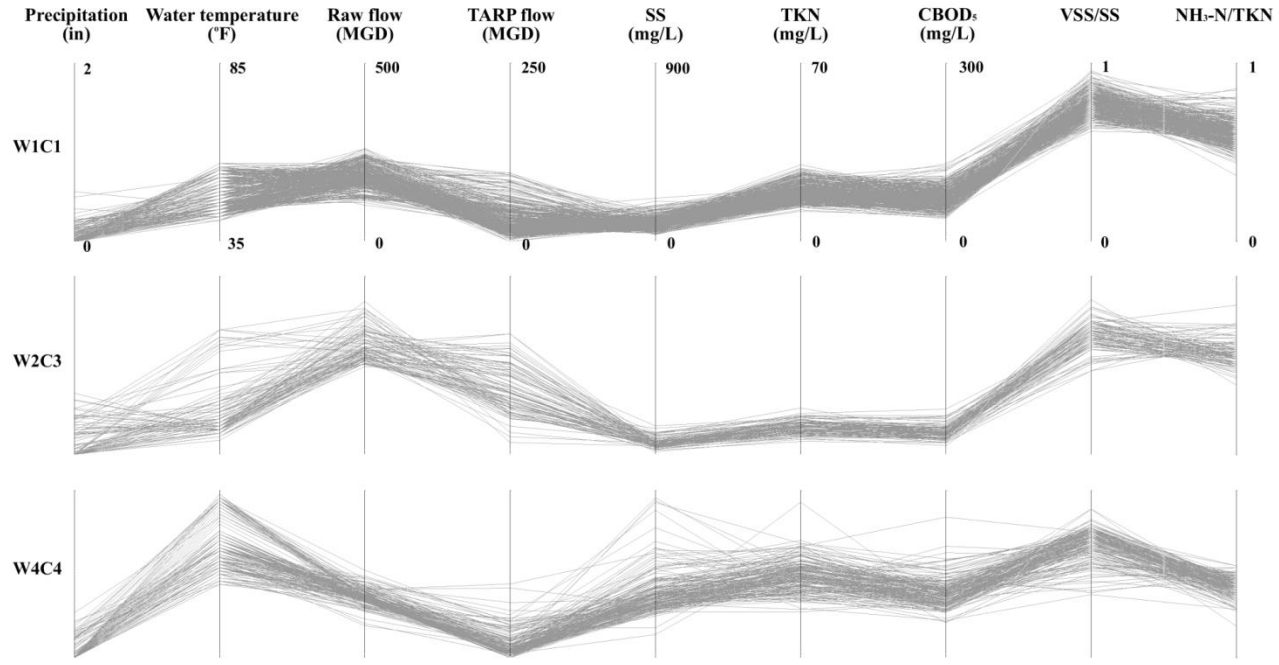


Figure 4. Parallel coordinates plot of selected significant influent scenarios. The ranges of each influent variable are determined based on their minimum and maximum values.

6. Discussion and Conclusion

Results of the analysis identify specific important scenarios that can be consulted for corresponding control strategies in the plant. For example, based on observed or predicted weather conditions, the corresponding range in composition conditions can be predicted from the associations between weather and composition clusters. Weather cluster 1 (W1) has a significant association only with composition cluster 1 (C1). As a result, on days with dry weather and cold temperatures the expected influent characteristics include relatively medium-to-low SS, TKN, and $CBOD_5$ concentrations and relatively high VSS/SS and NH_3-N/TKN ratios (medium VSS and high NH_3-N concentrations). Similarly, W5 is only significantly associated with C5, which means that low TKN and $CBOD_5$ concentrations, medium SS concentration, and low concentration ratios should be expected on days with large storms. In contrast with the unique weather-composition clusters of scenarios W1C1 and W5C5, other weather clusters can have more than two possible composition clusters. For example, W2 and W3 both have significant associations with either C3 or C5; and W4 can be associated with C1, C2, or C4. ASR values can be used to evaluate the probabilities of these scenarios. Probabilities for conditions represented by C3 and C5 are similar when weather conditions represented by scenario W2 exist because they have close ASR values (8.3 and 11.2). However, when weather conditions represented by scenario W3 occur, scenario C3 is more likely to happen because relative to scenario C5 it has a much larger ASR value (25 versus 12.8).

For planning WRP operations, it is also worthwhile to examine the duration times of the scenarios. Among the nine significant scenarios, W5C5 has the highest probability (85%) of lasting only one day (Figure 5). Alternatively, when there is a large storm with low concentrations (a scenario with 85% probability of lasting only one day), the probability of that storm continuing into the second day but ending before the third day is only about 12%. In contrast, scenarios that include W4 (W4C1, W4C2, and W4C4) last only one day and have relatively low probabilities, perhaps because dry, warm weather (W4) is the most common weather condition (44%). For example, scenario W4C4 has a low probability (34%) of one-day duration, but it has the longest duration (24 days), which is more than twice the maximum of any other scenarios (11 days for scenario W4C1). One-day duration scenarios that include C1 (W1C1 and W4C1) also have relatively low probabilities, but those involving C5 (W2C5, W3C5, and W5C5) are much more likely. The highest probability of a two-day duration event occurs with scenario W3C5 (30%), and the lowest probability is for W4C4 (12%). Because W1C1 has relatively lower probabilities for one- or two-day duration events, it has the highest probability of lasting longer than two-days (57%). The most unlikely scenarios with long duration times (> 2 days) are W5C5 (3%) and W2C5 (6%).

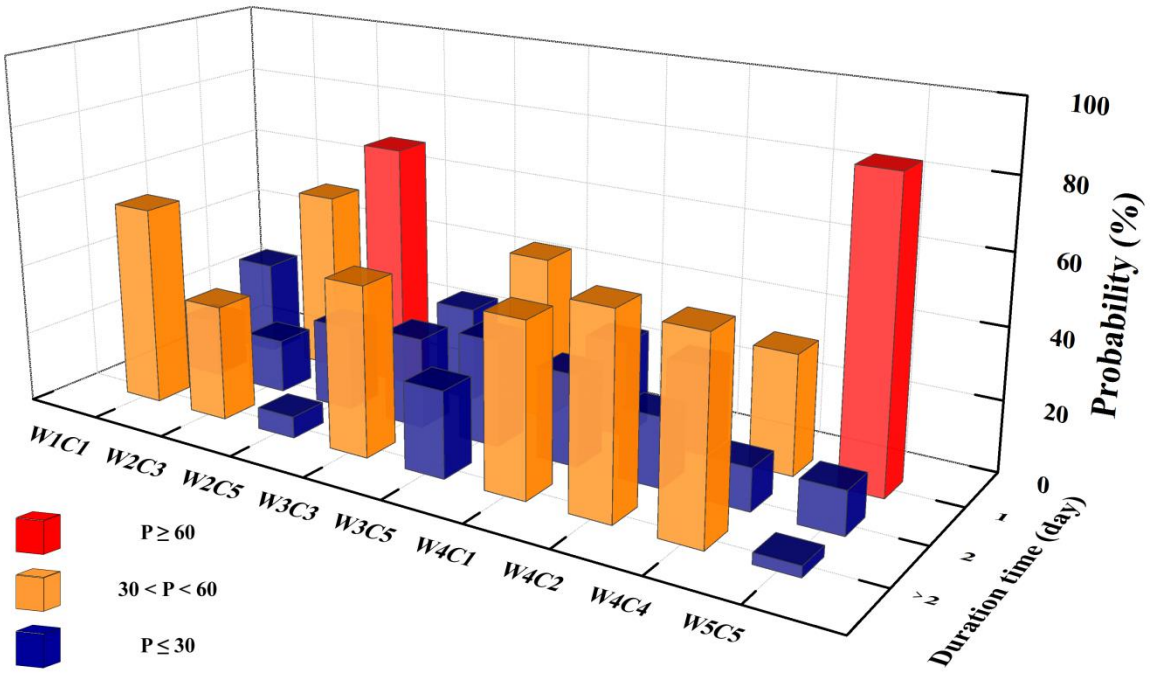


Figure 5. Nine influent scenarios and their probabilities of lasting one day, two days, and more than two days.

Information about influent scenarios can provide useful information for WRP operations; this study provides an assessment of significant scenarios at the Calumet WRP that we will use in future CPS applications. Cluster analysis and cross-tabulation were applied to characterize the influent based on ten years of historical data. Results from this study can be summarized as follows:

- Five weather clusters and five composition clusters were classified; nine significant combined influent scenarios were identified, accounting for 76% of the types of influent conditions.

- Two scenarios, W3C3 and W4C2, deserve special note. Scenario W3C3 (dry weather, median temperature conditions with low influent concentrations that immediately follow a wet weather day) has the highest ASR value. Scenario W4C2 (dry, warm weather with relatively medium influent quality) is the most frequent scenario and it has the second highest ASR value. Therefore, in planning for efficient WRP operations, these scenarios should be emphasized.

- Warm weather conditions mean more variation in influent quality. Relative to cold weather storms, warm weather storms are more likely to be large storms.

- For dry, cold-weather conditions (W1), the composition of the influent is more likely to be low TKN, $CBOD_5$, and SS concentrations and high NH_3-N/TKN and VSS/SS ratios (C1). For warm weather conditions with a large storm (W5), the composition of the influent is more likely to be low concentrations and concentration ratios (C5). The other scenarios exhibit a greater variety of composition conditions.

- Scenario W5C5 (warm weather with large storms and low influent concentrations) has a high probability (85%) of lasting only one day; scenario W1C1 (dry cold weather with low concentrations and high concentration ratios) has the highest probability (57%) of lasting more than two day.

7. Future Work

We expect there will be two directions for future work:

- A WRP process model will be developed to simulate the effects of the important scenarios identified in this study. Results from those simulations could improve our understanding of how to balance process resilience and possible energy savings.

- Although the other 16 scenarios identified in this study were not statistically significant, they did occur. Further study of those scenarios could help to understand their potential risks.

Acknowledgements

This study is part of project that is funded by the National Science Foundation (NSF) (Award Number: 1035894) in collaboration with the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC). The authors wish to thank Dr. Catherine O'Connor and Judith Moran, MWRDGC; for providing the data.

References

Abbas, O. A. (2008). Comparisons between data clustering algorithms. The International Arab Journal of information technology, 5, 3, 320-325.

Agresti, A. (2007). An Introduction to Categorical Data Analysis. The second edition. John Wiley & Sons. ISBN: 978-0471226185.

Akbar, T. A., Hassan, Q. K., & Achari, G. (2011). A methodology for clustering lakes in Alberta on the basis of water quality parameters. Clean – Soil, Air, Water, 39, 10, 916–924.

Albazzaz, H., Wang, X. Z., & Marhoon, F. (2005). Multidimensional visualisation for process historical data analysis: a comparative study with multivariate statistical process control. J. Process Contr., 15, 285–294.

Angiulli, F., Basta, S., & Pizzuti, C. (2006). Distance-based detection and prediction of outliers. Knowledge and Data Engineering, IEEE Transactions on, 18, 2, 145-160.

Astel, A., Tsakovskib, S., Barbieric, P., & Simeonovb, V. (2007). Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Res., 41, 4566–4578.

Brena, B. M., Arellano, L., Rufo, C., Last, M. S., Montano, J., Cerni, E. E., Gonzalez-sapienza, G., & Last, J. A. (2005). ELISA as an affordable methodology for monitoring groundwater contamination by pesticides in low-income countries. Environ. Sci. Technol., 39, 3896-3903.

CPS. (2010). CPS project on managing loosely coupled networked control systems with external disturbances: Wastewater processing. National Science Foundation award number: 1035894.
URL: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1035894
Last access: Feb. 10, 2014.

De la Vega, P. T. M., de Salazar, W. M., Jaramillo, M. A., & Cros, J. (2012). New contributions to the ORP & DO time profile characterization to improve biological nutrient removal. Bioresour. Technol., 114, 160–167.

Grieu, S., Traore, A., Polit, M., & Colprim, J. (2005). Prediction of parameters characterizing the state of a pollution removal biologic process. Eng. Appl. Artif. Intell., 18, 559–573.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. Biometrics, 29, 205-220.

Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. Neural Inf. Process Syst., 3, 281-288.

IBM. (2013). SPSS statistical (Statistical Package for the Social Sciences), the latest version 22.0 was released on August, 2013, developed by IBM (International Business Machines Corp.).
Website: http://www-01.ibm.com/software/analytics/spss/
Last access: Feb. 01, 2014.

Inselberg, A. (1997). Multidimensional detective. Information Visualization. Proceedings., IEEE Symposium on, 100–107.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recogn. Lett., 31, 8, 651-666.

Kamble, S. R. & Vijay, R. (2011). Assessment of water quality using cluster analysis in coastal region of Mumbai, India. Environ. Monit. Assess., 178, 321–332.

Mohamad, I. B., & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. Res. J. Appl. Sci., Eng. Technol., 6, 17, 3299-3303.

Mooi, E., & Sarstedt, M. (2011). A concise guide to market research. The process, data, and methods using IBM SPSS Statistics. Springer-Verlag Berlin Heidelberg. ISBN: 978-3-642-12540-9.

Mukhopadhay, A., Akber, A., & Al-Awadi, E. (2011). Evaluation of urban groundwater contamination from sewage network in Kuwait City. Water Air Soil Pollut., 216, 125–139.

MWRDGC. (2013). Raw data of the Calumet WRP from 2002 to 2011 was provided by the metropolitan water reclamation district of greater Chicago.

Nnane, D. E., Ebdon, J. E., & Taylor, H. D. (2011). Integrated analysis of water quality parameters for cost-effective faecal pollution management in river catchments. Water Res., 45, 2235-2246.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C. J. Mech. Eng. Sci., 219, 103-119.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". J. Comput. Appl. Math., 20, 53–65.