Identification of indispensable components for a better drinking water quality management: Tunis case of study

M. Hassen Baouab and Semia Cherif

ABSTRACT

In order to reduce the number of operations for the assessment of potable water treatment, principal component analysis and hierarchical clustering are applied to large databases of raw and treated water of three treatment plants with various processes. It appears that the measurements can be divided into three clear groups, with a correlation higher than 0.8. The first contains salinity, conductivity, water hardness, calcium, magnesium and chlorides. The second includes turbidity and organic matter. The third includes pH and alkalinity. Despite the disparities in water quality and in all the cases, three parameters were sufficient to represent all the routine measurements: conductivity, turbidity and pH, which can represent the three principal components of the data. It can reduce by two-thirds of the measurement and analysis, dropping from 6,960 to 2,088 analysis annually. The analysis on the principal axes of the individuals, represented by raw and treated water from the three treatment plants, reveals that the quality of the raw water seems more important than the type of treatment process, in the resulting quality of treated water. These results could be generalized and easily adopted by other treatment plants whatever the process. They could offer substantial savings of time, chemicals, electricity and longevity of the devices.

Key words | clustering, drinking water, principal component analysis, treatment, Tunisia

M. Hassen Baouab
Semia Cherif (corresponding author)
UR Chimie des Matériaux et de l'Environnement UR11ES25, ISSBAT,
Université de Tunis El Manar,
Tunis,
Tunisia
E-mail: *semiacherif@yahoo.fr*

INTRODUCTION

A water treatment process commonly comprises the steps of coagulation–flocculation–sedimentation–filtration. Coagulation is a fast mixing process where a coagulant is mixed with raw water. Rapid mixing leads to the formation of a sticky material called floc. Flocculants are added and slowly mixed to form larger flocs. Sedimentation occurs due to the higher specific gravity of the flocs. They settle in the bottom of the sedimentation tank allowing clarified water to go through the filtration process.

Coagulation flocculation is a conventional technique particularly suitable for surface water (Colin *et al.* 1986). This technique is considered the most critical process in drinking water treatment (Lamrini *et al.* 2005). Currently, in many water treatment plants (WTP), process control is generally accomplished through examining the quality of the produced water and adjusting the processes through an operator's own experience (Wu & Lo 2008).

Treating water to a satisfactory quality is important and thus the optimization of this process is surely beneficial. One of the main stages of optimization is the water quality monitoring and assessment. It is obvious that this situation necessarily leads to more frequent acquisitions and analysis of the monitoring criteria, generating a considerable amount of data, and optimization results typically improve when trained on more data. Generally, a number of simple but powerful techniques help to find statistically important factors and thus, improve conclusions on large data (Astel *et al.* 2006). In this context, many studies are based on principal component analysis (PCA), on the clustering analysis, or on both, applied to data on various fields such as food

doi: 10.2166/hydro.2017.070

(Cruz *et al.* 2013), electricity (Segreto *et al.* 2014) and genetics (Yeung & Ruzzo 2001). In the water field, some researchers used statistics (factor analysis techniques such as PCA and clustering), for example, for reservoir flood control optimization (Zhu *et al.* 2016), agricultural water management and irrigation (Valipour 2015a, 2015b, 2016a, 2016b), rainfall forecasting (Valipour 2016c), streamflow forecasting (Kim & Seo 2015) or pipe burst in water distribution systems (Jung *et al.* 2015). However, most of the researchers in the field apply those techniques to chemical data (Colin *et al.* 1986; Rosen & Lennox 2001; Parinet *et al.* 2004; Shrestha *et al.* 2008; Gvozdić *et al.* 2012; Cruz *et al.* 2013), this is called chemometric which aims at extracting as much information as possible through the application of statistics and other mathematical approaches to problems of chemistry (Duarte & Capelo 2006). Moreover, the results of chemometrics play a meaningful role in the assessment of variations in drinking water quality: Astel *et al.* (2006) and Colin *et al.* (1986) focused their research on drinking water quality using chemometrics (correlation matrix, PCA or cluster analysis) to better understand spatial and temporal variations. However, all these studies are focused on disinfection in the potable water process. Lamrini *et al.* (2005) determined the main parameters affecting the coagulant dosage but these studies applied PCA to raw water parameters only and to a unique treatment process. Fabris *et al.* (2015), through PCA, showed that organic and inorganic deposits in water distribution pipes are associated with the input water quality. Kaviarasan *et al.* (2015) used PCA and correlation coefficients to assess raw water quality and identify potable water area and polluted water areas. Bain *et al.* (2014) and Onda *et al.* (2012) used PCA with the aim of assessing multiple countries potable water safety focused on microbiological pollution risks.

Previous research focuses on only one type of water, either raw or treated water, in a specific river or treatment plant with a unique treatment process. However, this article stands out by the application of multiple statistical method (PCA and hierarchical clustering (HCL)) on potable water to both raw and treated water for three types of treatment processes: static settlement tank, lamella separator and pulsator settlement tank. It highlights the assessment parameters that are sufficient to represent all the others during the potable water treatment processes: all the measured parameters are reduced to an essential subset that represents the whole set of information. The findings can be an alternative to the numerous daily measures for the assessment of potable water treatment processes. Nevertheless, the measure of the other parameters or additional ones should continue when there is a need to have a better look at water quality and its variation. It enables WTPs operators to optimize the processes and to save time and cost. It is more convenient for operators and even for decision makers to have a quick preview on water quality in order to take the right decision in time in case of anomaly. A right decision can be taken by adapting the process to the new raw water quality in order to keep a good treated water quality. For example, the reductions of the control parameters to a few that can be controlled instantly (and not after the long hours needed by the jar test inter alia) allows the operators to use more appropriate quantities of process chemicals, and not the quantities corresponding to the day before. It can also raise equipment longevity (e.g. by preventing clogging of filters) and helps to reduce costs, especially in developing countries where very limited financial resources for equipment acquisition and maintenance are provided (Zugarramurdi *et al.* 1995).

DATA AND METHODS

Data

Three drinking WTPs, situated in Tunis and providing drinking waters to Tunis and surrounding areas, were studied. They included the usual main processes that are coagulation, flocculation, sedimentation and filtration. The difference between these plants is their sedimentation process: the first WTP (TP1) has a static settlement tank, the second one (TP2) has a pulsator settlement tank and the third (TP3) has a lamella separator. TP2 and TP3 have the same source of the raw water, different from TP1.

Three databases were used: a $2{,}102 \times 20$ matrix for the first treatment plant (TP1), a $2{,}088 \times 20$ matrix for the second treatment plant (TP2) and a $2{,}098 \times 20$ matrix for the third treatment plant (TP3). In each of these matrix, rows represents the day of sampling, from 2007 to 2012.

Columns represent the instrumental analysis values: ten parameters for raw water and the same number for treated water were measured. They are turbidity, salinity, conductivity, pH, M-alkalinity, water hardness, calcium, magnesium, chlorides, and organic matter. All these parameters were measured, none of them is calculated.

Turbidity was measured with a turbidimeter (±0.01 NTU), salinity and conductivity with a conductivity meter (±0.5%), pH with a pH-meter (±0.005%) and the other parameters with titrimetric determination method (±0.07% due to the use of an automatic burette) using H_2SO_4 for M-alkalinity, EDTA for water hardness, calcium and magnesium, $AgNO_3$ for chlorides, and $KMnO_4$ for organic matter. To avoid human error and guarantee good accuracy, analysis was carried out according to standards (such as NF-EN-ISO7027, NF-EN-ISO8467, NF-EN-ISO9297, NFT 90-016, NFT 80-008) and the measured values were constantly checked by operators and supervisors. Suspected analysis were repeated.

These ten parameters were chosen because potable WTPs require daily or occasional measurements of a number of parameters for the monitoring of the treatment processes. Based on literature, turbidity, conductivity, pH and concentration of some solutes (such as dissolved oxygen, alkalinity, water hardness and chlorine) are the most commonly measured parameters in drinking WTPs worldwide (Valentin 2000; Ratnaweera & Fettig 2015), generally with a daily frequency.

The data were collected over the period 2007–2012 from the three different WTPs in the hope that results could eventually be extrapolated for application to industrial treatment plant management.

Pre-processing of the data was required in order to avoid the effect of different variable scales (Cruz *et al.* 2013). Standardization tends to increase the influence of variables whose variance is small and reduce the influence of those whose variance is large. Furthermore, these procedures eliminate the influence of different units of measurement and render the data dimensionless (Astel *et al.* 2006). For PCA and HCL, Equation (1) was applied to the database for standardization:

$$\text{Xijstd} = \frac{(Xij - \mu j)}{\sigma j} \quad (1)$$

where Xij is the data on the i row and j column of the database matrix, the Xijstd is the corresponding standardized value, μj is the mean and σj the standard deviation of the values of the j column.

Method

The water parameters that significantly influence the variation of water quality during the treatment process are determined with factor analysis through PCA and HCL that are reduction and classification techniques.

PCA analyzes the proximity between variables in terms of correlation and the proximity between individuals in terms of similarities in behavior toward variables (Colin *et al.* 1986). Through PCA, a two-dimensional chart (Figure 1), with the axes representing two principal components (PC or factors) at a time, makes the data clusters easier to spot. In this chart, the i individuals or variables from the database are represented by a cloud of j points. The choice of individuals or variables for i depends on the objectives of the applied PCA. Every j has a portion of the initial information. The chart highlights the groups of correlated variables.

The total number of PC is equal to i. However, the first few components can capture a disproportionate amount of the original information: based on just a few PC we can reduce the observation space. Our choice of the sufficient number of PC will be based on Kaiser Criteria that recommend choosing a PC with eigenvalues higher than 1.0 (Jackson 1993).

The Varimax procedure recommended by Fabrigar *et al.* (1999) and Russell (2002) is used during PCA. It is the most common rotation option and it has generally been regarded as the best orthogonal rotation (Fabrigar *et al.* 1999). This rotation is done in order to have every variable associated to a factor (Abdi 2003). Therefore, this procedure yields results that make it easier to identify each variable with a single factor and make axes easier to interpret.

Cluster analysis is an unsupervised pattern recognition technique wherein the most similar samples are grouped into clusters. It is repeated until all the samples belong to a cluster and the distribution of clusters is represented on a chart called a 'dendrogram'. In this paper, the agglomerative hierarchical cluster analysis is applied according to the Ward method. The latter indicates that the distance between

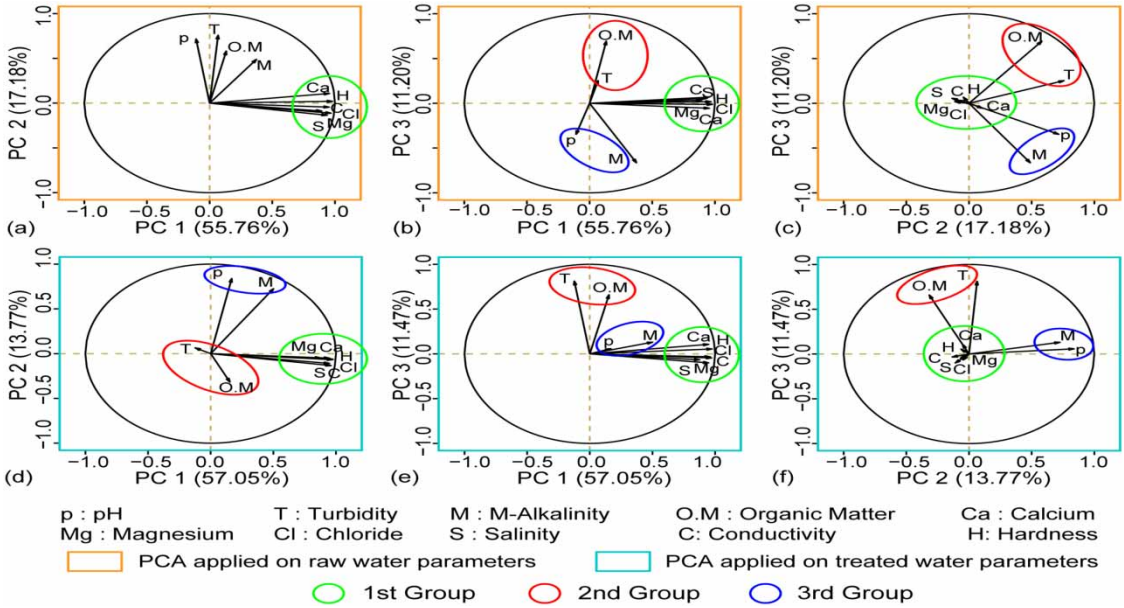

Figure 1 | PCA applied on TP1 data. (a)–(c) Raw water parameters; (d)–(f) treated water parameters.

two clusters is related to the increase in the sum of squares when merged. In HCL, at the beginning of the calculation, every point or variable is in its own cluster, so the sum of squares starts out at zero and then increases as clusters are merged. Ward's method aims to minimize this increase. Euclidean distance is used as a similarity metric to calculate the distance between samples. So the criterion for choosing the pair of clusters to merge at each step is based on the optimal value obtained by the Euclidian function.

In order to check multivariate analysis suitability, PCA and HCL are preceded by some tests applied to the selected databases: the determination of Kaiser–Meyer–Olkin (KMO) coefficient and the Bartlett's test.

The data analyses were performed using the software R with the function 'kmo' and 'bartlett.test' respectively for the Bartlett's test and the KMO test, and the package 'FactoMiner' for the PCA and HCL.

RESULTS AND DISCUSSION

Data description and tests

Three databases were analyzed. Each of them was composed of water quality parameters (turbidity, salinity, conductivity, pH, M-alkalinity, water hardness, calcium, magnesium, chlorides, and organic matter) for raw and treated water samples. Turbidity is an important component of the water treatment process (Morris & Knocke 1984). For raw water, the turbidity varied from 2.0 to 65.5 NTU with an average of 10.8 NTU for TP1 and from 1.9 to 328.0 NTU with an average of 21.2 NTU for TP2 and TP3. Concerning treated water, turbidity had an average of 2.3 for TP1, 1.2 and 1.1 NTU respectively for TP2 and TP3 (Table 1). The treated water was mixed with clarified water from other sources to meet local standards prior to distribution. The mixing with other sources occurred after the treated water parameters measures.

As PCA and HCL are non-parametric classification methods, they make no assumptions about the underlying statistical distribution of the data.

In this study, the KMO coefficients calculated for the entire databases were respectively 0.87 for the database of TP1, 0.85 for TP2 and 0.88 for TP3. The high value of the KMO coefficients, close to 1 (Parinet *et al.* 2004), indicates that the databases of the three treatment plants, TP1, TP2 and TP3, are suitable for factor analysis.

For the three databases, the significance level of the Bartlett's test (Parinet *et al.* 2004) was less than 0.05 indicating that there are significant relationships among variables.

Table 1 | The mean values of the studied water parameters for TP1, TP2 and TP3 treatment plants

		TP1		TP2		TP3	
		Raw water	Treated water	Raw water	Treated water	Raw water	Treated water
Turbidity	(NTU)	10.7	2.3	21.2	1.2	21.1	1.1
Salinity	(mg/L)	1,000.8	977.1	1,374	1,322	1,371	1,322.6
Conductivity	(µS)	1,417	1,378	1,931	1,858	1,926	1,861
pH		8	7.8	8	7.7	8	7.6
M-alkalinity	(mg/L)	116.2	111.6	117.3	107.8	117.2	106.5
Water hardness	(°fH)	41.0	40.0	51.2	48.7	51.0	49.0
Ca^{2+}	(mg/L)	100.2	97.4	123.5	119.1	121.2	118.9
Mg^{2+}	(mg/L)	38.7	37.2	49.4	46.1	49.5	46.6
Cl^{-}	(mg/L)	290.7	275.1	406.8	380.8	402.6	384.1
Organic matter	(mg O2/L)	4.9	4	5.5	4.1	5.5	3.9

Classification of water quality parameters based on PCA

The PCA applied on the TP1 raw water parameters (Table 2) showed that the three first principal components explained up to 84% of the total observation variance (PC1 55.8%; PC2 17.2% and PC3 11.2%). For the TP1 treated water parameters, it was 82.3%. For PCA applied to TP2 parameters for raw and treated water, the values were, respectively, 87.6 and 85.6%. For TP3, they are 85.1 and 86.5%, respectively. The chosen factors have eigenvalues higher than 1.0 (Kaiser Criteria) according to the results of Jackson (1993) that compares heuristical and statistical approaches to choose the adequate factors. All three TP1, TP2 and TP3 for both raw and treated water show eigenvalues for the three first principal components higher than 1.0.

By applying PCA on TP1 raw and treated water parameters (Table 2 and Figure 1) it appears that salinity, conductivity, water hardness, magnesium, calcium and chlorides are highly explained by the first principal component (PC1) for raw and treated water (with all variable contributions higher than 0.899), turbidity and organic matters are well explained by PC2 for raw water and by PC3 for treated water (with all variable contributions higher than 0.700). pH and M-alkalinity are well correlated to PC3 for raw water and to PC2 for treated water. All results from Figure 1 show clusters containing the parameters of salinity, conductivity, water hardness, magnesium, calcium and chlorides. This suggests that the variation of these parameters is consistently similar. From Figure 1(b), 1(e) and 1(f), it appears that both turbidity and organic matters, and pH and M-alkalinity, form their own separate clusters. This distribution is similar for the other two treatment plants (Table 2).

Regardless of the water type (raw or treated), the application of PCA to waters of TP2 and TP3 assigned all the parameters to three clusters (Figure 2): salinity, conductivity, water hardness, magnesium, calcium and chlorides (in green) are always correlated with each other, as it is for turbidity and organic matters (in blue) as well as for pH and M-alkalinity (in red).

The correlation between the parameters in each cluster (Table 2, Figures 1 and 2) allows the choice of one of them to represent the others. Conductivity is a simple measure and could represent the first cluster (green). Turbidity could represent the second cluster (red). The third cluster (blue) could be represented by pH as it is easily measured.

In this way, according to the study by PCA, conductivity, turbidity and pH parameters could be the representatives of potable water quality during the treatment process whatever the type of process: static settlement tank, lamella separator or even pulsator settlement tank.

Classification of water quality parameters based on HCL

The dendrograms that result from cluster analysis HCL for both raw and treated water parameters of TP1 show the same three distinct clusters as PCA (Figure 3). The green

Table 2 | PCA applied to raw and treated water parameters for different WTPs

	1st WTP (TP1)			2nd WTP (TP2)			3rd WTP (TP3)		
	Factors			Factors			Factors		
Raw water	**PC 1**	**PC 2**	**PC 3**	**PC 1**	**PC 2**	**PC 3**	**PC 1**	**PC 2**	**PC 3**
% of the total variance	55.76	17.18	11.20	60.42	14.98	12.18	60.98	13.40	10.72
Eigenvalues	5.58	1.72	1.12	6.04	1.50	1.22	6.10	1.34	1.07
Turbidity	0.002	**0.726**	0.340	−0.220	**0.802**	0.081	−0.178	**0.793**	0.064
Salinity	**0.948**	−0.013	−0.053	**0.949**	−0.195	−0.036	**0.936**	−0.232	−0.063
Conductivity	**0.956**	0.062	0.002	**0.951**	−0.190	0.049	**0.944**	−0.224	0.015
pH	−0.186	0.275	**0.731**	−0.135	0.168	**0.874**	−0.061	0.157	**0.925**
M-alkalinity	0.306	−0.084	**0.848**	0.340	−0.174	**0.804**	0.463	−0.431	**0.492**
Water Hardness	**0.973**	0.056	0.107	**0.969**	−0.139	0.141	**0.965**	−0.150	0.055
Calcium	**0.939**	0.088	0.196	**0.922**	−0.166	0.242	**0.912**	−0.267	0.111
Magnesium	**0.899**	0.005	−0.009	**0.941**	−0.090	−0.004	**0.907**	−0.007	0.004
Chlorides	**0.975**	−0.020	−0.006	**0.968**	−0.185	−0.012	**0.952**	−0.236	0.056
Organic matter	0.096	**0.905**	−0.095	−0.119	**0.883**	−0.055	−0.145	**0.873**	0.043
	Factors			**Factors**			**Factors**		
Treated water	**PC 1**	**PC 2**	**PC 3**	**PC 1**	**PC 2**	**PC 3**	**PC 1**	**PC 2**	**PC 3**
% of the total variance	57.05	13.77	11.47	59.50	14.99	11.07	62.05	13.44	10.96
Eigenvalues	5.71	1.38	1.15	5.95	1.50	1.11	6.20	1.34	1.10
Turbidity	−0.168	0.149	*0.793*	−0.220	0.116	*0.809*	−0.234	0.056	*0.785*
Salinity	**0.955**	0.064	−0.069	**0.952**	0.094	−0.077	**0.951**	0.173	−0.050
Conductivity	**0.969**	0.052	0.001	**0.955**	0.149	−0.027	**0.955**	0.196	−0.003
pH	0.003	***0.855***	−0.059	0.069	***0.878***	−0.027	0.123	***0.920***	−0.003
M-alkalinity	0.348	***0.820***	0.035	0.378	***0.807***	−0.125	0.471	***0.748***	−0.138
Water hardness	**0.967**	0.136	0.078	**0.963**	0.192	−0.018	**0.966**	0.193	−0.009
Calcium	**0.935**	0.150	0.127	**0.907**	0.310	0.018	**0.929**	0.263	0.025
Magnesium	**0.872**	0.115	−0.045	**0.941**	0.020	−0.068	**0.929**	0.047	−0.057
Chlorides	**0.968**	0.117	−0.020	**0.972**	0.121	−0.042	**0.954**	0.229	−0.015
Organic matter	0.182	−0.188	*0.700*	0.152	−0.292	*0.744*	0.201	−0.145	*0.792*

1st Group: Mineral elements (shown in bold); 2nd Group: Turbidity (shown in italic); 3rd Group: Alkalinity (shown in bold/italic).

cluster contains exclusively the parameters related to the mineral content of the waters: salinity, conductivity, magnesium, chlorides, water hardness and calcium. The red cluster contains two parameters: turbidity and organic matter. The blue cluster contains pH and M-alkalinity parameters.

The HCL results obtained for TP1 data are the same as those for TP2 and TP3 treatment plants: three clusters have emerged that could be represented by conductivity, turbidity and pH parameters. The results obtained by HCL are consistent with that of PCA: whatever the treatment process and its stage, three parameters that are conductivity, turbidity and pH can represent the water quality.

Generalization

The similarities in water quality of the three treatment plants are studied in order to generalize the results: the raw and the

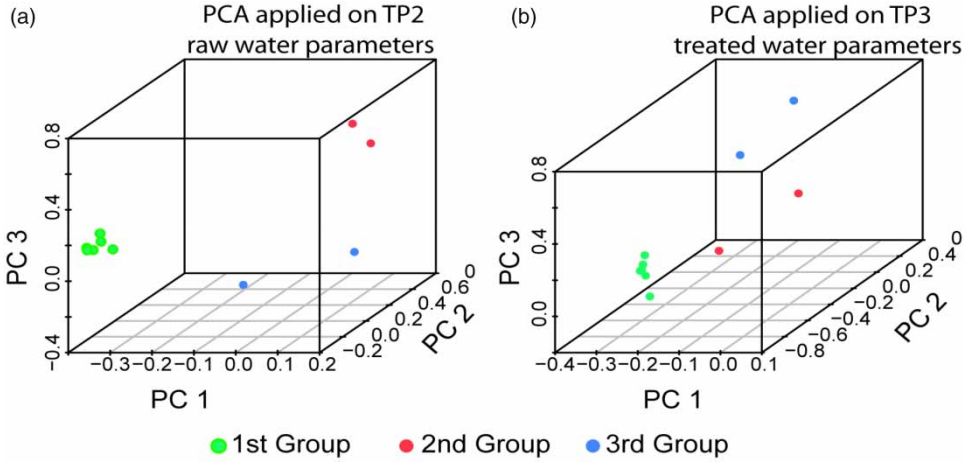

Figure 2 | PCA scores after application on (a) TP2 raw water parameters and (b) TP3 treated water parameters. Please refer to the online version of this paper to see this figure in color: http://dx.doi.org.10.2166/hydro.2017.070.

treated water quality of each studied plant are compared through PCA. First, the mean of the parameters measured in each treatment plant from the database detailed in Table 1 are merged. Then, PCA is applied to the new database and the first two principal components that explain 99.8% of the total variance of the observations are retained. The chart of the treatment plant raw water (TPraw) and treated water (TPtr) for the first two principal components (PC1 and PC2) shows three independent clusters: the first contains TP1 waters (raw and treated waters), the second contains TP2 and TP3 raw water and the third contains TP2 and TP3 treated water (Figure 4).

TP1 raw and treated waters are grouped together, apart from TP2 and TP3, probably because the TP1 water source is different from the TP2 and TP3 water source. TP2 and TP3 have the same raw water, and even if their treatment processes are different, TP2 and TP3 have similar treated water quality. More treatment plants should be investigated to confirm this hypothesis, these results nevertheless suggest that the quality of the raw water is more important than the

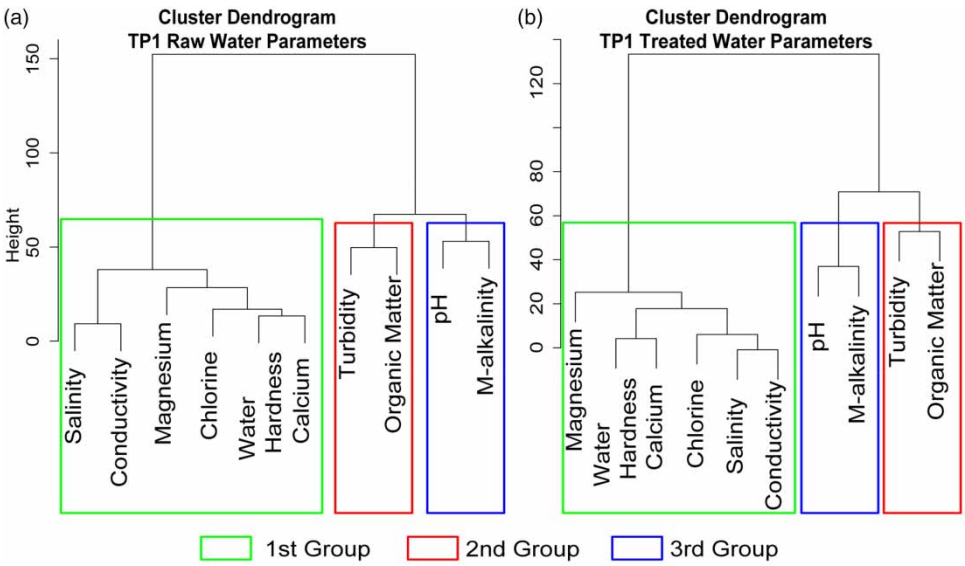

Figure 3 | Dendrogram of cluster analysis applied on TP1 database: (a) raw water parameters, (b) treated water parameters. Please refer to the online version of this paper to see this figure in color: http://dx.doi.org.10.2166/hydro.2017.070.

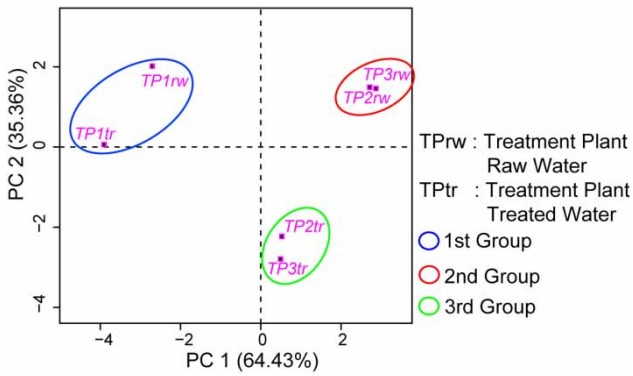

Figure 4 | Distribution of the treatment plant (TP) waters along the first two main axes PC1 and PC2.

type of treatment process, in the resulting quality of treated water.

DISCUSSION

All the parameters included in the first cluster (conductivity, salinity, magnesium, chlorides, water hardness and calcium) have the same variations during the water treatment processes. The correlation coefficient between conductivity and salinity is 0.98. This makes sense physically since salts break into positively and negatively charged ions when dissolving in water. These ions are conductors and thus conductivity increases suggesting that salinity should contribute to conductivity (CWT 2004). The correlation coefficient is 0.96 between water hardness and calcium and 0.90 between water hardness and magnesium. Water hardness is the sum of the molar concentrations of Ca^{2+} and Mg^{2+}. Correlation coefficients between calcium, magnesium, chlorides and conductivity two by two are higher than 0.80, in fact ions like Ca^{2+}, Mg^{2+} and Cl^{-} are chemical species responsible for the electrical conductivity as defined by Michael Faraday in 1830 (Mae-Wan 2010).

PCA and HCL showed that the variation of turbidity and organic matter parameters is correlated during the water treatment process. This fits the conclusion of the experiments carried out by the US Environmental Protection Agency on water quality that organic chemicals and biological contaminants (explaining the organic matters) have an effect on other water parameters including turbidity (Lambrou *et al.* 2012).

PCA and HCL place pH and M-alkalinity in the same cluster. Indeed, M-alkalinity directly measures the amount of any bicarbonate, carbonate, and hydroxide alkalinity present in water, depending on the pH value.

When PCA and Clustering are applied to treated water parameters then to raw water parameters, the group order obtained is different: after treatment, pH and M-alkalinity become correlated to the second PC and Turbidity and Organic matters become correlated to the third PC. This reflects the importance of adjuvants during the treatment, and more specifically during the coagulation–flocculation process. Indeed, the addition of coagulant during the coagulation process decreases the pH and the M-alkalinity of the water (Matilainen *et al.* 2010). The coagulant dose also has an effect on water turbidity and organic matter concentrations (Lee *et al.* 2001).

Other classification techniques may classify the studied data when the results given by PCA are difficult to interpret, e.g. Independent Component Analysis that is applied in many areas like signal and image processing (Shlens 2014). However, in this paper, PCA supported by HCL gives clear results (three clusters) and seems adequate for this case study.

The objective of PCA being to minimize the error between a projection and the original data, and then find a set of projections that maximize the variance of given data (Kwak 2008), our results are confirmed since, based on Table 1, the three first PCs explained up to an average of 85.2% of the total observation variance.

Lamrini *et al.* (2005) also studied the parameters influencing the treatment process; extracted TSS (total suspended solids), temperature, pH, conductivity and dissolved oxygen as independent parameters. Some of the parameters they extracted, such as pH, conductivity and dissolved oxygen (highly correlated to turbidity) agreed with our results. However, Lamrini *et al.* (2005) applied PCA on only 89 samples and only on raw water.

Additionally, many recent studies in the potable water field were interested in the chemicals introduced (coagulant or flocculant) during the treatment processes. In this regard, the quantities of coagulant and of flocculant used during the treatments are added to the database of TP1, but only on raw water parameters because these chemicals are introduced in the beginning of the treatment processes before another

PCA is applied. Three principal components explained 80.17% of the total variance of the observations. The charts (Figure 5) show that the clusters obtained previously appear again and contain the same parameters. Moreover, the injected quantities of chemicals are well explained by the second principal component (with a correlation coefficient of 0.89 for coagulant quantity and 0.83 for the flocculant quantity), the same as turbidity and organic matters. The two parameters of coagulant and flocculant quantities can be clearly included in the second cluster that already contains turbidity and organic matter. These results are the same when PCA is applied to TP2 and TP3 raw water databases including injected coagulant and flocculant quantities as new variables.

The inclusion of two more parameters to the databases does not influence the number and composition of clusters and still turbidity, pH and conductivity are the main parameters sufficient for the assessment of WTP performances. Adding new parameters (e.g. temperature and ultraviolet absorbance) would possibly influence the clusters' compositions and new clusters would appear which are not correlated to turbidity, conductivity or pH and the parameters that they represent. However, the parameters studied in this paper are the ones that are usually measured in WTPs, consequently it is unlikely that these results will change drastically. Even if these parameters can account for water quality in the present practices, additional analyses would surely enhance the quality of the available water in light of the increasing environmental issues.

CONCLUSION

The present study simplifies the assessment of the water quality during the water treatment process to three significant parameters which are conductivity, turbidity and pH. These parameters can be easily and quickly measured *in situ* and in a laboratory.

This result can be generalized and extrapolated since, in this study, PCA and HCL are applied to the databases of various WTPs wherein processes are different (static settlement tank, lamella separator and pulsator settlement tank) at different water treatment stages (raw water and treated water) and always leads to the same result: they divide the parameters into the three same clusters.

This approach can be a useful tool for the operators to quickly assess the treatment process or the water quality, and especially to detect anomalies or changes like unexpected changes of raw water during floods or fugitive pollutions. Additionally, if some specific nonstandard or non-routine action takes place, e.g. the addition to the process of activated charcoal (acting as an organic carbon parameter) due to unusual odor or color of the raw water or the addition of a prechlorination treatment to the process (change in chloride concentrations), it would be wise to measure some relevant parameters, besides the specified three. Then, after water quality stabilization and compliance with standards, water quality can be assessed based on the monitoring of the three extracted parameters.

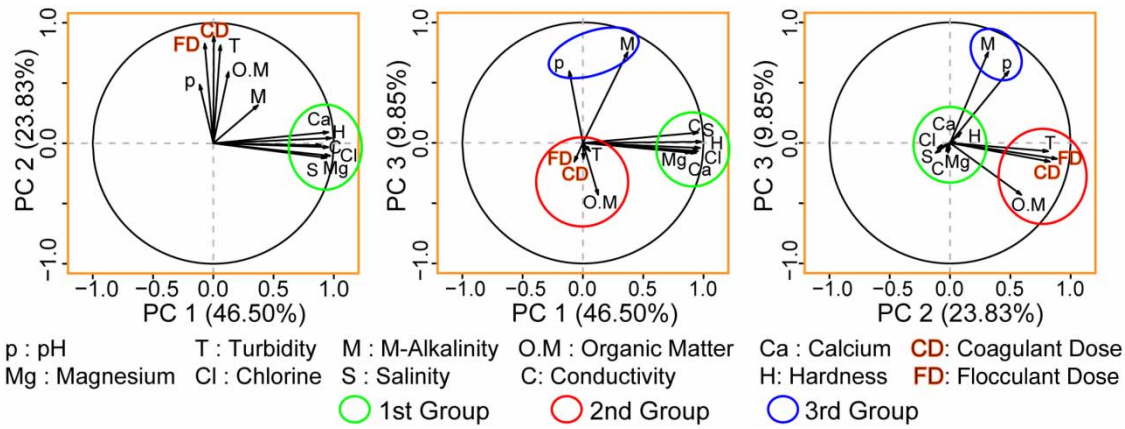

Figure 5 | PCA applied to TP1 data including coagulant and flocculant quantities.

Thus, most of the time, only three parameters could be monitored unless some extra actions or some occasional tests take place. The extrapolation of this paper's results are designed to be applied to local and worldwide potable WTPs monitoring (hence the inclusion of various treatment plant processes over a long time period). The reduction of the monitored parameters can be very useful.

Analyzing only the suggested parameters (conductivity, turbidity and pH) saves both time and money. Indeed, a simple calculation applied to our database shows that 10 parameters are assessed for each water type (raw and treated water) with about 348 sampling per year. Therefore, during one year, at least 6,960 analyses are conducted for the assessment of potable water treatment. Selecting only the three suggested parameters (instead of the current ten parameters), reduces the analysis to 2088, consistent with a reduction of more than one-third of the operations. Thus, time, money, operators, chemicals, electricity and device longevity can be saved or more frequent samples can be taken, thus improving the quality of the optimization.

This paper's results can also be a starting point for specific studies concerning water treatment optimization, for water modeling issues and for a fully automated monitoring system.

REFERENCES

Abdi, H. 2003 *Factor Rotations in Factor Analyses. Encyclopedia for Research Methods for the Social Sciences*. Sage, Thousand Oaks, CA, pp. 792–795.

Astel, A., Biziuk, M., Przyjazny, A. & Namieśnik, J. 2006 Chemometrics in monitoring spatial and temporal variations in drinking water quality. *Water Res.* **40** (8), 1706–1716.

Bain, R., Cronk, R., Hossain, R., Bonjour, S., Onda, K., Wright, J. & Bartram, J. 2014 Global assessment of exposure to faecal contamination through drinking water based on a systematic review. *Trop. Med. Int. Health* **19** (8), 917–927.

Colin, J. L., Dutot, A. L., Bablon, G. & Vie le Sage, R. 1986 Application de l'analyse des correspondances à des résultats d'essais de traitement d'eaux de surface. *Water Res.* **20** (6), 675–684.

Cruz, A. G., Cadena, R. S., Alvaro, M. B. V. B., Sant'Ana, A. S., Oliveira, C. A. F., Faria, J. A. F. & Ferreira, M. M. C. 2013 Assessing the use of different chemometric techniques to discriminate low-fat and full-fat yogurts. *LWT Food Sci. Technol.* **50** (1), 210–214.

C.W.T. Clean Water Team 2004 Electrical conductivity/salinity Fact Sheet, FS-3.1.3.0(EC). In: *The Clean Water Team Guidance Compendium for Watershed Monitoring and Assessment, Version 2.0*. Division of Water Quality, California State Water Resources Control Board (SWRCB), Sacramento, CA.

Duarte, A. C. & Capelo, S. 2006 Application of chemometrics in separation science. *J. Liquid Chromatogr. Relat. Technol.* **29** (7–8), 1143–1176.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. 1999 Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Meth.* **4** (3), 272.

Fabris, R., Denman, J., Braun, K., Ho, L. & Drikas, M. 2015 Surface analysis of pilot distribution system pipe autopsies: the relationship of organic and inorganic deposits to input water quality. *Water Res.* **87**, 202–210.

Gvozdić, V., Brana, J., Malatesti, N. & Roland, D. 2012 Principal component analysis of surface water quality data of the River Drava in eastern Croatia (24 year survey). *J. Hydroinform.* **14** (4), 1051–1060.

Jackson, D. A. 1993 Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74** (8), 2204–2214.

Jung, D., Kang, D., Liu, J. & Lansey, K. 2015 Improving the rapidity of responses to pipe burst in water distribution systems: a comparison of statistical process control methods. *J. Hydroinform.* **17** (2), 307–328.

Kaviarasan, M., Geetha, P. & Soman, K. P. 2015 Multivariate statistical technique for the assessment of ground water quality in Coonoor Taluk, Nilgiri District, Tamilnadu, India. *Indian J. Sci. Technol.* **8** (36). DOI: 10.17485/ijst/2015/v8i36/87535.

Kim, S. E. & Seo, I. W. 2015 Artificial neural network ensemble modeling with exploratory factor analysis for streamflow forecasting. *J. Hydroinform.* **17** (4), 614–639.

Kwak, N. 2008 Principal component analysis based on L1-norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **30** (9), 1672–1680.

Lambrou, T. P., Panayiotou, C. G. & Anastasiou, C. C. 2012 A low-cost system for real time monitoring and assessment of potable water quality at consumer sites. In: *Proceedings of the Sensors IEEE*, Taipei, Taiwan, pp. 1–4.

Lamrini, B., Benhammou, A., Karama, A. & Le Lann, M. V. 2005 A neural network system for modelling of coagulant dosage used in drinking water treatment. In: *Adaptive and Natural Computing Algorithms*, Springer, Vienna, pp. 96–99.

Lee, K. J., Kim, B. H., Hong, J. E., Pyo, H. S., Park, S. J. & Lee, D. W. 2001 A study on the distribution of chlorination by-products (CBPs) in treated water in Korea. *Water Res.* **35** (12), 2861–2872.

Mae-Wan, H. D. 2010 Les déplacements des ions qui dansent par le Dr. Mae-Wan Ho. Report of The Institute of Science in Society, 28 September 2010.

Matilainen, A., Vepsäläinen, M. & Sillanpää, M. 2010 Natural organic matter removal by coagulation during drinking water

treatment: a review. *Adv. Colloid Interface Sci.* **159** (2), 189–197.

Morris, J. K. & Knocke, W. R. 1984 Temperature effects on the use of metal-ion coagulants for water treatment. *J. Am. Water Works Assoc.* **76** (3), 74–79.

Onda, K., LoBuglio, J. & Bartram, J. 2012 Global access to safe water: accounting for water quality and the resulting impact on MDG progress. *Int. J. Environ. Res. Public Health* **9** (3), 880–894.

Parinet, B., Lhote, A. & Legube, B. 2004 Principal component analysis: an appropriate tool for water quality evaluation and management – application to a tropical lake system. *Ecol. Model.* **178** (3), 295–311.

Ratnaweera, H. & Fettig, J. 2015 State of the art of online monitoring and control of the coagulation process. *Water* **7** (11), 6574–6597.

Rosen, C. & Lennox, J. A. 2001 Multivariate and multiscale monitoring of wastewater treatment operation. *Water Res.* **35** (14), 3402–3410.

Russell, D. W. 2002 In search of underlying dimensions: the use (and abuse) of factor analysis in personality and social psychology bulletin. *Pers. Soc. Psychol. Bull.* **28** (12), 1629–1646.

Segreto, T., Simeone, A. & Teti, R. 2014 Principal component analysis for feature extraction and NN pattern recognition in sensor monitoring of chip form during turning. *CIRP J. Manuf. Sci. Technol.* **7** (3), 202–209.

Shlens, J. 2014 *A Tutorial on Principal Component Analysis.* Systems Neurobiology Laboratory, Salk Insitute for Biological Studies La Jolla, USA.

Shrestha, S., Kazama, F. & Nakamura, T. 2008 Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River. *J. Hydroinform.* **10** (1), 43–56.

Valentin, N. 2000 *Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable*. Thesis doctorate, l'Université de Technologie de Compiègne, France.

Valipour, M. 2015a Future of agricultural water management in Africa. *Arch. Agron. Soil Sci.* **61** (7), 907–927.

Valipour, M. 2015b Land use policy and agricultural water management of the previous half of century in Africa. *Appl. Water Sci.* **5** (4), 367–395.

Valipour, M. 2016a Variations of land use and irrigation for next decades under different scenarios. *IRRIGA* **1** (1), 262–288.

Valipour, M. 2016b How do different factors impact agricultural water management? *Open Agric.* **1** (1), 89–111.

Valipour, M. 2016c How much meteorological information is necessary to achieve reliable accuracy for rainfall estimations? *Agriculture* **6** (4), 53.

Wu, G. D. & Lo, S. L. 2008 Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Eng. Appl. Artif. Intell.* **21** (8), 1189–1195.

Yeung, K. Y. & Ruzzo, W. L. 2001 Principal component analysis for clustering gene expression data. *Bioinformatics* **17** (9), 763–774.

Zhu, F., Zhong, P. A., Xu, B., Wu, Y. N. & Zhang, Y. 2016 A multi-criteria decision-making model dealing with correlation among criteria for reservoir flood control operation. *J. Hydroinform.* **18** (3), 531–543.

Zugarramurdi, A., Parin, M. A. & Lupin, H. M. 1995 *Economic Engineering Applied to the Fishery Industry* (No. 351). Food & Agriculture Org. (FAO), Rome.

First received 1 July 2016; accepted in revised form 30 June 2017. Available online 24 August 2017