Groundwater

Groundwater Quality: Analysis of Its Temporal and Spatial Variability in a Karst Aquifer

by Roger Pacheco Castro[1], Julia Pacheco Ávila[2], Ming Ye[1], and Armando Cabrera Sansores[3]

Abstract

This study develops an approach based on hierarchical cluster analysis for investigating the spatial and temporal variation of water quality governing processes. The water quality data used in this study were collected in the karst aquifer of Yucatan, Mexico, the only source of drinking water for a population of nearly two million people. Hierarchical cluster analysis was applied to the quality data of all the sampling periods lumped together. This was motivated by the observation that, if water quality does not vary significantly in time, two samples from the same sampling site will belong to the same cluster. The resulting distribution maps of clusters and box-plots of the major chemical components reveal the spatial and temporal variability of groundwater quality. Principal component analysis was used to verify the results of cluster analysis and to derive the variables that explained most of the variation of the groundwater quality data. Results of this work increase the knowledge about how precipitation and human contamination impact groundwater quality in Yucatan. Spatial variability of groundwater quality in the study area is caused by: a) seawater intrusion and groundwater rich in sulfates at the west and in the coast, b) water rock interactions and the average annual precipitation at the middle and east zones respectively, and c) human contamination present in two localized zones. Changes in the amount and distribution of precipitation cause temporal variation by diluting groundwater in the aquifer. This approach allows to analyze the variation of groundwater quality controlling processes efficiently and simultaneously.

Introduction

Protection of groundwater resources is imperative in karst areas where karst aquifers are the only source of drinking water but vulnerable to contamination. Techniques of multivariate statistical analysis (e.g., cluster analysis, principal component analysis, factor analysis, and discriminants analysis) have been used widely to gain understanding of water quality with respect to spatial and temporal variability, hydrochemical facies, flow paths, and other factors that influence groundwater quality (Suk and Lee 1999; Thyne et al. 2004; Kim et al. 2005; Woocay and Walton 2008; Mohammadi 2009; Belkhiri et al. 2011; Nosrati and Eeckhaut 2012; Srivastava et al. 2012). Cluster analysis is a multivariate statistical technique used to classify a dataset into groups so that the data within each group are similar but different from data within other groups (Suk and Lee 1999; Davis 2002; Güler et al. 2002; Thyne et al. 2004; Shrestha and Kazama 2007; Everitt and Hothorn 2011). For temporal analysis, a common practice is to first apply cluster analysis sequentially to data of each sampling period, and then examine the clusters of each sampling period in order to understand water quality temporal processes (Suk and Lee 1999; Thyne et al. 2004; Hussain et al. 2008). However, understanding the spatial relation between the clusters and the temporal variation of the variables affecting water quality may not be straightforward, which renders the sequential cluster analysis inefficient to reveal temporal variation of groundwater quality.

This work presents an approach to use maps and boxplots of clusters for groundwater quality analysis. The proposed approach is based on the following observation. Given a set of clusters classified using Ward method with Euclidean distance, if water quality data at the same sampling location do not vary significantly in time, the data belong to the same cluster. This observation allows to identify the sample sites where groundwater quality changes in time and to analyze the temporal and spatial processes affecting groundwater quality. Instead of the sequential cluster analysis described above, cluster

[1]Geophysical Fluid Dynamics Institute and Department of Scientific Computing, Florida State University, Tallahassee, FL 32306.

[2]Corresponding author: Universidad Autónoma de Yucatán Facultad de Ingeniería, Av. Industrias no Contaminantes por Periférico Norte Apartado Postal 150 Cordemex, 97310 Mérida, Yucatán, México; pavila@correo.uady.mx

[3]Universidad Autónoma de Yucatán Facultad de Ingeniería, Av. Industrias no Contaminantes por Periférico Norte Apartado Postal 150 Cordemex, 97310 Mérida, Yucatán, México.

Article impact statement: Propose an innovative approach to analyze groundwater quality that reveal important information about its temporal and spatial variability.

Received November 2016, accepted May 2017.

doi: 10.1111/gwat.12546

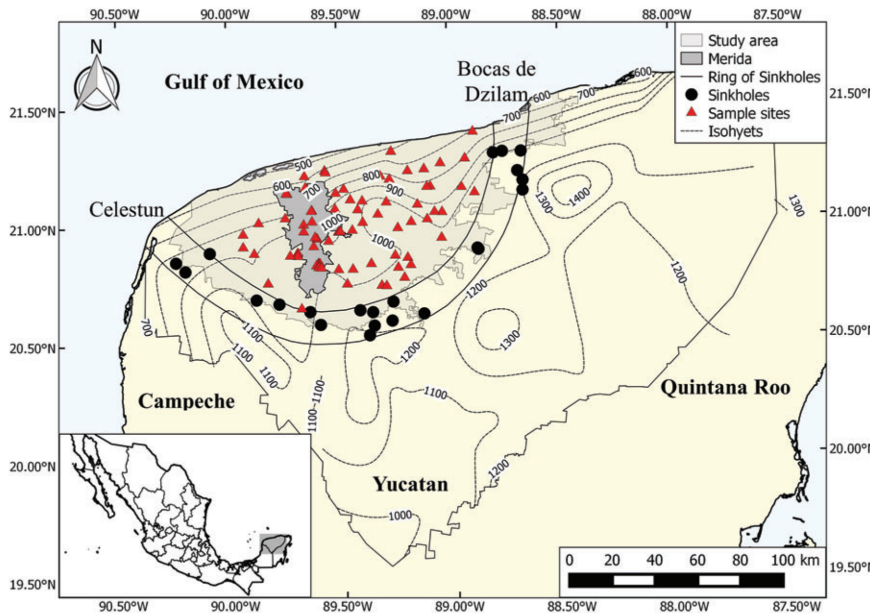

Figure 1. Location of the study area in the State of Yucatan along with average annual precipitation isohyets. The dots represent the locations of some sinkholes within the ring of sinkholes delimited by the two semicircles (Pérez-Ceballos et al. 2012). Contours of precipitation are adapted from PROCOMAR (2003).

analysis is applied to lumped data of all sampling periods. This leads to maps (snapshots) of spatially distributed clusters. Principal component analysis is used to verify the results of cluster analysis and to derive the variables that explained most of the variation of the groundwater quality data. The maps and boxplots of the major ions are used to analyze the spatial variation of the processes that govern groundwater quality at a given time. The snapshots show the spatial variation of water quality at different times, and the temporal processes variation can be then derived from the comparison of the snapshots. If water quality changes significantly with time at a given area, temporal variation of water quality processes can be revealed by comparing the chemical characteristics of the clusters to which that area belong. In other words, the proposed approach directly links cluster change with the temporal and spatial variation of processes controlling water quality. While applying cluster analysis to lumped data has been reported in literature (e.g., Vega et al. 1998; Wunderlin et al. 2001; Srivastava et al. 2012), the previous uses are limited to exploratory analysis with the objective of classifying water quality data, but do not analyze temporal relations among spatial distribution of clusters.

This procedure is applied to the water quality data collected from the karst aquifer in Yucatan, Mexico (Figure 1). The aquifer is the only source of drinking water for a population of nearly two million people. This aquifer is highly vulnerable to contamination, and groundwater in the upper part of the aquifer is not recommended for human consumption (Escolero et al. 2000; Marin et al. 2000; INEGI 2002, 2010). Therefore, understanding spatial and temporal variability of groundwater quality in Yucatan is important for protecting the groundwater resource and the population that depends on it. The issue of groundwater quality has been addressed from the chemical and bacteriological points of view in previous studies (Back and Hanshaw 1970; Doehring and Butler 1974; Pacheco et al. 2000, 2004; Perry et al. 2002; Delgado et al. 2010). Back and Hanshaw (1970) concluded that dissolution of carbonate rocks and salt water intrusion were the major processes controlling groundwater quality in the Yucatan peninsula. According to Perry et al. (2002), salt water intrusion extends more than 100 km from the gulf coast. Doehring and Butler (1974) found that the mixing with saline water may be enhanced by high pumping rates from supply wells. Pacheco et al. (2004) analyzed groundwater quality of water samples from supply wells in the Yucatan State, and concluded that groundwater quality was acceptable. However, they found that at some locations the concentrations of nitrate, chloride, total hardness, sodium, and cadmium exceeded the Mexican standards for drinking water. Domestic and municipal sewage, excreta and urine residues from domestic animals, and fertilizers have been identified as the major sources of groundwater pollutants (Doehring and Butler 1974; Pacheco and Cabrera 1997; Back 1999; Pacheco et al. 2001).

Cabrera et al. (2002) analyzed major ions of groundwater samples from a supply well field and its surrounding area located in the City of Merida, the capital city of the State of Yucatan. They concluded that, in the study area, there was no significant variation in groundwater quality between seasons (one dry season and one rainy season) or depths (shallow and 30 to 40 m deep). Pacheco et al. (2004) analyzed water samples from supply wells located throughout the entire state of Yucatan. They found that, in more than half of the towns within the State of Yucatan, the concentrations of two to three water quality variables exceeded the Mexican standards for drinking water. Delgado et al. (2010) analyzed groundwater samples from supply wells to examine whether groundwater can be used for irrigation. Based on a geostatistical analysis of effective salinity they divided the aquifer into seven zones, and concluded that groundwater in only one zone met the quality requirements for agricultural irrigation. Pérez-Ceballos et al. (2012) studied groundwater quality in the Ring of Sinkholes (the boundary of our current study area, Figure 1) at three periods (dry, rain, and cold-front) and three depths (0.5, 5.5, and 10 m). By applying multivariate statistical techniques they split the Ring of Sinkholes into three zones, and concluded that water quality in each zone did not vary with time.

The conclusions above were based on a limited amount of data, and the water samples used in the previous analysis did not have sufficient temporal resolution for analyzing the variation of groundwater quality temporal processes in the State of Yucatan. Recently, a campaign of groundwater sampling was conducted and a total of 288 water samples were collected from the Yucatan aquifer in four sampling periods over 2 years. Two sampling periods were for the dry season, and the other two for the rain season. This comprehensive dataset was of great value for advancing our understanding of groundwater quality for the State of Yucatan. The objective of this work is to analyze the groundwater quality data to determine the main variables influencing groundwater quality and their variations in space and time. The proposed approach is: 1) conduct cluster analysis for lumped data, 2) generate snapshots of groundwater quality clusters, 3) use principal component analysis to verify the validity of cluster analysis and to derive the variables that explain most of the variation of groundwater quality data, and 4) derive spatial and temporal variation of groundwater quality from the maps showing the temporal variations of the clusters and the boxplots of its major components. It is worth mentioning that this approach of using cluster analysis for lumped data is not limited to the groundwater quality data in Yucatan aquifer. It is applicable in other areas because no assumption is made specific to the aquifer of Yucatan characteristics.

Study Area and Groundwater Quality Data

The state of Yucatan, Mexico, occupies the central portion of the Yucatan Peninsula (Figure 1). The aquifer of Yucatan was developed in a carbonate platform, and it is composed mainly of limestone, marvel, and gypsum with high heterogeneity (Doehring and Butler 1974; Back 1999; INEGI 2002). The study area, known as the Hydrologic Region of the Ring of Sinkholes (Figure 1), is located within a semicircular formation of sinkholes, also known as the Ring of Sinkholes, in the northern part of the state of Yucatan (Perry et al. 1989; INEGI 2002). The aquifer is unconfined (except for a narrow band close to the coastline) and under-laid by saline water (Doehring and Butler 1974; Perry et al. 2002; INEGI 2010). The thickness of the aquifer is about 55 m at the south of Merida (Figure 1), the capital and the largest city of Yucatan hosting about half of the state population (Doehring and Butler 1974; Back 1999; INEGI 2002, 2010; Perry et al. 2002). The aquifer is highly vulnerable to contamination, and groundwater of the top 15 to 20 m of the aquifer does not meet the standards for human consumption (Escolero et al. 2000; Marin et al. 2000; INEGI 2002).

Groundwater samples were taken from the water supply wells located within the Ring of Sinkholes (Figure 1) at a depth ranging from 20 to 40 m. To understand the variability of the spatial and temporal processes affecting groundwater quality, a total of 288 groundwater samples were taken in four sampling periods between 2009 and 2011. The number of samples, for each sampling period, were 76, 63, 74, and 75, respectively. The first and third sampling periods were planned to be at the end of the rainy season and the second and fourth to be at the end of the dry season. According to Schmitter-Soto et al. (2002) the rainy season in Yucatan occurs during the months of June and October. The well locations for sampling period 1 can be found in Figure 1. Water samples were collected directly from the supply wells prior to water disinfection to avoid possible changes in water chemistry. For each sample the parameters pH, temperature, electric conductivity, total dissolved solids, and salinity were measured in the field using a multiparametric sonde Hach model Quanta. For the measurement of major ion concentrations samples were collected and preserved in plastic bottles for laboratory analysis. The measured ions were calcium (Ca^{+2}), magnesium (Mg^{+2}), sodium (Na^{+}), potassium (K^{+}), chloride (Cl^{-}), sulfate (${SO_4}^{-2}$), bicarbonate (${HCO_3}^{-}$), and nitrate (${NO_3}^{-}$). Major ions were determined using the standard methods described in APHA, AWWA, WEF (2005). Electrical-balance errors from the major ions in the samples were calculated as an indicator of data quality (Deutsch 1997). Although the common rule is that the electrical-balance errors should be less than 5% for groundwater samples (Hounslow 1995; Deutsch 1997), in this study a sample was considered good if the electrical balance was less than 10%. This value was chosen to take into account a systematic error that may be introduced by the use of titration techniques.

Multivariate Statistics

This section first explains briefly cluster and principal component analysis. The following explanations assume that the data set is composed of M observations $O_1, O_2, \ldots, O_M$, and N measured variables $X_1, X_2, \ldots, X_N$. Therefore, the i-th observation, $O_i = (x_{i1}, x_{i2}, \ldots, x_{1N})$, is a vector, where x_{ij} is the value of the j-th variable in the i-th observation. Each observation represents a sampling well at a given time, and the measured parameters for this study correspond to the physicochemical characteristics and major ion concentrations described in the previous section.

Hierarchical cluster analysis with the Ward method is an agglomerative technique. The method starts by treating each observation as a cluster. Each successive step, clusters are merged according to the Ward criterion, i.e., the minimization of the error sum of squares (ESS) defined bellow. Suppose that at step k, there are N_k clusters. Given cluster A_l with n_l observations ($l = 1, \ldots, N_k$), the ESS_l for A_l is defined as the sum of the squared Euclidean distance of each element in A_l from the mean value:

$$\text{ESS}_{A_l} = \sum_{i=1}^{n_l} \left(O_i - \overline{O}\right)\left(O_i - \overline{O}\right)' \tag{1}$$

where O_i ($i = 1, \ldots, n_l$) is the i-observation of cluster A_l, $\bar{O}$ the corresponding mean value of all the observations in A_l, and T indicates the transposed vector. The total ESS over all the clusters at this step is defined as

$$\text{ESS} = \sum_{l=1}^{N_k} \text{ESS}_{A_l} \tag{2}$$

At the beginning of the cluster algorithm, $\text{ESS} = 0$, as ESS_l is zero for all l. In the next step, all possible cluster combinations are considered and the corresponding ESS is calculated for each combination. According to the Ward's method, the pair of clusters that give the smallest increment in the value of ESS are merged. In other words, the Ward's criterion is to minimize ESS within the clusters at each step (Ward 1963; Davis 2002). The procedure of merging continues until only one cluster is formed. When the procedure is complete, a dendrogram is created, and it shows the degree of dissimilarity of the clusters. For the Ward method, the dissimilarity between two clusters is defined as $\sqrt{2\Delta\text{ESS}}$, where ΔESS is the increment in the ESS caused by merging two clusters, this means that if cluster C is generated by merging clusters A and B then $\Delta\text{ESS} = \text{ESS}_C - \text{ESS}_A - \text{ESS}_B$. In the dendrogram, the horizontal axis is for all the observations used in the analysis and the vertical axis shows the dissimilarity (Kaufman and Rousseeuw 1990; Davis 2002; Everitt and Hothorn 2011). By using the dendrogram, one can decide the number of clusters to use for further analysis. The decision on the number of clusters to use is subjective; however, it can be supported using other techniques (Thyne et al. 2004; Woocay and Walton 2008). Thyne et al. (2004) justified their decision on spatial coherence and inverse geochemical modeling. Woocay and Walton (2008) used the biplots of principal components to justify the selection of the number of clusters. In this study, while the number of clusters is determined empirically, we discuss a procedure that may help to determine the the optimal number of clusters to use. Biplots are used to verify our selection.

Principal component analysis is used to reduce the dimensionality of a problem by selecting a new set of variables, which has lower dimensionality, but account for most of the variance of the original variables. Principal component analysis is an orthogonal transformation of the system axes. The new set of variables, called principal components, are the axes of the transformed system and linear combinations of the original variables (Johnson and Wichern 2007; Everitt and Hothorn 2011). Suppose and Σ is the sample covariance of the measured variables $X_1, X_2, \ldots, X_N$. If λ_i and e_i ($i = 1, \ldots, N$) are the eigenvalues and eigenvectors of Σ, respectively, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ then the new variables (principal components) are given by

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{iN}X_N, \tag{3}$$

where e_{ij}, also called loadings, is the j-th components of eigenvector e_i. The y_i values of the principal components Y_i are called scores. The new set of variables are uncorrelated and $\text{var}(Y_i) = \lambda_i$ for $i = 1, \ldots, N$. Therefore, one can choose a subset of variables $\{Y_i, i = 1,2, \ldots, P\}$, with $P <= N$, that explain most of the variation of the original variables $\{X_j, j = 1,2, \ldots, N\}$. There are multiple ways to determine the number of principal components to be retained for further analysis (Kaiser 1960; Johnson and Wichern 2007). As the main objective of principal component analysis was to confirm the selection of the number of clusters, in this study we use only the first principal components that help to define the different clusters (Thyne et al. 2004). The number of principal components used for further analysis is based on the minimum number of components required to delineate the different clusters in the biplots. A biplot is a two-dimensional plot with two principal components as axes. The bottom and left axes are used for the scores, shown as a scatter plot, and the top and right axes are for the loadings, shown as scaled arrows from the origin (Woocay and Walton 2008). The loadings are scaled for visualization, the main interest is in their direction (Woocay and Walton 2008). A biplot allow a simultaneous analysis of the relation between principal components, the original variables, and the data (Woocay and Walton 2008).

The two techniques of multivariate statistical analysis have been widely used in hydrology. Suk and Lee (1999) divided the system under study into different hydrochemical regimes. They applied cluster analysis to factor scores to reduce the effects of misclassification caused by outliers and multicollinearity. Seasonal variation is deducted from the results of cluster analysis after applying the analysis to data of each season sequentially. Thyne et al. (2004) developed a methodology to characterize watershed hydrology. For understanding temporal variation, the authors applied the sequential method to runoff and baseflow conditions separately. Woocay and Walton (2008) applied sequentially principal component analysis and k-means cluster analysis to groundwater quality data for a better understanding of groundwater flow and evolution. The authors used biplots to understand the processes affecting the groundwater chemical composition. By the use of biplots and digital elevation maps of the principal components, flow paths, and interactions of groundwater with surface water and geologic features were inferred. Other studies have used the cluster analysis for lumped data of all the sampling periods, but their analysis is limited for classifying water samples into groups

(Vega et al. 1998; Wunderlin et al. 2001; Srivastava et al. 2012), not for analyzing temporal and spatial variation of groundwater quality from the distribution of the clusters.

Methods

This section describes the proposed approach for clustering analysis to understand the processes affecting spatial and temporal groundwater quality and their variation. The first step prior to the multivariate statistical analysis is to reduce the data size by eliminating the variables that do not provide important information of groundwater quality. The variable elimination is based on two criteria: the Pearson correlation coefficient and the coefficient of variation. When two variables have a high correlation coefficient (more than 90%) only one is kept. Variables highly correlated may introduce errors in the computations and interpretation of results for principal component analysis (Johnson and Wichern 2007). The variables with very low (less than 5%) coefficient of variation, in space and time, are considered to be almost constant and thus excluded from the statistical analysis. After removing the unimportant variables for the analysis, the data are log transformed and standardized. While the log-transformation is not necessary for the analysis, it helps make the data closer to normality and provides better statistical results, as suggested by Güler et al. (2002). The standardization removes the impacts of data units on the statistical analysis.

Hierarchical cluster analysis was applied to the log transformed and standardized data. Based on the observation that, if water quality data at the same sampling location do not vary significantly in time the data belong to the same cluster, cluster analysis was applied to the entire dataset. The Ward's method cluster together samples that have little water quality variation in time and space, allowing to derive spatial and temporal variation of the processes affecting the groundwater quality data. It is important to notice that the observation depends on the dissimilarity level chosen to form the clusters. If the dissimilarity is close to zero then the number of clusters is the same as the number of data and the observation is not true. In the other extreme if only one cluster is selected then all the data will be on the same cluster independent of the data values. Once the cluster analysis is complete, the dendrogram is used to select the number of clusters for further analysis. The decision of the number of clusters is empirical but justified later using principal component analysis (Thyne et al. 2004; Woocay and Walton 2008). Principal component analysis was applied to the log transformed data to verify the result of cluster analysis and to identify the main variables that explain most of the variation of groundwater quality. This was achieved by using biplots of the first principal components. The number of principal components selected for further analysis is based on the number of principal components required to distinguish the different clusters on the biplots. Based on the selected principal components, the main variables affecting the groundwater quality were obtained. After the validation of cluster analysis result, spatial distribution of clusters is plotted in maps. One map per sampling period was generated, and these maps are snapshots of the distribution of the water quality in the study area. All parameters were plotted using boxplots grouped by cluster. The boxplots provide information about the main characteristics of each cluster, and allows a direct comparison of differences among clusters. Using the maps and boxplots, spatial and temporal variation of the governing processes of groundwater were analyzed. At the end of this analysis we discuss the procedure followed to decide the optimal number of clusters to use.

The software R (R Core Team 2015; Rstudio Team 2015) was used to perform the statistical analysis and the statistical results are plotted using the packages of GGPLOT2 and RGDAL (Wickham 2009; Bivand et al. 2015). Other maps were generated using QGIS (QGIS Development Team 2015).

Results and Discussion

Several variables were excluded from the statistical analysis based on Pearson correlation coefficient and the coefficient of variation, as described above. The parameters pH and temperature were not considered for cluster analysis because the coefficient of variation for pH and temperature was less than 5% in all the sampling periods. On the other hand, the correlation coefficients between Cl^- and the variables of salinity, total dissolved solids, electrical conductivity and sodium were higher or equal than 0.90 for all the sampling periods and depths. Therefore, only Cl^- data were retained for further analysis. As a result, the multivariate statistical analysis was conducted for the following variables: Ca^{+2}, Mg^{+2}, K^+, $HCO_3{}^-$, Cl^-, $SO_4{}^{-2}$, and $NO_3{}^-$.

Cluster Analysis

Figure 2 shows the dendrogram of cluster analysis for the selected variables. For this study a dissimilarity value of about 15 was selected which splits the dendrogram into five clusters. This means that, the last two elements joined in a clusters had a dissimilarity value smaller than 15. No other two clusters are merged if the dissimilarity value was larger than 15. The selection on the number of clusters was based on a visual analysis of the dendrogram and supported with principal component analysis. However, at the end of this section we discuss a procedure that help to decide the optimal number of clusters. To verify the results of the cluster analysis, clusters were plotted into biplots using principal components as shown in Figure 3. While Figure 3a, for PC1 and PC2, shows that the data corresponding to clusters 3 and 5 are apparently overlapping, the two clusters have substantial difference in Figure 3b for PC1 and PC3. Therefore, three principal components are enough to verify cluster analysis in the sense that two clusters can be distinguished when they are plotted in the biplots. The first three principal components (PC1, PC2, and PC3) explained 75% of the total variation of the original dataset (39.63%, 18.72%, and 17.19% respectively). The biplots were also used to extract the main variables that explain most of groundwater quality data variance. The loadings of PC1 plotted in Figure 3 showed that the groundwater quality is dominated by Cl^- and SO_4^{-2}, whose loading values were −0.51 and −0.48, respectively. Similarly, the loadings of PC2 and PC3 indicated that NO_3^- and HCO_3^- were dominant factors, and their loading values were 0.68 and −0.69, respectively. Therefore, PC1 can be related to saltwater intrusion, which had been reported as a controlling factor of the groundwater quality in the study area (Back and Hanshaw 1970; Perry et al. 2002). PC2 can be related to contamination by sewage and agricultural fertilizers. This agrees with the finding of Pacheco et al. (2004) that nitrate concentrations were above the Mexican standards for drinking water in 11% of the analyzed samples from supply wells in the Yucatan aquifer. Back and Hanshaw (1970) reported local contamination of groundwater by sewage; however, they concluded that it was not a dominant factor of groundwater quality. Results of this study showed that contamination, by sewage and agricultural fertilizers, are now a dominant factor of groundwater quality. PC3 can initially be related to dissolution of the carbonate minerals in the Yucatan aquifer. Although Ca^{+2} (−0.2) and Mg^{+2} (−0.34) have low loadings in PC3 they have high loading (0.5 and −0.65, respectively) in PC4, which explained 11.9% of the total variation. Therefore PC3 and PC4 together highlight the karstic nature of the aquifer. Although only three components were required to explain the variation of the clusters, we decided to consider PC4 because of its importance in the karstic nature of the aquifer. It is noted that the two major ions, Ca^{+2} and HCO_3^-, characteristic of a karstic aquifer, were not related to the same principal component. This may be explained by the spatial distribution of the two major ions due to their genesis and chemical behaviors. Ca^{+2} in groundwater is mainly due to dissolution of carbonate rocks and have preference to form ion complexes with sulfates (Morel 1983; Deutsch 1997; Perry et al. 2002). HCO_3^- in groundwater has two sources, dissolved carbon dioxide and dissolution of carbonate rocks (Morel 1983; Appelo and Postma 2005; Ibanez et al. 2007). The interaction of groundwater with the carbonate rocks has been reported as a controlling factor of groundwater quality in the study area (Doehring and Butler 1974; Back and Hanshaw 1970). Ibanez et al. (2007) estimated that 0.77% of the CO_2 generated in soils ends in groundwater and concluded that this small amount has a significant impact in the groundwater chemical composition. The main sources of CO_2 identified by Ibanez et al. (2007) were respiration processes and decomposition of organic matter, which are dependent on the temperature and seasonal conditions. Therefore PC3 can be related to precipitation and PC4 to the dissolution of carbonate rocks. Back and Hanshaw (1970) reported that rain was not a significant controlling factor of the groundwater quality in Yucatan. However, this results suggest that rain is a dominant variable of groundwater quality. This result is further confirmed in the groundwater temporal variation section.

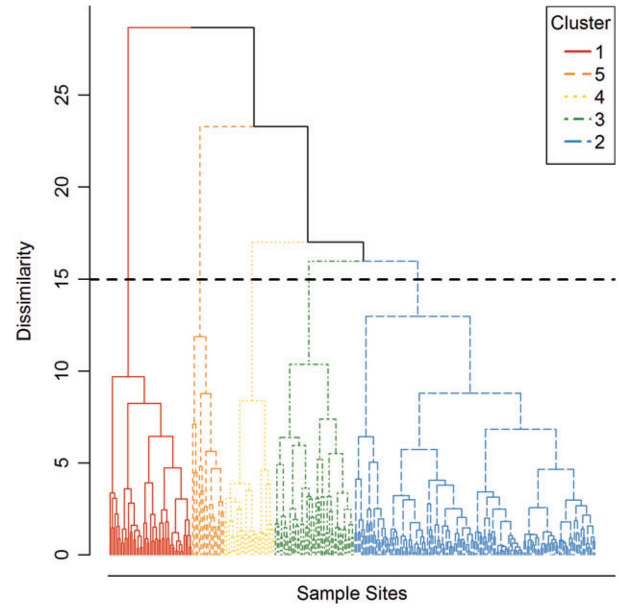

Figure 2. Dendrogram of cluster analysis applied to all the sampling periods and the seven major ions: Ca^{+2}, Mg^{+2}, K^+, HCO_3^-, Cl^-, SO_4^{-2}, and NO_3^-. The dashed line represent the level of dissimilarity chosen to identify the five clusters selected for further analysis.

Spatial Variation

The clusters identified above are plotted in Figure 4 to analyze the spatial and temporal variability of the groundwater quality governing processes. The box-and-whisker plots for all major ions, including sodium for comparison purposes, are presented in Figure 5. As shown in Figure 4, cluster 1 is located in the west boundary of the study area and at sampling locations close to the coastal line (Figure 4a, 4c, and 4d). For sampling period two (Figure 4b), only some data in cluster 1 are located close to the coast. This cluster is characterized by the highest values of all the ions but NO_3^- (Figure 5). The high values of Cl^-, Na^+, K^+, and SO_4^{-2} are attributed to saltwater encroachment (Perry et al. 2002; Pérez-Ceballos et al. 2012). The samples on the west boundary are also influence by groundwater rich in SO_4^{-2} coming from an evaporite region at the south border of Yucatan and Quintana Roo states which moves toward the west part of the study area through the Ticul fault (Perry et al. 2002). Cluster 2 corresponds to a zone located at the center of the study area with some elements at the east side of the study area (Figure 4). It is noted that this cluster has fewer elements for sampling period two than for other sampling periods. Figure 5 shows that the concentrations of Ca^{+2}, Mg^{+2}, Cl^-, Na^+, K^+, and SO_4^{-2} are smaller in cluster 2 than in cluster 1, suggesting that cluster 2 is less affected by saltwater intrusion and the groundwater rich in sulfates. The sampling locations of cluster 3 are those located in the east side the study area. The only exception is that the data of the second sampling period occupy a larger area extending towards the center of the study area. Figure 5 shows that, for cluster 3, the concentrations of Ca^{+2}, Mg^{+2}, Na^+, K^+, Cl^-, SO_4^{-2}, and HCO_3^- are

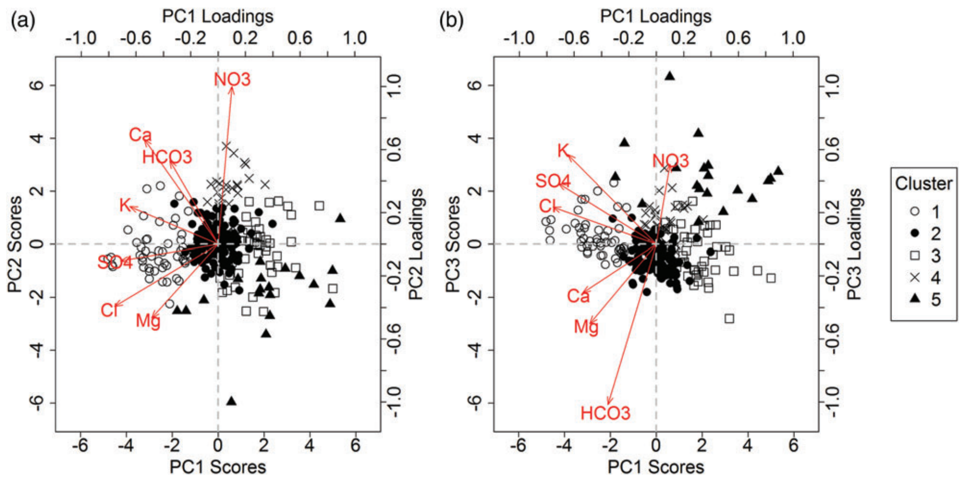

Figure 3. Biplots of the first principal components. (a) PC1 vs. PC2. (b) PC1 vs. PC3.

smaller than those of clusters 1 and 2, except that nitrate concentrations are slightly higher. The fact that cluster 3 has lower ion concentrations may be explained by the spatial distribution of average annual precipitation in the study area. As shown in Figure 1, most of the precipitation occurs in the east area of the State of Yucatan (INEGI 2002). Cluster 4 is located in two zones in the study area, one at the east and one at the west (Figure 4a, 4c, 4d). However, cluster 4 is located mainly in the west zone for the data of the second sampling period (Figure 4b). Cluster 4 is characterized by the highest values of NO_3^- and the lowest of Mg^{+2} (Figure 5). The zone at the west (Figures 4a, 4b, and 4c) is located over Merida, the biggest city in the State of Yucatan. Pacheco (2013) found that the east of the state had high annual loading of nitrogen due to extensive use of fertilizers, and the west had also high loading of nitrogen due livestock production, specifically pig farms. Therefore, this cluster can be associated with local contamination zones caused by the urbanization, disposal of untreated sewage, and excessive use of agricultural fertilizers (Pacheco and Cabrera 1997; Graniel et al. 1999; Pacheco 2013). Cluster 5 is present all over the study area during the second sampling period (Figure 4b), especially at certain locations occupied by the data of clusters 1 and 2. Note that this cluster is present only in the second sampling period. The average concentrations of Ca^{+2} and HCO_3^- are significantly lower than those of the other four clusters (Figure 5). These may be explained by the distribution of the precipitation in 2010, more details will be discussed below.

This analysis reveals the zones of influence of the different variables identified from principal component analysis. Cluster 1 shows the spatial extent of the sea water intrusion and the water coming from the south, rich in sulfates. However, it does not have the capability to separate the zones of influence of such processes. Clusters 2 and 3 have similar characteristics which are not affected by seawater intrusion or contamination. The average annual distribution of rain in the study area (Figure 1) causes the main differences between these clusters with cluster 3 having smaller concentrations of the ions. Thus, these zones represent the natural groundwater conditions of the aquifer only affected by the interaction of water with the rocks of the aquifer and the average annual distribution of precipitation. These processes were related to PC4 and PC3, respectively. Cluster 4 corresponds to the influence zones of the PC2 which was related to anthropogenic contamination. Cluster 5 that is present only in the second sampling period will be discussed in more detail the next section.

Temporal Variation

Figure 4 is used to analyze the temporal variation of groundwater quality by paying special attention to the changes of the cluster locations with time, as the temporal changes of the cluster locations reflect groundwater quality variation. There are some details about the precipitation that are important to mention and help to explain the temporal variation of clusters. The rainy season in Yucatan occurs between June and October (Schmitter-Soto et al. 2002). The Köppen-Geiger climate classification suggests that a month is dry if the total precipitation is less than 60 mm (Peel et al. 2007). Therefore, during 2010 the rainy season started in April and ended in September (Figure 6). Moreover, sampling period 2 planned to be at the end of the dry season was during the rainy season of 2010. Sampling period 3 was performed in dry months (October and November 2010); however, its chemical composition can be considered to be representative of the rainy season because the sampling was performed right after the rain season. Sampling period 4 is the only period performed in a dry season. Only 17 samples were taken during the first week of June 2011. The spatial distribution of the clusters are similar for sampling periods 1, 3, and 4 (Figure 4a, 4c, and 4d). Therefore, groundwater quality do not varies significantly over these sampling periods and the average chemical composition is given in Figure 5. Recall that the three sampling periods include two dry seasons and one rain season. This is considered to be the stable configuration of groundwater

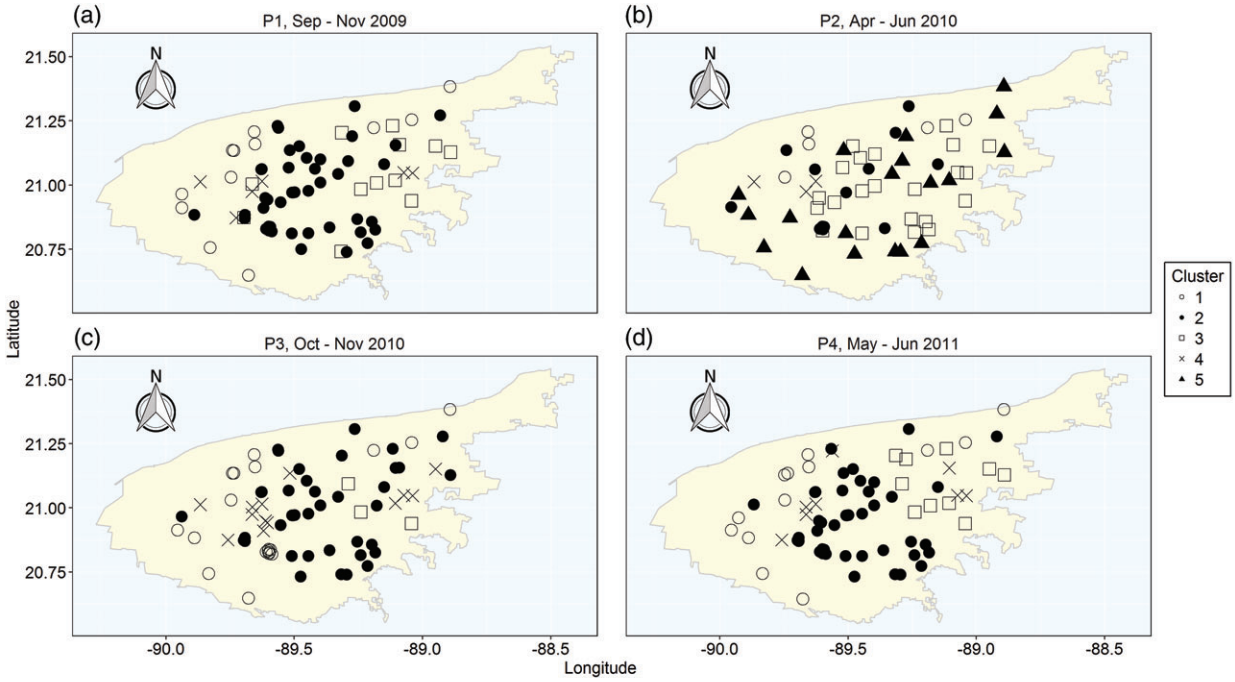

Figure 4. Spatial distribution of the five clusters selected for further analysis. (a) First sampling period, September to November 2009. (b) Second sampling period, April to June 2010. (c) Third sampling period, October to November 2010. (d) Fourth sampling period, May to June 2011.

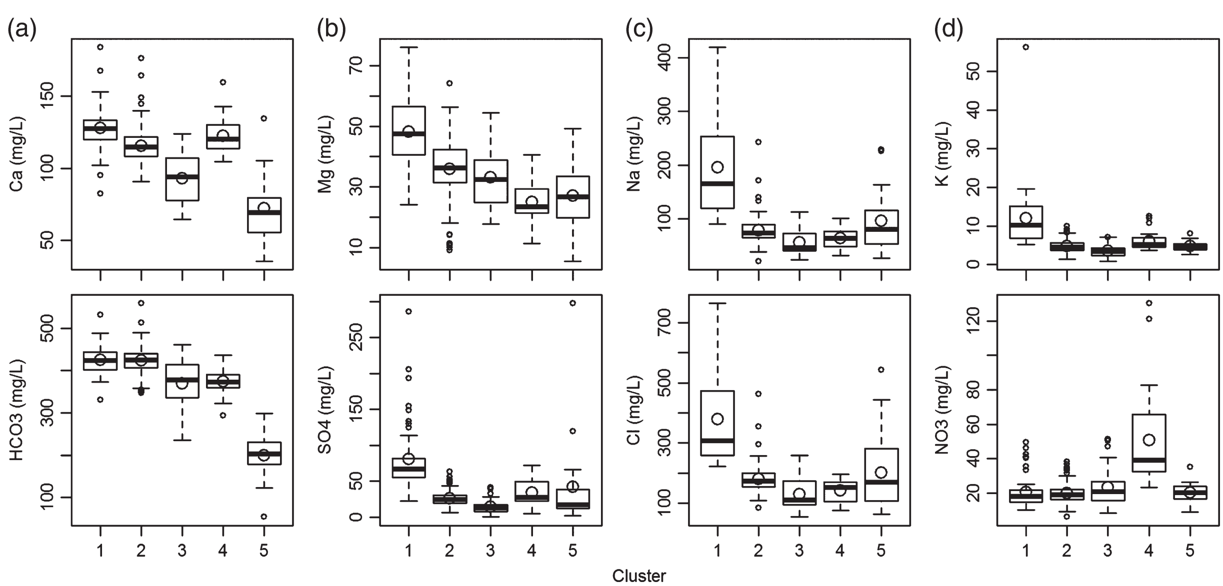

Figure 5. Box and whisker plots for the major ions. Sodium was not used for the analysis, it is presented for comparison purposes. The open circle represents the average parameter value.

quality of the aquifer. A significant change in the spatial distribution of the clusters is observed for the second sampling period. As shown in Figure 4b, clusters 1, 2 and 4 have fewer wells in the second sampling period than in the other three sampling periods. In addition, a new cluster (cluster 5) appears. Comparing sampling periods 1 and 2, some data in cluster 5 replaced data in clusters 1 and 2. In addition, in the east part of the study area, the location resulting in high nitrate concentrations in cluster 4 is occupied by data of low nitrate concentrations of cluster 3.

One hypothesis to explain the groundwater quality changes in sampling period 2 is that the precipitation during sampling period 2 was higher and caused the movement of water westwards. Corresponding, the data of cluster 3 spread to the west and data of cluster 1 were diluted, which forms the new cluster 5. The hypothesis is supported by the following facts. First, cluster 4 is absent in the east part of the study area for sampling period 2, and that the concentrations of Ca^{+2}, Mg^{+2}, Na^{+}, K^{+}, Cl^{-}, SO_4^{-}, and HCO_3^{-} in cluster 5 decreased by a factor

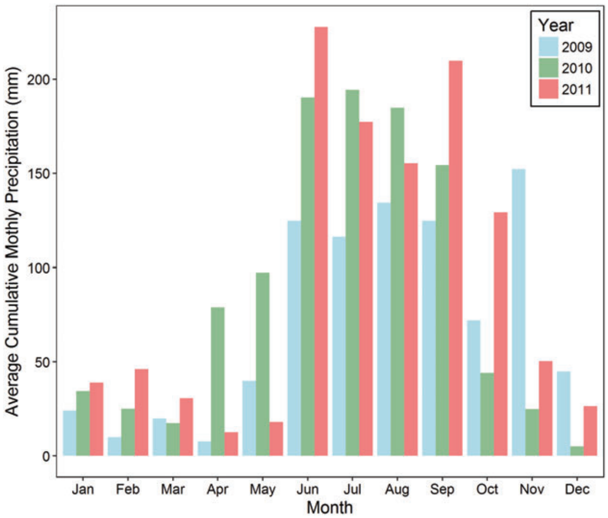

Figure 6. Average cumulative monthly precipitation in the State of Yucatan for the years 2009, 2010, and 2011.

of a half when comparing its concentrations with cluster 1. The only exception is the nitrate concentration, and this may be attributed to high nitrate concentrations in the top of the aquifer. Second, by examining the average cumulative monthly precipitation plotted in Figure 6 for each month during 2009, 2010, and 2011, we can see that sampling period 2 was carried out during the rainy season of 2010 and this may explain why the cluster 3 increases its influence towards the center of the study area and the dilution effect of cluster 5. However, the presence of cluster 4 at the west suggests that the distribution of the precipitation is also important. The distribution of the precipitation in April and June of 2010 is presented in Figure 7. The precipitation pattern shown in Figure 7 is different from the average annual precipitation, larger precipitation occurred in the west part of the Yucatan State. Therefore, cluster 5 is formed by groundwater samples from the east which had lower concentrations because of the amount of rain, and in the west because of the amount of precipitation and changes in the spatial distribution of the precipitation. Therefore, the amount and spatial patterns of precipitation are dominant variables of groundwater quality temporal variation in Yucatan. Cabrera et al. (2002) concluded that groundwater quality is not different in dry and rain seasons. Their study area was smaller (a zone at the south of Merida used to supply water to the city) and this behavior is also observed in sampling periods 3 and 4. However, because of the temporal and spatial resolution of the data used, this study leads to a different conclusion: precipitation cause variations in groundwater quality.

Optimal Number of Clusters

Finally, we discuss a procedure that may help to determine the optimal number of clusters. This empirical procedure starts with analyzing two clusters and sequentially increasing the number of clusters in the analysis. For this discussion Figures 2 to 4 can be used as reference for the dendrogram, biplots, and spatial distribution respectively. To start, we chose two clusters based on Figure 2. One cluster was cluster 1, and the other cluster was formed by merging clusters 2 to 5. The biplots showed that only the first principal component is required to justify this selection (Figure 3a). From the boxplots and the maps we identified the influence of the seawater intrusion, and all the properties of cluster 1 discussed in the previous section. We also derived the properties of the second cluster. To decide if the analysis should finish at this step, we did the same analysis for the case with three clusters. For this case we obtained the following clusters: cluster 1, cluster 5 and a cluster formed by merging cluster 2 to 4. Note that cluster 1 had the same elements, and thus all the information gained from the previous step about this cluster remains valid. At this step the first three principal components were required to justify this selection of clusters. As new information was gained in this step, we continued with the analysis with four and five clusters by gradually adding cluster 4, cluster 3, and cluster 2 to the analysis. Note that three principal components are still required for these steps. Adding a sixth cluster divided cluster 2 into two clusters with slight differences in calcium and magnesium. The selection of six clusters required four principal components given that principal component 4 was related to calcium and magnesium. We decided that five was the optimal number of clusters for this study because: 1) PC4 only explained 11.9% of the total variation of the data and 2) due to the karstic nature of the aquifer. The two clusters formed by splitting cluster 2 cannot be interpreted with the existent information, this may serve as the basis for future studies. We decided that five was the optimal number of clusters for this study.

Conclusions

The analysis of spatial patterns and temporal dynamics are important for the design of groundwater protection schemes, especially for karst aquifers that are vulnerable to contamination. An approach is developed to simultaneously analyze spatial and temporal variability of the processes governing groundwater quality. This approach was applied to the karstic aquifer of the state of Yucatan. Results of this work increase the knowledge about how precipitation and human contamination impact groundwater quality in Yucatan. Without anomalous precipitation events, the spatial variability of groundwater quality is stable and can be classified into four zones indicating the spatial influence of the different variables. The zone at the west (cluster 1) and the coastal area is characterized by the influence of sea water intrusion and the influence of water from the south rich in sulfates. The zone at the east (cluster 3) has the best water quality caused by the average annual distribution of precipitation. The zone in the middle (cluster 2) has higher concentrations of the all ions when compared to cluster 3 and there is no evidence of the affectation from seawater intrusion or anthropogenic contamination. Therefore clusters 2 and 3 are the zones with a natural composition of the water only influenced by

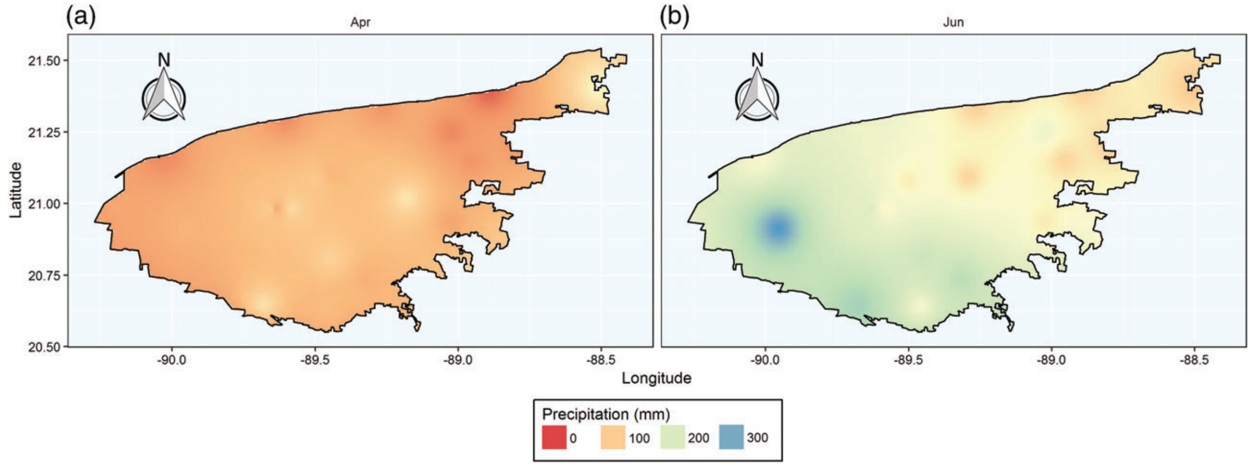

Figure 7. Distribution of the precipitation in the state of Yucatan. (a) April 2010. (b) June 2010.

carbonate rocks dissolution/precipitation and the average annual distribution of precipitation respectively. There are two localized contamination zones (cluster 4), one at the west and one at the east characterized by having the highest nitrate concentrations: one at the west under the biggest city of the Yucatan and one at the east under an agricultural zone. For the temporal variability the proposed approach allows to conclude that the groundwater quality is not steady in time. From the analysis of the snapshots of spatial variability of water quality, is concluded that the changes in time are influenced by the amount and distribution of precipitation. Those two factors cause the dilution of water samples belonging to other clusters and the formation of the cluster 5.

This study develops an approach to analyze the spatial and temporal variation of groundwater quality. The proposed approach of using cluster analysis for lumped data is proven to be useful for a simultaneous and efficient analysis of spatial and temporal variability of groundwater quality. We also discuss a procedure that can be used to decide the number of clusters to use. The proposed methodology is applicable to other areas given that no assumption is made about the groundwater in Yucatan.

Acknowledgments

J. P. A. would like to thank CONACYT and the Government of the State of Yucatan, Mexico for the financial support to the research project: YUC-2008-C06-108520. M. Y. was supported by NSF-EAR Grant 1552329. All the authors thank the Bruce Lafrenz and one anonymous reviewer for their insightful comments.

Authors' Note: The author(s) does not have any conflicts of interest or financial disclosures to report.

References

APHA, AWWA, and WEF. 2005. *Standard Methods for the Examination of Water and Wastewater*, 21st ed. Washington, D.C.: APHA-AWWA-WEF, 2005.

Appelo, C.A.J., and D. Postma. 2005. *Geochemistry, Groundwater and Pollution*, 2nd ed. CRC Press.

Back, W. 1999. The Yucatan Peninsula, Mexico. In *Karst Hydrogeology and Human Activities*, ed. D. Drew, and H. Hotzl, 13–30. Rotterdam, Netherlands: International Association of Hydrogeologists, A.A. Balkema.

Back, W., and B. Hanshaw. 1970. Comparison of chemical hydrogeology of the carbonate peninsulas of Florida and Yucatan. *Journal of Hydrology* 10: 330–362.

Belkhiri, L., A. Boudoukha, and L.A. Mouni. 2011. Multivariate statistical analysis of groundwater quality data. *International Journal of Environmental Research* 5, no. 2: 537–544.

Bivand, R., T. Keitt and B. Rowlingson. 2015. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.0-4. http://CRAN.R-project.org/package=rgdal (accessed June 2016).

Cabrera, A., J. Pacheco, E. Cuevas, J. Ramirez, M. Comas, and A. Camara. 2002. Hidrogeoquimica del Agua que Subyace a la JAPAY I, en Mérida, Yucatán, México. *Ingenieria* 6: 29–40.

Davis, J.C. 2002. *Statistics and Data Analysis in Geology*, 3rd ed. New York: John Willey & Sons.

Delgado, C., J. Pacheco, A. Cabrera, E. Batllori, R. Orellana, and F. Bautista. 2010. Quality of groundwater for irrigation in tropical karst environment: The case of Yucatan, Mexico. *Agricultural Waste Management* 97: 1423–1433.

Deutsch, W. 1997. *Groundwater Geochemistry. Fundamentals and Applications to Contamination*. Boca Raton, Florida: Lewis Publishers.

QGIS Development Team. 2015. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

Doehring, D.O., and J.H. Butler. 1974. Hydrogeologic Constraints on Yucatán's Development. *Science, American Association for the Advancement of Science* 186: 591–595.

Escolero, O.A., L.E. Marin, B. Steinich, and J. Pacheco. 2000. Delimitation of a hydrogeological reserve for a city within a Karstic aquifer: The Merida Yucatan example. *Landscape and Urban Planning* 51: 53–62.

Everitt, B., and T. Hothorn. 2011. *An Introduction to Applied Multivariate Analysis With R: Use R!*. Springer Science and Business Media.

Graniel, C., L. Morris, and J. Carrillo-Rivera. 1999. Effects of urbanization on groundwater resources of Merida, Yucatan, Mexico. *Environmental Geology* 37: 303–312.

Güler, C., G. Thyne, J. McCray, and K. Turner. 2002. Evaluation of graphical and multivariate statistical methods for classification of water quality data. *Hydrogeology Journal* 10: 455–474.

Hounslow, A. 1995. *Water Quality Data: Analysis and Interpretation*. Boca Raton, Florida.

Hussain, M., S. Ahmed, and W. Abderrahman. 2008. Cluster analysis and quality assessment of logged water at an irrigation project, eastern Saudi Arabia. *Journal of Environmental Management* 86, no. 1: 297–307.

Ibanez, J.G., M. Hernandez-Esparza, C. Doria-Serrano, A. Fregoso-Infante, and M.M. Singh. 2007. *Environmental Chemistry: Fundamentals*. New York: Springer Science and Business Media.

INEGI. 2010. *Censo de Población y Vivienda. 2010*. Yucatan, Mexico: Instituto Nacional de Geografia, Estadistica e Informatica. http://www3.inegi.org.mx/sistemas/mexicocifras/default.aspx?src=487&e=31 (accessed October 2014).

INEGI. 2002. *Estudio hidrologico del estado de Yucatan. Instituto Nacional de Geografia, Estadistica e Informatica*. del Estado de Yucatan, Mexico: Gobierno.

Johnson, R.A., and D.W. Wichern. 2007. *Applied Multivariate Statistical analysis*, 6th ed. Upper Saddle River, New Jersey: Pearson Education Inc.

Kaiser, H.F. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 1: 141–151.

Kaufman, L., and P.J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kim, J.H., R.H. Kim, J. Lee, T.J. Cheong, B.W. Yum, and H.W. Chang. 2005. Multivariate statistical analysis to identify the major factors governing groundwater quality in the coastal area of Kimje, South Korea. *Hydrological Processes* 19: 1261–1276.

Marin, L.E., B. Steinich, J. Pacheco, and O.A. Escolero. 2000. Hydrogeology of a contaminated sole-source karst aquifer, Mérida, Yucatán, Mexico. *Geofisica Internacional* 39: 359–365.

Mohammadi, Z. 2009. Assessing hydrochemical evolution of groundwater in limestone terrain via principal component analysis. *Environmental Earth Sciences* 59: 429–439.

Morel, F.M. 1983. *Principles of Aquatic Chemistry*. New York: John Wiley and Sons.

Nosrati, K., and M.V.D. Eeckhaut. 2012. Assessment of groundwater quality using multivariate statistical techniques in Hashtgerd Plain, Iran. *Environmental Earth Sciences* 65: 331–344.

Pacheco, J., A. Cabrera, and R. Pérez. 2004. Diagnóstico de la calidad del agua subterránea en los sistemas municipales de abastecimiento en el Estado de Yucatán, México. *Ingenieria* 8: 165–179.

Pacheco, A.J. 2013. Reserva hidrogeológica para el abastecimiento de agua con calidad para la zona metropolitana de Mérida, Yucatán. Final Report. Universidad Autónoma de Yucatán.

Pacheco, J., L. Marin, A. Cabrera, B. Steinich, and O. Escolero. 2001. Nitrate temporal and spatial patterns in 12 water-supply wells, Yucatan, Mexico. *Environmental Geology* 40: 708–715.

Pacheco, J., A. Cabrera, and L.E. Marin. 2000. Bacteriological contamination in the karstic aquifer of Yucatán, Mexico. *Geofisica Internacional Mexico* 39: 285–291.

Pacheco, J., and A. Cabrera. 1997. Groundwater contamination by nitrates in the Yucatan Peninsula, Mexico. *Hydrogeology Journal* 5: 47–53.

Peel, M.C., B.L. Finlayson, and T.A. McMahon. 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions* 4, no. 2: 439–473.

Pérez-Ceballos, R., J. Pacheco-Ávila, J.I. Euán-Avila, and H. Hernández-Arana. 2012. Regionalization based on water quality and physicochemical traits in the ring of cenotes, Yucatan, Mexico. *Journal of Cave and Karst Studies* 74, no. 1: 90–102.

Perry, E., J. Swift, J. Gamboa, A. Reeve, R. Sanborn, L. Marin, and M. Villasuso. 1989. Geologic and environmental aspects of surface cementation, north coast, Yucatan, Mexico. *Geology* 17, no. 9: 818–821.

Perry, E., G. Velazquez-Oliman, and L. Marin. 2002. The hydrogeoquality of the karst aquifer system of the northern Yucatan Peninsula. *Mexico International Geology Review* 44: 191–221.

PROCOMAR. 2003. Precipitacion media del estado de Yucatan. From Bitacora Ambiental del Programa de Ordenamiento Ecologico Territorial del Estado de Yucatan. SEDUMA, SEDESOL, SEMARNAT. http://www.bitacoraordenamiento.yucatan.gob.mx/ (accessed December 2015).

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/.

RStudio Team. 2015. *RStudio: Integrated Development for R*. Boston, Massachusetts: RStudio, Inc. http://www.rstudio.com/.

Schmitter-Soto, J., F. Comın, E. Escobar-Briones, J. Herrera-Silveira, J. Alcocer, E. Suarez-Morales, M. Elıas-Gutierrez, V. Dıaz-Arce, L. Marın, and B. Steinich. 2002. Hydrogeochemical and biological characteristics of cenotes in the Yucatan Peninsula (SE Mexico). *Hydrobiologia* 467, no. 1–3: 215–228.

Shrestha, S., and F. Kazama. 2007. Assessment of surface water quality using multivariate statistical techniques: A case of study of the Fuji river basin, Japan. *Environmental Modeling & Software* 22: 464–475.

Srivastava, P.K., D. Han, M. Gupta, and S. Mukherjee. 2012. Integrated framework for monitoring groundwater pollution using a geographical information system and multivariate analysis. *Hydrological Sciences* 57: 1453–1472.

Suk, H., and K. Lee. 1999. Characterization of a groundwater hydrogeochemical system through multivariate analysis: Clustering into groundwater zones. *Ground Water* 37, no. 3: 358–366.

Thyne, G., C. Güler, and E. Poeter. 2004. Sequential analysis of hydrochemical data for watershed characterization. *Ground Water* 42, no. 5: 711–723.

Vega, M., R. Pardo, E. Barrado, and L. Debán. 1998. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research* 32, no. 12: 3581–3592.

Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58, no. 301: 223–244.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Woocay, A.J., and Walton, J. 2008. Multivariate analyses of water quality: Surface and ground water interactions. *Ground Water* 43, no. 3: 437–449.

Wunderlin, A.M., M. Diaz, S. Ame, A. Pesce, M. Hued, and M.Á. Bistoni. 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquıća River Basin (Córdoba–Argentina). *Water Research* 35, no. 12: 2881–2894.