AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR016662

Water consumption patterns as a basis for water demand modeling

Noa Avni[1,2], Barak Fishbain[1,2], and Uri Shamir[2]

[1]Department of Environmental, Water, and Agriculture Engineering, Technion—Israel Institute of Technology, Haifa, Israel, [2]Faculty of Civil and Environmental Engineering, Department of Environmental, Water, and Agriculture Engineering, Technion—Israel Institute of Technology, Haifa, Israel

Key Points:
- Clustering facilitates regional water demand data analysis
- Clustering extracts meaningful insights on water demand generating factors
- Regional demand data analysis can improve large-scale water systems management

Supporting Information:
- Supporting Information S1

Correspondence to:
N. Avni,
noaavni@technion.ac.il

Citation:
Avni, N., B. Fishbain, and U. Shamir (2015), Water consumption patterns as a basis for water demand modeling, *Water Resour. Res.*, *51*, 8165–8181, doi:10.1002/2014WR016662.

Received 11 NOV 2014
Accepted 6 AUG 2015
Accepted article online 18 SEP 2015
Published online 17 OCT 2015

Abstract Future water demand is a main consideration in water system management. Consequently, water demand models (WDMs) have evolved in past decades, identifying principal demand-generating factors and modeling their influence on water demand. Regional water systems serve consumers of various types (e.g., municipalities, farmers, industrial regions) and consumption patterns. Thus, one of the challenges in regional water demand modeling is the heterogeneity of the consumers served by the water system. When a high-resolution, regional WDM is desired, accounting for this heterogeneity becomes all the more important. This paper presents a novel approach to regional water demand modeling. The two-step approach includes aggregating the data set into groups of consumers having similar consumption characteristics, and developing a WDM for each homogeneous group. The development of WDMs is widely applied in the literature and thus, the focus of this paper is to discuss the first step of data aggregation. The research hypothesis is that water consumption records in their original or transformed form can provide a basis for aggregating the data set into groups of consumers with similar consumption characteristics. This paper presents a methodology for water consumption data clustering by comparing several data representation methods (termed Feature Vectors): monthly normalized average, monthly consumption coefficient of variation, a combination of the monthly average and monthly variation, and the autocorrelation coefficients of the consumption time series. Clustering using solely normalized monthly average provided homogeneous and distinct clusters with respect to monthly consumption, which succeed in capturing different consumer characteristics (water use, geographical location) that were not specified a-priori. Clustering using the monthly coefficient of variation provided different, yet homogeneous clusters, clustering consumers characterized by similar variation trends that were closely related to consumer water use type. The concatenation of these two Feature Vectors provided further insight into the relationship between consumption patterns and variability of consumers. An autocorrelation Feature Vector provided results that can form a basis for constructing a time-series model that is based on a group of resembling time series. The approaches presented here are steps toward utilizing the increasing amount of available water consumption data and data analysis techniques to facilitate the modeling of water demands in larger and heterogeneous regions with sufficient resolution.

1. Introduction

Planning and management of regional water systems (RWS) is of great importance in the era of global climate change and rapid population growth. RWS planning and management encompass infrastructure capacity expansion, development plans, operation, and policy schemes (water rates, subsidies, regulation, etc.). Information relevant to future planning is often unavailable at the time of decision making. Thus, methods to estimate future driving forces affecting the RWS are fundamental if a robust and resilient management is desired. Future water demand is a main driving force in water system management [*Gleick et al.*, 2003]. Consequently, Water Demand Models (WDMs) have evolved in the past decades. WDMs may be used to gain insights on water consumption behavior or to forecast water demand as an input for decision-making processes in water distribution system (WDS) operation policy and infrastructure planning.

The literature on WDMs focuses on the urban and agriculture sectors. Urban demand modeling, as reflected in *House-Peters and Chang*'s [2011] review, is applied through a variety of methodologies (e.g., multivariate regression, Bayesian Maximal Entropy and Ordinary Least Square regression). The data used for developing

the various WDMs are often panel data, which is a set of explanatory variables such as climate records, water price, rate structure, age, education, and family size, considered as candidates to explain or forecast the water demand. Other WDM apply time-series analysis, which uses only historical records and reflect the inherent auto-correlation structure of the water use pattern over time [*Maidment and Parzen*, 1984; *Jain et al*., 2001; *Tiwari and Adamowski*, 2013].

Agriculture water demand modeling has been addressed mainly by an economical perspective, namely, estimating the water's economical value [*Howitt*, 1995; *Berger*, 2001; *Fisher et al*., 2002; *Medellín-Azuara et al*., 2012]. However, these models are not coupled with the physical WDS, namely, they do not specify where in the system water demands occur, and are thus less applicable to high-resolution, regional water demand modeling.

One of the biggest challenges in WDMs is the heterogeneity of the consumers served by the WDS. As the WDS size increases, additional types of consumers (i.e., connections to the grid) and consumption patterns come into play and should be accounted for in the WDM. Moreover, when a high-resolution, *regional* WDM is desired, accounting for regional heterogeneity becomes all the more important.

Several studies target water demand modeling in larger areas exist: *Babel et al*. [2007] apply multiple regression analysis to select a daily water demand function and relevant explanatory variables for water use in Kathmandu Valley (Nepal), having an area of 900 km^2; *Worthington et al*. [2009] apply multiple-regression to compute consumer price elasticity in 11 local governments in Queensland, Australia, based on tariff structure and precipitation data; and *Schleich and Hillenbrand* [2009] apply a log-log regression variant model to estimate the impact of economic, environmental, and social determinants on water price elasticity in 600 water supply areas in Germany.

These regional WDMs studies are few and often fit a single model to a relatively large area (e.g., supply area, a county, a basin), which ignores the region's heterogeneity of consumption. This may lead to a weak relationship between the explanatory variables and water demands. *Gutwein and Lang* [1993] predict agriculture water use in the Imperial Valley, CA, based on crop acreage and climatic variables data, with what the authors term as "limited success." *Franczyk and Chang* [2009] model Oregon State's municipal supply, irrigation, and total water withdrawals on a county-level scale, aiming to improve water forecasts using the degree of spatial correlation between counties. Despite of a certain improvement, correlations between explanatory variables and water demands on the county-level, were rather weak, and the authors suggest that "a more extensive analysis is needed for determining the relationship between municipal water withdrawals and other explanatory variables using long-term data with a finer spatial scale (e.g., metropolitan scale)."

Modeling regional water demands by fitting a single model to each small area requires the processing of large amounts of data (e.g., census-tract or household level), and is thus less applicable for regional scale models. On the other hand, over-aggregating the data may cause a loss of information if consumers with different consumption behaviors are lumped. Thus, the challenge in regional water demand modeling is to set a proper level of aggregation, i.e., one that reflects well the heterogeneity while maintaining a parsimonious representation.

Homogeneity, for the scope of this research, refers to similar behavior of a group of consumers (i.e., a connection to the regional water grid)—a similar consumption trend and/or similar response to the explanatory variables that is reflected in that trend. Even within a relatively small geographical region, several consumption patterns may exist. On the other hand, consumers in different areas may have a similar consumption pattern (e.g., two cities). Therefore, a geographical-based aggregation may not reflect well the regional heterogeneity [e.g., *Franczyk and Chang*, 2009; *Schleich and Hillenbrand*, 2009]. Another possible approach is consumer-type aggregation, that is, to aggregate the consumers based on their type of water use (e.g., farmers and domestic consumption). However, studies in the field of water demand modeling focus on single-type of consumer [e.g., *Olmstead et al*., 2007; *Lee and Wentz*, 2008; *Lee et al*., 2010] and do not address mixed water use. Nevertheless, applying consumer-type aggregation has several drawbacks: first, such an administrative classification (i.e., data available from the water authority) is often not available in a regional data set and even if it is available it will be too general to infer the consumer water consumption pattern (e.g., a classification of agriculture may be insufficient since farmers differ in their water use patterns depending on the type of agriculture they cultivate—field crops, orchards, or greenhouses); second, a

consumer may have several water use types (e.g., a rural settlement that uses water for irrigation *and* domestic consumption), which create a unique pattern. An example for the limitation of administrative consumer-type classification is presented in section 3.5.

This paper presents a two-step approach for developing regional WDMs: aggregating the regional data set to form homogeneous groups of consumers, and developing a WDM for each group. Thus, the final WDM is a set of models, each having a better fit for the data subset on which they are built. The approach depicted here is a step toward utilizing the increasing amount of available water consumption data and data analysis techniques to facilitate the modeling of larger, heterogeneous regions with sufficient resolution.

Techniques and methodologies for water demand modeling are widely available in the literature. Thus, the focus of this study is to discuss the first step—finding the proper level of aggregation, and to demonstrate its relevance to the next step—developing a WDM for each subgroup. The methodology for aggregating the data set is based solely on historical records of water consumption, and involves two main phases: First, the raw data are transformed into Feature Vectors (FVs) that highlight certain aspects of consumers behavior, and then a clustering procedure is performed with the objective of finding groups that are as distinct from each other as possible, with high similarity among group members.

The rest of the paper is organized as follows: Section 2 describes water consumption data characteristics and data representation metrics and methodology; section 3 presents the results of the experiment conducted using the different algorithms and Feature Vectors. A discussion and conclusions is given in sections 4 and 5.

2. Data and Methods

2.1. Data

A consumer, for the scope of this paper, is a monitored water connection from the Israeli national water grid (e.g., municipality, single farmer, regional council, etc.). The data set available for the research was a 19 year period (1994–2012) monthly data of 5141 consumer connections, distributed throughout the Israeli grid. The data are obtained from Mekorot, the Israeli national water company, which provides 70% of the water consumption in Israel.

The data set is composed of cities with varying sizes, farmers, industries, collective communities (Kibbutz), and agrarian and communal settlements, located throughout the National grid. It is observed that the monthly consumption patterns of domestic consumption (i.e., urban or residential settlements) is relatively stable, following the Israeli weather of hot summer (July–August), unstable weather during the spring and autumn (March–June and September–November, respectively), and moderate winter with high precipitation variability, which, despite of the country's small size, is highly location-dependent. The annual volume, however, may change due to natural demographical changes as well as government policy of regional development plans. The indoor versus outdoor water demand varies according to the residential structure, with some cities having larger green areas than others. The water demands of farmers presents a much higher variability, both on the monthly and annual scales. The monthly variability may be the result of varying precipitation and temperatures, while the annual variation may be due to structural changes (e.g., change in subsidies and water rates) and farmer decisions such as changes in crop type, use of greenhouses, orchards, and so on.

The water use of collective communities ranges from strictly domestic use to a mixture of industrial, domestic and agriculture water use, depending on the type of economy utilizes by the collective community. Agrarian settlements may have different types of land uses. In the past, the water used for agriculture in these settlements was metered with their domestic consumption, while in recent years the metering of the two water uses was separated. The communal settlements in Israel use water mainly for domestic consumption, and are characterized by private housing and public areas that require more outdoor irrigation.

From the full data set of water consumption, a smaller subset of 105 consumers for the years 2002–2007 was used for developing and testing the first of the two-stage methodology for data clustering. The reduction from the full data set stems from two constraints: First, during the preparation of the data we have encountered a problem with the records continuity: a consumer is represented via its unique number in the

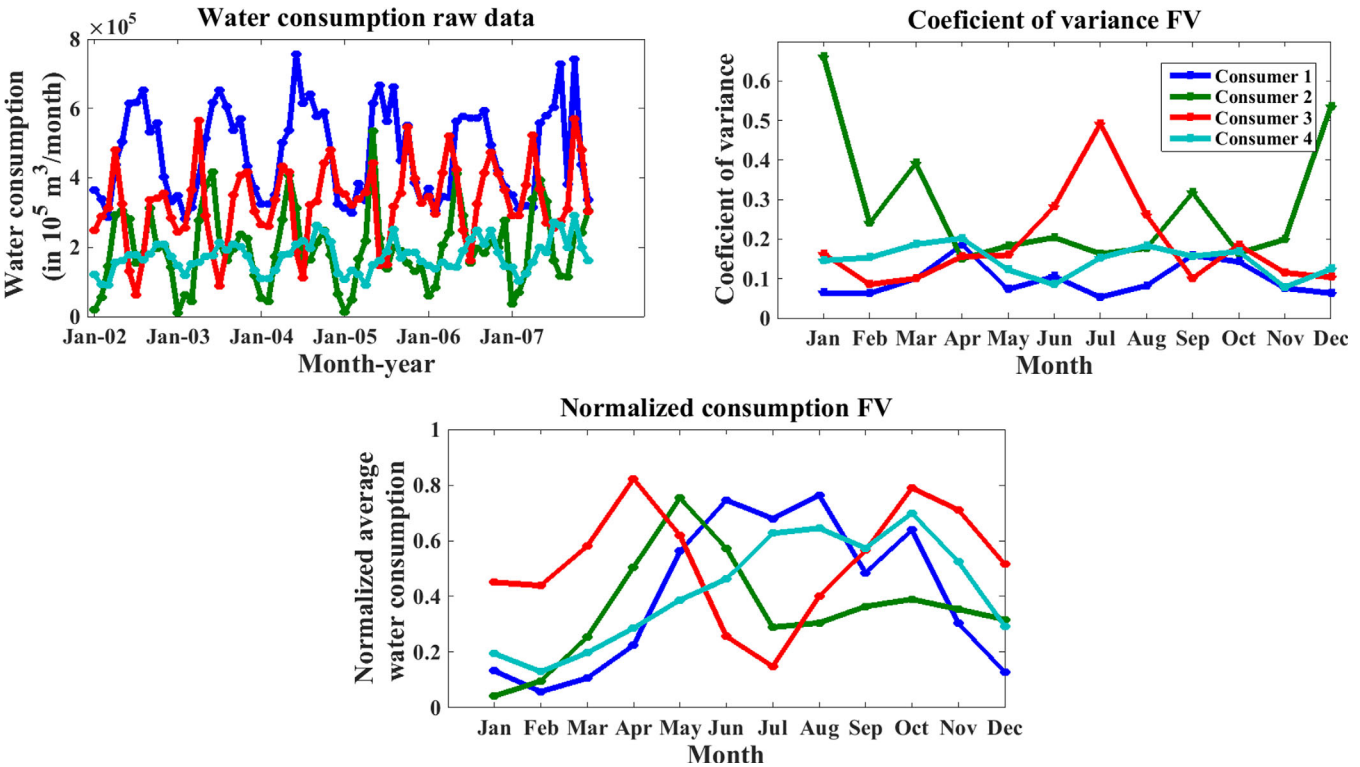

Figure 1. An example of raw data (a), Monthly normalized average consumption FV ($10^5 m^3$/month) (b), and coefficient of variation FV (c), of four consumers over 6 years (2002–2007).

National Water Company records. Due to regulatory changes in the past 15 years, many consumers either changed their number, or were grouped together with other consumers in the region under a water corporation—either an agricultural or a domestic one. Thus, it was impossible to keep track of all the stages without a scrupulous work of checking each consumer's history. Since this study aimed at presenting the approach in principle, we decided to reduce the data set to a manageable one, where changes in consumer ID could be tracked and corrected if necessary. Second, demonstration and discussion of the methodology requires a clear representation of the results, which we found to be clearer using the small data set.

2.2. Methods

Two steps precede the clustering of any data set: preprocessing of the raw data to form Feature Vectors (FVs) which serve as an input for the clustering algorithm, and determining the appropriate number of clusters—groups of consumers for the scope of this research.

2.2.1. Feature Vectors

When dealing with a large and heterogeneous data set of water consumption records, performing a clustering procedure on the raw data, may lead to clusters that do not meet clustering objectives [*Yang et al.*, 2013]. For example, differences in consumption magnitude may mask interesting patterns such a monthly variability or a recurring pattern, thus driving the clustering procedure toward forming clusters with similar consumption magnitudes. Figure 1a presents the raw water consumption (m^3/month) of four consumers from the data set, for a 6 year period (2002–2007). Although the raw data present a general seasonal trend, it is difficult to determine whether these consumers have similar consumption patterns (e.g., when do peaks and troughs occur) or not.

Therefore, a common practice in data analysis is to transform the raw data into FVs that highlight data characteristics relevant to the clustering objectives (i.e., finding groups of consumers with similar consumption pattern, or similar monthly variability). Even for a well-defined objective, several FVs can be tested, yielding different results [*Yang et al.*, 2013]. Therefore, the proposed methodology explores several FVs based on monthly statistical moments, each highlighting different characteristics of water consumption: normalized monthly average water consumption, monthly consumption coefficient of variation, a

concatenation of the two FVs, and the autocorrelation coefficients of the monthly water consumption time series.

2.2.1.1. Normalized Monthly Average

Assuming that the general consumption pattern is indicative of a certain response to the explanatory variables, an average of the monthly consumption can reflect a consumer's trend. Thus, the first FV explored is the monthly averaged normalized consumption (denoted NA-FV), which is constructed by: (*i*) normalizing the monthly consumption of each consumer over the 6 years of record to the range [0,1], and (*ii*) averaging the normalized consumption for each month to form a 12-entry FV. The NA-FV emphasizes peaks and troughs in the raw data, creating larger dynamic range on the scale [0,1], as can be seen in Figure 1a.

2.2.1.2. Monthly Coefficient of Variation

The variability in the monthly water consumption patterns often indicates water use type, and may thus indicate a certain response to water demand determinants. For example, a farmer cultivating field crops might show a high annual variability over the years of records (e.g., due to uncertainty in rainfall), compared with a residential area which is sensitive to water prices that affect its outdoor irrigation and has lower monthly consumption variability.

The coefficient of variation (CV) measures the variability of a data series independently of its measurement unit. Thus, the CV is used herein as a measure of a consumer's monthly variability. The CV Feature Vector (denoted CV-FV) is computed by dividing the standard deviation of the monthly consumption by the average monthly consumption, both based on the years of records available (Figure 1c). For example, the CV for January, based on a 6 year record (2002–2007) is given by equation (1):

$$CV_{Jan} = \frac{std(X_{Jan-02,\ldots,}X_{Jan-07})}{\text{average}(X_{Jan-02,\ldots,}X_{Jan-07})} \quad (1)$$

2.2.1.3. Concatenation of the Monthly Average and CV

The third FV evaluated was the NACV-FV, which is the concatenation of the NA-FV and CV-FV, forming a single 24-entry FV for each consumer, which provides information on both the consumer's consumption pattern and the inherent variability. A concatenation of FVs [e.g., *Yang et al*., 2013] increases the amount of information on the data instance. On the other hand, such a concatenation may add irrelevant information and should thus be used with caution.

2.2.1.4. Autocorrelation Coefficients

Consumers with similar water consumption patterns may be expected to present similar time series. Thus, another potential FV is the consumer autocorrelation (AC) function, which standardize the autocovariances for different lags of the time series [*Nelson*, 1973]. Figure 2 presents the correlogram (a graph of autocorrelation function of a time series [*Nelson*, 1973]) of the same four consumers. The four consumer correlograms show different AC between the various lags: e.g., consumers 2 and 3 have positive AC for lag-1 and negative AC for lag-2, whereas 1 and 4 have a positive AC for both lags. Also, the four consumers has a seasonal pattern, as reflected by the AC of lag-12.

Analyzing raw consumption data may reveal different responses of different consumers to the explanatory variables driving their water demand. The above illustrations in Figure 1 and Figure 2 show that constructing different FVs provides additional sensitivity that can better separate water consumers with respect to their response to explanatory variables of relevance to RWS management. For example, consumers 2 and 4 that presented a relatively similar bi-modal pattern with a peak in April/May and a trough in July (Figure 1b), but presents a rather different water consumption patterns in terms of variability (Figure 1c); consumers 2 and 3, despite their similar consumption pattern (reflected by their AC function, Figure 2), have different monthly variability trend, namely high variability in December and January, and a relatively flat variability throughout the year (0.1–0.2 monthly CV), respectively.

2.2.2. Selecting the Number of Clusters and Clustering Algorithm

A "good" clustering minimizes within-cluster dissimilarities (e.g., distances) and maximizes between-cluster dissimilarities. The quality of a clustering results dependents on the data itself (i.e., whether it contain distinguishable clusters) and the number of clusters (K) specified. Selecting the number of clusters is a challenging task since in most cases previous no knowledge exists on the number of underlying patterns in the data. Therefore, a heuristic or a trial and error procedure are used to select K.

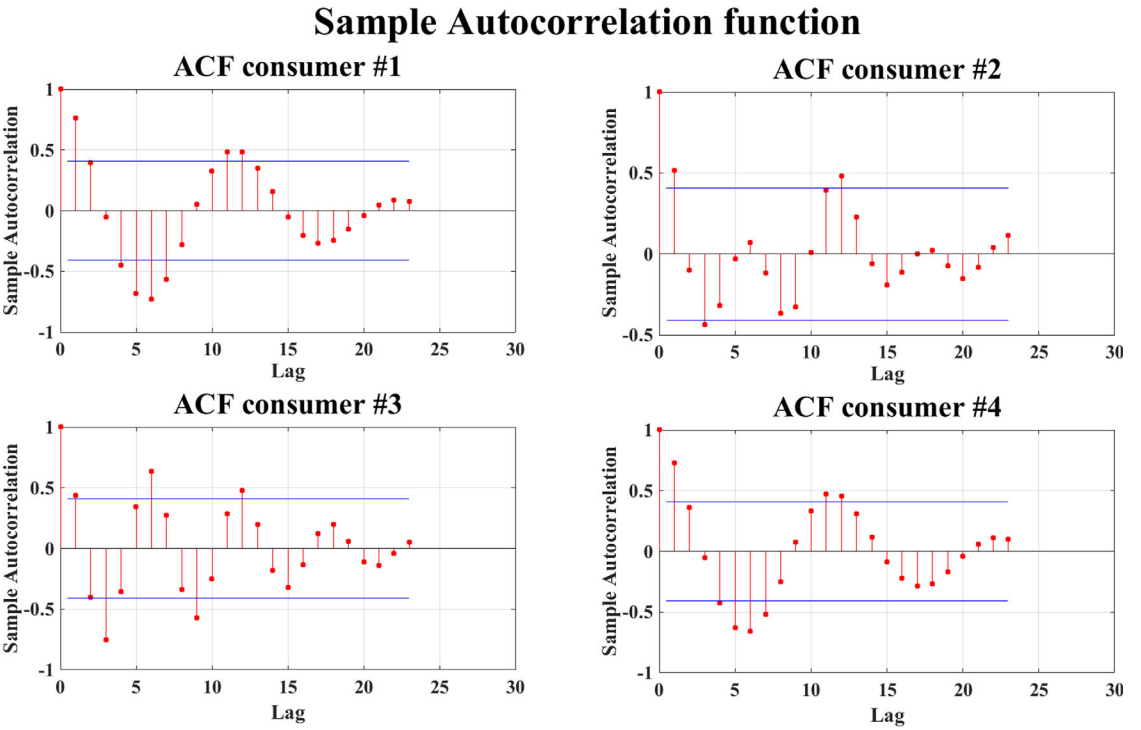

Figure 2. Example: Autocorrelation function (ACF-FV) for the four consumers in Figure 1.

A common tool for selecting *K* is the *silhouette plot* [*Rousseeuw*, 1987]. Silhouette plots represent the clustering results in terms of proximities between data set instances. Thus, they serve as a metric for quantifying the clustering structure of the data (i.e., are there well defined clusters in the raw or transformed data?) and the particular *K* used.

To construct the clusters' silhouette plots, the *silhouette widths, s*(*i*), measuring the suitability of each data instance to its cluster, are computed for each instance *i* [see *Rousseeuw*, 1987] and plotted against the cluster they were assigned to, as can be seen in Figure 3. *s(i)* values are on a [−1,1] scale, where $s(i) \approx 1$ denotes a well-classified data point, and $s(i) \approx -1$ implies that the point is most likely misclassified. A "good" clustering will have a "wide" silhouette (w.r.t. the ordinates).

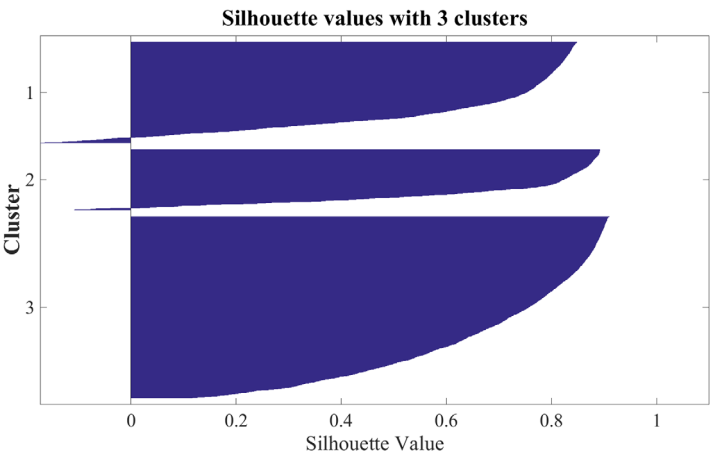

Figure 3. Example of a silhouette plot (www.mathworks.com). The ordinate represents the instances belonging to each cluster (1–3), ordered by a decreasing silhouette width (abscissa).

The three clusters in Figure 3 have relatively high *s(i)* values (higher than 0.6), with some of the instances in clusters 1 and 2 might be misclassified (negative *s(i)*). Thus, this data set presents a relatively good clustering structure.

The method of silhouette plots is useful when the instances' proximities are on a ratio scale (as in the case of Euclidean distances) and are known to work best in a situation with roughly spherical clusters. Due to the high variability in water consumption data, the silhouette width values obtained in the experiment were rather

low (average silhouette width of 0.3–0.4), showing a relatively moderate clustering structure. Moreover, the post analysis, where the clusters were inspected visually, showed that the silhouette width criterion favored K values which did not form sufficiently distinct clusters. Thus, the number of clusters was selected by evaluating K values in the neighborhood of the K value that seemed favorable according to the silhouette plot.

2.2.3. Clustering

The clustering method used is K-means clustering [*MacQueen*, 1967] which is an iterative algorithm that seeks to minimize an error term based on the distance between each data instance and K cluster-centroids. K-means is a standard method in unsupervised learning and is widely applied in various data analysis studies [*Bullinger et al*., 2004; *Jain*, 2010].

The K-means algorithm receives a dissimilarity matrix (i.e., distances between pairs of data instances) as an input. Thus, in order to apply K-means, a suitable distance function should be selected. Common metrics are the Euclidean distance, often used for points in Euclidean space [*Tan et al*., 2005]; the Mahalanobis distance that considers the spread of the points in the multidimensional space [*Benaichouche et al*., 2013; *Weinberger and Saul*, 2009]; and the cosine distance which is one minus the cosine of the angle between points and is commonly used for multidimensional data [*Yiakopoulos et al*., 2011]. This study proposes a general framework for developing regional WDMs and thus any clustering algorithm that uses a dissimilarity matrix (e.g., distances between pairs of data instances) as an input may serve in the clustering stage.

3. Results

3.1. Experimental Setup

The experimental part included two phases: (*i*) Clustering the data using four types of FVs; and (*ii*) discussing the applicability of clustering as a preprocessing stage in developing panel data or time series based regional WDMs through a postanalysis of the clustering results.

The four FVs were used as a representation of the raw data: the normalized monthly average consumption (NA-FV), the monthly coefficient of variation (CV-FV), a concatenation of the two FVs (NACV-FV), and the monthly autocorrelation function of the consumers (ACF-FV). The number of clusters, K, was selected using the silhouette width combined with a *visual analysis* of the consumption patterns assigned to each cluster. K was selected such that it provided sufficiently distinct clusters that were internally homogeneous (i.e., similar consumption patterns in each cluster and distinct patterns in different clusters). The distance function used with each FV was selected following the same criteria of creating distinct clusters having internal homogeneity.

The objective of the postanalysis was to examine the research hypothesis that clustering, done solely based on water consumption data, provides groups of consumers that are homogeneous with respect to explanatory variables such as climate (indicated by the geographical region), type of water use, etc. If so, then a model fit to each cluster may reflect better the relationship between the explanatory variables and the water consumption. To do so, the common characteristics of the consumers assigned to each cluster were analyzed.

3.2. Results for the Four Feature Vectors

Sections 3.1–3.4 present the clustering results using the four different FVs, along with a discussion of their relevance for extracting the explanatory variables that effect each group (panel data based models); Section 3.5 provides a comparison between aggregation based on water consumption data clustering and aggregation by administrative consumer type (i.e., as provided by the water company/authority).

3.2.1. Normalized Average (NA-FV)

Figure 4 presents the cluster silhouettes for $K = 5$ and $K = 6$, along with the average silhouette width of each cluster using the NA-FV. Clustering using $K = 5$ clearly provide a better cluster (see bar chart in Figure 4) and overall average silhouette width than $K = 6$: 0.4114 versus 0.277 and had less "misclassifications" (i.e., negative *s(i)* values). Nevertheless, the consumption patterns assigned to each cluster using $K = 5$ were not necessarily similar. For example, the consumers assigned to cluster #3 using $K = 5$ (Figure 5) have several different patterns: bimodal, and unimodal with a sharp peak during October–November or April–May (see supporting information Figure S1 for $K = 5$ clustering results). As stated earlier, the cluster silhouettes are helpful when the data contain well-defined clusters. The low average cluster silhouette values (<0.5), even for $K = 5$ shows

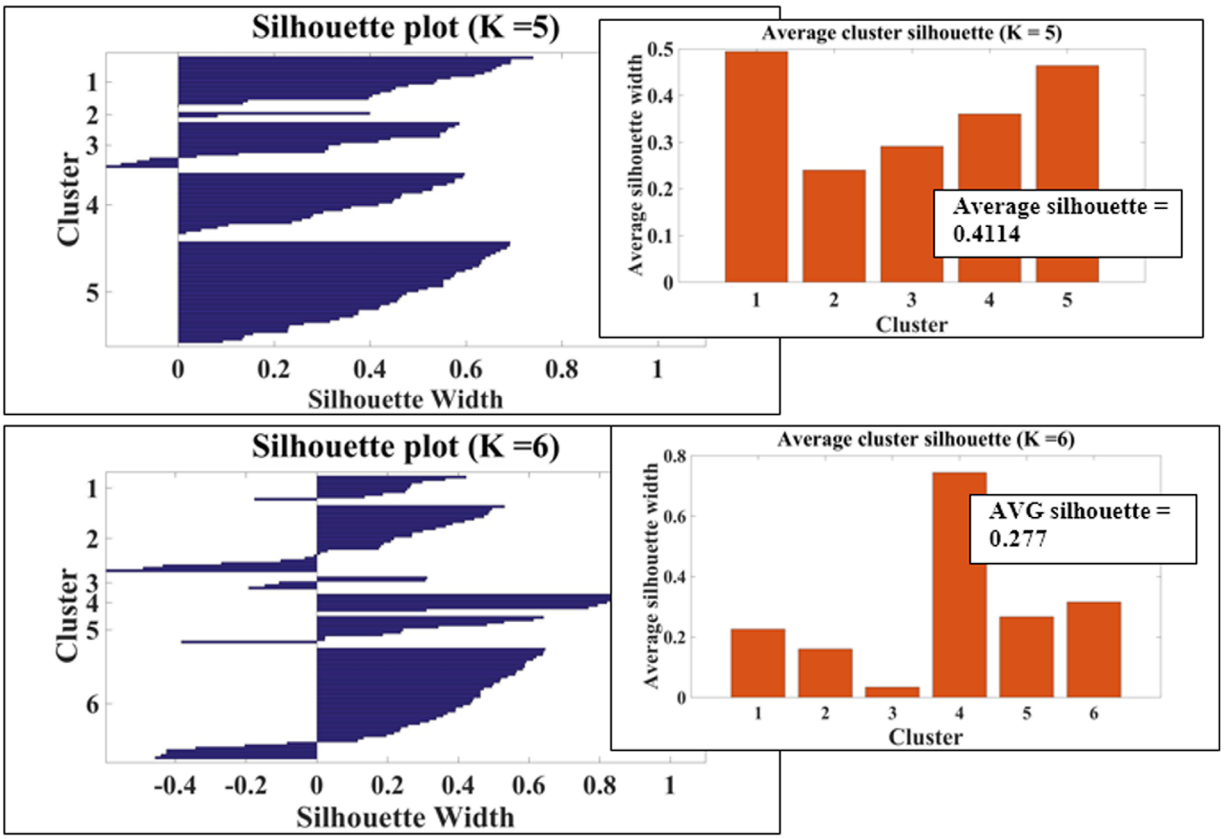

Figure 4. Silhouette plots for K = 5 and K = 6, with average cluster silhouette width (value).

that the clustering structure of the data is moderate. Therefore, the silhouette method should be used with cautious, and requires further validation by inspecting the patterns (consumers) assigned to each cluster.

Figure 6 presents the water consumption patterns of the consumers assigned to each cluster using $K = 6$. The patterns grouped together when using $K = 5$ were now divided into separate clusters, that is, when $K = 6$ was used the consumers in cluster #3 (Figure 5), having a sharp peak in April–May were assigned to

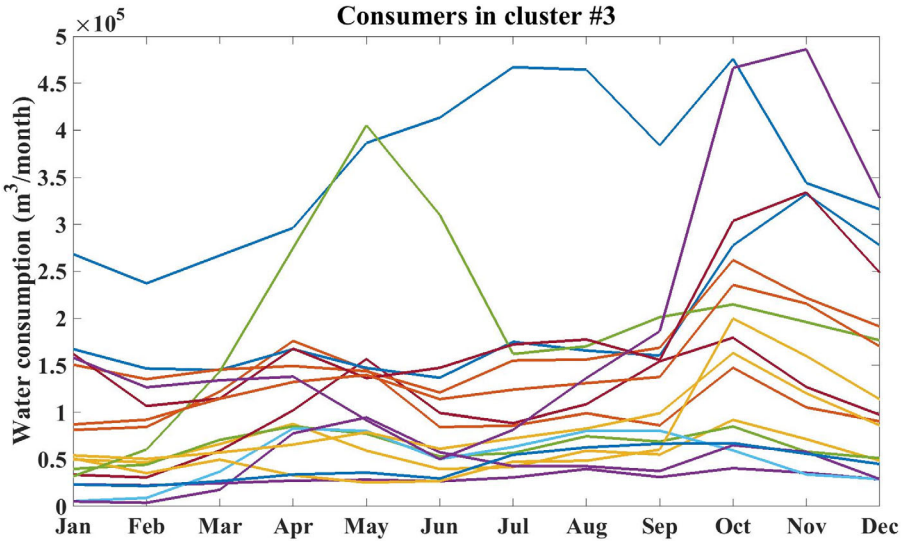

Figure 5. Water consumption (m^3/month) patterns for the 16 individual consumers in cluster #3 for K = 5 (NA-FV).

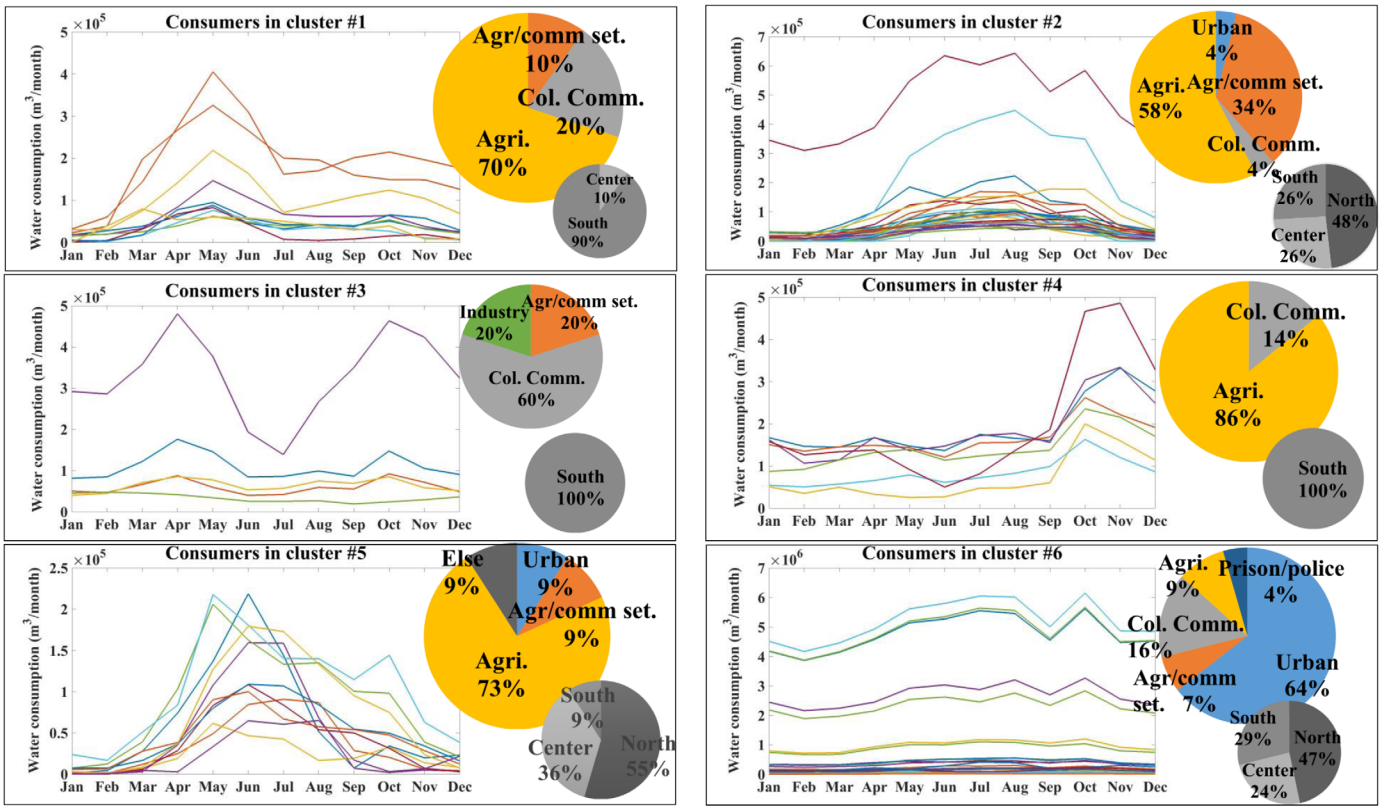

Figure 6. Water consumption (m^3/month) patterns for consumers in each cluster for K = 6, NA-FV. Each colored line corresponds to a single consumer (Note: for clarity of visualization, a different vertical scale is used in some of the figures); the pie charts present the common cluster characteristics—geographical location (greyscale) and main consumer types.

cluster #1. Similarly, the consumers with sharp increase in October–November were assigned to cluster #4. Thus, $K = 6$ was used for the NA-FV. The distance function used was the cosine function [*Yiakopoulos et al*., 2011], which provided clusters that were more homogeneous than those obtained by other distance measures (e.g., Euclidean or Mahalanobis).

The results for $K = 6$ show clear, distinct patterns assigned to each cluster. The consumers in clusters #2 and #6 have a gradual increase during February–August, followed by another peak in October; however, those in cluster #2 have a sharper increase, which justified the formation of two separate clusters for these consumers. Clusters #1 and #5 have earlier consumption peaks—during May and June, followed by another smaller peak in October. However, the decrease after the May/June peak for cluster #1 is more moderate than that in cluster #5, providing a justification for their assignment into a separate cluster. Cluster #3 has a bimodal pattern—with peaks during April and October, while cluster #4 has a unimodal pattern, with a distinct peak during October or November. The results show that clustering using the NA-FV follows the consumption patterns closely, i.e., normalizing and averaging the raw data extracts the features important for representing consumption trends.

The clustering was based on consumption data only, without recourse to information regarding the types of consumers and their expected behavior. Using the clustering results, it is instructive to note the common characteristics of the consumers that are grouped together. If the clustering points to homogeneous and distinct groups of consumers (i.e., sharing characters such as same geographic location and water use type), then by clustering, one can improve WDMs that are based on panel data. The following discussion will elucidate this idea.

In Israel, climate conditions may vary considerably. Thus, a clustering that reveals correspondence between consumption patterns and geographical location may indicate that the group's water consumption is influenced on climatic variables. Consequently, these variables may be relevant for modeling its water consumption. Similarly, different consumer types may have different response to socio-economic, demographic, and regulatory variables. Thus, revealing relationships between consumption patterns and consumer type may

indicate on sensitivity to these explanatory variables. In large data sets, a detailed classification (e.g., specific type of agriculture or residential structure) of each consumer is often not readily available and requires cross reference with other databases. If the consumer type can be revealed by clustering based solely on consumption patterns, this will facilitate the construction of panel data-based WDMs for regional WDS.

The pie charts in Figure 6 present the cluster composition with respect to geographical area and consumer type (a discussion of the consumer types particular to the Israeli water economy can be found in section 2.1). Clusters #1 and #5 have similar consumption patterns with a difference in timing of peak demands. The common characteristics analysis shows a majority (∼70%) of agriculture consumption in both clusters, which is in line with their pattern similarity. However, the consumers in cluster #1 are closely located in the Southern Negev region, while those in cluster #5 are distributed throughout the grid. This difference in geographical location may explain the difference in peak demand occurrence of these two groups.

Cluster #3 is composed mostly of collective communities and rural settlements, which are located around the Dead Sea. These consumers economy is based on crop fields and date palm plantation, and they are located in a region with very distinct climatic variables (precipition $<$50 mm/yr and summer temperatures of 32–39°C (www.ims.gov.il)). The majority of the consumers in cluster #4 are located in the Northern part of the Negev. However, unlike the consumers in cluster #1, located in the same region, these consumers use reclaimed wastewater, which is used for irrigating specific types of crop. The unique pattern of cluster #4, with a single peak during October–November, indicates the irrigation of different crops than those irrigated by the consumers in cluster #1. Consequently, these consumers can be affected differently by new water pricing policies or even change in fertilizer costs.

Nevertheless, it should be noted that similar average water consumption patterns are not always indicative of water use and consequently of the driving forces governing the water demand. The consumers in clusters #2 and #6 have similar geographical distribution and relatively similar consumption patterns. Still, their composition is different. Cluster #2 is composed mainly of Agriculture and Rural/collective communities (58% and 34%, respectively), and cluster #6 has a majority of urban water consumers (64%). And indeed the algorithm, using a sufficiently large K, was sensitive enough to distinguish between domestic and agriculture consumers, based solely on their normalized average consumption patterns.

The size of K is fundamental to the conclusions that can be extracted from the clustering results. $K = 6$ captures well the different patterns in the data set: while this is a relatively small number of clusters, each consumer group is different from the others both in its consumption pattern and common characteristics. Decreasing the size of K would mask phenomena revealed when a larger K value is used.

3.2.2. Coefficient of Variation (CV-FV)

Another potential FV is the monthly consumption variability, represented by the CV index. Consumers often differ in their monthly CV values. Thus, high, modest, or low monthly CV during certain periods of the year may indicate on consumers with different water use and thus suggest that different driving forces govern their water demand.

K-means for the CV-FVs was applied using the Euclidean distance function, which reflects the difference between the FVs better than the Cosine function used for the NA-FV. K was set to 5, using the same procedure used for NA-FV. While the best average cluster silhouette was obtained for K = 3 (See supporting information Text S2 and supporting information Figure S2) further increase to $K = 5$ increased the homogeneity of the clusters, as can be seen next, and thus $K = 5$ was used.

Figure 7 presents the consumer CV-FVs assigned to each cluster. Each insert graph represents the FV of a single consumer. The common cluster characteristics of each cluster are depicted next to each subplot. Clusters #1 and #2 have a relatively large CV amplitude, ranging from 0.2 to 2.4 and 0.3 to 2.4, respectively (Figures 7a and 7b). The consumers in cluster #1 have high variability from November to March, the rainy season in Israel; the consumers in cluster #2 have high variability during the same period, and higher variability during other months as well (CV≈0.4 versus CV≈0.2). The cluster characteristics show that these clusters are dominated by agricultural consumption (80% and 73%, respectively). However, the geographical location of the consumers differ—cluster #1 has mostly southern consumers, and cluster #2 has a majority of Northern consumers. Thus, the variations in consumption may be indicative on climate conditions related to water consumption.

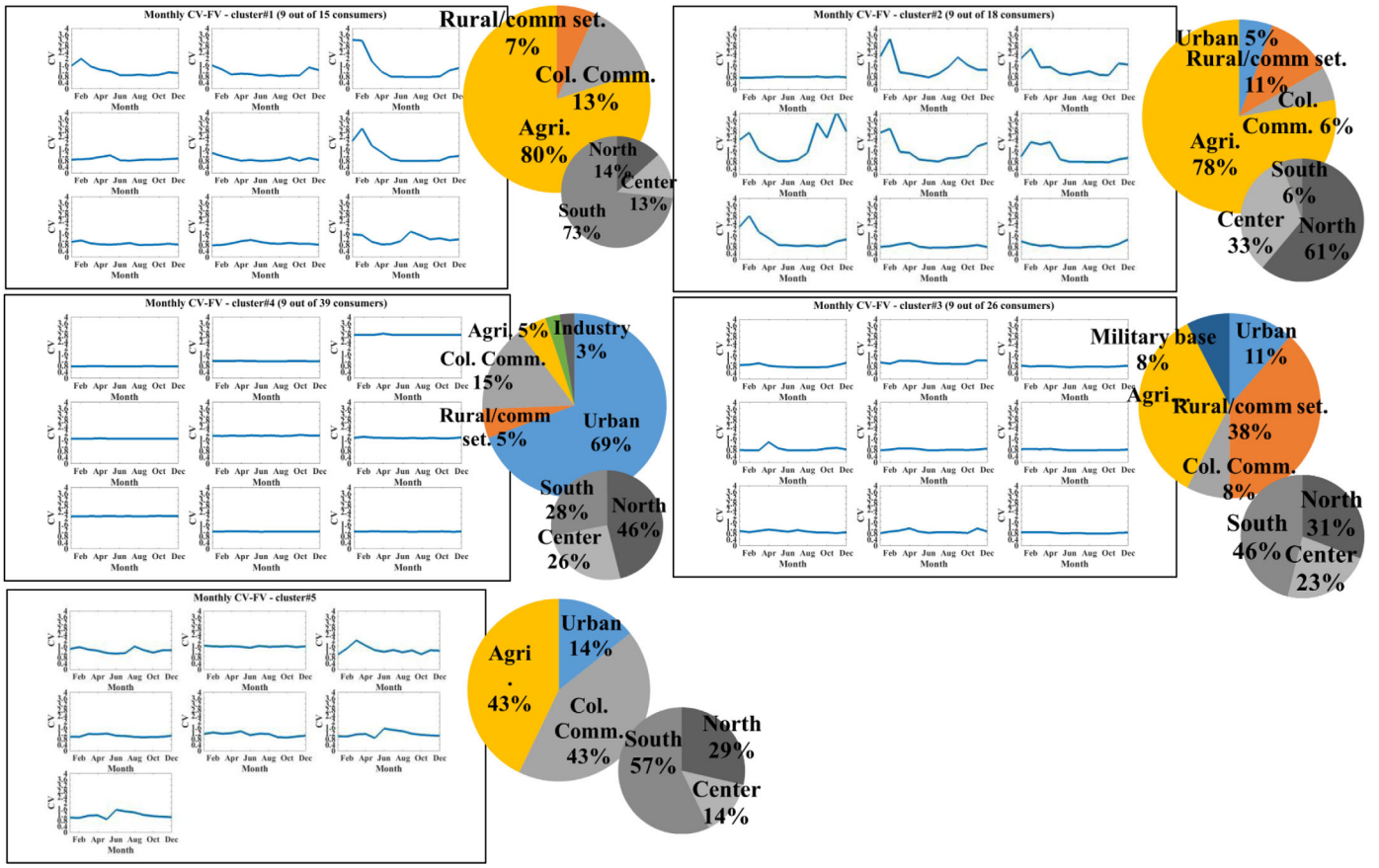

Figure 7. Clustering results for the CV-FVs, K = 5. The subfigures present the FVs assigned to each cluster (#1–#5), with each insert graph presenting a single consumer. For clarity of visualization, only part of the consumers in each cluster are presented. The *y* axis is the level of monthly coefficient of variation. The pie charts present the cluster characteristics—location and consumer types.

The variability of the consumers in cluster #3 ranges from 0.2 to 1.0, with peaks mostly during the spring. The CV amplitude and values are smaller than the previous cluster. The consumers in clusters #3 and 5# have a diverse consumer type composition. Cluster #3 has a majority of rural/communal settlements and agriculture (38% and 35%, respectively). Rural/communal settlements, as collective communities, often have multiple water uses (agricultural and domestic consumption). Therefore, their monthly variability can be similar. The consumers in cluster #5 are mostly farmers and collective communities, with slightly higher CV fluctuations. The patterns in this cluster are less homogenous, thus drawing a single line between its consumers in challenging. Cluster #4 have relatively low CV values ($<$0.4), and is composed mainly of urban consumers (70%), showing that domestic consumers have low variability compared with other types of consumers.

The CV-FV clustering showed a relationship between a consumer water use and its monthly variability, represented by the CV index. Variability may also indicate a homogeneity of geographical location, as seen in clusters #1 and #2 (large Agricultural areas in the pie charts).

3.2.3. Concatenation of the NA-FVs and CV-FVs (NACV-FV)

Combining FVs allow each FV to reveal a certain aspect of the data set [*Yang et al.*, 2013]. The clustering results with a FV created by concantenating the NA-FV and CV-FV (denoted NACV-FV) are presented next. The NA-FVs and the CV-FVs for this data set range between [0,1] and [0,2.4], respectively. To implement clustering based on the CV-FV, the CV was computed on the normalized data, thus all entries in the NACV-FV are between [0,1]. The NACV-FV provides insight into potential correlations between a consumer's monthly averages consumption and its variance. If a consumer has both similar average consumption pattern and similar variance pattern, it is expected to remain in the same cluster. However, if these patterns differ, the concatenation of these two statistics should give a more distinct representation for classification.

Figure 8 presents the two largest clusters (76% of the data set) formed by the NACV-FV: cluster #1 with 49 consumers (top) and cluster #5 with 31 consumers (bottom).

Cluster #1 has a majority of urban consumption (59%), but the other part is composed of consumers with various levels of agricultural consumption. These consumers are equally spread in all three geographical regions, thus this analysis is omitted here. This indicate that consumers of different types may have similar consumption characteristics. On the other hand, cluster #5 has a very low percentage of urban

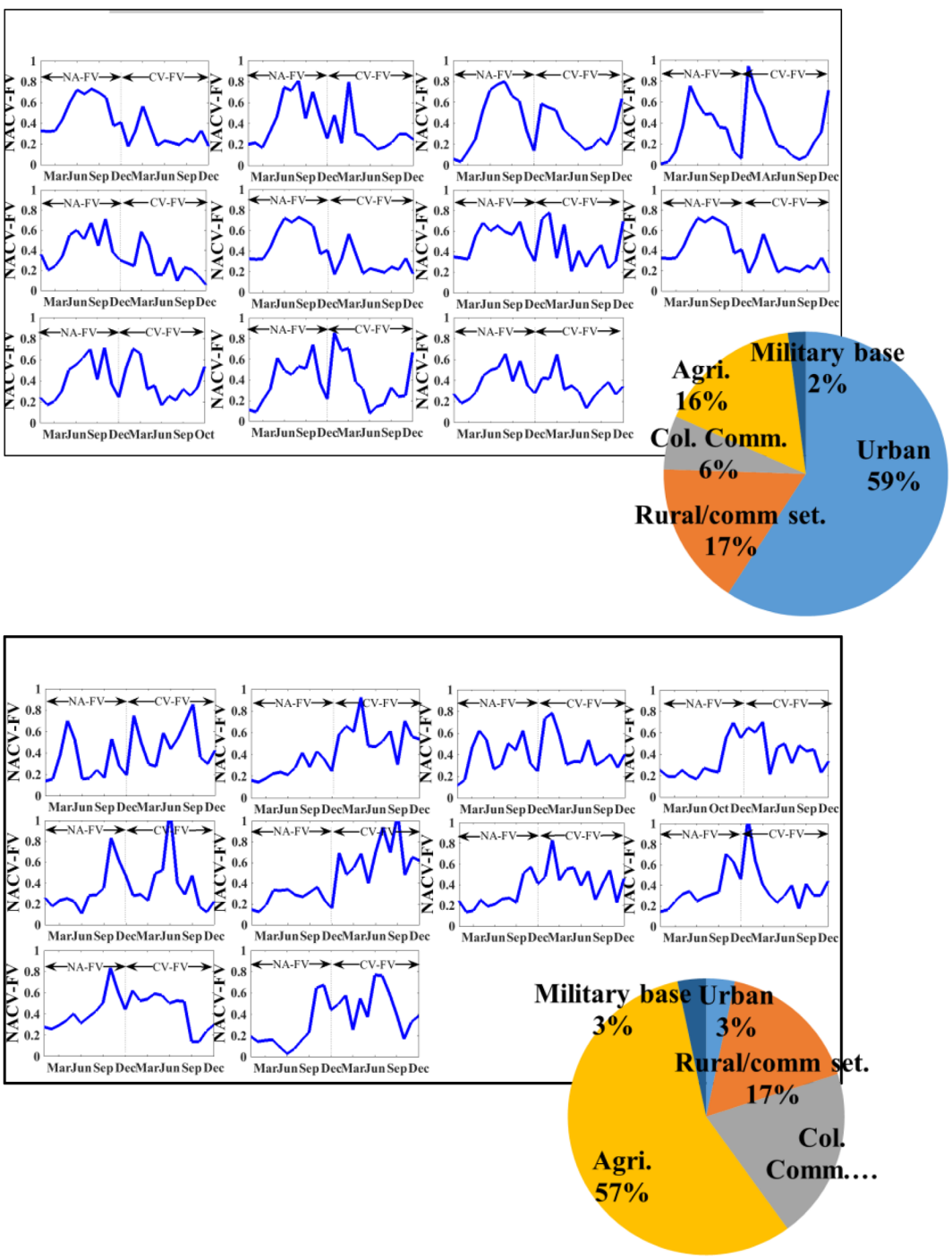

Figure 8. NACV-FV of consumers in cluster #1 (top, 11 out of 49 consumers) and #5 (bottom, 11 out of 31 consumers): The division by arrows above the graphs indicate the normalized monthly average (January–December) followed by the coefficient of variation index values for the same months.

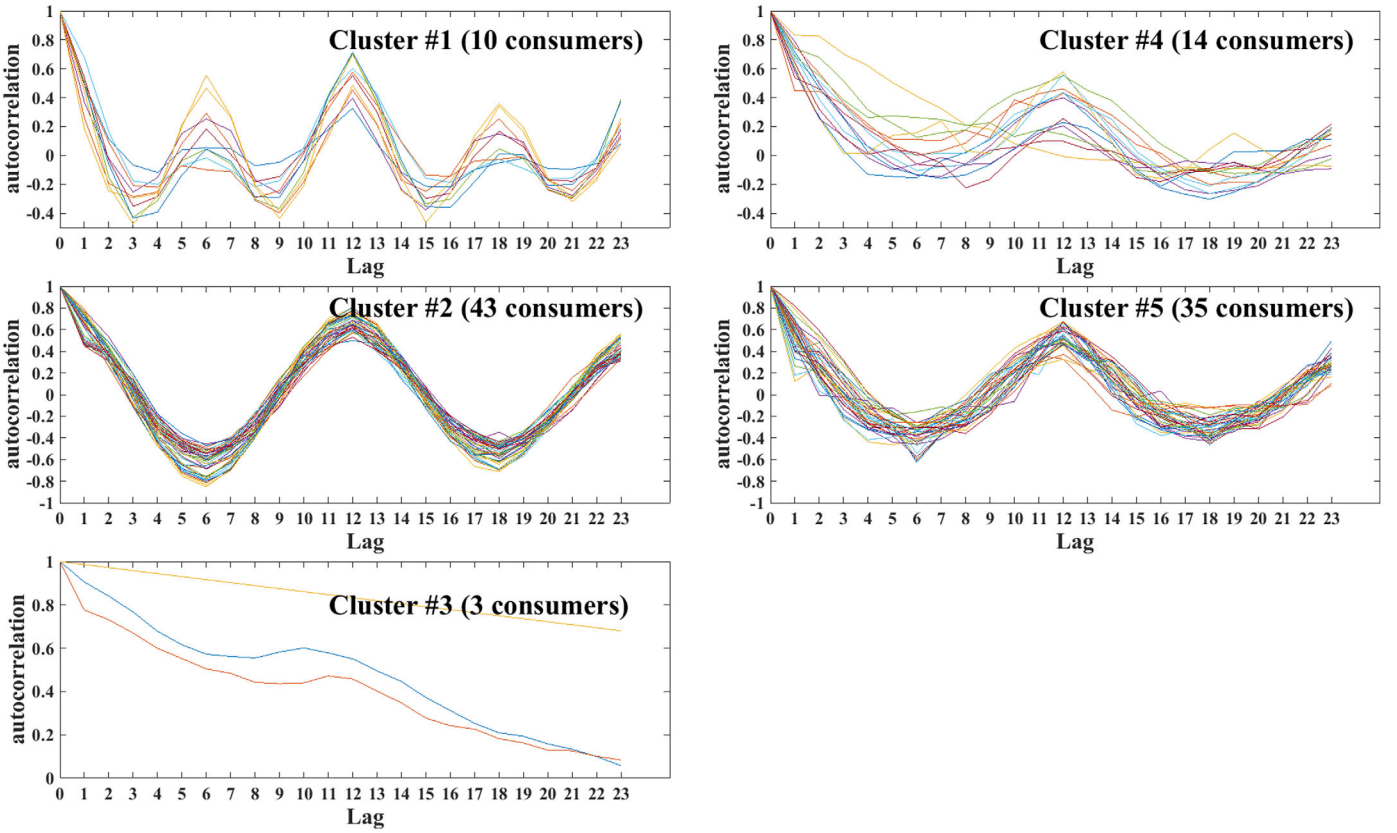

Figure 9. Clustering results for the ACF-FVs, K = 5. The subfigures present the autocorrelation function of the consumers assigned to each cluster (#1–#5).

consumption (3%), and a majority of strictly agricultural consumption (57%) accompanied by collective communities and rural/communal settlements that also have agricultural activities (20% and 17%, respectively).

Clusters 2–4 presented more homogeneity and a relationship between monthly average and monthly CV. For example, the consumer FVs assigned to cluster #4 are mainly agricultural consumers (77%). These consumers exhibit an interesting trend—when the consumption is high, the variability is low and vice versa (Figure 1c). Supporting information Figures S3–S7 contains the detailed NACV-FV results.

3.2.4. Time-Series Autocorrelation FV (ACF-FV)

Clustering based on the ACF-FV transformation was conducted using K = 5, which provided a better average overall Silhouette, compared with $K = 6$ (0.49705 versus 0.41922). Figure 9 presents the clustering results via the autocorrelation function (ACF) of the consumers assigned to each cluster.

Unlike the previous three FVs, the ACF-FV clustering did not create clusters with distinct common cluster characteristics (consumer type and geographical region): Urban water consumers were divided between four of the five consumers, as well as the agriculture consumers, both having mostly a distinct consumption pattern. The figures of the common cluster characteristics are presented in supporting information Figure S8.

3.2.5. Administrative Clustering

The postanalysis presented herein utilizes information on the specific consumer type. This raises the question why not use this classification as a basis for aggregation without bothering with data analysis procedures. The answer is twofold: first, the detailed consumer type used in this research (e.g., a Kibbutz with a specific agriculture type or the type of residential structure) is not readily available as an administrative classification, and was retrieved by directly investigating the specific type of each consumer using on-line resources, etc. Thus, a challenging (though feasible) process is required when a large data set is considered. The second and more serious consideration relates to the homogeneity of the consumers belonging to

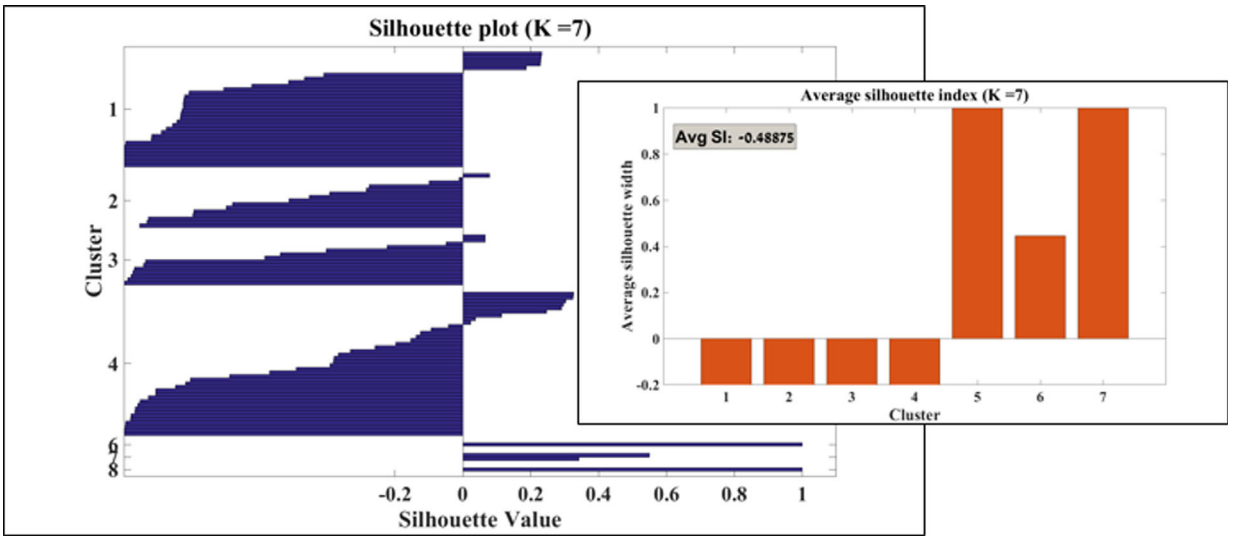

Figure 10. Silhouette plots and average cluster silhouette width for clustering by consumer type. Cluster 6 contains only 2 consumers and clusters 5 and 7 contain only one consumer.

each type: Figure 10 presents the silhouette plot and average silhouette index for the 7 consumer types available in the data set (Urban, rural/communal settlement, Kibbutz, agriculture, industry, police/military base, and "other"). Most of the clusters have a low silhouette index values (left), with a very low overall average silhouette width of −0.48875 (right). The high silhouette index of clusters 5 and 7 results from the fact that they include only one consumer, as well as cluster 6 which has only two consumers. For comparison, the overall average silhouette width of clustering using the NA-FV was 0.4114. Therefore, it can be seen that consumers classified as being nominally of the same type may have different consumption characteristics, which is only revealed via data analysis.

4. Discussion

The goal of clustering by water consumption patterns is to improve the demand information for data-driven regional WDMs. To do so, the research focused on different data representation methods that extract various water consumption features as a basis for regional water demand modeling.

Four Feature Vectors (FVs) were tested during the experimental stage, using a data set of 105 consumers. The normalized average (NA-FV), the Coefficient of variation (CV-FV), their concatenation (NACV-FV), and the water consumption time-series autocorrelation coefficients (ACF-FV). The clustering results using the four FVs provided different cluster compositions, in accordance to the features highlighted by each FV. The analysis of the advantages of using each FV, for the scope of this research, focused on the ability to reveal certain consumer characteristics, based solely on their monthly consumption characteristics (e.g., average, variability).

The results showed the strength of using clustering to separate a consumer data set in a way that reveals the explanatory variables dominating the water consumption of each group (e.g., geographical location, consumer type, main water use), without using this information directly. Using K-means with the NA-FV provided distinct and homogenous clusters with respect to the water consumption. Despite its simplicity, the NA-FV was sensitive enough to show relatively small differences in consumption patterns as belonging to different clusters, and succeeded in separating consumers of different types and geographical regions (thus operating under different climatic conditions). In addition, the NA-FV experiment showed that consumption patterns may serve as a good indicator for the type of agriculture applied by the consumer (e.g., cropland, orchards, or dates). In recent years, land-use analysis is widely applied, and this type of analysis provides a different angle on this subject.

The periodical variability of water demand gives another perspective of consumption patterns. A particular monthly variability pattern can indicate a specific water use and thus point to the demand generating factors that influences this consumer's demand. Using the coefficient of variation measure (CV-FV) resulted in a

different cluster composition from that obtained by using the NA-FV. The monthly variability gave a strong indication of the water use type, mostly differentiating between domestic and agriculture consumers. In addition, it gave a separation that revealed different consumer location (e.g., clusters #1 and #2, Figure 6). Thus, the CV-FV present an opportunity to increase the clustering algorithm's sensitivity to demand generating factors and therefore improve the ability to aggregate the data as a preliminary stage for developing WDMs.

The concatenation of monthly average and variability measures, the NACV-FV, gives another perspective to the data analysis. The NACV-FV provides, in one data structure, a relationship between the monthly average consumption and the monthly variability. The results created a different clustering composition that revealed homogeneity of consumption type, which showed that for some consumers, high variability may be linked to low monthly average consumption. The CV-FV offers increased flexibility in data representation, since several FVs can be combined and even weighted according to their relevance to the clustering objective. Nevertheless, its application should be performed with great care since the concatenation provides a less condensed data representation and features that provide a good separation (e.g., peaks in certain months) might get loss in the longer FV.

The ACF-FV provided distinct FVs for the five groups of consumers. However, in the postanalysis no significant common characteristics (i.e., consumer type and location) where revealed using this FV. Therefore, the autocorrelation function is inadequate as a FV if one wishes to divide the data set as a preliminary step for developing panel-data demand model. Nevertheless, the ACF-FV can form a basis for constructing a time-series model that is based on a group of resembling time series [*Dempster et al*., 1977; *Junninen et al*., 2004].

Aggregation based on administrative classification of consumer type could be inadequate or infeasible for aggregating large data set. First, often even if such a classification exists, it would be too general to truly reflect monthly water consumption behavior, while obtaining a more detailed classification (e.g., specific type of residential structure, or main economic activity) will require a strenuous process; second, as was demonstrated throughout this study, consumers of the same type may have different water consumption patterns (e.g., a Kibbutz or a communal settlement), pattern and variability-wise. To quantify this conclusion, administrative type aggregation was compared with aggregation based on the methodology presented herein (section 3.5). The average silhouette for the administrative classification was negative (−0.48875), meaning that most of the consumers could be better assigned, whereas, clustering based solely based on water consumption type obtained an average silhouette of 0.4114. Thus, clustering by transformation of water consumption data provides finer resolution on water consumption behavior than administrative data, even if it is highly detailed. The selection of K is essential for generating useful results. In this study, K was set based on the silhouette width, together with a visual inspection of the patterns or the FVs assigned to each cluster. The homogeneity of the clusters created based on the silhouette width was significantly improved when K was increased. Thus, this measure was found to be insufficient by itself to determine the number of clusters. However, it provided a guideline to the neighborhood of the suitable K.

The clustering procedure operates on a dissimilarity matrix, which is dependent on the type of FVs and distance function used. The FVs and distance matrix themselves may not always reflect well the real differences in pattern (or any other measure), and should thus be used as a guideline and not the optimal number of cluster. Another possible explanation to the failure of the silhouette plot is that this index is known to work best in a situation with roughly spherical clusters, which might not be the case here [*Rousseeuw*, 1987].

5. Conclusions

Water demand models (WDMs) are typically developed based on historical records of demand generating factors such as climatic conditions, water prices, land-use data, and socio-economic variables. The heterogeneity of consumers in a large region, such as a regional or a national water grid, poses a challenge for identifying the demand generating factors pertaining to the entire data set. Developing a WDM that considers the data set as a whole, might mask different consumer responses to driving forces. However, even dividing the data set based on some prior classification (industry, agriculture, etc.), may still mask heterogeneity in each subgroup.

To overcome this challenge, this paper presents an approach for regional WDM development which is based on (*i*) aggregating the data set based on various Feature Vectors as metrics to represent consumption characteristics (e.g., average consumption and variability), and (*ii*) developing a WDM for each relatively

homogeneous group. The proposed methodology is a necessary step in developing WDMs that are capable of representing and predicting water consumption of large and heterogeneous regions. The research results demonstrate the advantages of using different water consumption data transformations for clustering a water consumption data set into subsets that exhibit similar consumption patterns. The experiment showed that clustering, based *solely* on water consumption records, succeeded in grouping consumers with similar characteristics. Moreover, the approach was shown to be advantageous compared with clustering based on administrative classification.

The approach depicted here is a step toward utilizing the increasing amount of available water consumption data and data analysis techniques to facilitate the modeling of water demands in larger, heterogeneous regions with sufficient resolution. Still, in spite of the increased amount of data, available through on-going improvements in water metering and data storage, obtaining this data by the research community is often challenging and requires significant investment of time and effort.

The data set used for the research was obtained from the Mekorot, the Israeli Water Company. The main water demand characteristics that were explored herein were the average monthly water, reflecting water consumption pattern, the monthly coefficient of variation, reflecting monthly variability, and the autocorrelation coefficients of the water consumption time series. Despite of the local application, these Feature Vectors may be applicable to data sets in other parts of the world: Farmers, which are often highly dependent on climate conditions, can be identified by their monthly variability, as in the case of the Israeli data set; Domestic consumption, on the other hand, may be characterized by a relatively low variability and steady pattern (e.g., cluster #6, Figure 5) throughout the world. Nevertheless, the final selection of FVs, distance function, and number of clusters, should be done by experimenting with the data set and incorporating previous knowledge on the consumers served by the WDS (e.g., type of agricultural activity in the region and main water uses). In addition, this study focused on four FVs, while other FVs may be considered. For example, the monthly variance is a possible candidate, if one wishes to enhance the effect of monthly variability, as well as higher-order statistical moment.

The next steps of this research are the upscaling of the work to a regional level, and constructing a WDM for each of the groups formed by the clustering algorithm. The WDMs to be developed will consider the relevant panel data pertaining to each group, and autocorrelation factors pertaining to the time series characteristics.

Acknowledgments

The authors would like to thank the Glasberg-Klein, and the NY Metropolitan Research Foundation for their partial financial support of this research. The authors would also like to thank the Israeli Water Authority and Mekorot for providing the data used in this research. Data to support this article are from Mekorot, the Israeli National Water Company. Because of privacy and security issues, the data cannot be released. The software that facilitated this research was coded in *Matlab*® and is available upon request. To allow the execution of the code a synthetic data set with similar characteristics to the real-life one is also available with the software package.

References

Babel, M. S., A. Das Gupta, and P. Pradhan (2007), A multivariate econometric approach for domestic water demand modeling: An application to Kathmandu, Nepal, *Water Resour. Manage.*, *21*(3), 573–589, doi:10.1007/s11269-006-9030-6.

Benaichouche, A. N., H. Oulhadj, and P. Siarry (2013), Improved spatial fuzzy c-means clustering for image segmentation using PSO initialization, Mahalanobis distance and post-segmentation correction, *Digit. Signal Process.*, *23*(5), 1390–1400.

Berger, T. (2001), Agent-based spatial models applied to agriculture: A simulation tool for technology diffusion, resource use changes and policy analysis, *Agric. Econ.*, *25*(2-3), 245–260, doi:10.1111/j.1574-0862.2001.tb00205.x.

Bullinger, L., K. Döhner, E. Bair, S. Fröhling, R. F. Schlenk, R. Tibshirani, H. Döhner, and J. R. Pollack (2004), Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia, *N. Engl. J. Med.*, *350*(16), 1605–1616, doi:10.1056/NEJMoa031046.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc., Ser. B*, *39*(1), 1–38.

Fisher, F. M., S. Arlosoroff, Z. Eckstein, M. Haddadin, S. G. Hamati, A. Huber-Lee, A. Jarrar, A. Jayyousi, U. Shamir, and H. Wesseling (2002), Optimal water management and conflict resolution: The Middle East Water Project, *Water Resour. Res.*, *38*(11), 1243, doi:10.1029/2001WR000943.

Franczyk, J., and H. Chang (2009), Spatial analysis of water use in Oregon, USA, 1985–2005, *Water Resour. Manage.*, *23*(4), 755–774, doi:10.1007/s11269-008-9298-9.

Gleick, P. H., D. Haasz, C. Henges-Jeck, and S. Srinivasan (2003), *Waste Not, Want Not: The Potential for Urban Water Conservation in California*, Pac. Inst. for Stud. in Dev., Environ., and Security, Oakland, Calif.

Gutwein, B., and R. Lang (1993), Regional irrigation water demand, *J. Irrig. Drain. Eng.*, *119*(5), 829–847, doi:10.1061/(ASCE)0733-9437(1993)119:5(829).

House-Peters, L. A., and H. Chang (2011), Urban water demand modeling: Review of concepts, methods, and organizing principles, *Water Resour. Res.*, *47*, W05401, doi:10.1029/2010WR009624.

Howitt, R. E. (1995), A calibration method for agricultural economic production models, *J. Agric. Econ.*, *46*(2), 147–159, doi:10.1111/j.1477-9552.1995.tb00762.x.

Jain, A., A. K. Varshney, and U. C. Joshi (2001), Short-term water demand forecast modelling at IIT Kanpur using artificial neural networks, *Water Resour. Manage.*, *15*(5), 299–321, doi:10.1023/A:1014415503476.

Jain, A. K. (2010), Data clustering: 50 years beyond K-means, *Pattern recognition letters*, *31*(8), 651–666.

Junninen, H., H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen (2004), Methods for imputation of missing values in air quality data sets, *Atmos. Environ.*, *38*(18), 2895–2907, doi:10.1016/j.atmosenv.2004.02.026.

Lee, S. J., and E. A. Wentz (2008), Applying Bayesian Maximum Entropy to extrapolating local-scale water consumption in Maricopa County, Arizona, *Water Resour. Res.*, *44*, W01401, doi:10.1029/2007WR006101.

Lee, S. J., E. A. Wentz, and P. Gober (2010), Space-time forecasting using soft geostatistics: A case study in forecasting municipal water demand for Phoenix, Arizona, *Stochastic Environ. Res. Risk Assess.*, *24*(2), 283–295, doi:10.1007/s00477-009-0317-z.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. Le Cam and J. Neyman, 666 pp., Univ. of Calif. Press, Oakland.

Maidment, D. R., and E. Parzen (1984), Cascade model of monthly municipal water use, *Water Resour. Res.*, *20*(1), 15–23, doi:10.1029/WR020i001p00015.

Medellín-Azuara, J., R. E. Howitt, and J. J. Harou (2012), Predicting farmer responses to water pricing, rationing and subsidies assuming profit maximizing investment in irrigation technology, *Agric. Water Manage.*, *108*, 73–82, doi:10.1016/j.agwat.2011.12.017.

Nelson, C. R. (1973), *Applied Time Series Analysis for Managerial Forecasting*, 4th ed., Holden-Day, San Francisco, Calif.

Olmstead, S. M., W. Michael Hanemann, and R. N. Stavins (2007), Water demand under alternative price structures, *J. Environ. Econ. Manage.*, *54*(2), 181–198, doi:10.1016/j.jeem.2007.03.002.

Rousseeuw, P. J. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, *20*, 53–65, doi:10.1016/0377-0427(87)90125-7.

Schleich, J., and T. Hillenbrand (2009), Determinants of residential water demand in Germany, *Ecol. Econ.*, *68*(6), 1756–1769, doi:10.1016/j.ecolecon.2008.11.012.

Tan, P., M. Steinbach, and V. Kumar (2005), Introduction to Data Mining-Book, *Cluster Analysis: Basic Concepts and Algorithms*, First Edition, Pearson-Addison Wesley Higher Education publishers, 532–568.

Tiwari, M. K., and J. Adamowski (2013), Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models, *Water Resour. Res.*, *49*, 6486–6507, doi:10.1002/wrcr.20517.

Weinberger, K. Q., and L. K. Saul (2009), Distance metric learning for large margin nearest neighbor classification, *J. Mach. Learn. Res.*, *10*, 207–244.

Worthington, A. C., H. Higgs, and M. Hoffman (2009), Residential water demand modelling in Queensland, Australia: A comparative panel data approach, *Water Policy*, *11*, 427–441, doi:10.2166/wp.2009.063.

Yang, Y. T., B. Fishbain, D. S. Hochbaum, E. B. Norman, and E. Swanberg (2013), The supervised normalized cut method for detecting, classifying, and identifying special nuclear materials, *INFORMS J. Comput.*, *26*(1), 45–58, doi:10.1287/ijoc.1120.0546.

Yiakopoulos, C. T., K. C. Gryllias, and I. A. Antoniadis (2011), Rolling element bearing fault detection in industrial environments based on a K-means clustering approach, *Expert Syst. Appl.*, *38*(3), 2888–2911, doi:10.1016/j.eswa.2010.08.083.