Building an Integrated Water–Land Use Database for Defining Benchmarks, Conservation Targets, and User Clusters

Rebecca Dziedzic[1]; Katelyn Margerm[2]; Jeff Evenson[3]; and Bryan W. Karney, M.ASCE[4]

Abstract: Water utilities have large amounts of data at their disposal, which are seldom being used to their full potential. Integrating water billing records with land-use and demographic data organizes information and makes inherent correlations easier to understand, facilitating communication to stakeholders. This data was integrated for three Ontario (Canada) municipalities, Barrie, Guelph, and London. A summary tool was created, with proposed metrics and charts, that facilitates comparisons between cities, definition of benchmarks, and identification of targets for conservation. More than 60% of consumption in these cities is residential, and mostly lies below the Ontario average of 267 L/cap · day. Water user clusters were created through self-organizing maps, K-means, and hierarchical clustering, and selected according to their pseudo-F and Rand statistics. Users within the same or similar property codes were found to cluster together. The application of data-mining methods provides actionable information for utilities seeking to reduce demands and increase system sustainability. **DOI: 10.1061/(ASCE)WR.1943-5452.0000462.**

Author keywords: Water management; Conservation; Data analysis; Databases; Water demand; Land use.

Value of Integrating Data

"Divide and conquer," specialization, is a common motto in solving complex problems, which splits complex realities into a variety of disciplines, sectors, departments, etc. This means, however, that interactions and synergies between the segments are either ignored or downplayed. According to Hussey and Pittock (2012), three major barriers prevent greater integration between sectors and policy domains: data deficiencies (missing or disorganized), weak existing policies and frameworks (fragmented, inconsistent, lacking review), and cultural inertia/path-dependency (silo mentality). A consistent system for collecting data in an easily retrievable, standardized, and comprehensive fashion is instrumental in managing water demand (Cahill and Lund 2013). The present study focuses on data as a pathway to resolve the second and third types of barriers. It integrates water, land-use, and demographic data with the objective of facilitating both understanding and demand management.

The United Nations Environmental Programme (2012) reviewed worldwide applications of integrated approaches to water resource management and recognized the need for better information management, stating, "Information is the foundation of good decision-making and planning," with reference to integrated water resources management in Agenda 21 (United Nations Conference on Environment and Development 1992). Although progress has been slow (Muste 2013), this type of effort has been facilitated in recent years due to advances in technology and information exchange, fostering a greater commitment to initiatives in data collection (Maidment 2008). Bringing together data from different disciplines fills in gaps of information and knowledge (Muste 2013).

Boyle et al. (2011) stress the fact that utilities already have valuable data at hand. Utility billing data can be used to inform many types of management decisions, such as pricing, conservation marketing, and peak planning. The use of this data is supported by three characteristics: it is available to all utilities; it can be used to target specific customer groups with customized messages that are more cost-effective than broad public outreach programs; and it can enable an understanding of specific customers, leading to localized utility policies and strategies.

Using more detailed data can provide utilities with greater insights on to how water is consumed over space and time (Polebitski and Palmer 2010). Jorgensen et al. (2009) indicate that demographics, dwelling characteristics, and household composition all directly impact water consumption, conservation intention, trust, perceived behavioral control, perception, and habits. Polebitski and Palmer (2010), as well as Morales et al. (2011) joined utility billing data with census demographic and property appraisal data to forecast residential and non-residential water use, respectively.

Shandas and Parandvash (2009) suggest that, given current population growth and urban development, approaching water use through the lens of urban planning, namely the structural and demographic drivers of consumption, can improve the effectiveness of water conservation. Brooks (2006) defines water demand management operationally according to five motivators: (1) reducing the quantity or quality of water required for a specific use; (2) adjusting the nature of the task so it can be accomplished with less or lower-quality water; (3) reducing loss in quantity or quality of water in the distribution system; (4) shifting time of

[1]Ph.D. Candidate, Dept. of Civil Engineering, Univ. of Toronto, 35 St. George St., Toronto, ON, Canada M5S 1A4 (corresponding author). E-mail: re.dziedzic@mail.utoronto.ca

[2]Senior Engineering Researcher, Canadian Urban Institute, 555 Richmond St. W., Suite 402, Toronto, ON, Canada M5V 3B1.

[3]Vice President of Urban Solutions, Canadian Urban Institute, 555 Richmond St. W., Suite 402, Toronto, ON, Canada M5V 3B1.

[4]Professor, Dept. of Civil Engineering, Univ. of Toronto, 35 St. George St., Toronto, ON, Canada M5S 1A4.

Note. This manuscript was submitted on January 22, 2014; approved on May 15, 2014; published online on July 21, 2014. Discussion period open until December 21, 2014; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496/04014065(9)/$25.00.

use to off-peak periods; and (5) increasing the system's ability to operate during droughts. Other benefits include deferring and reducing capital works, reducing pumping costs, and increased flexibility of demand-side solutions in adjusting to changing circumstances (Sahely and Kennedy 2007). Although water demand management can, thus, be multifaceted, the focus here is on conservation.

The Ontario Water Works Association (2006) proposes a method for decreasing demand, which begins with evaluating the system and setting goals. Guidelines include the installation of certain efficiency devices towards the reduction of consumption by a preset percentage. Recommendations for selecting the strategy and target, however, are not given and may be arbitrarily determined by the utility. Morton (2011) suggests establishing a best practice range instead. According to the Morton (2011), there are two objectives in developing a benchmark: (1) determining the appropriate metric to be used, e.g., $m^3/m^2 \cdot$ year or $L/cap \cdot day$, and (2) determining the benchmark value based on the water consumption intensity across the dataset. The best practice range can be defined as a percentile, such as the first quartile of the dataset. Data mining is, thus, instrumental in establishing benchmarks for different sectors and monitoring improvement.

Dziegielewski and Kiefer (2010) suggest normalizing metrics for comparability. If metrics are being compared for a single utility over time, it should be sufficient to adjust the calculated metrics for weather conditions (temperature and rainfall). Annual changes in the number of users are accounted for by the scaling variable, such as population or building space. When metrics are compared across different utilities, it is recommended that all external factors that influence water consumption (outside the control of water users) should be considered. However, it is acknowledged that normalization of weather and other confounding factors across different utilities can be problematic. A practical approach is to use metered account-level information for homogeneous groups of customers and the same dimensions of water use (e.g., total annual, seasonal, nonseasonal).

Even though targets may not currently be set across utilities, provincial or national averages provide an important basis for comparison in a larger, yet similar, stage. Environment Canada (2010), from a survey of 530 Canadian municipalities, presents water-use rates by province. Average residential flow in Ontario is approximately 267 $L/cap \cdot day$, almost 20% lower than the national value of 327 $L/cap \cdot day$. Maas (2009) offers a blueprint for a comprehensive water conservation strategy, in which a target of 150 $L/cap \cdot day$[1] for Ontario municipalities is suggested. Albeit not at a policy level, this has been accepted by many Ontario utilities as a suitable goal.

Benchmarks for industrial, commercial, and institutional (ICI) water use are harder to find, since there is more variation, especially between different types of industries. Therefore, water-use data must be sorted according to property codes or industrial classification codes for better comparisons. Gleick et al. (2003) present water-use benchmarks for different industries in California by sector, specifically by Standard Industrial Classification codes. Water consumption is normalized by production or number of employees. South East Water (2006) provides an extensive list of benchmarks per sector based on surveys and literature review. The dates of the references, however, vary between 1997 and 2006, suggesting some of the data may no longer be current or relevant. The rate of obsolescence will depend on the sector, acceptance of these targets, and potential for improvement.

Previous research, as noted above, has stressed the need for conservation planning and underlined the importance of integrated approaches that combine information from different areas to fill in gaps of knowledge. Using practical models, with input variables that can be collected, monitored, and used by the utility, is key (Donkor et al. 2014). The present study integrates data that is readily available to most Canadian municipalities and their water utilities: water billing records, demographic census information, and structural data from property tax assessments. This information is used to build an integrated database to support decision-making, specifically demand management. Accordingly, benchmarks, conservation targets, and user clusters are defined.

Methodology

The data-mining method, on which this study is based, comprises six steps: (1) problem definition, (2) data preparation, (3) data exploration, (4) modeling, (5) evaluation, and (6) deployment. This process was repeated for three Ontario municipalities, and results were compared. The problem definition frames the succeeding steps by describing study objectives. Data preparation encompasses the process of cleaning and formatting the data, as well as building the integrated database. In data exploration, distributions, trends, and metrics are assessed and compared. The first benchmarks as well as targets for conservation are also established in this phase. Modeling furthers the analysis as user clusters are defined. These results are then evaluated and the model built for data clustering can be applied to different data sets.

Problem Definition

Many water utilities, such as the ones studied herein, bill their customers according to general sectors, i.e., residential, industrial, commercial, and institutional, or even at a flat rate for all users. Although Canadian municipalities have access to land use and demographic data from the Municipal Property Assessment Corporation (MPAC) and Statistics Canada (Statcan), respectively, most utilities only distinguish users by billing class. This classification is not descriptive, and does not allow for the definition of homogeneous groups of users, which are more suitable for analyzing trends, establishing benchmarks, and targeting conservation. Nonetheless, the more than 100 classes used by MPAC may prove to be excessive for utility needs, especially if different strategies are being used for each different type of consumer. Accordingly, the study seeks to define benchmarks and targets for water conservation, as well as water-user segments in three Canadian municipalities, Barrie, Guelph, and London, based on integrated water consumption, land use, and demographics data.

Data Preparation

The integrated databases were created by connecting data from an SQL server, through ODBC (Open Database Connectivity), for easy storage and updating. Database construction was completed as part of a research project funded by the Ontario Ministry of Environment with the Canadian Urban Institute (CUI). Cleaning and formatting of the billing records was subcontracted. This involved joining data from different months, and updating or rectifying inconsistent customer identifier or address formats. Spatial data, relating addresses to roll numbers, parcels, and dissemination blocks was joined by the CUI using their geographic information system (GIS) data. Roll numbers are the property identifiers used by MPAC for tax assessments. Parcels are pieces of land, generally equivalent to properties, and dissemination

blocks, equivalent to city blocks, are the smallest geographic areas for which Statcan releases population counts.

The integration of water consumption, land use, and demographic data involved connecting information from four different sources: water utilities, MPAC, Statcan, and CUI.

Six tables were generally integrated in the database, with their main variables listed below, either as categorical (c), or numerical (n):

- Customer information (utilities): customer ID (c), address (c), rate class (c);
- Billing data (utilities): customer ID (c), monthly water consumption (n);
- Address table (CUI), created using a series of spatial joins in GIS: address (c), roll number (c), parcel ID (c), dissemination block (DB) ID (c);
- Structural data (MPAC): roll number (c), year built (n), building footprint (n), property code (c);
- Parcel data (utilities): parcel ID (c), parcel area (n), building area (n); and
- Demographic data (Statcan): dissemination block ID (c), population count (n).

Customer information and billing data are collected for each user, with customer ID as the unique identifier. Billing data between 2006 and 2011 was collected. These datasets were linked by customer ID, and then summarized by address so that they could be combined with the address table. After these were joined, water consumption was associated with roll numbers, parcels, and DBs, and can, therefore, be integrated with land use and demographic data. After summarizing by roll number, this data was combined with the structural data. This was then summarized, either by parcel or by DB, to be joined to the parcel or demographic data, respectively.

There are, thus, two main tables as outputs of this integration process: a parcel table, with water and land use at the parcel level, and a DB table, with water, land use, and population count by DB. This can then be mapped, facilitating the visualization of spatial trends. Guelph, the smallest of the three municipalities had approximately 120,000 inhabitants in 2011, distributed in 35,000 parcels. Barrie had 150,000 inhabitants and 42,000 parcels (generally equivalent to properties), while London, more populated, had 360,000 inhabitants in 99,000 parcels.

Data integration can be a cumbersome process if attention is not paid to differences in formatting, missing values, data inconsistencies, nonunique identifiers, and varying levels of data summarization. That being the case, it is helpful to sketch the integration process beforehand and name queries and tables in a simple, yet descriptive, manner. In addition, after each query, as data is joined and summarized at different levels, results should be checked, i.e., water consumption, areas, and population. It is best not to compare totals to the previous iteration but to the initial tables, so that instead of a relative match percentage, which does not convey the total amount of data neglected, a net match percentage can be determined.

Because the tables contain data at different geographic levels, which can fully contain or overlap each other, as shown in Fig. 1, the order in which they are joined and summarized is important. Customer ID, address, and roll number are all point data. However, they are not absolutely equivalent. For instance, a condominium building may be considered as one customer to the water utility, but may have multiple addresses, corresponding to different units. A parcel, roughly a property, is a collection of roll numbers, and a DB, similar to a block, is a cluster of addresses. Even though a parcel is smaller than a DB, it may not be fully contained in one. Accordingly, joins have to be made between tables that are

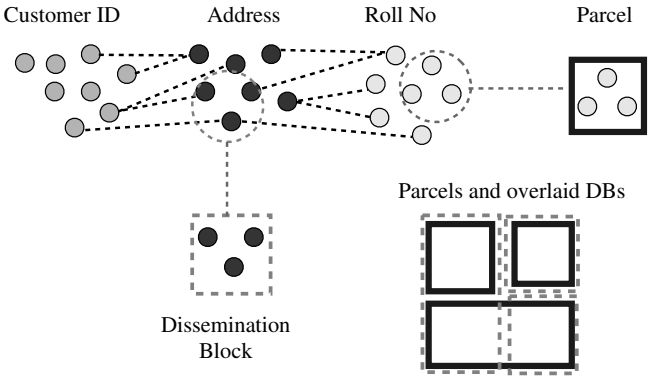

Fig. 1. Geographic levels of different data types and their spatial relations

summarized at the same level, in order to avoid duplicating data. Furthermore, special attention should be paid to the identifiers used for the joins, since they should have the same format. Addresses, for example, are generally recorded differently by utilities and MPAC. Inconsistencies in formats also occur between different years of billing data. The final match rate between water use and property codes was 90% or higher for all municipalities.

Missing roll numbers were substituted with unique dummy identifiers so that land use associated with blank identifiers, when summarized, would not be grouped together. As a result, the data can remain linked to its address despite the lack of a roll number. Because population counts are only released every four years, and water consumption data is reported monthly, demographic data was interpolated for the years with no census data, and only yearly per capita metrics were calculated. Population was assumed to increase geometrically. Land use data is also not released yearly. In this case, if the year built is more recent than the year of consumption, building footprint and unit (address) count were assumed to be zero. Whenever consumption was null, records were also neglected. Outliers (over three standard deviations away from the mean) were also removed before the modeling phase and represent 2% or less of entries. If unit count, building space, or property area were missing, values were modeled based on similar users with the nearest neighbor algorithm.

Water consumption data is collected every billing cycle, which is either monthly or bi-monthly for the given municipalities. However, different parts of the cities are metered at different times of the month. Therefore, attention must be paid to the difference between reading date and billing date, especially when seasonal use is being calculated. Although two regions may have been billed at the same time, that does not mean that the billed water consumption corresponds to the same period for both. The analyzed water data was separated by billing date, although seasonal use was calculated differently based on the billing cycle of the user.

Data Exploration

Because of this study's emphasis on integrating data, the amount of data and specific interest of the cities in visualizing correlations and targeting conservation, the data exploration focused on bivariate analyses. The main variables that were analyzed in this phase were monthly water use (m^3), unit count, building footprint (m^2), year built, property area (ha), population (cap), rate class, and property code. A spreadsheet-based summary tool comparing metrics from

all three cities was developed for the communication of results to water utilities, and for use in both result assessment and conveying this information to policy makers and users. When sharing these results outside the organization, however, care should be taken to avoid disclosing individual information. This implies classifications may have to be grouped, and if data is being mapped, individual properties should not be able to be singled out.

The summary tool graphically represents relations between user characteristics and demand, temporal (monthly, seasonally, and annually) trends of water use, as well as variations within and between water use metrics in property classes and sectors. It, thus, addresses questions regarding the significance of user characteristics to demand management, expectations of future use given changes to these characteristics, and key user types to target for conservation. The tool is most helpful to utility managers and planners, who can easily view and extract summarized information to plan conservation strategies and communicate with stakeholders.

The summary tool comprises three main components: (1) distributions, (2) trends, and (3) metrics. The first two present charts of total water use, whereas the latter has various metrics of normalized water consumption. The distribution of water consumption is presented in pie charts by sector and property code, as well as bar charts of the top water-consuming property codes and the percentage of total water use they represent together with the class coefficient of variation of normalized water consumption. Some examples of figures provided in the tool and the types of conclusions obtained are shown in the "Results" section.

Modeling

Given the objective of creating water user segments, the modeling phase applied different clustering techniques in grouping the data. Segments were created within each of the four larger main sectors: residential, industrial, commercial, and institutional. Furthermore, for this process to be easily replicable and results compared with other utilities, property codes were clustered through three methods: hierarchical clustering, K-means clustering, and self-organizing maps. According to Vesanto and Alhoniemi (2000), using prototype clusters reduces computational effort and noise since the prototypes are local averages of the data. The same methods were applied without using the property codes as cluster prototypes, and the results compared. Data-mining software with a user interface was used to facilitate the visualization and understanding of the process, as well as its modification.

Through a principal component analysis, in which water consumption was set as the target and population count, unit count, building space, and property area as attributes (components), it was found that all components equally explain the variations in water use. Population count was only included in the analysis of the residential sector, since it is an indicator of residential occupancy, not ICI. The attribute with the highest importance varies between sectors and cities. Correlations between the three parameters for ICI vary, generally being above 0.5, and reaching 0.95 in some instances. Because these correlations are not consistent, and occasionally low, no attributes were considered sufficiently similar to be excluded from the modeling phase for ICI. Population count, in the residential analyses, presented higher correlations with the other attributes—at least 0.7, and up to 0.99. Therefore, only unit count, building space, and property area were kept in the residential clustering as well.

The first clustering technique applied, hierarchical agglomerative clustering, minimizes the linkage, the dissimilarity within clusters. The algorithm is initiated with the data points as individual clusters, which are progressively merged at each step. This requires defining a type of dissimilarity. Ward's linkage, which minimizes total within-cluster variance based on weighted square distance between cluster centers, was used. K-means, the second method, assigns a specified number of centroids to the data set and minimizes the total intracluster distance (Tan et al. 2006). The number of clusters was optimized according to the silhouette index, which relates the average distance between a point and all other points within its cluster, as well as between the point and all points in other clusters. This approach generally outperforms other internal indices, and its performance is close to that of the best relative indices (Brun et al. 2007). The Euclidean distance was applied as the objective function, as recommended by Xu and Wunsch (2009).

In the third method, self-organizing maps, each node of data is connected to a vector by a specific weight, which defines the cluster. Adjusting the weights minimizes the distance between clusters, and the learning rate decays at each iteration (Vesanto and Alhoniemi 2000). The Gaussian function was applied for calculating the neighborhood kernel, as recommended by Lee and Verleysen (2002) to produce better results than the bubble function. The size of the map and the radius were chosen for a small number of clusters similar to what was found with the application of the two other methods.

Hillenmeyer (2005) points out issues for each of the methods. The hierarchical agglomerative algorithm is sensitive to outliers and might not produce optimal results, if a local, instead of a global, minima is found. Additionally, the generated trees might be too complex and hard to interpret. K-means clustering is a faster method. Nonetheless, results are sensitive to the choice of initial seeds. Self-organizing maps can be easier to visualize. However, the solution is sensitive to the starting structure, and there is no guarantee of convergence to representative clusters. These methods were not only applied to clustering the property codes and their average metrics, but to the parcel data (i.e., water use at property level) as well. In this case, due to the number of data entries, only a sample was used to create the clusters.

Evaluation

The results were compared among the three cities, for cross-validation, between the different clustering methodologies, and with or without property codes as the initial clusters. Clusters were validated internally with the pseudo-F statistic, proportional to the ratio between the sum of squares between clusters, and the sum of squares within clusters. For external validation, the Rand statistic was used in comparing the clusters created with parcel level data and property codes. The Rand statistic measures the proportion of pairs of vectors that agree by belonging either to the same cluster and property code or to different clusters and property codes (Brun 2007). It can vary between 0 and 1, with 1 being the highest score.

Deployment

In order to update the clusters with more recent data or model consumption in other municipalities, the data would need to be prepared, links in the database updated or modified, and queries rerun. With this, a new input file for the modeling stage would be created. Since the models will be run in data-mining software, deployment will use the same schemas created for modeling. For the evaluation of the clusters through the pseudo-F and Rand statistics, a program was written.

Results

Residential water use represents more than 60% of total consumption in all three municipalities, combined. The averages of the three cities are presented in Fig. 2. Overall, approximately 5% of the total water consumption was not assigned a property code in the database, due to formatting issues or outdated property information. Based on a search of the addresses of a sample of users with unassigned property codes, these were found to be from the ICI sectors. Therefore, ICI water use for the three cities represents more than 32% of total water use, the sum of industrial, commercial, and institutional use shown in Fig. 2(a).

Within the residential sector, single-family dwellings are responsible for consuming around 60% of water. This is the largest water-consuming property code across all municipalities. Most residential users in these cities consume more water than the provincial target of 150 L/cap · day, yet less than the Ontario and Canada averages of 267 and 327 L/cap · day. Within the commercial sector, shopping centers appear as a top water-consuming property code at around 20% of the total commercial use. Other common large commercial user types are hotels and large office buildings. Within the industrial sector, because the MPAC classification groups a variety of different industries under one property code, standard industrial properties, only general information is available. The more informative denominations show the high water consumption of specific uses, such as distilleries or breweries and water treatment stations. In the institutional sector, health and educational facilities use the most water. Due to their high combined water use, these categories of water users are prime targets for water conservation programs.

A decrease in residential consumption from 2006 to 2011 was observed for all three cities, Figs. 3(a–c). These trends are shown on different scales for each municipality in order to facilitate the visualization of the seasonal components. Values for the city of Barrie were extrapolated from 2009 to 2010 data because of issues in formatting the billing records. Residential water use during the winter remained fairly constant for all cities. The variation in yearly consumption is thus greatly explained by higher summer (peak) water use. Note that summer and winter each correspond to three months of consumption and their sum does not equal yearly water use. As expected, water consumption per capita has decreased due to the simultaneous decrease in gross use with an increase in population. Furthermore, the rates of decrease in per capita consumption are similar for all three cities. Residents have been reducing their water usage to such a degree that it counteracts the increase in population. Different trends, however, were observed for low, medium, and high density dwellings. There is also a reduction in water use for more recent building vintages, Fig. 4.

Residential consumption per unit, Fig. 3(f), has also decreased, although less intensely than per capita, Fig. 3(d), suggesting there

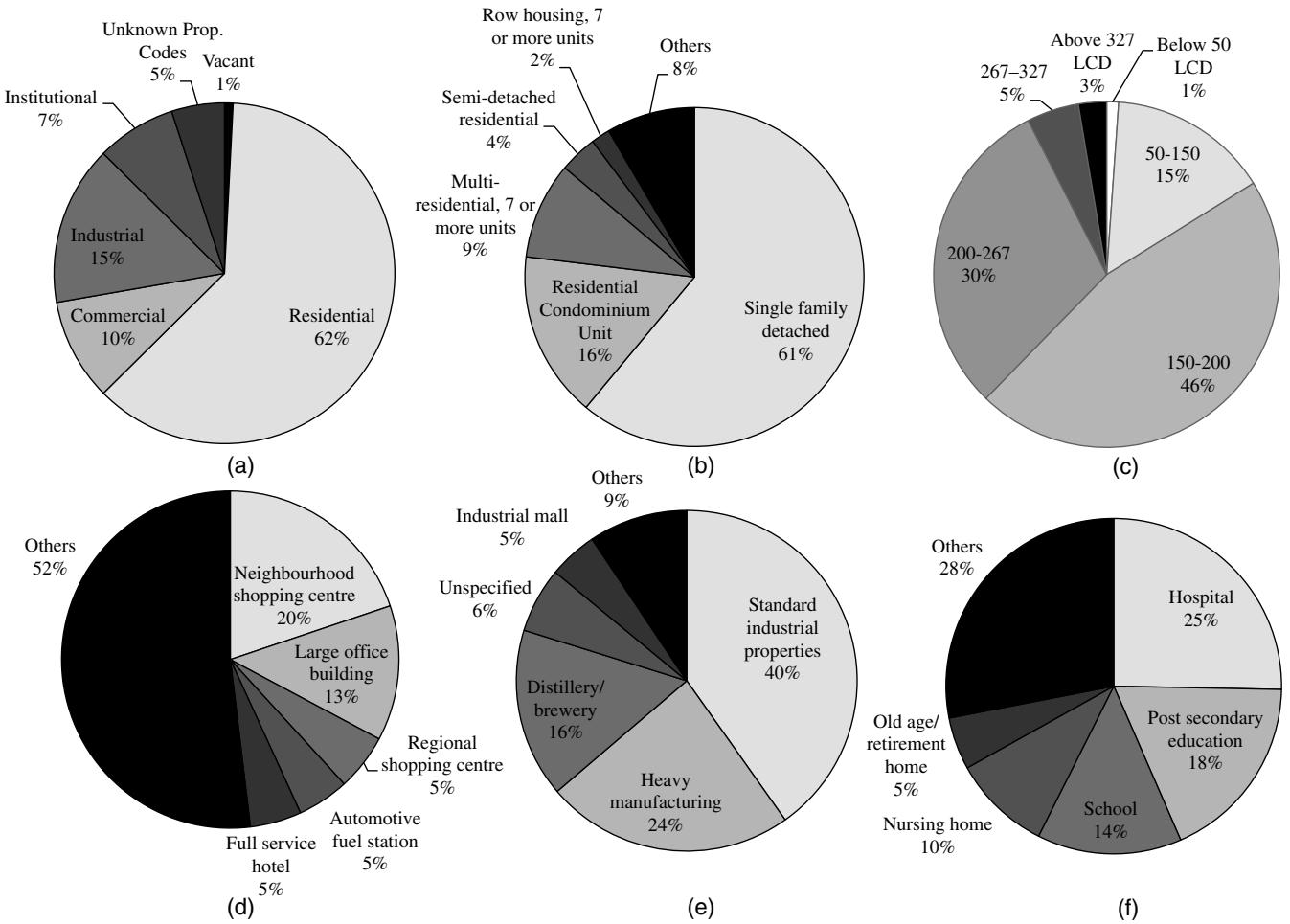

Fig. 2. Percentage of water use and users per segment (average 2011 water use for London, Barrie, and Guelph): (a) total water consumption per sector; (b) residential water consumption per property code; (c) residential water users per consumption range in liters per capita per day; (d) commercial water consumption per property code; (e) industrial water consumption per property code; (f) institutional water consumption per property code

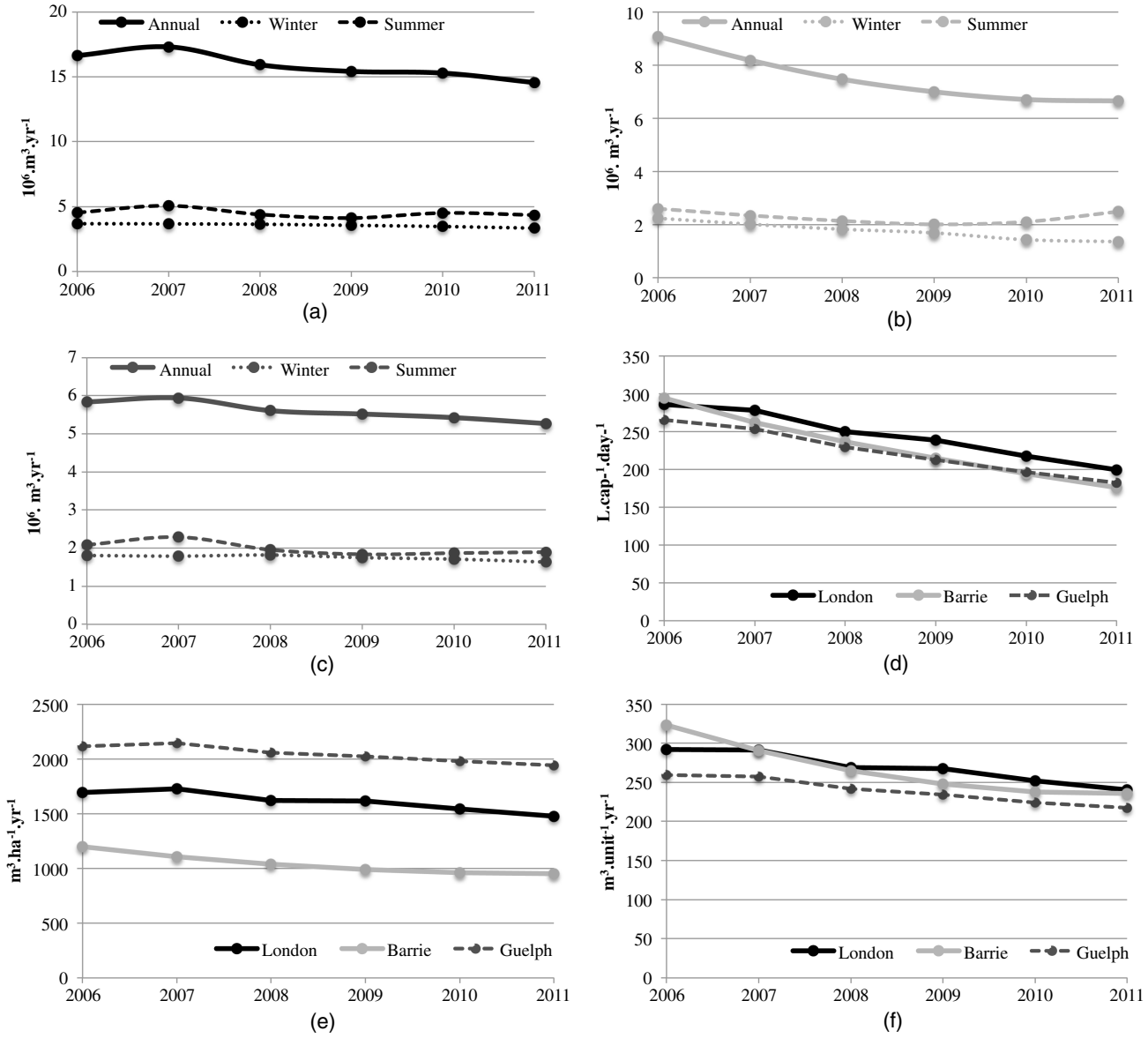

Fig. 3. Yearly and seasonal residential water consumption trends from 2006 to 2011: (a) total use in London; (b) total use in Barrie; (c) total use in Guelph; and in all three cities, (d) use per capita; (e) use per hectare; (f) use per unit

are now more residents per unit in these cities. Trends in water use per building space, not shown, are very similar to those per unit. Consumption per property area has generally decreased as well. For all three cities, residential trends in m^3/ha, Fig. 3(e), are similar to those simply in m^3, Figs. 3(a–c). Therefore, property area does not explain the variation in residential water use over time, but represents water use density. These trends are instrumental in planning, for forecasting future demands, as well as costs and revenue.

Fig. 5 displays two important characteristics of targets for conservation: (1) high water use, indicating a greater significance of the particular user type to the overall system, and potential for markedly decreasing total consumption, given the cumulative effect of various small modifications; and (2) high variation of water use metrics (m^3/m^2 or m^3/unit) within the class, evidencing the potential for improving practices. Only top water-using property codes in Barrie are shown in Fig. 5 as an example. In all three cities, within the residential sector, target property codes for conservation are single-family households, multiresidential units, condominium units, and row houses. Within the ICI sectors, targets are hospitals, shopping centers, schools, and restaurants. Because industrial property codes are less detailed, and different property types are grouped, it is more difficult to pinpoint targets. By identifying specific property types, communication regarding conservation programs can be tailored to emphasize particular efficiency devices or practices.

Given the detailed parcel-level or DB-level water-consumption data for sectors and property codes, benchmarks can be determined for each. As recommended by Morton (2011), benchmarks were defined as the 25th percentile for each group of users, a convention that accounts for the level of water consumption and user distribution. The frequency of the metrics is calculated according to their denominator, i.e., building space, unit count, property area, or population. For instance, the resulting benchmarks for residential consumption per capita are 172 L/cap · day, 149 L/cap · day, and 156 L/cap · day in London, Barrie, and Guelph, respectively.

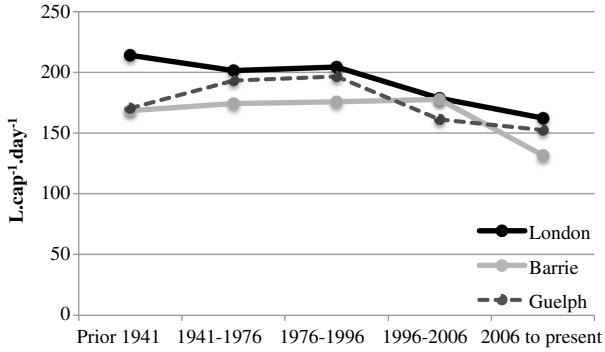

Fig. 4. Residential water consumption per capita by building vintage for Barrie, Guelph, and London

Naturally, as conservation ensues, benchmarks will change, motivating continuous improvement.

Clustering of parcel-level data indicated that users under the same property code tend to group together, i.e., they have similar water-use metrics. Therefore, users in the same class, expected to demand water for the same end uses, yet at different intensities, were shown to consume water similarly at the individual or building area level. However, there is not a clear separation between property codes. Only two to five clusters were formed within each sector, whether the process was initiated with parcel data or property codes. According to the pseudo-F statistic, hierarchical and K-means performed similarly for clustering parcel-level data. K-means performed the best in clustering property codes. Based on the Rand statistic, the hierarchical algorithm produces parcel clusters that best match property codes. In the residential sector, hierarchical clustering generated the best clusters as evaluated by the pseudo-F, and those most similar to property codes, according to the Rand statistic. The highest Rand values among all sectors were found for the residential clusters, between 0.94 and 0.96. This, however, is also due to the fact that one property code, single family dwellings, represents more than half of residential users.

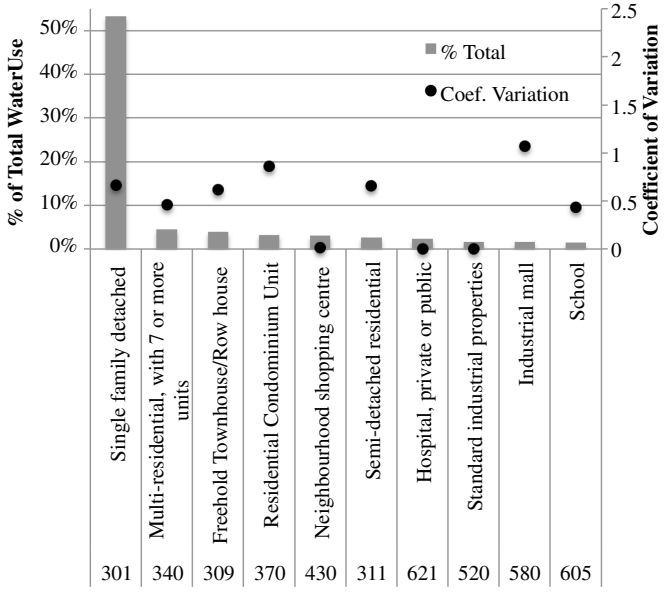

Fig. 5. Percentage of total water use and coefficient of variation of water use metrics for largest water using property codes in Barrie

Clusters of the parcel data of the ICI sectors correlated less to the property codes, since water use varies more in these classes, and the distribution of the metrics is more scattered. Values varied from 0.4 to 0.7, depending on the sector and the municipality. Although data was clustered at the property level in order to compare to the property code prototype clusters, data was still separated by sector. This distinction, however, was not confirmed by another cluster analysis, in which sectors were not separated in the input data. Although parcel data from identical property codes tend to cluster, sectors are not segregated through clustering. Therefore, separating water users into sectors does not reflect an inherent divide of the data, but facilitates user understanding.

Because the municipalities are composed not only of different property codes but also users that consume water at different levels, the clusters formed for each city differ. Although clusters differ, in all three cities, similar property codes tend to cluster, such as different types of multiresidential buildings, restaurants, shopping centers, hotels, retail stores, and nursing or retirement homes. Furthermore, within each sector a large cluster was formed, with different property codes, yet similar water metrics, and a few smaller clusters with similar types of properties. These smaller clusters, formed in one or more of the three cities, are listed below and whether their water use metrics are higher (h) or lower (l) than the other users in the "catch-all" cluster:

- Residential: residential properties with three or more self-contained units (h), cooperative housing (h), condominium units (l), other residential properties;
- Commercial: restaurants (h), shopping centers (h), hotels and motels (h), banks and similar financial institutions (l), retail (l), other commercial properties;
- Industrial: heavy nonautomotive manufacturing (h), automotive (h), private generating stations (h), water, wastewater, and waste treatment plants (h), distilleries and breweries (h), other industrial properties; and
- Institutional: hospitals (h), retirement and nursing homes (h), ambulance and police stations (h), museums and art galleries (h), other institutional properties.

Discussion

The method described herein for integrating water, land use, and demographic data is based on the data available to Canadian water utilities. However, it can be replicated for any utility where this data is collected. For those utilities which do not have access to such information, or which are beginning to plan their database, this study can assist in understanding the usefulness of different types of data, and determining which information should be collected. Altogether the study advocates a departure from the idea of simply collecting the maximum amount of information about the system, to instead collect data that can support measures for system improvement.

Data preparation, as was the case in the present study, can be the most time consuming step of the data-mining process, especially when integrating information from different sources. Data was provided at different spatial levels and in different formats. Because this type of research was not the application initially envisioned for the data, steps were not taken by the different data providers towards improving integration. A consensus between different providers would greatly facilitate the process. There should be a realization that data is not confined to sectors, and the exchange of information between departments and organizations is invaluable. Information should, thus, be collected and maintained

accordingly. Data should be easily understood by whoever may work with it.

Based on the difficulties faced in the present work, some recommendations for integrating data are:

- Request a list of identifiers of the data that is to be purchased;
- If subcontracting any part of the process be clear and specific about the procedure and expected results, because the employed professional might be an expert in data mining, but not water demand management;
- Sketch the structure of the database, input data, joins, queries, and output data, and update as needed;
- Define a descriptive nomenclature for the components of the database;
- Ensure data is formatted consistently throughout;
- Check match rates for each join or query;
- Summarize data to the desired spatial level, before joining; and
- Keep a log of issues encountered.

The applied clustering techniques divided the water users in all sectors into one large group and two to four smaller groups with similar property codes. This distinguishes high or low water-using clusters, which is instrumental in selecting targets for conservation. In order to better understand the segments of users, especially those grouped under one large cluster, and create more detailed clusters, further information could be added to the database, such as:

- Unit counts for all multiunit residential buildings;
- Unknown property codes;
- Standard industrial classification codes, which are more detailed than property codes;
- Participation in conservation programs;
- Fixture counts;
- Building or plumbing inspections;
- Temporary residents;
- Water metering; and
- User income.

Because the clustering analysis indicated that users under the same property code or similar property codes tend to cluster together, if utilities do not have other, more detailed, information available, these can be used to define segments of water users.

With the results from this data-mining process, Barrie, Guelph, and London have information to benchmark their water use internally and externally, target conservation, predict future consumption, review water rate structures, and improve communication with consumers and policy makers. The metrics and charts proposed can be used by other utilities for similar purposes and can add to a knowledge base for water systems seeking sustainable improvements. This proven methodology can also inform policy planners on potential metrics to be reported by utilities, targets to be set, and reasonable expectations.

Conclusions

Utilities have large amounts of data at their disposal, which are not being used to their full potential. By integrating this data, correlations can be found, and water use better understood. The proposed metrics allow for the comparison of normalized water consumption and promote further investigation into the causes of higher water use of certain customers within a given segment or property code. The correlations between water consumption, land use, and demographics confirm the importance of urban planning. Given these attributes, in all of the municipalities' sectors, few clusters were formed, one to six, comprising one larger group and smaller clusters of high or low water use. The clustering analysis also indicated that water users, especially residential, within the same or similar property codes tend to cluster together. Therefore, these can be used in segmenting water users if more data is not available.

Acknowledgments

This paper is partly based on case study research conducted within the project "Integrated Water Mapping: Enhancing Decision Support for Sustainable Water Planning with Municipal Data," by Cities Centre and the Canadian Urban Institute, and funded by the Ontario Ministry of Environment through the Showcasing Water Innovations program. However, the findings, interpretations, and conclusions in this document are entirely those of the authors, and should not be attributed to the aforementioned organizations. The authors acknowledge the inputs of Tom Weatherburn, Kathryn Grond, Wayne Galliher, Matt Feldberg, and Barry Thompson.

References

Boyle, C. E., Eskaf, S., Tiger, M. W., and Hughes, J. A. (2011). "Mining water billing data to inform policy and communication strategies." *Am. Water Works Assoc. J.*, 103(11), 45–58.

Brooks, D. B. (2006). "An operational definition of water demand management." *Water Resour. Dev.*, 22(4), 521–528.

Brun, M., et al. (2007). "Model-based evaluation of clustering validation measures." *Pattern Recognit.*, 40(3), 807–824.

Cahill, R., and Lund, J. (2013). "Residential water conservation in Australia and California." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000225, 117–121.

Donkor, E. A., Mazzuchi, T. A., Soyer, R., and Roberson, J. A. (2014). "Urban water demand and forecasting: Review of methods and models." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000314, 146–159.

Dziegielewski, B., and Kiefer, J. C. (2010). *Water conservation measurement metrics: Guidance report*, AWWA, Denver, CO.

Environment Canada. (2010). "Municipal water use report: Municipal water use, 2006 statistics." Gatineau, QC.

Gleick, P. H., et al. (2003). *Waste not, want not: The potential for urban water conservation in California*, Pacific Institute, Oakland, CA.

Hillenmeyer, M. (2005). "Machine learning." Stanford Univ., Stanford, CA.

Hussey, K., and Pittock, J. (2012). "The energy–water nexus: Managing the links between energy and water for a sustainable future." *Ecol. Soc.*, 17(1), 31.

Jorgensen, B., Graymore, M., and O'Toole, K. (2009). "Household water use behavior: An integrated model." *J. Environ. Manage.*, 91(1), 227–236.

Lee, J. A., and Verleysen, M. (2002). "Self-organizing maps with recursive neighborhood adaptation." *Neural Networks*, 15(8–9), 993–1003.

Morton, J. (2011). "Water management: A benchmark for Canadian office buildings." Real Property Association of Canada, Toronto, ON.

Maas, C. (2009). "H2Ontario: A blueprint for a comprehensive water conservation strategy." Water Sustainability Project, POLIS Project on Ecological Governance, Victoria, BC.

Maidment, D. R. (2008). "Bringing water data together." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)0733-9496(2008)134:2(95), 95–96.

Morales, M. A., Heaney, J. P., Friedman, K. R., and Martin, J. M. (2011). "Estimating commercial, industrial, and institutional water use on the basis of heated building area." *Am. Water Works Assoc. J.*, 103(6), 84–96.

Muste, M. V., et al. (2013). "End-to-end cyberinfrastructure for decision-making support in watershed management." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000289, 565–573.

Ontario Water Works Association (OWWA). (2006). "Water efficiency: A guidebook for small & medium-sized municipalities in Canada." Markham, ON.

Polebitski, A. S., and Palmer, R. N. (2010). "Seasonal residential water demand forecasting for census tracts." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000003, 27–36.

Sahely, H. A., and Kennedy, C. A. (2007). "Water use model for quantifying environmental and economic sustainability indicators." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)0733-9496(2007)133:6(550), 550–559.

South East Water (SEW). (2006). *Benchmarking Rep.*, Melbourne, Australia.

Shandas, V., and Parandvash, G. H. (2009). "Integrating urban form and demographics in water-demand management: An empirical case study of Portland, Oregon." *Environ. Plann. B Plann. Des.*, 37(1), 112–128.

Tan, P., Steinbach, M., and Kumar, V. (2006). "Cluster analysis: Basic concepts and algorithms." Chapter 8, *Introduction to data mining*, Addison-Wesley, Boston, MA.

United Nations Conference on Environment, and Development. (1992). "Agenda 21." Chapter 18, *Protection of the quality and supply of freshwater resources: Application of integrated approaches to the development, management and use of water resources*, Earth Summit.

United Nations Environmental Programme (UNEP). (2012). *The UN Water Status Rep. on the Application of Integrated Approaches to Water Resources Management.*

Vesanto, J., and Alhoniemi, E. (2000). "Clustering of the self-organizing map." *IEEE Trans. Neural Networks*, 11(3), 586–600.

Xu, R., and Wunsch, D. C. (2009). "Clustering." Chapter 2, *Proximity measures*, IEEE, Wiley, Piscataway, NJ.