Multivariate consumption profiling (MCP) for intelligent meter systems: a methodology to define categories and levels

J. L. Solanas and M. R. Cussó

ABSTRACT

Multivariate Consumption Profiling (MCP) is a methodology to analyse the readings made by Intelligent Meter (IM) systems. Even in advanced water companies with well supported IM, full statistical analyses are not performed, since no efficient methods are available to deal with all the data items. Multivariate Analysis has been proposed as a convenient way to synthesise all IM information. MCP uses Factor Analysis, Cluster Analysis and Discriminant Analysis to analyse data variability by categories and levels, in a cyclical improvement process. MCP obtains a conceptual schema of a reference population on a set of classifying tables, one for each category. These tables are quantitative concepts to evaluate consumption, meter sizing, leakage and undermetering for populations and groupings and individual cases. They give structuring items to enhance "traditional" statistics. All the relevant data from each new meter reading can be matched to the classifying tables. A set of indexes is computed and thresholds are used to select those cases with the desired profiles. The paper gives an example of a MCP conceptual schema for five categories, three variables, and five levels, and obtains its classifying tables. It shows the use of case profiles to implement actions in accordance with the operative objectives.

Key words | intelligent metering, leakage control, meter sizing, multivariate analysis, undermetering, water consumption profiles

J. L. Solanas (corresponding author)
AmbiTT,
SL., Bailen 50 at. 1ª,
08009 Barcelona,
Spain
E-mail: *j_solanas@telefonica.net*

M. R. Cussó
Departamento Ciencias Fisiológicas I,
University of Barcelona,
Barcelona,
Spain
E-mail: *mcusso@ub.edu*

INTRODUCTION

Water supply management is a complex activity involving the operation, maintenance and design of valuable facilities. Tariff systems, customer services, demand management, sizing and maintaining networks and meters, detection of useless consumption and undermetering prevention are the main functions in which information on consumption is important. Individual profiles and population patterns are needed to obtain well-founded predictions for water use.

Intelligent Meters (IM) add local computing functions to traditional meters, introducing new magnitudes derived from volume, time, and flow (Solanas 1996). In parallel, Automatic Meter Reading (AMR) performs quick and effective data readouts (Douglass 2005). IM plus AMR are permanently communicated to a computing centre.

doi: 10.2166/ws.2010.374

During the last decade, IM systems have been installed intensively in a wide range of uses and populations. For the first time in history we now have detailed data on consumption, but unfortunately the data remain dispersed in an excessive number of items.

At present, the data are used only for the purposes of billing, and use in management is limited to threshold control for some specific variables.

This paper proposes an efficient methodology to obtain a conceptual schema of a reference population as a set of classifying tables, one for each category. These tables are quantitative concepts to evaluate consumption, meter sizing, leakage and undermetering for populations and groupings and individual cases. They give structuring items to enhance "traditional" statistics and procedures (Farley & Trow 2003).

All the significant data from each new meter reading can be matched to the classifying tables. Then, a set of indexes is computed to select those cases with the desired profiles.

Experiences from other fields

Multivariate Analysis (MA) involves simultaneous analysis of many statistical variables (Mardia 1979; Bray & Maxwell 1985). MA applications on water use management have not yet been reported. Some experiences from other fields show a way to be followed:

- Some government departments and health care institutions use MA for census, migration, income and health studies (Aragon & Gesell 2003).
- Mass consumption, car industry, service management, telecommunications, energy and business have been the most advanced users (Held 1996; Jacobs & de Man 1996; MacDonald 2005).
- Because of deregulation and a newly competitive marketplace, some telecommunications and electricity companies have turned to sophisticated data analysis to solve fraud detection, customer retention, response to demand and dynamic maintenance (Neilson 2005).
- The use of IM and AMR systems has become standard in these fields (Blake & Kinsman 2005).
- Human health care and sports activity under medical control have recently started to use advanced statistical analysis (Cussó *et al.* 2006).

The aims of Multivariate Consumption Profiling

Multivariate Consumption Profiling (MCP) has been designed for application by water firms that:

- Have installed IM with AMR or other efficient readout procedures to obtain volume, time and flow data on a real time basis from a significant number of households.
- Wish to obtain well structured water use information to act on a timely basis.

This paper aims to describe:

- The methodology to calculate *category classifying tables.*
- The analysis of groups of cases and populations.
- The use of case profiles to implement *actions* in accordance with the *operative objectives.*

METHODS

MCP successively applies *Factor Analysis, Cluster Analysis and Discriminant Analysis* to the cases of a *reference population*.

- In Step 1, Factor Analysis is used to suggest which and how many tentative categories may be considered, and which associated variables are the most significant.
- In Step 2, Clustering Analysis finds groupings of cases that become intensity levels for each category.
- In Step 3, Discriminant Analysis confirms or rejects the efficacy of each associated variable for each category.

Using *cyclic refinements*, MCP classifies all cases into subsets of category levels, obtaining *classifying tables*, one for each category.

The classifying tables and other structuring tools integrate the *MCP conceptual schema* of a *reference population*.

We apply those tools to classify cases and calculate *case profiles* as sets of index evaluations.

Operative objectives apply *criteria* to select cases from an *external sample* with the desired profiles.

MCP concepts

- A *case* is a set of values corresponding to every *original variable* in Table 1 recorded at a *reading date* (or a time period) from *a meter in a household* (identified by its *serial number*).
- A *reference population* is a set of cases in a population on an IM metered water supply representing all the relevant features of the consumptions served.
- A *category* is a functional consumption property of all cases of a reference population. It is expressed by a set of *associated variables* selected from the ones in Table 1. Its intensity is graded in category levels.
- A *category level* is a graded group of cases characterised by their centroid (mean and standard deviation (SD)).
- A *v-l(R) category classifying table* is a set of the v associated variables by l levels obtained from a (R) reference sample.

Table 1 | MCP variables

N	The number of cases in a group.
QN	The nominal flow of a meter in cubic metres (m^3) per h.
Case identification key:	
REDT	The reading date or time period identifier.
SERN	The serial number of a meter in a household or meter identifier.
Original variables:	
Data computed locally at the meter as mean values (between two successive readings):	
CON/D	Consumption volume in m^3 per day
START/D	Number of times the meter starts per day
T%SLEEP	Time the meter is not measuring (%)
T% > QN/30	Time the meter is measuring at flow > QN/30 (%)
T% < QN/30	Time the meter is measuring at flow < = QN/30 (%)
V% < QN/100	Volume consumed at flow range: start flow to QN/100 (%)
V%QN100/67	Volume consumed at flow range: QN/100 to QN/67 (%)
V%QN67/30	Volume consumed at flow range: QN/67 to QN/30 (%)
V%QN30/3	Volume consumed at flow range: QN/30 to QN/3 (%)
V%QN3/1	Volume consumed at flow range: QN/3 to QN (%)
V%QN1 * 1.33	Volume consumed at flow range: QN to 1.33 * QN (%)
V%QN1.33 * 2	Volume consumed at flow range: 1.33 * QN to 2 * QN (%)
V% > QN * 2	Consumption volume at flow > 2 * QN (%)
Derived variables:	
Data computed from original variables specifically for this analysis:	
V% > QN/67	Consumption volume at flow > QN/67 (%)
V% > QN * 1.33	Consumption volume at flow > 1.33 * QN (%)
QMEAN	Computed as CON/D/(T% > QN/30 + T% < QN/30) * 100
CON/D/QN	Computed as CON/D/QN
QMEAN/QN	Computed as QMEAN/QN

- A *c-v-l(R) MCP conceptual schema* is the set of all *c* category classifying tables and the structuring tools designed to apply them.
- A *case profile* on a c-v-l(R) MCP conceptual schema is a set of *category continuous indexes* obtained by modelling the classifying tables.
- An *external sample* is a set of cases from an IM metered water supply to be evaluated by case profiling on a previously existing MCP conceptual schema.

MCP variables

Original data were in Structured Statistical CZ2000A Data format (http://www.contazara.net/). Variables have been restructured in a matrix format (cases in rows and variables in columns) to apply the statistical procedures. Statistical analysis was performed by using the Statistical Package for the Social Sciences (SPSS) syntax (http://www.spss.com/).

MCP variables are shown in Table 1.

Reference sample (R)

Table 2 gives the distribution of cases for the Reference sample (R), which was previously introduced by the authors to show the potential of multivariate data analysis for improving IM data systems (Solanas & Cussó 2006). This comprised 2,540 cases from a mixed urban and suburban population. Cases are from 1,252 different meters (or households). The cases from the same meter but

Table 2 | Reference sample (R). Number of cases and consumption per day classified by nominal flow

		QN										
		1.5	2.5	5	10	15	25	40	60	100	150	Total
N		1,559	191	272	303	81	62	22	38	5	7	2,540
CON/D	Mean	0.84	3.02	13.7	18	124	160	112	356	361	338	20.1
	SD	1.6	3.75	22.9	21.6	116	144	110	175	15.1	259	71.1

different dates have been used to study history of some specific households.

An initial sample of 2,241 real cases of a population with more than 100,000 metered consumptions was considered. These cases statistically represent all relevant features of premises served for the purposes of billing and tariff simulation. 299 real cases from another large town were added to complement very high flow cases. Three other external samples from other towns were used to control the consistency of the classifying results.

MCP methodology to build classifying tables

Step 1.–Factor analysis. Categories and associated variables

Factor Analysis (FA) (Kline 1994) is applied to describe variance between observed variables in terms of fewer unobserved variables called factors.

These unobserved variables, or factors, suggest concepts that can be associated to some observed variables in an inductive trial and error process. When these factors became consolidated concepts, we call them categories.

The observed variables are usually modelled as linear combinations of the factors, thus allowing analysis and measurement of the variance.

But in our case relationships are nonlinear and vary drastically for different levels of variables. That is why we combine FA with other statistical tools.

- Step 1.1.–We use FA to guess how many factors (categories) should represent the whole set of original variables, and which part of the total variance could be explained by them. In other words, the factors split total data variance into the most relevant parts.

Table 3 shows the % variance explained by each factor. By considering five main factors (categories) we explain 71.88% of cumulative variance.

- Step 1.2.–We use FA component analysis for each factor to find their associated variables. It is performed by a variance-maximising rotation of the linear variable space, taking into account all the variability in the variables (we use the Varimax-Kaiser method).

Table 4 shows the coefficients resulting from FA component analysis applied on (R) (when obtaining the rotated factor matrix if five main factors have been selected). Highest coefficients in each column indicate the main associated variables. The columns are tentative categories and the set of associated variables suggests some functional consumption property related to them:

Factor 1 suggests UNDERLOAD because of its relationships:
positive to: T% < QN/30, V% < QN/100, V%QN100/67; negative to V%QN30/3
Factor 2 suggests DEMAND because of its relationships:
positive to: CON/D, V%QN1 * 1.33, QMEAN
Factor 3 suggests LEAKAGE because of its relationships:
positive to: T% < QN/30, V%QN67/30; negative to T%SLEEP, V%QN3/1

Table 3 | Step 1.1.–Factor Analysis on (R) to find categories. Variance explained by a number of factors

Factor	% Variance	Cumulative %	Factor	% Variance	Cumulative %
1	23.76	23.76	7	6.38	85.27
2	19.76	43.51	8	6.16	91.43
3	11.42	54.93	9	5.33	96.77
4	8.55	63.48	10	1.91	98.68
5	8.39	71.88	11	1.32	100
6	7.01	78.89	12	6.12×10^{-12}	100

Table 4 | Step 1.2.–FA component analysis to find associated variables. Main associated variables have highest coefficients

	Factors				
Variables	**1**	**2**	**3**	**4**	**5**
CON/D	−0.098	**0.816**	0.328	0.063	0.114
START/D	0.15	−0.212	−0.211	**0.81**	−0.015
T%SLEEP	−0.473	−0.231	**−0.691**	**−0.343**	0.049
T% < QN/30	**0.66**	0.081	**0.641**	−0.088	−0.114
T% > QN/30	−0.184	0.31	0.266	**0.804**	0.095
V% < QN/100	**0.556**	−0.121	0.074	−0.025	0.061
V%QN100/67	**0.83**	−0.108	0.273	0.037	−0.015
V%QN67/30	0.089	0.015	**0.711**	−0.057	−0.08
V%QN30/3	**−0.781**	−0.435	0.062	−0.058	−0.011
V%QN3/1	−0.39	0.221	**−0.502**	0.028	**−0.479**
V%QN1 * 1.33	0.134	**0.559**	−0.27	0.001	−0.029
V%QN1.33 * 2	0.048	0.365	−0.172	0.058	**0.535**
V% > QN * 2	−0.04	0.081	−0.024	0.015	**0.796**
QMEAN	−0.086	**0.844**	0.115	−0.028	0.215
Variance explained (%)	23.76	19.76	11.42	8.55	8.39

Factor 4 suggests UNDERMETERING because of its relationships:

positive to: START/D, T% > QN/30; negative to T%SLEEP

Factor 5 suggests OVERLOAD because of its relationships:

positive to: V%QN1.33 * 2, V% > QN * 2; negative to V%QN3/1

Afterwards each category will be progressively refined and re-conceptualised or even renamed in search of the highest semantic and statistical consistency between categories and associated variables.

Step 2.–Clustering analysis. Levels for each category

Clustering Analysis (CA) (Everitt 1993) is the assignment of a set of cases into subsets (called clusters) so that observed values in the same cluster are similar in some sense.

We use Two-way CA to cluster the cases and also the variables. A likelihood distance measure has been chosen to determine the similarity of two cases when being compared. As CA reveals "natural" clusters of cases, we will use it to classify cases in levels within each category.

Table 5 shows the result of Step 2: the three clusters and their centroids obtained by applying CA to all (R) cases when the main four variables associated to UNDERLOAD are considered.

A cluster can be seen as a "cloud" in a multivariate space, as shown later in Figure 1. CA distinguishes different clusters of cases: each one has a centroid to which cases can be referred. The shortest likelihood distance determines to which part (cluster) each case is assigned. Clusters would be interpreted as intensity levels: "high", "moderate", "normal" are levels at which this category occurs in the population. A case is considered to have "normal" values of UNDERLOAD category variables if the likelihood distance to centroid of cluster 1 is shortest (and so on).

Step 3.–Discriminant analysis. Efficacy of associated variables

Discriminant Analysis (DA) (Huberty 1994) is applied to confirm, or reject, each associated variable.

DA attempts to find linear combinations of associated variables that best separate the groups of cases. These combinations are called discriminant functions. The procedure chooses a first function that will separate the groups

Table 5 | Step 2.–Cluster Analysis on (R) reveals groupings of cases (levels). For four main variables associated to UNDERLOAD

		Centroids							
		T% < QN/30		V% < QN/100		V%QN100/67		V%QN30/3	
Cluster	**N**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
3	279	65.21	24.76	17	23.18	17.4	16.27	20.71	19.63
2	481	64.43	32.28	8.14	7.32	76.2	14.47	2.47	4.28
1	1,780	6.8	12.69	1.47	2.6	2.49	4.32	48.7	25.9
Total	2,540	24.07	32.84	4.43	9.97	18	29.82	36.91	29.44

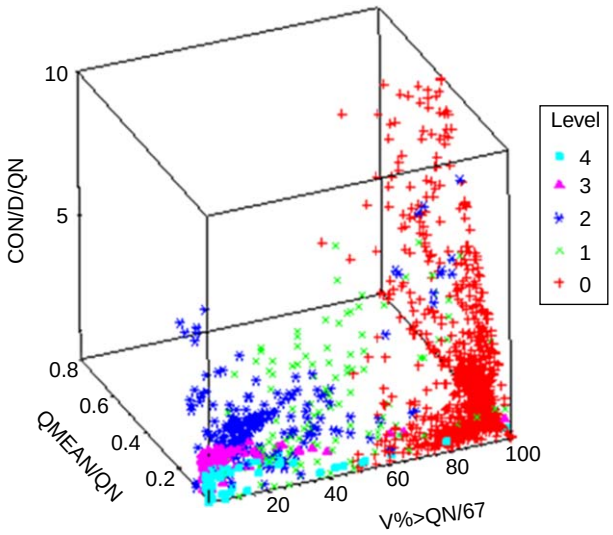

Figure 1 | Three-dimensional graph showing UNDERLOAD level groupings for all R cases.

as much as possible. It then chooses a second function that is both uncorrelated with the first function and provides as much further separation as possible, and so on.

Our aim is to confirm only those associated variables with the highest efficacy to distinguish category levels. To do so we apply DA to each category and its associated variables. Tools designed to evaluate the results obtained by DA can quantify and compare the efficacy of each associated variable.

Table 6 gives DA results of Step 3: it is applied to all four variables initially associated to UNDERLOAD:

(a) The degree of classification success is 96.2% (number of cases classified correctly);
(b) If we calculate the Structure Matrix coefficients (within-group correlations of each variable for each discriminant function), we have a low coefficient (−0.240) for V%QN30/3.

Table 6 | Step 3.–Discriminant Analysis applied to UNDERLOAD associated variables. Structure Matrix coefficients show usefulness of only three variables

	T% < QN/30	V% < QN/100	V%QN100/67	V%QN30/3	Cases correctly classified
4 ass. var.	0.693	0.520	0.939	−0.240	96.2%
3 ass. var.	0.733	0.543	0.950		95.3%

If we only take the other three variables:

(a) Cases correctly classified are still 95.3%;
(b) We improve the Structure Matrix coefficients for the other three variables.

Each centroid is now better differentiated from any other (there is less ambiguity when assigning an individual case to a level) and more category levels would be identifiable.

When varying our partial view on the whole set of MCP variables by taking only three variables, we can "see" them as more consistent and it will be easier to distinguish each group of cases.

Now we have to go back to Step 2 and apply CA again to reevaluate UNDERLOAD with only three variables because V%QN30/3 has not been confirmed.

Step 2. (second time) Clustering analysis. Levels for each category

Table 7 shows the three clusters and their centroids when performing CA with three associated variables. If we compare Tables 5 and 7, we can see that V%QN30/3 is no longer an associated variable and the statistical distributions have improved. Both category and level concepts are reinforced.

We have now a tentative 3-3(R) UNDERLOAD classifying table.

More levels will be considered in the next cyclic steps.

Successive improvements by a cyclic process

The improvement must be performed by repeating the cycle:

- Step 1.–Varying the number and/or nature of factors (categories) and/or associated variables; Trying new

Table 7 | Step 2. (second time) Cluster Analysis on (R) to obtain three levels. For the other three variables associated to UNDERLOAD (V%QN30/3 has been excluded because it was not sufficiently useful)

		Centroids					
UNDERLOAD		T% < QN/30		V% < QN/100		V%QN100/67	
Levels	N	Mean	SD	Mean	SD	Mean	SD
3	551	64.78	32.75	13.8	17.48	69.9	22.92
2	453	48.09	22.75	4.93	5.594	10.5	10.39
1	1,540	2.51	4.64	0.95	1.66	1.79	3.48
Total	2,540	24.07	32.84	4.43	9.97	18	29.82

derived variables to standardise values (for example: by dividing by QN or accumulating some original variables as shown in *derived variables* in Table 1);

- Step 2.–Varying the number of clusters (either subdividing or associating them); controlling the statistical independence of clusters;
- Step 3.–Analysing the significance of associated variables for each category.

In this inductive process, we improve previous knowledge of the field with our new experience in using data analysis tools. The results become increasingly consistent over the successive cycles. The process finishes when no more improvement is apparent.

RESULTS

Classifying tables

Once the process has finished, we have the final categories, levels and associated variables.

Table 8 shows the five 3-5(R) category classifying tables.

The MCP conceptual schema

The *category classifying tables* have a structure with three layers: categories, levels and associated variables that can be used by tools and procedures.

The *5-3-5(R) MCP conceptual schema* applies such structuring concepts on the General Linear Model. This "traditional" statistical tool incorporates a big number of well established statistical models and hypothesis tests. They can be used to analyse cases, groupings and populations and find consumption patterns for water use prediction.

Three-dimensional graph

This is a useful tool when building category levels. We can "see" all groupings and identify singular cases (those not consistent with other similar cases). When considering more associated variables, this can be done by taking the three most significant variables, or those containing new variables that we are studying in detail. This suggests potential changes in variables or categories.

With a set of three-dimensional graphs for (R) cases, one for each category, this is again useful to study an external sample in detail by matching new cases over the previous (R) groups. This is very practical when rapid comprehension of results is paramount.

Figure 1 shows a three-dimensional graph on UNDERLOAD level groupings for all R cases of our *5-3-5(R) MCP conceptual schema*.

Continuous indexes for each category and variable

We model five *category continuous indexes*, say DEM, OVLD, UNLD, LEAK, UNME, as sets of weighted linear functions (one for each level) on a range of 0–10 (very low to extra high); values >5 require attention.

Continuous indexes allow better classification of results when evaluating the intensity of categories and variables to compare cases or groupings.

Example of case profile

Table 9 gives a simple *case profile*. It gives the values for category indexes and the associated variables of each category, showing that:

DEM = 6.16	DEMAND is moderately high mainly because of QMEAN value.
OVLD = 7.96	OVERLOAD is high because of V% $>$ QN $*$ 1.33 value.
UNLD = 1.05	UNDERLOAD is very low because of V% $>$ QN/67 value.
LEAK = 0.12	LEAKAGE is extremely low mainly because of T% $<$ QN/30 value.
UNME = 6.75	UNDERMETERING is moderately high mainly because of T%SLEEP value.

Example of actions to implement operative objectives

Table 10 gives the case profiles for the top ten cases: an *external sample* of 214 readings is classified by UNLD in decreasing order.

The *operative objective* was to consider the resizing of underloaded meters. An example of *actions* and a simple criterion would be:

Table 8 | 5-3-5(R) MCP conceptual schema classifying tables

DEMAND		**CON/D (m^3/d)**		**QMEAN (m^3/h)**		**T% > QN/30 (%)**	
Level	N	Mean	SD	Mean	SD	Mean	SD
4	23	572.21	53.01	23.25	2.9	62.47	29.01
3	57	267.62	98.52	20.73	13.34	39.76	26.84
2	238	81.47	50.77	4.72	4.42	28.95	24.49
1	439	10.56	5.31	0.78	0.85	18.5	23.89
0	1,783	0.98	1.20	0.26	0.13	13.39	12.13
Total	2,540	21.33	70.95	1.44	4.9	16.77	19.83
OVERLOAD		**CON/D/QN (coef.)**		**QMEAN/QN (coef.)**		**V% > QN * 1.33 (%)**	
Level	N	Mean	SD	Mean	SD	Mean	SD
4	35	7.11	11.31	1.43	1.17	91.44	10.03
3	33	6.28	6.38	0.94	0.71	85.96	12.44
2	103	6.22	7.11	0.62	0.44	19.85	19.72
1	227	5.92	3.35	0.32	0.19	1.11	3.52
0	2,142	0.62	0.63	0.12	0.08	0.29	1.82
Total	2,540	1.49	3.1	0.18	0.29	3.52	15.4
UNDERLOAD		**CON/D/QN (coef.)**		**QMEAN/QN (coef.)**		**V% > QN/67 (%)**	
Level	N	Mean	SD	Mean	SD	Mean	SD
4	99	0.47	0.28	0.034	0.013	8.1	17.17
3	138	0.88	0.28	0.053	0.016	7.44	12.96
2	237	1.73	1.19	0.09	0.043	21.48	20.05
1	118	1.96	1.49	0.102	0.077	46.06	23.2
0	1,958	1.52	3.48	0.225	0.321	94.68	8.02
Total	2,540	1.49	3.1	0.187	0.292	77.53	33.9
LEAKAGE		**T%SLEEP (%)**		**T% < QN/30 (%)**		**START/D (sts/d)**	
Level	N	Mean	SD	Mean	SD	Mean	SD
4	118	0.22	0.45	99.17	1.22	4.29	10.64
3	40	1.31	1.86	94.52	2.99	21.35	37.11
2	157	6.54	6.22	84.57	7.82	49.77	74.43
1	386	9.67	14.3	48.68	25.75	36.17	61.64
0	1,849	79.23	22.63	7.46	12.76	147.21	191.9
Total	2,540	59.34	38.1	24.07	32.84	115.8	174.11
UNDERMETERING		**T%SLEEP (%)**		**V% < QN/100 (%)**		**QMEAN/QN (coef)**	
Level	N	Mean	SD	Mean	SD	Mean	SD
4	20	51.22	37.78	80.19	19.37	0.064	0.092
3	121	19.69	26.13	28.50	10.20	0.043	0.098
2	365	17.92	19.01	6.88	5.40	0.075	0.035
1	492	13.64	19.76	3.38	4.98	0.192	0.212
0	1,552	86.77	12.8	1.33	2.35	0.219	0.339
Total	2,540	59.33	38.1	4.43	9.97	0.184	0.287

Table 9 | Case profile READT 0704 SERN AE002156

DEM	CON/D (m^3/d)	QMEAN (m^3/h)	T% > QN/30 (%)			
6.16	67.00	13.11	21.26			
OVLD	UNLD	CON/D/QN (coef.)	QMEAN/QN (coef.)	V% > QN * 1.33 (%)	V% > QN/67 (%)	
7.96	1.05	1.45	0.09	79.53	90.33	
LEAK	UNME	T%SLEEP (%)	T% < QN/30 (%)	START/D (sts/d)	V% < QN/100 (%)	QMEAN/QN (coef.)
0.12	6.75	78.74	0.00	263.64	4.18	0.09

Action: Consider resizing to decrease undermetering if UNLD > 5 & UNME > 5.

See first two cases: lower QN would decrease undermetering and improve the metering range.

Action: Consider resizing to decrease leakage if UNLD > 5 & LEAK > 5.

See the following three cases: lower QN would improve the metering range.

Action: Inform clients of possible useless consumption if UNLD < 5 & LEAK > 5.

See the other five cases: clients have continous consumption at low flows, but there is significant consumption at higher flows.

DISCUSSION

The efficacy of classifying tables

The efficacy of classifying tables can be evaluated by DA tools in terms of number of cases classified correctly. It can be improved:

Table 10 | External sample classified by UNLD index. The top ten cases

REDT	SERN	OVLD	UNLD	CON/D/QN	QMEAN/QN	V% > QN * 1.33	V% > QN/67	
704	*CE0058*	**0.02**	**8.72**	0.45	0.09	0.00	5.54	
705	*CE4312*	**0.08**	**8.37**	0.52	0.11	0.00	6.52	
702	*3C0314*	**0.01**	**8.2**	6.48	0.27	0.00	20.29	
702	*CE3147*	**0.08**	**8.19**	3.38	0.14	0.00	2.50	
708	*3E0346*	**0.98**	**5.12**	3.11	0.13	0.00	33.78	
702	*3C0218*	**1.72**	**4.4**	8.98	0.37	0.00	66.93	
702	*3C0392*	**2.57**	**3.92**	2.39	0.10	0.00	69.63	
710	*3E0346*	**1.78**	**3.58**	4.50	0.19	0.00	62.37	
702	*3C0239*	**3.1**	**3.52**	7.25	0.30	0.00	79.41	
702	*3A0341*	**1.96**	**3.1**	9.47	0.39	0.00	82.47	
REDT	**SERN**	**LEAK**	**UNME**	**T%SLEEP**	**T% < QN/30**	**START/D**	**V% < QN/100**	**QMEAN/QN**
704	*CE0058*	**1.49**	**6.79**	78.81	3.36	268.20	41.98	0.09
705	*CE4312*	**1.82**	**6.77**	80.36	4.68	182.83	38.79	0.11
702	*3C0314*	**9.94**	**0.19**	0.00	85.44	0.00	0.00	0.27
702	*CE3147*	**9.64**	**4.77**	0.09	93.52	1.00	85.00	0.14
708	*3E0346*	**9.5**	**0.22**	0.00	91.78	0.20	0.03	0.13
702	*3C0218*	**9.82**	**0.3**	0.00	47.20	0.00	0.03	0.37
702	*3C0392*	**9.11**	**5**	0.00	88.60	0.00	2.00	0.10
710	*3E0346*	**9.23**	**0.22**	0.12	79.51	1.86	0.03	0.19
702	*3C0239*	**9.44**	**0.22**	0.01	69.07	0.11	0.00	0.30
702	*3A0341*	**9.52**	**0.18**	0.02	30.79	0.57	0.00	0.39

(1) By adding more associated variables to each category. We have compared results based respectively on 5-3-5 and 5-5-5 (R) MCP conceptual schemas, that is, by considering five associated variables instead of three for each category. Most improvement has been obtained for UNDERLOAD (90.80 to 97.10%) and UNDERMETERING (83.60 to 93.50%).
(2) By introducing more sophisticated derived variables.

As an example, it has been found that V% < QN/100 * QMEAN/QN improves results as an UNDERMETERING associated variable.

More associated variables and more sophisticated derived variables obtain better quantitative results and give more semantic elements for a better conceptual interpretation. They could be the object of future communications.

Comparing populations and samples

To compare pairs of samples (reference samples or external samples) we use the classical Welch–Student's *t* test (Welch 1947). We calculate *t* test parameters for every pair of centroids for all variable levels of each pair of samples.

MCP results obtained from other large samples of well balanced mixed populations give similar classifying tables. Sizing or other particular management features do not much affect the span and structure of classifying tables and derived indexes. FA, CA and DA groupings and parameters are highly dependent on case distance to the centroids but not on case "density" (the number of cases by cluster does not imply the mean and SD of centroids).

As counterexamples, when we compare (R) to irrigation consumption samples, we observe significant deviations for underload, leakage and undermetering levels.

Those results are not shown herein, but the authors invite the reader to contact them for more details.

A more "universal" schema

Water metering practices differ according to climate and tradition. Agglomerations with scarce water sources tend to support more costly devices such as IM in order to manage consumption.

Data from samples with a wider range of consumption habits need to be processed to ensure a "universal" conceptual schema.

MCP methodology can be applied to any IM significant sample to compare results with those of Table 8 and thus help in introducing Multivariate Analysis to water supply management.

CONCLUSIONS

Some leading water companies have used Intelligent Meter systems for many years and have performed systematic data retrieval. But taking full advantage of these technologies is difficult because the data are inherently repetitive and sparse. Most of them are not processed, and their potential is lost.

The MCP methodology is based on multivariate analysis tools (Factor Analysis, Cluster Analysis and Discriminant Analysis). It constructs an MCP *conceptual schema* consisting of a set of classifying tables and other related tools. It extends concepts that are currently used in water supply management.

Demand, overload, underload, leakage and undermetering became quantitatively defined categories. *Actions* based on them can be applied using easily defined *operative objectives*.

The method could be extended to populations with other consumption habits in order to obtain more "universal" tools.

MCP offers a new method based on *multivariate statistics* that is particularly useful for analysing the properties, habits, and trends of the populations served.

REFERENCES

Aragon, S. & Gesell, S. 2003 A patient satisfaction theory and its robustness across gender in emergency departments: a multigroup structural equation modeling investigation. *Am. J. Med. Qual.* **18**(6), 229–241.

Blake, M. & Kinsman, C. 2005 *Establish AMR and Real-time Pricing Simultaneously to Maximize Economic Benefits*, AMRA Autovation Annual Symposium.

Bray, J. H. & Maxwell, S. E. 1985 *Multivariate Analysis of Variance*. Sage Publications Inc., Thousand Oaks, CA.

Cussó, R., Guiu, M. & Solanas, J. L. 2006 Natural Groupings and Modelling in Sports Medicine. *BIFI. II International Congress: From Physics to Biology: The Interface Between Experiment and Computation*. Zaragoza, Spain.

Douglass, J. 2005 *Advanced Metering*, Extension Energy Program, Washington State University, Washington, USA.

Everitt, B. S. 1993 *Cluster Analysis*. John Wiley & Sons. Inc., New York.

Farley, M. & Trow, S. 2003 *Losses in Water Distribution Networks; A Practitioner's Guide to Assessment Monitoring and Control*. IWA Publishing, London.

Held, J. R. 1996 Clusters as an economic development tool: beyond the pitfalls. *Econ. Dev. Q* **10**(3), 249–261.

Huberty, C. J. 1994 *Applied Discriminant Analysis*. John Wiley & Sons, New York.

Jacobs, D. & de Man, A. P. 1996 Clusters, industrial policy and firm strategy: a menu approach. *Technol. Anal. Strateg. Manage.* **8**(4), 425–437.

Kline, P. 1994 *An Easy Guide to Factor Analysis*. Routledge, London.

MacDonald, G. 2005 *Hydro One Comments on OEB's Draft Smart Meter Implementation Plan*. Hydro One Networks Inc., Toronto, Ontario.

Mardia, K. V. 1979 *Multivariate Analysis*. Academic Press, New York.

Neilson, R. 2005 *New Wave of Performance-Based Regulation Transforms AMR Business*, AMRA Autovation Annual Symposium.

Solanas, J. L. 1996 *Intelligent Water Meters Are Here*, AMRA Autovation Annual Symposium.

Solanas, J. L. & Cussó, R. 2006 *Efficient Process and Analysis for Intelligent Metering*. IWA World Water Congress, Beijing, China.

Welch, B. L. 1947 The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34**(1–2), 28–35.