Evaluation of statistical sampling for the assessment of residential consumption totals in water distribution networks

A. Fortunato, C. Arena and M. R. Mazzola

ABSTRACT

The paper provides insights into stratified sampling, a standard statistical technique that may be employed to assess domestic water use in water distribution networks. The basic idea is to use only a few meters to provide inference on the total water consumption of a network or of a district metered area through the knowledge of some additional stratification variables, such as household typology, size and occupants number. Since any sampling procedure assumes that the variance of the variable at stake is known, either a suitable amount of past consumption data is necessary, or a specific preliminary survey must be carried out, in order to define the sampling plan. An application with real consumption data from a small municipality in Sicily (Italy) shows that number of occupants for each household is sufficient to design an effective sampling plan and that the methodology can be successfully applied in the technical practice, thus allowing a dramatic reduction of the number of customer meters to be read in order to quantify total water consumption compared to standard practice based on the reading of all meters.

Key words | statistical sampling, water demand, water distribution networks

INTRODUCTION

Developing water balances is a basic step to assess the level of water losses in district metered water distribution networks and is hence preliminary to the selection of asset management options such as leak detection and control, pipe rehabilitation or substitution. Water losses can in fact be appraised as the difference between water entering the network, or a district metered area (DMA), and authorised consumption (IWA 2005). While there are generally few entering points that are relatively easy to meter, authorised consumption, and in particular its billed component, is instead the aggregation of a large number of metered deliveries. Albeit conceptually straightforward, the water balance methodology has hence its major drawback in the metering of water uses, as the metering system can be (inhomogeneously) old or even absent in some instances, or it can prove difficult to organise extensive and detailed surveys of consumption data, whose format and information attached are in most cases designed for accounting purposes, rather than for technical ones. In addition, the amount of time required to implement such procedures may be incompatible with a short-term strategy of losses monitoring. Smart metering solutions for household water consumptions are now developing and full-scale applications to towns are starting to be recorded, as for instance, in Malta. However, it still requires quite large investments that cannot be afforded by all organisations. For such reasons, water balances are often given up in the diagnosis of water losses in a water distribution network in favour of bottom-up approaches such as minimum night flow (MNF) measurement (Farley & Trow 2003).

The MNF in a DMA, usually occurring between 2:00 and 4:00 a.m. (Thornton 2005), is a meaningful indicator of leakage issues. During this period, authorised consumption is at a minimum and leakage is at its maximum, reaching the maximum share of the total flow entering the DMA due to the highest pressure occurring during the
day. Leakage is signalled by the high ratio between MNF and legitimate night water use. The latter cannot nevertheless be measured accurately in the DMA since it includes a large number of connections, therefore it is generally evaluated either by using guidelines derived from many surveys undertaken in various parts of the world (South African Water Research Commission 1999) or specific techniques which include conducting an assessed night use study (Cheung et al. 2010; Loureiro et al. 2010; Alkassem et al. 2013). Furthermore, if it is known that there is an exceptional night use within the zone, this must also be estimated or measured by carrying out meter readings during the minimum night period (Thornton 2005).

Considering the drawbacks of the MNF methodology, it can hence be of interest to explore the option to provide only few representative selected customers with a perfectly working and updated metering system, and to use it for the assessment of total water consumption, thereby exploiting the potential of the water balance technique.

Curiously enough, the idea has received little attention both from the research community and from water utilities: after the pioneering work by Hanke & Mehrez (1979) on micro-data of peak hour water use, although the importance of sampling domestic water use has been recognised and illustrated (Mays 2004), little research and field work has been produced on the subject: Zhang & Brown (2005) investigated domestic water consumptions in Beijing and Tianjin by using stratified sampling method based on the predominant housing typologies, Arreguin-Cortes & Ochoa-Alejo (1997) apply stratified sampling procedures directly to losses measurements, while Speight et al. (2004) suggest stratified sampling methodology as a suitable procedure for water quality assessment in water distribution systems and proposed stratification criteria that are then evaluated via a synthetic data set. Other handbooks (e.g. US National Research Council 2002) also recommend stratified statistical sampling for water use estimation but at a state-wide scale and for the various water sectors in place (e.g. irrigation, commercial, industrial, nuclear power, etc.). To some degree, such reduced set of experiences is in contrast with the considerable development of studies and tools for modelling time series of domestic consumption (e.g. Buchberger & Wu 1995; Magini et al. 2008; Mohamed & Aysha 2010; Zhai et al. 2012).

In this paper we work with micro-data of residential water use to assess the aggregated water use of a whole network or DMA by means of statistical sampling, in order to investigate its applicability and effectiveness in the technical practice.

The paper is organised as follows. After a brief review on the theoretical background of sampling, a case study is introduced dealing with the estimation of total domestic water consumption in a small municipality in Sicily (Italy) adopting the stratified sampling technique.

**METHODS**

Statistical sampling (Cochran 1977; Cocchi 2006) allows us to obtaining of information on a given characteristic $\eta$ of a population $\Omega$ of $N$ elements through a subset (sample) of $\Omega$, formed by $n$ sampling units ($n < N$). $\eta$ may be the water consumption of the $N$ domestic connections of a DMA or a municipality. Sampled values will be denoted with $y$ so as to distinguish them from population character $\eta$.

The information on the studied characteristic consists of a synthetic value $f(\eta)$, for example the mean or the total, assessed using as estimator the same function of the sampled values $h(y) = f(y)$. The issue is to quantify the number of sampling units to draw such that the function of the sampled values $h(y)$ can be assumed to approximate the synthetic value $f(\eta)$ with given precision $\epsilon$ and confidence level $1 - \alpha$.

In statistical terms this means that $f(\eta)$ will fall within the interval $[h(y) - \epsilon, h(y) + \epsilon]$ with a probability of $1 - \alpha$.

Sampling techniques vary according to the way in which the $n$ sample units are selected out of the population. Simple random sampling, for instance, is a method of selecting sample units such that all the possible samples have the same chance to be drawn. In stratified random sampling the population is instead preliminarily subdivided into $M$ non-overlapping subpopulations $\Omega_k$ (called groups or strata) of size $N_k$ ($1 \leq k \leq M$), and the subsamples of size $n_k$ are drawn randomly from the corresponding strata. Some basic definition and results of sampling theory are reported in the following (Cocchi 2006).

Denote with $m(\eta)$, $\sigma^2(\eta)$ and $s^2(\eta) = N/(N-1) \cdot \sigma^2(\eta)$ the population mean, the population variance and the commonly used unbiased estimator of population variance, respectively. In stratified sampling, strata means
where

\[ m(\eta_k) = \frac{1}{N_k} \sum_{i=1}^{N_k} \eta_i \]

and strata variances \( \eta^2(\eta_k) = \frac{1}{N_k} \sum (\eta_i - m(\eta_k))^2 \), within-strata variance \( \eta^2(\eta) \) and between-strata variance \( \eta^2(\eta) \) are also defined:

\[ \eta^2(\eta) = \frac{1}{N} \sum_{k=1}^{M} N_k \eta^2(\eta_k) \]

\[ \eta^2(\eta) = \frac{N}{N - M} \eta^2(\eta) \]

\[ \eta^2(\eta) = \frac{1}{M} \sum_{k=1}^{M} N_k \eta^2(\eta_k) \]

Within-strata variance, \( \eta^2(\eta) \), is the weighted average of the group variances, \( \eta^2(\eta_k) \), the weights being the groups sizes, while between-strata variance, \( \eta^2(\eta) \), is a measure of the dispersion of groups means, \( m(\eta_k) \), around the population mean, \( m(\eta) \).

It is possible to show that the following additive decomposition of population variance holds (Cocchi 2006)

\[ \eta^2(\eta) = \eta^2(\eta) + \eta^2(\eta) \]

Stratified sampling methodology consists of choosing a stratification that maximises the distance among different groups (between-strata variance) thus minimising within-stratum variance: since within-stratum variance drives the required sample size, reducing it through appropriate stratification criteria will entail a significant reduction of the overall sample size.

In random sampling the number of sampling units \( n_{\text{random}} \) necessary to obtain an estimation of the population’s mean with precision \( \epsilon \) at a confidence level \( 1 - \alpha \) is

\[ n_{\text{random}} = \left( \frac{\epsilon/z_{\alpha/2}}{s(\eta)} + \frac{1}{N} \right)^{-1} \]

where \( s(\eta) \) is the population variance that can be substituted by its unbiased estimator \( s(\eta) \) and \( z_{\alpha/2} \) is a given percentile of the distribution (assumed symmetric, e.g. Gaussian) of standardised estimator \( h(y) \).

In stratified sampling \( n_{\text{strat}} \) is instead given by the following:

\[ n_{\text{strat}} = \left( \frac{N(\epsilon/z_{\alpha/2})^2}{\sum_{k=1}^{M} N_k s^2(\eta_k)} + \frac{1}{N} \right)^{-1} \]

where within-stratum variances, \( s^2(\eta_k) = \frac{1}{(N_k - 1)} \sum (\eta_i - m(\eta_k))^2 \), appear in lieu of the total population variance and hence \( n_{\text{strat}} \) will always be less than \( n_{\text{random}} \).

When the sample size is small and/or the exact distribution of the sample means is unknown, the t-Student distribution can be used instead of the standard normal distribution, yielding conservatively larger samples than those provided by the standard normal distribution due to its thicker tails. An issue arising when using the t-Student distribution is that the appropriate number of degrees of freedom is equal to \( n_{\text{strat}} - 1 \), being \( n_{\text{strat}} \) unknown; however \( n_{\text{strat}} \) can be iteratively assessed starting from the value yielded by the normal distribution.

If sampling fractions \( f_k \) are the same for each stratum, that is, if \( f_k = n_k/N_k = n/N = f \), then stratified sampling is defined proportional. In optimal stratified sampling \( n_{\text{strat}} \) is instead subdivided among the different strata in such a way that the \( n_k \) minimise the estimator variance (on which \( \epsilon \) depends), thus maximising estimation precision. Variance minimisation yields (Cocchi 2006)

\[ n_k = \frac{n_{\text{strat}}}{M} \frac{s(\eta_k) N_k}{\sum_{k=1}^{M} N_k s(\eta_k)} \]

subsamples sizes \( n_k \) are then proportional to subpopulation variances \( s(\eta_k) \) and to dimensions \( N_k \).

If the quantity to estimate is the population’s total, all the preceding equations still apply, with a precision of \( \epsilon \cdot N \).

Any sampling procedure assumes that the population variance is known, thus implying that a suitable amount of preliminary data is available for its estimation. In the case of stratified random sampling such preliminary information should be larger than for random sampling, since the available data will be subdivided between the groups. In addition, stratified sampling requires appropriate stratification criteria.
to be defined. The stratification criteria will be then applied to the set of preliminary data and group variances will be estimated. In the case of domestic water uses, reasonable stratification criteria are the household typology (e.g. detached houses or semidetached houses, with or without garden, flats, etc.), the number of users per connection (which seems to be the most meaningful variable) and quite probably also the income. These are after all some of the basic explanatory variables used in demand characterization studies, such as demand function and elasticity assessment (e.g. Young 2005).

Unfortunately, apart from past water consumptions, such data are not always readily available to the water utilities, whose databases simply contain the contracts holders’ data, without further information about households sizes, number of occupants, etc., so extra-utility, preliminary surveys are necessary.

**RESULTS AND DISCUSSION**

**Case study description**

The practical applicability of sampling methodologies has been checked in a full-field real data case study. The annual water consumptions, from 1995 to 2008, of each residential connection (2074 overall) of Santo Stefano di Quisquina, a small town in Sicily of about 5000 inhabitants, were collected and for each year the average water use per connection was calculated by summing up the annual consumption of each connection and dividing the sum by the number of connections. These values were then scaled to a daily basis by dividing them by 365, as in Table 1.

The town shows a declining demography, with an average population reduction of 0.7%/year. Clearly the number of connections oscillates, as some new ones are opened and some are closed. For this study, however, we only considered the connections that have been active along the whole 1995–2008 period. Given the limited population dynamics, these connections represent the vast majority of the whole residential consumption of the town and their composition in terms of number of users per connection (Table 2) can be assumed to stay constant along the years.

In Santo Stefano di Quisquina there are only individual connections since every household is equipped with its own meter, even if it is located in one of the few, small condominiums in town. All the meters are class B according to the international metrological standards. The age of the meters is considerably variable and ranges from 0 to 30 years, thus making it difficult to classify them according to the

### Table 1: Statistical data of water consumption of the case study

<table>
<thead>
<tr>
<th>Year</th>
<th>Population water use total, ( t(\eta) ) [m³/day]</th>
<th>Population water use mean, ( m(\eta) ) [l/conn./day]</th>
<th>Population water use standard deviation, ( s(\eta) ) [l/conn./day]</th>
<th>Population water use variance, ( s^2(\eta) ) [l²/conn.²/day²]</th>
<th>Population water use variation coeff., CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>787.8</td>
<td>379.9</td>
<td>201.8</td>
<td>40,706.1</td>
<td>0.531</td>
</tr>
<tr>
<td>1996</td>
<td>778.0</td>
<td>375.1</td>
<td>195.9</td>
<td>38,379.2</td>
<td>0.522</td>
</tr>
<tr>
<td>1997</td>
<td>775.5</td>
<td>373.9</td>
<td>198.7</td>
<td>39,499.0</td>
<td>0.531</td>
</tr>
<tr>
<td>1998</td>
<td>745.4</td>
<td>359.4</td>
<td>190.8</td>
<td>36,398.0</td>
<td>0.531</td>
</tr>
<tr>
<td>1999</td>
<td>769.3</td>
<td>370.9</td>
<td>196.9</td>
<td>38,766.0</td>
<td>0.531</td>
</tr>
<tr>
<td>2000</td>
<td>779.0</td>
<td>375.6</td>
<td>199.3</td>
<td>39,737.8</td>
<td>0.531</td>
</tr>
<tr>
<td>2001</td>
<td>746.9</td>
<td>360.1</td>
<td>191.7</td>
<td>36,732.7</td>
<td>0.532</td>
</tr>
<tr>
<td>2002</td>
<td>777.2</td>
<td>374.7</td>
<td>199.6</td>
<td>39,853.0</td>
<td>0.533</td>
</tr>
<tr>
<td>2003</td>
<td>777.3</td>
<td>374.8</td>
<td>199.2</td>
<td>39,663.4</td>
<td>0.531</td>
</tr>
<tr>
<td>2004</td>
<td>782.0</td>
<td>377.1</td>
<td>200.8</td>
<td>40,337.5</td>
<td>0.533</td>
</tr>
<tr>
<td>2005</td>
<td>796.1</td>
<td>385.9</td>
<td>203.9</td>
<td>41,556.6</td>
<td>0.531</td>
</tr>
<tr>
<td>2006</td>
<td>819.4</td>
<td>395.1</td>
<td>209.8</td>
<td>44,022.7</td>
<td>0.531</td>
</tr>
<tr>
<td>2007</td>
<td>837.2</td>
<td>403.7</td>
<td>214.4</td>
<td>45,957.2</td>
<td>0.531</td>
</tr>
<tr>
<td>2008</td>
<td>835.2</td>
<td>402.7</td>
<td>213.8</td>
<td>45,702.8</td>
<td>0.531</td>
</tr>
</tbody>
</table>
recent MID European Directive. From 1995 to 2006 meters were read yearly by staff from EAS, the regional utility managing water services. From 2006 onwards the service has been operated by the municipality, and meters are read by municipality's staff.

Santo Stefano features a remarkable uniformity in household typology and dimension (two-three storey buildings without garden, each storey forming one household), so that these variables are not likely to be effective to explain consumption variance. Income could be a relevant stratification variable, but it is difficult to obtain income micro-data for so many customers. For these reasons, in this work the number of occupants per connection was adopted as the only stratification variable. To obtain this information, the customers database has been cross-checked with the municipal civil registry in order to link each connection with the corresponding number of users in a strictly anonymous way: in this application this was possible because the water service is operated by the municipality that also holds the civil registry.

As clearly shown in Table 1, the average daily consumption per connection varied over the years from 359.4 l/conn./day to 403.7 l/conn./day. Considering that the average number of occupants per connection is 2.4, the resulting per capita daily water consumption ranged between 150 and 168 l/capita/day. However, the most relevant aspect from the standpoint of sampling method application is that the internal population variability of water consumptions remained nearly constant, as shown by the yearly values of the coefficient of variation, which ranged from 0.522 to 0.533. Stratification was then performed on the basis of the statistical population of average daily water uses of connections, pooling all data between 1995 and 2007 (Table 2) and leaving the 2008 data for validation of the sampling procedure, as described in detail below. Table 2 contains all the stratum statistics necessary to apply the methodology.

The extremely small water use standard deviation of the first stratum is partially due to the fact that, as said before, for each connection the average daily water use, pooling all data between 1995 and 2007, was assumed, thus smoothing the variability. Furthermore, the 655 connections belonging to stratum 1 are themselves rather uniform as far as consumption is concerned, as they mainly correspond to single, retired, persons.

### Sampling plan design

Assuming one wants to assess the average daily use per connection in 2008 with precision $\epsilon$ of $\pm1.5$ l/day (corresponding to 0.4% of the actual average daily use per connection) and confidence level $1-\alpha$ of 99%, from the data of Table 2, using Equation (7), a total number of sampling units $n_{\text{strat}} = 71$ was obtained using $t_{\alpha/2} = 2.6490$; if applying Equation (7) with $z_{\alpha/2} = 2.5758$, $n_{\text{strat}}$ would be equal to 67. This sampling size, divided among strata according to Equation (8) and rounded off to the nearest integer number, increases to 76, resulting in a sample fraction $n/N$ of 3.6% (Table 3). The sampling size of stratum
Table 3 | Subdivision of total sample size among the strata

<table>
<thead>
<tr>
<th>Number of users per connection</th>
<th>Sample size per stratum, ( n_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>( \geq 6 )</td>
<td>3</td>
</tr>
<tr>
<td>Total sample size</td>
<td>76</td>
</tr>
</tbody>
</table>

6 should be 1, but it has been opted for a minimum sample size of 3 for each stratum.

It is worthwhile highlighting that random sampling (Equation (6)) would yield, for the same precision and confidence level, a sample size of 2038, corresponding to a sample fraction of 98.4%. While this latter value depends on the very high precision and confidence level required (for instance, obtaining \( \epsilon = 20 \text{l/conn./day} \) with \( 1 - \alpha = 99\% \), would require an \( n_{\text{random}} = 505 \), corresponding to a sample fraction of 24.3%), it nonetheless shows the importance of selecting appropriate stratification criteria and the effectiveness of the criterion adopted in this specific case study: admittedly, the availability of databases containing information about the number of occupants per connection is, in many cases, out of the reach of water utilities. Other criteria are however adoptable: for instance, in areas with different dwelling typologies (condominiums, apartment houses, etc.), the number of households per connection could be used, although it is expected to provide strata with higher variance and ultimately larger sampling sizes.

Validation of the sampling procedure

To validate the sampling procedure, 1000 different random samples of \( n_{\text{strat}} = 76 \) water uses were drawn from the 2008 average daily consumptions data set: this means that for each of the 1000 random samples, \( n_k \) connections are selected randomly from the corresponding stratum, according to the sampling plan of Table 3, together with their average daily consumption in 2008; each sample of 76 water consumptions is then used to infer \( m(y) \), the total average daily consumption of all the residential connections in Santo Stefano di Quisquina.

\( m(y) \) ranged between 400.0 l/conn./day and 404.5 l/conn./day, with a mean value over the 1000 samples, \( E(m(y)) \), of 402.7 l/conn./day and a standard deviation of 0.75 l/conn./day. The 99% confidence interval of \( E(m(y)) \) consequently resulted 402.7 ± 1.9 l/conn./day. It should be pointed out that \( E(m(y)) \) equals exactly the observed average daily consumption per connection in 2008, as reported in the last row of Table 1.

In addition, the confidence interval of \( m(y) \), with a confidence level of 99%, has been further calculated for each of the 1000 samples: it included the ‘true’ value of \( m(\eta) \) in 98.7% of the cases. The width of \( E(m(y)) \)'s confidence interval and the percentage of correctness of the single estimates are partly due to the number and specific selection of the samples, but also to the fact that in this case study the sample mean’s distribution is not normal, as it was checked through a statistical test.

CONCLUSIONS

The paper has discussed the basic issues involved in designing a sampling plan for domestic water use in water distribution networks. Possible applications of this methodology include the periodic assessment of the level of water losses through the drawing of water balances at the scale of DMA, real-time leakages control (by means of a network of metering devices conveniently placed and read remotely) as well as water consumption models calibration. In fact, the major drawback of the water balance methodology is that the metering system of water uses can be old, imprecise or even absent in some cases.

The review of the main results of sampling theory shows that stratified sampling always allows a great reduction of sample size, compared to simple random sampling, with sampling fractions depending on consumption variances, estimates’ precision and confidence level. The paper also highlights the need for water use data in order to preliminary estimate the overall population’s and groups’ variances, and to perform the stratification. The most relevant stratification criteria for domestic connections are household size and typology and, above all, the number of people per connection. This latter piece of information is admittedly not always readily available to water utilities.
A full-field real data application of the sampling technique to the domestic consumptions of a small town in Sicily, Italy, which provided results very close to the ones predictable by the theory, shows that it can be applied in technical practice, even adopting as only stratification criterion the number of users per connection. Incidentally, it turns out that using stratified sampling only 76 sampling units need to be monitored in order to estimate the annual average daily use per connection of 2074 connections with a confidence level of 99% and the precision of ± 1.5 l/day. Random sampling would instead yield 2040 sampling units. Such extremely good performance of stratified random sampling is however due to the availability of a good set of previous consumption data for all the connections, to be employed for the calculation of population and strata variances, and to the availability of the probable best possible stratification variable, namely the number of occupants per household: the unavailability of this information and the use of a less informative criteria such as apartment size would certainly result in larger sample sizes, while the unavailability of good consumption data series would imply the need for the execution of preliminary surveys to assess population and strata variances and may affect the confidence in results.

It should be highlighted that sampling methods are profitable also when meters are not missing: the time required to implement data consumption surveys can in fact be incompatible with the needs of the asset planning process and with a quasi-real-time strategy of losses monitoring; besides, meters readings would be not synchronised, thus yielding less precise results than those based on few, homogeneous, precise and reliable measures, possibly to be provided by smart metering systems.

REFERENCES


First received 7 March 2014; accepted in revised form 19 September 2014. Available online 30 September 2014.