Missing values imputation in a load curves sample: An approach combining time series and survey sampling techniques.

Anne De Moliner*, EDF R&D (Clamart, France), anne.de-moliner@edf.fr

In the near future, tens of millions of load curves measuring the electricity consumption of French households in small time intervals (probably half hours) will be available. All these collected load curves represent a huge amount of information which could be exploited using sampling techniques. In particular, the total consumption of a specific customer group (for example all the customers of an electricity supplier) could be estimated using sampling methods. Unfortunately, data collection, like every mass process, may undergo technical problems at every point of the metering and collecting chain resulting in missing values. This problem is similar to the non response phenomenon in surveys: it reduces the accuracy of the estimators and may generate bias. In order to minimize these problems, we have to impute missing values. Two types of imputation methods are usually implemented in that context: what we will call "static" methods coming from sampling theory (for example regression imputation) consisting in imputing the missing value using the consumption of other units at the same instant and possibly auxiliary information and what we will call "dynamic" methods coming from time series theory (for example exponential smoothing) imputing the missing value using the consumption of the same unit at other instants. Both types of methods leave aside a part of the information (the information of the unit for static methods and the information of the instant for dynamic methods). The aim of this communication is to present new imputation methods taking into account simultaneously and as efficiently as possible the so-called static and dynamic information. As we use more information, these combined methods are expected to perform better than static or dynamic ones. Two ways of combining the information (linear regression of real values on estimated values for the respondents, and minimization of the estimated imputation error) are tested. For these methods, the optimal combination depends on the length of the missing values series and on the position of the missing value in the series. The variances of the combined estimators are estimated using a population bootstrap. The methods are compared to each other and to static and dynamic ones, on real datasets.

Key Words: non response, electricity, industry