

Data-Stream and multi-level statistical data recovery strategy: an application to data-stream from electrical communicating meters¹

Alain Dessertaine²

Electricité de France Research & Development, e-mail: alain.dessertaine@edf.fr

Abstract: This communication is aimed to make a state of the art of statistical tools from the survey sampling theory and their application and adaptation in the context data recovery from 34 millions of communicated meters installed from now to 2016 by Electricite De France. A statistical structure of multi-level data recovery is proposed, based on several phases sampling design to allow the use of appropriate recent survey sampling techniques like balanced sampling or calibration approach using Hilbertian variables, or calibration on cumulative distribution function. Those interests will be illustrated by examples using a population of 35000 electrical consumption curves to show how to manage and anticipate some data-transmission delays, and theirs impacts to build real-time estimations.

Keywords: Survey Sampling, Functional data analysis, balanced sampling, data-stream

1. Introduction - Why a several phases strategy

The installation planned from now to 2016 of more than 34 millions of communicating meters for all of EDF's consumers should allow a better analysis of the EDF customers' spending patterns. Indeed, these meters will be used like sensors in order to measure all the load curves of each customer to very fine temporal granularities (measuring until the minute, even the second). Research work is underway to develop an information system to manage and make best use of this information. They address a joint use of technical summaries and analysis of data streams and tools and techniques of sampling and estimation discussed and developed in the framework of the Survey Sampling theory.

The large volume of data to be processed will require very strong constraints in terms of transmission and data storage. Some studies are being done to propose space-time sampling strategies of streams to be used ; in order to take into account some preliminary constraints to our future treatments, we will refer to these first studies. So, we know that we will not be able to preserve (nor to collect) the whole of the curves of consumption at the step second or the step minute for all customers. Let us imagine as a working hypothesis that we will manage a panel with a great number of customers, and that we will work on specific summaries of theirs consumption curves, according to the complexities of their own process of consumption. We could imagine that we will be able to collect curves with different temporal granularities (Chiky and al, 2008) with a data stream sampling.

We can also assume that the needs of users require different degrees of fineness of the data. So we can imagine a scheme for the development of various output streams of information

²Address of correspondence: 1, avenue du Général de Gaulle - 92141 Clamart Cedex - France

with summaries which could be limited to some "simple" queries, or should necessitate development and management of curves samples. Thus, some flows can be sampled upstream of load curves database (possibly temporary database) to limit transmission (via a adapted management data transmission event delays from the meters, for example) or to limit the potential storage (especially in the context of temporal data with very fine granularity). Then, downstream, or from the load curves database we called "Access panel", summaries from specific applications or space-time curves samples could feed some new database or new queries, which we shall call by language default "Statistics-Datamart". For example, we could elaborate a specific Datamart developed for the forecasting, using the principle of forecast by aggregation / disaggregation show in Dessertaine (2008) or in Poggi and al (2008); in this context, it will be necessary to retain samples of curves with historical sufficiently long to develop models in each cluster specifically designed for customers from their load curve and their predictability. Then, it will be necessary to conduct sampling in successive step. This succession of sampling is well known as two or more phases sampling.

We will identify in this communication some tools from the surveys sampling theory that may be useful to treat data-stream from the meters. We will specifically look at the two or more phases sampling design, the unequal probabilities and balanced sampling designs. We illustrate these techniques in their likely applications at Electricite De France.

Let us consider a population U consisting of N elements labeled $i=1, \dots, N$ (the EDF portfolio, for example). For each customer i , one (or more) communicating meter(s) measures at each time t the value of a variable C , which takes the value $c_{i,t}$. We can also look at a variable curve measured on an interval $[T, T + P]$, which will then take the following functional form $C_i(t)$. We then define a sampling design to build a sample S of n elements i among the N of the population U with a probability of selection choosed by the statistician, with an algorithm Sampling within these probabilities and a sampling strategy used in the hope of control "at best" sampling errors, such as stratified sampling design, two or more phases sampling design, or unequal probabilities sampling design, for example - to a broad view of these algorithms, see the book by Yves Tille (2006). Some adaptation are current to use these tools with functional data, like in Cardot and al (2009).

2. Two phases sampling design

The two or more phases sampling strategy presented in the introduction of this document requires us to detail the several phases sampling design. First, we consider only two phases design. After a formal presentation of this design and its use in general practice, we will describe in more detail its relevance to our sampling strategy. Several scenarios can be envisaged:

- The technical or costs constraints related to data transmission and storage will probably necessitate partial recovery of them, especially if we could clearly recover very fine time granularities data (a given information each minute, even each second). A space-time sampling may then be useful in this phase. Such samples are referred to in Chiky et al (2008) and some ideas about specific's estimation with non-synchronous measures are discussed in Dessertaine (2007).
- Some algorithms, analyzes and designs used on datamarts statistical issues identified above may be particularly difficult, even impossible to implement on all data (or in a too

big sample, like the database named "access panel"). Also, a smaller sample may be preferred. It will be good enough if it randomly brews the entire population, and if it will have a good performance in terms of accuracy. At this level, we can implement balanced samples, as presented in Deville et al (2004) and Chauvet et al (2006), in second phase of a space-time sample.

For real-time estimation of some indicators (or near real-time), we can anticipate how to manage the transfer of data to not obstruct the system. Thus, even if we work on a meters sample, we can develop this sample to built an efficient sub-sample (using old data like auxiliary information in a specific unequal balanced sampling design) to recover the data in a shorter sample and estimate our indicators in real time.

3. Balanced sampling design using Hilbertian auxiliary Information

The last two scenarios presented above may use two techniques already developed and widely used in practice surveys, the balanced and/or unequal probabilities sampling design.

A balanced sampling design (with equal or inequal probabilities) has the important property that the Horvitz-Thompson estimators of the totals for a set of auxiliary variables are equal to the totals we want to estimate. Therefore, the variances of all the variable of interest are reduced, depending on their correlations with the auxiliary variables.

The balanced samples are random sample built under constraints to respect the balanced equations. Only samples within these constraints can be learned. In the N -size sample space, samples verifying the balanced constraints belong to an affine subspace of dimension $N-p$ (with p , the number of auxiliaries variables used). This geometric property allowed the development of a specific algorithm, called the cube method (Chauvet et al, 2006) to select a sample that is perfectly balanced if possible, otherwise select a sample of the most balanced possible if perfect balance is impossible, and in all cases respect the inclusion probabilities choosed by the statistician.

Such a sample can be applied in the case of a second phase sample (from the "access panel" defined above). Indeed, in most developing datamarts statistics, balancing on the phenomena of the temporal power observed will be appropriate to maintain a performant sample.

Work presented in Dessertaine (2007) has shown interest in wavelet-transform curves $C_i(t)$ and curve $C(t)$ (sum of each curves on the population) so as to balance the sample on the approximation coefficients. In this paper, a example set to judge the relevance of this approach. Thus from a population developed with a set of 2300 curves measured each half-hour over a period of 1 month around the Christmas holidays in the year 2000, the details of balanced samples were ever better than obtained with an unequal probability sample of the same size. Gains recorded in this work to estimate the total curve in the same month in 2001 are particularly important over the estimation period. So the sampling errors on the unequal probabilities varied around 2.9 percent, while the balanced strategy corresponding to a balancing of the 64 approximation coefficients of level 5, after a wavelet decomposition of Haar, varied around 0.44 percent, a decrease of about 80 percent of the sampling error.

4. Conclusion - extensions

This paper described some technics used in the survey sampling practice, and showed the interest of their potential used of a multi-level structure Statistical data recovery from the future 34 millions communicating meters which will be installed at all EDF's customers.

Some extensions are current :

- We could enrich the sample data with auxiliary variables, such as temperature and Historical data available on the entire population. We will try to improve our estimators with non-parametric functional estimation adapted to the framework of the survey theory.
- We will take into account the presence of non-response for some elements by reweighting techniques, such as calibration, developed in Deville et al (1992), and in Ren (2000).
- We will take into account the presence of outliers greatly disrupting the robustness of our estimates.

References

- Cardot H. and Josserand E., (2009), Sondages stratifiés pour données fonctionnelles : allocation optimale et bandes de confiance asymptotiques, *XLIeme journées de statistique*, SFdS, Bordeaux, France (Mai 2009).
- Chauvet G. and Till Y., (2006), A fast algorithm of balanced sampling, *Journal of Computational Statistics*.
- Chiky R., Cubille J., Dessertaine A., Hebrail G. and Picard M.L., (2008), Gestion de panel dans un environnement de flux de données, *Revue des nouvelles technologies de l'information - actes EGC 2008*.
- Dessertaine A., (2006), Sampling and Data-Stream : some ideas to built balanced sampling using auxiliary hilbertian informations, *56th ISI Conference : IPM56 - New methods of sampling*, Lisbonne, Portugal (22-29 August 2007)..
- Dessertaine A., (2008), Non-parametric load forecasting by "aggregation/disaggregation", *International Symposium of forecasting*, Nice, France (22-25 june 2008).
- Dessertaine A., (2008), Echantillonnage et Flux de données : Estimation de courbes de consommation électrique a partir de données fonctionnelles, in *"Methodes de sondage : applications aux enquetes longitudinales, a la sante, aux enquetes electorales et aux enquetes dans les pays en developpements"*, Dunod.
- Deville J.C. and Tille Y., (2004), Efficient balanced sampling : the cube method, *biometrika*, vol 91, pp 893-912
- Misiti A., Misiti Y., Oppenheim G. and Poggi J.M., (1998), Methode d'ondelettes en statistique : introduction et exemples, *Journal de la Societe française de Statistique*, vol 139,4, pp. 3-29.
- Misiti A., Misiti Y., Oppenheim G. and Poggi J.M., (2008), Optimized Clusters for Disaggregated Electricity Load Forecasting, *Compstat 2008 Proceedings*, pp. 225-232, Porto, Portugal (August 2008).
- Ren R., (2000), Utilisation d'information auxiliaire par calage sur fonction de repartition, *these de doctorat*, Universite Paris-Dauphine.
- Sarndal C.E., Swensson B. and Wretman J., (1992), Model assisted survey sampling, Springer-Verlag, New-York.
- Tille Y., (2006), Sampling algorithms, Springer-Verlag, New-York.