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

Anne De Moliner[1]

[1] *EDF R&D, 1 avenue du général de Gaulle 9210 Clamart, France, anne.de-moliner@edf.fr*

Abstract

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 reduces the accuracy of the estimators and may generate bias and in order to minimize these consequences, we have to impute missing values. Two types of imputation methods are usually implemented in this context: what we will call "static" methods coming from the 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 the time series theory (for example exponential smoothing) imputing the missing value using the consumption of the same unit at other instants. The aim of this communication is to present new imputation methods taking into account simultaneously 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. The optimal combination depends on the length of the missing values series and on the position of the missing value in the series. Two ways of combining the information (linear regression of real values on estimated values for the respondents and minimization of the global estimated imputation error) are compared to each other and to static and dynamic ones, on real datasets. The variances of the combined estimators are estimated using a population bootstrap.

Keywords: big data, energy, industry, nonresponse

1. Introduction

In the next few years in France, tens of millions of smart meters will be deployed and will collect the individual load curves of residential and business customers at short time steps (probably half hours). This deployment will result in a huge increase in the amount of available data for energy suppliers such as EDF (Electricité de France) and power grid managers. However, it may be complex to stock and exploit such a large quantity of information, therefore it will be relevant to use sampling techniques to estimate load curves of specific customer groups (e.g. market segments, possessors of a specific equipment or clients of an energy supplier). Most studies are performed at an aggregated level so we usually don't need to preserve the coherence of the individual curves.

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 very similar to nonresponse in survey sampling: it deteriorates the accuracy of the estimators and may generate bias if the clients affected by missing values are different from the clients with complete curves. In order to reduce the impact of this phenomenon, an often used solution consists in imputing the missing

values, i.e. filling in the gaps with values as pertinent as possible.

For this purpose, two types of methods are commonly applied:

- Techniques used in survey methodology, consisting in carrying out the imputation independently, instant by instant, using the consumption of other units from the sample at the considered time to determine the imputed value. This class of methods includes for example regression imputation, ratio imputation and class mean (documented for example in Ardilly (2005)) or donor based imputation such as hot deck imputation (Andridge and Little (2010)) Later in the article we will refer to these methods as "static" methods.
- Times series based techniques, consisting in completing each curve independently by exploiting the information provided by non missing values of the curve and that we will refer to as "dynamic" methods. We can cite for example Holt-Winters triple exponential smoothing (Winters (1960)) or more basic methods such as linear interpolation or historical imputation (using the value observed a week before). These methods are commonly used by EDF statisticians because, contrary to the static ones, they preserve the internal coherence of each curve. Nevertheless they can create biases if the nonresponse occurrence process is not independent from the value of the electric consumption.

With both types of techniques, some information is left aside (information contained in the rest of the curve for static methods and information provided by the rest of the sample at the considered instant for dynamic ones). In this paper, we will propose and try out two different ways of combining one or more static imputation methods (also called static estimators) and one or more dynamic imputation methods (dynamic estimators) to exploit more information and to enhance the quality of estimations.

These methods, developed in the context of electrical consumption studies can thus be applied to any curves sample.

2. Methods

We consider the problem of estimating the global electric consumption of a finite population consisting in *N* customers at different times $t \in [1, T]$ *:*$(Y_t)_{t=1..T}$, with

$$Y_t = \sum_U y_{it},$$

where y_{it} denotes the consumption of unit *i* at the time *t*. We select a sample *s* of size *n* according to a given sampling design *p(s)*. Let $\pi_i = P(i \in s)$ and $w_i = \frac{1}{\pi_i}$ denote the Horvitz-Thomson weight. In the absence of nonresponse, a standard estimator of Y_t is:

$$\hat{Y}_t = \sum_s w_i y_{it}$$

Let r_t and o_t denote respectively the subsets of respondents and non-respondents[1] at time *t* and let y_{it}^* denote the imputed value for unit *i* at time *t*. The imputed estimator in presence of nonresponse becomes:

$$\hat{Y}_t^* = \sum_{r_t} w_i y_{it} + \sum_{o_t} w_i y_{it}^*$$

The missing value layout concept:

[1] In an attempt to be concise, we will use the classical survey sampling terminology and refer to units without or with a missing value at time t as "respondent" or "nonrespondent" at time t, even if the occurrence of missing values does not result from the will of the client.

The relative performances of imputation methods depend on the length of the missing values sequence and on the position of the considered missing value in the sequence: it seems intuitive that dynamic methods will perform better for short gaps or near the extremities of the gap whereas these considerations have no influence on static methods which are implemented independently for each instant, regardless of the rest of the curve.

Therefore, we define here the concept of "layout" of a missing value, which in this article will refer simultaneously to the length of the sequence the missing value belongs to and to its location in the sequence. We propose here to find the optimal combination of static and dynamic methods for each missing value layout.

To be perfectly thorough we should define one layout for each length and each location in the sequence (e.g. "third missing value in a sequence of 12") but in order to limit computation time, we will gather together some of the layouts (e.g.: "isolated missing value", "missing value in a sequence of 3"," missing value near the extremity in a sequence of 10",…).

In particular, we will gather in a unique layout all the missing values of « too » incomplete curves, that is to say curves with a missing values rate over a threshold defined by the statistician. These curves do not contain enough information to use a dynamic method so we will only apply a static one.

Combination methods:

As mentioned in the previous section, the efficiency of dynamic methods depends on the missing value layout but is not impacted by the missing values rate at the considered instant whereas the efficiency of static methods depends only on the features of the instant and not on the missing value layout. Moreover static imputation can be improved by grouping units into homogenous imputation classes according to some auxiliary information**. In order to take all these elements into account we will determine the optimal combination of static and dynamic methods independently for each triplet (instant x layout x imputation class).**

We propose the following procedure for combining the estimators:

- 1. For each missing value of the dataset, estimate the imputed values for the static and dynamic estimators.
- 2. For each instant, simulate missing values of each layout for each respondent, and collect the predicted values given by static and dynamic estimators for these simulated gaps.
- 3. For each triplet (instant x layout x imputation class), assess the quality of the two methods by comparing the imputed values to the real ones on the respondents' dataset and, depending on their relative performance, determine the optimal combination.
- 4. Apply these combinations to real missing values.

The most intuitive way to carry out the third step of this procedure is to use a linear combination of the estimators. In that paper, we will present two ways of finding the combination coefficients: linear regression and global imputation error estimation.

The first proposed combination method is to fit the linear regression of the real value on the predicted values estimated by each estimator (static and dynamic) on the respondents' dataset. As mentioned in step 2, in order to find these predicted values we have to simulate missing value sequences (for each layout) and to collect the predictions.

For a given triplet (instant x layout x imputation class), assume the model:

$$Y_i = a + s\hat{y}_i^s + d\hat{y}_i^d + \varepsilon_i$$

Where $\hat{y}_i^s$ denotes the static prediction for unit *i* (unique for all layouts) and $\hat{y}_i^d$ the dynamic one (depending on the layout). We will then assume that $(\varepsilon_i)_i$ are iid white noises and fit this model using Ordinary Least Squares (OLS). It will give us the coefficients $\hat{a}, \hat{s}, \hat{d}$ for each triplet which will be used in step 4 to impute each missing value: $y_i^* = \hat{a} + \hat{s}\hat{y}_i^s + \hat{d}\hat{y}_i^d$. The time indicator *t* is omitted to lighten the notations.

This method seems intuitive and natural and its implementation is also quite easy. Moreover, the presence of the intercept in the regression can correct potential biases of the dynamic estimator. Furthermore, we can assume that the resulting prediction will not be too bad at the individual level because the regression is specifically conceived to minimize individual prediction errors. However, this method is based on an individual optimality criterion and not on a global one, so we can't be sure that it also produces the best estimator for the aggregated curve on the whole sample and yet we often want to be as precise as possible at the global level. For this reason, we have developed a second combination method based on an aggregated quality criterion, presented in the next paragraph.

For a given triplet (instant x layout x imputation class), we want to determine the combination parameter $\varphi \in [0,1]$ such as the combined imputed value $y_i^* = \varphi\hat{y}_i^s + (1 - \varphi)\hat{y}_i^d$ has "the best performance as possible" at the aggregated level according to a "well chosen" criterion.
Let ε_i^e denote the imputation error for method *e (e=s or d)* and unit *i* (at time *t*): $y_i = \hat{y}_i^e + \varepsilon_i^e$. The ε_i^e can be calculated for each respondent.
Let's define the "estimated global imputation error" of an estimator *e (e=s or d)* as :

$$\hat{L}(e) = \sum_{i \in r_c} \frac{(1-\hat{p}_i)}{\hat{p}_i} w_i^2 (\varepsilon_i^e)^2 + \sum_{i \neq j \in r_c^2} \frac{(1-\hat{p}_i)}{\hat{p}_i} \frac{(1-\hat{p}_j)}{\hat{p}_j} w_i w_j \varepsilon_i^e \varepsilon_j^e$$

Where $\hat{p}_i$ denotes the estimated presence probability of unit *i* (estimated by building homogenous nonresponse groups and then estimating the presence rate at the considered instant for each group) and r_c the respondents' subset (of imputation class *c)*. Time and configuration indicators are omitted to lighten the notations.

The lower this estimated global imputation error is, the better the estimator is. Indeed, each individual error is weighted proportionally to its sampling weight and its missingness probability $(1 - \hat{p}_i)$ and the $\frac{1}{\hat{p}_i}$ is an expansion term used to extrapolate results on the respondent subset to the whole sample. The interaction of two errors (e.g. if a dynamic method has a constant bias) is also taken into account thanks to the second term.

We will choose the combination parameter $\hat{\varphi} = \frac{\hat{L}(d)}{\hat{L}(s)+\hat{L}(d)}$ and impute the value:

$$y_i^* = \hat{\varphi}\hat{y}_i^s + (1 - \hat{\varphi})\hat{y}_i^d$$

We can actually demonstrate that, as the dynamic and static error are independent, this parameter $\hat{\varphi}$ minimizes an estimator of the mean square error of $\hat{Y}_t^*$ conditionally to the observed sample and nonresponse.

Contrary to the linear regression, this combination method aims to optimize the imputation at a global level. Its main drawback is that it doesn't include any intercept, so the estimated total after imputation could be biased if the dynamic imputation method is biased. However, we can reasonably think that, if its bias is too big, the dynamic method will have a small weight in the combined estimator.

The extension of these two combination methods to more than one dynamic and/or static estimators is straightforward.

Variance estimation:

To estimate the variance of the total after imputation, we have developed a variance estimation method based on Booth's population bootstrap (1994).

We build a superpopulation by replicating x_i times the curve of each unit i of the sample (including the incomplete ones) where $x_i = floor(\pi_i)$ and then, as the weights are not always integers, we complete the population with a simple random sampling. Then we draw with replacement a large number of resamples of size n (the original sample size) with respect to the original sampling design (stratifications, inclusion probabilities, balance,…). Next, we impute independently the missing values on each resample using the chosen combining method and calculate the total after imputation for each resample. Finally the observed variability of that imputed total (at each instant) among the resamples provides an estimator of the variance.

Then, in order to take into account the randomness of the nonresponse mechanism, we need to add a corrective term. Indeed, the global variance can be decomposed into:

$$V = E_r\left(V_p\left(\hat{Y}^*\middle|a\right)\right) + V_r(E_p(\hat{Y}^*|a)) = V_1 + V_2$$

Where E_r, V_r denote the expectation and variance with respect to the nonresponse mechanism and E_p, V_p the expectation and variance with respect to the sampling design. The bootstrap procedure gives an estimation of V_1, so we have to add an estimate of V_2. Assuming that the nonresponse probability is constant by class, V_2 can be estimated by:

$$\hat{V}_2 = \sum_K (1 - \widehat{p_k}) \sum_{r_k} w_i \sum_{c=1}^{C} \pi_c (y_i - \hat{y}_{i,c}^*)^2$$

With $\hat{y}_{i,c}^*$ the imputed value of i for layout c, K and C respectively the number of imputation classes and layouts, π_c the probability of layout c if the value is missing.

3. Simulations on a real dataset

We worked on a sample of 770 complete load curves with a measure every half hour for two weeks. For each curve, the electricity tariff and previous yearly consumption of the client is available. We created six imputation classes by separating the clients into three homogenous consumption groups for each of the two tariffs. In order to focus on the impact of nonresponse, we worked on the whole datasets instead of drawing samples.

On that curves, we simulated missing data sequences of size 1, 2, 3, 4, 24 and 48, occurring randomly with a homogeneous probability for every client and every instant, for a total nonresponse rate of 10%. Then we added simultaneous missing values sequences of length 1 and 48 for 10% of the units. This missing data simulation process was repeated 80 times on the original sample.

We tested the following imputation methods: mean class imputation, regression on the previous year consumption, linear interpolation, triple exponential smoothing, triple exponential smoothing and regression imputation combined by linear regression, triple exponential smoothing and regression imputation combined by global imputation error estimation.

4. Results

For each imputation method, and on each sample, we measured the mean and quantiles of the relative estimation errors over the instants $t \in [1..T]$: $REE_t = \frac{|Y_t - \hat{Y}_t|}{Y_t}$ (global level) and $REE_{i,t} = \frac{|y_{it} - \hat{y}_{it}|}{\sum_{\tau=1}^{T} y_{i\tau}}$ (individual level) and then calculated the mean of those quantities (mean and quantiles of REE) over all the samples:

	Global level REE (%)						**Individual level REE (%)**					
Method	Mean	Median	q10 %	q25 %	q75 %	q90 %	Mean	Median	q10 %	q25 %	q75 %	q90 %
1.Class mean	0,89	0,73	0,14	0,34	1,27	1,85	75	49	9	22	93	158
2.Régression imputation	0,85	0,70	0,13	0,33	1,21	1,78	63	43	8	20	77	123
3.Linear interpolation	1,23	1,01	0,18	0,47	1,75	2,54	65	34	5	13	77	159
4. Smoothing	0,83	0,68	0,12	0,31	1,18	1,77	56	**31**	**5**	**13**	66	131
5.Combination (regress)	0,72	0,59	0,11	0,28	1,04	1,52	**53**	33	6	15	**62**	110
6.Combination (global crit)	**0,71**	**0,59**	**0,11**	**0,28**	**1,02**	**1,49**	**53**	34	6	16	**62**	**108**

The individual results were also detailed for each sequence length. We also analysed the combination coefficients in the second method and, logically, the weight of the static estimator increases when the length of the sequence grows.

5. Analysis and discussion

In our tests, static methods performed better than dynamic ones at the global level and at the individual level for long gaps (≥24 missing values), whereas dynamic ones were better at an individual level for short gaps. Combined methods perform better that static or dynamic ones at the global level and at the individual one for long gaps. The combination using the global criterion seem to give slightly better results. The calculation times are quite reasonable (< 1 hour by sample) and are mainly due to the dynamic imputation process.

6. Conclusions

We have presented here an imputation procedure using an adjusted combination of static and dynamic imputation methods for each instant, class of units and missing values sequence length. Two ways of combining have been presented: the regression, conceived to minimize the individual imputation error, and a global criterion which is recommended if we are only interested in the total estimator.

In both cases, this combination reduces the precision loss due to missing values at the global level and at the individual one for long gaps more than static or dynamic methods alone. The variance of the estimated total can be estimated using a bootstrap procedure.

However, if we need to make online imputation on very large datasets, the dynamic imputation and so the combined one can be very long (because each curve is treated separately) and it may be reasonable to only use static imputation if we only want to work at an aggregated level.

References

[1]Ardilly, P. (2005), "Sampling Methods", Springer Verlag
[2]Andridge R, and Little, R (2010), "A Review of Hot Deck Imputation for Survey Non-response", *Int Stat Rev.*,78, 40-64
[3] Winters, P. R. (1960), "Forecasting sales by exponentially weighted moving average". *Management Science*, 6, 324-342.
[4] Booth J.G., Butler R.W. and Hall P. (1994), "Bootstrap Method for finite population" *JASA* 89,1282-1289