Exploiting auxiliary data: Random Forest Regression estimator

Luis Sanguiao, INE (Spanish NSI), luis.sanguiao.sande@ine.es

Abstract

Suppose we have, in addition to our survey data, some auxiliary data A that are related to our target variables, but we don't know how. The source of the data could be, for example, administrative data or even Big Data, and it covers an important subset of the target population. An important question arises: how to exploit the data to improve our estimations? The increase of the accuracy would allow a reduction in the sample size of the survey.

We propose a new regression estimator based on random forests: the Random Forest Regression Estimator. Being similar to the GREG estimator (Särndal et al., 1992) it has the same advantages that random forest regression has over linear regression. In particular, the relationship between A and target variables is learnt directly from data and both discrete and continuous explanatory variables can be used directly. This also means that the method easily accommodates a change in the structure of the auxiliary data.

The estimator is nearly unbiased and we give an approximate estimator of its variance for stratified random sampling. It is important to note that the estimator remains unbiased even if the model is poor. Several simulations are run, with both synthetic and real data to show its performance in practice.

Keywords: Multivariate auxiliary information, Sample size reduction, Machine Learning, Imputation, Random Forest, Regression estimator

1. Introduction

Suppose we have a finite population $\mathcal{P} = \{u_1, \cdots, u_N\}$. In $\mathcal{P}$ we want to estimate the totals of variable x, so we can take a sample $\mathcal{S} = \{j_1, \cdots, j_n\}$ and the Horvitz-Thompson estimator (Horvitz and Thompson, 1952)

$$\widehat{X} = \sum_{i \in \mathcal{S}} \frac{x_i}{\pi_i} \tag{1}$$

which is unbiased. We also have unbiased estimators for its variance.

Now suppose we have some auxiliary data known for each element of the population $\mathcal{A} = \{a_1, \cdots, a_N\}$ where a_i are vectors of numbers and/or factors. The auxiliary data is also related somehow to population target variables, so we should be able to improve X estimation with the help of $\mathcal{A}$. We have two main options:

1. Use $\mathcal{A}$ to choose a sampling design (usually stratified sampling) that makes variance smaller. There are some drawbacks. Since you are using the same sample for all target variables, your estimator will be optimum only for one variable (or not optimum but a *good* estimator for several variables). Another drawback is that you may have to redesign the sampling from time to time.

2. Use $\mathcal{A}$ after sampling. Since a is related to x, you may *predict* x_i from a_i. Some examples of this are calibrated estimators, such as GREG estimator. These estimators are (nearly) unbiased and their variance can be approximately estimated.

Of course they are not exclusive. It is usual to combine GREG estimator with stratified sampling.

We propose an estimator, similar to GREG but based on Random Forest Regression (Breiman, 2001) instead linear regression. We will focus on stratified simple random sampling. Suppose we have a stratified design with L strata, and let denote N_h the size of the stratum h and n_h the size of the sample in the same stratum.

Definition 1.1. For a stratified simple random sample, the Random Forest Regression estimator is given by

$$\hat{X}_{RFRE} = \sum_{i=1}^{N} \hat{x}_i + \sum_{h=1}^{L} \sum_{i=1}^{n_h} N_h \frac{x_{hi} - \hat{x}_{hi}}{n_h} \tag{2}$$

where $\hat{x}_i$ is the random forest regression estimation of x_i when $i \notin \mathcal{S}$ and the *out-of-bag* estimation of x_i when $i \in \mathcal{S}$.

Note that the first summand is a synthetic estimator and the second summand is an stratified estimator but for the error of the model instead the variable.

This estimator has several advantages compared to the GREG estimator:

- The statistician does not need to make any assumption about the model.
- Any change on the model or on the auxiliary data available, is automatically detected and corrected by the algorithm (as far as possible)
- Dependencies far from being linear do not suppose a problem.
- It works out-of-the-box with both discrete and continuous explanatory variables.

- Irrelevant variables tend to be discarded so there is little harm in including a moderate amount of them in $\mathcal{A}$.

Of course the performance of the RFR estimator may be worse than performance of the GREG estimator, but this is unlikely to happen for all variables in general since you have one model per variable. Its disadvantages compared to the GREG are:

- There is no weighting system: RFR estimator is non linear.
- The model is a blackbox.

Even in the absence of weighting system, there is an obvious interpretation of the estimator: as has been already noted, we are elevating the errors of the model to the population aiming to correct the model bias. From this point of view we are still using the design weighting system.

2. Estimator properties

The following theorem shows that the RFR estimator for stratified random sampling has good properties.

Theorem 2.1. *The RFR estimator is nearly unbiased and an approximate estimator of its variance is given by*

$$\hat{V}_{RFR} = \sum_{h=1}^{L} N_h^2 (1 - \frac{n_h}{N_h}) \frac{s_{e,h}^2}{n_h} \tag{3}$$

where $s_{e,h}^2$ is the sample variance of the out-of-bag error in each stratum h.

Proof. See (Sanguiao, 2018). □

In (Sanguiao, 2018) it is also given an unbiased version both for the estimator and the variance estimator. In this work the approximate versions are used, because

- The estimator in Definition 1.1 is easily computed from the output of existing random forest implementations.
- The bias is very small, specially for the estimator (a bit higher for the estimator of the variance).

- The first implementation with the unbiased formulae, turned out to be too slow to run the simulations in a reasonable time (for a single estimation the speed would be more than enough).

3. Simulations

Several simulations have been done, to test the performance of the estimator. To make variances a bit smaller, we have chosen the estimator of the mean. The estimations are compared to the known (since it is a simulation) population mean. The estimators of its variances are compared to the variance of the simulation estimations. This is so because RFR estimator exact variance involves lots of random forest models which is computationally unaffordable.
The main results shown are:

Bias of the estimator It is the difference between the mean of the estimations and the population mean. It is shown even for unbiased estimators as a reference for negligible bias. The units are in percentage of the population mean.

Variance of the estimator It is actually an approximation, taking the sample variance of the simulation estimations. The units are in percentage of the sample mean variance for simple random sampling.

Bias of the estimated variance It is the difference between the mean of the estimations of the variance and the approximated variance of the estimator.

Variance of the estimated variance Again an approximation taking the sample variance of the simulation variance estimations.

3.1. Synthetic data with simple random sampling

The population size has been taken to be 1000 and 100000 samples of size 100 are extracted.
The first subset has been generated as follows:

- A follows a Poisson with $\lambda = 15$.
- B follows a lognormal with $\mu = 3$ and $\sigma = 1$.

- $C = 3A - 2B + e$ where e is uniform on $(0, 100)$.

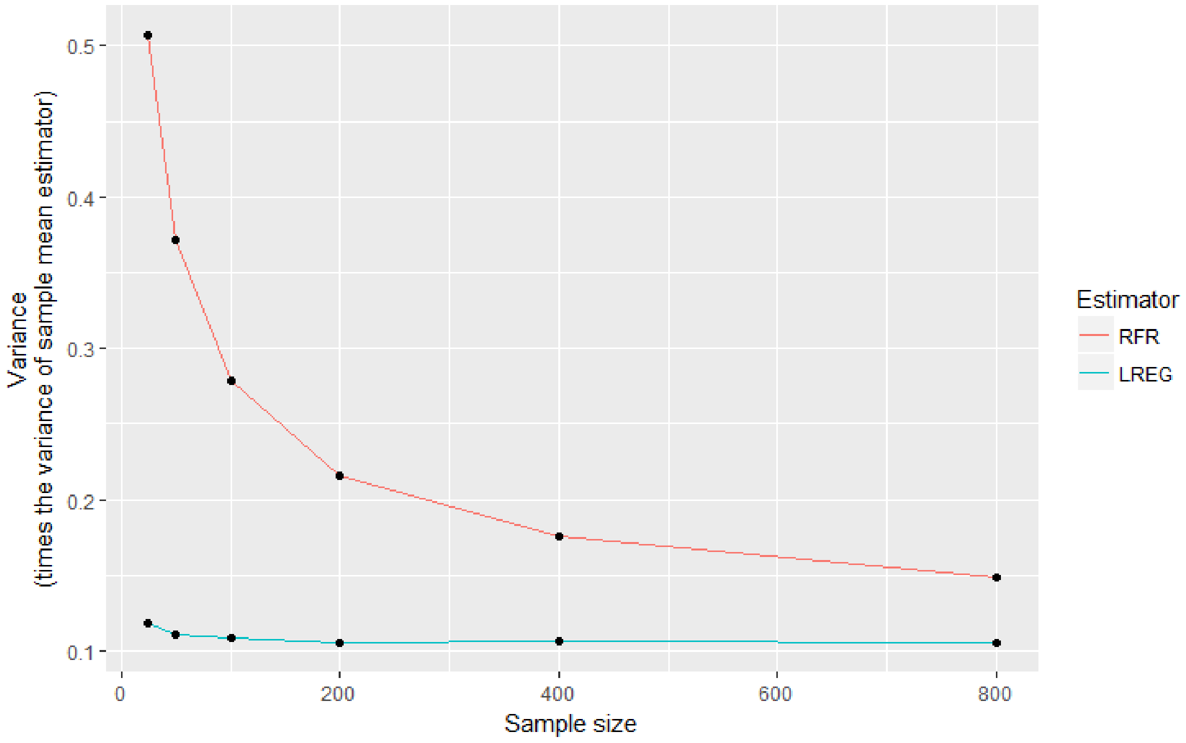
A and B are supposed to be known for the whole population and we have a simple random sample of C. Since the data follows a linear model, the best performance is obtained with the linear regression estimator (see (Cochran, 1977) for example), as obvious. Still, the performance of the RFR estimator is quite good: it has negligible bias and reduces greatly the variance of the sample mean estimator.

Table1. Synthetic data, linear model

Estimator	RFRE	Linear Regression	Sample mean
Bias (estimator)	0.01	0.11	-0.07
Variance (estimator)	27.88	10.87	100
Bias (estimated variance)	-2.92	-3.33	0.59
Variance (estimated variance)	10.97	0.07	100

Several similar simulations were run for different sample sizes. In the following figure we can see how the RFR estimator evolves when you increase the sample size.

Figure 1. Evolution of the variance of the regression estimators for linear model



Note that this variance decrease is not the usual $o(\frac{1}{n})$ rate, since we are dividing by the sample mean variance, thus removing that effect. In fact, we can see that the linear regression estimator variance is an horizontal line. This additional decrease of the variance is because the algorithm “learns” from data. The more the data, the more it learns and the better it models. With enough data its difference from the “real” model would be negligible, which can also be seen in Figure 1. Note also that even if we change the population size, the graph would remain the same.

The second subset is also a linear model, but logarithms have to be taken and we will not do it. It is a simple example of what would happen if the model were not linear (or if the analyst misses the transformation). Its description:

- A and B follow a lognormal with $\mu = 0$ and $\sigma = 1$.
- $C = 3\log(A) - 2\log(B) + e$ where e follows a normal with $\mu = 0$ and $\sigma = 0.1$.

As expected, this time the performance of the linear regression estimator is not that good. Sample mean is an unbiased estimator, so the big number in the bias means that we would need more iterations to reduce it because of the high variance of the variable. Anyway, the bias of the linear regression estimator is considerable. On the other hand, RFR has no appreciable bias and reduces variance a lot more.

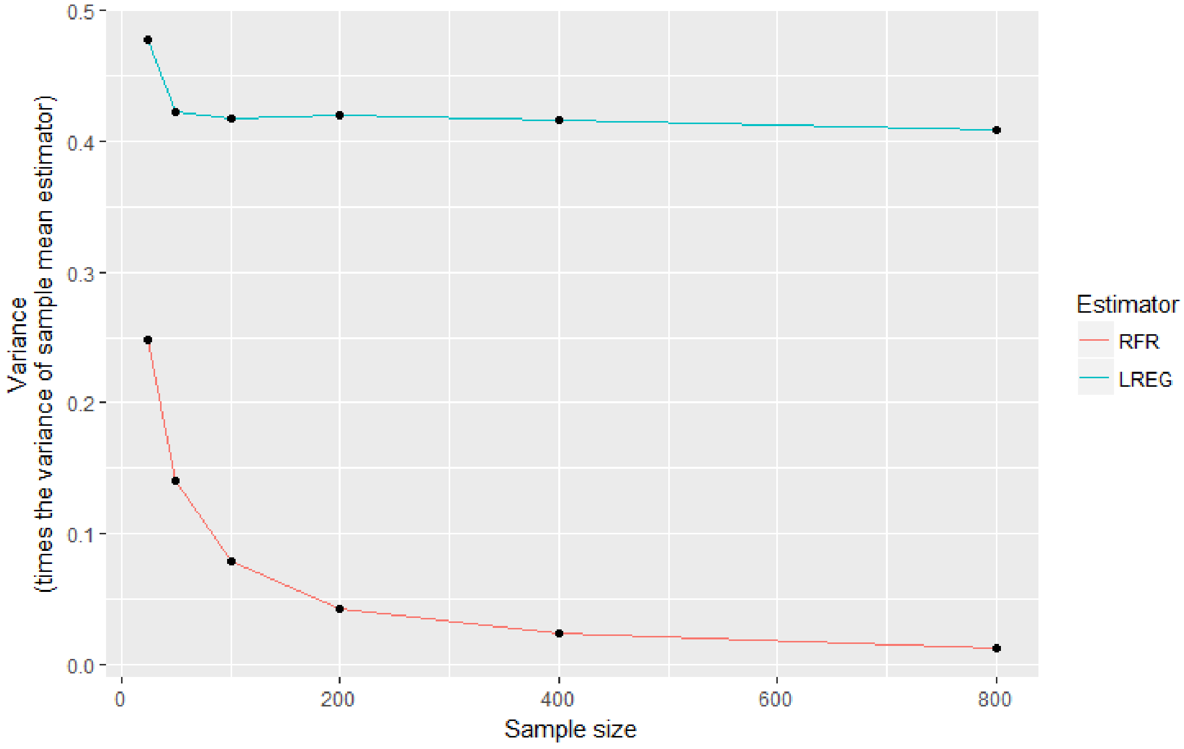
Table2. Synthetic data, non linear model

Estimator	RFRE	Linear Regression	Sample mean
Bias (estimator)	0.10	1292.15	-20.14
Variance (estimator)	7.86	41.70	100
Bias (estimated variance)	-4.86	-20.70	-0.20
Variance (estimated variance)	0.98	44.37	100

The formula (3) is approximate in a way that a small negative bias is always present. That’s why in both simulations the bias of the estimated variance is negative. The use of the unbiased formula would avoid this small bias.

The graphs of the evolution of the variance of the regression estimators are quite similar in shape to those of the previous data set. The main difference is that RFR estimator graph seems to have been “shifted” downwards.

Figure 2. Evolution of the variance of the regression estimators for non linear model



From a machine learning point of view, what happens in Figure 2 (and also in Figure 1), is that linear regression learns faster than Random Forest. This is because linear regression is parametric with few parameters and Random Forest is non parametric. But this also means that Random Forest will keep learning when linear regression gets stuck, what can also be seen in Figure 2.

3.2. Real data with simple random sampling

The population has been built from a sample of the Structural Business Statistics survey and some administrative data available from corporate tax (46 variables). This time the population size is 60948 and 10000 samples of size 6000 are taken. The variable estimated is Total Expenses and there is no correspondent administrative variable. The linear regression estimator uses Turnover (a better model might be found).

The bias continues being inappreciable and the variance improvement in this case is more moderate, but still important. Bias is more problematic for the linear regression estimator unless you are very careful at modeling (we were not).

While it would be possible to obtain a linear model with similar variance (bias could be problematic though) to the RFR estimator, it would represent a lot more work in

Table3. Real data, simple random sampling

Estimator	RFRE	Linear Regression	Sample mean
Bias (estimator)	0.24	2.77	0.32
Variance (estimator)	60.98	78.73	100
Bias (estimated variance)	-0.97	-24.02	1.83
Variance (estimated variance)	82.82	87.76	100

modeling. And if you use GREG, you have to use the same regressors for all variables, so chances are that you would obtain a slightly worse variance.

3.3. Real data with simple stratified sampling

The formulae (2) and (3) both suppose simple stratified sampling, so this example is similar to the previous one but stratified design is used (with minimum variance allocation). The strata are based in the number of employees. The new population size is 55777 and we take 10000 samples of size 5000. This time we estimate Total Personnel Expenses, again without administrative counterpart. Linear regression estimator is not included in the comparison, we had enough with the previous simulations, so we include the correspondent estimators for simple random sampling. In the RFR estimator with simple random sampling we included stratum as a regressor (we can, because Random Forest deals with discrete data).

Table4. Real data, simple stratified sampling and SRS

Estimator	RFRE str	Sample mean str	RFRE	Sample mean
Bias (estimator)	-0.08	0.01	-0.01	-0.04
Variance (estimator)	40.65	86.01	13.60	100
Bias (estimated variance)	-0.89	0.56	-2.05	-0.57
Variance (estimated variance)	2875.59	12300.93	29.88	100

There are two interesting things to note from Table 4. First, while RFR stratified estimator has better accuracy than stratified sample mean, it is beaten by RFR under simple random sampling! This is so, because:

- The allocation is optimal for the target variable, but RFR is based on the error of the model. We would need minimum variance allocation for the error, which is not possible. Proportional allocation would probably be more appropiate.
- Since the strata variable is one of the regressors of the model (and a lot more

regressors are used), we are already including information from the stratification in the model, so little gain in accuracy should be expected anyway.

The second one is the poor performance of the variance estimator in stratified sampling. Even the RFR estimator gets higher variance than sample mean with simple random sampling. The reason is that the sample size is very low for some strata.
Perhaps it is not a good stratification, and a lot more of tests should be done, but a possible conclusion would be that it does not worth it to use sophisticated stratified designs to improve accuracy when you use the RFR estimator. Of course, stratified design is still useful if you want to estimate means or totals in the strata.

4. Conclusions and possible uses

Random Forest Regression estimator is good at improving the accuracy of our estimations when we have auxiliary data to exploit. One of its main advantages is that you don't need to analyze the data, the Random Forest algorithm does it for you, and in a standard way. Thus, it is specially useful when the dependencies with the auxiliary data are expected to be complex and/or the metadata for the auxiliary data are poor. Both problems can be found in administrative data and, specially, in Big Data. As a consequence of the accuracy increment we can reduce the sample size and therefore the response burden.

It was originally intended for SBS surveys as an alternative to a complex and difficult to maintain imputation process. The Random Forest model is also complex, of course, but the analyst is kept safe from this complexity. Also, the RFR estimator provides formulae to remove the bias and to estimate the variance, what may suppose an issue with imputation.

It is important to notice, that there are several slightly different decision trees implementation which would allow us to define alternative versions of the RFR estimator, with some advantages:

- Oblique decision trees would improve accuracy a lot, specially when there are linear dependencies. Probably, the linear regression estimator would no longer outperform perceptibly the RFR estimator.

- Multivalued decision trees would allow us the conservation of linear identities, which is sometimes an issue if you do not use a weighting system based estimator.

A possible alternative use of the estimator, is to estimate under non response. You still can not remove the non response bias, but with enough auxiliary data a good model would reduce it considerably. Moreover, even if the response subpopulation might be quite different from then non response subpopulation, its model error might well be a lot more similar. We expect to develop this approach to non response in future works.

5. References

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Cochran, W. (1977). *Sampling Techniques*, 3rd edn, Wiley, New York.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**: 663–685.

Sanguiao (2018). A design unbiased model assisted estimator for finite populations.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*, Springer, New York.