ESTIMATION OF INTERNET QUALITY MEASUREMENTS FOR BRAZILIAN PUBLIC SCHOOLS FROM A SELF-SELECTED SAMPLE

Marcelo Pitta[1], Thiago Meireles[1], Pedro Luis do Nascimento Silva[2]

meireles@nic.br

[1] Regional Centre for Studies on the Development of the Information Society (Cetic.br) –São Paulo, Brazil.

[2] SCIENCE – Society for the Development of Scientific Research.

The spread and use of the Internet have been changing the way companies, governments, health facilities, schools, and other institutions offer services, products, and establish relationships with their customers—similar to the changes in human relationships. In this new and rapidly changing scenario, the use of the Internet and technology in schools provides an instrument for diversifying content, increasing the use of interactive education methodologies, and has been a major subject of debate among scholars.

Regardless of the importance and effectiveness of using such technologies in schools, the use of the Internet is affected by its quality. In this respect, this paper presents an application to weigh internet quality measurements from a self-selected sample of Brazilian public elementary schools aiming to estimate for the full population of such schools. The measurements are provided by SIMET, a software/firmware developed by CEPTRO.br, a department of NIC.br (Brazilian Network Information Center), to evaluate the quality of Internet connection according to five distinct aspects: upload, download, jitter, latency, and packet loss.

The use of SIMET was part of a government program called Connected Education, introduced by the Brazilian Ministry of Education in 2017. Once a school installs the SIMET meter, automatic collection of internet quality measurements begins. The program is not compulsory, so the resulting data collected by SIMET meters come from a self-selected, non-probability sample.

To obtain estimates of Internet quality for the target population—the entire set of public schools in Brazil—we experimented with methods to estimate pseudo-weights. The main goal was to produce estimates that would represent the population of all public elementary schools in Brazil, based on a large, self-selected

sample of public elementary schools that participated in the Internet Traffic Measurement System (SIMET) program.

Since the population of public elementary schools is known and its characteristics are measured by the Brazilian Annual School Census, the approaches we tried aimed to estimate pseudo-weights that would, within margins of error, reliably reflect the known characteristics collected in the annual school census.

The process of pseudo-weighting estimation followed two steps:

(a) Determine which part of the school's population can be represented by those that have SIMET installed. In this step, detailed below, we found that federal public schools do not participate in the program. These schools, fewer than a thousand in Brazil, have very specific characteristics that set them apart from the rest of the public schools. Since no federal schools provide SIMET data, these schools were therefore excluded from the target population.

(b) Construct pseudo-sampling weights for schools that have SIMET to provide estimates of the quality of internet for the population established in (a).

Different approaches were tested:

- Naïve (consider the available sample as if it was a simple random sample from the set of public schools);
- Randon Forest for the indicator of providing SIMET measurements;
- Lasso regression or the indicator of providing SIMET measurements;
- Random Forest + Logistic regression for the indicator of providing SIMET measurements.

The process of estimation of pseudo-weights suffered from the fact that, in this case, the problem deals with a large database from the annual school census: more than 100,000 records and more than 300 potential predictor variables. To evaluate each of the listed methods, six statistics were calculated. These statistics are presented below:

- Sum of absolute deviations for categorical variables

$$SDA_c = \sum_{a=1}^{A} \sum_{j=1}^{K_a} \left(\left| \hat{p}_j^a - p_j^a \right| \right)$$

- Sum of Relative Absolute Deviations for Categorical Variables

$$SDAR_c = \sum_{a=1}^{A} \sum_{j=1}^{K_a} \left(\frac{\left| \hat{p}_j^a - p_j^a \right|}{p_j^a} \right)$$

where a is a categorical variable existing in the dataset, A is the total of categorical variables in the dataset, j is a category of the categorical variable a existing in the dataset, K_a is the total of categories of the categorical variable a in the dataset, and p_j^a is the proportion of records reporting category j for a variable a in the Census, while $\hat{p}_j^a$ is the corresponding estimate from a weighted sample of available records after applying pseudo-weights obtained by each one of the alternative approaches considered. We also computed summary measures for numeric census variables, namely:

- Sum of Absolute Deviations for Numerical Variables $$SDA_n = \sum_{b=1}^{B} \left| \hat{q}_j^b - q_j^b \right|$$
- Sum of Relative Absolute Deviations for Numerical Variables $$SDAR_n = \sum_{b=1}^{B} \left(\frac{\left| \hat{q}_j^b - q_j^b \right|}{q_j^b} \right)$$

where: b is the numerical variable in the dataset, B is the number of numeric variables in the dataset, and q_j^b is the mean of variable b for all Census records, and $\hat{q}_j^b$ is he corresponding estimate from a weighted sample of available records after applying pseudo-weights obtained by each one of the alternative approaches considered.

Var_c is the number of categorical variables for which the 95% confidence interval for a proportion of one of its categories does not contain the proportion observed in the Census; and

Var_n is the number of numerical variables for which the 95% confidence interval for the mean does not contain the observed Census mean.

These summary statistics were used to compare the estimation performance of each method. Additionally, as required for disseminating the results, the pseudo-weights were calibrated to the known totals of schools by state (27 levels), area (rural/urban), and administrative jurisdiction (municipal, state). Among all the methods tested, the combination of Random Forest and Logistic Regression showed the best results, though some bias was still present in the estimates as shown in table 1 below.

Table 1: Summary statistics for applied methodologies

Strategy → / Statistics ↓	Naïve (SRS)	RF	LASSO	RF+Logistic (5 numeric & 160 categorical)
SDA_c	2052.05	272.77	758.87	93.30
SDAR_c	134.70	28.44	67.00	14.03
SDA_n	1308.40	156.40	337.46	23.98
SDAR_n	60.83	6.61	11.74	1.94
Var_c	142	123	165	47
Var_n	100	75	86	5

Source: Brazilian Schools' Census, SIMET measurements.

The estimation process resulted in a dataset that allows the estimation of Internet quality measurements for all public schools—with Internet and computers. Updating the estimates using the process of data analysis and pseudo-weight estimation is necessary if new schools are included in the SIMET measurement process, as well as in the target population of public schools.

References

Dever, J. A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*, 15.

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2 (6).

Elliott, M. R., & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, 32 (2), 249–264. https://doi.org/10.1214/16-STS598

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data. *Wiley Series in Probability and Statistics*.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9, 11. https://doi.org/https://doi.org/10.1186/1471-2105-9-307

Valliant, R. (2019). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, 8 (2), 231–263. https://doi.org/10.1093/jssam/smz003

Valliant, R., & Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. Sociological Methods & Research, 40 (1), 105–137. https://doi .org/10.1177/0049124110392533