STAT 410 Course Project: Investigating Sampling Techniques on Sulfur Oxide Air Pollutants

Negin Bolkhanian, Helen Jung, Forrest Paton, Rongjuan Wang

April 07, 217

Abstract

The purpose of this paper is to analyze companies operating in Canada and their 2014 Sulfur Oxide emission. Particularly, we estimate the amount of total SOX emissions in Canada. Three different sampling methods were used to identify which estimators performed the best: Simple Random Sampling(SRS), Ratio Estimation, and Stratification. The results show that Ratio estimation and SRS perform better than stratification. Ratio Estimation has the least variance and therefore the smallest mean square error. SRS produces accurate estimates, but the estimate has more variability when compared to ratio estimation.

1 Introduction

Among many dangers, the presence of human-released pollutants in the atmosphere interact to create smog and acid rain.

When air pollutants gather as dense clouds of particulate matter (smog), symptoms such as upper respiratory infections, nausea and eye irritation can occur [1].

Air pollutants can also react with water molecules in the atmosphere to form acid rain. When acid rain falls, chemical properties of soil, streams and lakes are altered. This chemical change has adverse effects on the ecosystem, harming plants, animals, and other wildlife.

Human-released pollutants stem from activities such as burning fuels for energy, heat production, transportation and various industrial processes. Naturally occurring sources of air pollutants also exist, however their contribution is often rare and uncontrollable. [1]

Air Pollutant Emissions indicators track emissions from human-related sources of sulfur oxides (SOX), nitrogen oxides (NOX), volatile organic compounds (VOCs), ammonia (NH3), carbon monoxide (CO) and fine particulate matter (PM2.5) [2]. In this paper we analyze companies operating in Canada and their 2014 SOX emission contributions. Data used is published by Environment and Climate Change Canada and licensed under the government of Canada's Open Data Initiative. We use three different sampling methods to identify which estimators perform the best: Simple Random Sampling, Ratio Estimation, and Stratification.

[1] Sources of naturally occurring air pollutants include volcanic eruptions and emissions of volatile organic compounds from vegetation.

In this project we are interested in the estimator τ, the total SOX emissions from monitored companies in Canada.

2 Data

Our data set contains 168 companies that contributed to sulfur oxide emissions in Canada for 2014 [3]. Along with the measured Sulfur output, the data contains spatial covariates: City, Province, Longitude, Latitude, as well as Nitrogen Oxide emissions. To control for company size/ output we performed a logarithmic transformation on the Sulfur and Nitrogen columns. Figure 1 shows the distribution of Sulfur emissions by province.

3 Methods

3.1 SRS (Simple Random Sampling)

Our first method is a random sample without replacement (SRS). The first step is finding an appropriate sample size. Using n_0 from Thompson (2012):

$$n_0 = \frac{z^2\sigma^2N^2}{d^2}$$

we found an appropriate sample size of 21. Figure 2 is a simulation of mean SOX estimates vs sample size.

We see from our simulation, the sample mean seems to converge to the true mean at a sample size of 50. At sample size equal to 25 the bootstrap mean seems to become random about the true mean. Which provides evidence that our sample size is appropriate.

Our estimator of interest is τ the population total of Sulfur Oxide emissions in Canada.

$$\tau = \sum_{n=1}^{N} y_i$$

and variance:

$$var(\tau) = N(N-n)\frac{\sigma^2}{n}$$

We bootstrapped 1000 samples of size 21, calculating these estimators. Results are discussed in the next section

3.2 Ratio Estimate

For Ratio Estimation, we used the nitrogen oxides measurements (NOX) as our auxiliary variable. First we need to check if there's a positive proportional correlation between NOX and SOX. Assuming a relationship exists (in our case there was, we fitted a linear model: SOX ~ NOX, p <.01) between SOX and NOX the population ratio R is defined to be:

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\tau_y}{\tau_x}$$

The estimated total sulfur oxide emission can be obtained from:

$$\tau_r = N\mu_r = r\tau_x$$

$\hat{\tau_r}$ is approximately unbiased since the sample size is sufficiently large and SRS is used for sampling. We used bootstrapping techniques, with the number of replicates = 1000. We assume that the auxiliary data (NOx) is known. Figure 4 shows the histogram of bootstrapped estimates of total population. Results are shown in the next section.

3.3 Stratified

For our last method, we used stratified sampling. The companies were put into four stratums according to their geographical location: We bootstrapped 1000 samples from the four

Table 1: Stratums

Stratum	Provinces	N_h(Stratum Size)	n_h(Sample size)
Atlantic	NL, PEI, NS, NB	24	2
Central	QC, ON	77	14
Prairie	MB, SK, AB	50	7
West Coast	BC	16	2

stratums according to the design above. n_h was calculated using optimal allocation. I.e. we choose the sample sizes based on the relative stratum size and variance. Our estimator was then the total sum of each stratum. The Northwest Territories was left out since the contribution was not significant to produce meaningful results.

$$\hat{\tau} = Atlantic + Central + Prairie + WestCoast$$

This estimator was run for 1000 samples.

4 Results

Along with the estimates of τ we calculated the mean square error, bias, SE and a 95% confidence interval for our estimator. Results are placed in Table 2. We see that SRS and Ratio Estimation produced the closest estimates to our population total τ. However ratio estimation had the smallest mean square error, bias and confidence interval. Ratio estimation produced the best results which is intuitive since we could reduce the variance of our estimator by including more information. Specifically the auxiliary variable (NOX) was linearly correlated SOX, which led to decreased variance of our estimator τ. SRS and stratification produced very similar results. Our estimator using stratification was slightly biased, we believe this is because of our small n_h (sample size within strata). As some stratums sampled more brought down the estimator.

Table 2: Simulated Sampling results

Model	τ	MSE	$BIAS$	$St.Error$	$95\%CI$
TRUE VALUE	1189.8	-	-	-	-
SRS	1189.1	2985.2	-0.672	54.65	(1079, 1299)
Ratio Estimation	1190.5	627.9	0.67	25.05	(1140, 1238)
Stratified	1184.1	2732	-5.75	51.98	(1080, 1287)

5 Conclusions

Our sampling methods provided interesting insight on an issue that remains important to monitor. We found that estimating total SOX emissions in Canada, a SRS design was effective. This is interesting since SRS is an easy design to implement. If we want more accuracy, ratio estimation provides better results. However, more data is needed to get estimates (NOX auxiliary variable) which can complicate and add extra costs. Auxiliary variables are not always available, which is needed to perform ratio estimation. It's also interesting to note that stratifying by geographical regions did not provide a better estimates than a simple SRS design. More data on the specific companies and other partitioning methods might provide better results. For example some clustering method such as k-means could be run first, then the data could be stratified using those groups.

References

[1] Sulfur Dioxide Basics. (2016, August 16). Retrieved April 02, 2017, from https://www.epa.gov/so2-pollution/sulfur-dioxide-basics#effects

[2] Government of Canada, Environment and Climate Change Canada. (2017, March 10). Environment and Climate Change Canada - Environmental Indicators - Air Pollutant Emissions. Retrieved April 02, 2017, from https://www.ec.gc.ca/indicateurs-indicators/default.asp?lang=en&n=E79F4C12-1

[3] Total particulate matter emissions by facility, Canada, 2014. (n.d.). Retrieved April 02, 2017, from http://maps-cartes.ec.gc.ca/indicators-indicateurs/TableView.aspx?ID=13&lang=en

6 Appendix

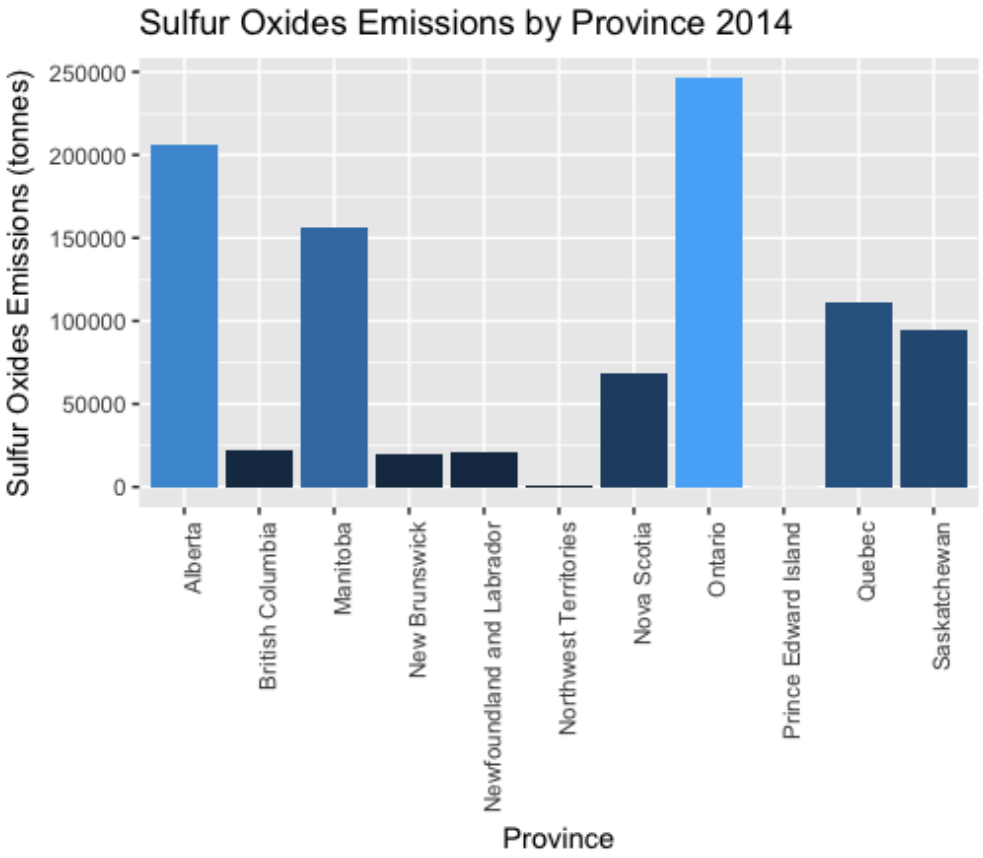

Figure 1: Provincial/ Territorial contributions to Sulfur Emissions

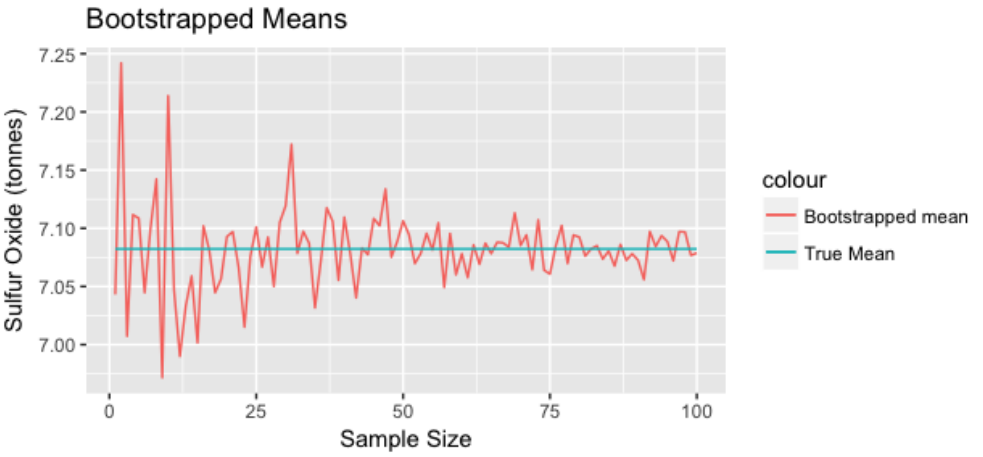

Figure 2: Bootstrapped Means for various sample sizes

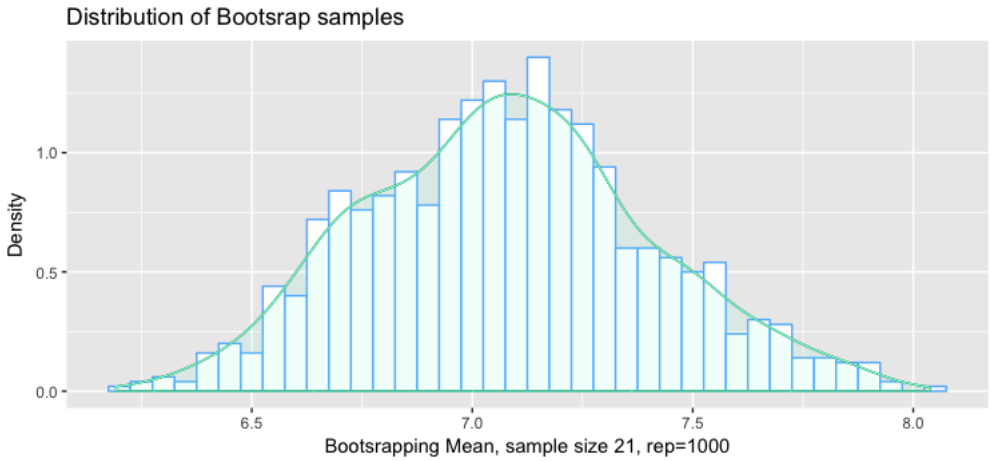

Figure 3: Histogram of bootstrapped mean estimates SRS

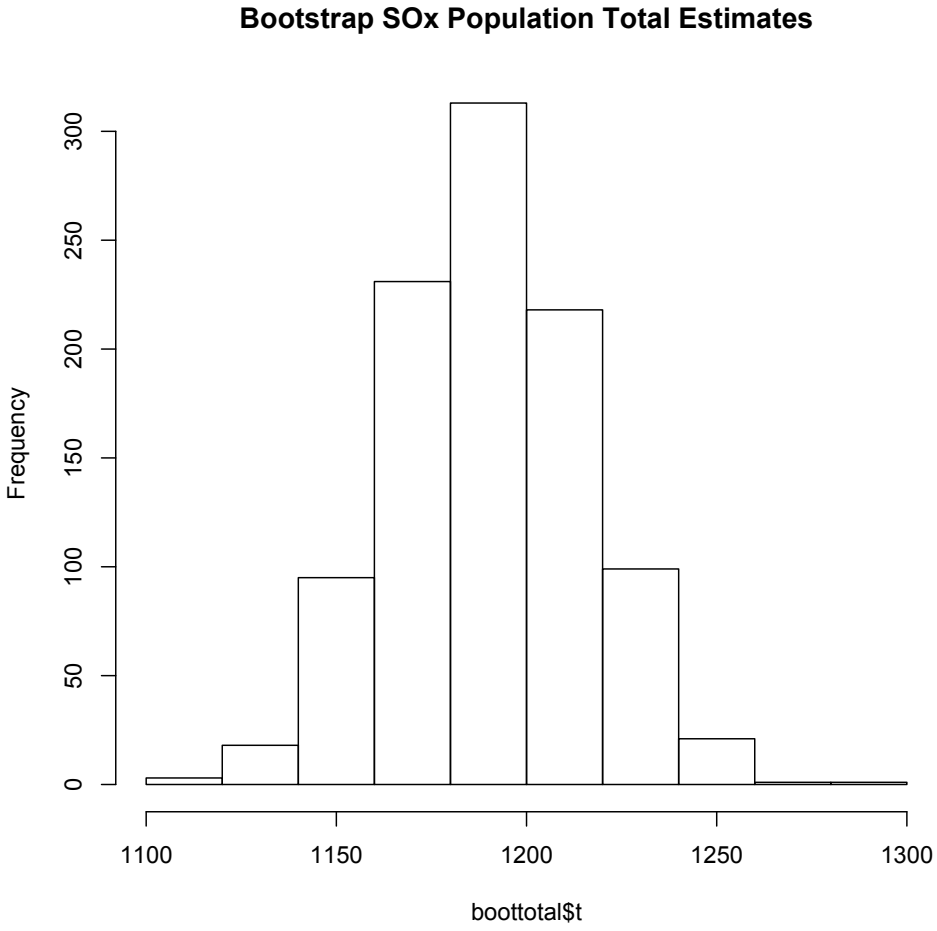

Figure 4: Histogram of bootstrapped total estimates using ratio estimation