SSC Annual Meeting, May 2023

Proceedings of the Survey Methods Section

MEASURING THE ACCURACY OF A PREDICTION FOR A FINITE POPULATION MEAN

Abel Dasylva[1], Jean-François Beaumont[2], Keven Bosa[3] and Guillaume Maranda[4]

ABSTRACT

For timeliness and cost reduction, a statistical agency may predict a mean by modeling future responses based on past responses and fixed covariates, e.g., predicting the yield of crops based on remote sensing and agro-climatic data. Such a prediction may resort to different methodologies, from simple linear predictors to complex predictors based on machine learning techniques, including random forests, boosted trees or deep learning. When doing so, the loss can be measured by the mean square error to compare different predictors or report the prediction accuracy. However estimating the mean square error of a predicted mean is challenging with complex predictors. To address this issue without being tied to a specific prediction methodology, it is proposed to bootstrap the past and future responses based on the residuals while holding the covariates fixed.

KEY WORDS: accuracy, uncertainty quantification

RÉSUMÉ

Pour l'actualité et la réduction des coûts, une agence statistique peut prédire une moyenne en modélisant les réponses futures selon des réponses antérieures et des prédicteurs, dont les valeurs sont fixées, par exemple en prédisant le rendement des cultures à l'aide de données de télédétection et d'ordre agro-climatique. Une telle prédiction peut faire appel à diverses méthodologies, des prédicteurs linéaires simples jusqu'aux prédicteurs complexes, basés sur des techniques d'apprentissage automatique, dont des forêts aléatoires, des arbres de décision optimisés ou l'apprentissage profond. Dans ce cas, on peut mesurer la perte par l'erreur quadratique moyenne afin de comparer différents prédicteurs ou rapporter l'exactitude de la prédiction. L'estimation de l'erreur quadratique moyenne d'une moyenne prédite est cependant difficile avec des prédicteurs complexes. Pour répondre à ce problème sans se contraindre à une méthodologie de prédiction spécifique, nous proposons un bootstrap des réponses passées et futures basé sur les résidus, tout en fixant les valeurs des prédicteurs.

MOTS CLÉS : exactitude, mesure de l'incertitude

1 INTRODUCTION

For timeliness, a statistical agency may predict a total by modeling future responses based on past responses and fixed covariates. For example, Statistics Canada predicts the yield of crops based on remote sensing and agro-climatic data (Statistics Canada 2020). Such predictions must be published with a measure of their accuracy or uncertainty for the data users, even when they are based on sophisticated machine

[1] Abel Dasylva, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, abel.dasylva@statcan.gc.ca

[2] Jean-François Beaumont, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, jean-francois.beaumont@statcan.gc.ca

[3] Keven Bosa, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, keven.bosa@statcan.gc.ca

[4] Guillaume Maranda, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, guillaume.maranda@statcan.gc.ca

learning techniques. However, cross-validation (Hastie et al. 2001, chap. 7.10) and other similar approaches (Lei et al. 2018) are inadequate for two reasons. The first reason is that they assume that the covariates are random, while data users have come to expect a measure of the uncertainty that is conditional on the observed covariates. The second reason is that they focus on the uncertainty of a single prediction. To address the problem without being tied to a specific prediction methodology, it is proposed to bootstrap the past and future responses based on the residuals while holding the covariates fixed. The remaining sections describe the notation and assumptions, methodology, simulations and conclusions, in this order.

2 NOTATION AND ASSUMPTIONS

We consider two finite populations, where each unit is associated with fixed covariates and a continuous response and the responses from the different units are mutually independent within each population and across the two populations. For $t = 1, 2$, population t has N_t units, where unit i is associated with the *fixed* covariates $\boldsymbol{x}_{ti}$, the response Y_{ti} and the weight a_{ti} such that $\sum_{i=1}^{N_t} a_{ti} = 1$. When predicting the yield of a crop, a unit is a field and the weight is proportional to the field area. Let $\mu(.)$ and $\sigma^2(.)$ denote the common mean and variance functions for both populations, i.e., $\mu\left(\boldsymbol{x}_{ti}\right) = E\left[Y_{ti}\right]$ and $\sigma^2\left(\boldsymbol{x}_{ti}\right) = var\left(Y_{ti}\right)$ for $t = 1, 2$ and $i = 1, \ldots, N_t$. The responses are supposed to be of the form

$$Y_{ti} = \mu\left(\boldsymbol{x}_{ti}\right) + \sigma\left(\boldsymbol{x}_{ti}\right)\epsilon_{ti}, \tag{1}$$

where the errors $\epsilon_{11}, \ldots, \epsilon_{1N_1}, \epsilon_{21}, \ldots, \epsilon_{2N_2}$ are independent and identically distributed with a zero mean and a unit variance.

In the first population, the covariates and responses are observed. In the second population, only the covariates are observed. Using the data from the first population, the mean function is estimated by $\widehat{\mu}(.)$ and the mean $\overline{Y}_2 = a_{21}Y_{21} + \ldots + a_{2N_2}Y_{2N_2}$ is predicted by $\widehat{\overline{Y}}_2 = a_{21}\widehat{\mu}(\boldsymbol{x}_{21}) + \ldots + a_{2N_2}\widehat{\mu}(\boldsymbol{x}_{2N_2})$. Our goal is to evaluate the mean square error of this prediction, i.e., $E\left[\left(\widehat{\overline{Y}}_2 - \overline{Y}_2\right)^2\right]$.

3 METHODOLOGY

It is proposed to generate bootstrap responses in each population, while all the covariates are held fixed. Let $\left[Y_{ti}^{(b)}\right]_{1\leq i\leq N_t}$ denote the bootstrapped responses for population t in repetition b. Also let $\overline{Y}_2^{(b)}$ and $\widehat{\overline{Y}}_2^{(b)}$ denote the actual mean and the predicted mean for the second population in this repetition. Then, the mean square error may be estimated by

$$\widehat{E}\left[\left(\widehat{\overline{Y}}_2 - \overline{Y}_2\right)^2\right] = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\overline{Y}}_2^{(b)} - \overline{Y}_2^{(b)}\right)^2, \tag{2}$$

where the bootstrap mean is $\overline{Y}_2^{(b)} = a_{21}Y_{21}^{(b)} + \ldots + a_{2N_2}Y_{2N_2}^{(b)}$. The above estimator may be computed because the actual mean of the second population is known in each bootstrap repetition. When the variance function is known and the errors are known to have a standard normal distribution, the bootstrapped responses may be generated according to

$$Y_{ti}^{(b)} = \widehat{\mu}(\boldsymbol{x}_{ti}) + \sigma(\boldsymbol{x}_{ti})\epsilon_{ti}^{(b)},\ t = 1, 2,\ i = 1, \ldots, N_t, \tag{3}$$

where $\epsilon_{ti}^{(b)}$ is drawn according to the standard normal distribution. When the error distribution is unknown, the bootstrap errors may be based on resampling the residuals $(Y_{11} - \widehat{\mu}(\boldsymbol{x}_{11}))/\sigma\left(\boldsymbol{x}_{11}\right), \ldots, (Y_{1N_1} - \widehat{\mu}(\boldsymbol{x}_{1N_1}))/\sigma\left(\boldsymbol{x}_{1N_1}\right)$ (Efron 1979). The prediction model is applied to the bootstrap responses from the first population to obtain the bootstrap estimate $\widehat{\mu}^{(b)}(.)$ of the mean function and the predicted mean $\widehat{\overline{Y}}_2^{(b)} = a_{21}\widehat{\mu}^{(b)}(\boldsymbol{x}_{21}) + \ldots + a_{2N_2}\widehat{\mu}^{(b)}(\boldsymbol{x}_{2N_2})$.

4 SIMULATIONS

Setup: The simulations are based on crop yield data for spring wheat in Alberta, Manitoba and Saskatchevan, from 2001 to 2018, including $N_2 = 35$ observations in 2018 and $N_1 = 584$ observations in prior years. The data includes five covariates about the vegetation and the temperature. Three scenarios are considered where the covariates are based on the actual covariates and the responses are generated according to Eq. 1. The simulations are based on 1,000 repetitions and 1,000 bootstrap samples in each repetition. In the first scenario, the mean function is linear (i.e., $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$), the variance is constant, and $\boldsymbol{\beta}$ and σ^2 are chosen by fitting a linear model with homoschedastic variance estimated to the actual data. The prediction model is linear, and the bootstrap procedure is based on a linear model or a random forest, where the latter is built with the R package Ranger (Wright & Ziegler 2017) using 2 or 5 splitting variables (i.e., mtry=2,5 in Ranger) and the default number of trees, which is 500. When fitting the random forest, the minimum node size (i.e., min.node.size in Ranger) is selected through a search on a grid, which includes all the values between 5 and 70 by increment of 5 as well as the values 100, 150, 200 and 300. In the second scenario, the mean function is based on fitting a random forest with 2 splitting variables to the actual data (including the responses). The prediction model is based on a random forest, and the bootstrap procedure is based on a linear model or a random forest that is parametrized as in the first scenario. The third scenario is identical to the second scenario, except that the prediction model is linear. In all the scenarios, the variance is constant and known and ϵ_{ti} has a standard normal distribution. The performance of the estimated mean square error of the predictor is evaluated by the relative difference when compared to its Monte Carlo mean square error.

Results: The results appear in Figs. 1-3, where the yellow dot indicates the relative bias of the estimated mean square error.

In the first scenario, the linear bootstrap estimates the mean square error with a small variance and a moderate relative bias (-8.8%), while the random forest bootstrap estimates the mean square error with a large bias and a large variance. In the latter case, the estimator has a better performance when using 5 instead of 2 splitting variables, even if the bias and variance remain large. These results may indicate that there are too few data points.

In the second scenario, the random forest bootstrap estimates the mean square error with a small variance and a small relative bias (5.5%), when using two splitting variables. However, with five splitting variables, the variance is large as well as the relative bias (128.4%). With the linear bootstrap, the variance is large as well as the relative bias (345.6%).

In the third scenario, the random forest bootstrap has a small relative bias (-5.2%) with two splitting variables, but a large relative bias (37.6%) with five splitting variables. In both cases, the variance is large and there are extreme values, which explains why the yellow dot is outside the box in the plot when using five splitting variables. As for the linear bootstrap, it has a large relative bias (-30.3%) but a small variance.

Overall, the best performance is obtained when the bootstrap model reflects the actual distribution of the responses. With the random forest bootstrap, the variance may be large.

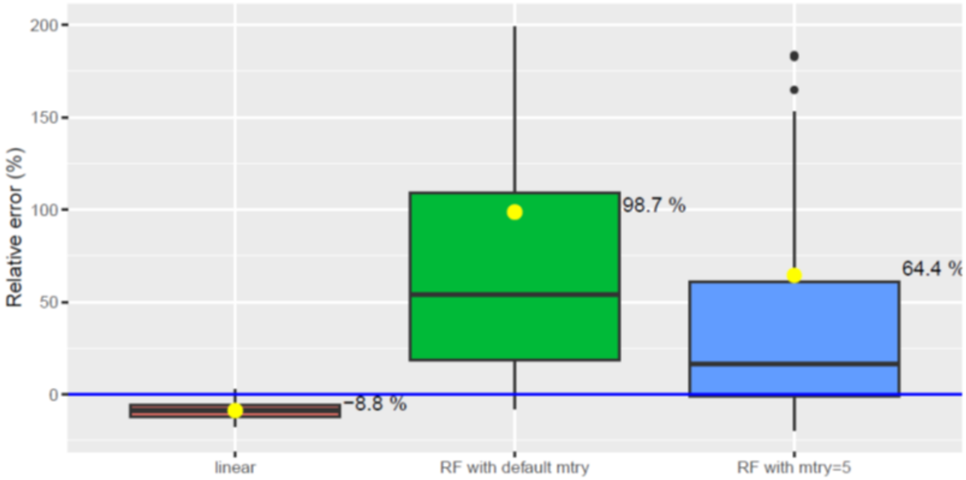

Figure 1: Box plots for the first scenario.

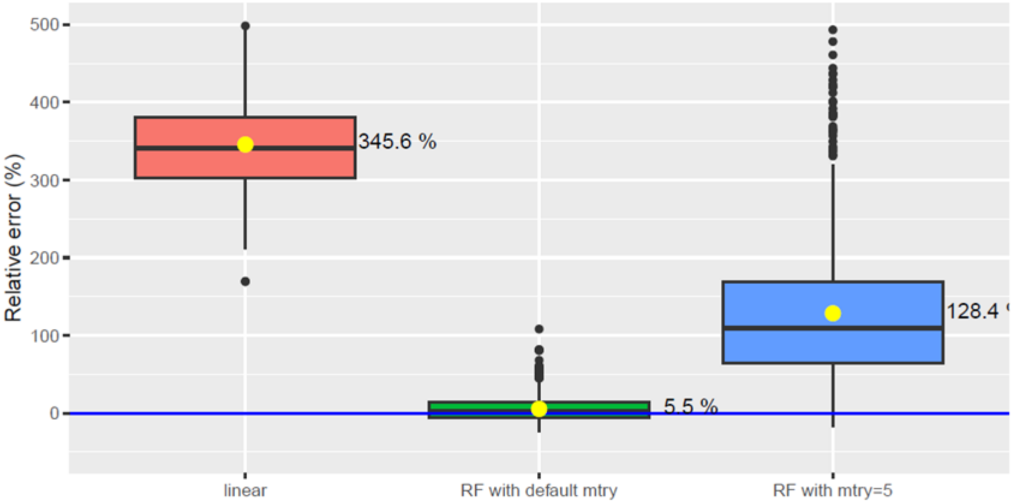

Figure 2: Box plots for the second scenario.

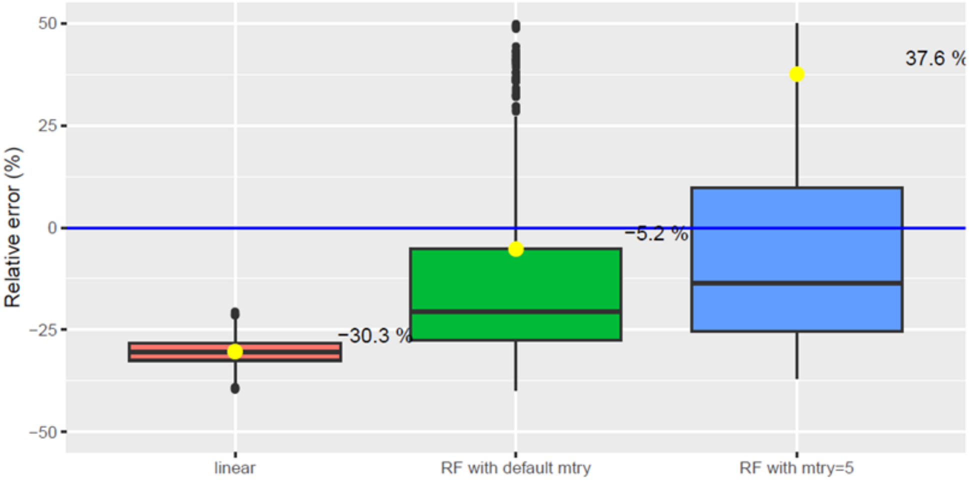

Figure 3: Box plots for the third scenario.

5 CONCLUSIONS AND FUTURE WORK

A bootstrap methodology was proposed to estimate the mean square error of a predicted mean when conditioning on the observed covariates. In the simulations, the resulting estimator performs better when the bootstrap and population models are close. Also, when the bootstrap is based on a random forest, the variance may be large, which may be an indication that there are too few data points. Future work will experiment with other nonlinear mean functions and with larger datasets.

6 DISCLAIMER

The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that may not reflect those currently implemented by the Agency.

REFERENCES

Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *Annals of statistics* **7**, 1–26.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The elements of statistical learning*, Springer, New York.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. & Wasserman, L. (2018), 'Distribution-free predictive inference for regression', *Journal of the American Statistical Association* **113**, 1094–1111.

Statistics Canada (2020), 'An integrated crop yield model using remote sensing, agroclimatic data and crop insurance data', `https://www.statcan.gc.ca/en/statistical-programs/document/3401_D2_V1`. Accessed: 2023-04-01.

Wright, M. & Ziegler, A. (2017), 'ranger: A fast implementation of random forests for high dimensional data in C++ and R', *Journal of Statistical Software* **77**(1), 1–17.