WE PREDICT CONFLICT BETTER THEN WE THOUGHT! TAKING TIME SERIOUSLY WHEN EVALUATING PREDICTIONS IN BINARY-TIME-SERIES-CROSS-SECTION-DATA

GÖKHAN ÇİFLİKLİ

London School of Economics and Political Science

NILS W. METTERNICH

University College London

Abstract. Efforts to predict civil war onset, its duration, and subsequent peace have dramatically increased. Nonetheless, by standard classification metrics the discipline seems to make little progress. Some remedy is promised by particular cross-validation strategies and machine learning tools, which increase accuracy rates substantively. However, in this research note we provide convincing evidence that the predictive performance of conflict models has been much better than previously assessed. We demonstrate that standard classification metrics for binary outcome data are prone to underestimate model performance in a binary-time-series-cross-section context. We argue for temporal residual based metrics to evaluate cross-validation efforts in binary-time-series-cross-section and test these in Monte Carlo experiments and existing empirical studies.

1. Introduction

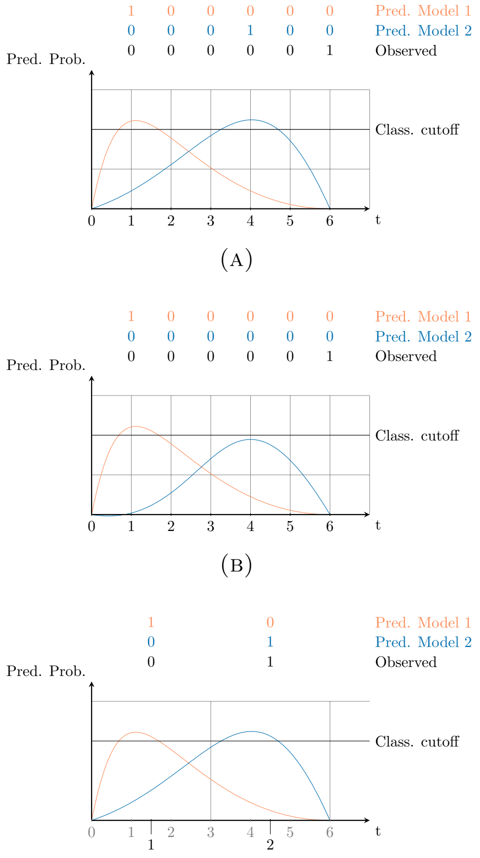
The prediction of conflict and peace dynamics has become a centerpiece of academic output in the field of international relations and conflict studies (e.g. Gurr and Lichbach, 1986; O'brien, 2002; Schrodt, 2006; Goldstone et al., 2010; Weidmann and Ward, 2010; Schneider, Gleditsch and Carey, 2011; De Mesquita, 2011; Ward et al., 2013; Gleditsch and Ward, 2013; Hegre et al., 2013; Bell et al., 2013; Brandt, Freeman and Schrodt, 2014; Muchlinski et al., 2016). Prediction has received this attention because the discipline is increasingly a) emphasizing predictive performance over p-value statistics, b) relying on robust models with external validity, and c) responding to policy makers' demands for meaningful forecasts. We contribute to this literature by demonstrating that standard prediction performance measures have underestimated the sensitivity (true-positive rate) of conflict models in the context of Binary-Time-Series-Cross-Section data.

2. Prediction in Binary Time Series Cross Section Data

Our main argument is that prediction metrics applied to BTSCS data need to embrace the established insight that BTSCS models are a special case of duration models (Beck, Katz and Tucker, 1998). BTSCS data are simply a form of time-series-cross-section data with a binary dependent variable (e.g. conflict onset; termination) instead of continuous outcome. Binary outcome variables are common, especially in International Relations (Beck, Katz and Tucker, 1998). While predictive performance of BTSCS models is usually assessed by identifying correctly predicted years, months, or weeks, we argue that predictive performance needs to account not only for cross-sectional, but also temporal prediction residuals. To illustrate the need to focus on temporal residuals, we provide three interlinked conceptual problems of ignoring the time dimension when evaluating the predictive performance of BTSCS models.

2.1. Temporal Residual Problem. In binary outcome models classification residuals are frequently treated binary: the outcome is either correctly predicted (1) or not (0) depending on whether predicted values or below or above a classification cutoff (e.g. 0.5). Brier scores that calculate the average distance between predicted values and the true outcome and visualization such as separation plots (Greenhill, Ward and Sacks, 2011) try to alleviate this problem, but they cannot account for residuals in time. That implies that a positive prediction shortly before an event is realized is treated similar to a positive prediction further away from the actual realization.

To illustrate this, consider two models predicting the onset of war in a particular country. Figure 1a visualizes such an instance where time is displayed on the x-axis and predicted probabilities on the y-axis. The orange model and the blue model make predictions for six time periods and actual conflict is observed in the sixth time period. Given a pre-defined classification threshold the orange model predicts conflict in the first time period, while the blue model crosses the threshold in the fifth time period. According to standard classification metrics, the orange and the blue model in Figure 1a perform identically: both models have 0 sensitivity (true-positives) (0/1) and 0.8 (4/5) specificity (true-negatives). But when taking into consideration how far off in time the model predictions are, we would probably prefer the blue model (one year off) over the orange model (five years off).



(A)

(B)

(C)

FIGURE 1. (A) Temporal Residual Problem: Blue and orange model have same sensitivity and specificity, despite the blue model's ability to predict a positive event closer to actual realization. (B) Global Threshold Problem: Blue model is preferred according to specificity. However, blue model should not be preferred based on specificity, but because highest prediction is closer to actual realization. (C) Modifiable Temporal Unit Problem: Panel similar to (A) except for time resolution. Simply modifying the temporal unit creates a perfect model.

2.2. Global Threshold Problem. Second, the evaluation of prediction performance in binary classification hinges on a global classification cutoff. In cross-sectional data this is a sound assumption, because observations are not nested and only a global cutoff provides a meaningful quantity. However, in BTSCS data we have a clear nesting structure (e.g. Ward et al., 2013). This nesting structure allows us to calculate not only global cutoff criteria, but also unit specific cutoffs. Figure 1b provides a visualization of the global threshold problem. As in the previous example, imagine two models predicting the onset of war with the difference that the blue model is slightly below the classification threshold. According to the classification threshold the orange model should be preferred over the blue model. However, this global classification ignores the important information that the predicted probability of the blue model peaks just before the actual onset.

2.3. Modifiable Temporal Unit Problem. BTSCS data can be interpreted as discrete duration models. This implies that duration is not measured continuously, but by a temporal unit due to the resolution of the data. Thus, the temporal unit is to some degree arbitrary, which raises a modifiable *temporal* unit problem similar to the more widely-known modifiable areal unit problem (Openshaw, 1984) in spatial analysis. The core of the problem can easily be proven when decreasing the limit of the temporal unit approaches $\lim_{t \to 0}$. As the temporal unit gets smaller, it becomes exceedingly difficult to predict an event at an exact point in time. While there might be a 'sweet-spot' unit to predict wars, it becomes clear that it is probably harder to predict the onset of war at very high time resolutions.

Standard binary classification metrics were developed for cross-sectional data and therefore ignore the modifiable temporal unit problem in BTSCS data. To further illustrate this point consider Figure 1c which only differs from Figure 1a by aggregating the time scale. Instead of six time periods, we now observe units in two time periods. The models still make the same underlying predictions, but because the time scale has changes the blue model's predictions are suddenly 'perfect'. Hence, just by changing the time scale the blue model sensitivity and specificity shifted from 0 to 1 and 0.8 to 1, respectively.

We suggest that the solution to the temporal residual problem and the modifiable temporal unit problem in BTSCS classification is to implement classification metrics that take into account the continuous temporal distance from prediction to observed outcome of interest. We also propose unit specific or local classification thresholds to tackle the outlined global threshold problem.

3. Residual based, Local Assessment of Predictive Performance in BTSCS data (RLC)

We propose to assess BTSCS prediction models' sensitivity through residual based, local assessment of classification performance. We outline such an approach, which compares predictive model performance taking into account temporal residuals.[1]

Thus to assess the temporal dimension of predictive performance, we first generate predicted values for each observation $\hat{y}_{it}$, where i is a unit that is observed over time t and $f(.)$ is a function

[1] This approach also allows for the comparison of BTSCS models on different time scales. For example, if a monthly model provides predictions that are on average, e.g., 11 months away from the actual onset, it can be compared to a yearly model that makes an average prediction error of 13 months.

or algorithm mapping predictors into the outcome variable y:

$$\hat{y}_{it} = f(X\beta + f(t)) \tag{1}$$

Our residual performance criterium is local as we normalize each units' prediction by its maximum prediction such that $max(q_i) = 1$:

$$q_{it} = \frac{\hat{y}_{it}}{max(\hat{y}_i)} \tag{2}$$

Finally, we calculate for each unit that has experienced the outcome of interest, how many time units the highest normalized prediction of i ($q_{it} = 1$) is from the actual realization of $Y_{it} = 1$. The residual based, local assessment of classification performance provides the mean value of this residual.

$$RLC = \frac{T_{(Y_{it}=1)} - T_{(q_{it}=1)}}{\sum Y_{it} = 1} \quad \text{if } Y_{i.} = 1 \tag{3}$$

Thus, we arrive at a measure indicating how many time units on average the highest prediction of a unit i is away from the actual realization.

4. Monte Carlo Study

In this section, we demonstrate that standard classification approaches ignore temporal residuals in BTSCS models. Different to standard Monte Carlo studies, our focus is on the model's ability to predict the outcome of interest (Y) and not the recovery of point estimates (e.g. β's or σ's). We compare different cross-validation approaches that have been implemented and known to effectively increase accuracy rates in BTSCS data (Muchlinski et al., 2016). The data generating process underlying the Monte Carlo simulations is provided by Hendry (2014), proxying a Cox-proportional hazard data generating process (with a squared function of time[2]) with time-varying covariates.

Algorithm 1: Summary of Monte Carlo study to compare prediction evaluation metrics.

```
Simulate Cox proportional hazard data (squared function of time) with algorithm provided by
  Hendry (2014); n=100, 500, 1000 (average t=17.2 → average sample size= 1720, 8600, and
  17200)
For each n generate i=1000 data sets
for data set $i = 1000$ do
    for cross-validation approach in k; k=10-Fold, 10-Fold No Time, 10-Fold Downsampled,
      Forward Rolling Origin do
        for model in m; m=logit,random forest do
            Train $m$ model on data set $i$ using cross-validation approach $k$
            Test logit and random forest model on $i + 1$ (for train i=1000 test i=1)
            Calculate Sensitivity, Specificity, Accuracy, and RLC
        end
    end
end
```

[2] see results for cubic DGP in the Appendix

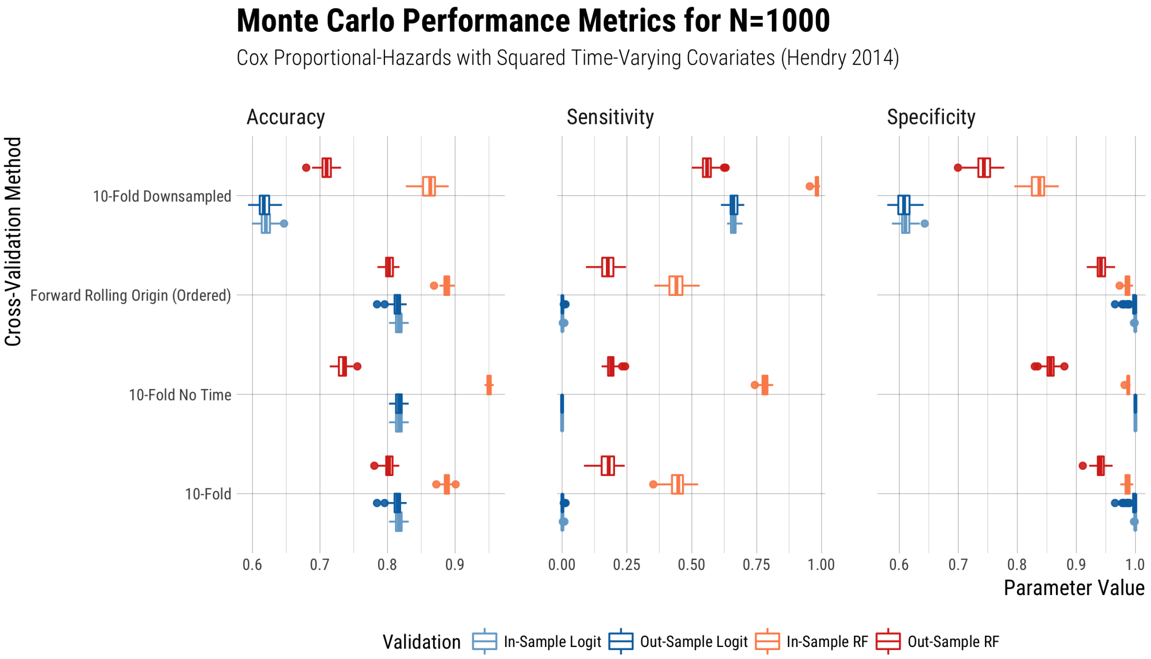


Figure 2. Monte Carlo simulation metrics for N=1000 with Logit and Random Forest models. From left to right, panels shows model accuracy, sensitivity, and specificity. Monte Carlo simulations demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well in regards to sensitivity.

We conduct Monte Carlo experiments for two different non-monotonic hazards (squared and cubic functions of time) and run 1000 Monte Carlo iterations with 100, 500, and 1000 unique units i. On average each i is observed for $t = 17.2$ leading to an average sample size of 1720, 8600, and 17200, respectively. Ten percent of the observations are right-censored and their realizations are not observed in the dataset. Please refer to Algorithm 1 for an overview of the monte carlo study.

We implement two predictive models for the purpose of this study. First, logistic regression modela with the following model specification, $Y = f(X\beta + t + t^2 + t^3)$. Second, random forest models which are gaining popularity in conflict studies (Muchlinski et al., 2016). Here we set the input variables to: $Y = f(X\beta + t + t^2 + t^3)$ while growing 25 trees that randomly selects three variables to split on.[3]

Time-dependent data can challenge sample independence assumptions in cross-validation approaches (Arlot and Celisse, 2010). Evidence in larger samples implies that standard CV approaches are appropriate, but we also implement forward-rolling origin CV by establishing an 'origin' at which the forecast is based which then rolls forward in time (Hyndman and Athanasopoulos, 2014). In Figure 2, we present results for 10-fold CV, 10-fold CV that excludes time variables (t, t^2, t^3), and

[3] We utilize the `ranger` package (Wright and Ziegler, 2017), a fast `C++` implemention ported to `R`, via the `caret::train` framework.

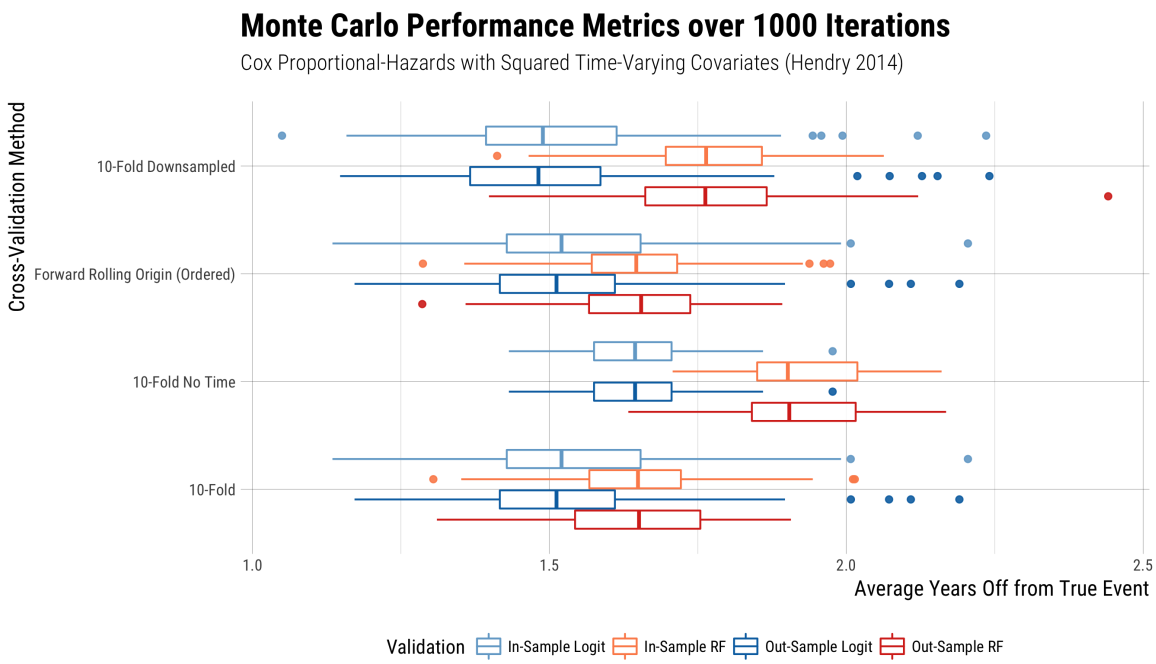


Figure 3. Monte Carlo simulation RLC metrics for N=1000 with Logit and Random Forest models. Panel shows that different to the models' sensitivity. The RLC demonstrates that predictions are fairly close to the true outcome, and that variation between CV methods and estimation methods are less pronounced.

forward rolling origin CV. In addition, we take into account the inherent class-imbalance in BTSC data by 10-fold down-sampling the dependent variable as suggested by (Muchlinski et al., 2016).

Figure 2 displays results of the Monte Carlo experiment with N=1000. On the y-axis, we find the different cross-validation approaches and the respective in- and out-sample performance of logit and random forest models. We provide all other results in the Appendix. Except for the the downsampled data (compare Muchlinski et al., 2016), we find the typical pattern in BTCSC models with high specificity and low sensitivity.

However, when looking at the RLC of all models in Figure 3, we actually find smaller differences between the different CV methods and also fairly precise predictions about the time until an event occurs. This motivates us to consider existing empirical studies on conflict onset and duration to assess the degree to which standard metrics have underestimated the prediction performance in this context.

5. Predicting Conflict

We replicate and evaluate three influential studies in the civil war literature for onset (Fearon and Laitin, 2003; Hegre and Sambanis, 2006)[4] and duration (Wucherpfennig et al., 2012). In line with the Monte-Carlo study, all three studies display similar tendencies representative of BTSCS

[4] Data and model specification as provided by replication material from (Muchlinski et al., 2016)

data: high specificity and very low sensitivity as shown in Figure A11. If we assess the studies using standard evaluation methods, we find that their outsample sensitivity (truepositive rate) are between 0.13 and 0.27 with some models (especially downsampled logit model) above 0.5. However in Figure A12, the RLC metrics demonstrate that models' ability to predict the time until conflict onset and termination is much closer to the observed outcome than the model sensitivity would suggest. Additionally, large differences in sensitivity do not translate into reducing the overall RLC score. In fact, despite the lower sensitivity of some random forest models compared to the downsampled logit models, in the Hegre and Sambanis (2006) and Wucherpfennig et al. (2012) models their prediction are only off by about two years. In the best outsample performance, the random forest algorithm using 10-fold cross-validation is only 1.69 years off on average from predicting the true event.

6. Conclusion

We show that the standard performance metrics originated from cross-sectional data ignore temporal residuals in BTSCS applications. We identify three potential pitfalls for why this is the case. Implementing prediction metrics that account for temporal residuals, such as our proposed RLC approach, reveal that standard sensitivity metrics are prone to underestimate prediction performance in BTSCS data. We hope that this study motivates further work on prediction evaluation that takes into account temporal, but also spatial residuals in binary, count, and continuous outcome models.

References

Arlot, Sylvain and Alain Celisse. 2010. "A survey of cross-validation procedures for model selection." *Statistics surveys* 4:40–79.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable." *American Journal of Political Science* 42(2):1260–1288.

Bell, Sam R, David Cingranelli, Amanda Murdie and Alper Caglayan. 2013. "Coercion, capacity, and coordination: Predictors of political violence." *Conflict Management and Peace Science* 30(3):240–262.

Brandt, Patrick T, John R Freeman and Philip A Schrodt. 2014. "Evaluating forecasts of political conflict dynamics." *International Journal of Forecasting* 30(4):944–962.

De Mesquita, Bruce Bueno. 2011. "A New Model for Predicting Policy Choices Preliminary Tests." *Conflict Management and Peace Science* 28(1):65–87.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review* 97(1):75–90.

Gleditsch, Kristian Skrede and Michael D Ward. 2013. "Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes." *Journal of Peace Research* 50(1):17–31.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder and Mark Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 54(1):190–208.

Greenhill, Brian D., Michael D. Ward and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." *American Journal of Political Science* 55(4):991–1003.

Gurr, Ted Robert and Mark Irving Lichbach. 1986. "Forecasting Internal Conflict A Competitive Evaluation of Empirical Theories." *Comparative Political Studies* 19(1):3–38.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand and Henrik Urdal. 2013. "Predicting Armed Conflict, 2010–2050." *International Studies Quarterly* 57(2):250–270.

Hegre, Håvard and Nicholas Sambanis. 2006. "Sensitivity Analysis of Empirical Results on Civil War Onset." *Journal of Conflict Resolution* 50(4):508–535.

Hendry, David J. 2014. "Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers." *Statistics in Medicine* 33(3):436–454.

Hyndman, Rob J and George Athanasopoulos. 2014. *Forecasting: principles and practice*. Melbourne: OTexts.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24(1):87–103.

O'brien, Sean P. 2002. "Anticipating the good, the bad, and the ugly an early warning approach to conflict and instability analysis." *Journal of Conflict Resolution* 46(6):791–811.

Openshaw, Stan. 1984. The modifiable areal unit problem. In *Geo Abstracts University of East Anglia*.

Schneider, Gerald, Nils Petter Gleditsch and Sabine Carey. 2011. "Forecasting in International Relations: One Quest, Three Approaches." *Conflict Management and Peace Science* 28(1):5–14.

Schrodt, Philip A. 2006. Forecasting conflict in the Balkans using hidden Markov models. In *Programming for Peace*. Springer pp. 161–184.

Ward, Michael D., Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz and Simon Weschle. 2013. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction." *International Studies Review* 15(4):473–490.

Weidmann, Nils B. and Michael D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54(6):883–901.

Wright, Marvin and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software, Articles* 77(1):1–17.

Wucherpfennig, Julian, Nils W. Metternich, Lars-Erik Cederman and Kristian Skrede Gleditsch. 2012. "Ethnicity, the state, and the duration of civil war." *World Politics* 64(1):79–115.

Appendix for 'We predict conflict better then we thought! Taking time seriously when evaluating predictions in Binary-Time-Series-Cross-Section-Data'

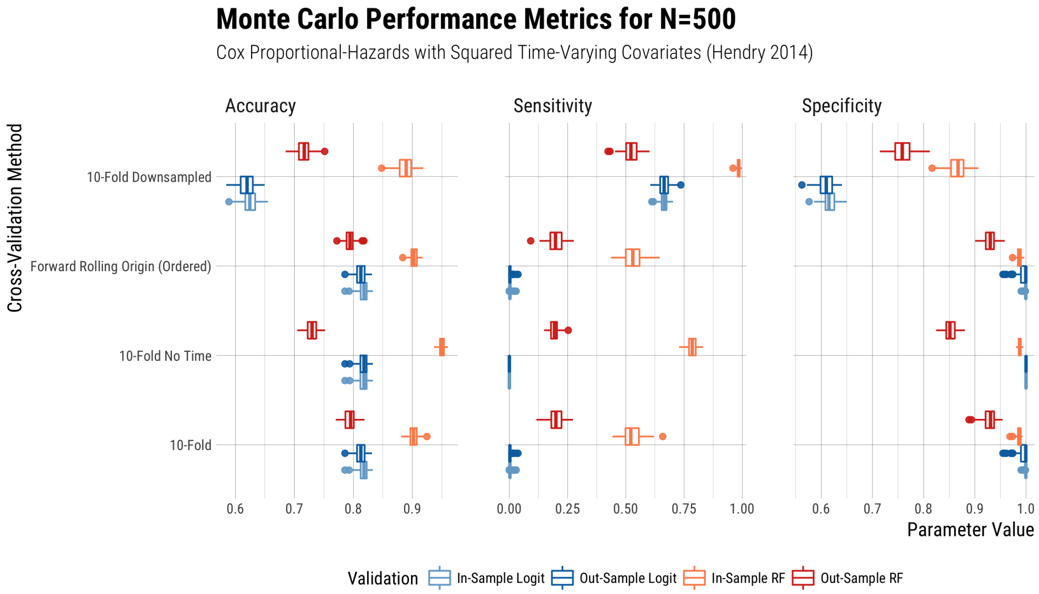


Figure A1. Monte Carlo simulation metrics for N=500 with Logit and Random Forest models. DGP specified with squared function of time. From left to right, panels shows model accuracy, sensitivity, and specificity. Monte Carlo simulations demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well in regards to sensitivity.

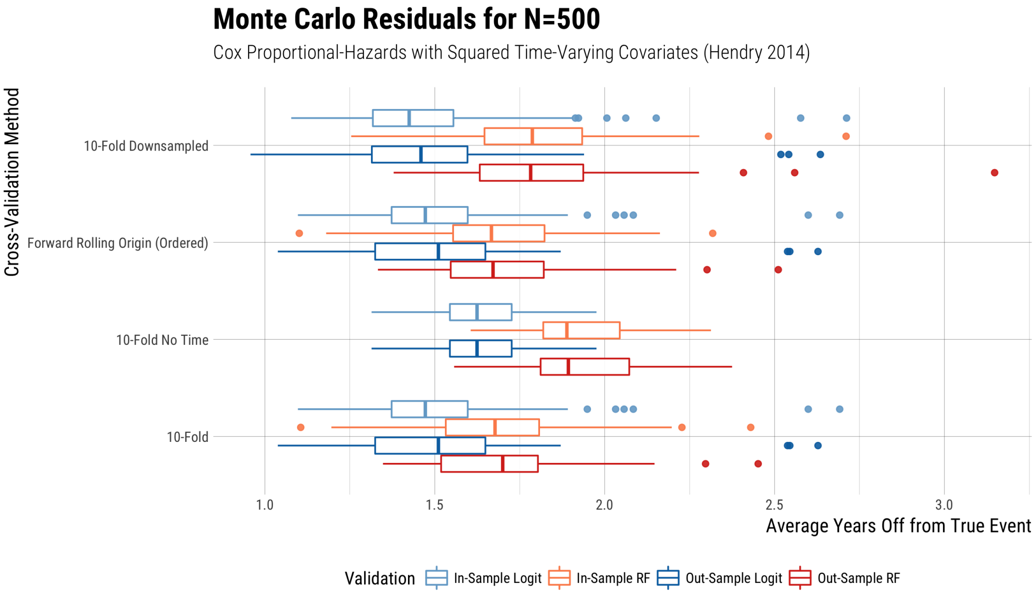


Figure A2. Monte Carlo simulation RLC metrics for N=500 with Logit and Random Forest models. DGP specified with squared function of time. Panel shows that different to the models' sensitivity. The RLC demonstrates that predictions are fairly close to the true outcome, and that variation between CV methods and estimation methods are less pronounced.

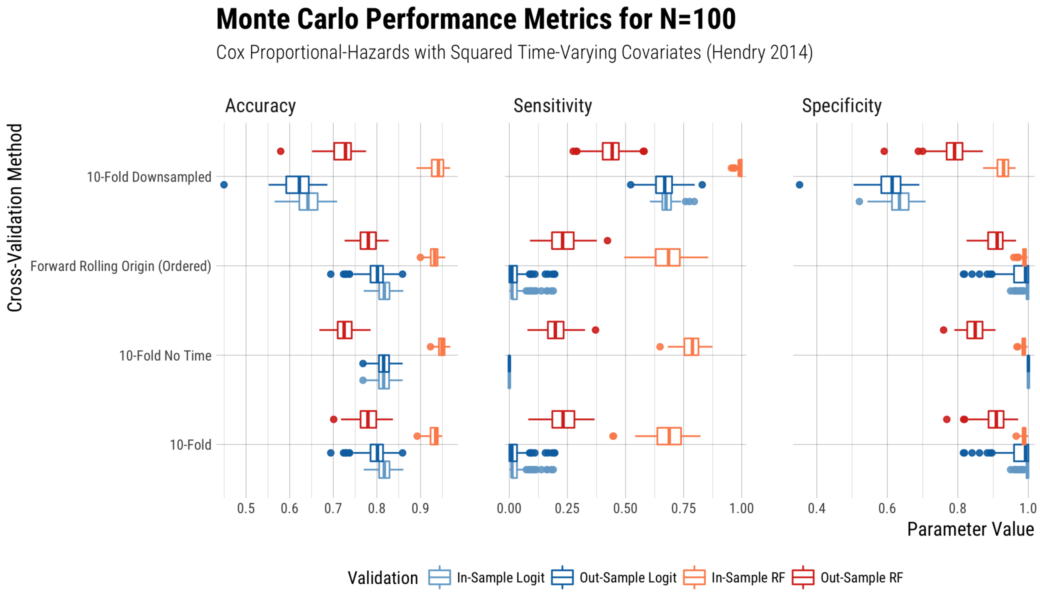


Figure A3. Monte Carlo simulation metrics for N=100 with Logit and Random Forest models. DGP specified with squared function of time. From left to right, panels shows model accuracy, sensitivity, and specificity. Monte Carlo simulations demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well in regards to sensitivity.

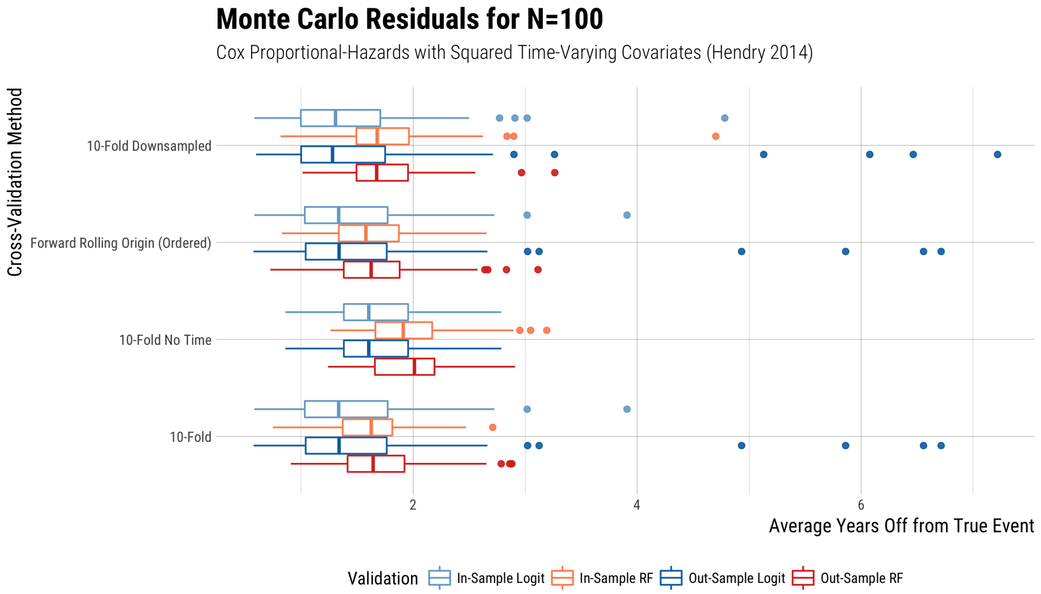


Figure A4. Monte Carlo simulation RLC metrics for N=100 with Logit and Random Forest models. DGP specified with squared function of time. Panel shows that different to the models' sensitivity. The RLC demonstrates that predictions are fairly close to the true outcome, and that variation between CV methods and estimation methods are less pronounced.

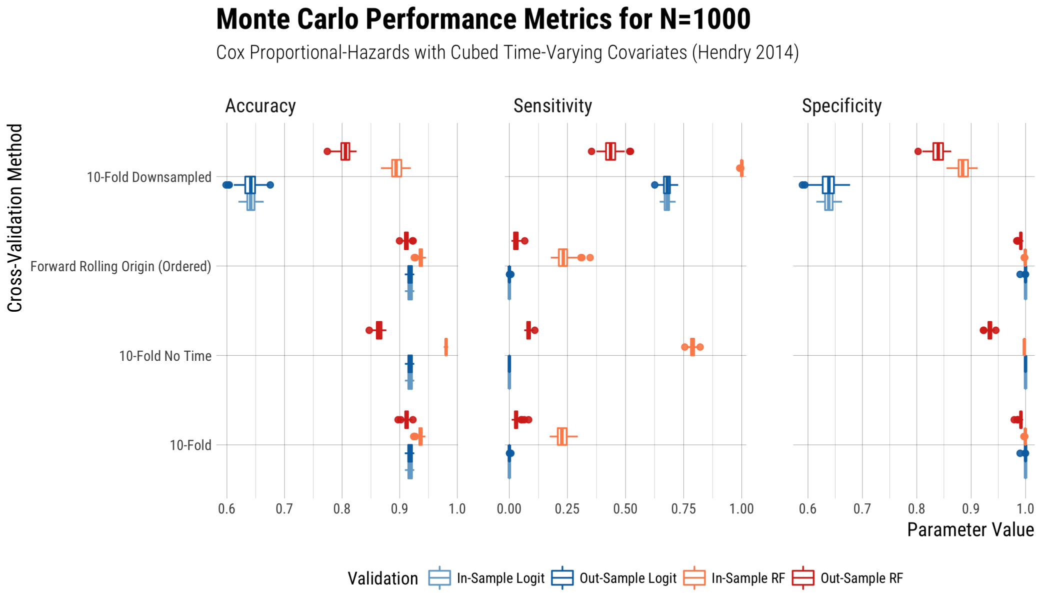


Figure A5. Monte Carlo simulation metrics for N=1000 with Logit and Random Forest models. DGP specified with cubic function of time. From left to right, panels shows model accuracy, sensitivity, and specificity. Monte Carlo simulations demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well in regards to sensitivity.

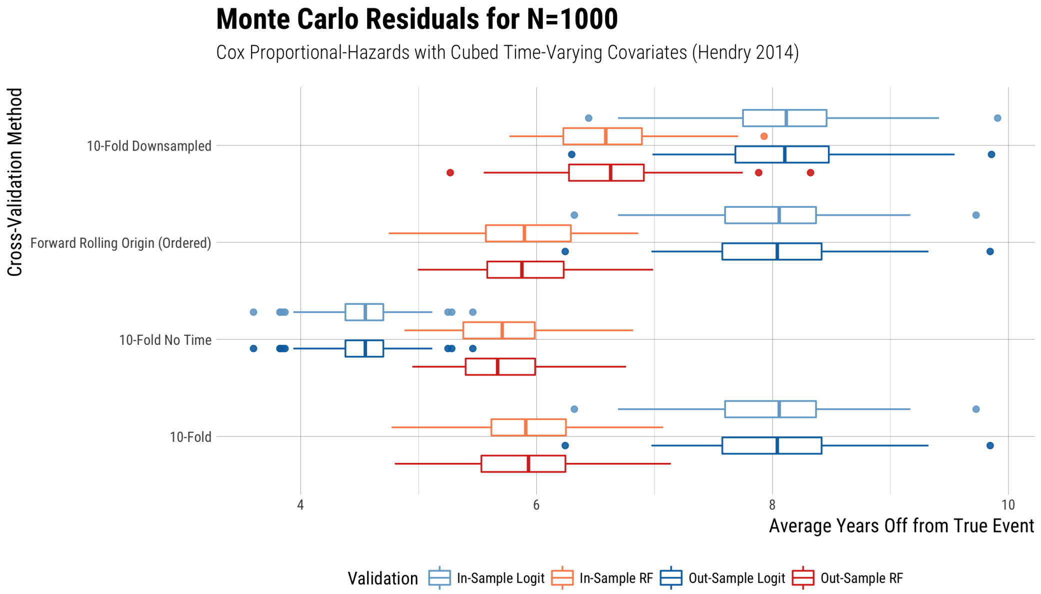


Figure A6. Monte Carlo simulation RLC metrics for N=1000 with Logit and Random Forest models. DGP specified with cubic function of time. Panel shows that different to the models' sensitivity. The RLC demonstrates that predictions are fairly close to the true outcome, and that variation between CV methods and estimation methods are less pronounced.

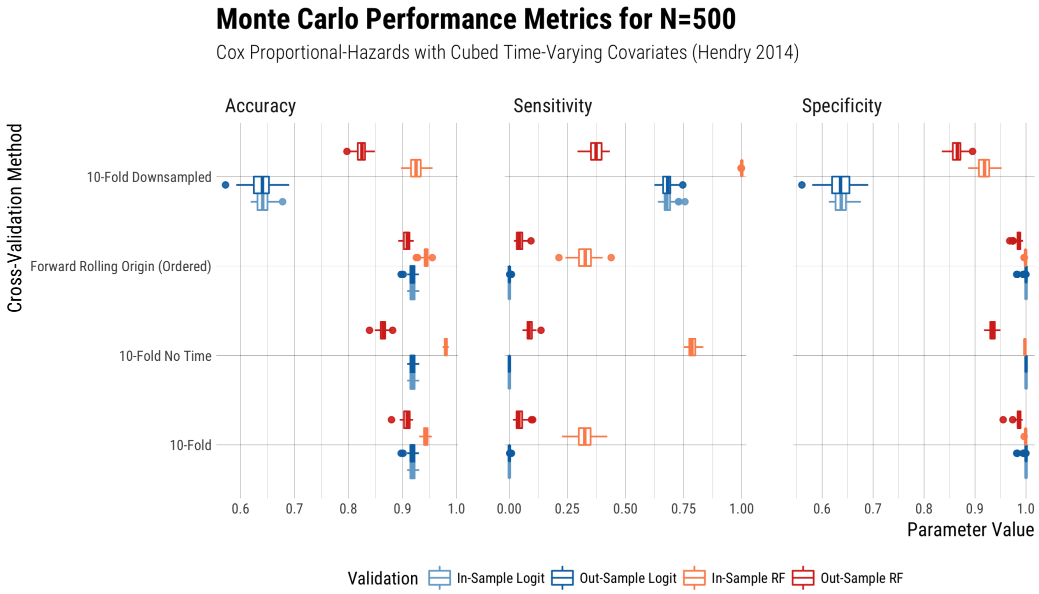


Figure A7. Monte Carlo simulation metrics for N=500 with Logit and Random Forest models. DGP specified with cubic function of time. From left to right, panels shows model accuracy, sensitivity, and specificity. Monte Carlo simulations demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well in regards to sensitivity.

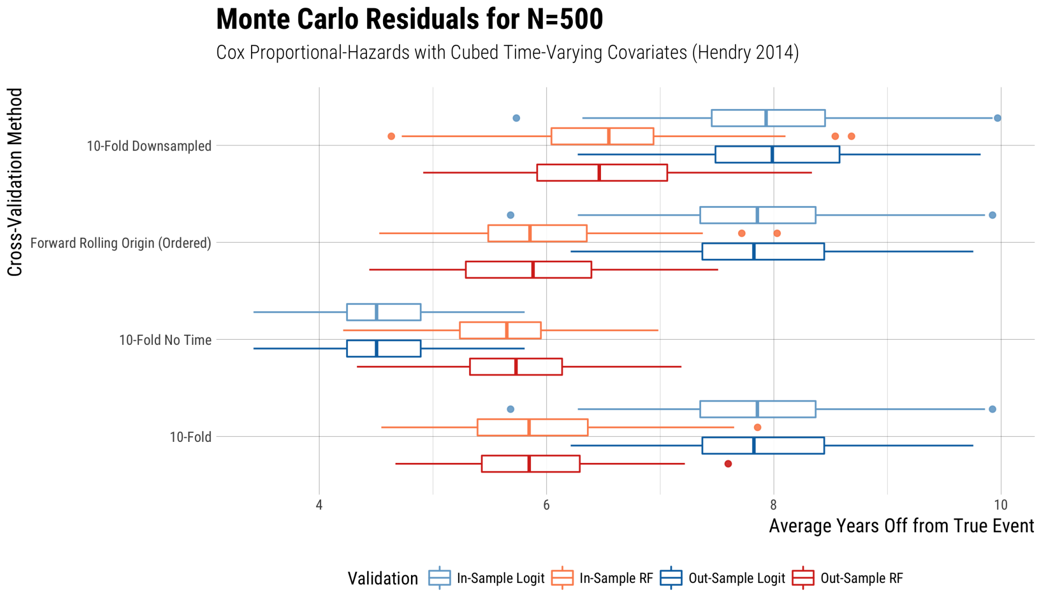


Figure A8. Monte Carlo simulation RLC metrics for N=500 with Logit and Random Forest models. DGP specified with cubic function of time. Panel shows that different to the models' sensitivity. The RLC demonstrates that predictions are fairly close to the true outcome, and that variation between CV methods and estimation methods are less pronounced.

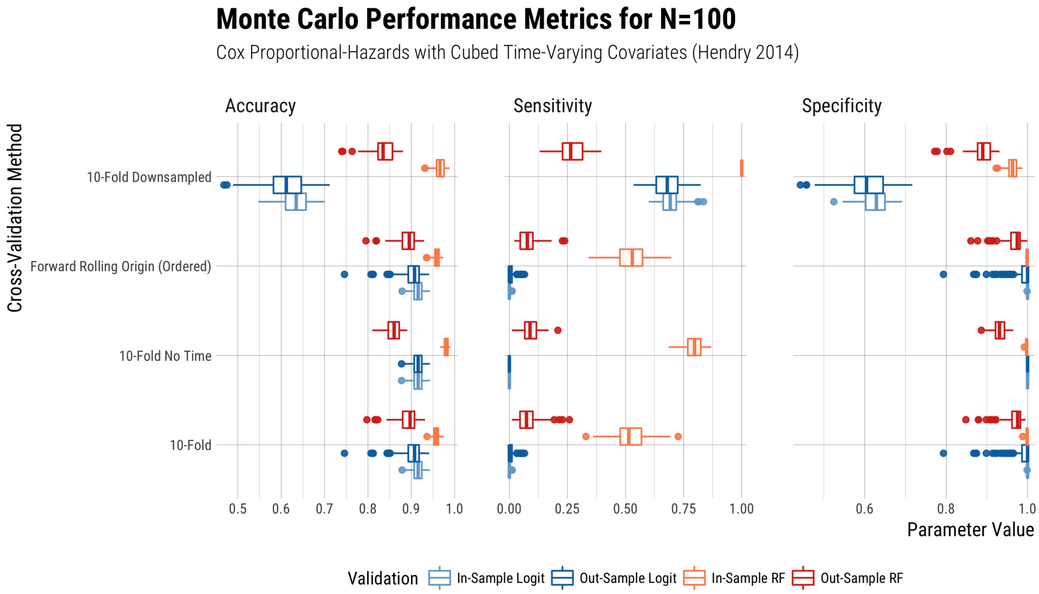


FIGURE A9. Monte Carlo simulation metrics for N=100 with Logit and Random Forest models. DGP specified with cubic function of time. From left to right, panels shows model accuracy, sensitivity, and specificity. Monte Carlo simulations demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well in regards to sensitivity.

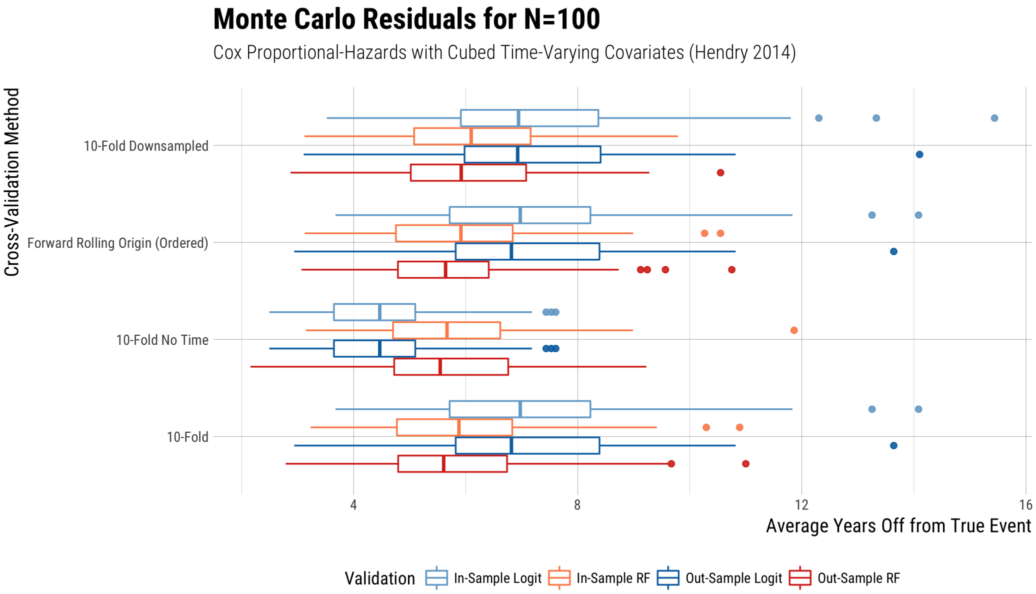


Figure A10. Monte Carlo simulation RLC metrics for N=100 with Logit and Random Forest models. DGP specified with cubic function of time. Panel shows that different to the models' sensitivity. The RLC demonstrates that predictions are fairly close to the true outcome, and that variation between CV methods and estimation methods are less pronounced.

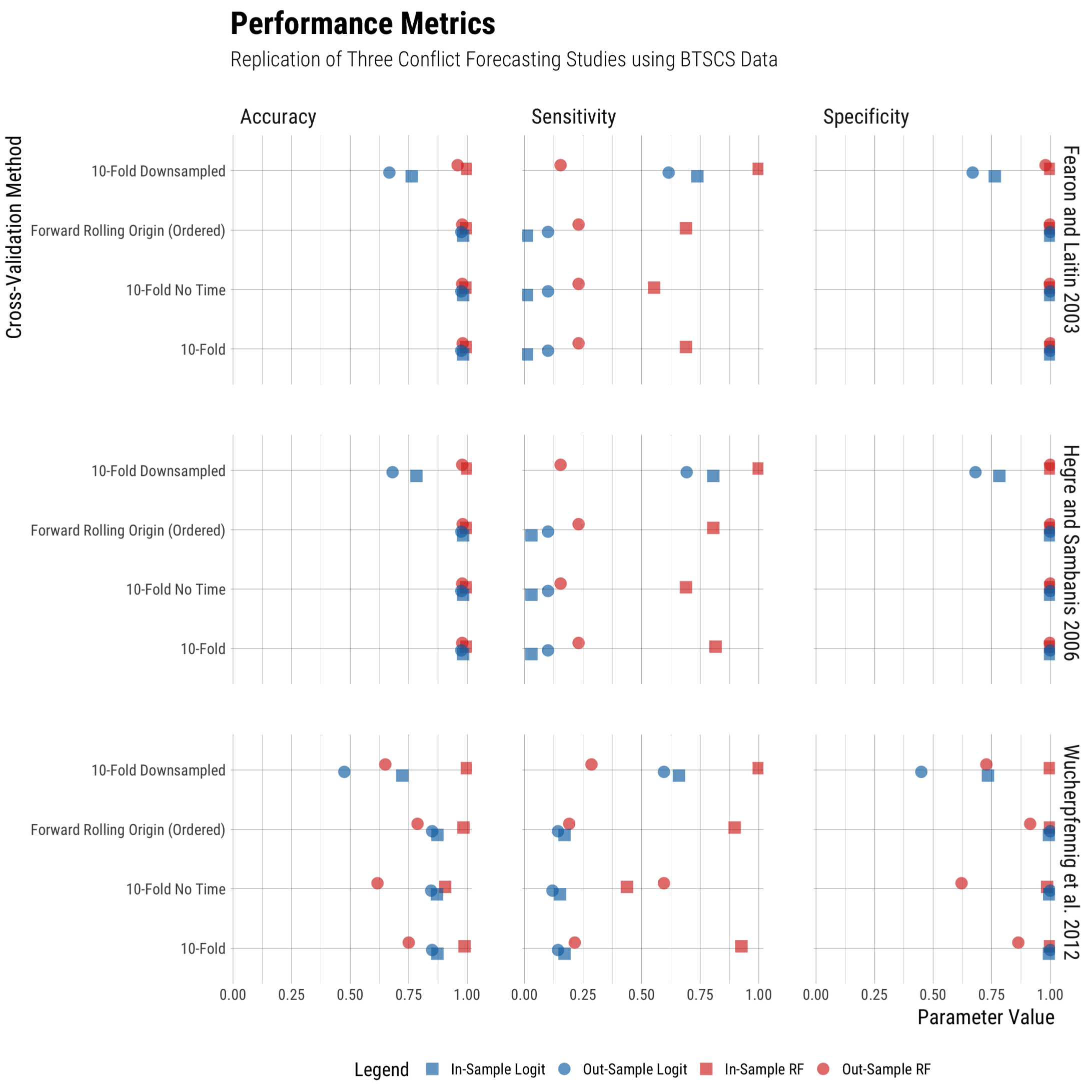


Figure A11. Prediction metrics for Fearon and Laitin (2003); Hegre and Sambanis (2006); Wucherpfennig et al. (2012) with Logit and Random Forest models. From left to right, panels shows model accuracy, sensitivity, and specificity. Similar to the Monte Carlo simulations the empirical examples demonstrate that higher accuracy rates are driven by high specificity rates. Only 10-fold downsampled CV performs reasonably well out-of-sample in regards to sensitivity.

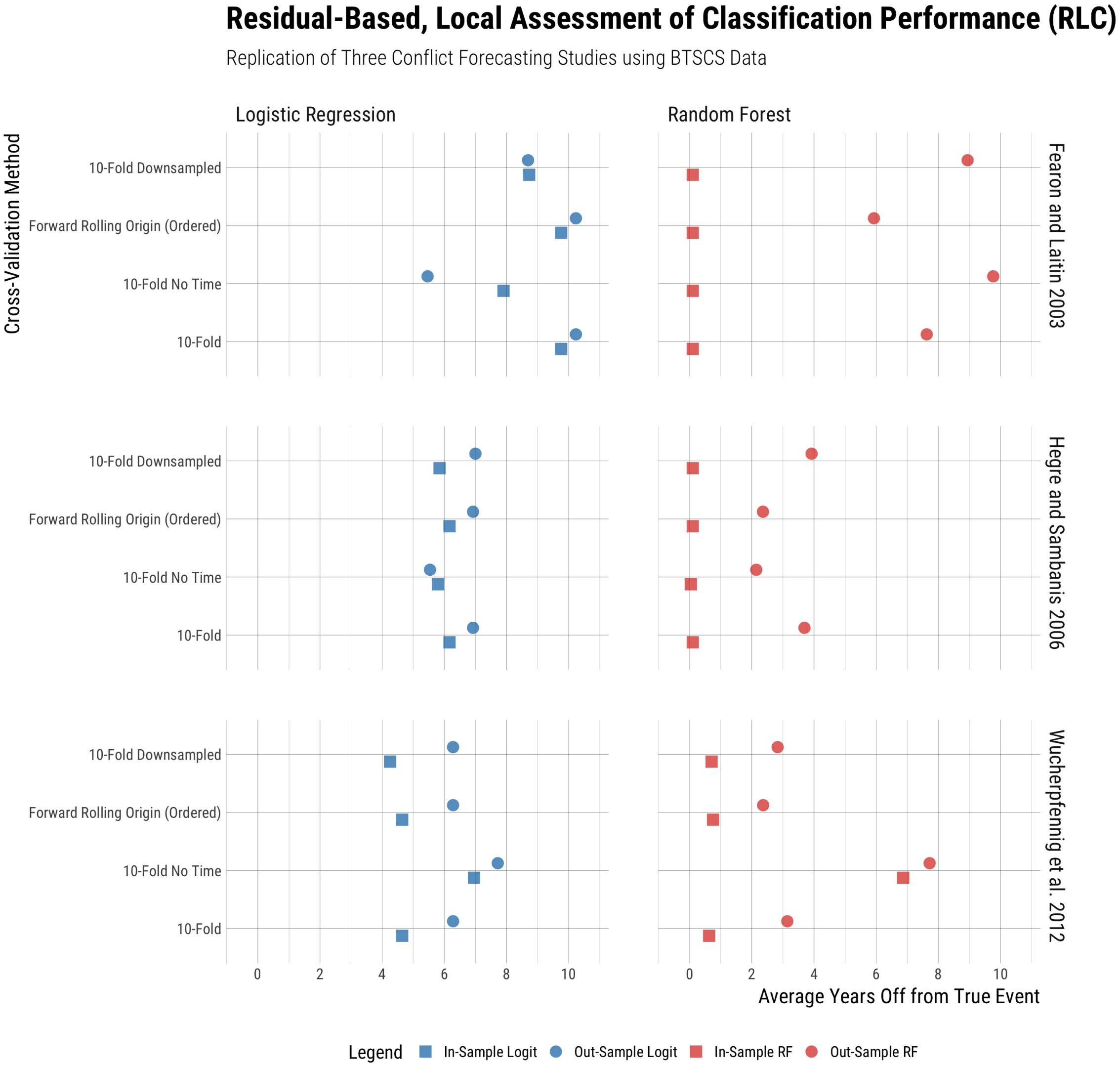


Figure A12. RLC performance of three replicated publications on conflict dynamics. Top panels show logistic regression performance. Bottom panels correspond to random forest models. Results demonstrate that mean out-sample predictions perform better than expected from sensitivity results. This is pertains especially to random forest results.