Taking time seriously when evaluating predictions in Binary-Time-Series-Cross-Section-Data

Gökhan Çiflikli
UCL

Kit Rickard
UCL

Sigrid Weber
UCL

Nils W. Metternich
UCL

Abstract

Efforts to predict civil war onset, its duration, and subsequent peace have dramatically increased. Nonetheless, by standard classification metrics the discipline seems to have made little progress. Although some remedy is promised by particular cross-validation strategies and machine learning tools, which increase accuracy rates substantively, predictions over time remain challenging. In this research note we provide evidence that the predictive performance of conflict models is plagued by temporal residual error. We demonstrate that standard classification metrics for binary outcome data are prone to underestimate model performance in a Binary-Time-Series-Cross-Section context when temporal prediction error is high. We approach this problem as a Modifiable Temporal Unit Problem and propose to evaluate the predictive performance of this type of model in differently sized temporal windows. While retaining the ability of models to leverage disaggregated data for prediction, we provide a parsimonious aggregation approach that allows researchers to evaluate the time frame in which predictive models perform best. We demonstrate this procedure in Monte Carlo experiments and with existing empirical studies.

Keywords: keywords, comma-separated, not capitalized.

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; Schrodtt 2006; Goldstone et al. 2010; Weidmann and Ward 2010; Schneider et al. 2011; De Mesquita 2011; Ward et al. 2013; Gleditsch and Ward 2013; Hegre et al. 2013; Bell et al. 2013; Brandt et al. 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 mak-

ers' demands for meaningful forecasts. We contribute to this literature by demonstrating that standard prediction performance measures do not take into account temporal prediction error in the context of Binary-Time-Series-Cross-Section (BTSCS) data. We argue that accounting for potential temporal prediction error in BTSCS prediction models should be standard practise for researchers using discrete temporal models. This allows researchers to understand how predictive performance of their BTSCS model varies across different aggregation windows.

2. Prediction in Binary Time Series Cross Section Data

Our main arguments rests on the well established insight that the BTSCS model is a special case of a duration model (Beck et al. 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 a continuous outcome. Binary outcome variables are common, especially in International Relations (Beck et al. 1998). While predictive performance of BTSCS models is usually assessed by identifying correctly predicted years, months, weeks or even days in cross-sectional units, we argue that predictive performance needs to account for the temporal window used to generate predictions. To illustrate the need to focus on temporal aggregations, we provide two interlinked conceptual problems of ignoring the time dimension when evaluating the predictive performance of BTSCS models: the Temporal Residual Problem and the Modifiable Temporal Unit Problem. These two problems are inherently linked in BTSCS models as the Temporal Residual Problem describes the inability of common predictive performance metrics to account for temporal predictive errors, while the Modifiable Temporal Unit Problem

demonstrates that especially small temporal resolutions are likely to exacerbate the Temporal Residual Problem. At the same time, linking the Temporal Residual Problem to the Modifiable Temporal Unit Problem allows us to propose a simple aggregation procedure to assess the predictive model performance across time windows.

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 are 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 et al. 2011) try to alleviate this problem, but they

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) Modifiable Temporal Unit Problem: Panel similar to (A) except for time resolution. Simply modifying the temporal unit creates a perfect model.

cannot account for residuals in time. This implies that a positive prediction shortly before a realized event is treated similar to a positive prediction further away from the actual realization. The main reason for the temporal residual problem to occur is a mismatch between the data generating mechanism of duration data, where the data generation error refers to the time dimension, and classification metrics that evaluate prediction error only cross-sectionally.

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 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). If time periods were years and the events were the outbreak of conflict, this example reflects a situation where one model predicts conflict in 1991, the other in 1995, and conflict breaks out in 1996.

2.2. Modifiable Temporal Unit Problem. An additional problem arises from the fact that BTSCS data are discrete duration models. Instead of measuring duration continuously, data are split into temporal units. However, as data resolution or 'granularity' increases (Li et al. 2016), scholars have greater discretion over the choice of temporal scale (Meentemeyer 1989; Bettini et al. 1996). This has led to a growing literature on the so-called Modifiable Temporal Unit Problem (MTUP) which reflects the question of how the choice of the temporal scale affects statistical inference (e.g. Freeman 1989; Shellman 2004; Alt et al. 2001).¹

For BTSCS data, observations are aggregated from temporal units (continuous time) into temporal segments (discrete time) as binary incidence variables. When measuring processes of change, more frequent sampling returns a more detailed data, while coarser granularity results in sparser sampling regimes, less detailed data, information loss, and the risk of masking causal effects (Alt et al. 2001; Freeman 1989; Hornsby and Egenhofer 2002).² However, the measure of a scale - be it spatial or temporal - is rarely an a priori given fact (Koch and Carson 2012) and frequently ex-post determined by frequency of measurement or conventions in the discipline (Sheppard and McMaster 2008, 7).

A simple example of how the choice of time segments matter, is to think of the time segment $\Delta t \rightarrow 0$. As the temporal segments becomes smaller, it becomes increasingly difficult to predict an event at an exact point in time. One can imagine, for example, the difficulty of predicting the outbreak of a war to the second as opposed to the year. Standard binary classification metrics that were developed to evaluate prediction performance on cross-sectional

¹Issues of scale selection have received more attention in geographic research, where the more widely known Modifiable Areal Unit Problem (MAUP) focuses specifically on spatial scales (Openshaw and Taylor 1979).

²See also Cheng and Adepeju (2014) who identify three problems when discretizing time into temporal units: temporal aggregation, segmentation and boundary effects. First, the researcher chooses an arbitrary temporal aggregation rule to split time into regularly spaced units. Second, when observations are converted from a fine time interval to a coarser interval, the data are aggregated into a summary statistic (such as count, mean or variance). This can be problematic because summary statistics vary depending on the starting point of the temporal interval. The segmentation effect reflects the problem of how to choose the starting point of temporal intervals. Third, researchers face a temporal boundary problem as researchers identify the start and end points of a time series with little to no theoretical justification.

data ignore the Modifiable Temporal Unit Problem in BTSCS data.

To further illustrate this point, consider Figure 1b 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 changed the blue model's predictions are now 'perfect'. Hence, by changing the time scale the blue model sensitivity and specificity shifted from 0 to 1 and 0.8 to 1, respectively.

3. Predictive Performance for Multiple Time Windows

In approaching the identified Temporal Residual Problem and Modifiable Temporal Unit Problem, we suggest a procedure similar to space-time scan statistics used to detect spatio-temporal phenomena (SaTScan 2007; Schutte and Donnay 2014; Braithwaite and Johnson 2016; O'Loughlin and Witmer 2011). We propose the aggregation of predicted probabilities according to basic probability laws, which allows us to assess models at different temporal resolutions while preserving detailed data from more frequent sampling.

Following our approach, the performance of prediction models in BTSCS data should be evaluated across varying temporal windows post-estimation. Prior-to-estimation approaches aggregate data to different discrete time units and estimate the respective models. An example is the aggregation of spatial point data, such as conflict events, to areal units before estimating models. However, aggregation necessitates the loss of information through aggregation rules on both the right- and left-hand side of the equation. Our approach leverages independent and identically distributed as random (i.i.d.) assumptions in BTSCS approaches to aggregate predictions post-estimation. We know from the BTSCS literature on binary outcome models that we can treat observations i.i.d. once we account for time dependency on the right-hand side of the equation. For example, in the context of the well studied logit model this implies that we can treat the probability of an event occurring in $y_{i,t}$ given a set of covariates $x_{i,t}$ as independent from $y_{-i,t}$ and $y_{i,-t}$ if we account for the time dependency through the inclusion of a variable or a set of variables $f(t)$ capturing this dependency.

$$P(y_{i,t} = 1 | x_{i,t}) = \frac{1}{1 + e^{-(x_{i,t}\beta + f(t))}}$$

Assuming that i.i.d. is not violated, we can aggregate the predicted probabilities ($\hat{p}_{i,t}$) through probabilistic rules relying on the independence assumption. More specifically, we define t as a time point expressed in the current discrete time unit $t = 1, \dots, T$. In order to aggregate to a time window Δ_t , let us assume that d is the size of time window Δ_t and that $M = \frac{T}{d}$ is the number of aggregation windows that can be established given T and d .

Thus, aggregation necessitates the mapping of a point in time t_j into an interval $\Delta_j = [t_0, t_0 + d), \dots, [t_0 + (M-1)d, t_0 + Md)$. Unit t_j falls into aggregation Δ_j if $t_0 + (m-1)d \leq t_j < t_0 + md$. In this case we can say $t_j \in \Delta_j$. Given the independence assumption, the probability that at least one event takes place in unit i in time aggregation window Δ_j is 1 minus the probability of no event in any of the time periods that fall into Δ_j . Because the probabilities are assumed to be independent, we obtain the combined probability of no event in Δ_j by multiplying the individual probabilities at each time point. For example, if the probability of an event is 0.8 at t_1 and 0.6 at t_2 , when both periods are aggregated to a new time window (Δ_j) the probability of any event occurring (p_{i,Δ_j}) is 0.92.³

$$\hat{p}_{i,\Delta_j} = 1 - (\prod(1 - \hat{p}_{i,t_j \in \Delta_j}))$$

³1 - ((1-0.8) x (1-0.6))

This puts us in a position to evaluate predictions at a greater granularity without compromising information loss compared to prior-to-estimation approaches. It also allows us to evaluate model performance at differently sized Δ s and report improving or declining predictive performance at higher aggregation levels. We propose that routinely implementing this approach in BTSCS models mitigates the described Temporal Residual Problem as well as the Modifiable Temporal Unit Problem. We recommend to assess models with both the Area under the Receiver Operating Characteristic curve (AUC-ROC), as well as the Area under the Precision-Recall curve (AUC-PR), as they provide a basic understanding of model performance across different classification cutoffs. In addition, the AUC-PR provides valuable insights to model performance in highly class-imbalanced data, which many of the BTSCS data have by definition.

4. Monte Carlo Study

In this section, we demonstrate that standard classification approaches for BTSCS models ignore the Temporal Residuals Problem and are sensitive to the Modifiable Temporal Unit Problem. 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). The data generating process underlying the Monte Carlo simulations is provided by [Hendry \(2014\)](#), proxying a Cox-proportional hazard data generating process with time-varying covariates by relying on a transformation of random variables, which are generated according to a truncated piecewise exponential distribution ([Zhou 2001](#); [Leemis 1987](#)).

We conduct Monte Carlo experiments for two different non-monotonic hazards (squared and cubic functions of time) and run 100 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.

We implement two predictive models for the purpose of this study. The first is a logistic regression model with the following model specification, $Y = f(X\beta + t + t^2 + t^3)$. Logistic regression models have been most commonly used in political science applications to BTSCS ([Greenhill et al. 2011](#)) and are likely to remain an important approach to classification problems in International Relations and Conflict Studies. However, researchers are increasingly turning to machine learning approaches when faced with classification problems in BTSCS. Especially for purely predictive purposes, machine learning tools are gaining popularity, which is reflected in current scholarship ([Muchlinski et al. 2016](#)). Therefore, we also implement a set of random forest models to demonstrate that the outlined challenges for BTSCS classification are not method dependent. 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.⁴ We utilize 10-fold cross-validation for training the logistic and random forest models. After training the models with one of the 100 generated data sets for each sample size, we conduct out-of-sample predictions on a newly generated data set of the same sample size.⁵

⁴We utilize the `ranger` package ([Wright and Ziegler 2017](#)), a fast C++ implementation ported to R, via the `caret::train` framework.

⁵Time-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.

Figure 2: Out of sample Monte Carlo simulation for $N=100$, $N=500$, and $N=1000$ with Logit Regression models. Left panels report the Area under the ROC-Curve for increasing temporal aggregation of predictions, while the right panels refer to the Area under the Precision-Recall Curve. Upper panels pertain to sample size $N=100$, middle panels $N=500$, lower panels refer to sample size $N=1000$.

4.1. Logistic Regression Results

Figure 2 displays results of the Logistic Regression Monte Carlo experiment with $N=100$, $N=500$, and $N=1000$. The plots refer to out-of-sample performance of models trained by 10-fold cross-validation. On the x-axis of each panel, we find increasing temporal aggregation

Figure 3: Out of sample Monte Carlo simulation for $N=100$, $N=500$, and $N=1000$ with Random Forest models. Left panels report the AUC-ROC for increasing temporal aggregation of predictions, while the right panels refer to the AUC-PR. Upper panels pertain to sample size $N=100$, middle panels $N=500$, lower panels refer to sample size $N=1000$.

resolution. The value of one refers to temporal unit of estimation. The y-axis in each panel refers to the AUC-ROC or AUC-PR at each temporal aggregation.

Independent of sample size, we find that the AUC-ROC increases as the temporal aggregation becomes greater. However, the gains in AUC-ROC level off after the initial increase in temporal aggregation windows. Because we are faced with highly class-imbalanced data, this

implies that the ability to predict zeros increases initially, but the rate at which it does decreases with greater aggregations. In contrast, the ability of the models to correctly identify time periods that experience an event (AUC-PR) increases at a steady rate over aggregations. In this simulation this is not surprising as the data generating process consistently links observations closer to the realized outcome with higher values in X . Hence, e.g., at the greatest aggregation (two time periods) the ability of the model to correctly classify an event becomes very likely.

4.2. Random Forest Results

Figure 3 provides the results for the random forest models experiment. Similar to Figure 2, Figure 3 provides out-of-sample performance of models trained by 10-fold cross-validation. The results are almost identical to the logistic regression experiment with slightly lower AUC-ROC values and slightly higher AUC-PR values. The marginal difference between logistical regression and random forest predictions is of course due to the low complexity data generating mechanism used in the Monte Carlo simulations. The overall picture that emerges is that the ability to identify zeros in class-imbalanced class data with time dependencies increases with time aggregation and then levels off, while the ability to correctly identify ones is greatest when using the highest aggregation, which means aggregating to a cross-section of the data. Again, this is expected when the data generating process consistently links higher value in X with realizations over time. Furthermore, these findings are in line with our expectation that existing prediction metrics are unable to account for the Temporal Residual Problem and the Modifiable Temporal Unit Problem. These theoretical insights motivate us to consider existing empirical studies on conflict onset and duration to assess the degree to which standard metrics ignore the temporal dimension in prediction performance.

Table 1: Example of grouped duration data structure.

	Spell ID	Country ID	Year	Time ID	Event
1	A.1	A	1960	1	0
2	A.1	A	1961	2	0
3	A.1	A	1962	3	1
4	A.2	A	1966	1	0
5	A.2	A	1967	2	0

5. Predicting Conflict

We evaluate data from three studies in the civil war literature. We choose these data because the related studies are influential in this field, but more importantly because they are conducted at different temporal resolutions: the year (Hegre et al. 2013), the month (Chiba and Gleditsch 2017), and the day (Weisiger 2016). It is important to note that we do not replicate the the studies, but merely use their data and similar model specifications using generic logistic and random forest models.⁶

⁶Using data from these studies to re-estimate more simplistic models (with same input variables) demonstrates the general procedure we propose to evaluate the predictive power over different temporal windows. We do not aim to improve on the predictive performance of the three assessed studies. The original models

Hegre et al. (2013) predict changes in global and regional incidences of yearly armed conflict for the 2010-2050 period using dynamic multinomial logit model estimation.⁷ We re-estimate their model drawing on the same yearly dataset of 178 countries between 1945 and 2000.

Chiba and Gleditsch (2017) build on an earlier study by Buhaug et al. (2014) and draw on a monthly dataset of conflict onset and termination from January 1996 to December 2012. The key innovation of this study is the inclusion of dynamic information on interactions between different actors within states, which the authors measure with political event data from the Integrated Conflict Early Warning System (ICEWS) (Boschee et al. 2015). We focus on the first of their two-way transition model-transitions from war to peace—which they estimated as a logit model.

Table 2: Summary of data drawn from empirical studies.

	Study	Scale	Period	Dependent variable	Events	Observations	Spells	Mean spell duration
1	Hegre et al. (2013)	Year	1945-2000	Conflict onset	116	6359	283	22.46996
2	Chiba and Gleditsch (2017)	Month	1995-2015	Conflict onset	145	37765	320	118.01562
3	Weisiger (2016)	Day	1848-2003	Coalition end	85	40085	85	471.58824

Finally, prediction studies in this field rarely make predictions at low temporal resolutions such as the week or the day. Instead, we use data from a study by Weisiger (2016), who tests a theory of why countries abandon allies in war. Weisiger (2016) conducts a survival analysis where the unit of analysis is the day of country membership in a war coalition in the period 1848-2003. Survival analyses, as noted by Beck et al. (1998), is similar to a corrected logit analysis of BTSCS data as they both analyse grouped duration data and account for temporal dependence between observations.

Due to the similarities in the underlying data structure necessary to conduct survival analyses and BTSCS analyses, Weisiger (2016)’s data was already setup as grouped duration data. In order to replicate the studies of Hegre et al. (2013) and Chiba and Gleditsch (2017) as BTSCS analyses, we included independent IDs for each spell in their data. Spells are periods of no conflict and end either at the end of the study period (right censored) or when the country experiences an event (in this case, a conflict or leaving a wartime coalition), as shown in Table 1. Country *A* experiences an event in year 1962. It re-enters the dataset after the conflict in 1966 with a new spell ID (*A.2*). In both conflict datasets, periods of active conflict are omitted from the analysis. For example, in the Hegre et al. (2013) data, Burundi’s first spell begins in 1962 when it gained its independence from Belgium. This ends in 1965 when conflict breaks out. Burundi’s second spell starts in 1970 and ends in 1972. In total, Burundi has four spells. The same spell structure is in Chiba and Gleditsch (2017)’s data. For example, the USA’s first spell starts at the beginning of the study period in January 1995. It lasts 81 months until September 2001, when the US invaded Afghanistan. It’s second spell starts in January 2015 and lasts until the end of the study period as it does not experience conflict

yield significantly better predictions than our streamlined prediction approach. Hegre et al. (2013) for example report that their more complex final prediction model has an area under the ROC of 0.937 for the years 2007-2009 (out-of-sample prediction) while we only reach a fraction of this performance with our procedure. Out of sample, Chiba and Gleditsch (2017)’s model for the onset of territorial conflict has an AUC of 0.932 and 0.868 for government onset.

⁷Hegre et al. (2013) first estimate a set of models for the 1970-2000 period and predict conflict for 2001-2009. The following simulation procedure used to make predictions of conflict in the 2010-2050 period is designed to model the endogeneity between predictions. The predictions are endogenous as a predicted conflict in one country will affect subsequent predictions in said country but also in neighboring countries.

Logistic regression models

Figure 4: Prediction metrics depending on different temporal aggregations for the respective data (Chiba and Gleditsch 2017; Hegre et al. 2013; Weisiger 2016). Logistic regression models with lagged independent variables. The right panels display the AUC-ROC for each temporal window; the left panels show the AUC-PR depending on the temporal window.

again.

We re-estimate 1) logistic regression models and 2) random forest models with 10-fold cross-validation. To provide robust estimates we use lagged independent variables in all models

and we cross-validate the random forest models using 10-fold cross-validation before predicting out-of-sample. To evaluate the out-of-sample predictions across different temporal windows, we calculate the aggregated predicted probability that an event occurs relying on the independence assumption as outlined above. This allows us to use fine-grained data for prediction but to evaluate the performance across varying time aggregations.⁸ Our results are summarised in Figure 4 (logit models) and Figure 5 (random forest models). The plots demonstrate how the temporal aggregation (x-axis) affects the AUC-ROC (left panels) and the AUC-PR (right panels).

Figure 4 displays the results of our estimations using logistic regression models. In general, we find a similar pattern across all three empirical studies. We find that temporal aggregation increases the AUC-ROC, suggesting that using more coarse temporal windows improves predictive performance. For example in the case of [Hegre et al. \(2013\)](#), we find that a 25 year time period yields an AUC-ROC value of 0.4 while making yearly predictions only yields an AUC-ROC value of 0.25. With some variation, the steady improvement of predictive performance indicates that all three models poorly predict temporal dynamics and predominantly identify cross-sections that are more likely to see conflict.

When dealing with highly class-imbalanced data, the AUC-PR can be more informative of a model's predictive performance than the AUC-ROC ([Davis and Goadrich 2006](#)). Since class-imbalanced data, in which an event rarely occurs, can be common in International Relations (e.g. occurrence of war) and a feature of BTSCS data, we present the aggregation results for the AUC-PR in the right panels. The Precision-Recall curve shows the trade-off between the precision - the number of correct positive predictions - and the recall - the correct positive predictions made out of all positive predictions - and does not depend on true negatives. Our results suggest that the Precision-Recall also increases for more coarse temporal windows when re-estimating the data of [Weisiger \(2016\)](#), [Chiba and Gleditsch \(2017\)](#) and [Hegre et al. \(2013\)](#). This again suggest that we can maximise predictive power by identifying realistic temporal windows to use for prediction.

In Figure 5, the replication results are presented for random forest models that were trained and cross-validated before out-of-sample prediction. With some small exceptions, the random forest models yield a similar result to the logistic regression models: temporal aggregation improves predictive performance.

6. Conclusion

Current prediction metrics focus on cross-sectional errors, while the data generating process of duration models (BTSCS models being a special case of such) explicitly generate uncertainty in temporal realization. We identify two conceptual frameworks: the Temporal Residual Problem and the Modifiable Temporal Unit Problem that allow us clearly point to challenges of current prediction metrics. We further attempt to draw attention to these problems through Monte Carlo simulations and then by estimating a set of logistic regression and random forest models on existing data in the context of conflict studies.

Relying on independence assumptions, we suggest that researchers should evaluate the predic-

⁸To train the random forests and make predictions, we keep the original yearly, monthly, and daily data structure and only aggregate the predicted probabilities. A different workflow would be to first aggregate the data, then train and predict. We show results for this procedure in Figure A1 (logistic regression) and Figure A2 (random forest) in the appendix.

Random forest models

Figure 5: Prediction metrics depending on different temporal aggregations for the respective data (Chiba and Gleditsch 2017; Hegre et al. 2013; Weisiger 2016). Random forest models with lagged independent variables. The left panels display the AUC-ROC for each temporal window; the right panels show the AUC-PR depending on the temporal window.

tive performance of their models for different temporal windows post-estimation. We propose a parsimonious approach to do so by aggregating predictions that rely on fine-grained data to higher temporal segments. We demonstrate this procedure in Monte Carlo simulations

and using data of three empirical studies. As expected, these demonstrations show that ignoring the temporal residual and modifiable temporal unit problem leads to the predictive performance increasing at higher levels of aggregation. This is particularly important where aggregation to only slightly larger time windows can lead to doubling in the AUC-PR or where some aggregations show clear jumps in predictive performance (see Figure 4 and Figure 5).

We encourage scholars to implement our post-estimation aggregation in the spirit of model improvement (Colaresi and Mahmood 2017) rather than model selection. For example, if AUC-ROC and AUC-PR increase quickly after temporal aggregation, we can be confident that our models include variables that might not necessarily predict the right day, month, or year, but that they are able to track time dynamics that are relevant to the prediction of the respective outcome. However, if the AUC-ROC and AUC-PR only slowly improve across temporal aggregations, this implies that our models include variables that are picking up on cross-sectional structural variation, but are missing out on explaining temporal variation. If this is the case, researchers might want to decide to change their data aggregation pre-estimation.⁹

Hence, we believe that scholars can benefit from an assessment in which temporal window their models reliably predict events and how well their models account for temporal dynamics, which might inform the inclusion of additional variables, a re-assessment of the discrete time unit, and communication to policy makers. The approach we present in this research note can be readily applied in predictive studies using BTSCS data to gain a better understanding of the temporal prediction performance of our models. 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

- Alt, J. E., King, G., and Signorino, C. S. (2001). Aggregation among binary, count, and duration models: Estimating the same quantities from different levels of data. *Political Analysis*, 9(1):21–44.
- Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.
- Beck, N., Katz, J. N., and Tucker, R. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science*, 42(2):1260–1288.
- Bell, S. R., Cingranelli, D., Murdie, A., and Caglayan, A. (2013). Coercion, capacity, and coordination: Predictors of political violence. *Conflict Management and Peace Science*, 30(3):240–262.
- Bettini, C., Wang, X. S., and Jajodia, S. (1996). A general framework and reasoning models for time granularity. pages 104–111.

⁹In the Appendix Figure A1 and Figure A2, we provide insights to pre-aggregating and estimating the data used in our main analysis. Using the maximum values per time period when aggregating, the prediction problem for identifying true outcomes becomes easier, which is reflected especially in higher AUC-PRs.

- Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., and Ward, M. D. (2015). ICEWS Coded Event Data. *Harvard Dataverse*, V2.
- Braithwaite, A. and Johnson, S. (2016). Space-time modelling of insurgency and counterinsurgency in Iraq. In Wilson, A. G., editor, *Global dynamics*, volume 49 of *Wiley series in computational and quantitative social science*, pages 195–213. John Wiley & Sons, Chichester, West Sussex, UK.
- Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2014). Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30(4):944–962.
- Buhaug, H., Cederman, L.-E., and Gleditsch, K. S. (2014). Square pegs in round holes: Inequalities, grievances, and civil war. *International Studies Quarterly*, 58:418–431.
- Cheng, T. and Adepeju, M. (2014). Modifiable temporal unit problem (mtup) and its effect on space-time cluster detection. *PloS one*, 9(6):e100465.
- Chiba, D. and Gleditsch, K. S. (2017). The shape of things to come? expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research*, 54(2):275–297.
- Colaresi, M. and Mahmood, Z. (2017). Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193–214.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.
- De Mesquita, B. B. (2011). A new model for predicting policy choices preliminary tests. *Conflict Management and Peace Science*, 28(1):65–87.
- Freeman, J. R. (1989). Systematic sampling, temporal aggregation, and the study of political relationships. *Political Analysis*, 1:61–98.
- Gleditsch, K. S. and Ward, M. D. (2013). Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, 50(1):17–31.
- Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J., and Woodward, M. (2010). A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208.
- Greenhill, B. D., Ward, M. D., and Sacks, A. (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, T. R. and Lichbach, M. I. (1986). Forecasting internal conflict a competitive evaluation of empirical theories. *Comparative Political Studies*, 19(1):3–38.
- Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., and Urdal, H. (2013). Predicting armed conflict, 2010–2050. *International Studies Quarterly*, 57(2):250–270.

- Hendry, D. 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.
- Hornsby, K. and Egenhofer, M. J. (2002). Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):177–194.
- Koch, A. and Carson, D. (2012). Spatial, Temporal and Social Scaling in Sparsely Populated Areas- Geospatial Mapping and Simulation Techniques to Investigate Social Diversity. *GLForum*, pages 44–53.
- Leemis, L. M. (1987). Variate generation for accelerated life and proportional hazards models. *Operations research*, 35(6):892–894.
- Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., and Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. 115:119–133.
- Meentemeyer, V. (1989). Geographical perspectives of space, time, and scale. *Landscape ecology*, 3(3-4):163–173.
- Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103.
- O’Brien, S. 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, S. and Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. in n. wrigley and r. bennett, eds., statistical applications in the spatial sciences. london: Pion press.
- O’Loughlin, J. and Witmer, F. D. (2011). The localized geographies of violence in the north caucasus of russia, 1999–2007. *Annals of the Association of American Geographers*, 101(1):178–201.
- SaTScan, K. M. (2007). v. 7.0. 2: Software for the spatial and space-time scan statistics. *Information Management Services, Inc.*
- Schneider, G., Gleditsch, N. P., and Carey, S. (2011). Forecasting in international relations: One quest, three approaches. *Conflict Management and Peace Science*, 28(1):5–14.
- Schrodt, P. A. (2006). Forecasting conflict in the balkans using hidden markov models. In *Programming for Peace*, pages 161–184. Springer.
- Schutte, S. and Donnay, K. (2014). Matched wake analysis: finding causal relationships in spatiotemporal event data. *Political Geography*, 41:1–10.
- Shellman, S. M. (2004). Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*, 12(1):97–104.
- Sheppard, E. and McMaster, R. B. (2008). *Scale and Geographic Inquiry: Nature, Society, and Method*.

- Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., and Weschle, S. (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, N. B. and Ward, M. D. (2010). Predicting conflict in space and time. *Journal of Conflict Resolution*, 54(6):883–901.
- Weisiger, A. (2016). Exiting the coalition: When do states abandon coalition partners during war? *International Studies Quarterly*, 60(4):753–765.
- Wright, M. and Ziegler, A. (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.
- Zhou, M. (2001). Understanding the cox regression models with time-change covariates. *The American Statistician*, 55(2):153–155.

Appendix

Affiliation:

Nils W. Metternich
University College London
London, UK
E-mail: n.metternich@ucl.ac.uk
URL: <https://www.ucl.ac.uk/~uctqnm0>

SocArXiv Website <https://socopen.org/>
SocArXiv Preprints <https://osf.io/preprints/socarxiv>
Preprint *Submitted:* October 9, 2020
<https://osf.io/preprints/socarxiv/tvshu/>
Accepted: October 9, 2020

Logistic regression models (first aggregated)

Figure A1: Prediction metrics for [Chiba and Gleditsch \(2017\)](#); [Hegre et al. \(2013\)](#); [Weisiger \(2016\)](#) when first aggregating the data to different temporal windows and then predicting values using logistic regression. The data is aggregated by taking the maximum value of each variable per temporal window. The right panels display the area under the ROC curve for each temporal window; the left panels show the area under the Precision-Recall curve depending on the temporal window.

Random forest models (first aggregated)

Figure A2: Prediction metrics for [Chiba and Gleditsch \(2017\)](#); [Hegre et al. \(2013\)](#); [Weisiger \(2016\)](#) when first aggregating the data to different temporal windows and then predicting values using random forest. The data is aggregated by taking the maximum value of each variable per temporal window. The right panels display the area under the ROC curve for each temporal window; the left panels show the area under the Precision-Recall curve depending on the temporal window.