, 1–7 ()
DOI:

Dynamic Matrix Factor Model for GDELT Data

Ama AMPADU-KISSI-OWUSU, Rong CHEN, Kathyrina DOWLIN, Yuefeng HAN, Junyi LI, Zebang LI, Ruofei MAO, Han XIAO, Minge XIE, Ruofan YU, and Zheshi ZHENG

Abstract

Geopolitical events exert profound influence on global affairs, shaping the political and economical courses of nations and societies. Understanding these events and being able to accurately predict any future events are critical for effective policy-making, risk assessment, and strategic planning. This paper uses a dynamic factor model for matrix-valued time series as the main prediction model. The outcomes of the main model are further combined with the predictions from several simpler models, and it effectively produces high prediction accuracy for highly heterogeneous count data. The paper leverages the Global Database of Events, Language, and Tone (GDELT) dataset that contains records of events found in broadcast, print, and web news sources around the world. The data consists of weekly frequencies of 20 (aggregrated) types of events from 260 countries and regions, for 215 weeks. Notably, the proposed model and its prediction averaging approach achieved the best performance among all participating methods in the 2022 ATD program challenge sponsored by NSF on the same data set.

1. INTRODUCTION

The research presented here is motivated by the data challenge held by the Algorithm for Threat Detection (ATD) program, sponsored by the National Science Foundation (NSF) that took place in 2022. The aim of the competition is to develop a forecaster for predicting national-level geopolitical event counts given historical data. We developed a meta predictive inference procedure, using a dynamic factor model for matrix-valued time series as the main model assisted with model averaging techniques. Our development is effective for highly heterogeneous count data, and it achieved the best performance on predicting new events in holdout testing data among all participating methods in the competition.

In the realm of political science and international relations, the integration of advanced computational techniques has revolutionized the analysis of global event data. The Global Database of Events, Language, and Tone (GDELT) stands as a comprehensive repository, encompassing diverse sources that provide a holistic view of international events and their associated metadata. Researchers have increasingly relied on GDELT as a pivotal resource for comprehending global political, social, and economic phenomena [5]. For instance, in their work, Galla and Burke (2018) leverages machine learning models to predict social unrest using GDELT data [4]. Their analysis highlights the importance of news articles' negative sentiment in predicting major civil unrest events. By examining news articles captured by GDELT, the study identifies social factors and events that precede large-scale unrest at both state and county levels. Similarly, [5] demonstrates how GDELT data can be used to track and forecast political instability across different regions of the world. Their work shows the potential of GDELT data for providing real-time insights into global political dynamics In addition to conflict prediction, GDELT data has been employed to understand political communication and media influence. [1] explores how media attention and public interest, similar to data captured by GDELT, influence financial markets. By examining the correlation between media coverage and stock market movements, they highlight the impact of global news on economic activities.

The GDELT data used in the competition contains weekly counts of 20 types of events from 260 countries and regions. For each week, the observations are naturally represented by a 260 × 20 matrix. The dataset is thus a times series of matrices. The recent development of matrix time series models [3, 8, 7] provides a valuable tool for modeling intricate interactions within each component of the time series. These models extend traditional univariate and vector time series approaches, enabling the capture of evolving dependencies between multiple variables over time.

There are two major approaches of modeling matrix and tensor time series. Employing a multi-linear autoregressive structure, one can effectively represent and analyze the temporal dynamics inherent in the matrix time series with forecasting capability [3]. On the other hand, matrix and tensor factor models offer a potent framework for dimension reduction and the extraction of latent variables from high-dimensional data, though it lacks the forecasting capability [8, 7]. The dynamic matrix factor model [9] combines the advantages of both approaches by employing the factor model to reduce the dimension, and specifying the autoregressive structure of the factor process to allow for predictions. Within the domain of time series analysis, dynamic factor models thus facilitate the discernment of underlying structures governing both cross-sectional and temporal patterns. The aforementioned models have found success in a spectrum of fields, including finance, economics, transportation and others. They are especially suited to account for the unique characteristics of multi-dimensional datasets like GDELT.

We explore the use of matrix time series models and model-averaging techniques in constructing a predictive model for the complex GDELT dataset in the competition.

1.1 The Dataset

The dataset under consideration is the GDELT (Global Dataset of Events, Language, and Tone) dataset, a public database that contains data on geopolitical events around the globe. It uses the CAMEO coding system to record the geopolitical events on where the they took place, the actors, sources and the types of events are recorded. The data used in this project is at a national level, and contains weekly aggregated event frequencies, based on event records found in broadcast, print, and web news sources around the world. There are 20 event codes/types (”protests”, ”threats”, ”providing aid”, ”engaging in diplomatic cooperation”, ”assaults”, etc) and 260 country/regions. The dataset spans from 2014 to 2018 with total 215 weekly observations. Specifically, a matrix of counts is observed every week: each row of the matrix corresponding to a country/region and each column an event type.

A major challenge of analyzing the GDELT data is the heterogeneity that exists among and within countries and events. The data exhibits a large and diverse variation not only within individual countries but also when comparing different countries to each other. The variability in the data is extensive and makes it a particularly challenging dataset to analyze and work with. Figures 1 and 2 are two snapshots of the data, showing the vast disparities within/between the countries with the highest frequencies, and those with the least frequencies.

The goal of the 2022 ATD Data Challenge is to create a multivariate forecaster capable of predicting national-level geopolitical event counts. Given past historical event counts, the goal is to predict the number of each region-event (i.e. every cell of the matrix) for the next k weeks. The host of the competition split the data into a training dataset (215 weekly data points) which was made available at the start to be used to develop and investigate the appropriate models, and a holdout dataset which was used to independently evaluate the performance of the forecasters.

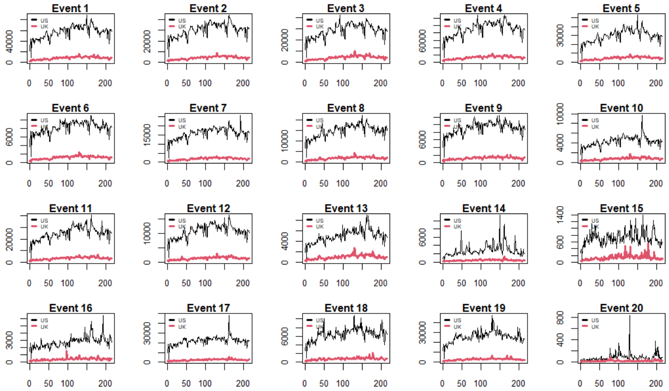

Figure 1: Time series plots of the 20 events[1] for US and UK, two countries with high event frequencies.

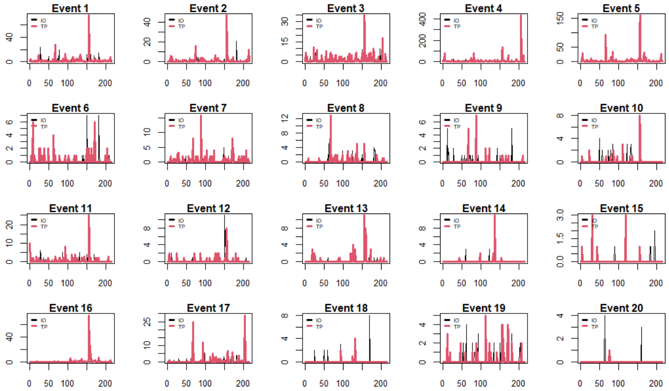

Figure 2: Plots of 20 events for British Indian Ocean Territory (IO) and East Timor (TP), two countries with low event frequencies.

[1] The events are: 1. Make Public Statement, 2. Appeal, 3. Express Intent to Cooperate, 4. Consult, 5. Engage in Diplomatic Cooperation, 6. Engage in Material Cooperation, 7. Provide Aid, 8. Yield, 9. Investigate, 10. Demand, 11. Disapprove, 12. Reject, 13. Threaten, 14.

Table 1. Illustration of Data in a matrix form

Event	AA	AC	AE	AF	AG	AJ	...	ZI
Make Public Statement	0	4	180	635	33	41	...	207
Appeal	0	4	72	252	27	2	...	108
⋮	⋮	⋮	⋮	⋮	⋱	⋮		
Use Unconventional Mass Violence	0	0	0	0	2	0	...	1

1.2 Data Preprocessing

The trend component in the proposed forecasting model plays a critical role in capturing the underlying patterns or long-term trends present in the geopolitical event data. We use the exponential smoothing method to estimate the trend for each individual component time series. Specifically, the trend component $\boldsymbol{V}_t = (V_{t,ij})$ is calculated for each region-event pair (i, j) using:

$$V_{t,ij} = \alpha X_{t,ij} + (1 - \alpha) V_{t-1,ij},$$

where: $V_{t,ij}$ represents the trend component for country i and event j at time t, $X_{t,ij}$ is the observed count of the (i, j)-th pair at time t, and $0 \leq \alpha \leq 1$ is the smoothing parameter, which determines the weight assigned to the current observation compared to the previous trend value. A smaller α gives more weight to historical observations, leading to a smoother trend estimation.

The trend component $\boldsymbol{T}_t$ is crucial in separating the long- and short-term fluctuations of the data, enabling the model to focus on capturing the underlying dynamics in the stationary component for more accurate forecasting. The exponential smoothing parameter is set to 0.125 based on empirical considerations in this analysis. The detrended data will be modeled as detailed in the next section.

2. METHODOLOGY

Let $\boldsymbol{X}_t$ be the observed (possibly detrended) data matrix at time t of dimensions $d_1 \times d_2$, where d_1 and d_2 are the number of countries/regions and events respectively, $t = 1, 2, \ldots, T$,

$$\boldsymbol{X}_t = \begin{bmatrix} X_{t,11} & \ldots & X_{t,1d_2} \\ \vdots & \ddots & \vdots \\ X_{t,d_1 1} & \ldots & X_{t,d_1 d_2} \end{bmatrix}.$$

A naive approach to modeling this data is fitting each individual series using a univariate time series model, or fitting a Vector Autoregressive model (VAR) to each country (the counts over different events become a 20-dimensional vector). These approaches ignore the complex correlation across the events and/or regions hence is less accurate. We employ a model that preserves the matrix structure of the data to improve intepretability and predictability.

2.1 Dynamic Matrix Factor Model

We use the Dynamic Matrix Factor Model (DMFM) proposed in [9]. It combines matrix factor model for dimension reduction and the matrix AR model for incorporating the temporal dynamics into the factor process. This approach allows us to produce accurate predictions for high dimensional matrix time series data.

The underlying premise of the DMFM is rooted in the idea that a select number of unobserved dynamic factors are the primary drivers behind the observed co-movements in the matrix time series. These latent factors, in turn, follow a time series process, specifically represented by the Matrix Autoregressive (MAR) model. The primary motivation for embracing the DMFM lies in its efficacy, enabling us to harness information from the entire d_1 by d_2 variable matrix while leveraging only r_1 by r_2 factors.

For the time series data $\boldsymbol{X}_t$, where $t = 1, \ldots, T$, the DMFM is characterized by:

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{Q}_1 \boldsymbol{F}_t \boldsymbol{Q}_2' + \boldsymbol{E}_t \\ \boldsymbol{F}_t &= \boldsymbol{A}_1 \boldsymbol{F}_{t-1} \boldsymbol{A}_2' + \boldsymbol{Z}_t, \end{aligned} \tag{2.1}$$

where $\boldsymbol{F}_t$ is an $r_1 \times r_2$ unobserved matrix of common fundamental factors; $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ denote the $d_1 \times r_1$ and $d_2 \times r_2$ front and back loading matrices respectively; $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are time series coefficient matrices of dimensions $r_1 \times r_1$ and $r_2 \times r_2$ respectively; $\boldsymbol{E}_t = (E_{t,ij})$ is a $d_1 \times d_2$ error matrix and $\boldsymbol{Z}_t = (Z_{t,ij})$ is a $r_1 \times r_2$ error matrix. We assume $\boldsymbol{E}_t$ and $\boldsymbol{Z}_t$ are independent matrix white noise processes.

This model framework allows us to simultaneously capture the temporal dependencies inherent in the factor structure and the interactions among the observed variables, enhancing our ability to make informed inferences about the underlying dynamics of the system.

2.2 Parameter Estimation

Following [9], the parameters in model (2.1) can be estimated in two steps. The first step involves the estimation of the coefficients of the factor model, using the procedure in [8]. The core concept of the approach involves computing the autocovariance of the time series data, followed by the construction of a matrix-form Box-Ljung type statistic. In the context of the matrix factor model and the assumption of white idiosyncratic noise, the loading matrices are directly connected to the space defined by such a matrix.

Protest, 15. Exhibit Force Posture, 16. Reduce Relations, 17. Coerce, 18. Assault, 19. Fight, 20. Use Unconventional Mass Violence.

Specifically, we adopt the estimation procedure of [8] by using only the lag-1 sample autocovariance matrices. Let

$$\hat{\boldsymbol{M}}_1 = \frac{1}{T-1}\sum_{t=1}^{T-1} \boldsymbol{X}_t \boldsymbol{X}_{t+1}'. \tag{2.2}$$

To estimate the column space of $\boldsymbol{Q}_1$, denoted $\mathcal{M}(\boldsymbol{Q}_1)$, we use $\hat{\boldsymbol{Q}}_1 = \{\hat{\boldsymbol{q}}_{1,1}, \ldots, \hat{\boldsymbol{q}}_{r_1,1}\}$ consisting of the r_1 left singular vectors of $\hat{\boldsymbol{M}}_1$ corresponding to its leading singular values. Similarly, the estimation of $\boldsymbol{Q}_2$ involves applying the same procedure to the transposes of $\boldsymbol{X}_t$.

Given the estimates of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, it follows that

$$\hat{\boldsymbol{F}}_t = \hat{\boldsymbol{Q}}_1' \boldsymbol{X}_t \hat{\boldsymbol{Q}}_2.$$

The next step is to estimate the coefficients of the matrix AR component of the model (the second part of (2.1).) Here, we follow the estimation procedure in [3]. Assuming the covariance matrix of the error matrix $\boldsymbol{Z}_t$ takes the structure

$$\mathrm{Cov}(\mathrm{vec}(\boldsymbol{Z}_t)) = \Sigma_c \otimes \Sigma_r,$$

the log likelihood under normality is

$$\begin{aligned} &- r_1(T-1)\log|\Sigma_c| - r_2(T-1)\log|\Sigma_r| \\ &- \sum_t \mathrm{tr}(\Sigma_r^{-1}(\boldsymbol{F}_t - \boldsymbol{A}_1\boldsymbol{F}_{t-1}\boldsymbol{A}_2')\Sigma_c^{-1}(\boldsymbol{F}_t - \boldsymbol{A}_2\boldsymbol{F}_{t-1}\boldsymbol{A}_2')'). \end{aligned}$$

Ignoring the estimation error in $\hat{\boldsymbol{F}}_t$ and plugging them into the likelihood function, the MLE can be found by iteratively updating one of the following, while keeping the other three fixed:

$$\boldsymbol{A} \leftarrow \left(\sum_t \hat{\boldsymbol{F}}_t \Sigma_c^{-1} \boldsymbol{B} \boldsymbol{F}_{t-1}'\right)\left(\sum_t \hat{\boldsymbol{F}}_{t-1} \boldsymbol{B}' \Sigma_c^{-1} \boldsymbol{B} \hat{\boldsymbol{F}}_{t-1}'\right)^{-1},$$

$$\boldsymbol{B} \leftarrow \left(\sum_t \hat{\boldsymbol{F}}_t' \Sigma_c^{-1} \boldsymbol{A} \hat{\boldsymbol{F}}_{t-1}\right)\left(\sum_t \hat{\boldsymbol{F}}_{t-1}' \boldsymbol{A}' \Sigma_c^{-1} \boldsymbol{A} \hat{\boldsymbol{F}}_{t-1}\right)^{-1},$$

$$\Sigma_c \leftarrow \frac{\sum_t \boldsymbol{R}_t' \Sigma_r^{-1} \boldsymbol{R}_t}{r_1(T-1)},$$

$$\Sigma_r \leftarrow \frac{\sum_t \boldsymbol{R}_t \Sigma_c^{-1} \boldsymbol{R}_t'}{r_2(T-1)},$$

where $\boldsymbol{R}_t = \hat{\boldsymbol{F}}_t - \boldsymbol{A}\hat{\boldsymbol{F}}_{t-1}\boldsymbol{B}'$.

2.3 Prediction under DMFM

Prediction is carried out in a straightforward manner. For the h-step ahead prediction, we initially forecast $\tilde{\boldsymbol{F}}_t(h)$ using the MAR model, and subsequently incorporate it into the factor model to obtain $\tilde{\boldsymbol{X}}_t(h)$. Specifically, let $\mathcal{F}_t$ be the σ-field generated by $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\}$. Under the model:

$$\mathbb{E}(\boldsymbol{X}_{t+1}|\mathcal{F}_t) = \boldsymbol{Q}_1 \mathbb{E}(\boldsymbol{F}_{t+1}|\mathcal{F}_t)\boldsymbol{Q}_2'.$$

where $\mathbb{E}(\boldsymbol{F}_{t+1}|\mathcal{F}_t) = \boldsymbol{A}_1\boldsymbol{F}_t\boldsymbol{A}_2'$. Again, ignoring the estimation error in $\hat{\boldsymbol{F}}_t$, then the one step ahead prediction is

$$\begin{aligned} \tilde{\boldsymbol{X}}_t(1) &= \boldsymbol{Q}_1 \tilde{\boldsymbol{F}}_t(1) \boldsymbol{Q}_2' \\ &= \boldsymbol{Q}_1 \boldsymbol{A}_1 \hat{\boldsymbol{F}}_g \boldsymbol{A}_2' \boldsymbol{Q}_2' \\ &= \boldsymbol{Q}_1 \boldsymbol{A}_1 \boldsymbol{Q}_1' \mathbf{X}_t \boldsymbol{Q}_2 \boldsymbol{A}_2' \boldsymbol{Q}_2'. \end{aligned}$$

Similarly, the h-step ahead predictor is given by

$$\tilde{\boldsymbol{X}}_t(h) = \boldsymbol{Q}_1 \tilde{\boldsymbol{F}}_t(h) \boldsymbol{Q}_2',$$

where $\tilde{\boldsymbol{F}}_t(h) = \boldsymbol{A}_1 \tilde{\boldsymbol{F}}_t(h-1) \boldsymbol{A}_2'$.

2.4 Additional Models

The GDELT dataset displays significant heterogeneity, as countries and regions vary greatly in size, and the frequency of events can differ substantially by the nature of their types, as seen in Figures 1 and 2. It is therefore necessary to incorporate alternative models since a single DMFM model may not be able to capture and account for the nuanced patterns and complexities inherent in the data. By considering a range of simpler models, we aim to identify the special series that cannot be modeled by DMFM in a unified model, especially those series with extremely low counts. This approach allows us to gain a deeper understanding of the intricate dynamics at play, ultimately leading to more accurate and meaningful predictions.

Specifically, three additional prediction methods are considered, based on 1) a pure exponential smoothing, 2) an univeraite ARIMA(1,1,1) model and 3) a trivial prediction (always predict the future value by zero). The main reason for incorporating the trivial prediction is due to the fact that a substantial number of countries have zero event counts for some of the 20 events, as shown in Figure 3. The ARIMA(1,1,1) model is particularly effective for capturing the possible trend and more complex temporal dependence for an individual series. It is an essential and simple tool for forecasting in time series literature [2].

2.5 Prediction Averaging

In our ensemble approach, a weighted averaging prediction scheme is implemented to enhance the accuracy by capturing the different strengths of the different prediction methods. The weights used for averaging are based on the performance of each prediction methods. Specifically, they are determined by the inverse of the Mean Absolute Scaled Error (MASE, defined in Section 3.2) for each prediction method. The MASE measures the forecast accuracy, and method with a lower MASE value is assigned a higher weight. The weights are normalized to ensure that they sum

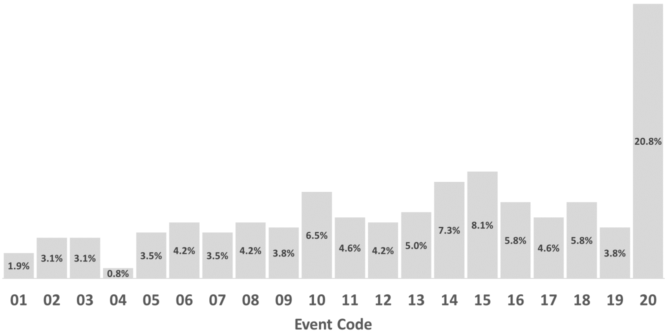

Figure 3: The proportion of countries with zero frequency for each event.

to 1. Specifically, for each model m and horizon $h = 1, 2, 3, 4$, let $E_{m,ij,T}(h)$ represent the rolling MASE of calculated over a window of length 100 till the end of the training dataset for country i and event j, the weight $w_{m,ij,T}(h)$ is determined as:

$$w_{m,ij,T}(h) = \frac{[E_{m,ij,T}(h)]^{-1}}{\sum_{m=1}^{M}[E_{m,ij,T}(h)]^{-1}}.$$

Subsequently, we obtain the final weighted predicted value of $X_{ij,T+h}$ as:

$$\hat{X}_{ij,T}(h) = \sum_{m=1}^{M} w_{m,ij,T}(h)\tilde{X}_{ij,T}^{(m)}(h). \tag{2.3}$$

This aggregation technique leverages the strengths of each individual prediction method, resulting in an ensemble prediction that often outperforms any single model in isolation.

3. MODEL RESULTS & DISCUSSION

3.1 The algorithm

Algorithm 1 outlines the sequential steps taken to fit the data and compute the h-step ahead prediction. In the case of the Dynamic Matrix Factor Model (DMFM), after prediction, the trend components are reintegrated to yield a forecast of the original series.

Algorithm 1: Proposed Forecasting Algorithm

```
Data: GDELT data in matrix form: $\boldsymbol{X}_{t,ij}, t = 1, \dots, T$
Result: $h$-step-ahead prediction: $\boldsymbol{X}_T(h), h = 1, 2, \dots$
for each pair $(i, j)$ do
    Estimate trend $\boldsymbol{V}_t$ as in Section (1.2).
    Let $\boldsymbol{X}_t^* = \boldsymbol{X}_t - \boldsymbol{V}_t$.
    Obtain trend prediction $\boldsymbol{V}_T(h) = \boldsymbol{V}_T$ for $h = 1, 2, \dots$
Estimate $\hat{\boldsymbol{Q}}_1$ and $\hat{\boldsymbol{Q}}_2$ as in Section (2.2) using the
  detrended data $\boldsymbol{X}_t^*$
Estimate $\hat{\boldsymbol{F}}_t = \hat{\boldsymbol{Q}}_1' \boldsymbol{X}_t^* \hat{\boldsymbol{Q}}_2$
Estimate time series coefficient matrices $\hat{\boldsymbol{A}}_1$ and $\hat{\boldsymbol{A}}_2$
for $h = 1, 2, \dots$ do
    Calculate $\tilde{\boldsymbol{F}}_T(h) = \hat{\boldsymbol{A}}_1 \hat{\boldsymbol{F}}_T(h-1) \hat{\boldsymbol{A}}_2'$
    Calculate $\tilde{\boldsymbol{X}}_T(h) = \hat{\boldsymbol{Q}}_1 \tilde{\boldsymbol{F}}_T(h) \hat{\boldsymbol{Q}}_2'$
    Let $\hat{\boldsymbol{X}}_T^{(1)}(h) = \tilde{\boldsymbol{X}}_T(h) + \boldsymbol{V}_T(h)$
for each country, $i = 1, \dots, d_1$ do
    for each event, $j = 1, \dots, d_2$ do
        Fit ARIMA(1,1,1) on the original data $\boldsymbol{X}_t$
        Obtain ARIMA model $h$-step ahead predictions
          $\hat{\boldsymbol{X}}_T^{(2)}(h)$
for $h = 1, 2, \dots$ do
    Let $\hat{\boldsymbol{X}}_T^{(3)}(h) = \boldsymbol{V}_T(h)$
for $h = 1, 2, \dots$ do
    Let $\hat{\boldsymbol{X}}_T^{(4)}(h) = 0$
for $h = 1, 2, \dots$ do
    Calculate the MASE for each prediction method
    Calculate model weights as in Section (2.5)
    Obtain the prediction $\hat{\boldsymbol{X}}_T(h)$ using the weighted
      prediction average (2.3).
```

3.2 Performance Metrics

The predictions are evaluated using two metrics adopted by the organizers of the ATD challenge. Specifically, predictions are evaluated using the following two standard metrics and the performance is compared against baseline models on a holdout dataset. We express their definitions in a slightly more general form, by allowing the metrics to be evaluated over any time window from $t_0 + 1$ to t_1, while the

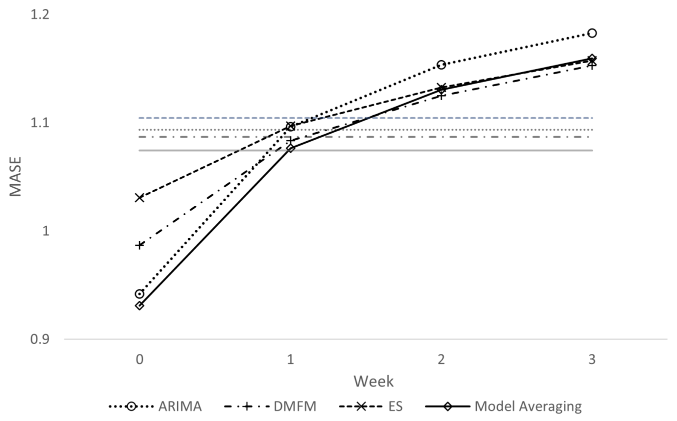

Figure 4: Prediction performance comparison. The horizontal lines represent the average performance of each model.

errors shown in Figure 4 are calculated on the entire holdout dataset.

- Mean Absolute Scaled Error (MASE)

$$\mathrm{MASE}(h) = \frac{\sum_{t=t_0+1}^{t_1} \sum_{ij} |\boldsymbol{X}_{ij,t+h} - \hat{\boldsymbol{X}}_{ij,t}(h)|}{\sum_{t=t_0+1}^{t_1} \sum_{ij} |\boldsymbol{X}_{ij,t+h} - \boldsymbol{X}_{ij,t}|}.$$

- Root Mean Square Error (RMSE)

$$\mathrm{RMSE}(h) = \sqrt{\frac{1}{(t_1 - t_0) d_1 d_2} \sum_{t=t_0+1}^{t_1} \sum_{ij} (\boldsymbol{X}_{ij,t+h} - \hat{\boldsymbol{X}}_{ij,t}(h))^2}.$$

3.3 Results

Comparing the predictive performances of different methods, we note that no single method outperforms others in all horizons. For one-step ahead prediction, the ARIMA model performs the best. The DMFM outperforms all other models in two- to four-week ahead prediction, and consequently has the best overall performance. The exponential smoothing method outperforms the ARIMA model in longer prediction horizons and is close to the performance of the DMFM for $h \geq 2$. These observations provide a strong motivation to consider prediction averaging approach in order to capitalize on the strengths of each method while mitigating their weaknesses.

Figure 4 shows the overall out-of-sample rolling prediction MASE of different prediction methods. Clearly the prediction averaging approach provides the best overall prediction performance. For longer term prediction, DMFM actually performs slightly better than the model averaging approach, though the model averaging approach automatically adapts to different prediction horizons. The constant lines show the average prediction MASE over the four prediction horizons.

In the broader competition context, a holdout dataset, mirroring the size of the data used for modeling, is employed to assess the algorithm's comprehensive predictive capabilities. Notably, the algorithm detailed in this paper demonstrates markedly superior performance compared to numerous benchmark models that are put forth. Table 2 is a snapshot of the performance of the final model as it compares to the baseline models developed by the competition organizers. MASA and RMSE are defined above. SPEC is Stock-keeping-oriented Prediction Error Costs [6] and "Column Wins" shows the number of times a prediction method has the best MASE and RMSE for each individual country/event averaged over time. Results from other competitors are not shown. DMFM is the winner in all categories.

Team	MASE	RMSE	SPEC	# Col Wins
DMFM	0.92	50.06	9117.64	2981
TFT[2]	0.98	52.94	25553.63	1428
EWMA[3]	1.01	53.42	2485.89	694
DeepAR[4]	1.02	54.42	15916.29	406
Croston[5]	1.02	53.68	10574.11	1107
PredictLast[6]	1.06	57.51	2090.55	809
NBEATS[7]	1.08	57.71	133555.47	720
PredictMean[8]	1.88	77.71	135313.43	1210

Table 2. Leaderboard of model performance under different metrics. Models other than 'DMFM' represent baseline models that were presented by competition organizers.

4. CONCLUSION

In this paper we use a prediction averaging approach to obtain accurate prediction of high dimensional geopolitical event counts, based on the dynamic matrix factor model as the main method, and several other simpler prediction methods as supplement for dealing with heterogeneity in the data. For building the DMFM model, a data preprocessing step is used to remove the underlying trends in the time series before fitting the DMFM. This procedure plays a crucial role in ensuring that the DMFM is based on stationary data, which in turn enhances the accuracy and reliability of our forecasts. DMFM uses a matrix factor model for effective dimension reduction and a matrix AR model on the latent factors to capture temporal dynamics and to introduce prediction capability. The model maintains the matrix form of the data hence retains the crucial information on

[2] Temporal Fusion Transformer - Deep learning model integrating historical data dynamically.
[3] Exponential Weighted Moving Average with emphasis on recent observations.
[4] Probabilistic forecasting with autoregressive recurrent networks.
[5] Forecasting method for intermittent demand.
[6] Uses the last observed value as the forecast.
[7] Neural network approach for interpretable time series forecasting.
[8] Uses the mean of historical values for forecasting.

row classification (the countries) and column classification (the event types). It allows effective modeling of the complex inter-relationship among the event types and countries, while maintaining model simplicity and interpretability. The estimation follows a two-step procedure which is simple and fast, without complex optimization procedures. In addition to the primary model, we explore the inclusion of simpler models to account for the heterogeneity within the dataset. A simple prediction average approach is used. This comprehensive strategy allows us to harness the strengths of various prediction methods, ensuring a robust and accurate forecasting process that accommodates the intricacies of the data.

A significant avenue for future research lies in the investigation of dynamic matrix factor models tailored for count time series data. The model employed in this study operates under the assumption of continuous Gaussian data, which does not always hold true. A specialized model designed to accommodate positive integer data has the potential to yield more accurate forecasts and insights. This area of exploration could lead to significant advancements in modeling techniques for count-based time series data.

REFERENCES

[1] Alanyali, M., Moat, H. S. and Preis, T. (2016). Quantifying the Relationship Between Financial News and the Stock Market. *Journal of Economic Interaction and Coordination* **11**(1) 67–87. https://doi.org/10.1007/s11403-014-0151-6.

[2] Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis, Forecasting and Control*. Holden Day: San Francisco.

[3] Chen, R., Xiao, H. and Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics* **222**(1, Part B) 539–560. https://doi.org/10.1016/j.jeconom.2020.07.015.

[4] Galla, D. and Burke, J. (2018). Predicting Social Unrest Using GDELT. In *MLDM 2018: Machine Learning and Data Mining in Pattern Recognition* 103–116. Springer. https://doi.org/10.1007/978-3-319-96133-0_8. https://doi.org/10.1007/978-3-319-96133-0_8.

[5] Leetaru, K. and Schrodt, P. A. (2013). GDELT: Global data on events, location, and tone. *ISA Annual Convention*.

[6] Martin, D., Spitzer, P. and Kühl, N. (2020). A New Metric for Lumpy and Intermittent Demand Forecasts: Stock-keeping-oriented Prediction Error Costs. *CoRR* **abs/2004.10537**. 2004.10537.

[7] Rong Chen, D. Y. and Zhang, C. -H. (2022). Factor Models for High-Dimensional Tensor Time Series. *Journal of the American Statistical Association* **117**(537) 94–116. https://doi.org/10.1080/01621459.2021.1912757. https://doi.org/10.1080/01621459.2021.1912757.

[8] Wang, D., Liu, X. and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics* **208**(1) 231–248. https://doi.org/10.1016/j.jeconom.2018.09.013.

[9] Yu, R., Chen, R., Xiao, H. and Han, Y. (2024). Dynamic Matrix Factor Models for High Dimensional Time Series. *arXiv preprint arXiv:2407.05624*.

Ama Ampadu-Kissi-Owusu. Department of Statistics, Rutgers University, USA.
E-mail address: aka117@stat.rutgers.edu

Rong Chen. Department of Statistics, Rutgers University, USA.
E-mail address: rongchen@stat.rutgers.edu

Kathyrina Dowlin. Rutgers University, USA.
E-mail address: ked167@scarletmail.rutgers.edu

Yuefeng Han. Department of Applied and Computational Mathematics and Statistics, Notre Dame University, USA.
E-mail address: yuefeng.han@nd.edu

Junyi Li. Department of Statistics, Rutgers University, USA.
E-mail address: jl2707@stat.rutgers.edu

Zebang Li. Department of Statistics, Rutgers University, USA.
E-mail address: zl326@stat.rutgers.edu

Ruofei Mao. University of California, San Diego, USA.
E-mail address: alexmrf6439@gmail.com

Han Xiao. Department of Statistics, Rutgers University, USA.
E-mail address: hxiao@stat.rutgers.edu

Minge Xie. Department of Statistics, Rutgers University, USA.
E-mail address: mxie@stat.rutgers.edu

Ruofan Yu. Department of Statistics, Rutgers University, USA.
E-mail address: ruofanyu@stat.rutgers.edu

Zheshi Zheng. Department of Biostatistics, University of Michigan, USA.
E-mail address: zszheng@umich.edu