Terrorist Attacks Time Series and Spatial Analysis

Yuan Gao, Qi Wang, Yizheng Wang, Darien Zhang

4/21/2017

Introduction

This project aims to explore and analyze the dataset from Global Terrorism Database that contains details on every terrorist activity since 1970. There are two parts of the project. The first part is creating time series models for aggregated count data for terrorist activities in each region of the world. We aggregated the incidents to yearly and monthly basis count data and built ARIMA models with seasonal trends to forecast the number of attacks in the future years. The second part focuses on the spatial pattern of these terroristic attacks in Iraq, which is a country that suffered the most from terrorism: over 18,000 attacks took place in Iraq from 1970 to 2015. In addition, we incorporated the civilian deaths dataset from the Iraq Body Count website, which could be used to conduct inference on the number of attacks and the total civilian deaths from violence. In order to capture the spatial patterns, several spatial models including SAR, CAR and their corresponding bayesian version models have been implemented. In the GTD dataset, there is a categorical variable which is called `success`. Success of a terrorist strike is defined with respect to the tangible effects of the attack. Based on this variable, we also fit a spline model which can be used to predict which areas of Iraq have high success rate in terrorist attacks.

Data

The Global Terrorism Database (GTD) is maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START). The database collects from media articles and electronic archives, and to a lesser extent, existing data sets, secondary source materials such as books and journals, and legal documents. Each row represents one terrorist incident, and the variables include incident date, incident information, incident location, attack information, weapon information, target information, and perpetrator information. There are 156,772 incidents ranging from 1970 to 2015. The GTD defines a terrorist attack as the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation. In order to consider an incident for inclusion in the GTD, all three of the following attributes must be present: the incident must be intentional, the incident must entail some level of violence or immediate threat of violence, and the perpetrators of the incidents must be sub-national actors. Incidents occurring in both the same geographic and temporal point will be regarded as a single incident, but if either the time of occurrence of incidents or their locations are discontinuous, the events will be regarded as separate incidents.

We also incorporated another source of data for the spatial analysis from Iraq Body Count. Iraq Body Count (IBC) maintains the world's largest public database of violent civilian deaths. Its data drawn from cross-checked media reports, hospital, morgue, NGO and official figures or records. IBC can allow user to specify the number of civilian deaths for each province. Hence, we are able to cumulatively sum the number of civilian deaths for each province of Iraq and use these data to fit models.

Methodology

Time Series Analysis

The AR(1) process is

$$AR(1): \quad y_t = \delta + \phi\, y_{t-1} + w_t$$

A moving average process is similar to an AR process, except that the autoregression is on the error term.

$$MA(1): \qquad y_t = \delta + w_t + \theta\, w_{t-1}$$

An ARMA model is a composite of AR and MA processes,

$$\begin{aligned} ARMA(p,q): & \\ & y_t = \delta + \phi_1\, y_{t-1} + \cdots \phi_p\, y_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t_q} \\ & \phi_p(L) y_t = \delta + \theta_q(L) w_t \end{aligned}$$

We used Autoregressive integrated moving average (ARIMA) to build our models, which is an extension of an $ARMA$ model to include differencing of degree d to y_t before including the autoregressive and moving average components. Differencing is $\Delta y_t = y_t - y_{t-1}$, and $\Delta^d\, y_t$ is repeated applications of this operator. [1]

$$ARIMA(p,d,q): \qquad \phi_p(L)\ \Delta^d\, y_t = \delta + \theta_q(L) w_t$$

Spatial Analysis

Originated from time series analysis, *Simultaneous Autoregressive (SAR)* model and *Conditional Autoregressive (CAR)* model have been applied and extended by different fields such as econometrics, geography and medical statistics. Analogy to *Autoregressive model*, the CAR and SAR models include spatial neighboring observations. Here we fit the SAR and CAR model by both Frequentist and Bayesian way to discover spatial patterns of terroristic attacks in Iraq.

Similar to kernel regression and k-nearest-neighbors regression, *Smoothing Spline* model is able to flexibly estimate underlying regression function $f(x)$. We used *Smoothing Spline* model to discover how "successful terroristic attack" distributes in Iraq.

SAR and CAR

For every spatial data point s, *Simultaneous Autoregressive* assumes:

$$y(s) = \phi \sum_{s'} \frac{W_{s\,s'}}{W_{s\,\cdot}} y(s') + \epsilon$$

$$y \sim \mathcal{N}(0,\ \sigma^2\,((I - \phi W)^{-1})((I - \phi W)^{-1})^t)$$

Conditional Autoregressive assumes:

$$y(s)|y_{-s} \sim \mathcal{N}\left(\sum_{s'} \frac{W_{s\,s'}}{W_{s\,\cdot}} y(s'),\ \sigma^2\right)$$

$$y \sim \mathcal{N}(0,\ \sigma^2\,(I - \phi W)^{-1})$$

Where W is weight matrices and σ^2 is variance.

Spline

Spline provides a flexible way of estimating the underlying regression function $r(x) = E(Y|X = x)$. It estimates values using a mathematical function that minimizes overall surface curvature. This results in a smooth surface that passes exactly through the input points. In the spatial context, it can predict ridges and valleys in the data and is the best method for representing the smoothly varying surfaces(Childs, 2014). The general spline formula is as follows:

$$S(x, y) = T(x, y) + \sum_{j=1}^{N} \lambda_j R(r_j),$$

where N is the number of points, λ is the coefficients, and r_j is the distance from the point (x,y) to the j^{th} point. $T(x, y)$ and $R(r)$ represent tension option and regularized option, respectively. They are defined differently and really depend on the specific context.

Result

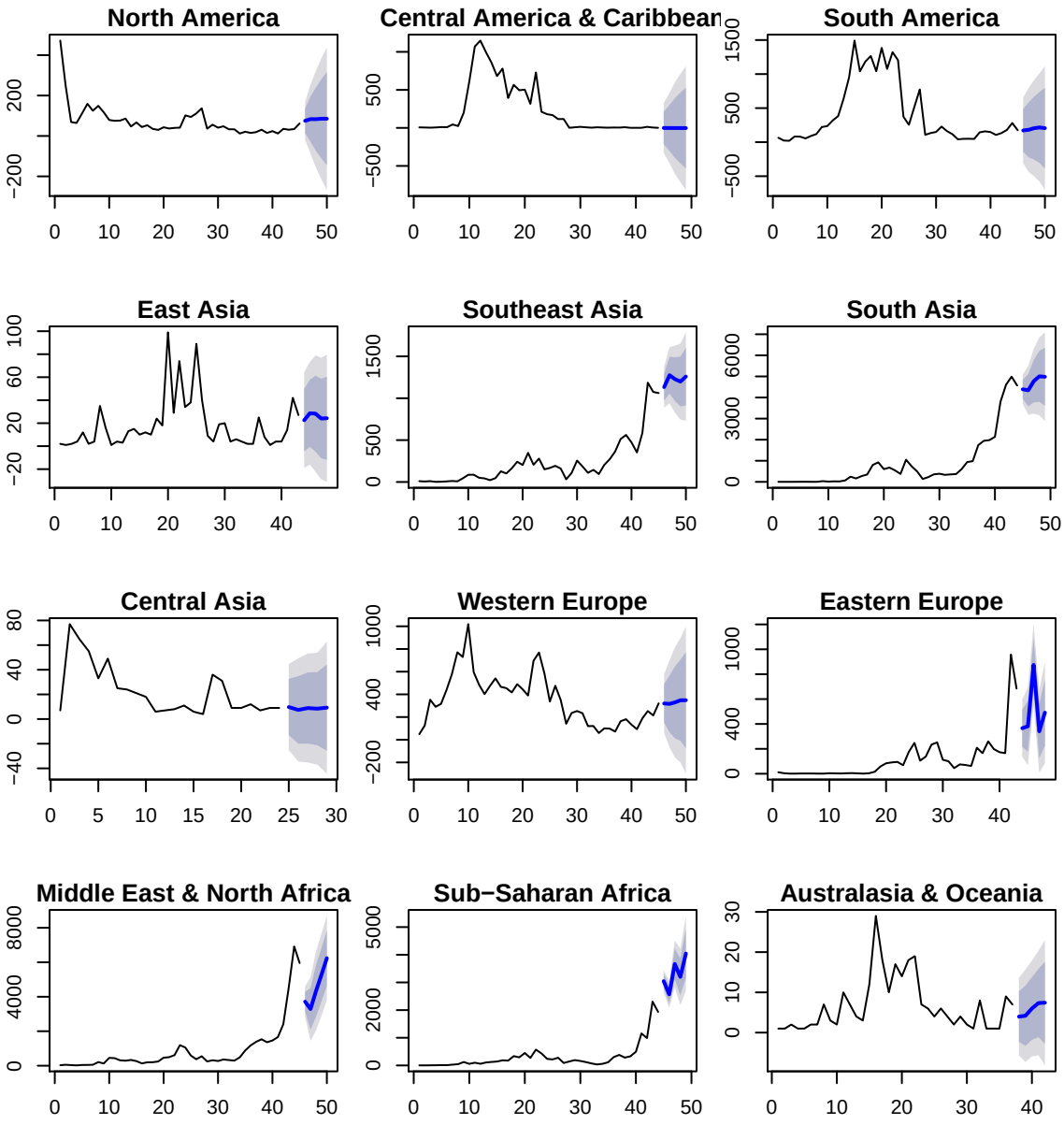
Part I: Regional Time Series Analysis

The first part of the project is to build ARIMA models for all twelve regions on aggregated counts of the incidents and perform forecasts for the next several years. We performed the analysis on both yearly and monthly aggregated count data.

Looking at the plots of the original data in Appendix, we can see that there are obvious drifts in the original time series. For fulfilling stationarity, we have visualize the difference of the original data. After differencing, data does seem to appear to have a mean at 0. Corresponding ACF and PACF are also plotted for modeling. Periodograms are plotted, and they can be used to identify the dominant periods (or frequencies) of a time series. Periodogram is a helpful tool in spectral analysis to detect cyclical behavior. Typically the time series are observed in time domain, but it can also be transformed to the frequency domain by Fourier Transformation. If there is significant peak at certain frequency, there may be significant periodicity in the series.

Based on observing the ACF, PACF and periodogram, by grid searching on a small range, here we have presented the best models. For the models based on yearly data, they turn out to have relatively low AIC values and the residuals are not autocorrelated. Thus, the models are good enough for us to use for forecast. For the monthly data, the residuals are not behaving as nicely looking at the acf and pacfs, and further tuning or more complicated models could be used to enhance the performance in the future. All of these plots can be found in the Appendix.

The plot below shows the yearly forecasted trends for the regions, and they seem to match up with our expectations of the number of terrorist attacks in the region. For instance, due to influx of refugees to Western Europe and complexity in the international politics of the region, an increase in the number of terrorist incidents could be expected in Western Europe as indicated by the model. Since we do not have ground truth on the actual terrorist attack counts in the future, we cannot validate our prediction. We also need to keep in mind that terrorist attacks have complicated motives and are triggered by many elements and the models are just a quick look at the trends to provide some insights in the matter.



Yearly Forecast

Part II: Iraq Spatial Analysis

Spatial Autocorrelation

Before we apply any statistical model to detect spatial patterns of terroristic attacks in Iraq, we use *Moran's I* and *Geary's C* to measure spatial autocorrelation of terroristic attack observation in Iraq.

Values of Moran's I range from -1 to $+1$. Negative values indicate negative spatial autocorrelation and positive values indicate positive spatial autocorrelation. A zero value indicates a random spatial pattern.

The value of Geary's C lies between 0 and 2. 1 means no spatial autocorrelation. Values lower than 1 demonstrate increasing positive spatial autocorrelation, whilst values higher than 1 illustrate increasing negative spatial autocorrelation.

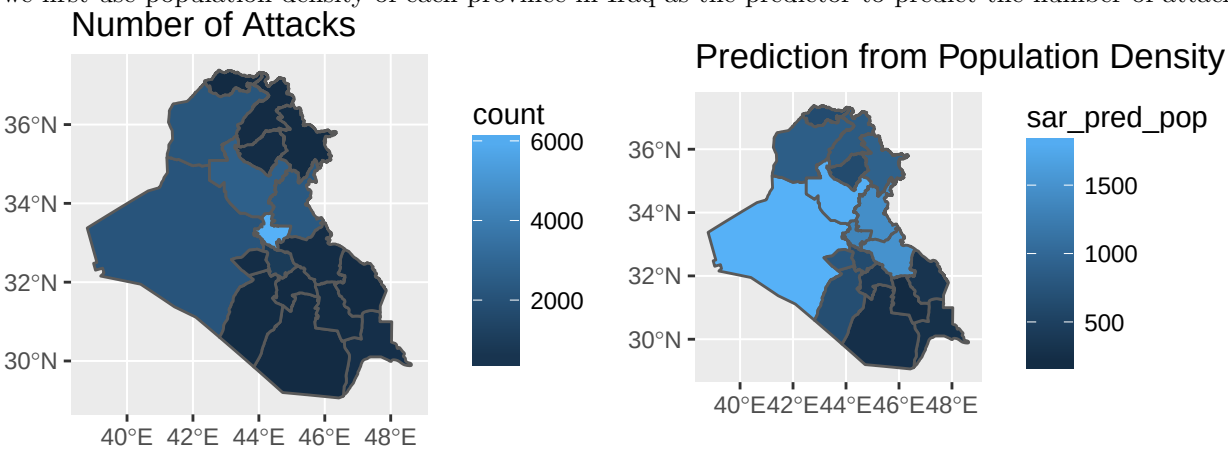
Table 1: Measurement of Spatial Autocorrelation

Measurement	Value
Moran's I	0.27
Geary's C	0.77

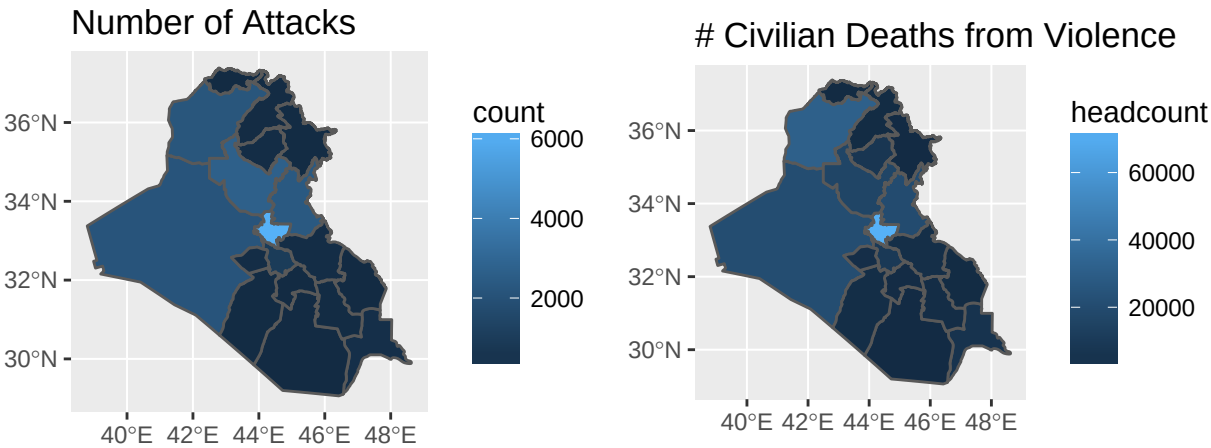
In our Iraq terroristic attack dataset, we calculate its Moran's I and Geary's C. Table.1 shows there is spatial pattern among the observations. Therefore, we decide to fit the *CAR*, *SAR* and *Spline* models to explore terroristic attacks in Iraq.

Sar model with population density as predictor

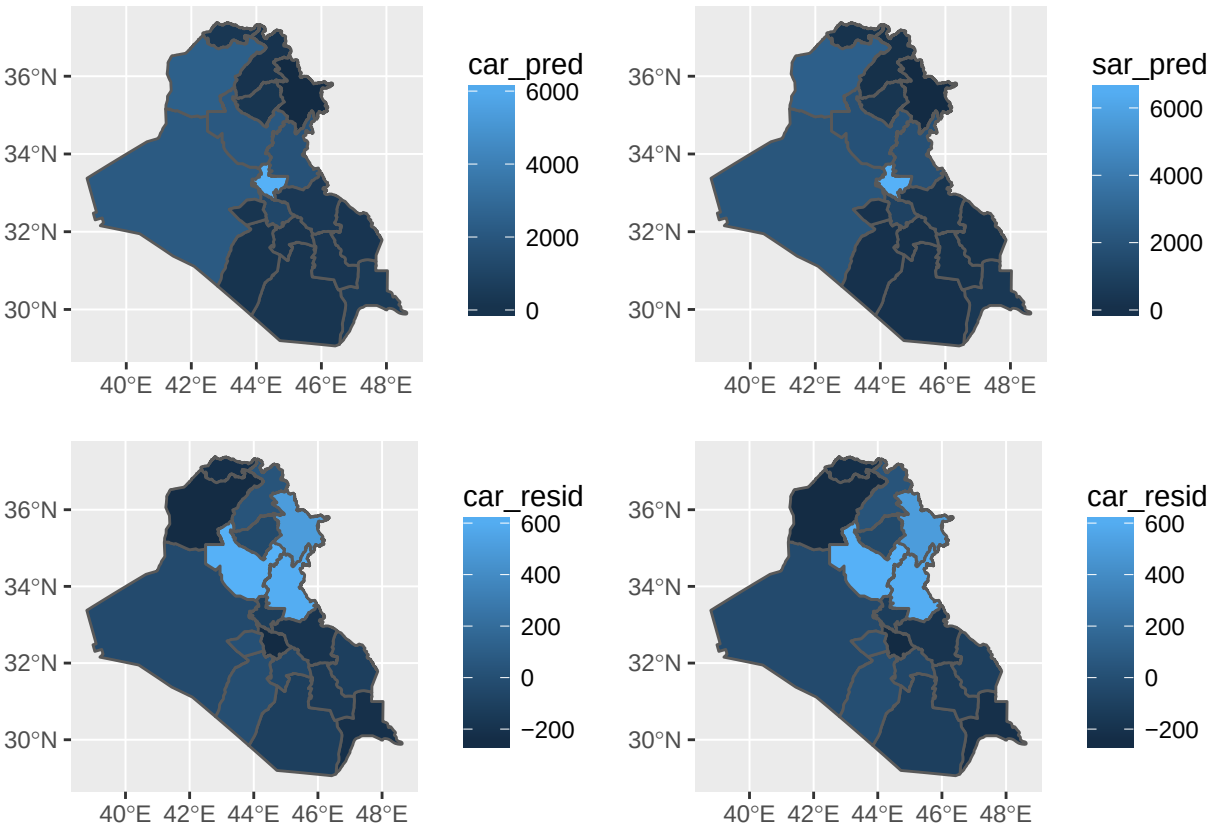
In the GTD dataset, most of the columns are categorical variables, which can not be used as predictors in this context. Based on common knowledge, we speculate that most of the terroristic attacks took place in high population density area in order to reach their goal: attracting more attention from others. Therefore, we first use population density of each province in Iraq as the predictor to predict the number of attacks.



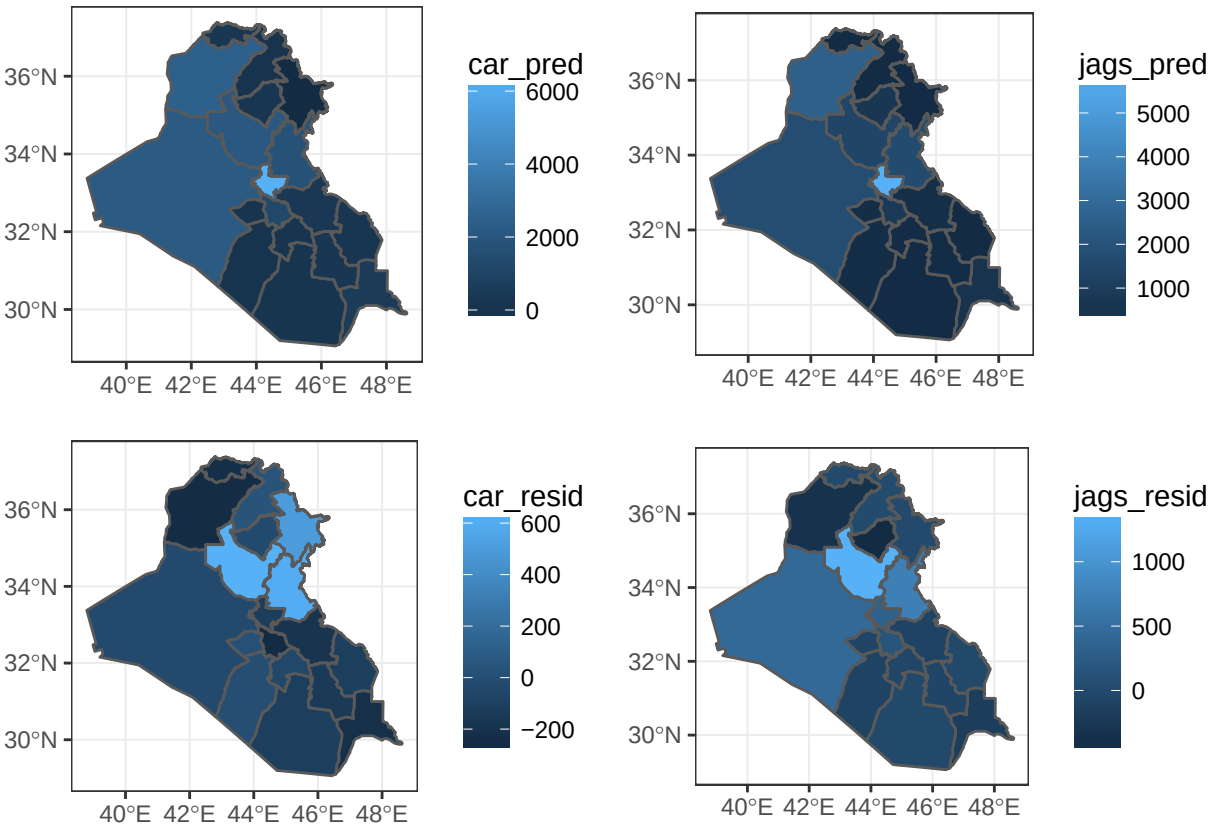
Prediction Based on Civilian Death



Given there is spatial pattern among terroristic attacks in Iraq, we use the number of civilian deaths from violence to predict the number of attacks in Iraq. In this case, we apply CAR and SAR using both Frequentist and Bayesian approaches to predict.



As the graphs show, the predicted maps are similar to the observed attack map. However, the predicted maps from the CAR and SAR models by Frequentist approach are more accurate than the models by Bayesian approach. Table.2 shows Car model using the Frequentist approach performs best among all models in terms of RMSE. If considering Moran's I of Residual and Geary's C of Residual, SAR model by the Frequentist approach performs best overall, since there is almost no hidden spatial pattern in its residual.



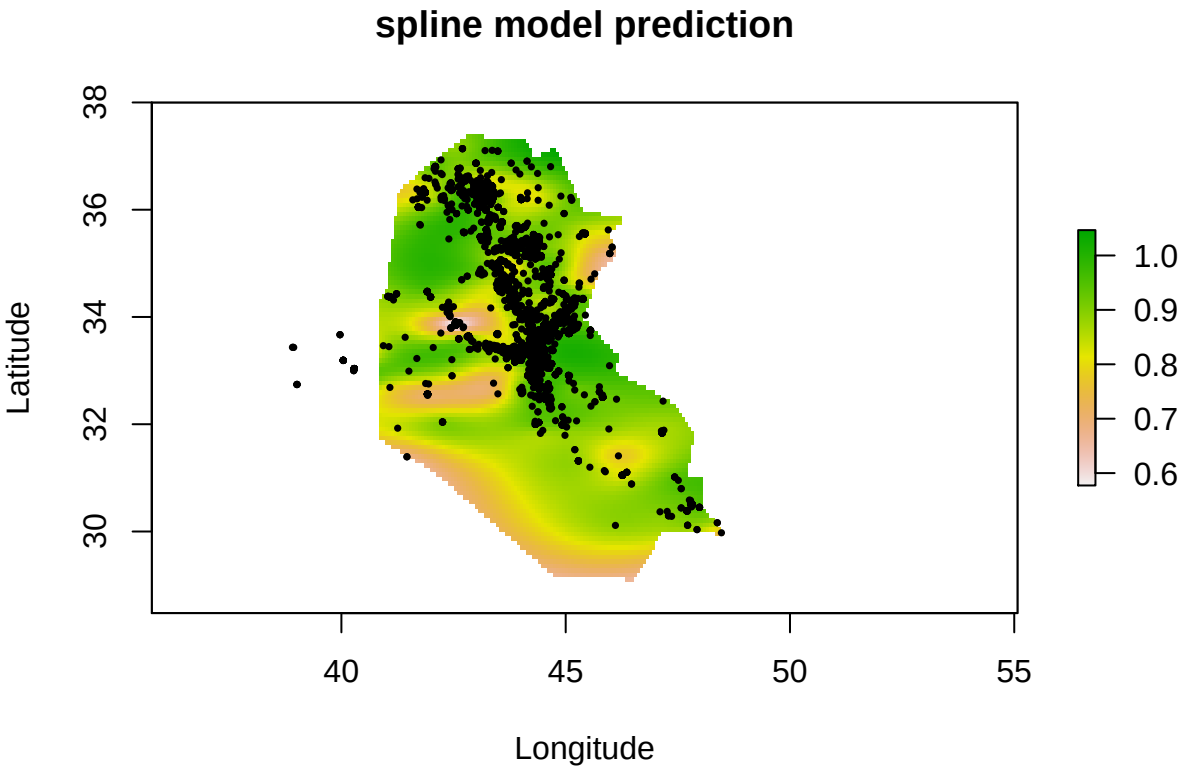
However, we can still observe some hidden spatial pattern from Moran's I of Residual and Geary's C of Residual in the JAGS models. Moreover, the RMSE of the JAGS models is also more than 40% higher than that of Frequentist models. We think it might be caused by limited iteration. The performance of the JAGS model might be better, if we add more iterations.

Table 2: RMSE, Moran's I of Residual and Geary's C of Residual

Model	RMSE	Moran's I of Residual	Geary's C of Residual
SAR	295.02	0.01	1.11
CAR	273.29	0.30	0.76
SAR JAGS	422.51	-0.32	1.46
CAR JAGS	413.75	0.05	1.14

Spline model results

As mentioned before, spline model is like bending a sheet of rubber so that it passed through the points while minimizing the total curvature of the surface. Compared to the SAR and CAR models, the spline model can capture the spatial pattern more precisely. Especially for predicting ridges and valleys, the spline will perform much better than SAR and CAR model. In this project, spatial patterns for these terroristic attacks are highly similar to the rugged terrain since these attacks always take place in the area where it has high density of either population or energy.



Instead of trying to explain the spatial patterns of terroristic attacks using splines, we use spline to predict the spatial pattern of success rate of attacks across the Iraq. As can be seen in the above plots, most of the attacks concentrating in the center of Iraq have high success rates. Not surprisingly, the capital city, Baghdad, is also located in the center of Iraq. Only a small amount of attacks took place in the south of Iraq but with high success rate. However, although a lot of attacks took place on the northern provinces compared to the southern region, the success rate is relatively low.

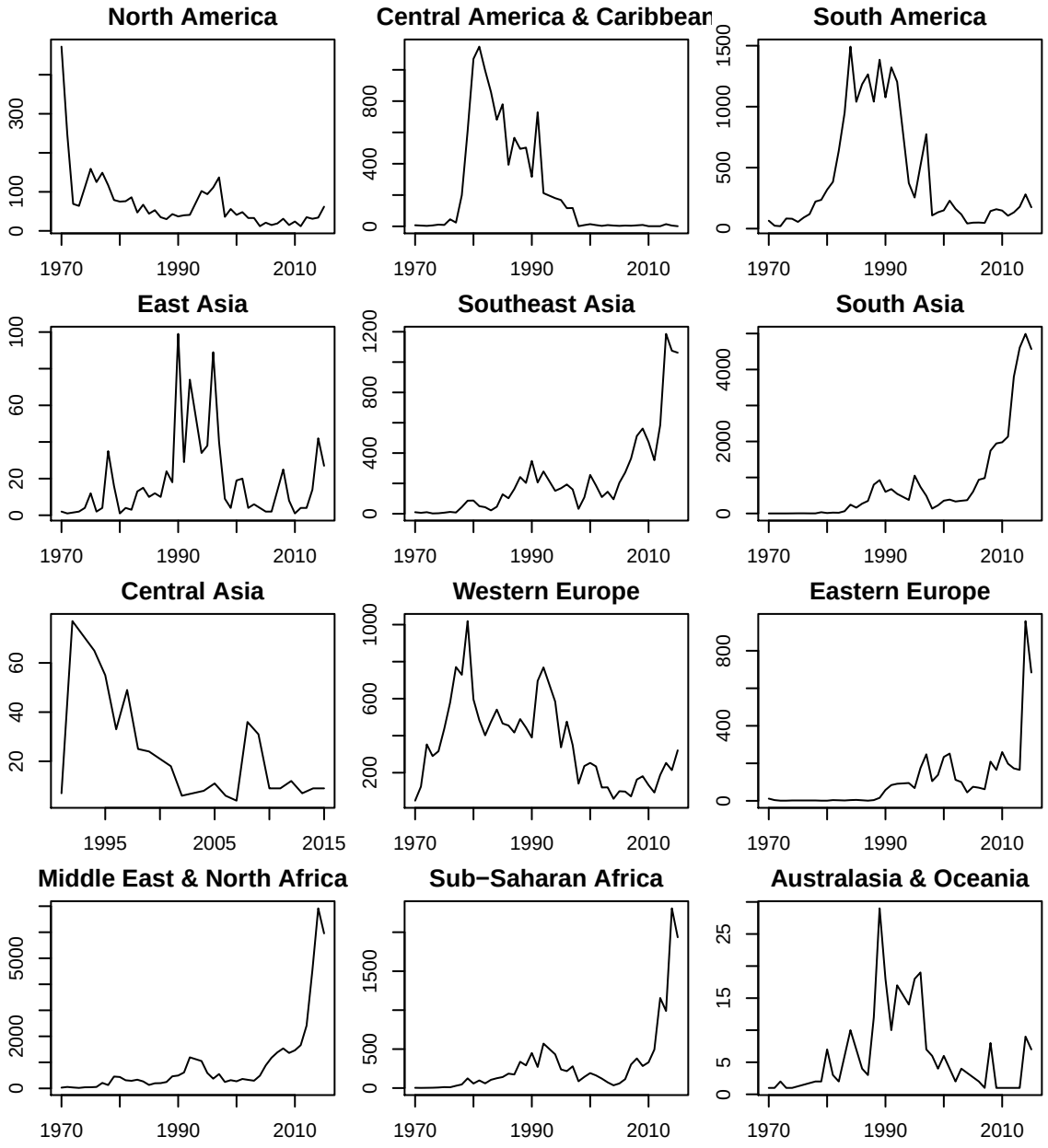
Conclusion and Discussion

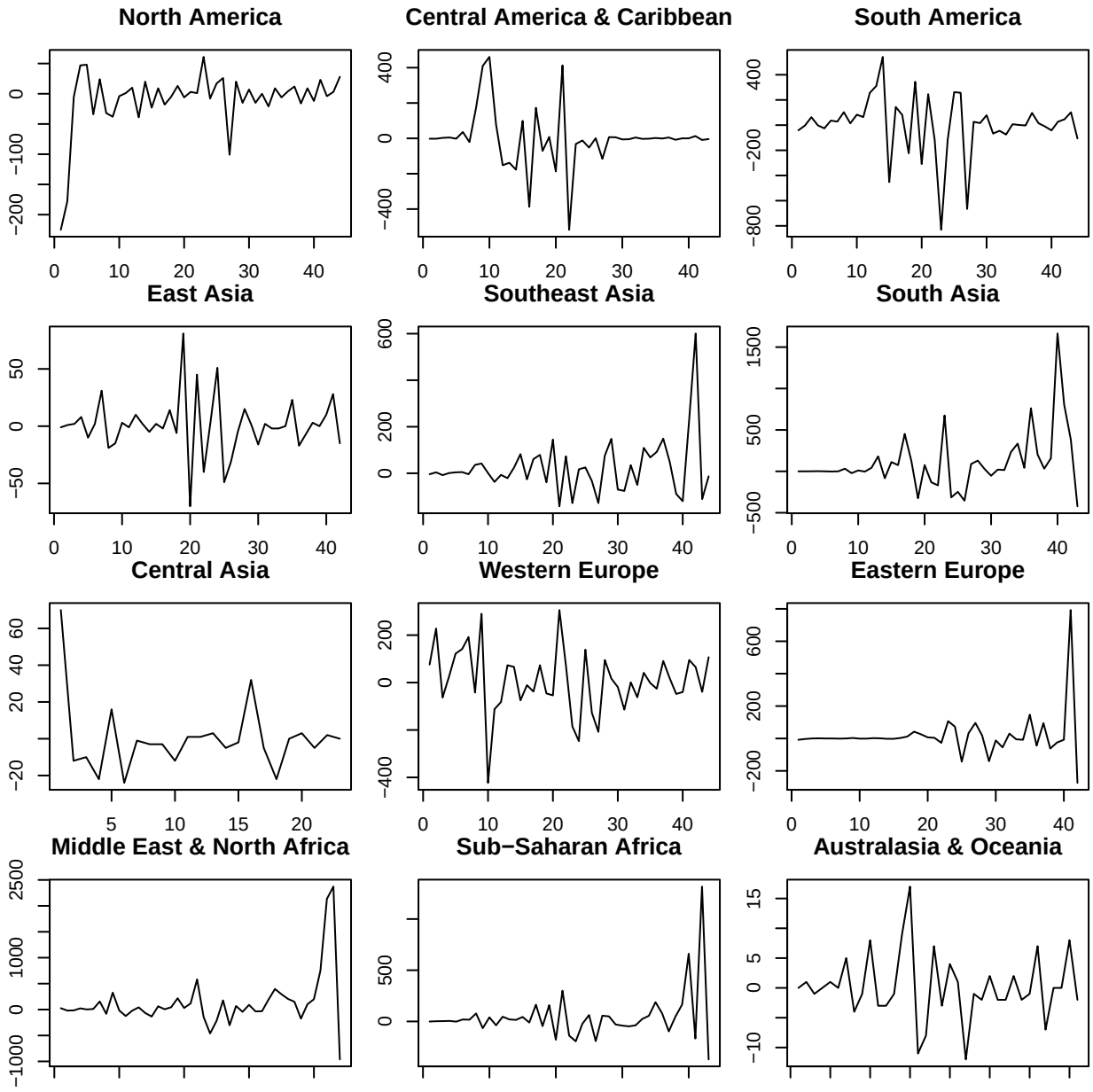
For this project, we utilized methods learnt from this class using both time series and spatial models and applied them on the terrorist attack dataset from Global Terrorism Database. We used ARIMA models to forecast attack counts for different regions of the world and were able to obtain models that match the current global political situations. In the spatial context, both frequentist and bayesian versions of CAR and SAR model can capture the spatial patterns of terroristic attacks. In addition, we used a spline model to predict the success rate of a terrorstic attack based on the observed data. In the future, we could try bayesian version of spline model to produce predications. Then we can compare the results to the regular spline model, since the regular spline model is not good at uncertainty prediction. This project sheds some light on the complicated situation of terrorist attacks with models that are easy to intepret. To improve the results of our models, we can take further steps such as incorporating outside sources of data in the analysis and further tuning of the models.

Reference

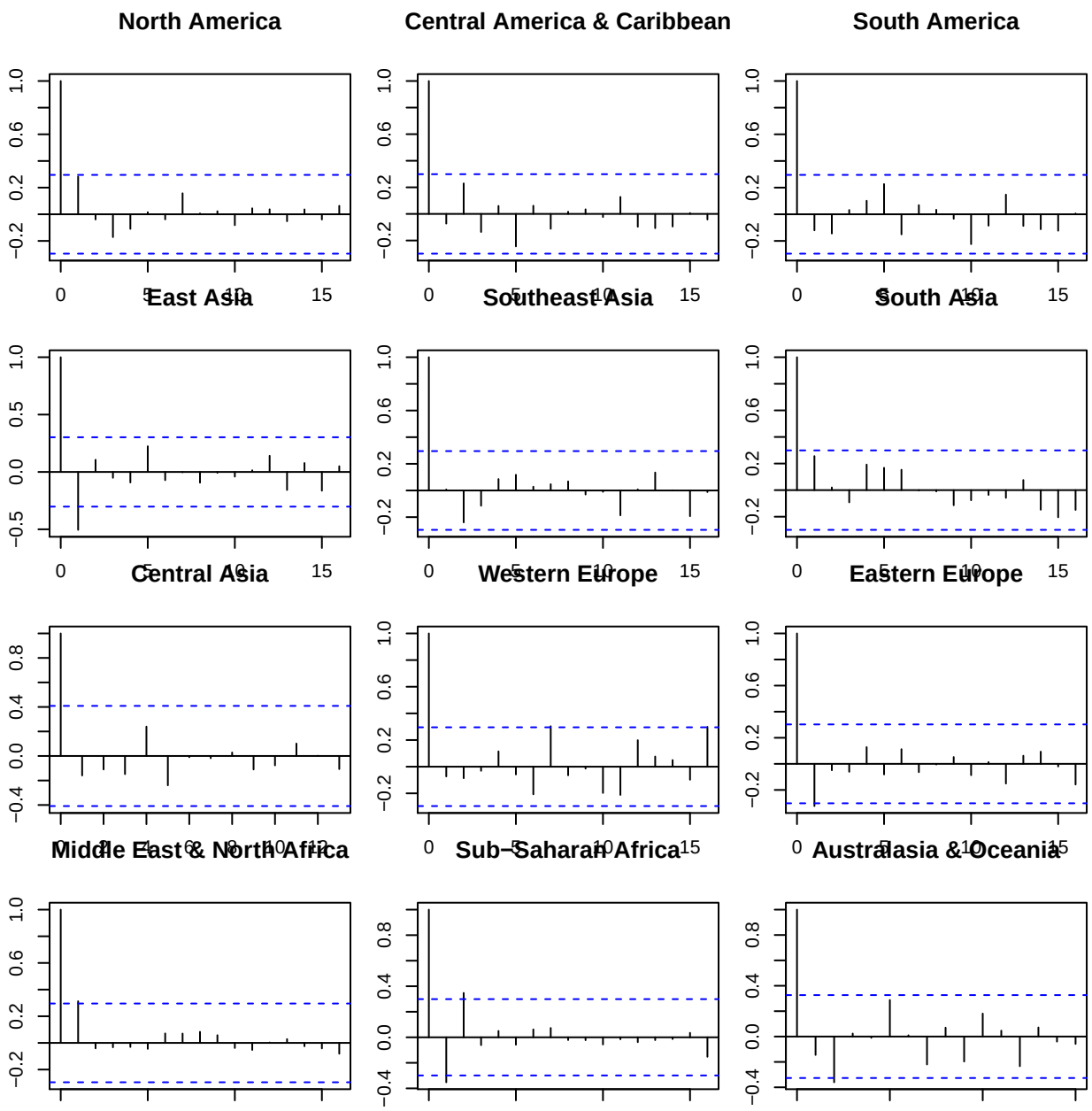
[1] Rundel, Colin. “ARIMA Models.” STA644. Duke University. Feb 8, 2017. [2] Coline Childs, “Interpolating Surfaces in ArcGIS Spatial Analyst”, ESRI Education Services Sep 2004

Appendix

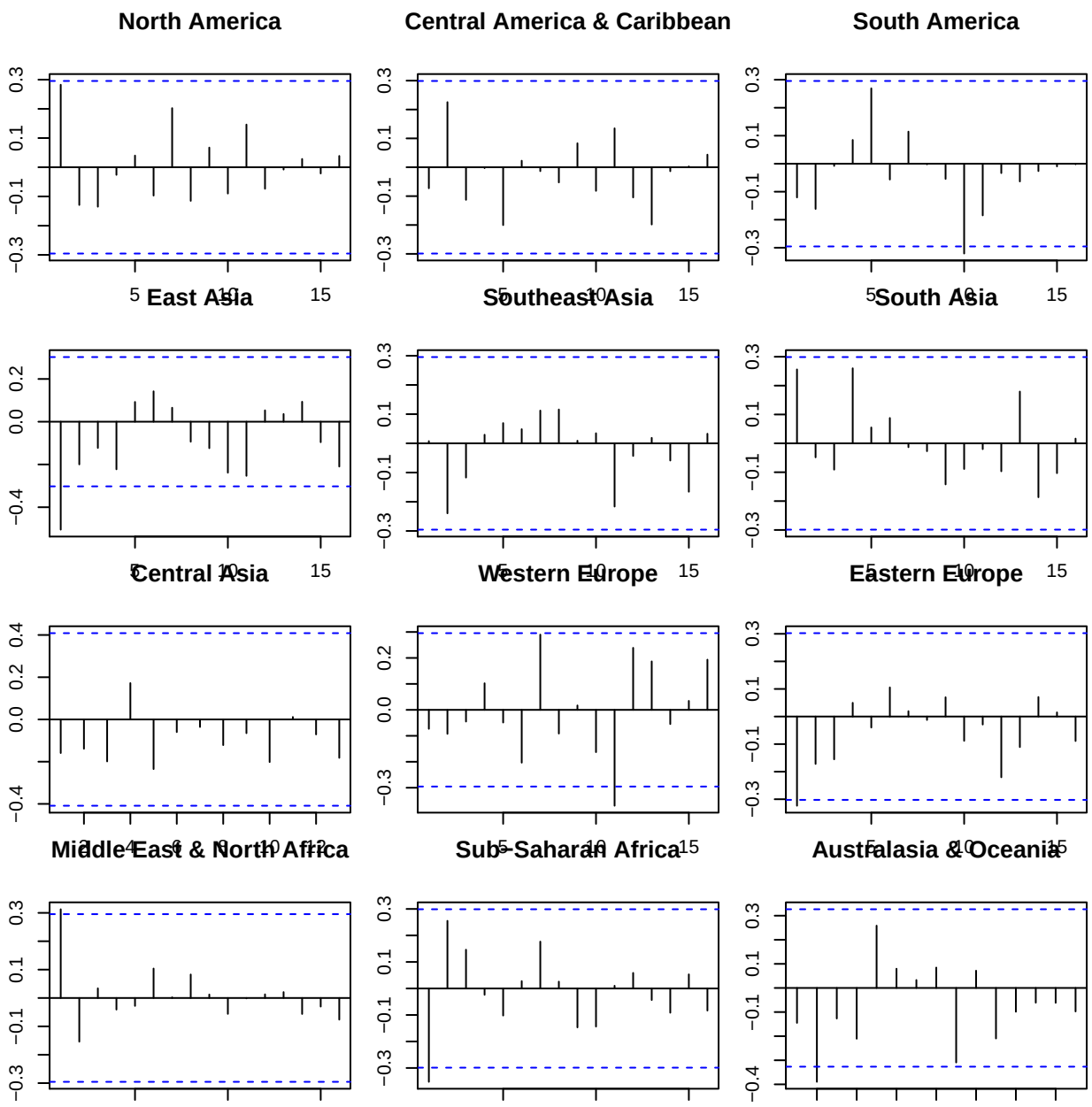

North America
Central America & Caribbean
South America
East Asia
Southeast Asia
South Asia
Central Asia
Western Europe
Eastern Europe
Middle East & North Africa
Sub−Saharan Africa
Australasia & Oceania
Yearly Count



Differenced Data

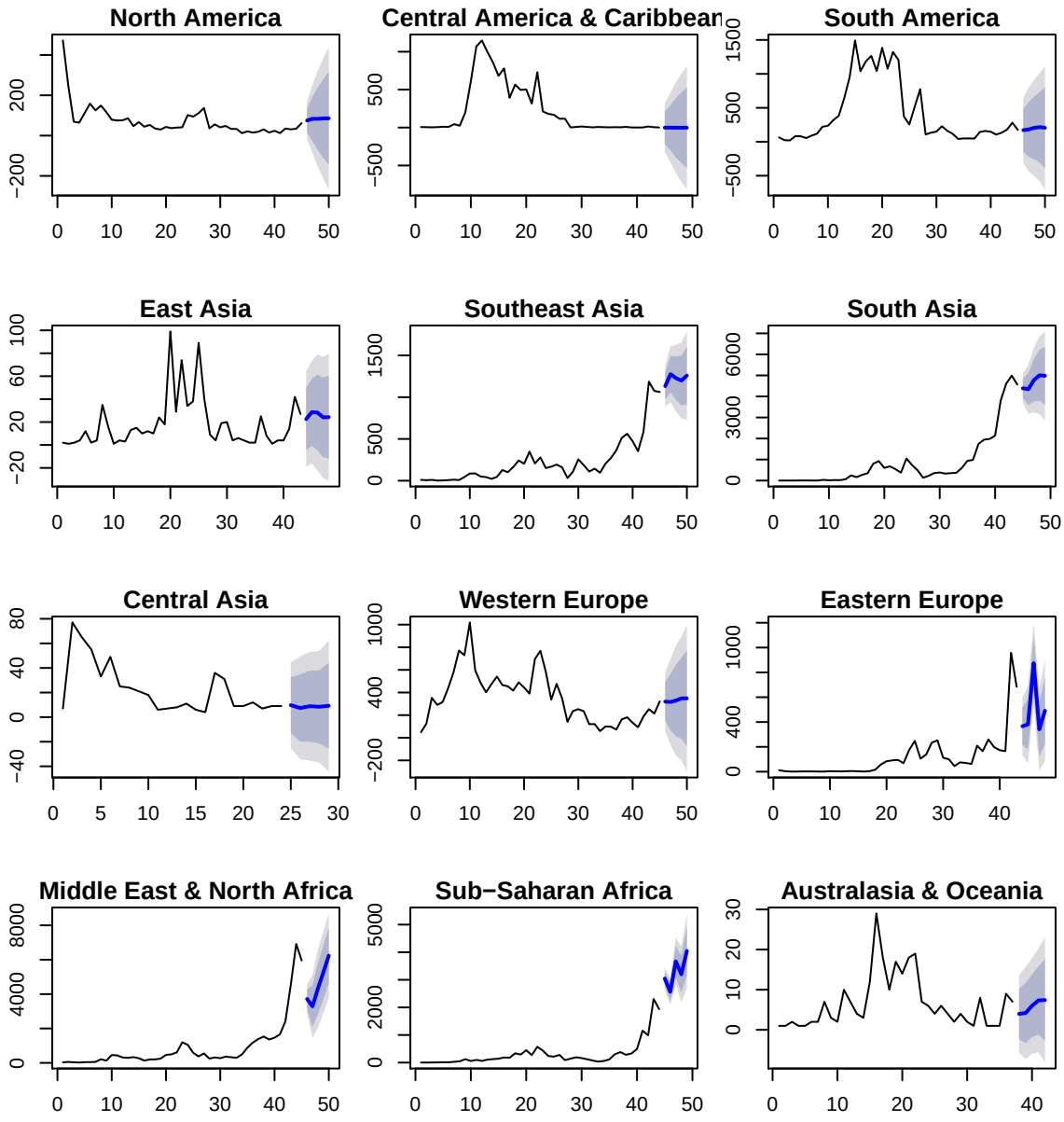

North America
Central America & Caribbean
South America
East Asia
Southeast Asia
South Asia
Central Asia
Western Europe
Eastern Europe
Middle East & North Africa
Sub−Saharan Africa
Australasia & Oceania

acf

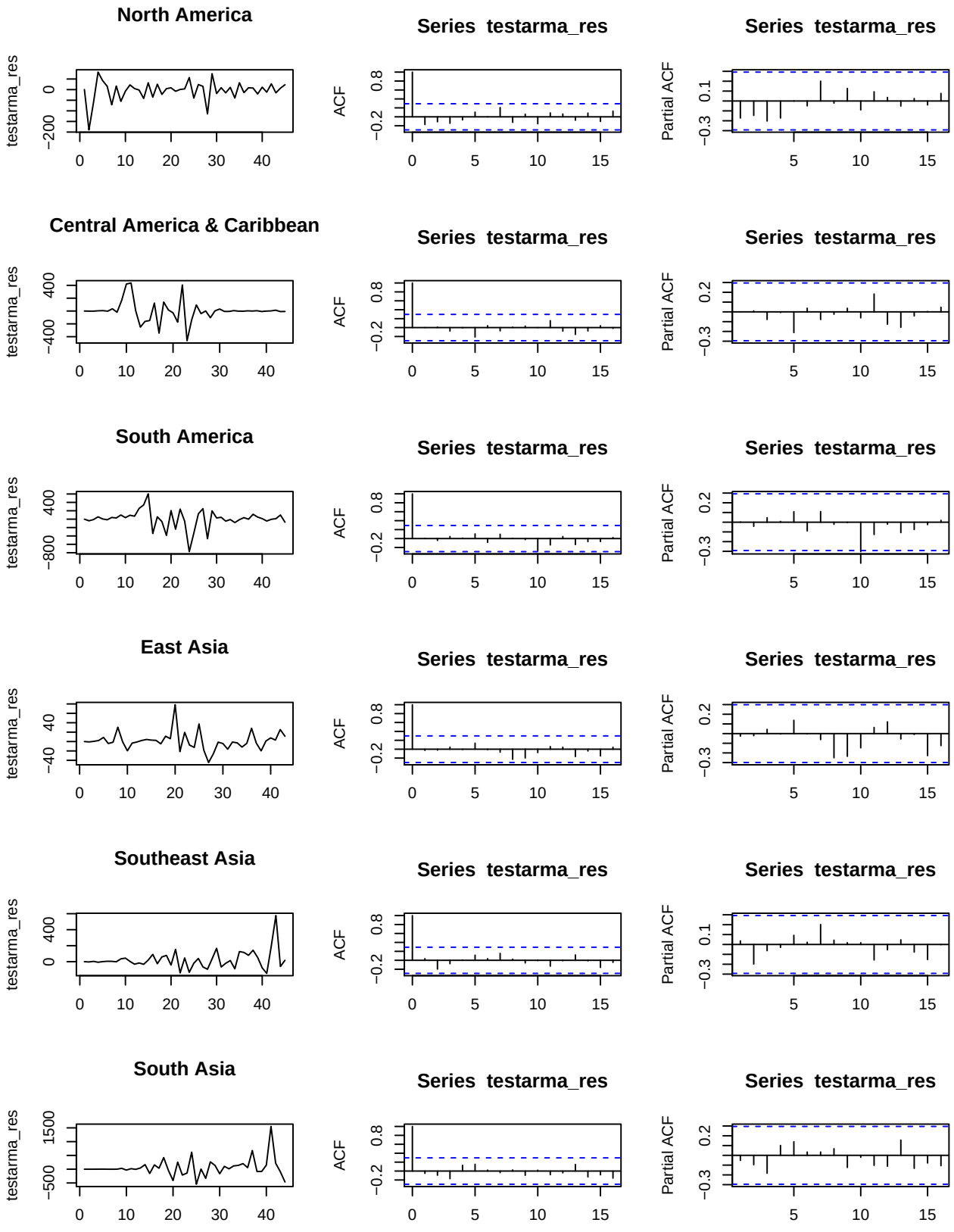


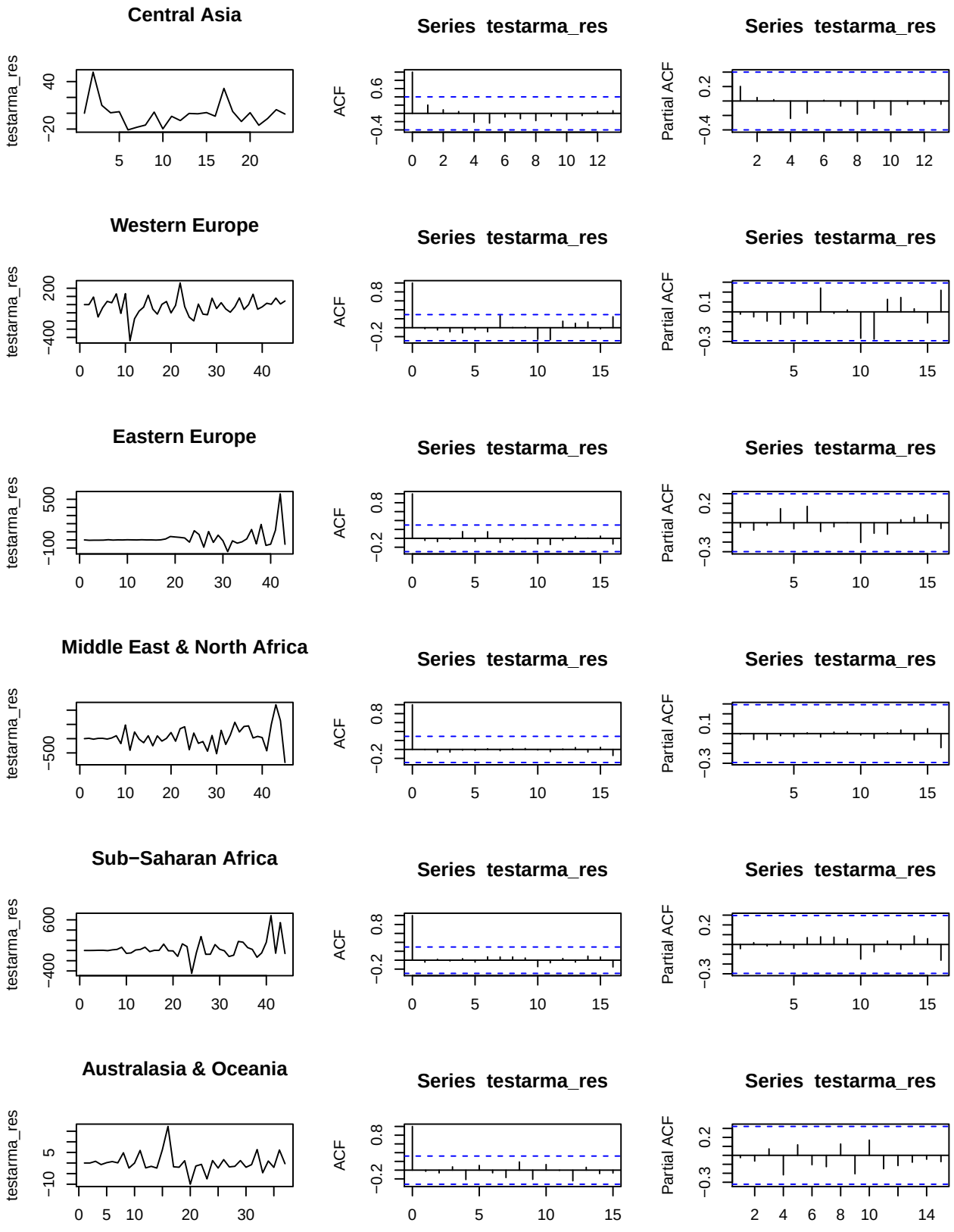
pacf

Periodogram

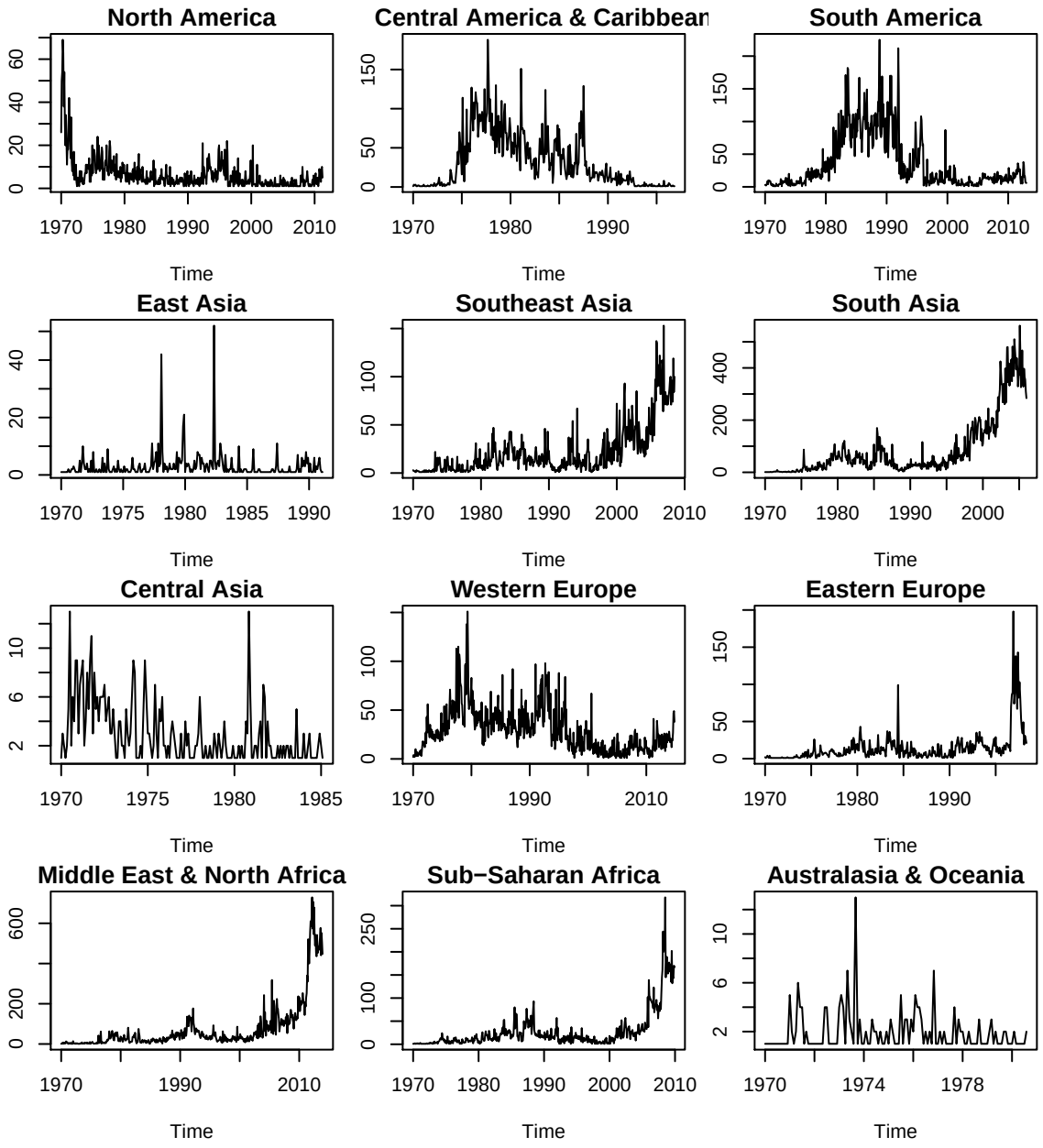


Yearly Forecast

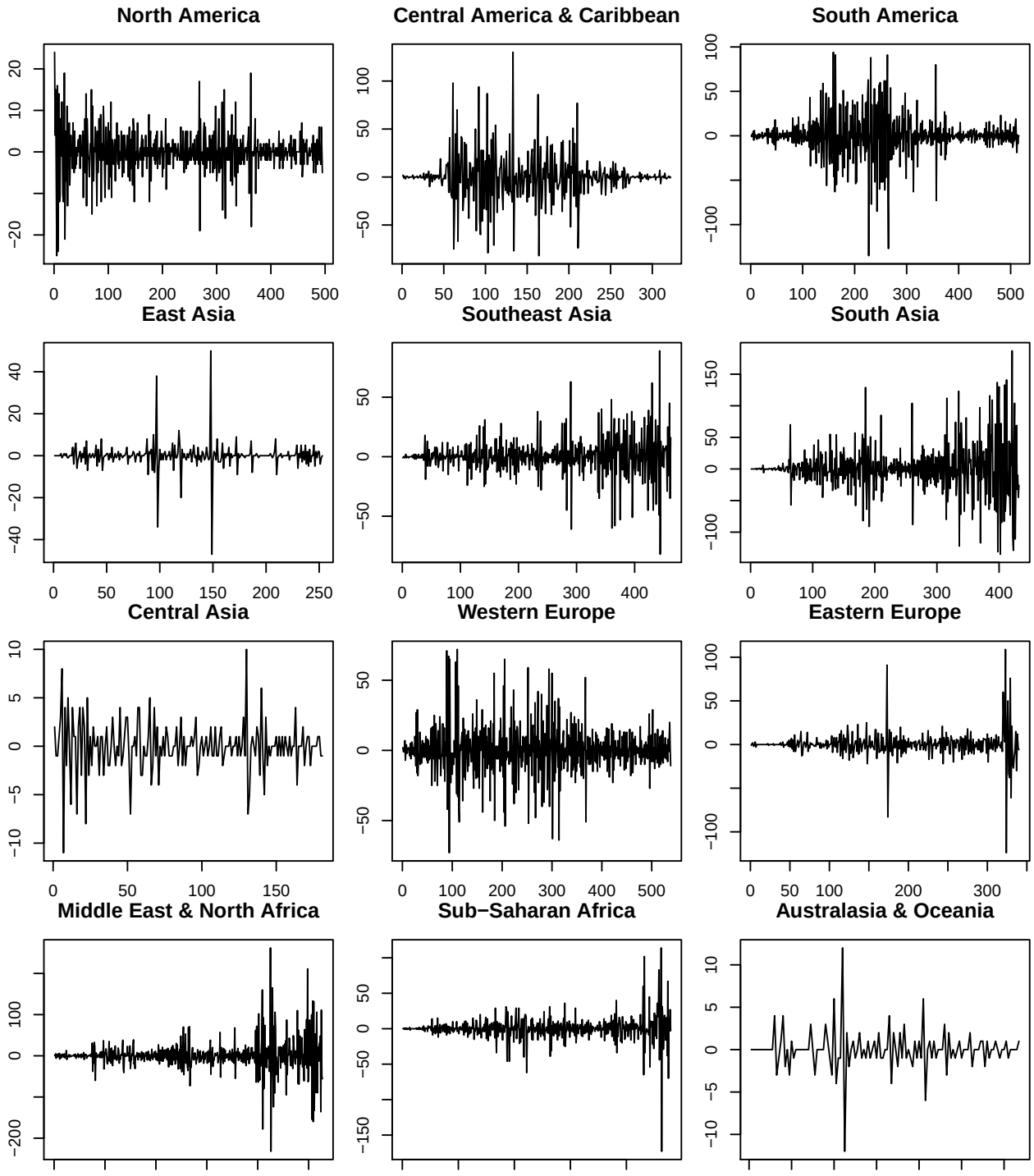

North America
Series testarma_res
Series testarma_res
Central America & Caribbean
Series testarma_res
Series testarma_res
South America
Series testarma_res
Series testarma_res
East Asia
Series testarma_res
Series testarma_res
Southeast Asia
Series testarma_res
Series testarma_res
South Asia
Series testarma_res
Series testarma_res
testarma_res
ACF
Partial ACF


Central Asia
Series testarma_res
Series testarma_res
Western Europe
Series testarma_res
Series testarma_res
Eastern Europe
Series testarma_res
Series testarma_res
Middle East & North Africa
Series testarma_res
Series testarma_res
Sub−Saharan Africa
Series testarma_res
Series testarma_res
Australasia & Oceania
Series testarma_res
Series testarma_res
testarma_res
ACF
Partial ACF

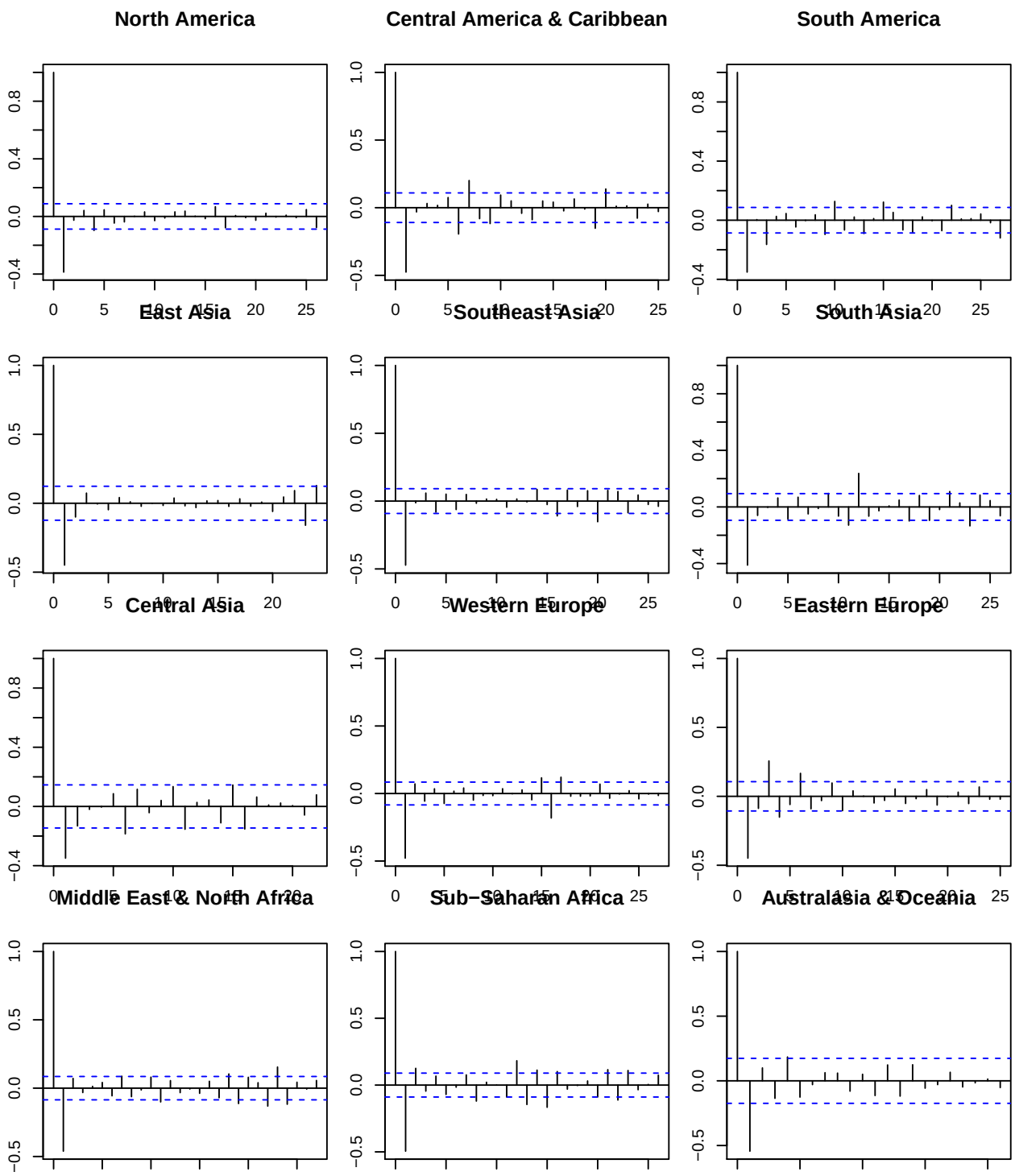
Residuals



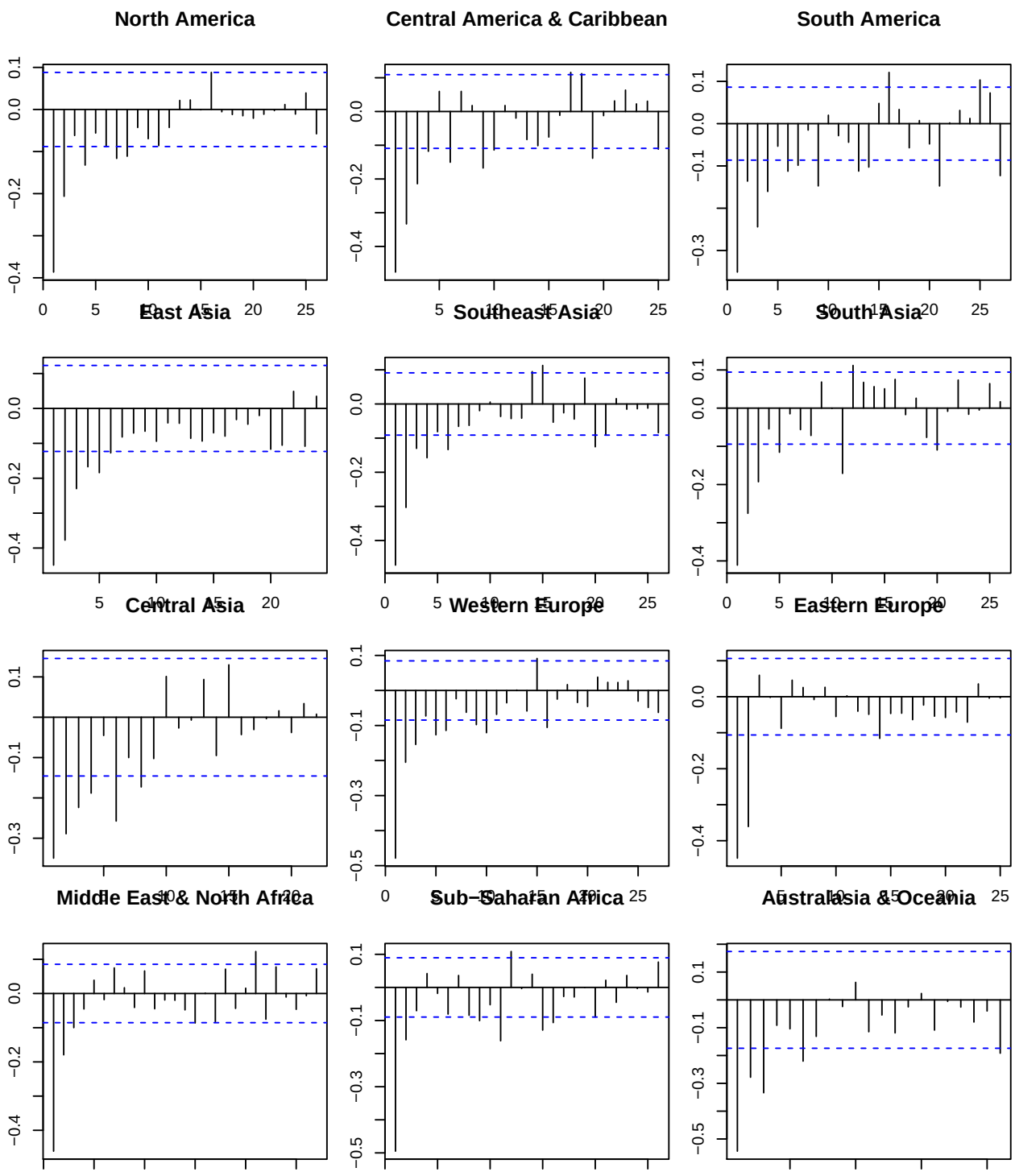
Monthly Original Data



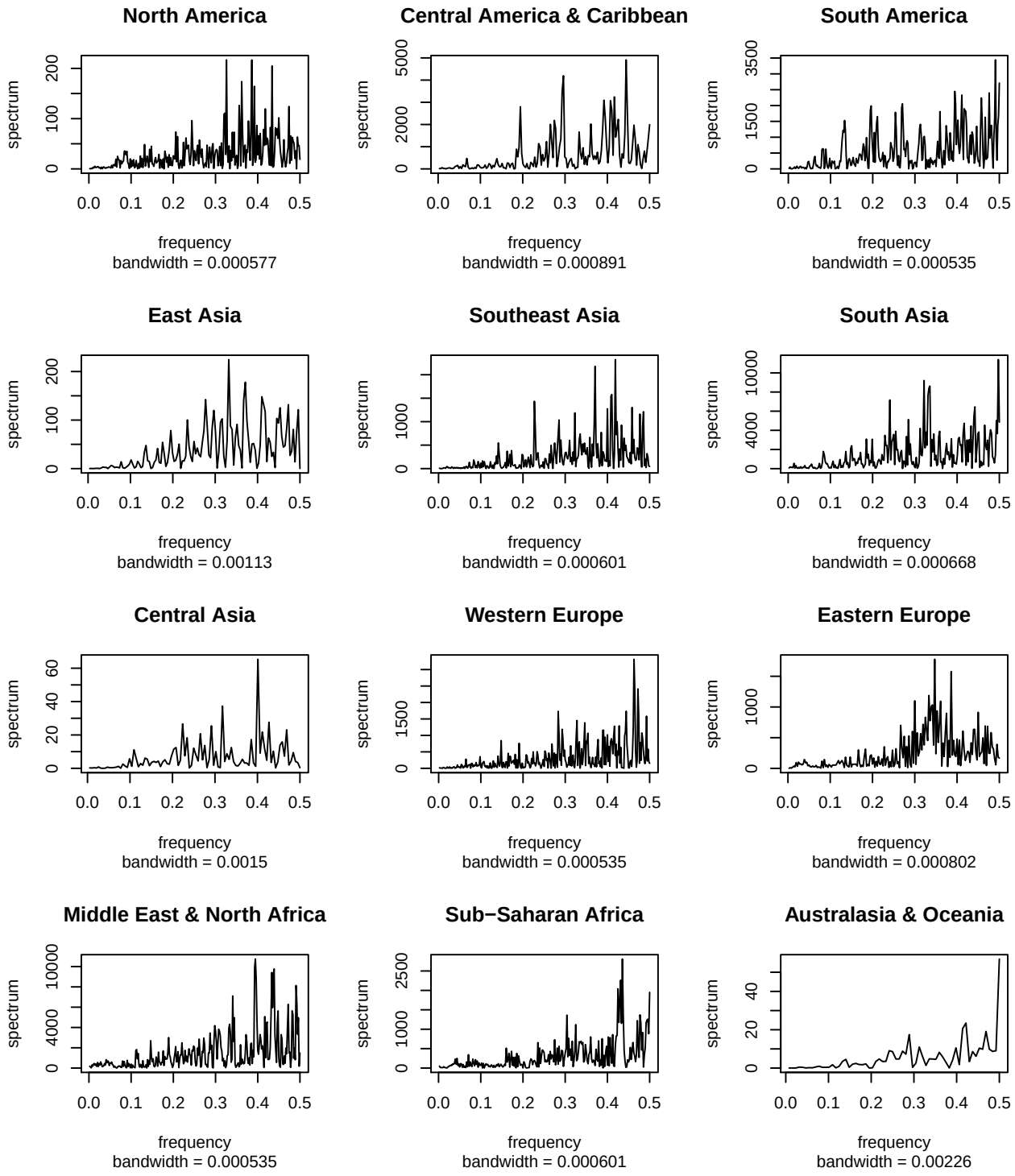
Differenced Data


North America
Central America & Caribbean
South America
East Asia
Southeast Asia
South Asia
Central Asia
Western Europe
Eastern Europe
Middle East & North Africa
Sub−Saharan Africa
Australasia & Oceania

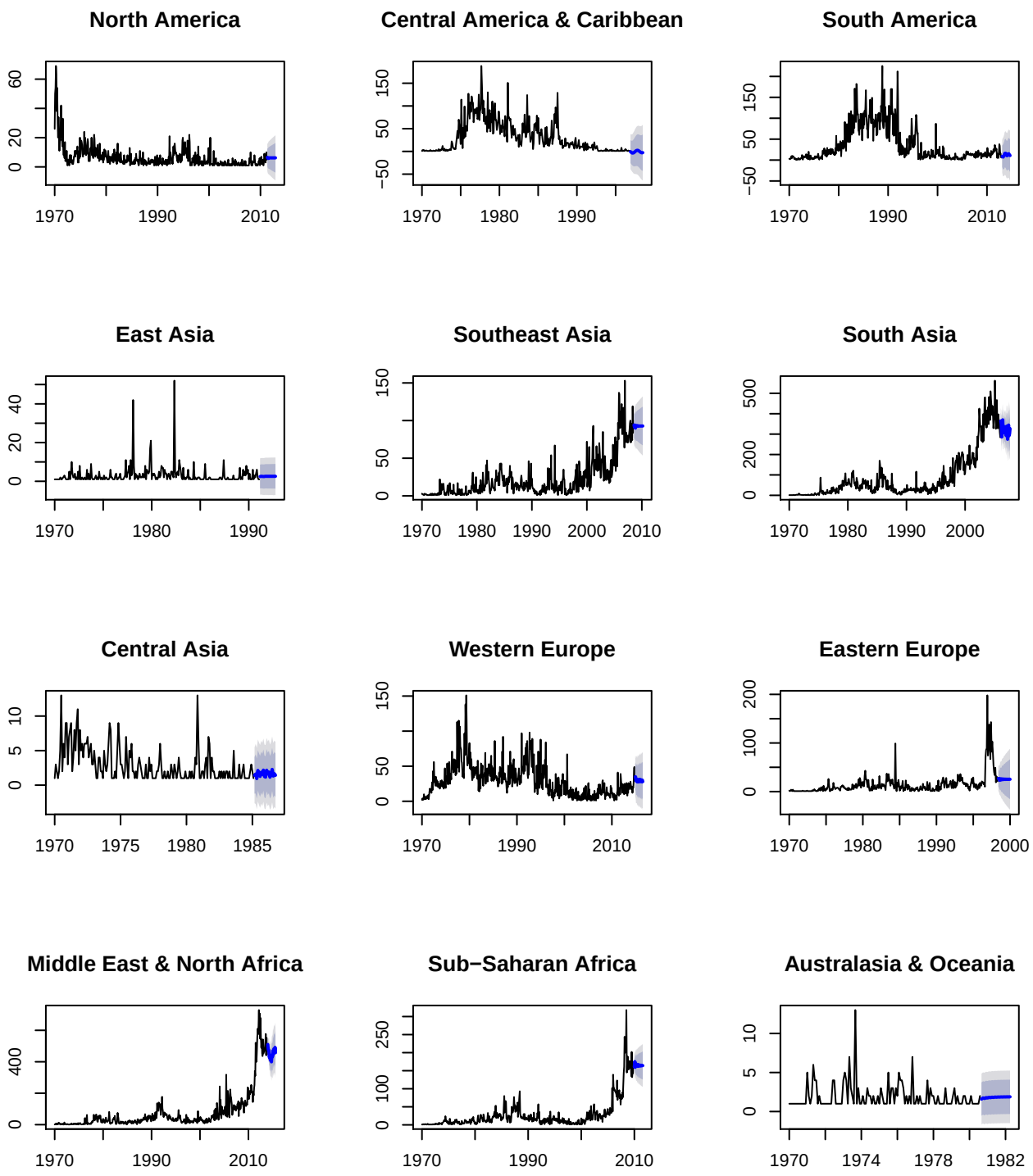
acf


North America
Central America & Caribbean
South America
East Asia
Southeast Asia
South Asia
Central Asia
Western Europe
Eastern Europe
Middle East & North Africa
Sub−Saharan Africa
Australasia & Oceania

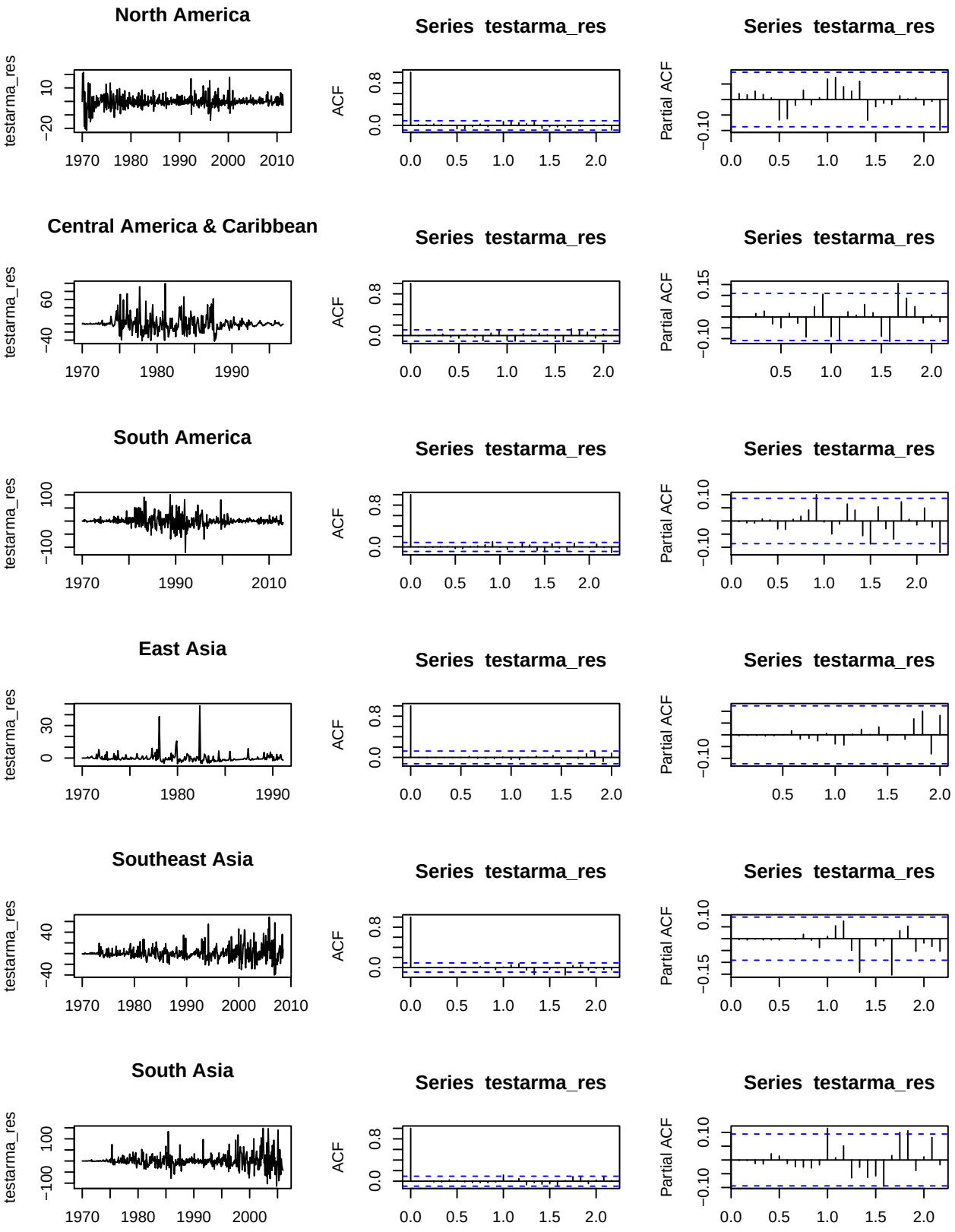
pacf

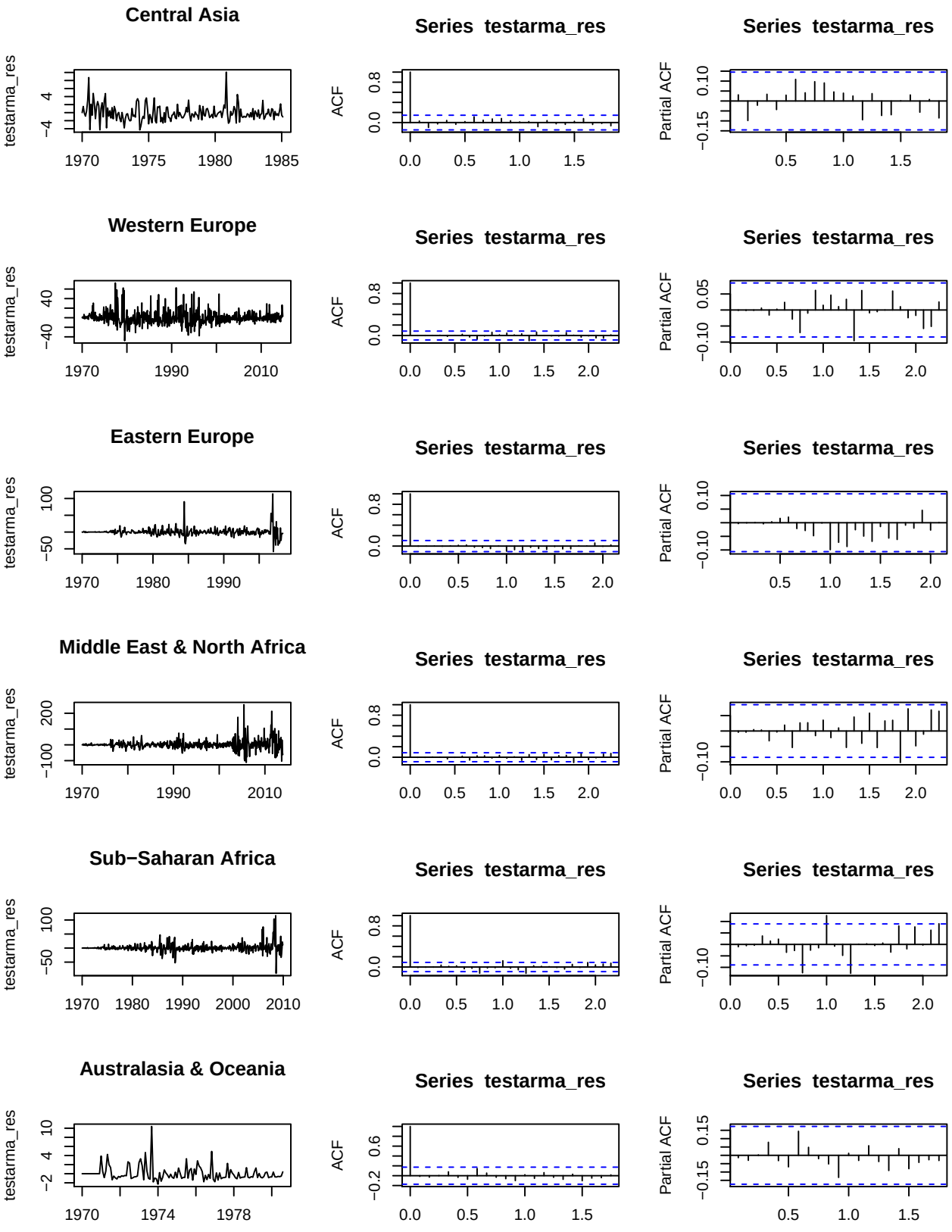


Periodogram



Forecast - Monthly


North America
Series testarma_res
Series testarma_res
Central America & Caribbean
Series testarma_res
Series testarma_res
South America
Series testarma_res
Series testarma_res
East Asia
Series testarma_res
Series testarma_res
Southeast Asia
Series testarma_res
Series testarma_res
South Asia
Series testarma_res
Series testarma_res
testarma_res
ACF
Partial ACF


Central Asia
Series testarma_res
Series testarma_res
Western Europe
Series testarma_res
Series testarma_res
Eastern Europe
Series testarma_res
Series testarma_res
Middle East & North Africa
Series testarma_res
Series testarma_res
Sub−Saharan Africa
Series testarma_res
Series testarma_res
Australasia & Oceania
Series testarma_res
Series testarma_res
testarma_res
ACF
Partial ACF

Residuals