Supplemental material for "Granger causality and time series regression for modelling the migratory dynamics of influenza into Brazil"

Aline Foerster Grande[1], Guilherme Pumi[1,2] and
Gabriela Bettella Cybis[1]

December 2022

The material contained herein is supplementary to the article named in the title and published in SORT-Statistics and Operations Research Transactions Volume 46(2).

---

[1] Programa de Pós-Graduação em Estatística - Universidade Federal do Rio Grande do Sul.

[2] Corresponding author. E-mail: guilherme.pumi@ufrgs.br

Introduction

Due to page restrictions, in this document we present some simple plots and graphs related to the data studied in its parent paper. Please refer to the parent paper for more details regarding the data and context for the results presented here.

Is it necessary to use covariates?

In Section 3.2, we apply the concept of Granger causality to determine whether data from other regions may be useful in predicting the incidence of influenza in Brazil. The results of the regression models applied to the data show that using covariates produces reasonably accurate forecasts. But a question that remains to be answered is how a simple autoregressive model performs compared to the models explored in Section 5. To examine this question, in this section we fit a simple AR model to the influenza incidence data in Brazil (see Section 4.1 for details).

Using the Box and Jenkins approach to time series modeling, a simple stationary AR(2) model was capable to model the data. The relevant results are presented in Table 1. A Ljung-box test applied to the squared residuals of the fitted AR(2) model points toward the absence of an ARCH effect (p-value 0.08).

Table 1. *Fitted AR(2) model for the influenza incidence data in Brazil.*

	Estimate	Std. Error	z stat.	$\Pr(>\|z\|)$
intercept	260.42	57.291	4.546	$< 10^{-5}$
ϕ_1	1.0180	0.0823	12.376	$< 10^{-15}$
ϕ_2	-0.3909	0.0821	-4.759	$< 10^{-5}$

Log-likelihood: -848.2 AIC: 1704.41

Ljung-Box test (df = 18): p-value = 0.9752

Roots of the characteristic polynomial: $1.3023 \pm 0.9287i$

Absolute value: $|1.3023 \pm 0.9287i| = 1.5995$

Although the model identification and estimation present no difficulty, it is well known that out-of-sample forecasted values for models presenting very short memory, such as the AR(2), converge very fast to a constant. This is the case of the fitted AR(2) model as it can be seen in Figure 1 where we present the observed values (black) along with 11-steps ahead forecasts (red). As expected, the fitted AR(2) model does a poor job at forecasting the out-of-sample dynamics of the data, converging very fast to a constant. The mean square error for the 11-step ahead forecast is 50,372, more than 2.5 times higher than the worst result presented in Section 5. The in-sample mean squared error is also high: 56,677. In conclusion, the introduction of covariates in the model significantly improves forecasting.

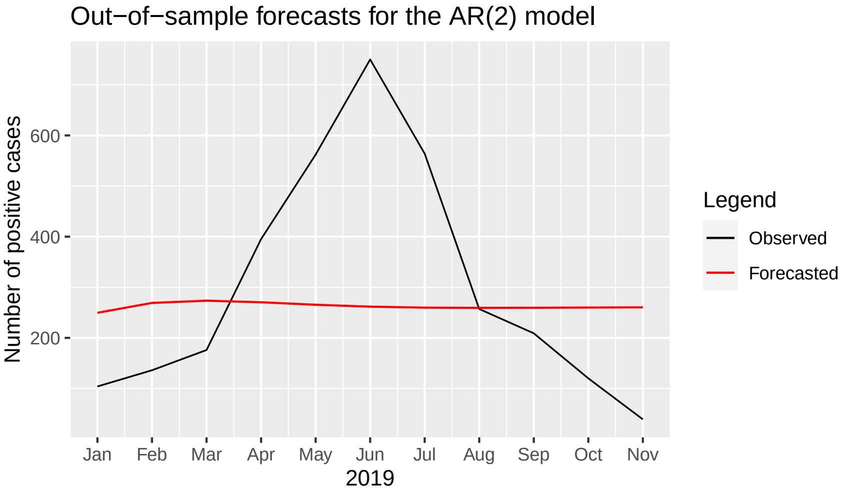


Figure 1. *11-steps ahead forecasts for the fitted AR(2) model (red) compared to the observed values (in black).*

List of countries in each geographic region for genetic diversity data

Table 2. *List of countries with sequence data in each geographic region for H3N2 genetic diversity measures. Countries in bold have more than 100 sequences.*

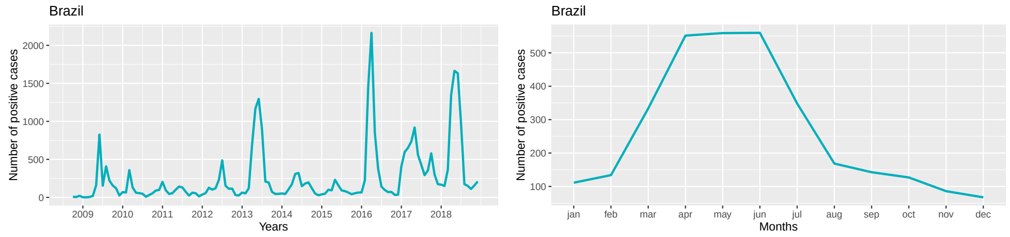
Region	Countries
Asia	**Singapore, China, Hong Kong, Japan, Thailand, Malaysia, South Korea, Viet Nam, India, Iran,** Saudi Arabia, Indonesia, Israel, Jordan, Philippines, Kuwait, Taiwan, Georgia, Afghanistan, Lebanon, Kyrgyzstan, Bangladesh, Cambodia, State of Palestine, Bahrain, Nepal, Sri Lanka, United Arab Emirates, Iraq, Myanmar, Bhutan, Qatar, Turkey
North America	**USA, Canada, Nicaragua**, Mexico, Panama, El Salvador, Dominican Republic, Honduras
Global	**USA, Australia, Canada, Switzerland, Singapore, Japan, China, Hong Kong, Nicaragua, New Zealand, Germany, Chile, Thailand, Denmark, South Korea, India, Russia, Peru, Malaysia, Netherlands, United Kingdom, Italy, Iran,** Brazil, Cambodia, Czech Republic, Philippines, Viet Nam, Taiwan, Kenya, Mexico, France, Guam, Israel, Finland, Jordan, Sweden, Kuwait, Saudi Arabia, Uganda, Georgia, South Africa, Bahrain, Egypt, Spain, State of Palestine, Lebanon, Colombia, Norway, Tunisia, Bangladesh, Indonesia, Kyrgyzstan, Ethiopia, Panama, Uruguay, Belgium, Bolivia, Sri Lanka, Djibouti, Chad, United Arab Emirates, Austria, Bhutan, Iraq, Nepal, Senegal, Dominican Republic, Hungary, Serbia, Bosnia and Herzegovina, Afghanistan, El Salvador, Ghana, Honduras, Kosovo, Myanmar, Luxembourg, New Caledonia, Qatar, Turkey

Table 3. *List of countries with sequence data in each geographic region for H1N1 genetic diversity measures. Countries in bold have more than 100 sequences.*

Region	Country
Asia	**Singapore, China, Japan, India, Thailand, Taiwan, Malaysia, Iran, Hong Kong, Viet Nam,** South Korea, Turkey, Jordan, Nepal, Saudi Arabia, Oman, Sri Lanka, Afghanistan, Kuwait, Israel, Bahrain, Mongolia, Indonesia, Kazakhstan, Myanmar, Turkmenistan, Lebanon, State of Palestine, Kyrgyzstan, Bangladesh, Cambodia, Philippines
North America	**USA, Canada, Nicaragua, Mexico,** Dominican Republic, Haiti, El Salvador, Puerto Rico,
Global	**USA, United Kingdom, Singapore, China, Japan, Finland, Canada, Brazil, Taiwan, Russia, India, Nicaragua, Mexico, Iran, Hong Kong , Australia, Viet Nam, Malaysia, Greece, New Zealand, Chile,** Germany, South Korea, Denmark, Estonia, Turkey, Argentina, Czech Republic, Netherlands, Kenya, Cambodia, Norway, Saudi Arabia, France, Guam, Peru, Oman, Egypt, Uganda, Spain, Belgium, Kuwait, Poland, Jordan, Nepal, Puerto Rico, Italy, Dominican Republic, Mongolia, Paraguay, Afghanistan, Indonesia, Austria, Hungary, Philippines, Serbia, Tunisia, Sri Lanka, Ireland, Kazakhstan, New Caledonia, Colombia, Bahrain, Ghana, Myanmar, Sweden, Bolivia, Senegal, South Africa, State of Palestine, Ecuador, El Salvador, Kyrgyzstan, Nigeria, Switzerland, Djibouti, Ethiopia, Luxembourg, Solomon Islands, Venezuela, Bangladesh, Fiji, Zambia, Israel, Lebanon, Tonga, Turkmenistan, Ukraine, Haiti.

Time series plots

Figure 2(a) presents the time series of positive flu cases in Brazil. From the plots, it can be seen that the series does not appear to be stationary, due to a very distinct seasonal pattern. Figure 2(b) shows the number of average positive cases each month in Brazil. The months of April, May and June (close to winter in Brazil) are the ones presenting the highest incidence of influenza. In addition, the months with the lowest number of cases of influenza are November, December and January, which corresponds to summertime.

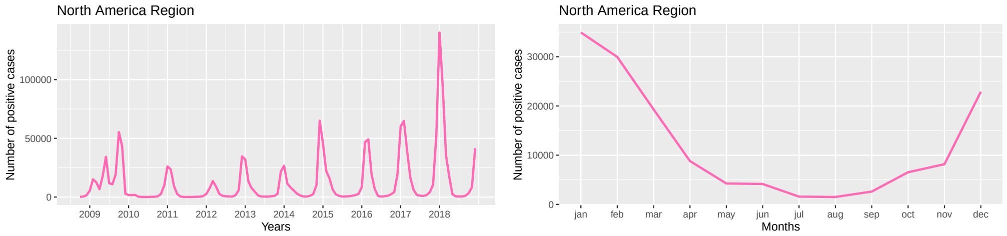


(a) Time series

(b) Monthly average

Figure 2. *Time series and monthly average of flu incidence in Brazil.*

Figure 3(a) shows the time series of positive flu cases in the North America Region. The series presents a very distinct seasonality with peaks occurring at the beginning/end of the year (winter season). Figure 3(b) illustrates the number of average positive cases each month in the North America Region. December, January, February and March are the months with the highest number of influenza cases, which, not coincidentally, are the peak of the winter in the northern hemisphere.

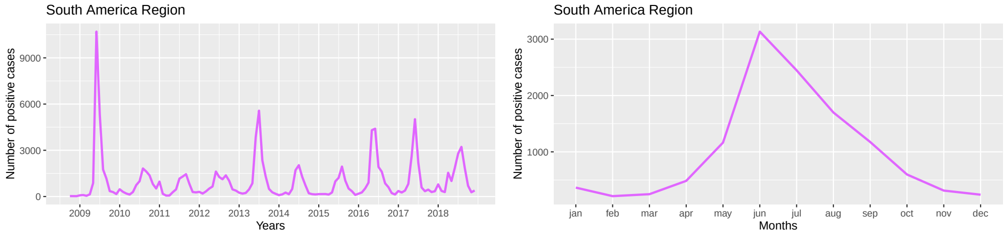


(a) Time series

(b) Monthly average

Figure 3. *Time series and monthly average of flu incidence in North America.*

Figure 4(a) shows the time series of positive flu cases in the South America Region. Like the others, the time series shows a remarkable annual seasonality, and is, therefore, not stationary. Figure 4(b) shows the monthly average of positive cases in the South America Region. It can be seen that, as winter approaches the southern hemisphere, the number of cases increases. The months with the highest incidence of influenza are June, July and August (winter season).

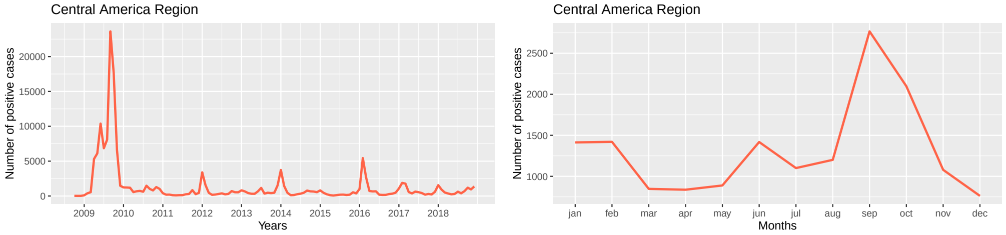


(a) Time series

(b) Monthly average

Figure 4. *Time series and monthly average of flu incidence in South America.*

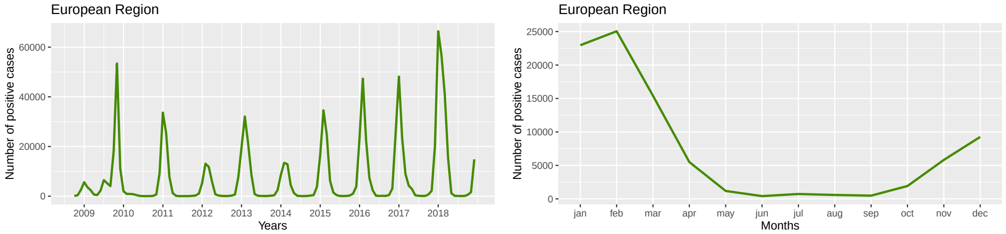
Figure 5(a) shows the time series of positive flu cases in the Central America Region. Notice that between the years 2009 and 2010 the incidence of influenza is much higher than the other years. The swine flu (H1N1) is the main reason for this peak, since Mexico is considered the origin and epicenter of this epidemic. In addition, the time series presents a distinct annual seasonality. Figure 5(b) shows the monthly average of positive cases each month in the Central America Region. The graph shows a peak in the incidence of influenza in September.



(a) Time series (b) Monthly average

Figure 5. *Time series and monthly average of flu incidence in Central America.*

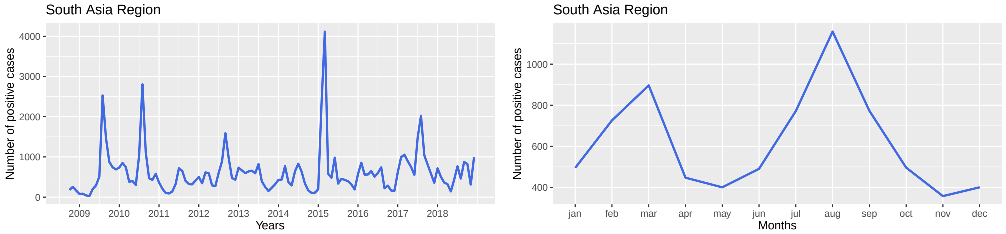
Figure 6(a) shows the time series of positive flu cases in the European Region. Analogously to the previous cases, the series presents a clear annual seasonality. Figure 6(b) presents the monthly average of positive cases by month in the European Region. The plot shows a behavior similar to that of the North America Region, since the months with highest incidence of influenza are December, January, February and March (winter season).



(a) Time series (b) Monthly average

Figure 6. *Time series and monthly average of flu incidence in Europe.*

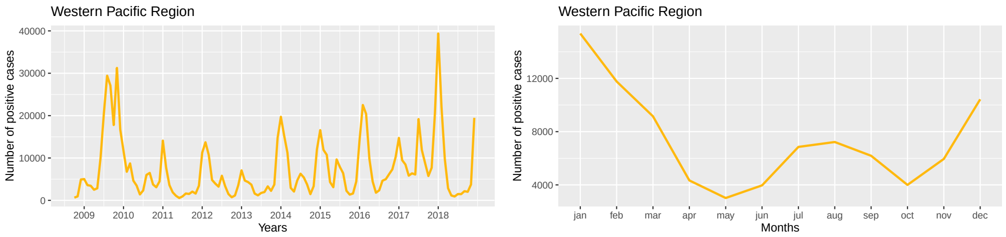
Figure 7(a) shows the time series of positive flu cases in the South Asia Region. The series behavior is similar to the other ones. Figure 7(b) shows the the monthly average of positive cases for each month in the South Asia Region. Incidence of influenza peaks in March and August.



(a) Time series (b) Monthly average

Figure 7. *Time series and monthly average of flu incidence in South Asia.*

Figure 8(a) shows the time series of positive flu cases in the Western Pacific Region. As with the others, the series behavior shows strong annual seasonality. Figure 8(b) illustrates the monthly average of positive cases for each month in the Western Pacific Region. Although this region contains countries in the northern and southern hemisphere (for example, China and Australia), the graph shows the typical seasonal behavior of the northern hemisphere, since the vast majority of data comes from China. Because of this, the number of positive cases is greater in the months of December, January, February and March.



(a) Time series

(b) Monthly average

Figure 8. *Time series and monthly average of flu incidence in Western Pacific.*

The correlation matrix (Figure 9) aims to describe the association between the different regions considered in the work. In this case, the variables are the incidence of influenza in the seven regions. It is interesting to notice that the incidence in Brazil does not present any significantly correlation with the other regions. However, there is a positive correlation between the North America Region and the European Region (0.79), between the North America Region and the Western Pacific Region (0.70) and between the European Region and the Western Pacific Region (0.67).

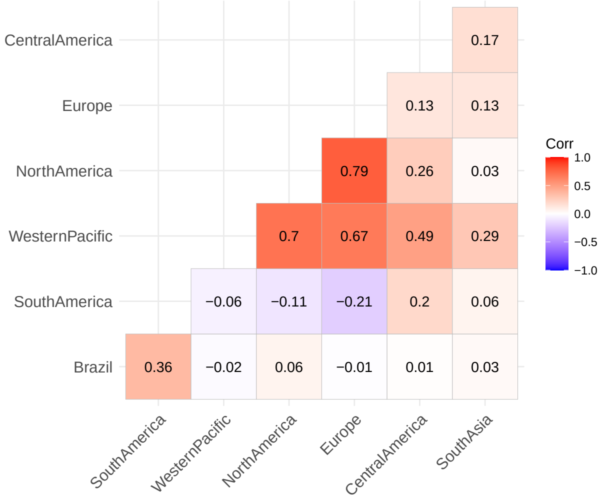


Figure 9. *Correlation matrix of influenza incidence in the different regions considered.*

The correlations between the incidences can be confirmed through a graphical analysis, which aims to compare the time series of the different regions. Figure 10 present a time series plot comparing the incidence in the Europe and North and South America Regions. It can be seen that the Europe and North America series have a very similar behavior and this justifies the high correlation between them (0.79). However, the time series behavior in South America is very different from the others, with peaks occurring in different epochs with a much smaller magnitude. As a result, its correlation with the European Region is only -0.21, and with North America, -0.11.

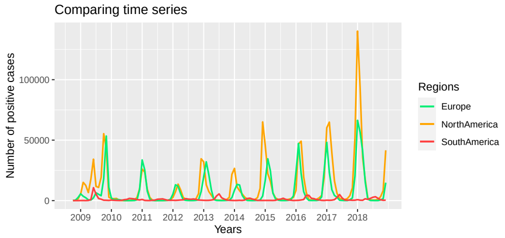
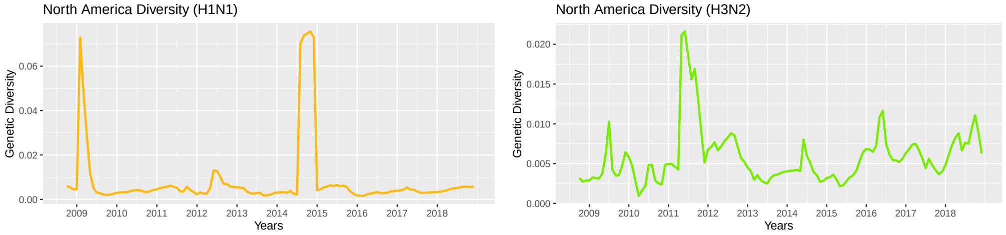


Figure 10. *Time series plot comparing the Europe, North and South America regions.*

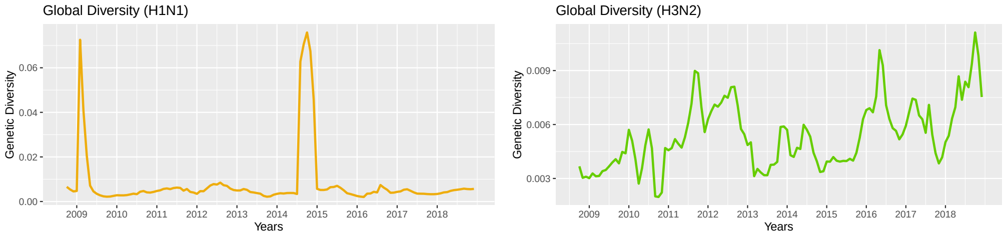
Figure 11 shows the genetic diversity of the H1N1 and H3N2 viruses in the North America Region. In Figure 11(a) (H1N1) we observe a high genetic diversity in 2009, due to the swine flu pandemic. There is a peak in 2014 that, according to **?**, was a period in which the circulating strains of the H1N1 virus caused unusually high levels of the disease in middle-aged adults, since the mutation of that period (2013-2014) was very particular, avoiding the immune responses in the group of adults. In addition, we observe that the dimension of the genetic diversity of the H1N1 virus is greater than that of the H3N2 virus. In Figure 11(b) (H3N2) we observe that there is a peak in the year 2011 and a small growing trend from the year 2013, but no identifiable seasonality.



(a) H1N1 (b) H3N2

Figure 11. *North America quarterly genetic diversity time series for the H1N1 and H3N2 viruses.*

Figure 12 shows the global genetic diversity of the H1N1 and H3N2 viruses. Figure 12(a), relative to the H1N1 subtype, shows a great increase in diversity in 2009, coinciding with the swine flu pandemic. There is a peak in diversity in 2014, reflecting the same process seen in North America. Furthermore, when comparing the global diversity graph with that of North America (H1N1), we notice that both graphs are very similar, partly consequence of the fact that a large portion of the global data on genetic diversity comes from North America.



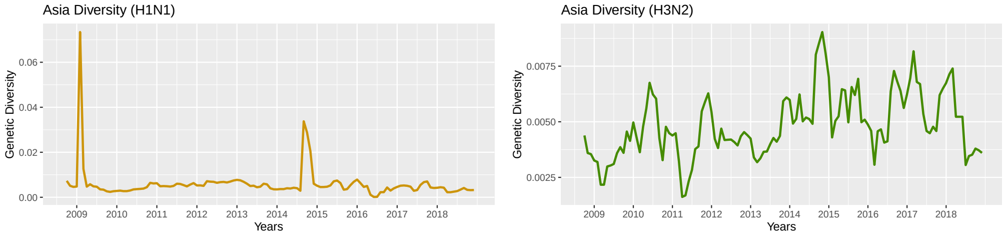
(a) H1N1

(b) H3N2

Figure 12. *Quarterly time series of global genetic diversity for the H1N1 and H3N2 viruses.*

Figure 12(b) shows the global diversity of the H3N2 virus. A growing trend is perceived over the years, indicating a possible adaptability of the virus. However, the impact is less than that of North America (H3N2), as the size of global genetic diversity is smaller. The series does not show constant mean and variance over time, indication that the time series behavior is non-stationary.

Figure 13 shows the genetic diversity of the H1N1 and H3N2 viruses in Asia. In Figure 13(a) (H1N1) there is a behavior similar to the other regions (North America H1N1 and global H1N1), except for the year 2014, where the peak of Asian genetic diversity is lower. Furthermore, the series does not present any trend or seasonality. In Figure 13(b) of the diversity of the H3N2 virus, there is a small increasing trend on the genetic diversity over time, which may again indicate an adaptability of the virus. However, the dimension of diversity in Asia is smaller when compared to the diversity of North America (H3N2).



(a) H1N1

(b) H3N2

Figure 13. *Asia genetic diversity quarterly time series for the H1N1 and H3N2 viruses.*