SORT 46 (2) July-December 2022, 161-188 DOI: 10.2436/20.8080.02.122

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]

Abstract

In this work we study the problem of modelling and forecasting the dynamics of the influenza virus in Brazil at a given month, from data on reported cases and genetic diversity collected from previous months, in other locations. Granger causality is employed as a tool to assess possible predictive relationships between covariates. For modelling and forecasting purposes, a time series regression approach is applied considering lagged information regarding reported cases and genetic diversity in other regions. Three different models are analysed, including stepwise time series regression and LASSO.

MSC: *62P10, 62J05, 62J07, 92D30.*

Keywords: *Flu, time series regression, variable selection, genetic diversity, Granger causality.*

1. Introduction

Caused by the influenza virus, the flu is one of the most prevalent diseases in Brazil and worldwide, infecting about 10% of the world's population every year and causing a toll between 250,000 and 500,000 deaths annually (Barr et al., 2010; Rambaut et al., 2008). It is characterized by an acute infection of the respiratory system. Common symptoms are cough, fever, headaches, throat and muscle pain (Eccles, 2005; Rambaut et al., 2008). Due to its severity, the World Heath Organization (WHO) actively surveys the virus through the Global Influenza Surveillance and Response System (GISRS) Network. The patterns of influenza incidence are influenced by seasonality and the emergence of new variants – new types of virus that infect humans for the first time thus managing to spread further due to reduced immunity in the population. According to the

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

[2] Corresponding author.

Received: January 2022
Accepted: May 2022

WHO, extensive vaccination against influenza is the most effective measure for its prevention (Barr et al., 2010). Public vaccination policies, therefore, become a fundamental agent in preventing serious epidemics and reducing the death toll from influenza.

Severe influenza cases require hospitalization and intensive care, including the need for artificial respirators. With the emergence of COVID-19, such precious assets have become scarce in many countries. Thus, moving forward, the forecast of influenza cases can help guide public health care systems in allocating resources and planning for seasonal concomitance of the two diseases.

A global dispersion process is responsible for seeding new variants that drive yearly influenza epidemics through most of the world. A frequent pattern is that new lineages affect the northern hemisphere first during the winter season. These variants tend to arrive latter in regions of the southern hemisphere such as South America and Oceania (Lemey et al., 2014; Petrova and Russell, 2018). This movement, if mathematically well described and statistically well modelled, has the potential to allow for predictions for the incidence of influenza, as well as the description of the strains expected to circulate in Brazil from data collected in Europe, Asia, and the United States during the winter season in the northern hemisphere. Such a forecast could be of great value for planning and implementation of public vaccination policies to reduce potential epidemics and minimize deaths due to influenza in Brazil.

In this paper, we study the problem of forecasting the number of influenza cases in Brazil at a given time t from data on influenza cases, as well as data related to genetic diversity, collected in other regions of the globe in preceding months.

2. The influenza virus

There are three common types of influenza viruses, influenza A, B and C, the first two being responsible for seasonal epidemics. The evolutionary dynamics of influenza A is composed by rapid mutation, natural selection and frequent rearrangement (Rambaut et al., 2008). Of the three types of viruses, type A is the one with the highest replication capacity in humans. Most of its cases occur in winter and in countries with temperate climates.

Influenza A type is subdivided into subtypes according to their surface proteins hemagglutinin (H1, H2 and H3) and neuraminidase (N1, N2). In this study, our goal is to investigate the behaviour of the two most recurrent subtype of influenza A, H1N1 and H3N2.

The H1N1 subtype appeared in 1918 causing the Spanish flu pandemic, one of the most deadly pandemics in history, affecting about a quarter of the world's population and responsible for tens of millions of deaths (Garten et al., 2009). The H1N1 flu virus reappeared in 1977 and subsequently its epidemics showed lower mortality rates when compared to the H3N2 epidemics (Rambaut et al., 2008). Then, in 2009 a pandemic of H1N1 occurred, widely known as the swine flu pandemic. The virus was first reported in Mexico, spreading across the world in the following months and infecting anywhere

between 700 million and 1.4 billion of people (Rambaut and Holmes, 2009). After the 2009 pandemic, the H1N1 virus continued circulating, being responsible for annual seasonal outbreaks with high mortality rates in Brazil. The new phylogenetic groups (of origin) of the H1N1 virus, seem to appear in the northern hemisphere, arriving in Brazil only in the seasonal outbreak of the following year (Silva, 2015).

The H3N2 subtype emerged in 1968 as the third pandemic of the 20th century called the Hong Kong flu and has dominated seasonal influenza A virus epidemics in recent years (Ibiapina, Costa and Faria, 2005). Born et al. (2016) found that the strains of the seasonal influenza A(H3N2) epidemics in South America are powered by a continuous introduction of viral variants from other geographic regions, especially from North America, and an extensive viral exchange among South American countries. They also found that the subtype tends to arrive in Brazil from neighbouring countries in South America, mainly through its south-east region.

2.1. Migratory dynamics

The source-sink model for global flu circulation states that tropical regions are the origin of new seasonal mutations. Genetic diversity is generated in these original populations, and then advances to the northern and southern hemispheres. Additionally, China is identified as the most likely epicentre of the flu A virus (Rambaut et al., 2008). More recent phylogeographic studies have found that there is substantially more viral flow between locations, and that the pattern does not adhere strictly the source-sink model. However the trunk of the phylogenetic tree, which represents the viral lineage that persists over time, was placed reliably, most of the time, in China, Southeast Asia and India. Viruses circulating in other locations do not usually last more than a season or two before being replaced by new lineages originating from the trunk (Petrova and Russell, 2018; Lemey et al., 2014; Bedford et al., 2010). Furthermore, strains are generally first spread to North America and Europe and only later to South America (Russell et al., 2008).

Influenza epidemics in temperate regions of the northern hemisphere typically occur between the months of November and March and in the southern hemisphere from May to September. Seasonality patterns in the tropics vary more according to location (Petrova and Russell, 2018). In Brazil, Almeida, Codeço and Luz (2018) identified that different regions have varying seasonality patterns, with stronger seasonality signals closer to the coast, peaks happening earlier towards the north of the country and later in the year in the southern region. Born (2013) studied the phylogeography of influenza in Brazil, identifying as unlikely that the origin of a new variant be located in Brazil. Additionally the main gateway for the H3N2 flu virus in the country would be the Southeast region followed by the South and Northeast regions. The existence of such global patterns which are repeated somewhat consistently throughout the years can be seen as motivation for seeking to forecast Brazilian incidence as a function of reported cases in other countries in the previous months.

3. Methods

3.1. Granger causality

In this study, the Granger causality method will be used to study the migratory dynamics of influenza (Granger, 1969). This method aims to determine the causal direction between two variables, stipulating that X_t Granger-causes Y_t if past values of X_t help to predict the present value of Y_t, and may provide better results than considering only the past Y_t. More specifically, it is a way of verifying whether a time series helps in predicting another series through VAR modelling. To use this method, the series need to be matched.

3.1.1. Vector Autoregressive Models (VAR)

The vector autoregressive model (VAR) is an extension of the autoregressive models (AR) and its objective is to model a vector time series considering only their past values (Sims, 1980). Mathematically, a k-dimensional $\boldsymbol{Y}_t$ stochastic process is said to be a VAR(p) process if it can be written as

$$\boldsymbol{Y}_t = c + A_1\boldsymbol{Y}_{t-1} + A_2\boldsymbol{Y}_{t-2} + \cdots + A_p\boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where $c \in \mathbb{R}^k$ is a vector of constants (intercepts), $A_1, \cdots, A_p$ are $k \times k$ matrices and $\boldsymbol{\varepsilon_t}$ is a k-dimensional error term.

3.1.2. Granger causality

The idea behind Granger causality (for univariate time series) is to consider the model

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i Y_{t-i} + \sum_{j=1}^{m} \alpha_j X_{t-j} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\varepsilon}_t$ denotes white noise. We say that X_t Granger-causes Y_t if past values of X_t help to predict the Y_t. In view of (1), to test whether X_t Granger-causes Y_t the following test can be performed:

$$H_0 : \alpha_1 = \cdots = \alpha_m = 0 \text{ vs. } H_1 : \alpha_s \neq 0, \text{ for at least one } s \in \{1, \cdots, m\}.$$

In the above test, rejection of the null hypothesis is considered evidence that X_t Granger-causes Y_t.

3.1.3. Granger Causality and Stationarity

Before applying the Granger causality method, it is necessary to check whether the series are stationary or not. A preliminary graphical analysis can assist in this matter. The absence of visible deterministic trends and/or apparent seasonality are indications of stationary behaviour. However, they are usually not enough for decision making,

which should preferably be done through appropriate tests such as the widely applied Augmented Dickey-Fuller (Dickey and Fuller, 1979) or the Phillips-Perron (Phillips and Perron, 1988) test. In these tests, the null hypothesis is that the time series has at least one unit root (i.e., the series is non-stationary) and the alternative hypothesis is the absence of unit roots. In this way, the time series will be considered stationary if the null hypothesis is rejected.

3.1.4. Granger causality for non-stationary series

One way to apply the Granger causality method if the series are not stationary is to use the Toda and Yamamoto procedure, introduced by Toda and Yamamoto (1995), which comprises the following steps:

1. Check whether the series cointegrate. Two series cointegrate if they have the same integration order, say m, and if the residual of regression from one series to the other are stationary, which can be determined using a test such as the Phillips-Perron.

2. Adjust a VAR(p) model.

3. Apply the Wald Test. In order to do so, it is necessary to fit a VAR($p+m$) model to the data. This model will certainly present several non-significant variables, given the previous steps, but this is not a problem since this model will not be used directly - it is only a device to guarantee the asymptotic theory. Rejection of the null hypothesis is evidence towards the existence of Granger causality in the tested direction.

Granger causality is a concept applied in many fields. In economics, Farias and Sáfadi (2010) employed Granger causality to study the relationship among the main stock exchanges in the world, showing how markets behave with each other and analyzing whether a market has a strong influence on the others. In agronomy, Diniz et al. (2009) studied whether certain agricultural and socio-economic variables (such as cattle and demographic density) Granger-cause deforestation in the Amazon. In biology, Chen et al. (2018) study the causal relationship between cases of influenza in humans and air pollution in Taiwan. The results indicated that pollution Granger-causes flu cases in the elderly group (over 64 years old).

3.2. Variable selection in regression models

Variable selection is a central topic in regression models involving many covariates. In this section we review some of the available techniques for variable selection which will be used here.

3.2.1. Stepwise regression

Stepwise regression is an automatic tool that aims to select the most influential independent variables in a given model. It is an iterative method that adds or removes variables according to a given selection criterion. The most popular types of stepwise regression are the forward-stepwise and backward-stepwise. In this paper, the backward-stepwise selection method was preferred due to the model size. We consider the p-value based stopping criterion, which selects variables according to their Wald statistics, eliminating non-significant terms based on the magnitude of their p-values (higher p-values are preferred in removing terms), in order to obtain a model for which all variables are significant.

3.2.2. LASSO

The LASSO (least absolute shrinkage and selection operator) regression is a penalty method that aims to provide smaller and more parsimonious models (Hastie, Tibshirani and Friedman, 2009). The penalty is applied to the coefficients to decrease the number of parameters and, consequently, reduce the dimension and uncertainty in the model. It is a regression method that aims to reduce the dimensionality and improve the accuracy of the forecast and the interpretability of the resulting model.

4. Data

In this section we provide detailed information regarding the data used in our study. The Supplementary Material presents a detailed exploratory analysis of the data.

4.1. Number of positive flu cases

The data for number of positive flu cases was taken from FluNet, an online tool maintained by the World Health Organization (WHO, 2020) whose objective is to aggregate influenza virological surveillance data, launched in 1997. FluNet data comes from weekly country reports of the number of tested cases, the number of positive cases and the type of virus. Typically the reported data refer to data collected in a few reference centres in each country, and do not represent the actual flu incidence data. Since the number of positive cases are expected to correlate with influenza incidence, for the purpose of this paper it is considered as a proxy for incidence. Thus, such data will be referred to here as influenza incidence data. For this project, data from H1N1 and H3N2 influenza were collected from January 2008 to November 2019, but due to missing data problems in 2008, the data used in the analysis cover the period from October 2008 to November 2019. However, for modelling purposes, we only use data from October 2008 to December 2018, which yields a sample size $n = 123$, while data from January 2019 to November 2019 is reserved for out-of-sample forecasting purposes.

Table 1. *Aggregated regions.*

Region	Countries
Europe	Belgium, Switzerland, Spain, Estonia, Germany, Ireland, Israel, Italy, Latvia, Netherlands, Norway, Poland, Russian Federation, Slovenia, Sweden, Turkey, Denmark, United Kingdom of Great Britain, Northern Ireland
North America	Canada and United States
South America	Argentina, Bolivia, Chile, Colombia, Ecuador, French Guiana, Paraguay, Peru
Central America	Costa Rica, Cuba, Dominican Republic, El Salvador, Guatemala, Honduras, Jamaica, Mexico, Nicaragua, Panama
South Asia	India, Thailand, Indonesia, Bangladesh, Bhutan, Nepal, Sri Lanka
Western Pacific	China, Japan, Australia, Republic of Korea, Singapore, Malaysia, Vietnam, New Caledonia, Philippines, Cambodia, Lao People's Democratic Republic

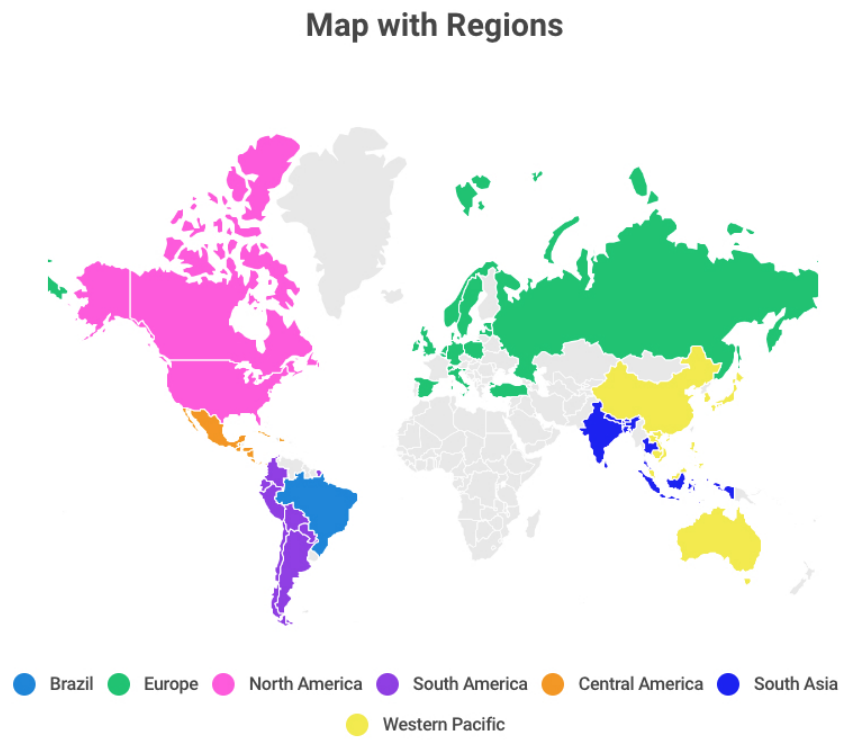


Figure 1. *Map with the regions considered in the study.*

For simplicity, we aggregate data geographically into regions based on WHO regions, with the exception of Brazil which is the focus of this study. The following locations were considered: Brazil, North America, South America (without Brazil), Central America, Europe, South Asia and Western Pacific. Note that each region is composed of

a certain number of countries, responsible for reporting their data in the database. However, due to local characteristics, several countries had a high amount of missing data, often above 50%, resulting in useless local data for our purposes. In order to make the analysis feasible, all countries presenting more than 50% missing data were excluded in the construction of the database for the respective region. Table 1 and Figure 1 present the configuration of each region after applying this criterion.

The time series techniques that we used require that the time series do not contain any missing data. To resolve this, we applied an imputation method, which aims to fill in the missing data using the following criteria: multiply the average regional number of positive flu cases of the respective week by the proportion that the respective country represents in the region. In some cases, however, it happened that the week had missing data in all countries, and this was resolved by imputing it through the average between the previous and subsequent weeks. After imputation, the data were aggregated monthly.

4.2. Genetic diversity

For the genetic diversity data, viral RNA sequences were collected from the NCBI Influenza Virus Database, which compiles a comprehensive assortment of influenza sequences generated by research groups around the world (NCBI, 2020). The genetic dataset was assembled considering all complete chromosome 4 (hemagglutinin gene) sequences from human influenza A viruses in the database, from all continents and in the interval from October 2008 to September 2019. The data were retrieved in March 08, 2020. This resulted in a total of 16,008 H1N1 sequences and 15,418 H3N2 sequences. As the goal was to measure genetic diversity of viral populations, H1N1 and H3N2 sub-types were treated separately. Influenza B sequences were excluded from this study because of insufficient data at many time points.

Genetic diversity is a population measure that seeks to quantify viral variability, allowing for comparisons over time or between populations. For the sake of simplicity, throughout this paper let the genetic diversity of a viral population be defined as the average genetic distance between all sequences in the population. The distance measure considered here is the K80 distance (Kimura, 1980), which is based on nucleotide substitutions, thus the sequences must be aligned so that individual mutations can be identified. Due to the size of our dataset, the online tool MAFFT (Yamada, Tomii and Katoh, 2016) was used to generate the alignments.

The aligned sequences were then used to build a distance matrix between individual sequences in the dataset. This resulted in a symmetric $n \times n$ matrix $\mathbf{D}$ with entries $d_{i,j}$ denoting the genetic distances between the sequences i and j, where $i, j \in \{1, \cdots, n\}$.

Finally, genetic diversity was computed for temporally and geographically defined sub-populations, by averaging over all relevant entries in the distance matrix. As suggested in Jesus (2018), virus diversity was assessed using a quarterly moving average scheme, since a three-month window size best captured smooth diversity fluctuations over time for these data. This calculation was made for each month in the range from October 2008 to September 2019, separately for H1N1 and H3N2, and for the following

regions: Asia, North America and global (all continents). Other regions were not considered due to insufficient data. See the Supplementary Material for a list of the countries comprising each region. Similarly to the incidence data, for modelling purposes we only consider data from October 2008 to December 2018 ($n = 123$), while data from January 2019 to November 2019 are reserved for out-of-sample forecasting purposes. All code was written in `R` (version 4.0.0, R Core Team, 2020) and is available (along with the relevant data) at github.com/AlineFoersterGrande/Flu_Paper.

5. Results

In this section we present the results of our analysis. We separate the different analyses by technique.

5.1. Granger causality

5.1.1. Positive influenza counts

In this section we present a Granger causality analysis of the number of positive flu cases in Brazil considering data from the other regions. Our main interest is to verify if the number of cases in Brazil can be explained by the recent historical data from other regions. In all situations, the data were considered non-stationary due to the clear seasonal pattern present (the time series plots are presented in the Supplementary Material). The global task resulted in 66 comparisons. The p-values presented in this section were corrected for false discovery rate, implemented through the function `p.adjust` in R (R Core Team, 2020). On performing step 2 of Toda and Yamamoto's procedure, we consider $p = 6$ as the maximum lag to adjust the VAR(p) model to the data.

Table 2. *Granger causality results for the number of flu cases - Brazil case.*

Null hypothesis	p-value	Lag
North America Region does not Granger-cause Brazil	0.84	–
European Region does not Granger-cause Brazil	**0.10**	3
Central America Region does not Granger-cause Brazil	0.99	–
South America Region does not Granger-cause Brazil	**0.05**	2
South Asia Region does not Granger-cause Brazil	0.78	–
Western Pacific Region does not Granger-cause Brazil	0.98	–

Table 2 presents the results of the Granger causality analysis. We conclude that among all regions, only Europe (at 10% significance) and South America (at 5% significance) Granger-cause Brazil. This result suggests that the historical data on the number of cases of flu in the European and South America Regions are helpful in predicting the present value of the incidence in Brazil.

As presented in Section 2.1, the trunk of the phylogenetic tree and source of most seasonal variation for influenza is associated to the Asian continent, particularly China,

from where the virus frequently migrates to the northern hemisphere. With this in mind, we perform a Granger causality analysis to verify if the Western Pacific Region (the region that contains the majority of the data from Asia), Granger-causes the regions in the northern hemisphere.

Table 3 presents the significant cases. We find that the Western Pacific Region Granger-causes the regions of Europe and South Asia. Therefore, it can be said that the historical data related to the number of flu cases in the Western Pacific Region helps to predict the present incidence of the South Asian and European Regions.

Table 3. *Granger causality results for the number of flu cases - Western Pacific case.*

Null hypothesis	p-value	Lag
Western Pacific does not Granger-cause South Asia	0.05	3
Western Pacific does not Granger-cause Europe	0.04	3

Note that, although there is no direct evidence that the number of flu cases in the Western Pacific Region Granger-causes the incidence in Brazil, there is an indirect effect of the Pacific Region in Brazil, since the Pacific Granger-causes the European Region which in turn, Granger-causes the incidence in Brazil. This indirect effect was not directly detected because the Granger causality analysis is not transitive. Finally, we investigate whether any region Granger-causes another region. The results are presented in Table 4.

Table 4. *Granger causality results for the number of flu cases - all regions.*

Null hypothesis	p-value	Lag
Central America does not Granger-cause Europe	0.05	3
Central America does not Granger-cause Western Pacific	0.00	3

Note that the results in Table 4 indicate that the incidence in Central America Granger-causes the incidence of the European and Western Pacific Regions. The most likely justification for the Granger causality of Central America in other regions is the occurrence of the swine flu (H1N1) in the years 2009 and 2010. Mexico is considered the origin of the swine flu pandemic, which justifies the increase in the incidence of influenza first in the Central America and then in the other regions.

5.1.2. Genetic diversity

In this section, the data described in subsection 4.2 (genetic diversity) are used to examine the influence among different regions under the prism of Granger causality. The initial interest is to verify whether the number of flu cases in Brazil can be explained by the past genetic diversity data from other regions. Granger causality tests were used to assess whether H1N1 and H3N2 genetic diversity in North America, Asia and around the globe (termed All) affect Brazilian incidence. The results are shown in Table 5.

Table 5. *Granger causality results considering genetic diversity data as covariate and number of cases in Brazil as response.*

Null hypothesis	p-value
North America (H1N1) does not Granger-cause Brazil	0.85
All (H1N1) does not Granger-cause Brazil	0.84
Asia (H1N1) does not Granger-cause Brazil	0.84
North America (H3N2) does not Granger-cause Brazil	0.69
All (H3N2) does not Granger-cause Brazil	0.42
Asia (H3N2) does not Granger-cause Brazil	0.76

It can be seen that no genetic diversity Granger-causes Brazil, that is, the measurement of genetic diversity does not help in predicting the present value of the incidence of influenza in Brazil. Given these results, a second analysis was performed to verify whether the incidence of other regions can be explained by the genetic diversity data. Table 6 presents the all statistically significant results.

Table 6. *Granger causality results considering the genetic diversity data as covariate and number of cases or genetic data in other regions as responses.*

Null hypothesis	p-value	Lag
North America (H1N1) does not Granger-cause South Asia (cases)	0.05	3
Asia (H1N1) does not Granger-cause Central America (cases)	0.00	6
North America (H1N1) does not Granger-cause Asia (H1N1)	0.01	2
All (H1N1) does not Granger-cause North America (H1N1)	0.00	2
All (H1N1) does not Granger-cause Asia (H1N1)	0.00	2

We conclude that the genetic diversity of Asia (H1N1) helps in predicting the incidence of influenza in the Central American Region. Furthermore, it shows that North American genetic diversity (H1N1) Granger-causes the South Asian cases and H1N1 diversity on Asia.

5.2. Time series regression approach

In this section we present time series regression analysis of the data presented in Sections 4.1 and 4.2. A classical ARMA approach to model incidence data is presented in the Supplementary Material.

5.2.1. Number of positive flu cases

We start by considering the data described in Section 4.1 (number of positive flu cases) to represent the flu incidence in Brazil (denoted by B_t), in Europe (E_t), in North America (A_t), in Central America (C_t), in South America (S_t), in South Asia (s_t) and in Western

Pacific (W_t) at time t. We applied historical data of the regions considered in the last 11 months (lags), denoted by $B_{t-1},\cdots,B_{t-11}$ for Brazil, $A_{t-1},\cdots,A_{t-11}$ for the North America and similarly for other regions. We also considered a covariate μ_t representing the monthly average number of positive flu cases in Brazil at time t, $t \in \{1,\cdots,123\}$, calculated as

$$\mu_t = \frac{1}{\#I_t}\sum_{k\in I_t} B_k,$$

where, I_t denotes the set of time indexes in $\{1,\cdots,123\}$ corresponding to the same month as t and $\#I_t$ denotes the cardinality of I_t. Notice that only observed values were used to calculate μ_t. For modelling purposes, B_t was the response variable, while lagged data from Brazil and all other regions were used as covariates.

5.2.2. Modelling

Due to the large number of explanatory variables, we considered three different methods to fit the data, chosen because of their ability to perform variable selection. The first one was the stepwise backward regression method based on p-values with significance level 0.1, denoted simply by Stepwise model. We also applied the LASSO model considering two main schemes for model selection: first, based on cross-validation (leave-one-out), denoted LASSO CV, which yielded a model with 13 covariates plus the intercept;

Table 7. *Estimation results for the fitted Stepwise, LASSO 5 and LASSO CV models.*

Variables	Stepwise	LASSO 5	LASSO CV
Intercept	-0.843	60.821	44.089
B_{t-1}	0.84700	0.52843	0.68871
B_{t-2}	-0.34248	–	-0.15364
A_{t-2}	-0.00414	–	–
A_{t-4}	–	0.00019	0.00086
C_{t-1}	0.01515	–	0.00095
C_{t-3}	-0.02490	–	-0.01308
E_{t-1}	0.00535	–	0.00180
E_{t-2}	0.01117	0.00548	0.00650
E_{t-3}	–	0.00242	0.00084
E_{t-4}	0.00642	–	0.00230
E_{t-8}	0.00255	–	–
μ_t	–	0.01431	–
s_{t-4}	–	–	-0.00173
s_{t-9}	–	–	0.00066
W_{t-7}	–	–	-0.00158
W_{t-9}	–	–	0.00138

and second, since this model is somewhat large, a more parsimonious alternative using the "one-standard-error" rule (Hastie et al., 2009, section 7.10), selecting a model with five variables denoted by LASSO 5. Table 7 presents the covariates selected by each model and their fitted values. The intercept was always kept.

Notice that variables B_{t-1} (number of positive cases in Brazil at time $t-1$) and E_{t-2} (number of positive cases in Europe at time $t-2$) are the only variables present in all fitted models. Another way of interpreting the results is by analysing the coefficients of each variable present in the final model. It shows the direction of the impact that the explanatory variables have on the response variable B_t. For example, the explanatory variable B_{t-1}, which is included in all models, has a positive coefficient. This indicates that as the number of cases in Brazil at time $t-1$ increases/decreases, so does the number of cases in Brazil at time t. Also noteworthy is that the variable μ_t appears only in the LASSO 5 model.

5.2.3. Forecast

After modelling, we perform an in-sample and out-of-sample forecast exercise for the data considering each fitted model. Data from October 2008 to December 2018 were used for modelling purposes, while data from January 2019 to November 2019 were reserved for out-of-sample comparison. Hence, the forecast horizon in all cases is $h = 11$ steps ahead.

Notice that, since we are using a time series regression approach with several past values of regressors entering in the final model, these values must be updated if out-of-sample forecast values are to be obtained (that is, in order to obtain future values of the response variable, we need future values for the covariates as well). In order to do that we employed two approaches. The first approach employed is known as h one-step ahead forecast. In this case, for each incremental step ahead we updated the covariates with their observed values. This is only useful for small forecast horizons or to forecast short run dynamics, as in the case of flu data.

In the second approach, known simply as h-steps ahead forecast, we did not use any knowledge about future values of the covariates. Instead we forecasted their values using some plausible method. Of course, there are several ways to do that. We forecasted future values of the covariates by using their monthly average calculated from the observed data, including Brazil. This second approach can be employed for forecasting in practice. Figures 2 to 4 show both, the in-sample and out-of-sample (for $h = 11$) one-step ahead forecast values (the regressor values are updated at each step, including out-of-sample) compared to the observed ones (in black).

It can be seen that both in-sample and out-of-sample predictions appear to be reasonable for all models, except in a few epochs, such as the year 2011, where the models predicted a peak that did not occur, or the first half of 2015, where the peak was overestimated by all models. Since we are using a non-restricted time series approach to model the data, we obtain a few negative values for the incidence, located at the valleys. These negative values are not considered a problem because the main focus of the study of flu

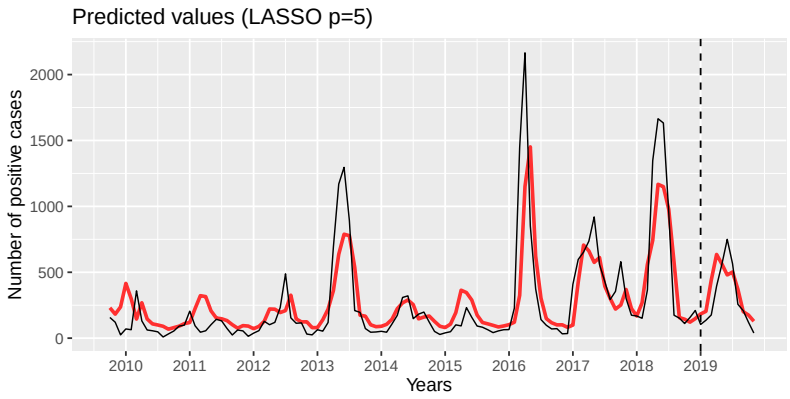


Figure 2. *In-sample and out-of-sample one-step ahead forecasts for the LASSO 5 model.*

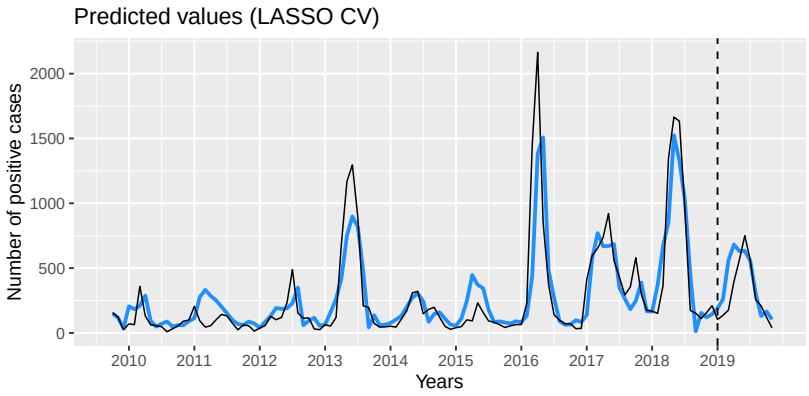


Figure 3. *In-sample and out-of-sample one-step ahead forecasts for the LASSO CV model.*

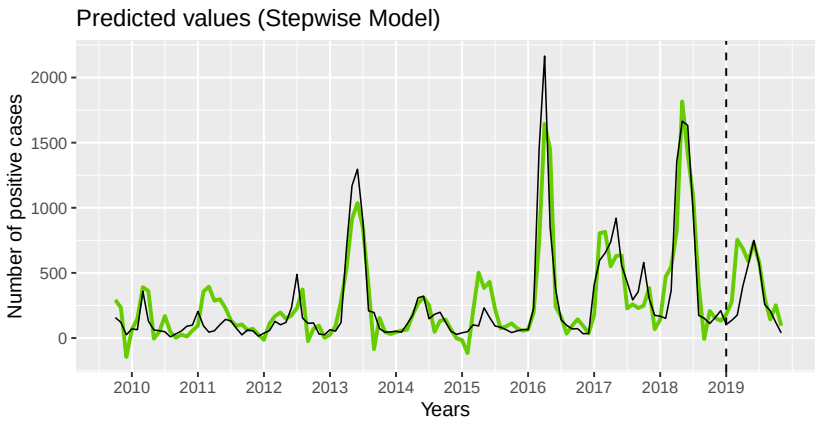


Figure 4. *In-sample and out-of-sample one-step ahead forecasts for the Stepwise model.*

pandemics are the peaks of the curve, not its valleys. Table 8 presents in-sample and out-of-sample mean square error (MSE) and mean absolute percentage error (MAPE) of forecasting. The best results in each case are highlighted in red.

Table 8. *Mean square error and mean absolute percentage error of the in-sample and 11 one-step ahead forecasts for each of the 3 fitted models.*

Measures/Models	Stepwise	LASSO 5	LASSO CV
MSE (in-sample)	31586.7	52788.1	38354.5
MSE (out-of-sample)	43228.2	21805.3	25860.9
MAPE (in-sample)	91.2	105.2	81.5
MAPE (out-of-sample)	81.2	64.3	68.7

The Stepwise model and the LASSO CV were the best performers in-sample, while for out-of-sample, the LASSO 5 model performed best both in terms of MSE and MAPE.

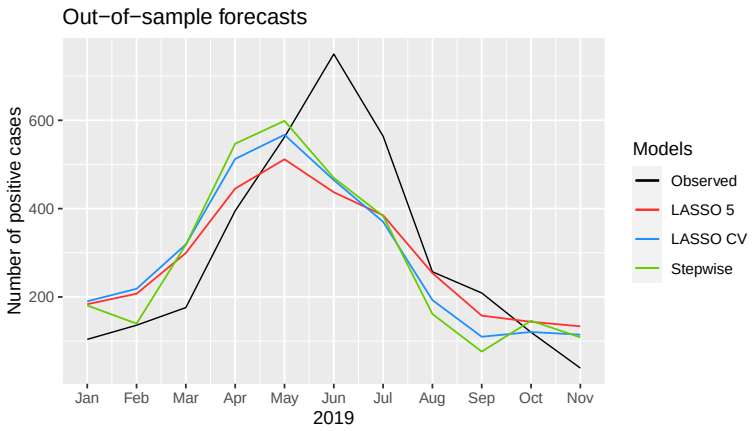


Figure 5. *11-steps ahead forecasts for the different fitted models compared to the observed values (in black).*

In a second moment, we analyze the out-of-sample h-steps ahead forecast ability of the fitted models, for $h \in \{1, \cdots, 11\}$. Figure 5 presents the forecasted values of each fitted model along with the observed values (in black). From Figure 5 we observe that all models overestimate the number of positive cases until May/April, missing the peak that occurred in June and underestimating the number of cases from June to October. For comparison purposes, Table 9 presents the mean square error for h-steps ahead forecasts for each model. The best results in each forecast horizon are highlighted in red. The model presenting the overall best performance was the LASSO 5, which presented the lowest MSE in 7 out of the 11 forecast horizons considered, followed by the Stepwise which presented overall smallest MSE in the remaining 4 forecast horizons. Interestingly, the LASSO 5 uniformly outperforms the LASSO CV in all forecast horizons.

This poor forecasting performance of the LASSO CV may be attributed to overfitting. Stepwise and LASSO CV presented similar performances.

Table 9. *Mean squared error of the h-steps ahead forecast for each fitted model. The best forecast in terms of MSE for each horizon is presented in red.*

Horizon/Models	Stepwise	LASSO 5	LASSO CV
1-step ahead	5904.13	6346.41	7417.52
2-steps ahead	2958.38	5712.86	7110.19
3-steps ahead	8736.86	8895.71	11549.42
4-steps ahead	12323.75	7305.05	12114.87
5-steps ahead	10129.28	6349.04	9697.23
6-steps ahead	21543.97	21599.59	21646.22
7-steps ahead	23130.21	23092.44	23917.95
8-steps ahead	21381.31	20206.98	21442.05
9-steps ahead	20962.64	18254.56	20149.43
10-steps ahead	18934.49	16485.37	18134.53
11-steps ahead	17653.39	15798.62	17005.14

5.2.4. Genetic diversity

In this section we consider both the number of positive flu cases and the genetic diversity data (described in sections 4.1 and 4.2) to characterize the flu incidence in Brazil. We apply similar notation to Section 5.2.1. The genetic diversity in North America at time t is denoted by N_t for the H1N1 subtype and n_t for the H3N2 subtype, in Asia by P_t for the H1N1 subtype and p_t for the H3N2, and M_t (H1N1) and m_t (H3N2) denote the global genetic diversity. Again the response variable is taken as B_t while lagged variables related to other regions, including genetic data, will be used as covariates.

We aim to explain the incidence of influenza in Brazil at time t (B_t) by using the historical incidence and genetic diversity data of the regions considered in the last 6 months (lags). The same three methods of Section 5.2.1 were considered. Table 10 presents the selected covariates and their respective coefficients, for each model.

From Table 10, we observe that the incidence variables that appear in all three models are B_{t-1} (number of positive cases in Brazil with one lag), E_{t-2} (number of positive cases in Europe with two lags) and A_{t-4} (number of positive cases in North America with four lags). Furthermore, when analyzing the variables related to genetic diversity, we observe that the covariates P_{t-4} (genetic diversity of the H1N1 flu in Asia with four lags) and P_{t-5} (genetic diversity of the H1N1 flu in Asia with five lags) appear in two of the tree models. After model fitting, we proceed with an in-sample and out-of-sample forecast analysis similar to the one presented in Subsection 5.2.3. To perform the out-of-sample analysis, it is necessary to forecast future values of covariates entering the model. Covariates related to incidence are forecasted in the same way as in Subsection 5.2.1. The genetic

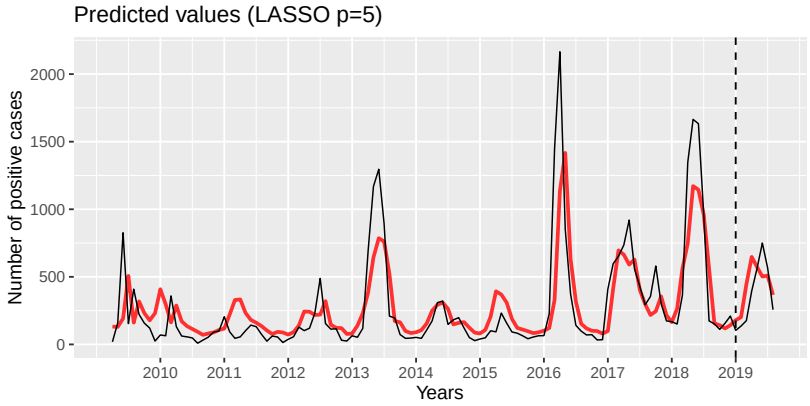


Figure 6. *In-sample and out-of-sample one-step ahead forecasts for the LASSO 5 model.*

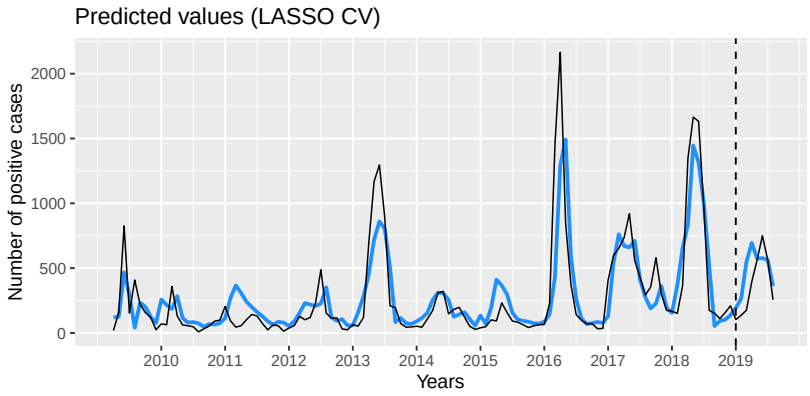


Figure 7. *In-sample and out-of-sample one-step ahead forecasts for the LASSO with cross-validation model.*

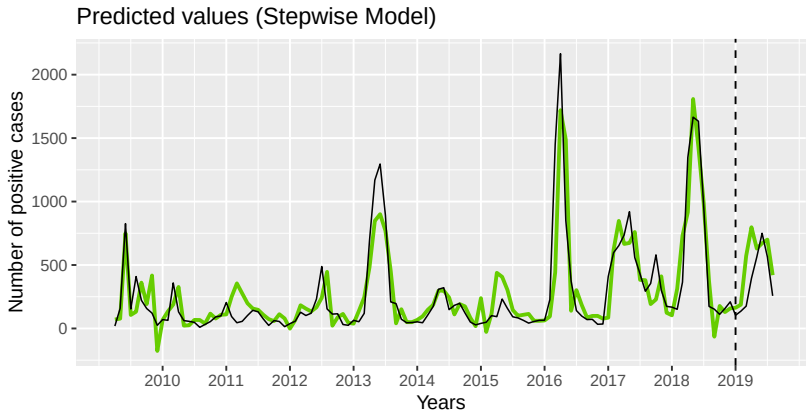


Figure 8. *In-sample and out-of-sample one-step ahead forecasts for the Stepwise model.*

Table 10. *Estimation results for the fitted Stepwise, LASSO 5 and LASSO CV models.*

Variables	Stepwise	LASSO 5	LASSO CV
Intercept	38.789	54.872	55.460
B_{t-1}	0.85717	0.50499	0.60024
B_{t-2}	-0.34293	–	-0.03383
B_{t-3}	–	–	-0.04411
A_{t-4}	0.00286	0.00037	0.00166
C_{t-2}	0.02204	–	–
C_{t-3}	-0.03545	–	-0.00946
S_{t-2}	0.02604	–	–
E_{t-1}	–	–	0.00161
E_{t-2}	0.00893	0.00535	0.00609
E_{t-3}	–	0.00219	0.00086
E_{t-4}	–	–	0.00053
s_{t-4}	–	–	-0.00957
W_{t-6}	–	–	-0.00094
μ_t	–	0.07586	0.05940
M_{t-5}	–	–	-837.80
P_{t-4}	6482.74	–	3827.28
P_{t-5}	-9221.79	–	-3156.14

diversity time series, however, do not present any evident trend or seasonality, as in the incidence data (see the time series plots presented in the supplementary material). Hence, the same approach of considering monthly averages is not adequate for the genetic diversity data. To overcome this difficulty, we consider a static approach: future values of genetic data are forecasted considering the average of the respective data observed from January to December, 2018. Figures 6 to 8 show the one-step ahead forecasted values in and out-of-sample for each model along with the observed values (in black).

Table 11. *Mean square error and mean absolute percentage error of forecast for each model.*

Measures/Models	Stepwise	LASSO 5	LASSO CV
MSE (in-sample)	34567.0	55306.2	42026.3
MSE (out-of-sample)	47623.4	27781.2	36757.7
MAPE (in-sample)	91.2	110.9	89.8
MAPE (out-of-sample)	66.5	53.1	66.7

Note that, in general, the in-sample and out-of-sample predictions appear to be reasonable for all considered models. Some peaks, such as the ones in years 2013, 2016 and 2018, are underestimated by the models, while others, such as 2011 and 2015 are overes-

timated. To compare the models regarding their predictive abilities, Table 11 presents the MSE and MAPE for the one-step ahead forecast for each model, in and out-of-sample. The best results in each case are highlighted in red.

Analogously to the results obtained in Section 5.2.1, analyzing the MSE we observe that the Stepwise model and the LASSO CV are the best perform in terms of in-sample forecast while the LASSO 5 is the best performer out-of-sample. Again, the main source of forecast error are a few peaks in the data not very well identified by any of the models, more noticeably, 2011, 2015 and 2017.

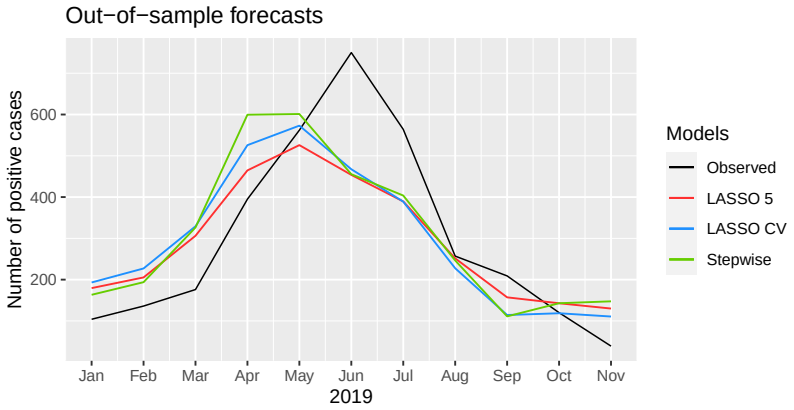


Figure 9. *11-steps ahead forecasts for the different fitted models compared to the observed values (in black).*

Table 12. *Mean squared error of the h-steps ahead forecast for each fitted model. The best forecast in terms of MSE for each horizon is presented in red.*

Horizon/Model	Stepwise	LASSO 5	LASSO CV
1-step ahead	3512.23	5701.45	7945.86
2-steps ahead	3428.68	5255.09	8122.44
3-steps ahead	9823.78	9153.06	13233.38
4-steps ahead	17825.10	8076.48	14200.18
5-steps ahead	14567.40	6721.95	11384.56
6-steps ahead	26605.43	20228.37	22796.37
7-steps ahead	26475.39	21693.13	23918.67
8-steps ahead	23179.94	18985.19	21037.13
9-steps ahead	21671.71	17174.71	19696.15
10-steps ahead	19556.80	15509.01	17726.74
11-steps ahead	18847.90	14851.20	16581.80

In a second step, we analyse the models' predictive capabilities considering forecast horizons from 1 to 11-steps ahead, in the same spirit as in Subsection 5.2.1. Figure 9

presents the forecasted values for the different models, as well as the observed values (in black). We observe that all models overestimate the number of positive cases until May/April, missing the peak that occurred in June and underestimating the number of cases from June to October. A comparison between the models is presented in Table 12, where we present mean squared error for each considered h-steps ahead forecast. The results show that no model uniformly outperforms all others. The model with the best results was the LASSO 5, which displayed the lowest MSE in 9 out of the 11 forecast horizons considered. Again the LASSO CV is uniformly outperformed by LASSO 5, while compared to Stepwise, the LASSO CV wins in middle to long horizons.

5.2.5. Comparison of forecasts

We now compare the results presented in Subsections 5.2.1 and 5.2.4. The interest lies in comparing the models with only incidence data with the models with incidence and genetic diversity data, regarding their predictive power. In the graphs below, the term "Incidence" will be used for models containing only incidence data while the term "Genetic" will be used for models considering incidence and genetic diversity data. Figure 10 shows the MSE obtained in the out-of-sample forecast for all models. It can be seen that in all cases the models based on "Incidence" presented more accurate forecasts (lower MSE). Furthermore, the LASSO 5 proved to be the overall best model in terms of prediction capabilities in all cases.

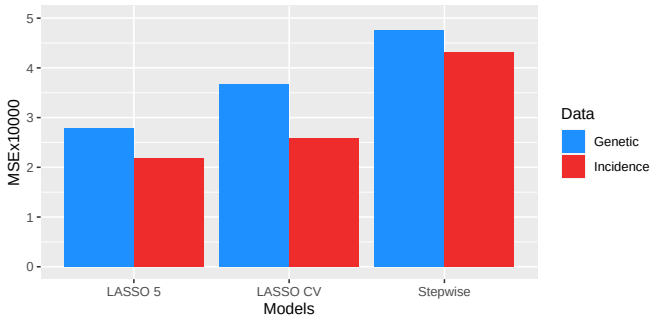
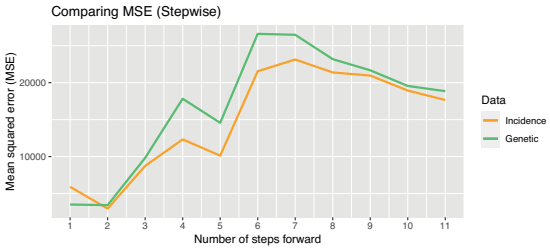


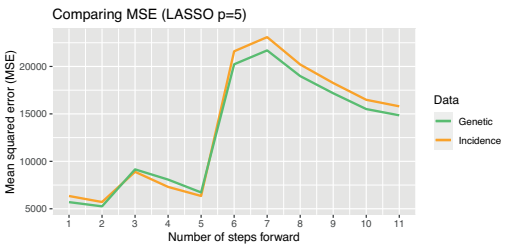
Figure 10. *Comparison of the mean squared errors (MSE) between the incidence data and the genetic data in the 11 one-step ahead forecasts.*

Figure 11 presents time series plots of the MSE for h-steps ahead forecast for each model considering the incidence and genetic data. For the Stepwise model, the out-of-sample forecasts produced using the incidence data present smaller MSE in all horizons but $h = 1$. For the LASSO, the models based on the genetic data presented smaller MSE in the long run, that is, for all horizons $h \geq 6$ for the LASSO 5 and $h \geq 7$ for the LASSO CV. In the short run, for the LASSO CV, the model based on incidence data performs best, while there is no clear pattern in the case of the LASSO 5 model. Ultimately, this indicates that including genetic diversity data, at least as measured here, does not seem to add much predictive value to the models. However, there are many other approaches to

assess genetic diversity that can be explored and might prove more valuable for incidence modelling.



(a) Stepwise



(b) LASSO 5

(c) LASSO CV

Figure 11. *Comparison of the h-steps ahead forecasts mean squared errors between the incidence and genetic data for the Stepwise (upper panel), LASSO 5 and LASSO CV (lower panel) models.*

5.3. Residual analysis

Residual analysis is of paramount importance in time series analysis, being performed after model identification and fitting. In this section we present a residual analysis related to the models fitted in the previous sections, focusing mainly in portmanteau and normality tests. Observe, however, that the only model that actually requires a residual analysis is the Stepwise, as it is the only one based on p-values. Nevertheless, for the sake of exploration, we shall proceed with the residual analysis for all models. To assess the presence of correlation in the residuals, we perform the widely applied Ljung-Box test (Ljung, 1986). Recall that the null hypothesis for the Ljung-Box test is that all correlations up to a specified lag m are null. In this analysis we consider $m = 20$. We also test the residuals for normality by using Shapiro-Wilk's test (Shapiro and Wilk, 1965), for which the null hypothesis is that the tested sample comes from a normally distributed population.

The Ljung-Box test's results for the residuals of all models presented in Sections 5.2.1 (flu incidence) and 5.2.4 (genetic data) are presented in Table 13. From the results we conclude that in all cases the residuals present no correlation up to lag $m = 20$, at any reasonable significance level.

Table 13. *p-values of the Ljung-Box test applied to the fitted models' residuals with* $m = 20$.

Dataset	Models		
	Stepwise	LASSO 5	LASSO CV
Incidence	0.9561	0.2096	0.8212
Genetic	0.9922	0.3805	0.4770

As for the Shapiro-Wilk test, it is clear from the in-sample forecasts (Figures 2 to 4 and Figures 6 to 8) that the residual will present outliers due to underestimation of peak values. These outliers may substantially affect the Shapiro-Wilk test. To minimize this effect, we removed some of the outliers by using two hard thresholds: we eliminate any points with magnitude larger than 400 (threshold 1) and 200 (threshold 2), in absolute value. Table 14 summarizes the results by presenting the p-values of the Shapiro-Wilk test with and without the removal of outliers, along with the number of outliers removed in each case. From the results we observe that the residuals of all models reject the null hypothesis in the Shapiro-Wilk test with very small p-values. The Stepwise model for the incidence data is the only one that do not reject the null hypothesis in the Shapiro-Wilk's test after applying threshold 1, which trimmed out only 5 points. The LASSO CV model for all data and Stepwise with genetic data did not reject at the 0.05 significance level the null hypothesis in Shapiro-Wilk's test after applying threshold 2, at the cost of removing several points. The Shapiro-Wilk's test applied to the residuals from the LASSO 5 model rejected the null hypothesis in all cases.

Table 14. *p-values for the Shapiro-Wilk test applied to the complete residuals and upon removing points with magnitude greater than 400 and 200, in absolute value. The number of points removed for each threshold applied is presented in parenthesis.*

Dataset	Threshold	Models		
		Stepwise	LASSO 5	LASSO CV
Incidence	complete	< 0.0001	< 0.0001	< 0.0001
	400	0.3012(5)	< 0.0001(9)	0.0003(5)
	200	–	0.0089(23)	0.1881(19)
Genetic	complete	< 0.0001	< 0.0001	< 0.0001
	400	0.0046(4)	< 0.0001(9)	< 0.0001(6)
	200	0.2357(26)	0.0015(28)	0.0952(25)

Finally, another important diagnostic is the homoscedasticity of the residuals for the fitted Stepwise model, the only one of our procedures that relies on distributional assumptions for model selection. Figure 12 presents simple time series plot and observed vs. fitted values for the residuals obtained from the Stepwise model considering the Incidence and Genetic data. From the time series plot we observe a clear increase in variance in both residuals, also evident in the residual versus fitted models. These findings are corroborated by Breusch-Pagan and White's tests (see Greene, 2012, section 11.4)

(not shown). The presence of heteroscedasticity in the model's residuals may affect the p-values obtained from Wald's test, implying that the fitted model may be incorrectly specified in the sense that the procedure may have excluded important variables, included unimportant ones, or both. This, however, does not diminish its applicability as a predictive model.

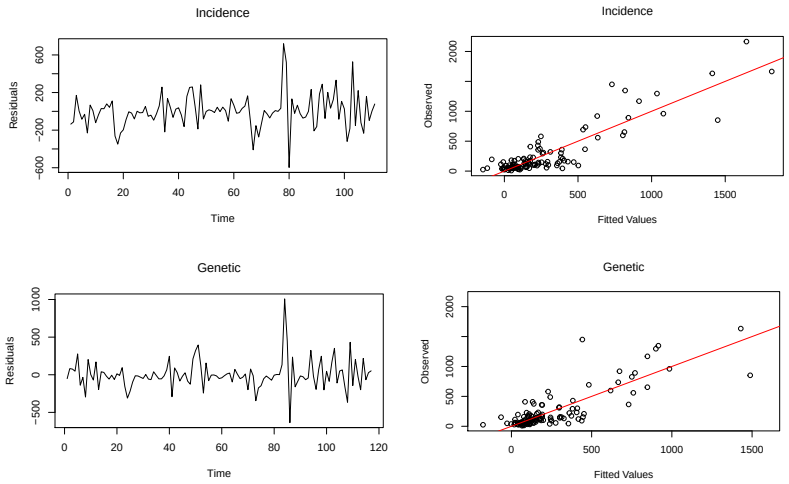


Figure 12. *Time series plot (left panel) and observed vs. fitted value (right panel) for the residuals obtained from the fitted Stepwise model. Plots related to incidence are shown in the upper panel while genetic ones are shown in the lower panel.*

6. Discussion

In this paper we considered the problem of modelling and forecasting the incidence of influenza virus in Brazil at a given month t. Here, FluNet positive flu counts were used as a proxy for incidence. The objective is to use temporal information (flu historical time series data) to model the number of cases in Brazil based on recent data on the number of cases and the genetic diversity observed in other regions. Incidentally, the study also sheds light on the migratory dynamics of the influenza virus from North America and Europe to Brazil.

In Section 5.1 (Granger causality analysis) we found evidence that past values of influenza incidence in the European and South American Regions help to predict the present value of influenza incidence in Brazil. We also discovered evidence of an indirect effect of the Western Pacific Region and Central America in Brazil. These results are intriguing when considering updating vaccines in Brazil with data related to strains from Europe from previous seasons.

As for the time series regression approach (Section 5.2), it was found that only two variables are present in all considered models, namely: B_{t-1} (number of positive flu cases in Brazil with one lag), and E_{t-2} (number of positive cases in Europe with two

lags), while A_{t-4} (number of positive cases in North America with four lags) was present in five out of six models. It is interesting to note that most predictors from northern hemisphere regions appear with lags of 3–5, possibly capturing seasonal properties of the dynamic. Additionally, while Asian genetic diversity measures appear as relevant predictors, the global genetic diversities do not.

The proposed models were also evaluated regarding their forecast capabilities. Considering h-steps ahead out-of-sample forecast, in both analysis of Sections 5.2.1 and 5.2.4, the model that overall best predicted the incidence of influenza in Brazil (in terms of MSE) in the short run was the Stepwise and in the middle to long run, the LASSO with 5 variables. The LASSO CV model performed poorly in all cases. This might be a consequence of overfitting since the LASSO CV is the one with most variables included among the considered models.

The Covid19 pandemic has largely impacted human global circulation and, consequently, the global dynamics of influenza transmission. Some lineages have remained present in local circulation and others have all but disappeared (such as B/Yamagata). Overall, the FluNet numbers of positive cases have drastically decreased. It is expected that once circulation returns to prepandemic levels influenza cases will rise again, however it is still unclear to what degree the previous transmission patterns will be reestablished or if we will see new dynamics. It has even been argued that we might see more severe influenza epidemics due to changes in immunity related to low circulation periods (Dhanasekaran et al., 2021).

Ultimately, it is likely that influenza incidence will once more be largely determined by a global dynamic, and thus modelling the Brazilian cases based on the number of cases in other regions will remain relevant. Furthermore, this same approach might prove valuable to other countries, particularly those in the global south, similarly placed in the global dynamics.

Overall, our results for short and long run forecasts ($h = 1$ and $h = 11$ steps ahead) were fairly good. Together with the relationships outlined by the Granger-causality analysis they help shed light on the global determinants of influenza incidence in Brazil. Time will tell if the particular predictors selected here will remain relevant, and in this sense, this work can be seen as historical record to be compared with the postpandemic dynamics. Nevertheless, the overall approach highlights a modelling concept which can potentially be useful in the development of public health policies regarding epidemic management and immunizations.

Acknowledgments

Aline F. Grande and Guilherme Pumi gratefully acknowledge the support of CNPq and FAPERGS. Gabriela B. Cybis gratefully acknowledges the support of the Serrapilheira Institute (grant number Serra-G1709-18939). The authors are also grateful to Rafaela Gomes de Jesus for helping with the genetic diversity data assembly.

References

Almeida, A., Codeço, C., and Luz, P. M. (2018). Seasonal dynamics of influenza in Brazil: the latitude effect. *BMC infectious diseases*, 18(1):1–9.

Barr, I. G., McCauley, J., Cox, N., Daniels, R., Engelhardt, O. G., Fukuda, K., Grohmann, G., Hay, A., Kelso, A., Klimov, A., Odagiri, T., Smith, D., Russell, C., Tashiro, M., Webby, R., Wood, J., Ye, Z., and Zhang, W. (2010). Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: Basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009-2010 Northern Hemisphere season. *Vaccine*, 28(5):1156–1167.

Bedford, T., Cobey, S., Beerli, P., and Pascual, M. (2010). Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS pathogens*, 6(5):e1000918.

Born, P. S. (2013). *Análises filogenéticas e filogeográficas dos vírus influenza A(H3N2) papel do Brasil no cenário de dispersão global e ajuste temporal entre as cepas vacinais e os vírus circulantes no período de 1999 a 2012*. PhD thesis, Instituto Oswaldo Cruz.

Born, P. S., Siqueira, M. M., Faria, N. R., Resende, P. C., Motta, F. C., and Bello, G. (2016). Phylodynamics of influenza A(H3N2) in South America, 1999-2012. *Infection, Genetics and Evolution*, 43:312–320.

Chen, C. W. S., Hsieh, Y.-H., Su, H.-C., and Wu, J. J. (2018). Causality test of ambient fine particles and human influenza in Taiwan: Age group-specific disparity and geographic heterogeneity. *Environment International*, 111:354–361.

Dhanasekaran, V., Sullivan, S., Edwards, K., Xie, R., Khvorov, A., Valkenburg, S., Cowling, B., and Barr, I. (2021). Human seasonal influenza under covid-19 and the potential consequences of influenza lineage elimination. *Research Square (preprint)*.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Diniz, M. B., Junior, J. N. O., Neto, N. T., and Diniz, M. J. T. (2009). Causas do desmatamento da Amazônia: uma aplicação do teste de causalidade de Granger acerca das principais fontes de desmatamento nos municípios da Amazônia Legal brasileira. *Nova Economia*, 19(1):121–151.

Eccles, R. (2005). Understanding the symptoms of the common cold and influenza. *The Lancet Infectious Diseases*, 5(11):718–725.

Farias, H. P. and Sáfadi, T. (2010). Causalidade entre as principais bolsas de valores do mundo. *RAM. Revista de Administração Mackenzie*, 11(2):96–122.

Garten, R. J., Davis, C. T., Russell, C. A., and et. al. (2009). Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. *Science*, 325(5937):197–201.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438.

Greene, W. H. (2012). *Econometric Analysis*. Pearson Education Limited, 7 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Ibiapina, C. C., Costa, G. A., and Faria, A. C. (2005). Influenza A aviária (H5N1) - a gripe do frango. *Jornal Brasileiro de Pneumologia*, 31(5):436–444.

Jesus, R. G. (2018). Caracterização e visualização da diversidade genética do vírus influenza ao longo do tempo. Monografia (Bacharel em Estatística), UFRGS (Universidade Federal do Rio Grande do Sul), Porto Alegre.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., et al. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS pathogens*, 10(2):e1003932.

Ljung, G. M. (1986). Diagnostic testing of univariate time series models. *Biometrika*, 73(3):725–730.

NCBI (2020). National Center for Biotechnology Information. Last accessed 12 June 2020.

Petrova, V. N. and Russell, C. A. (2018). The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16(1):47–60.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. and Holmes, E. (2009). The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Currents*, 1:RRN1003.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619.

Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., and Smith, D. J. (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Silva, P. C. R. (2015). Dinâmica molecular dos vírus Influenza A (H1N1) pandêmico em cinco anos de circulação no Brasil. Master's thesis, Instituto Oswaldo Cruz.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Toda, H. Y. and Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66(1-2):225–250.

WHO (2020). FluNet. Last accessed 12 June 2020.

Yamada, K. D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data−reexamination of the usefulness of chained guide trees. *Bioinformatics*, 32(21):3246–3251.