Medeiros, Marcelo C.; Street, Alexandre; Valladão, Davi; Vasconcelos, Gabriel; Zilberman, Eduardo

Working Paper

Short-term Covid-19 forecast for latecomers

Texto para discussão, No. 670

Provided in Cooperation with:
Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro

Suggested Citation: Medeiros, Marcelo C.; Street, Alexandre; Valladão, Davi; Vasconcelos, Gabriel; Zilberman, Eduardo (2021) : Short-term Covid-19 forecast for latecomers, Texto para discussão, No. 670, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Departamento de Economia, Rio de Janeiro

This Version is available at:
https://hdl.handle.net/10419/249718

TEXTO PARA DISCUSSÃO

No. 670

Short-Term Covid-19 Forecast for Latecomers

Marcelo Medeiros
Alexandre Street
Davi Valladão
Gabriel Vasconcelos
Eduardo Zilberman

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

Short-Term Covid-19 Forecast for Latecomers

Marcelo C. Medeiros[1*] Alexandre Street[2] Davi Valladão[3]
Gabriel Vasconcelos[4] Eduardo Zilberman[1]

[1]Department of Economics, Pontifical Catholic University of Rio de Janeiro
[2]Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro
[3]Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro
[4]Department of Economics, University of California – Irvine

February 14, 2021

Abstract

The number of Covid-19 cases is increasing dramatically worldwide, with several countries experiencing a second and worse wave. Therefore, the availability of reliable forecasts for the number of cases and deaths in the coming days is of fundamental importance. We propose a simple statistical method for short-term real-time forecasting of the number of Covid-19 cases and fatalities in countries that are latecomers – i.e., countries where cases of the disease started to appear some time after others. In particular, we propose a penalized (LASSO) regression with an error correction mechanism to construct a model of a latecomer in terms of the other countries that were at a similar stage of the pandemic some days before. By tracking the number of cases in those countries, we forecast through an adaptive rolling-window scheme the number of cases and deaths in the latecomer. We apply this methodology to four different countries: Brazil, Chile, Mexico, and Portugal. We show that the methodology performs very well. These forecasts aim to foster a better short-run management of the health system capacity and can be applied not only to countries but to different regions within a country, as well.

Keywords: Covid-19, LASSO, Forecasting, Pandemics

Acknowledgements: The authors wish to thank the Associate Editor and two anonymous referees for very helpful comments, and CAPES, CNPq and FAPERJ for partial financial support. Finally, the authors are extremely grateful to all the `Covid19Analystics` team for the great work and the many enlightening discussions.

*Corresponding author. Associate Professor, Department of Economics, Pontifical Catholic University of Rio de Janeiro. Address: Rua Marquês de São Vicente 225, Rio de Janeiro, RJ, Brazil, 22451-900. email: mcm@econ.puc-rio.br

1 Introduction

Being able to forecast accurately the number of Covid-19 cases and deaths in the very short-run, say the next few days or weeks, is crucial to manage properly the health system. Depending on the next days pressure on the health system capacity, one can make more informed decisions on how to allocate hospital beds and ventilators, on whether to set more field hospitals, on whether to train more health workers, and so on and so forth.

In this paper, we propose a statistical method to forecast in real-time the very short-run evolution of the number of Covid-19 cases and deaths in countries that are latecomers. Given that these latecomers were hit by the Covid-19 pandemic only after other countries, we can use information from these other countries when they were at a similar stage of the pandemic, a few days or weeks before. In particular, we propose a penalized Least Absolute Selection and Shrinkage Operator (LASSO) regression proposed by Tibshirani (1996), to construct an error correction model (ECM) of a latecomer in terms of the other countries. The idea behind the ECM model is to adjust the short-run dynamics of the latecomer to departures from the equilibrium between the latecomer country and its peers. By tracking the number of cases in those countries, we can forecast the short-run number of cases in the latecomer. The forecasts for the number of deaths are constructed as a linear regression on the number of cases. As the pandemic evolves, one can run the model on a daily basis, and in a adaptive rolling window scheme, to obtain updated forecasts for the next days. The model is easily estimated and confidence intervals can be computed in a straightforward manner by simulation techniques.

The rolling window (adaptive) scheme is important to acknowledge the dynamic nature of the pandemic and to attenuate the effects of outliers and potential structural breaks (due to, for example, more or less testing after a given period, policy changes, start of vaccination campaigns in some countries, or changes in the relations between countries used as explanatory variables and the latecomer). Nonetheless, it is important to emphasize that despite this attenuation, one might expect a worsen of the forecasts a few days after a structural break as the model adapts. Hence, and needless to say, the use of the proposed forecasting method should be complemented with evaluations on how the pandemic is evolving.[1]

We apply the methodology to four different countries: Brazil, Chile, Mexico, and Portugal. We show that it has been performing very well in forecasting the out-of-sample number of cases and deaths up to the next 14 days during the full year of 2020. The number of cases used correspond to those that are detected by the health system, which is the proper measure to track if the concern is to evaluate the impact on its capacity.

Tracking the evolution of the Covid-19 has been posing several challenges. The proposed method overcomes some of them. First, standard epidemiological models used to track the evolution of an epidemic are hard to discipline quantitatively to a new disease. Despite the enormous effort worldwide to understand transmission, recovery and death rates, many pa-

[1] In the case of Brazil and other less developed countries, for example, the proposed method might not anticipate the acceleration in cases and fatalities after the Covid-19 reaches areas with high urban density that lack proper sanitation. But as the model adapts, we expect to get more reliable forecasts under this new stage of the pandemic's evolution.

rameters one needs to calibrate remain uncertain (Atkeson, 2020), and behavioral responses of individuals as well as containment policies should affect these parameters (Eichenbaum, Rebelo, and Trabandt, 2020).[2] The proposed forecasting method, instead, has the advantage of being model-free, and makes projections based solely on available data.

Second, even if possible to discipline those epidemiological models reliably, they speak to the evolution of the infected population. From the perspective of managing health resources, the relevant figure is the number infected individuals that end up pressuring the health system. Note that a lot of individuals who end up being infected are asymptomatic or do not need to access the health system. Hence, the evolution of the virus among the sub-population that actually needs health care, which possesses certain characteristics that differ from the rest of the population, might be different from the evolution in the whole population. The proposed method avoid this problem by forecasting directly the number of infected individuals who are detected by the health system or specific parts of it that are of interest to be monitored, e.g., specific regions within a country.

Finally, alternative methods to track the evolution of the Covid-19, and forecast the pressure on health resources, such as massive testing, are expensive and unavailable to many countries. The methodology and our codes are immediately and cheaply reproducible to any latecomer that tracks the number of Covid-19 cases (and deaths). Note also that the proposed methodology can be as well applied to different regions within a country. This is particularly useful in large countries as Brazil, where the disease have hit distinct regions with delays.[3]

The aforementioned challenges are even harder to overcome in poor and developing countries, mostly latecomers, due to the lack of high-quality research, reliable data and limited budget. By tracking the very short-run evolution of the number of Covid-19 cases (and deaths) in real-time, we hope that this methodology can be useful to inform policymakers and the general public. In the authors' point of view, an adaptive and accurate data-driven forecast is critical to foster better management of the health system, especially in those countries that lack proper resources.

1.1 Main Takeaways

The ECM method proposed in this paper provides forecasts for cases and deaths with lower mean absolute percentage errors (MAPE) than a benchmark model. Namely, a simple quadratic trend regression that has been shown to be quite precise for short-term forecasts of Covid fatalities (Coroneo, Iacone, Paccagnini, and Monteiro, 2020).

According to the test of superior predictive ability put forward by Giacomini and White (2006), the ECM model statistically outperforms the benchmark for most horizons considered.

[2]Epidemiologists and researchers from other fields rushed to improve those models and provide simulations on the spread of the disease, some of them taking into account counteracting policy and/or behavioral responses. A very incomplete list includes Berger, Herkenhoff, and Mongey (2020), Kucharski, Russell, Diamond, Liu, Edmunds, Funk, Eggo, et al. (2020), Walker, Whittaker, Watson, et al. (2020), Wu, Leung, and Leung (2020), and Bastos and Cajueiro (2020).

[3]We post on a daily basis updated forecasts for Brazil, methodology updates, and codes. The domain is https://covid19analytics.com.br/, and one can check there for updates.

Furthermore, the superiority of the ECM model is even more pronounced when we look at fatalities. Our findings are very similar over the four countries considered.

More specifically, for Chile, the ECM delivers MAPEs for Covid-19 cases between 0.525% (1-day-ahead) and 6.166% (14-days-ahead) while the benchmark's MAPEs range between 0.755% and 6.188%. For Brazil, the ECM's MAPEs lie between 0.685% and 5.532%, whereas the ones from the benchmark vary between 1.170% and 7.366%. For Mexico, our approach delivers MAPEs ranging from 0.337% and 3.372% while the benchmark yields MAPEs between 0.582% and 3.541%. Finally, for the case of Portugal, the MAPEs are 0.336% versus 0.973% for one-steap ahead and 6.107% versus 8.236% for 14-steps ahead.

Now, turning to deaths, the differences are even more evident. In the case of Chile, our proposed model delivers MAPEs as low as 1.192% for one-day-ahead forecasts as apposed to 2.514% from the benchmark alternative. For 14-steps ahead, the MAPE from the ECM approach is 9.704% and the one from the benchmark is 18.472%. For Brazil, the MAPEs for one-step-ahead are 0.739% (ECM) and 1.302% (benchmark). For 14-days-ahead the figures are 5.854% (ECM) and 7.629% (benchmark). For Mexico, the ECM MAPEs range from 0.957% (one-day-ahead) to 4.385% (14-days-ahead), whereas the benchmark model provides MAPEs between 1.062% (one-day-ahead) and 5.541% (14-days-ahead). Finally, for Portugal, the MAPEs are 0.376% (ECM) versus 1% (benchmark) for one-day-ahead and 3.921% (ECM) and 8.219% (benchmark) for 14-days-ahead.

It is important to highlight that as the pandemic evolves, the performance of the ECM method improves and the MAPEs decrease substantially. Finally, due to the rolling window scheme adopted in this paper, the model is able to rapidly adapt to new scenarios without the need of any change in its structure.

1.2 Comparison to the Literature

Since the outbreak of the Covid-19 pandemic, a large number of papers dealing with short-term forecasts of cases and deaths counts has been published in a wide collection of academic journals. The models range from different versions of epidemiological compartmental models to pure statistical and machine learning approaches. The models can be as simple as a pure trend regression or as complicated as deep learning neural networks. Nevertheless, a number of studies provide strong evidence that is quite difficult to beat the simplest alternatives. Our approach keeps the simplicity of several statistical models, explores equilibrium relations between a latecomer country and its peers, and shows robustness against many breaks in the dynamics of the series over the year of 2020.

Coroneo, Iacone, Paccagnini, and Monteiro (2020) compare the predictive accuracy of forecasts for the number of fatalities produced by several forecasting teams and collected by the United States Center for Disease Control and Prevention (CDC) and a simple benchmark alternative. The set of models include both statistical (dynamic growth model) and compartmental approaches (different versions of SEIRD models). The authors find that a simple quadratic trend regression outperforms all the alternatives for horizons up to one week ahead. For longer

horizons, some of the models sometimes outperform the simple benchmark. However, the authors show that the ensemble of models outperform all the other alternatives. Similar quadratic trend models are also considered with some adjustments by Jiang, Zhao, and Sha (2020) and Li and Linton (2021). Due to the previous satisfactory performance of the quadratic trend model, we also adopt it as a benchmark specification in this paper.

Hendry (2020) consider a flexible trend model and apply it to forecast confirmed cases and deaths in a large number of countries. As ours, their model has no epidemiological component. For confirmed cases, the authors report MAPEs between 0.4% and 2.1% for one-day-ahead and from 1.7% to 7.6% for four-days-ahead. In the case of death counts, the MAPEs are higher. Although not directly compared to our MAPEs, as we are not analysing the same countries, the MAPEs reported here are lower than the ones in the above mentioned paper. See also Petropoulos, Makridakis, and Stylianou (2020) for a similar approach to Hendry (2020).

Nonlinear machine learning methods such as, Long Short-Term Memory and Deep Neural Networks, Random Forests and Support Vector Machines, have also been considered to forecast cases and death counts in the short-run. For example, Ribeiro, da Silva, Mariani, and Coelho (2020) estimate a vast amount of statistical and machine learning methods to forecast future cases in Brazil. Not only their models are much more complex than the ones considered here, but their MAPEs range between 0.87%–3.51%, 1.02%–5.63%, and 0.95%–6.90% for one, three, and six-days-ahead forecasts, respectively. These numbers are systematically larger than the MAPEs from our ECM specification. Other examples of application of nonlinear machine learning models are Zeroual, Harrou, Dairi, and Sun (2020), Chimmula and Zhang (2020), and da Silva, Ribeiro, Mariani, and Coelho (2020), among many others.[4]

1.3 Organization of the Paper

In addition to this Introduction, this paper is organized as follows. Section 2 presents the methodology. Section 3 gives some guidance to practitioners. Section 4 describes the results. Finally, Section 5 concludes.

2 Methodology

Let $\tau = 1, 2, \ldots, \mathcal{T}$, represents the number of days after the 100th confirmed case of Covid-19 in a given country/region. Define y_τ as the natural logarithm of the number of confirmed cases τ days after the 100th case of the disease in this specific country/region. In addition, let $\boldsymbol{x}_\tau$ be a vector containing the natural logarithm of the number of reported cases for p other countries also τ days after the 100th case has been reported and a quadratic trend, i.e, $\boldsymbol{x}_\tau$ also includes τ and τ^2. The idea is that, in the regular time scale, $\boldsymbol{x}_\tau$ may be ahead of time of y_τ. For example, in France and Spain, the 100th was reported on February 29 and March 2, respectively. On the other hand, in Brazil, a latecomer, the 100th case was confirmed only on March 14. Therefore,

[4]None of these papers provide convincing evidence of the superiority of complicated machine learning models to simpler alternatives.

the idea is to use, for instance, data from France on February 29 and Spain on March 2 to explain the number of cases in Brazil on March 14. Note that we do not claim any causal link between the p countries and the latecomers. Our proposal explores the fact the evolution of the disease in different countries share similar patterns.[5]

The statistical approach considered in this paper is a simple error correction model (ECM) which maps $\boldsymbol{x}_\tau$ into y_τ as:

$$\Delta y_\tau = \Delta \boldsymbol{x'}_\tau \boldsymbol{\pi} + \gamma \left(y_{\tau-1} - \boldsymbol{x}'_{\tau-1} \boldsymbol{\beta} \right) + u_\tau, \tag{2.1}$$

where u_τ is zero-mean random noise with variance σ^2, $\Delta y_\tau = y_\tau - y_{\tau-1}$, $\Delta \boldsymbol{x}_\tau = \boldsymbol{x}_\tau - \boldsymbol{x}_{\tau-1}$, and $\boldsymbol{\pi}$, γ, and $\boldsymbol{\beta}$ are unknown parameters to be estimated. As can be seen in Figure 1, the number of cases in τ-scale in different countries display a strong common exponential trend. The logarithm transformation is important to turn the exponential trend into a linear one.

The model is estimated in two steps. In the first step, we estimate $\boldsymbol{\beta}$ in a long-run equilibrium model:

$$y_\tau = \boldsymbol{x}'_\tau \boldsymbol{\beta} + \varepsilon_\tau, \tag{2.2}$$

where ε_τ is a zero-mean second-order stationary error term.

Due to the limited amount of data and the large dimension of $\boldsymbol{x}_t$ as compared to the sample size, we use the least absolute and shrinkage operator (LASSO) to recover the parameter vector. The goal of the LASSO is to balance the trade-off between bias and variance and is a useful tool to select the relevant peers in an environment with very few data points. Therefore, the estimator of the unknown parameter $\boldsymbol{\beta}_\mathcal{T}$ in equation (2.2) is defined as:

$$\widehat{\boldsymbol{\beta}}_\mathcal{T} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{K} \sum_{\tau=\mathcal{T}-K+1}^{\mathcal{T}} \left(y_\tau - \boldsymbol{x}'_\tau \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2.3}$$

where K is the number of days in the estimation window, and $\lambda > 0$ is the penalty parameter. Theoretical justification for the use of LASSO to estimate the parameters in this setup with trends can be found in Masini and Medeiros (2019).

Once we estimated equation (2.2), we proceed to a second step estimating the ECM by Ordinary Least Squares (OLS) with the variables selected in the first step with the LASSO. The final prediction h–step ahead from $\mathcal{T}$ reads as:

$$\widehat{y}_{\mathcal{T}+h} = \Delta \boldsymbol{x}'_{\mathcal{T}+h} \widehat{\boldsymbol{\pi}}_\mathcal{T} - \widehat{\gamma}_\mathcal{T} \boldsymbol{x}'_{\mathcal{T}+h-1} \widehat{\boldsymbol{\beta}}_\mathcal{T} + (1 + \widehat{\gamma}_\mathcal{T}) \, \widehat{y}_{\mathcal{T}+h-1}, \tag{2.4}$$

where $\widehat{y}_{\mathcal{T}+h-1}$ is the forecast for the previous day. Confidence intervals were obtained through simulation by assuming that the error term u_τ in (2.1) is normally distributed.

The intuition behind the proposed ECM is to model the dynamics and the reactions to departs from the equilibrium between y_τ and $\boldsymbol{x}_\tau$: the disease behaves somehow in a similar

[5]These similarities are also explored in compartmental models. See also, Carroll, Bhattacharjee, Chen1, Dubey, Fan, Gajardo, Zhou, Müller, and Wang (2020).

fashion in different countries. Note that we do not claim a causal link, for instance, from the cases in Germany to the cases in Brazil due to mobility among these two countries. What the model explores is that the evolution of diseases like Covid-19 seems to share similar patterns in different locations. Furthermore, as the first-stage LASSO regression is a model selection tool, if our hypothesis of common dynamics among countries is not valid, the LASSO will not select any country to explain the latecomer behavior and/or the residuals of the first stage LASSO regression present statistical evidence of non-stationarity.

Our interest relies on the forecasts for number of cases in levels not in logs: $Y_\tau := \exp(y_\tau)$. Therefore, for the horizon $\mathcal{T}+h$, the forecasts are constructed as:

$$\widehat{Y}_{\mathcal{T}+h} = \widehat{\alpha}_{\mathcal{T}}\, e^{\widehat{y}_{\mathcal{T}+h}}, \tag{2.5}$$

where $\widehat{\alpha}_{\mathcal{T}} = \frac{1}{K}\sum_{\tau=\mathcal{T}-K+1}^{\mathcal{T}} \exp(\widehat{u}_\tau) := \frac{1}{K}\sum_{\tau=\mathcal{T}-K+1}^{\mathcal{T}} \exp\left(y_\tau - \widehat{y}_\tau\right)$ is a correction which is essential to attenuate the induced bias when we take the exponential of the forecasted value of $y_{\mathcal{T}+h}$; see Wooldridge (2019). Note that, in the τ-scale, the peers are "in the future" and we can plug-in actual values of $\boldsymbol{x}_{\mathcal{T}+h}$ to construct our forecasts. Note also that a rolling estimation window of $K = 28$ days induce an adaptive forecasting framework suitable to capture the dynamic nature of the pandemic and to attenuate the effects of outliers and potential structural breaks. Finally, in order to give more weight to the newest observations, we inflated the data by repeating the last four observations where the last observation is repeated four times with a linear decay for the observations before.

It is worth emphasizing that this model is only a local approximation of a more complex and dynamic process. Therefore, its best use relies on fresh and updated data, and the rolling window scheme takes care of that. Although the model has been providing to have excellent adherence, the proposed forecasting method may be complemented with indexes, such as proxies for social distancing, to guide evaluation of the future dynamics affecting the number of cases and deaths. However, the inclusion of other regressors such as Google mobility has not showed to improve the quality of the forecasts.

3 Guide to Practice

The implementation of the proposed forecasting method requires the choice of three tuning parameters: the penalty term in the first-stage LASSO regression (λ), the length of the estimation window (K) and the data inflation mechanism.

The penalty term is selected by the Bayesian Information Criterion (BIC) as discussed in Medeiros and Mendes (2016). The degrees of freedom of the LASSO are determined by the nonzero estimated coefficients. Cross-validation can be also used to determine the penalty parameters. However, we prefer the BIC in order to avoid any extra computational burden.

The estimation window length (K) and the inflation scheme for the most recent observations can be estimated in a rolling window process. Before computing the actual forecasts, one could select these tuning parameters from a rolling window using previous data and selecting the

values that minimize the out-of-sample error measure (MAPE for example). However, this procedure gives us the best model for past data, especially because we need a significant number of windows that go back several weeks to obtain stable estimates. Although a procedure like this could lead to some local improvements, it could also lead to situations where the forecast explodes, especially when a very small K is selected with no data inflation. To avoid unnecessary data mining that could lead to unreliable results, we choose to use a fixed value of $K = 28$ and the inflation scheme for the four most recent lags. We understand that this sample size is enough to get a satisfactory model given the number of variables and it is not too big to include many structural breaks. The inflation scheme is just to give more weight to the most recent data, which is likely to be more similar to future data in the short-run. Small changes in the inflation strategy do not affect the overall results. However, no inflation yields higher errors.

Another important point is to check if the first-stage errors are stationary. This can be conducted by common unit-root tests. If the null hypothesis of unit-root is not rejected, the first stage is clearly misspecified and the forecasts will not be reliable. In this paper, we run unit-root tests after every LASSO estimation and there is no evidence of misspecification.

Finally, the ECM methodology proposed here is flexible enough to include other regressors, such as, for example, mobility data. However, in our experience the inclusion of such data did not bring any evident improvement in the performance of the model.

4 Results

4.1 Data and Results

We used the John Hopkins compiled data[6] for all countries with Covid-19 cases and the Brazilian Ministry of Health official data.[7]. The data are organized in epidemiological time, i.e., the time dimension represents the number of days after the 100th case.

The models were computed on a rolling window scheme with 28 in-sample observations per window. For each country, the model estimation started when the number of confirmed cases of Covid-19 reached 20,000. The last in-sample day for every country was December 17, 2020, which makes December 31, 2020 the day of the last out-of-sample forecast. The start date for each country was May 2nd for Chile, April 11th for Brazil, May 1st for Mexico and April 19th for Portugal. As described in Section 3, we setting the window length to 28 days turned out to be a good trade-off between the quality of the in-sample adjustment and robustness to potential structural breaks.

Figure 1 illustrates the evolution of Covid-19 cases in several countries. The data is displayed in epidemiological time, i.e., the x-axis represents the number of days since the first registered case. It is clear from the figure that some countries are ahead of epidemiological time than others.

[6]John Hopkins data available at https://github.com/CSSEGISandData/COVID-19
[7]Brazilian Ministry of Health data available at https://covid.saude.gov.br/

4.2 Forecasting results

4.2.1 Mean Absolute Percentage Errors

In other to compare the performance of the ECM model proposed in this paper we consider the benchmark model as described in Coroneo, Iacone, Paccagnini, and Monteiro (2020). The model is a simple quadratic trend regression defined as:

$$y_\tau = \alpha_0 + \alpha_1 \tau + \alpha_2 \tau^2 + \epsilon_\tau, \tag{4.1}$$

where ϵ_τ is a zero-mean error term. As mentioned before, although this bechmark is amazingly simple, it proved to be quite precise for short-term forecasts.

Tables 1 and 2 present the forecasting results for the full out-of-sample period. The tables show the MAPE for forecasting horizons of 1 to 14 days ahead of the Covid-19 accumulated number of cases (Table 1) and deaths (Table 2). Values in parenthesis are p-values for the Giacomini & White test for superior predictive ability (Giacomini and White, 2006). The null hypothesis of the test is that both forecasts have the same MAPE.

We start by comparing the models with respect to the forecasts for case counts. For Chile, the ECM overperforms the benchmark in 11 out of 14 horizons. However, the differences are statistically significant only for one-day-ahead. For Brazil, the results are much more favorable to the ECM model as it has lower MAPEs than the benchmark for all horizons and the differences are all significant. For Mexico, the ECM is also superior to the benchmark for all horizons, but the differences are significant in seven out 14 cases. Finally, for Portugal, the benchmark performs poorly for all horizons. The differences in performance of the ECM and the benchmark are statistically different in all horizons up to 12-days-ahead.

Turning the attention now to fatalities, the ECM model is clearly superior to the benchmark and the differences are much more pronounced. For Chile, Brazil, and Portugal the ECM is better than the benchmark for all horizons and the differences are all statistically significant. For Mexico, the ECM overperforms the benchmark in 13 out of 14 horizons, and for all horizons greater than four-days-ahead the differences are statistically significant.

In order to analyze how the errors unfold over the evolution of the pandemic, we plot rolling MAPEs over 14 days in Figures 2–5 for cases forecasts and in Figures 6–9 for deaths. We report only results for selected horizons. It is clear from the figures, that both models improve over time. In some cases the reductions be larger than 50%. For Chile, the gains of the ECM over the benchmark are more evident during the months of July and August when we look to cases. On the other hand, for deaths, the ECM is better than the benchmark for all windows. In case of Brazil, the superiority of the ECM forecasts for cases are more concentrated in the beginning of the sample, when the benchmark performs very poorly. A similar pattern is visible in the case of deaths. In the case of Mexico and for forecasts for case counts, the ECM is systematically superior to the benchmark over the sample and when we consider the one-day-ahead forecasts. For the other horizons, the gains are more concentrated in the beginning and in the end of the sample. Equivalent conclusions emerge when we look at the forecasts for deaths. In the case of

Portugal, the benefits of using the ECM instead of the benchmark are clearer during the first half of the sample.

In order to complement the analysis and to reconcile the results presented in Tables 1 and 2 and in Figures 2–9, we compute the frequency of days when the ECM has a lower absolute percentage error than the benchmark and the median of the ratio of the daily absolute errors of the ECM and benchmark specification over the forecasting sample. Figures 10–13 report, for each forecasting horizon, the frequency of days over the sample when the daily absolute percentage error of the ECM is smaller than the one from the benchmark alternative. The upper panel in the figures present the results for cases whereas the lower panel shows the numbers of deaths. To quantify these gains, Figures 14–17 present, for each horizon, the median of the ratios of the daily absolute errors. A number less than one favors the ECM model. As before, the upper (lower) panels concern cases (deaths). We prefer the use the median instead of the mean to avoid potential effects of outliers.

For Chile, the one-day-ahead forecasts of ECM are the best ones in 72.17% (61.75%) of the days when cases (deaths) are considered. These numbers drop to 46.52% (53.91%) one the forecasting horizon is set to 14 days. For almost all the horizons the proportion of days when the ECM is better than the benchmark is larger than 50%. From the analysis of the results in Figure 14, it is clear that the ECM is better than the benchmark for almost all horizons when cases are considered. When fatalities are analyzed, the ECM is always superior.

For Brazil, the results are less favourable to the ECM. Both Figures 11 and 15 point to the superiority of the benchmark. However, when looking at the results in Tables 1 and 2 above, we may reach a different conclusion. Therefore, it is important to uncover the drivers to the best MAPE of the ECM over the entire out-of-sample period. The reason for this finding is that the ECM is way superior to the benchmark during the first 100 days of the sample. This was the period when the number of cases and deaths in Brazil was accelerating the most.

For Mexico, the results are very supportive to the ECM specification. The ECM is superior to the benchmark in more than 50% of the days in almost every case considered in the analysis. The median ratios of the absolute percentage errors are always bellow one, when Covid-19 cases are considered. For deaths, the ratios are bellow one for horizons larger than four-days-ahead.

Finally, for Portugal, the superiority of the ECM draws attention. For all horizons considered, the ECM is better than the benchmark in terms of number of days with lower errors as well as in terms of the relative magnitude of the absolute percentage errors.

4.2.2 Diagnostic Tests

We report two diagnostic tests. First, Figure 18 illustrates the empirical distribution of the estimated coefficient of a first-order autoregressive, AR(1), model estimated with the residuals from the first-stage LASSO regression. The distribution is over all the rolling windows and each one of the four countries analyzed in this paper. It is clear from the figure that apart from a single case, all the estimates are bellow one in absolute value. Unit-root tests also strongly reject the null of unit roots in all but one case. This specific negative case is related to a huge

outlier in the data which distort the estimation of the AR coefficient and, consequently, the unit-root test. Based on this analyze we are quite confident that out methodology is adequate for the present data.

The second diagnostic is related to the data inflation heuristic. Table 4 presents the MAPEs of the ECM with data inflation divided by the MAPEs of the ECM without data inflation. Numbers lower than one favors the inflation heuristic. For Brazil, Chile, and Portugal, it is clear that data inflation is superior to no inflation at all. For Mexico, we see improvements when the forecasts for cases are considered but not deaths. Changing the number of observations to inflate seems to have no significant effect and these extra results can be obtained upon request.

4.2.3 Variable Selection

Finally, it is important to understand which variables are being selected by the LASSO during the first stage of the methodology. Table 3 shows the frequency of selection of each variable over the rolling windows. Mexico is the latecomer country where each variable in the pool is selected at least once. Portugal seems to be the country with the most parsimonious model. Note also the frequency of selection of each variable differs from country to country.

5 Conclusion

In this paper, we propose a statistical model to forecast in the very short-run the reported number of cases and deaths by the Covid-19 in countries/regions that are latecomers. We believe this is a useful tool to inform health management. Nonetheless, structural breaks might worsen forecasts a few days after such breaks. So the use of this tool should complemented with other external information, such as proxies for social distancing, to guide subjective or objective assessments on potential dynamic changes on the pandemic's evolution. We hope to keep improving the model by improving the methodology and incorporating more information. And we aim to keep forecasts, methodology and codes updated on a daily basis at https://covid19analytics.com.br/.

References

Atkeson, A. (2020): "How Deadly Is COVID-19? Understanding The Difficulties With Estimation Of Its Fatality Rate," Working Paper 26965, National Bureau of Economic Research.

Bastos, S., and D. Cajueiro (2020): "Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil," Discussion paper, arXiv.org.

Berger, D., K. Herkenhoff, and S. Mongey (2020): "An SEIR Infectious Disease Model with Testing and Conditional Quarantine," Working Paper 597, Federal Reserve Bank of Minneapolis.

Carroll, C., S. Bhattacharjee, Y. Chen1, P. Dubey, J. Fan, A. Gajardo, X. Zhou, H. Müller, and J. Wang (2020): "Time dynamics of COVID-19," *Scientific Reports*, 10.

Chimmula, V., and L. Zhang (2020): "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Caos, Solitons & Fractals*, 135.

Coroneo, L., F. Iacone, A. Paccagnini, and P. Monteiro (2020): "Testing the predictive accuracy of Covid-19 forecasts," Working Paper 20/10, University of York.

da Silva, R., M. Ribeiro, V. Mariani, and L. Coelho (2020): "Forecasting Brazilian and American Covid-19 cases based on artificial intelligence coupled with climatic exogenous variables," *Caos, Solitons & Fractals*, 139.

Eichenbaum, M., S. Rebelo, and M. Trabandt (2020): "The Macroeconomics of Epidemics," Working Paper 26882, National Bureau of Economic Research.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

Hendry, J. D. J. C. D. (2020): "Short-term forecasting of the coronavirus pandemic," *International Journal of Forecasting*, forthcoming.

Jiang, F., Z. Zhao, and X. Sha (2020): "Time series analysis of Covid-19 infection curve: A change-point perspective," *Journal of Econometrics*, forthcoming.

Kucharski, A., T. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. Eggo, et al. (2020): "Early dynamics of transmission and control of COVID-19: a mathematical modelling study," *The Lancet Infectious Diseases*.

Li, S., and O. Linton (2021): "When will the Covid-19 pandemic peak?," *Journal of Econometrics*, 220, 130–157.

Masini, R., and M. Medeiros (2019): "Counterfactual Analysis With Artificial Controls: Inference, High Dimensions and Nonstationarity," Working Paper 3303308, SSRN.

Medeiros, M., and E. Mendes (2016): "ℓ_1-Regularization of High-dimensional Time-Series Models with Non-Gaussian and Heteroskedastic Innovations," *Journal of Econometrics*, 191, 255–271.

Petropoulos, F., S. Makridakis, and N. Stylianou (2020): "Forecasting confirmed cases and deaths with a simple time series model," *International Journal of Forecasting*, forthcoming.

Ribeiro, M., R. da Silva, V. Mariani, and L. Coelho (2020): "Short-term forecasting Covid-19 cumulative confirmed cases: Perspectives for Brazil," *Caos, Solitons & Fractals*, 135.

Tibshirani, R. (1996): "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Walker, P., C. Whittaker, O. Watson, et al. (2020): "The Global Impact of COVID-19 and Strategies for Mitigation and Suppression," Discussion paper, Imperial College London.

Wooldridge, J. (2019): *Introductory Econometrics: A Modern Approach*. Cengage Learning, 7th edn.

Wu, J., K. Leung, and G. Leung (2020): "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study," *The Lancet*, 395(10225), 689–697.

Zeroual, A., F. Harrou, A. Dairi, and Y. Sun (2020): "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study," *Caos, Solitons & Fractals*, 140.

Table 1: Cases Forecasting Mean Absolute Percentage Error

The table shows the forecasting mean absolute percentage error (MAPE) for forecasting horizons of 1 to 14 days ahead of the Covid-19 accumulated number of cases. The models were computed on a rolling window scheme with 28 in-sample observations per window. For each country, the model estimation started when the number of cases of Covid-19 reached 20,000. The last in-sample day for every country was December 17, 2020, which makes December 31, 2020 the last out-of-sample forecast. The start date for each country was May 2nd for Chile, April 11th for Brazil, May 1st for Mexico and April 19th to Portugal. Values in parenthesis are p-values for the Giacomini & White test. (Giacomini and White, 2006).

Cases: Forecasting Mean Absolute Percentage Errors

Chile

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	0.755	0.950	1.163	1.381	1.618	1.908	2.292	2.721	3.192	3.695	4.245	4.839	5.484	6.188
ECM	0.526	0.852	1.065	1.276	1.490	1.755	2.130	2.633	3.141	3.689	4.282	4.872	5.494	6.166
	(0.024)	(0.275)	(0.503)	(0.410)	(0.313)	(0.281)	(0.226)	(0.651)	(0.933)	(0.677)	(0.602)	(0.603)	(0.705)	(0.827)

Brazil

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	1.170	1.530	1.921	2.317	2.707	3.130	3.576	4.034	4.528	5.092	5.664	6.245	6.802	7.366
ECM	0.685	1.205	1.550	1.787	2.012	2.204	2.450	2.804	3.173	3.629	4.110	4.567	5.039	5.532
	(0.000)	(0.001)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Mexico

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	0.582	0.762	0.939	1.112	1.282	1.457	1.665	1.903	2.152	2.405	2.670	2.930	3.213	3.541
ECM	0.337	0.594	0.773	0.951	1.078	1.221	1.407	1.645	1.908	2.166	2.483	2.763	3.105	3.372
	(0.000)	(0.185)	(0.011)	(0.208)	(0.016)	(0.029)	(0.028)	(0.026)	(0.110)	(0.094)	(0.395)	(0.235)	(0.747)	(0.332)

Portugal

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	0.973	1.306	1.666	2.061	2.498	2.980	3.511	4.079	4.688	5.331	6.004	6.709	7.453	8.236
ECM	0.336	0.591	0.855	1.134	1.392	1.671	2.003	2.406	2.881	3.413	3.979	4.612	5.305	6.107
	(0.001)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.007)	(0.014)	(0.037)	(0.130)	(0.445)

Table 2: Deaths Forecasting Mean Absolute Percentage Error

The table shows the forecasting mean absolute percentage error (MAPE) for forecasting horizons of 1 to 14 days ahead of the Covid-19 accumulated number of deaths. The models were computed on a rolling window scheme with 28 in-sample observations per window. For each country, the model estimation started when the number of cases of Covid-19 reached 20,000. The last in-sample day for every country was December 17, 2020, which makes December 31, 2020 the last out-of-sample forecast. The start date for each country was May 2nd for Chile, April 11th for Brazil, May 1st for Mexico and April 19th to Portugal. Values in parenthesis are p-values for the Giacomini & White test. (Giacomini and White, 2006).

Deaths: Forecasting Mean Absolute Percentage Errors

Chile

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	2.514	3.349	4.259	5.268	6.325	7.427	8.565	9.703	10.906	12.212	13.601	15.077	16.672	18.472
ECM	1.192	1.687	2.206	2.794	3.389	4.032	4.686	5.393	6.060	6.668	7.411	8.190	8.945	9.704
	(0.000)	(0.000)	(0.001)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)

Brazil

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	1.302	1.695	2.100	2.540	2.987	3.455	3.951	4.465	4.999	5.555	6.094	6.609	7.117	7.629
ECM	0.739	1.111	1.413	1.664	1.924	2.194	2.508	2.907	3.368	3.841	4.317	4.761	5.266	5.854
	(0.004)	(0.008)	(0.011)	(0.011)	(0.009)	(0.006)	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)

Mexico

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	1.062	1.317	1.606	1.915	2.239	2.590	2.961	3.356	3.735	4.113	4.486	4.830	5.166	5.541
ECM	0.957	1.353	1.551	1.659	1.690	1.788	2.121	2.588	3.026	3.302	3.542	3.710	4.006	4.385
	(0.396)	(0.743)	(0.660)	(0.033)	(0.002)	(0.001)	(0.004)	(0.017)	(0.043)	(0.034)	(0.023)	(0.013)	(0.020)	(0.021)

Portugal

Days ahead	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Benchmark	1.000	1.353	1.747	2.179	2.646	3.147	3.681	4.250	4.847	5.471	6.124	6.801	7.500	8.219
ECM	0.376	0.517	0.693	0.874	1.084	1.290	1.543	1.799	2.068	2.372	2.741	3.123	3.507	3.921
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

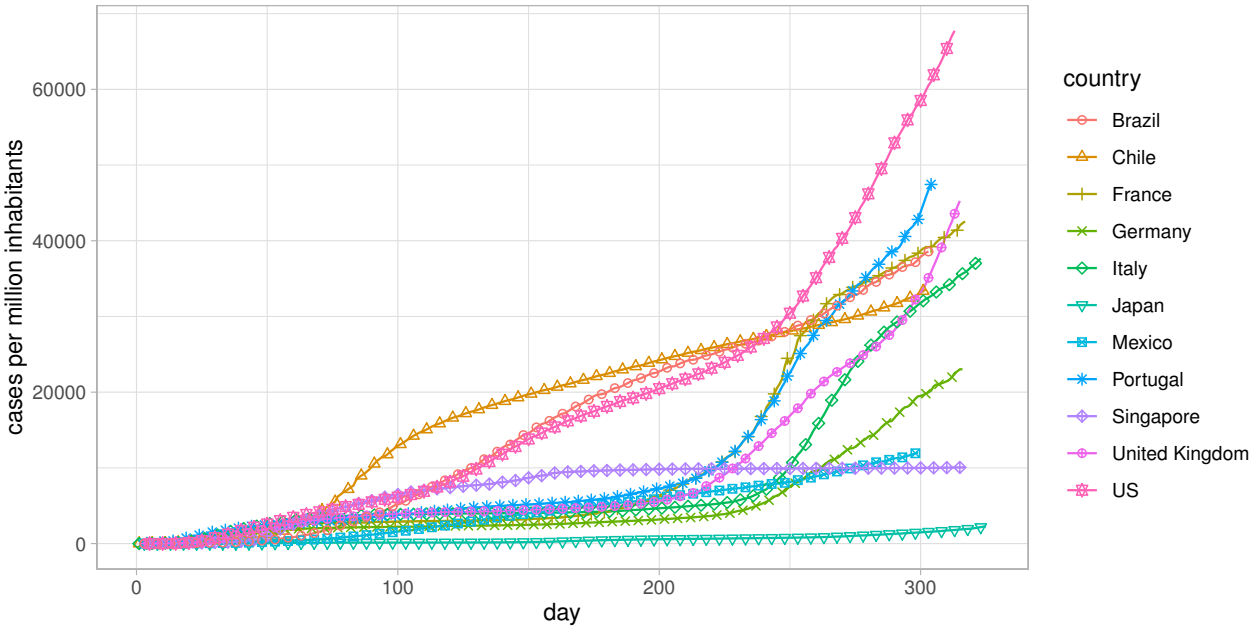

Figure 1: Evolution of cases in different countries in epidemic time.
The figure illustrates the evolution of the cases of Covid-19 in different countries according to the epidemic calendar, i.e., the x-axis represents days from the first confirmed case of Covid-19. It is clear that come countries are in front of others in epidemic time.

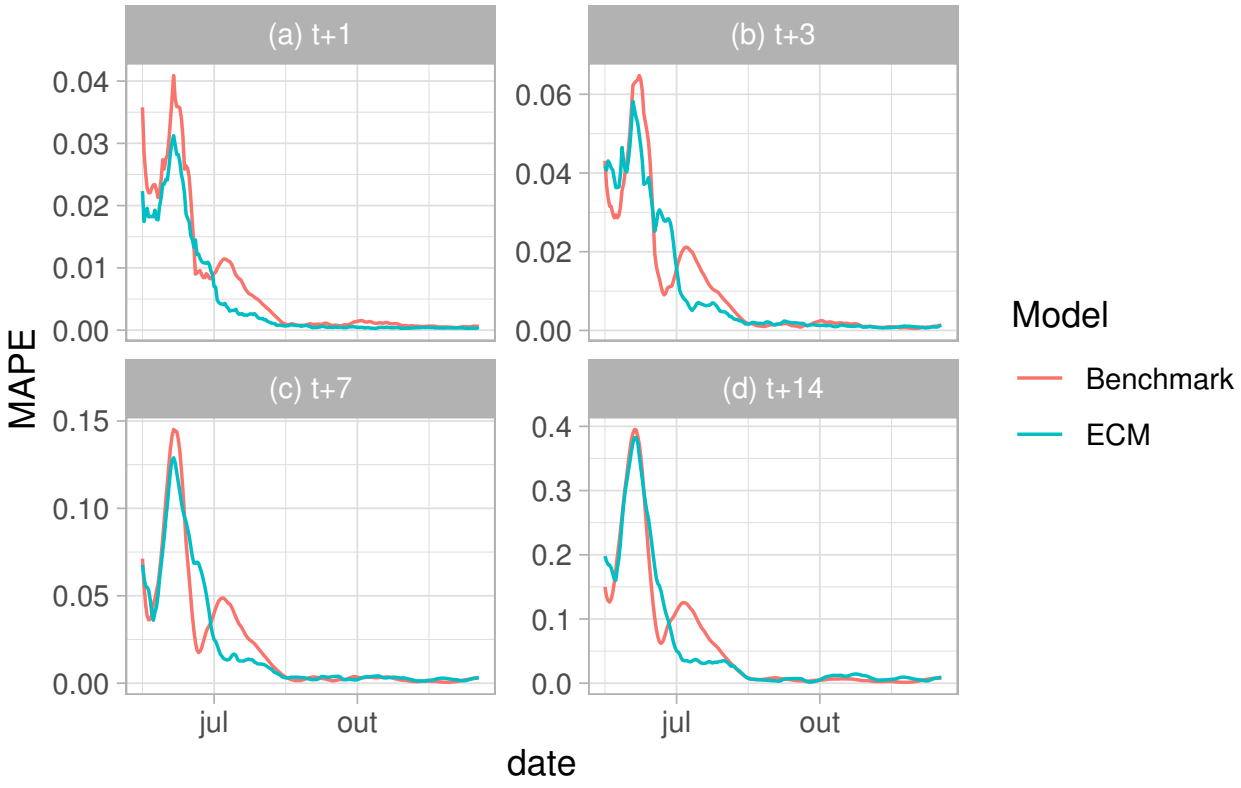

Figure 2: Cases Rolling Mean absolute percentage error - Chile.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

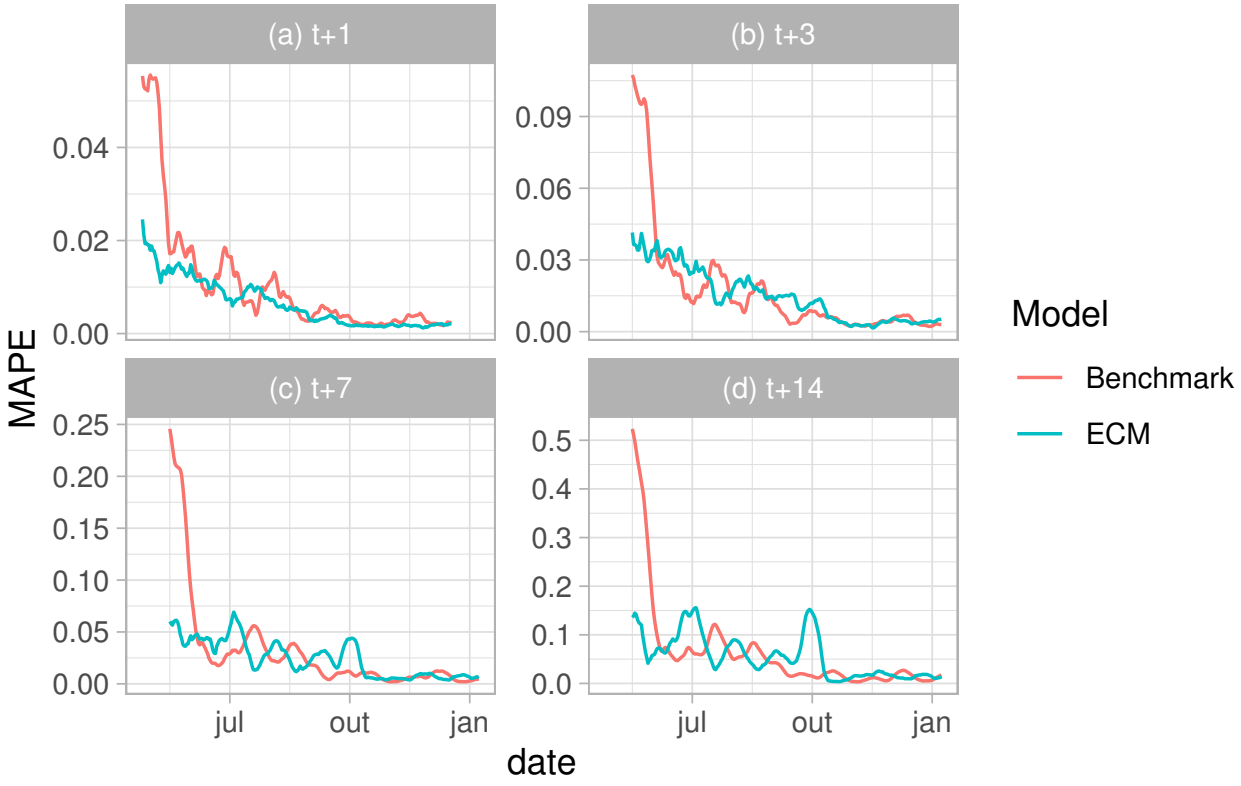

Figure 3: Cases Rolling Mean absolute percentage error - Brazil.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

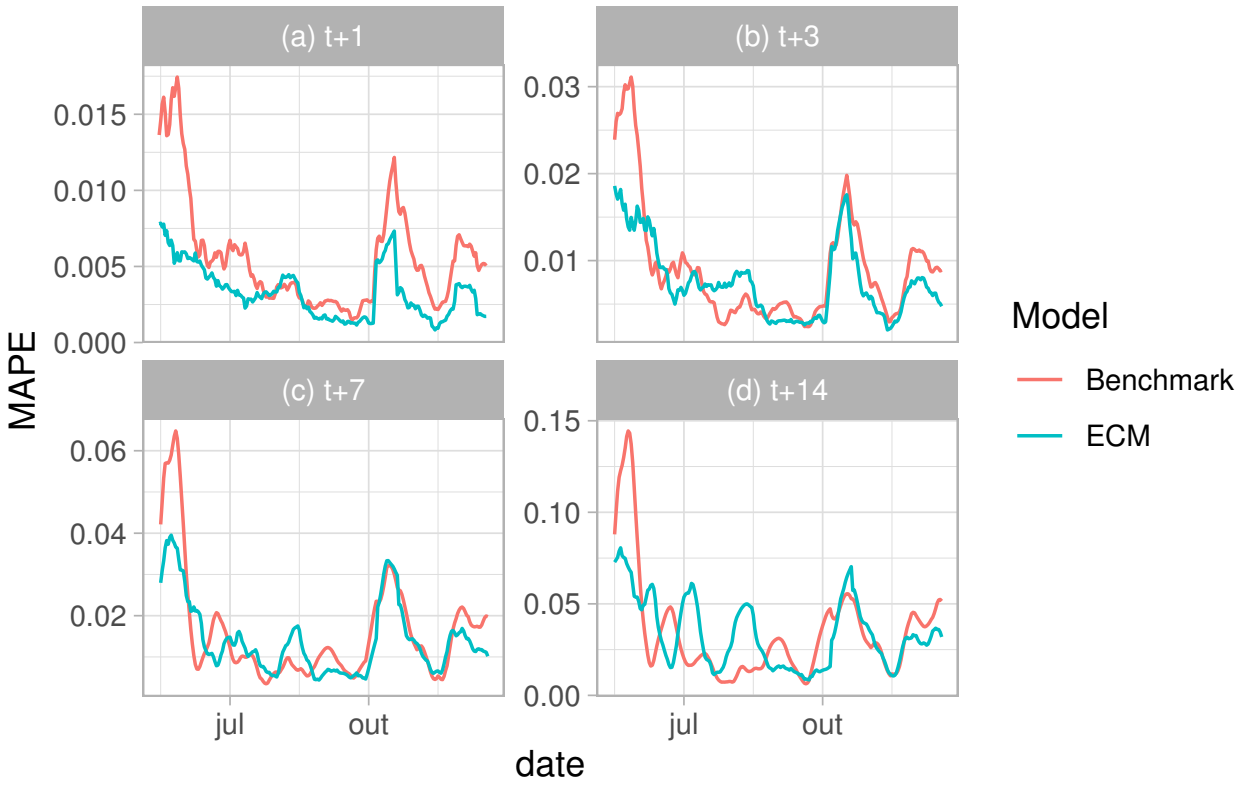

Figure 4: Cases Rolling Mean absolute percentage error - Mexico.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

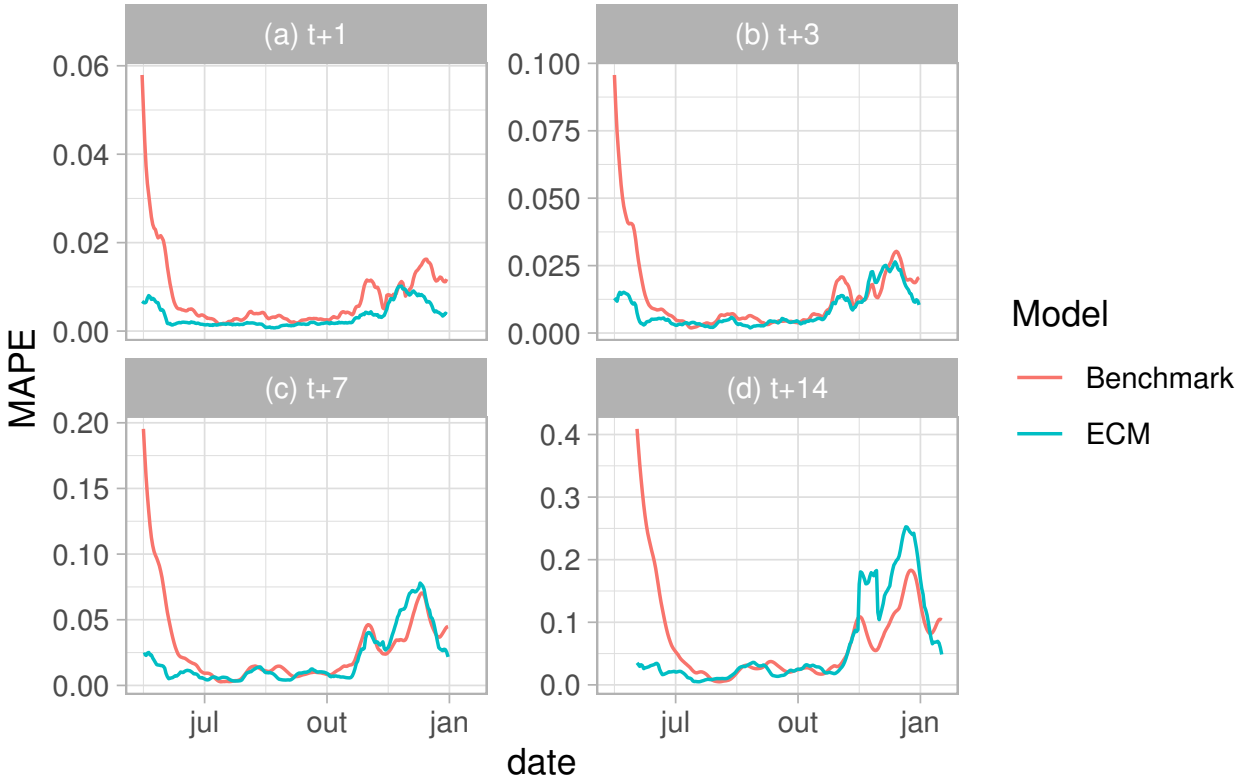

Figure 5: Cases Rolling Mean absolute percentage error - Portugal.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

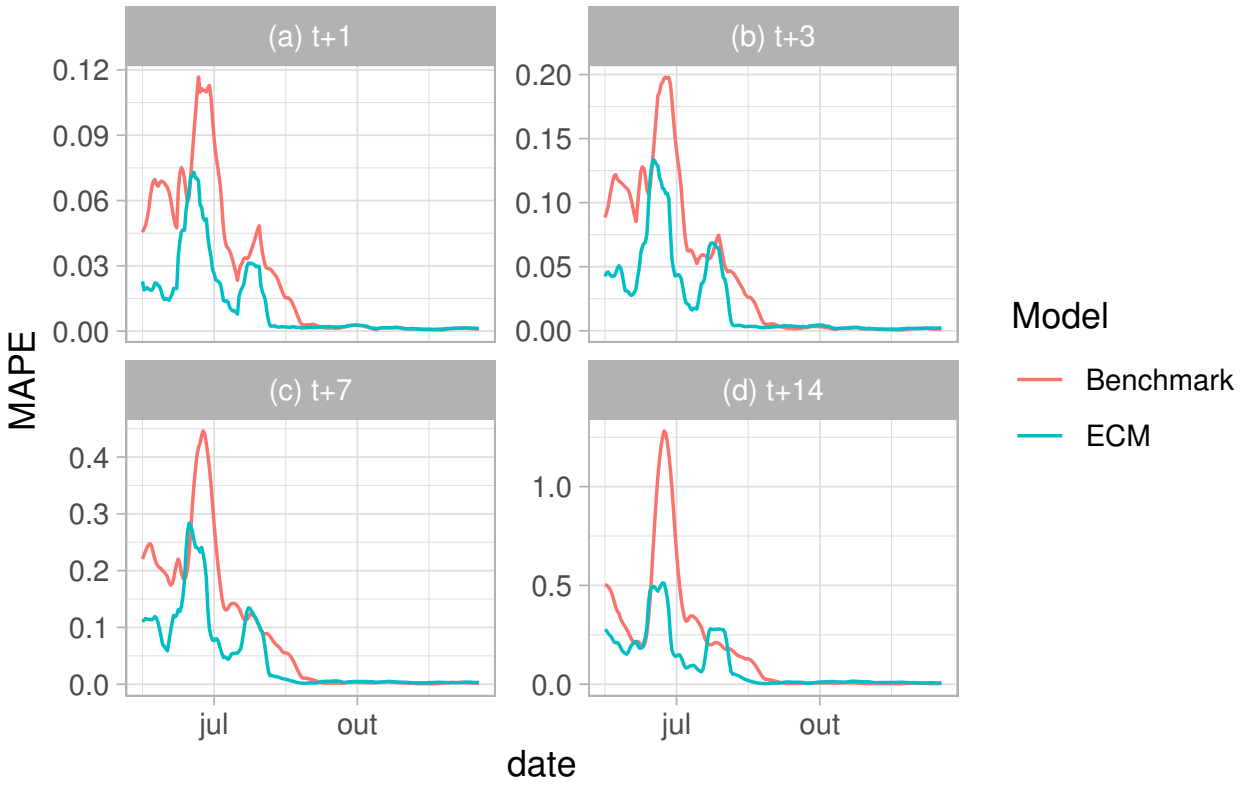

Figure 6: Deaths Rolling Mean absolute percentage error - Chile.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

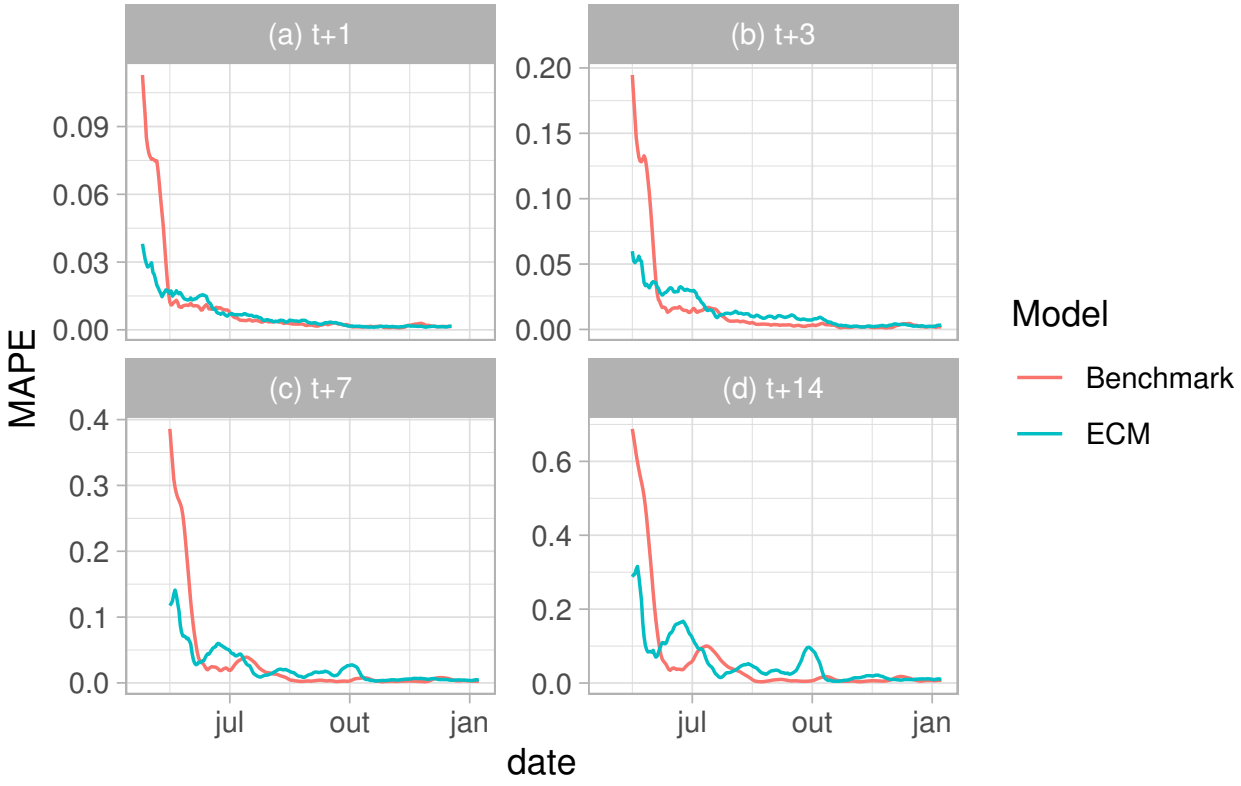

Figure 7: Deaths Rolling Mean absolute percentage error - Brazil.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

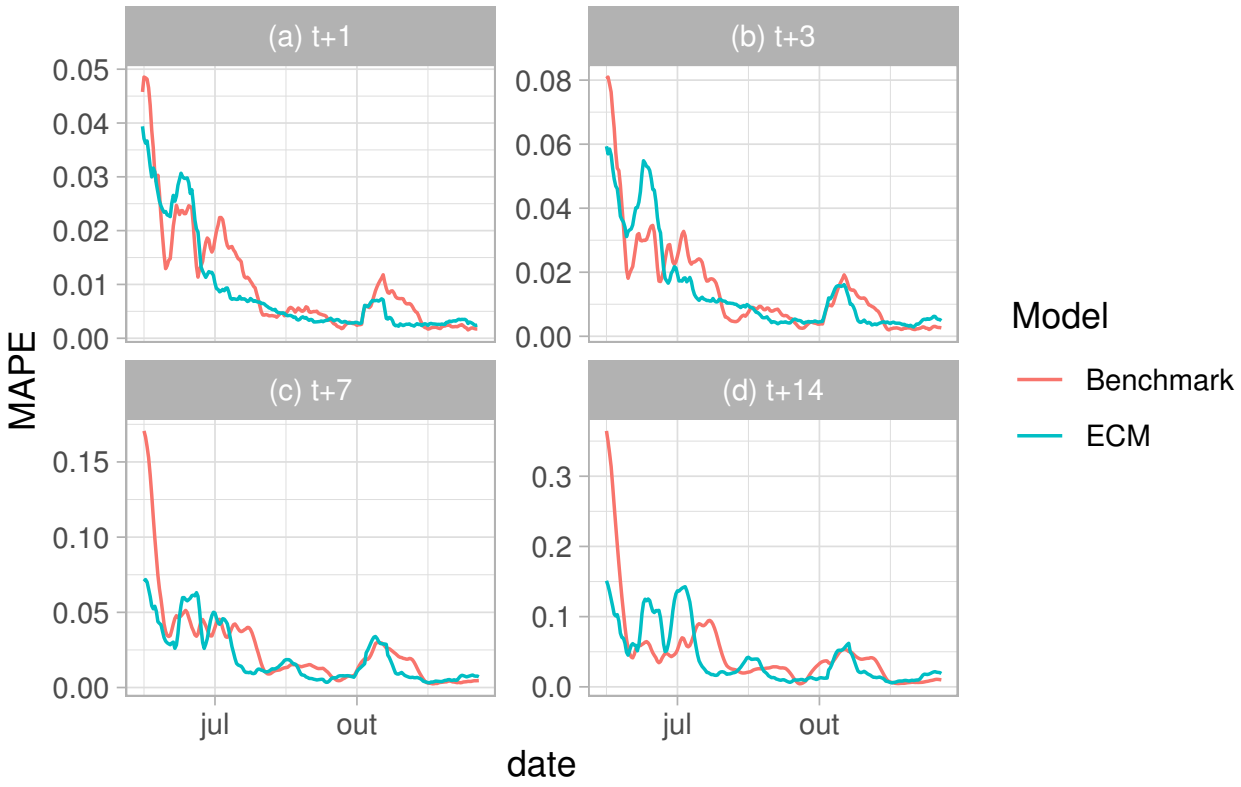

Figure 8: Deaths Rolling Mean absolute percentage error - Mexico.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

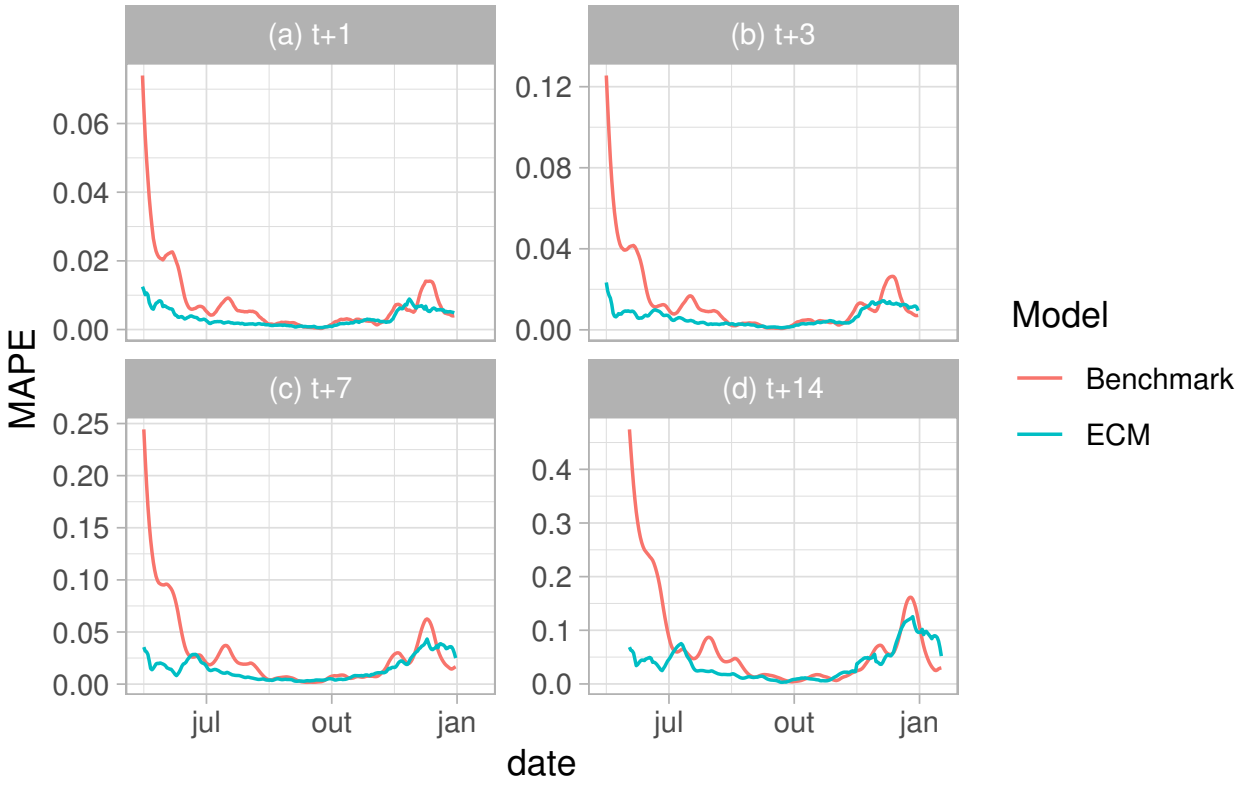

Figure 9: Deaths Rolling Mean absolute percentage error - Portugal.
The figure illustrates, for different horizons, the Mean Absolute Percentage Error (MAPE) computed over rolling windows with 14 observations.

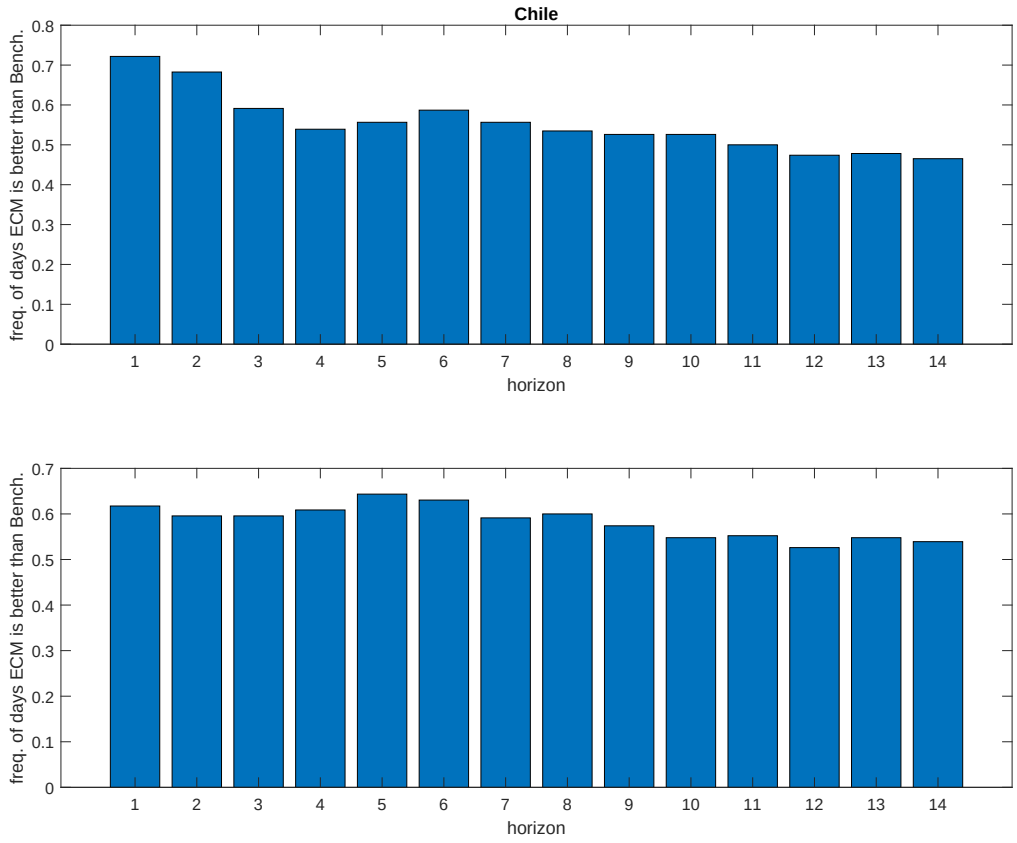

Figure 10: Chile: Frequency of days when ECM is better than the benchmark

The figure illustrates for Chile and for different horizons, the frequency of days when the absolute percentage error of the ECM is smaller than the one from the benchmark specification. Upper panel refers to cases. Lowe panel refers to deaths.

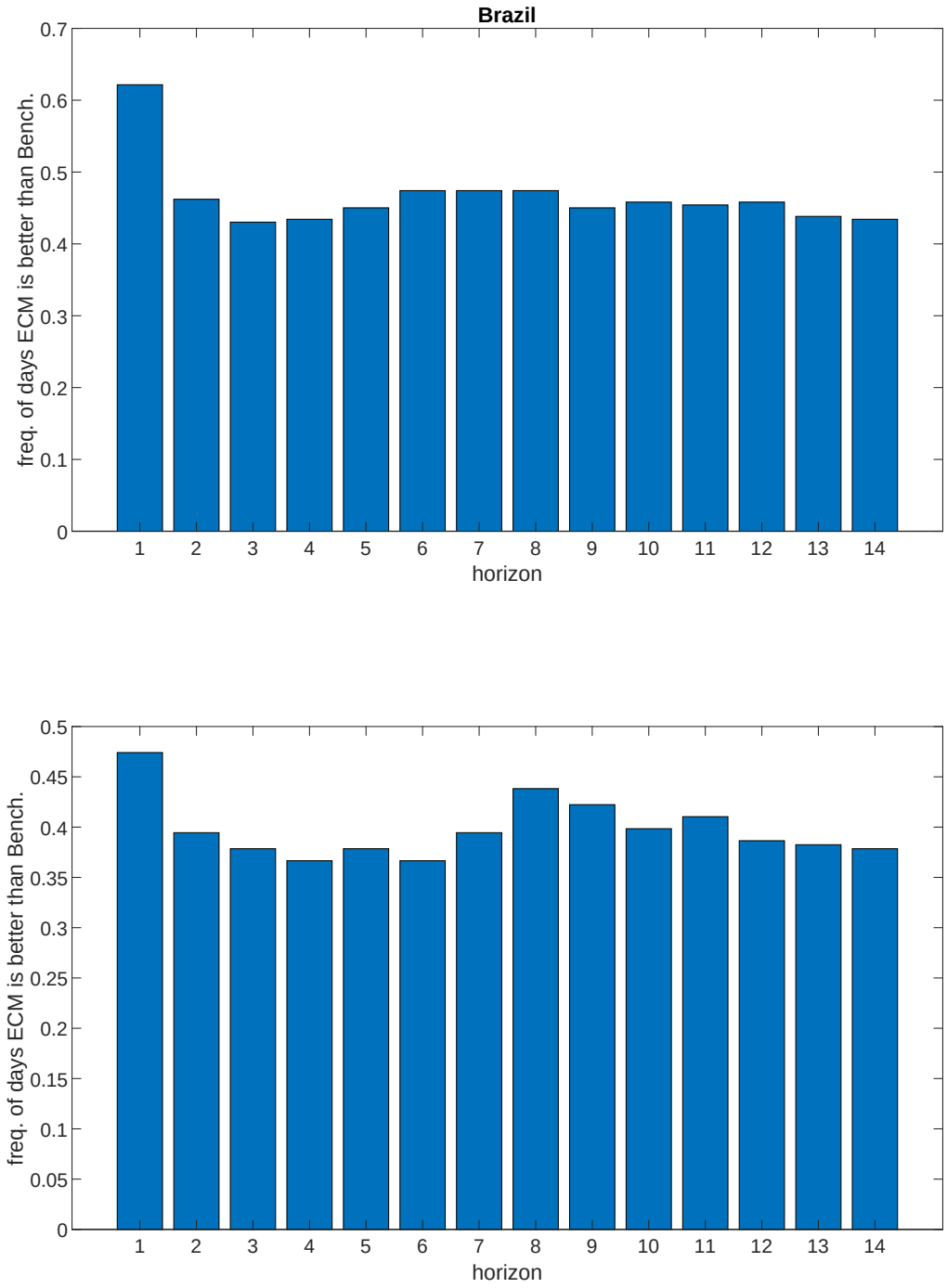

Figure 11: Brazil: Frequency of days when ECM is better than the benchmark

The figure illustrates for Brazil and for different horizons, the frequency of days when the absolute percentage error of the ECM is smaller than the one from the benchmark specification. Upper panel refers to cases. Lowe panel refers to deaths.

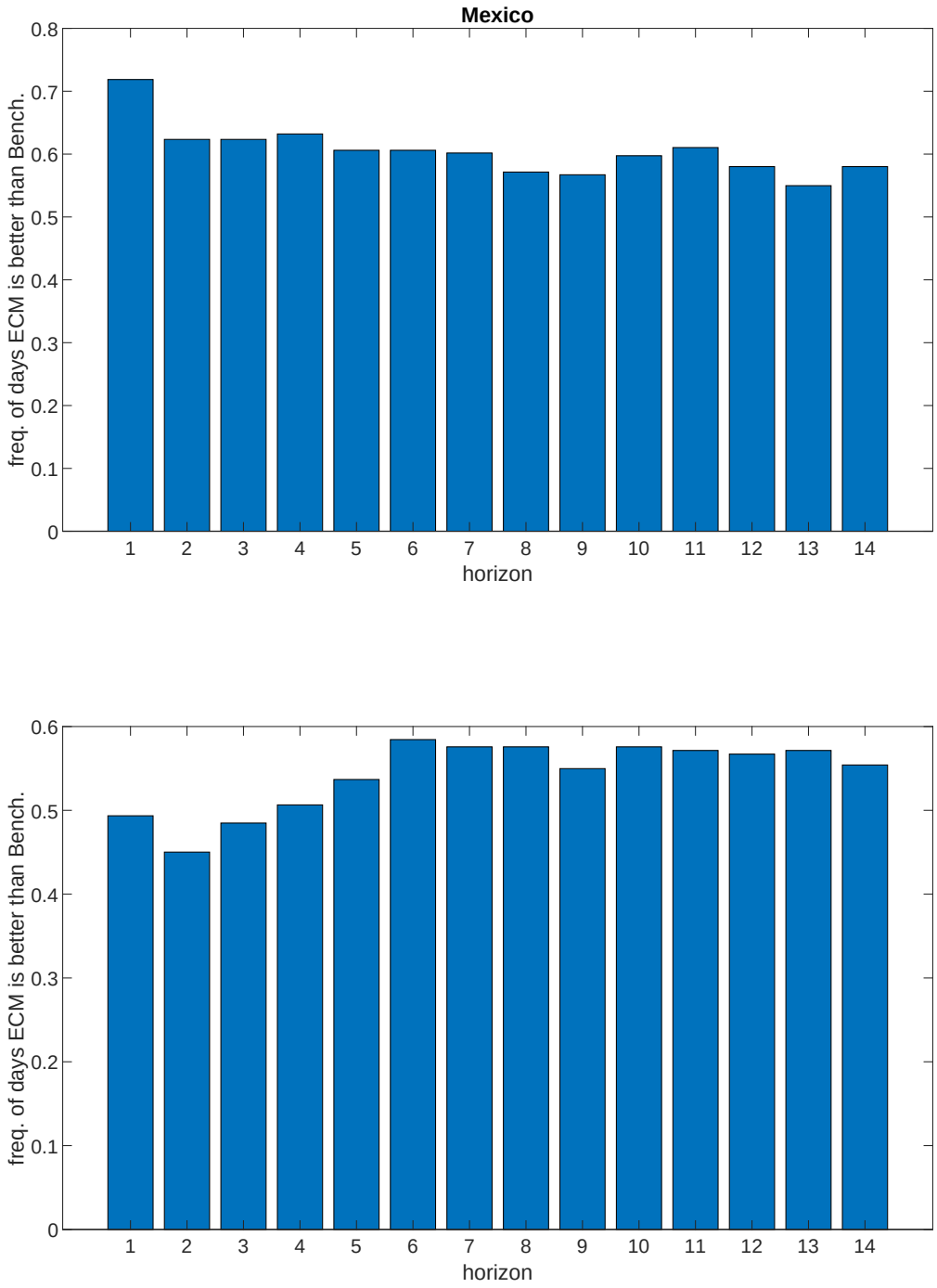

Figure 12: Mexico: Frequency of days when ECM is better than the benchmark

The figure illustrates for Mexico and for different horizons, the frequency of days when the absolute percentage error of the ECM is smaller than the one from the benchmark specification. Upper panel refers to cases. Lowe panel refers to deaths.

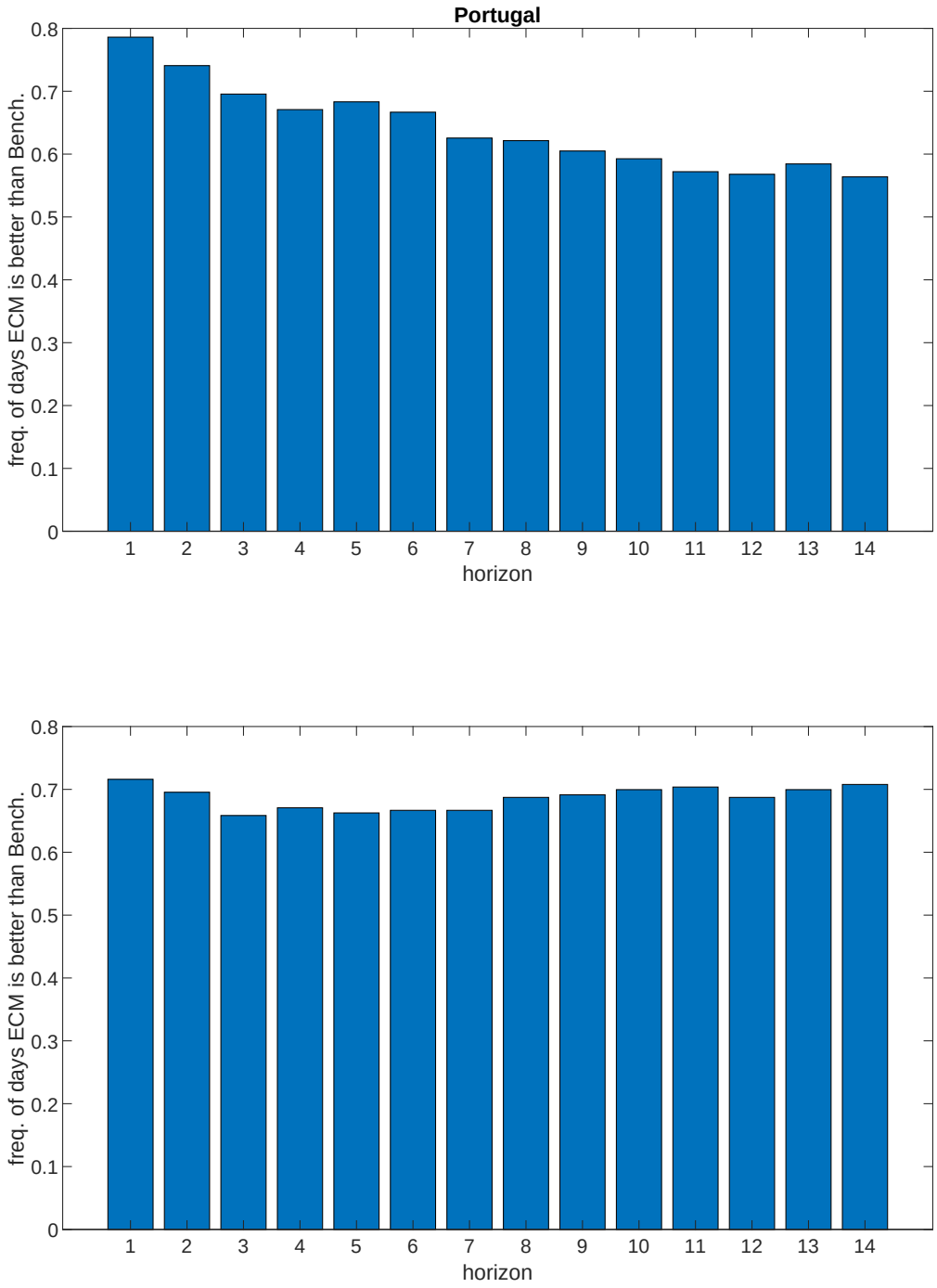

Figure 13: Portugal: Frequency of days when ECM is better than the benchmark

The figure illustrates for Portugal and for different horizons, the frequency of days when the absolute percentage error of the ECM is smaller than the one from the benchmark specification. Upper panel refers to cases. Lowe panel refers to deaths.

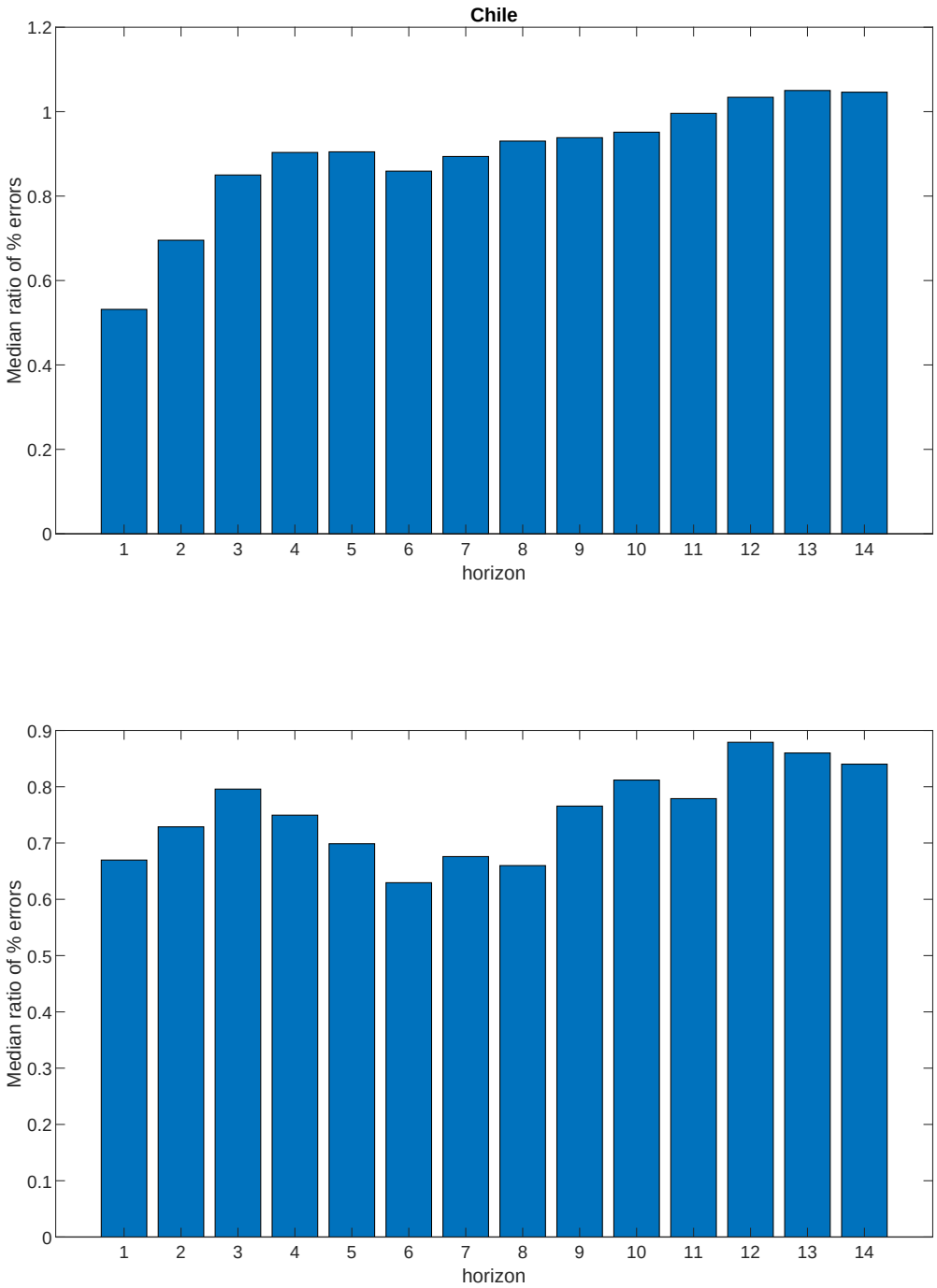

Figure 14: Chile: Median percentage error ratios

The figure illustrates for Chile and for different horizons, the median of the daily rations between the absolute percentage error of the ECM and the benchmark specifications. Upper panel refers to cases. Lowe panel refers to deaths. Numbers less than one favors the ECM model.

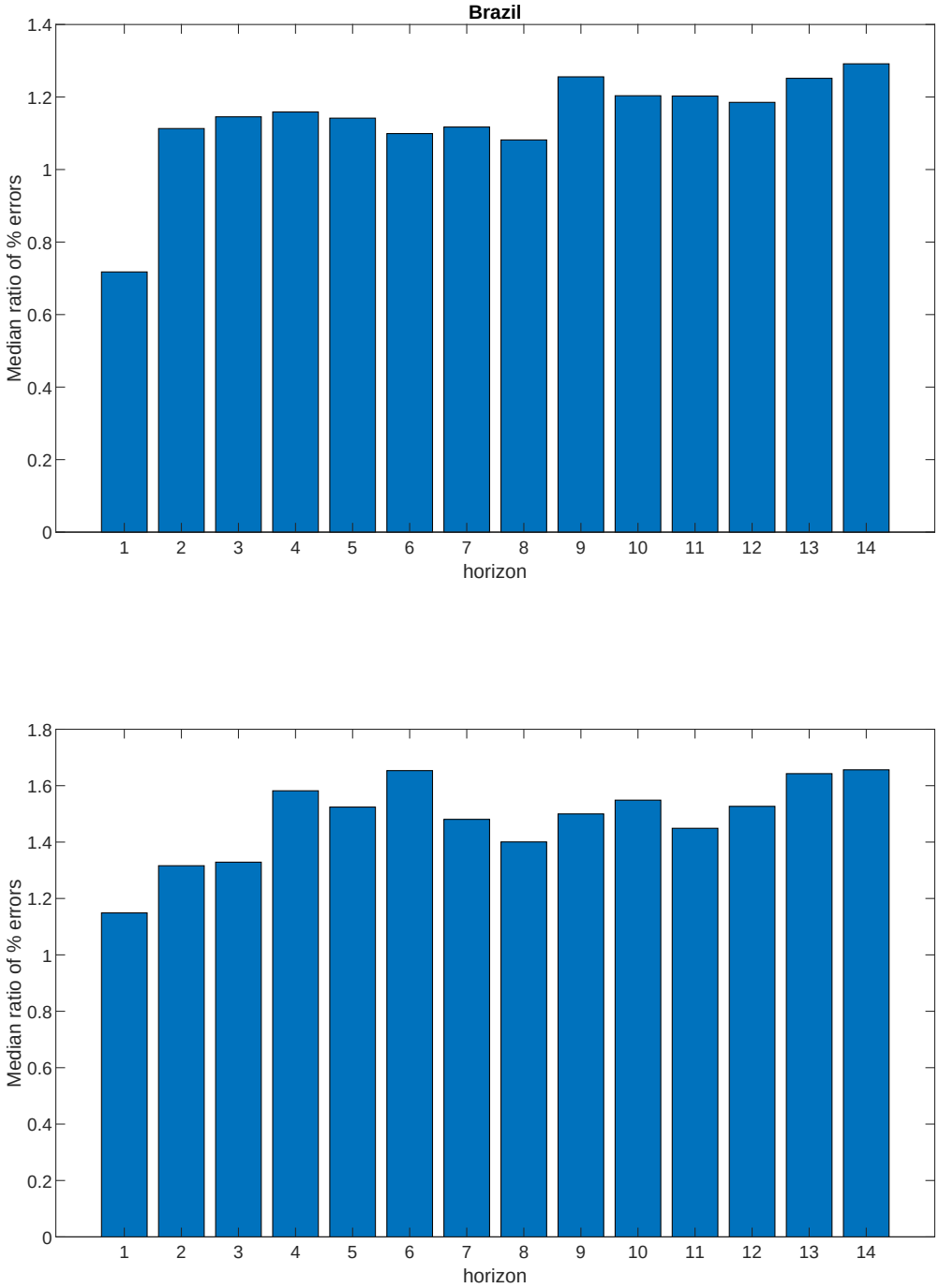

Figure 15: Brazil: Median percentage error ratios

The figure illustrates for Brazil and for different horizons, the median of the daily rations between the absolute percentage error of the ECM and the benchmark specifications. Upper panel refers to cases. Lowe panel refers to deaths. Numbers less than one favors the ECM model.

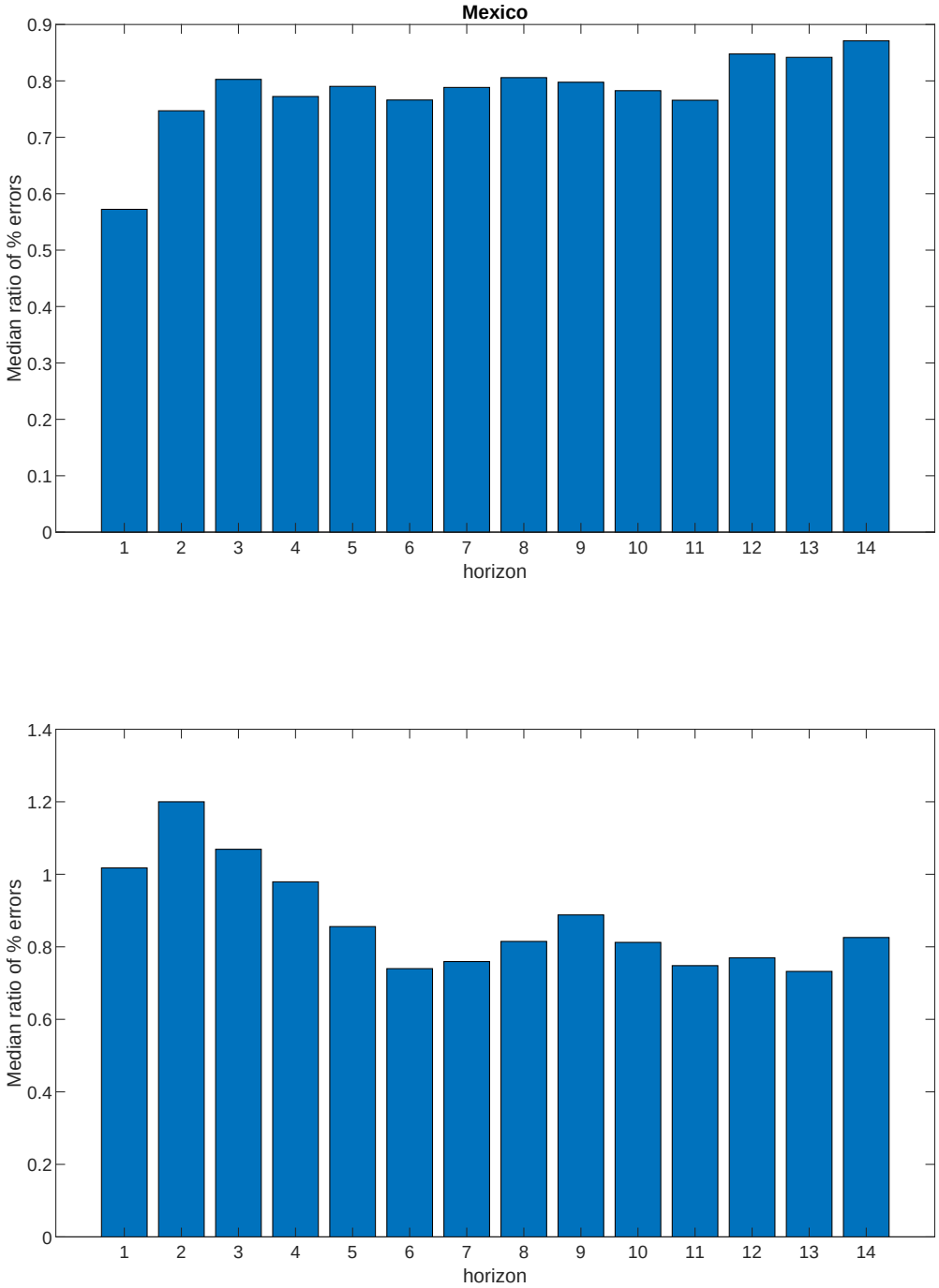

Figure 16: Mexico: Median percentage error ratios

The figure illustrates for Mexico and for different horizons, the median of the daily rations between the absolute percentage error of the ECM and the benchmark specifications. Upper panel refers to cases. Lowe panel refers to deaths. Numbers less than one favors the ECM model.

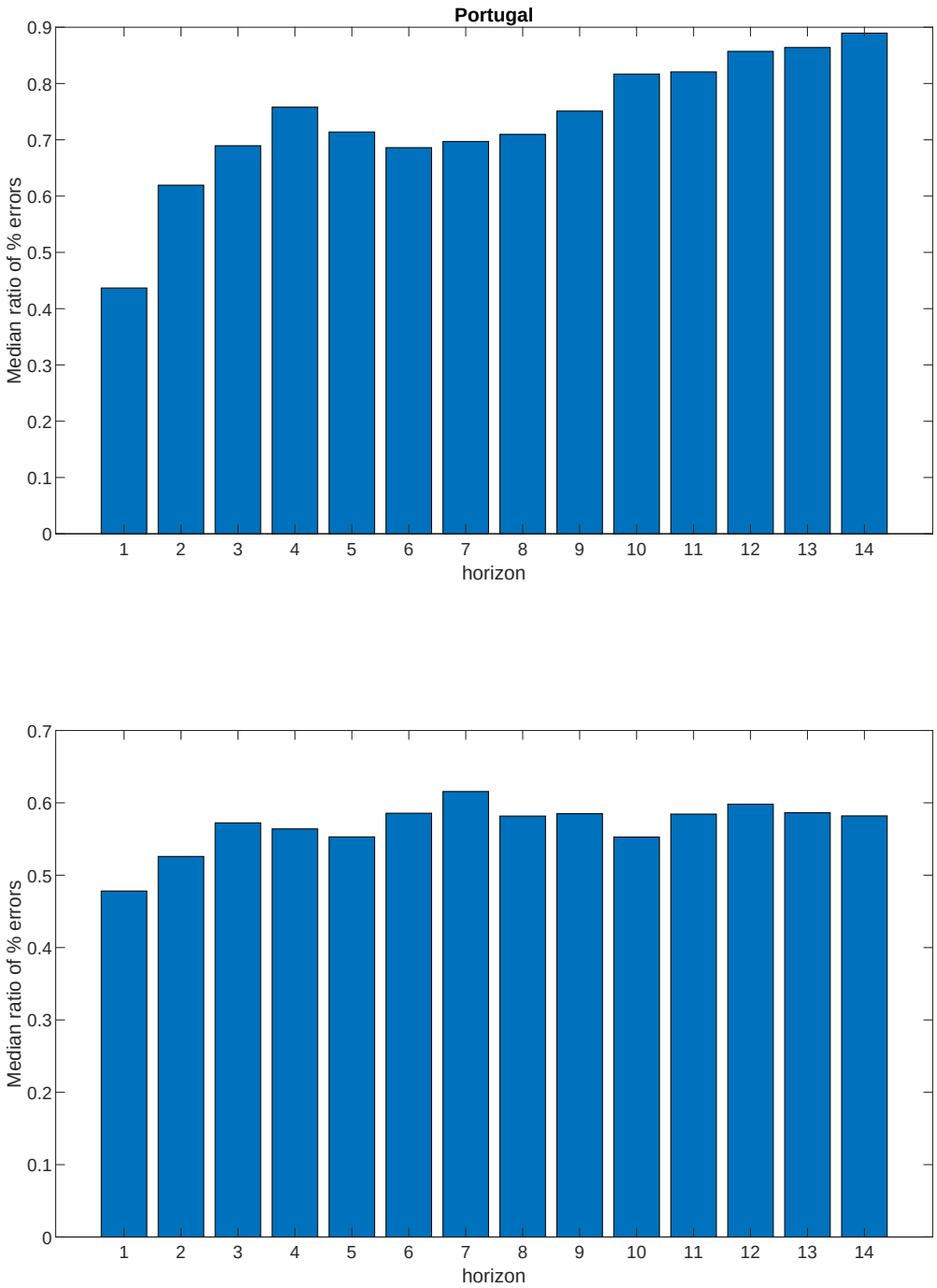

Figure 17: Portugal: Median percentage error ratios

The figure illustrates for Portugal and for different horizons, the median of the daily rations between the absolute percentage error of the ECM and the benchmark specifications. Upper panel refers to cases. Lowe panel refers to deaths. Numbers less than one favors the ECM model.

Table 3: Proportion of times each variable is selected by the LASSO

The table shows the frequency of times each variable is selected by the LASSO in the first-stage regression.

Target	t	t^2	France	Iran	Italy	Japan	South Korea	Singapore	Germany	Spain	United Kingdom	US
Brazil	0.52	0.07	0.50	0.64	0.70	0.37	0.53	0.57	-	-	-	-
Chile	0.63	0.16	0.55	0.55	0.38	0.45	0.32	0.69	0.25	-	-	-
Mexico	0.41	0.09	0.08	0.66	0.17	0.31	0.32	0.25	0.39	0.29	0.33	0.67
Portugal	0.54	0.42	-	0.69	0.69	0.40	0.52	-	-	-	-	-

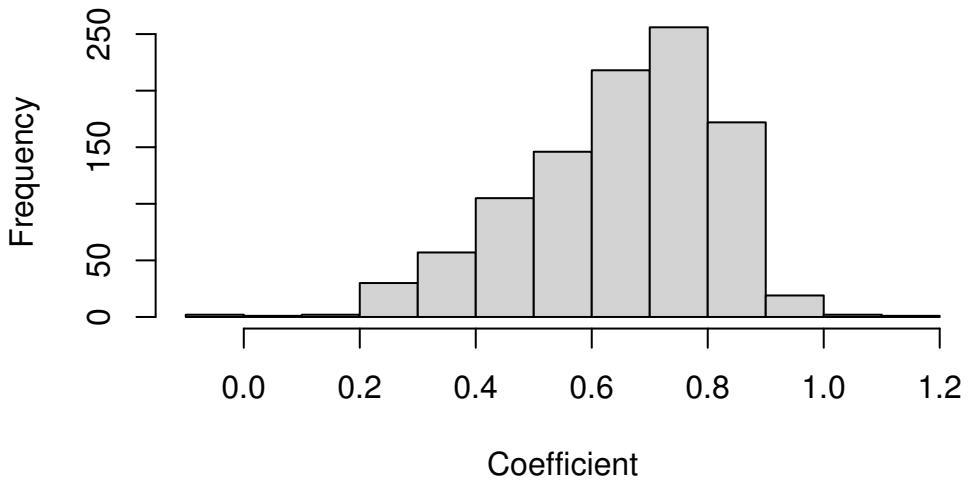

Figure 18: First-stage Residual AR Coefficients

The figure illustrates the empirical distribution of the estimated AR coefficients of an AR(1) model estimated with the residuals of the first-stage LASSO regression.

Table 4: Effects of data inflation.

Forecasting MAPEs of the ECM with data inflation divided by the forecasting MAPEs of the ECM without data inflation. Numbers lower than one favors the inflation heuristic.

	Country							
	Brazil		Chile		Mexico		Portugal	
horizon	cases	deaths	cases	deaths	cases	deaths	cases	deaths
1	0.868	1.040	0.828	0.942	0.915	1.009	0.807	0.880
2	0.950	0.999	0.899	0.912	0.971	1.036	0.814	0.833
3	0.971	0.964	0.847	0.915	0.989	1.044	0.834	0.820
4	0.944	0.925	0.828	0.914	0.989	1.066	0.860	0.822
5	0.907	0.889	0.827	0.915	0.974	1.083	0.873	0.852
6	0.870	0.854	0.817	0.925	0.946	1.052	0.878	0.865
7	0.857	0.840	0.818	0.922	0.948	1.051	0.883	0.875
8	0.859	0.825	0.834	0.922	0.959	1.054	0.892	0.873
9	0.849	0.827	0.842	0.925	0.966	1.052	0.900	0.861
10	0.854	0.825	0.848	0.929	0.974	1.060	0.916	0.855
11	0.866	0.820	0.852	0.927	0.992	1.064	0.932	0.862
12	0.864	0.811	0.850	0.931	0.991	1.056	0.947	0.872
13	0.865	0.811	0.844	0.929	1.001	1.056	0.958	0.878
14	0.860	0.813	0.837	0.941	0.983	1.050	0.970	0.882