 doi:10.20944/preprints202206.0086.v1

Article

Leveraging Geographically Distributed Data for Influenza and SARS-CoV-2 Non-Parametric Forecasting

Pablo Boullosa [1], Adrián Garea [1], Iván Area [2], Juan J. Nieto [3], Jorge Mira [1,4]*

[1] Departamento de Física Aplicada, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.;
[2] Universidade de Vigo. Departamento de Matemática Aplicada II. E.E. Aeronáutica e do Espazo. Campus de Ourense. 32003 Ourense, Spain.;
[3] Instituto de Matemáticas, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain. [4] Instituto de Materiais (iMATUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
* Correspondence: jorge.mira@usc.es

Abstract: The evolution of some epidemics, as influenza, shows common patterns both in different regions and from year to year. On the contrary, epidemics like the novel COVID-19 show quite heterogeneous dynamics and are extremely susceptible to the measures taken to mitigate their spread. In this paper we propose empirical dynamic modeling to predict the evolution of influenza in Spain's regions. It is a non-parametric method that looks into the past for coincidences with the present to make the forecasts. Here we extend the method to predict the evolution of other epidemics at any other starting territory and we test also this procedure with Spanish COVID-19 data. We finally build influenza and COVID-19 networks to check possible coincidences in the geographical distribution of both diseases. With this, we grasp the uniqueness of the geographical dynamics of COVID-19.

Keywords: non-parametric modeling; flu; influenza; COVID-19; SARS-CoV-2; Empirical Dynamic Modeling; forecasting

1. Introduction

Influenza (or flu) is an infectious respiratory disease caused by the influenza virus. It results in an estimated 250 000 to 650 000 deaths annually [1]. Until 2020, influenza epidemics showed a seasonal recurrence (Figure 1**a**), with a yearly pattern of exponential growth of infections, a marked peak, and a similarly prompt decay of new cases (Figure 1**b**). The width and magnitude of each outbreak varied moderately across years. These overall regularities [2] allowed a moderate success of different modeling approaches in forecasting outbreak magnitude and duration [3–6] .

In late 2019 the emergence of the COVID-19 (the infectious respiratory disease caused by the SARS-CoV-2 virus) global pandemics disrupted the seasonal influenza pattern [7]. The fast spread and severity of the disease led to the enforcement of strong distancing measures around the globe, which halted the propagation of other, less aggressive respiratory viruses such as influenza's [8]. Contrary to the regularities observed in flu epidemics, COVID-19 data is very erratic. On the one hand, we do not know yet whether it will become a stationary disease – with patterns of infection growth and decay similar to those of Flu. If that were the case eventually, we have not observed the new epidemics for a sufficiently long time yet as to infer repeated trends. On the other hand, while influenza is dealt with casually, the countermeasures to tackle COVID-19 have disrupted its natural cycle. When these counter-measurements were relaxed, new outbreaks emerged. This resulted in an irregular train of *waves* that often overlapped in time. As measures were dictated by an array of authorities (from local to supra-national), these waves differed wildly across regions – even within a same country (Figure 1**c**). Some biological aspects of SARS-CoV-2 contributed to the disarray: it presents a long incubation period [9] during

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

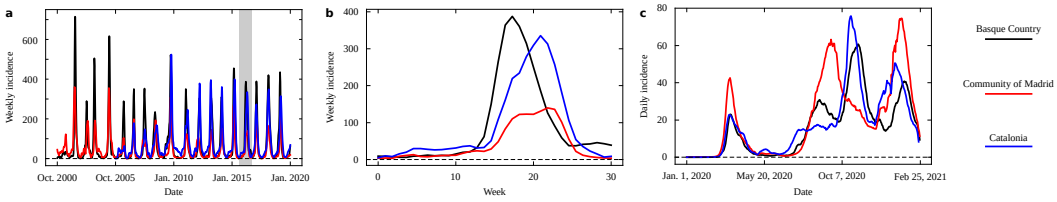

Figure 1. Empirical time series of the influenza and SARS-CoV-2 epidemics. a Examples of historical time series of flu over 20 years in three Spanish regions: Basque Country (black), Community of Madrid (red), Catalonia (blue). Data from each year (spanning from week 40 of a given year to week 20 of the next year) have been concatenated omitting the warm season (during which incidence is negligible). The gray area is expanded in **b** to show the yearly exponential raise, peak, and fall that characterizes the influenza cycle. Here, data from 2015/16. **c** Evolution of the SARS-CoV-2 pandemics in the same three regions shows the pattern of waves within a single year, which are not always in phase across regions.

which an infected person does not show symptoms; a large fraction of people suffer an asymptomatic version of the disease, but they can propagate the virus [10]; and, also, large outbreaks have been attributed to single *super-spreaders* who infected up to hundreds of people during a single event [11]. Under such conditions, tracking the exact timing of each infection is difficult – which resulted in unreliable epidemiological time series.

The very mathematical nature of epidemic dynamics also hinders its forecasting. A popular approach to this problem are compartmental models. In them, a population is coarse-grained into broad classes (e.g. Susceptible, Infected, and Recovered) and simple rules are established to regulate the flows between compartments. Typically, Infected people move into the Recovered class at a given rate, while they infect Susceptible individuals (moving them to the Infected compartment) with some probability. These rates and probabilities can be empirical model parameters inferred from the data. Such simple models can correctly capture broad qualitative aspects – e.g. a phase of exponential growth or the existence of herd-immunity thresholds (when such critical fraction of the population has been infected, the outbreak remits spontaneously). However, these models are notably bad at forecasting real-life scenarios. On top of all the troubles affecting data quality mentioned above, the phase of exponential growth in epidemic dynamics constitutes an important limiting factor. As it happens in deterministic chaos [12], small errors or uncertainties in the data are magnified exponentially by the epidemic dynamics themselves. Unlike chaotic systems, epidemic peaks can be fairly stereotypical; but the effect of exponentially magnified errors is enough to prevent a systematic and precise forecast of an outbreak's magnitude and duration [13]. This exponential factor in epidemic dynamics results in broad ranges or parameters that can fit correctly past data, but that are compatible with wildly diverging future behaviors.

In this framework, we have turned our attention to Empirical Dynamic Modeling (EDM) [14], a form of non-parametric modeling that looks at past examples of how a dynamical process (e.g. an epidemic) looked like, and uses them to forecast how it might unfold. Traditional EDM applications take a specific historic time-series (e.g. epidemic data of influenza in a French region [15]), then look at more recent data (say, the last 5 weeks of influenza cases in that same region) and find instances of the past series that resemble the new data. The known progress of the closest historical matches is used as an estimate for the evolution of the current situation. This avoids fitting empirical data to highly sensitive exponential dynamics that magnify small errors. Instead, it relies on bounded, repeated trends of a same kind.

Because of the issues around SARS-CoV-2 data outlined above, EDM seems still an unfit technique for the ongoing pandemic. The strong regularities seen in influenza mitigate some of these problems, making it an excellent test-bench for this approach. In an attempt to improve forecasting for the current pandemics, we expanded the classical functioning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

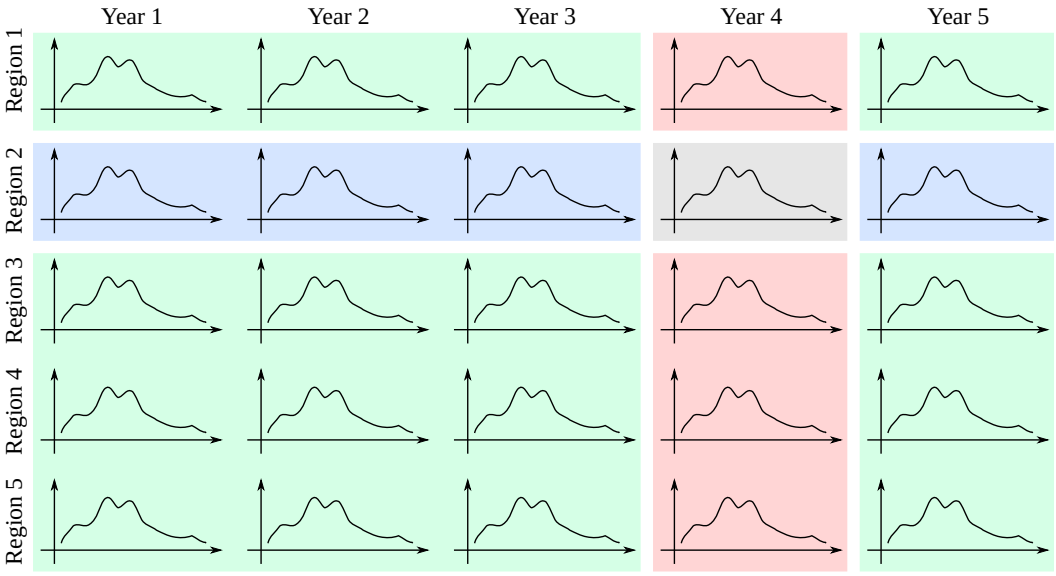

Figure 2. Illustration of different data pools for EDM. In grey, the example we want to forecast, the testing series. In blue, the method we call *classic*, which uses as library of patterns all the examples of the same region. In pink, the method *annual*, which uses as library all the examples from the same year. In green, the method *pool*, where we use the biggest library, taking all the series that are not from the same region or the same year.

of EDM to leverage geographically distributed data. Instead of using a single historic time series to predict what might happen in a specific region, we studied a set of areas on which epidemics unfolded simultaneously, and allowed historical series from each-other to serve as examples of how the dynamics might progress. We first used influenza data from different Spanish regions to study a controlled scenario of known regularity. Through this controlled case, we quantified how much of an improvement our approach is with respect to earlier applications of EDM that did not leverage similar dynamics in geographically distributed data. We explore briefly the more difficult case of SARS-CoV-2, in which data remains scarce and heterogeneity across regions is more pronounced. Our approach is moderately successful in capturing some features of the epidemics in different Spanish regions, while it fails in some important aspects. We propose that pooling geographically distributed data might speed up the gathering of recurring patterns, thus enhancing forecasting methods (beyond EDM) if COVID-19 becomes a seasonal disease. Finally, we turned EDM on its head to infer relationships between epidemic dynamics across different Spanish regions and over time. Since EDM uses past examples to forecast future dynamics, we quantified how often dynamical patterns from a region and year served as an estimate for each other's unfolding. Thus, we derive networks that illustrate epidemiological patterns across regions and correlations between flu strains from different years. These might offer relevant information to track which are the closest patterns that new epidemics follow.

2. Materials and Methods

2.1. Data and data preprocessing

We obtained time series of influenza cases from the Spanish National Center for Epidemiology (Instituto de Salud Carlos III). Out of 19 Spanish regions, we gathered data for 17 (which include 15 autonomous communities and the autonomous cities of Ceuta and Melilla). Data spans from 2000 to 2020, with some regions starting off at different times (as summarized in Table A1). Each time series starts at week 40 of a year and ends on week 20 of the next year – thus skipping the warm season in which influenza is uncommon.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

Raw data consists of weekly incidence per 100 000 inhabitants. This was smoothed with a 3-week moving average to mitigate sampling effects.

We obtained COVID-19 time series from the same Instituto de Salud Carlos III. This data spans from the beginning of the epidemic to April 19 2021. Raw data consists of the cumulative cases of COVID-19 in each Spanish region (all the Autonomous Communities, and the independent cities Ceuta and Melilla). We smooth the data with a 7-days moving average to mitigate sampling effects. From here, we derive daily incidence for each region and report it as number of cases per 100 000 inhabitants.

Let us note the time series of weekly (for flu) or daily (for COVID-19) new cases as $x(t)$, where $t \in \{T_0, \ldots, T_{end}\}$ is a discrete index in units of weeks or days respectively. Given data up to some time t, our task is to attempt a forecast for $x(t' > t)$. Following the literature on EDM (our tool of choice), we will work with the discrete derivative of $x(t)$:

$$\Delta x(t) \quad \equiv \quad x(t) - x(t - \Delta t). \tag{1}$$

We will base our forecast on this variable instead of on $x(t)$. This choice removes the effect of short-term linear auto-correlations [14].

2.2. Empirical Dynamic Modeling

Two approaches to forecasting stand out in the literature. On the one hand, modeling based on agents or equations try to capture the underlying causal processes behind a phenomenon (in our case, epidemic dynamics). Such causality is encoded by parameters – e.g. the likelihood that a contagious person infects someone else, or that he/she recovers from the disease. This approach allows us to understand a process and to test hypotheses to control it, but it is very sensitive to the array of problems discussed in the introduction. Non-parametric modeling, on the other hand, foregoes any attempt at understanding the mechanisms behind a phenomenon. These methods are more pragmatic – blindly seeking to extract as much useful information as possible to predict future scenarios. Little care is put on distilling this information into simple operating principles.

Empirical Dynamic Modeling (EDM) [14] is a non-parametric forecasting technique. EDM builds a library of dynamical patterns observed in the past history of a time series. Then, an ongoing situation is matched to examples from this library. The evolution of the matching examples becomes an estimator of how the current situation might progress. This method has been used in epidemiology under alternative names, such as the "Method of Analogues" [15].

To the best of our knowledge, all applications of EDM base their forecast for an ongoing time series on examples drawn from its own past history. This does not help much for the current SARS-CoV-2 global pandemics, for which at most one year of very irregular data exists. However, the epidemics has unfolded simultaneously throughout the world, effectively generating parallel samples of the same process. Can we leverage this geographically distributed information? To do so, we expand EDM's library of examples not only to the past of some ongoing dynamics, but to ongoing processes across different regions.

Let there be a short time series:

$$Y \quad \equiv \quad [y(t - n_L + 1),\, y(t - n_L + 2),\, \ldots,\, y(t)]. \tag{2}$$

This usually consists of the last n_L data points of an ongoing process, of which we wish to forecast the immediate future. The length of this short series (n_L) will be chosen as explained below. Additionally, let there be a set of longer time series:

$$\tilde{Y}^r \quad \equiv \quad [\tilde{y}^r(T_0^r),\, \ldots,\, \tilde{y}^r(T_{end}^r)]; \tag{3}$$

where the superscript r labels all available regions, and each corresponding series runs between times T_0^r and T_{end}^r (which might differ between regions). Let us call $\tilde{Y} \equiv \{\tilde{Y}^r\}$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

to the collection of all such series from all regions. $\tilde{Y}$ constitutes our library of dynamic patterns, upon which we will base our forecast for the future of Y.

To build this forecast, we first search for dynamical patterns in $\tilde{Y}$ that resemble Y. We compute the Euclidean distance between Y and each stretch of length n_L within $\tilde{Y}$:

$$d^r_{t'} = \sqrt{\sum_{i=1}^{n_L} \Big(y(t-n_L+i) - \tilde{y}^r(t'-n_L+i)\Big)^2}, \tag{4}$$

where we have selected each suitable stretch from region r and labeled it with its ending time t'. Note that both Y and each stretch of length n_L within the library are a point in an n_L-dimensional space. Equation 4 tells us how close to Y each point in the library is in this abstract space. Of all the examples available we select the n_B closest neighbors to Y. In this paper (and following suit with EDM literature [14]) we take

$$n_B \equiv n_L + 1. \tag{5}$$

We note the selected neighbors as $\{\hat{Y}_i;\, i = 1, \ldots, n_B\}$. Alongside these examples we store h time steps into their future, such that:

$$\hat{Y}_i \equiv \{\hat{y}_i(t'_i - n_L + 1), \ldots, \hat{y}_i(t'_i), \hat{y}_i(t'_i + 1), \ldots, \hat{y}_i(t'_i + h)\}. \tag{6}$$

Note that the index t'_i labels time differently within each selected example. We estimate the future h time steps ahead of the last point in Y as a weighted sum over the nearest neighbors:

$$\hat{y}(t+h) \equiv \frac{1}{Z}\sum_{i=1}^{n_B} \omega_i \hat{y}_i(t'_i + h); \tag{7}$$

where:

$$Z \equiv \sum_{i=1}^{n_B} \omega_i \tag{8}$$

and:

$$\omega_i \equiv e^{(-\beta d_i/d_0)}, \tag{9}$$

with d_i the Euclidean distance (as per Equation 4) to the i-th neighbor, d_0 the distance of the closest neighbor, and β a meta-parameter to be set as indicated below. Note that if $\beta = 0$, equation 7 renders just the mean of the evolutions, while for $\beta > 0$ we assign more importance to points closer to Y. In the limit $\beta \xrightarrow{\infty}$ only the closest neighbor contributes.

2.3. Meta-parameters, performance evaluation, and cross-validation

Back to our empirical time series, we will base our forecast on the observed $\Delta x(t)$ as defined in subsection 2.1 – so our library $\tilde{Y}$ will consist of all such time series for a set of regions. Let us label the time series of region r as $\Delta X^r(t) \equiv \{\Delta x^r(T_0), \ldots, \Delta x^r(T_{end})\}$, and say we wish to produce a forecast for this time series from some time $t \in \{T_0, \ldots, T_{end}\}$ onwards. First of all, this region will be removed from $\tilde{Y}$ (we would be cheating otherwise). Then, we take the shorter time series $Y \to \{\Delta x(t - n_L + 1), \ldots, \Delta x(t)\} \subset \Delta X^r(t)$, and proceed as indicated above. Note that this consists of the last data point before our forecast begins and the $n_L - 1$ previous time-steps. We can repeat this process for all possible t to evaluate repeatedly how well EDM performs on a given region.

Before we can do that, EDM presents two meta-parameters, n_L and β, that need to be assigned a numerical value to operate. The meta-parameter n_L determines the length of Y and that of the patterns within $\tilde{Y}$ that Y is compared to. In principle, we wish to take n_L as large as possible to find the most informative matches within $\tilde{Y}$. In the ideal case

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

we would find a pattern that completely matches $\Delta x^r(t')$ for all $t' < t$. In a realistic setup, an informative pattern might match up to a certain window in the past, and then diverge wildly from our ongoing time series. If n_L is too large, we risk missing good patterns because of this – so we need to balance a tradeoff. The other meta-parameter, β, determines a non-linearity in weighting the neighbors of Y – as explained above. There is no principled way to set these parameters beforehand, so we tune them using cross-validation on all the data sets.

To implement this cross-validation and measure the performance of our method, we use the correlation coefficient between our forecast and the empirical time series. Say we have generated a forecast of how Y will behave h time units (weeks for flu, days for COVID-19) ahead, and that we have done this for all possible $t \in \{T_0 + n_L, \ldots, T_{end} - h\}$ for a region's time series. Then:

$$\rho^r(h) \equiv \frac{\mathrm{Cov}[y(t+h), \hat{y}(t+h)]}{\sigma[y(t+h)]\sigma[\hat{y}(t+h)]} \tag{10}$$

measures the correlation coefficient ($\rho^r(h) \in [-1, 1]$) between the forecast and the actual data in this region; noting that $y(t+h) \equiv \Delta x(t+h)$, $\hat{y}(t+h)$ is given by equation 7, and $\mathrm{Cov}[\cdot,\cdot]$ and $\sigma[\cdot]$ indicate covariance and standard deviation respectively. We can obtain an average performance:

$$\rho(h) \equiv \langle \rho^r(h) \rangle_r, \tag{11}$$

where $\langle \cdot \rangle_r$ indicates average over regions.

We find optimal values of EDM's meta-parameters by repeatedly evaluating all regions with fixed n_L and β (thus obtaining $\rho^r(h; n_L, \beta)$ and $\rho(h; n_L, \beta) \equiv \langle \rho^r(h; n_L, \beta) \rangle_r$), and selecting the combination (n_L^*, β^*) that renders a largest correlation. Optima values of the meta-parameters might depend on the forecasting horizon h, thus $n_L^* \equiv n_L^*(h)$ and $\beta^* \equiv \beta^*(h)$. In our experiments, we evaluated EDM's performance for $n_L = 1, \ldots, 18$, and $\beta \in \{0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 5\}$.

3. Results

3.1. Pooling geographically distributed information enhances EDM performance on influenza data

We carried out a series of numerical experiments to test the performance of EDM with and without pooling together geographically-distributed information. This subsection reports results for flu data. Each experiment was carried out for a series of conditions that we label *pool*, *classic* and *annual*.

In the *pool* condition we separated our influenza data series by seasons. To build forecasts for a region in a given season, the pattern library $\tilde{Y}$ included past and future seasons from all regions (including the one being forecast), while data from any region and same season was removed from $\tilde{Y}$. Note, first, that including such future examples of the season being evaluated is standard in EDM [14]. We expect that the causality between a year and the next one is fairly broken. Second, imagine that an important region would present some idiosyncratic dynamics during a season, which is later replicated in some other areas. This trend could serve as an indicator of what might happen in those adjacent regions with some delay. If we would include all data from a given season, EDM could draw the inference in the opposite direction as well (using data of adjacent regions to forecast dynamics that had played out some weeks ahead). This is why, to be on the safe side, we removed all data from all regions for the season being forecast.

In condition *classic*, the library of patterns contained only examples from past and future seasons of a given region – which is how EDM was originally conceived [14], and how it has been applied, e.g., to forecasting flu trends in the past [15]. In condition *annual*, the library of patterns consisted of all the contemporary examples of a given region, ignoring all the examples from different years. This condition is proposed to measure the similarity between series from different regions in the same year, as the dominant flu strain will be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

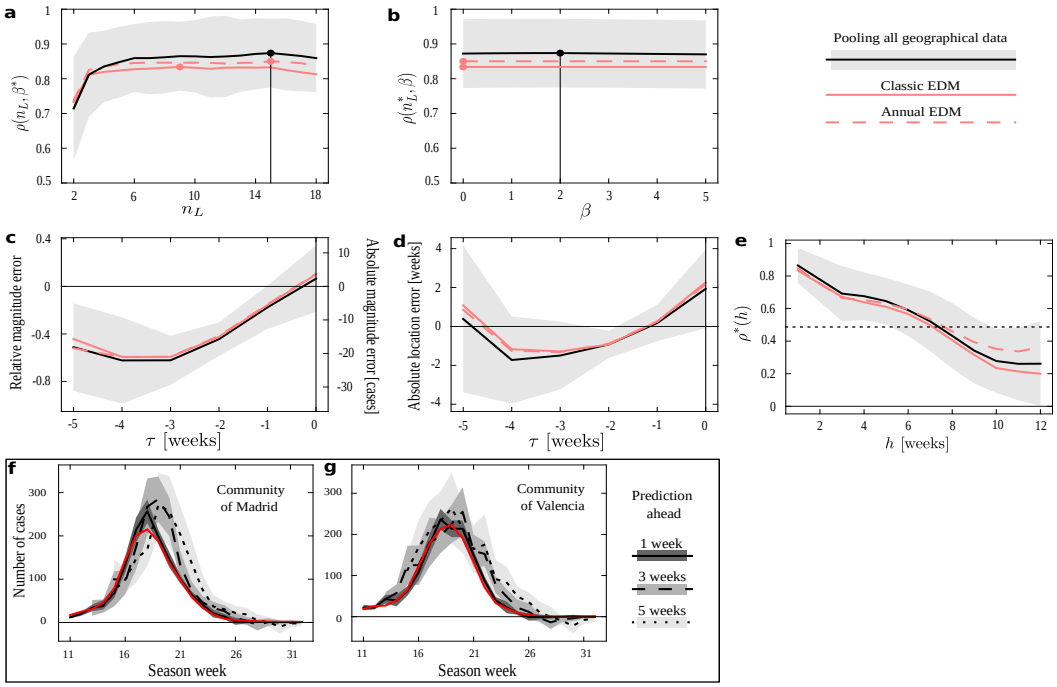

Figure 3. Pooling geographically distributed information for influenza forecast. a-e Results of different numerical experiments for conditions *pool* (solid black curves, with shading indicating standard deviation), *classic* (solid red) and *annual* (dashed red). Filled circles in **a-b** mark the location of optimal meta-parameters for each protocol. The optima for *pool* are also marked by vertical solid lines. Solid horizontal lines in **c-f** mark the 0 of the vertical axis. Solid vertical lines in **c-e** mark the location of the peak in time. Dotted horizontal line in **f** marks $\rho^*(h) = 0.5$. **a** EDM performance (as measured by correlation between data and forecast) as a function of n_L with fixed, optimal β^*. **b** EDM performance as a function of β with fixed, optimal n_L. **c** Average error in forecasting the peak magnitude. **d** Average error in forecasting the peak location. **e** EDM performance as we attempt to predict more time ahead. **f-g** Examples of how forecast become worst as we attempt to predict with more anticipation. Real data (solid red curves) is compared to forecasts derived with one week (solid black), three weeks (dashed black), or five weeks (dotted black) of anticipation. The various shadings indicate standard deviation of the estimated quantity. **f** Forecasts for the Community of Madrid. **g** Forecasts for the Community of València.

the same and there may the effect of a region could be transmitted to a neighboring one within the same season. In Figure 2 we represent the three conditions.

In Figure 3 we show EDM performance under the different conditions in a series of numerical experiments. If we keep $h = 1$ fixed, we are simply trying to predict the next amount of new cases following the available data. We see the performance on this task with optimal β^* and varying n_L in Figure 3**a**; and for optimal n_L^* and varying β in Figure 3**b**. In both plots, condition *pool* outperforms all others in almost all the ranges explored and, most importantly, it does so for the optimal $n_L^* = 15$ and $\beta^* = 2$ (even for the optima derived independently for all other conditions, marked by filled red circles). Such optima would be the meta-parameters with which we should operate if we tried to forecast new time series not present in our data set – and in all cases the results suggest that we should use condition *pool*. Condition *annual* performs slightly better than the *classic* EDM, and both fall below *pool*. This demonstrates an overall advantage of pooling together epidemiological data across regions. This result might have been expected, since the *pool* protocol provides us with more data in our training set. But it is not trivial that dynamics across regions (and, notably, having discarded series of a same season) would be informative to each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

other – each could have been affected by idiosyncratic factors such as population density, demographic structure, differences between urban and rural dynamics, etc.

The most important features that we would like to predict of an epidemic episode are how many people will be affected and how long it will last. The maximum height and location in time of the peak are a first proxy. To study how well we can forecast this, in a second experiment we aligned the data from all seasons taking each peak as a temporal reference. Then, we looked at how good the forecast of this peak was if EDM only had data until τ time units (weeks in the case of flu) before.

Figure 3**c** shows the average error (as relative and absolute magnitudes) that EDM makes in predicting the peak's height. Figure 3**d** shows the error (in weeks) in predicting when the epidemics will reach its maximum. We appreciate that all protocols present quite similar curves. Thus, while *pool* produces better forecasts in average (as shown above), our results suggest that the uncertainty in predicting the magnitude and end of an epidemic process cannot be alleviated by more abundant data. This is in line with recent research [13] that shows how behind epidemic processes lie mathematical mechanisms that make them inherently unpredictable. Unfortunately, our non-parametric method cannot circumvent such problems.

Looking at these plots with greater detail, we see how errors become smaller as we get closer to the actual peak – as might be expected (but see the case for COVID-19). The smallest average error in magnitude happens as the data up to the very time of the peak is considered ($\tau = 0$), while location is better predicted a week before the peak happens ($\tau = -1$). Error changes signs from negative to positive, meaning that EDM progresses (in average) from underestimating to overestimating. The forecast at $\tau = -5$ (which is the furthest from the peak that we can study with the available data) is more accurate than some others for peak magnitude and than many others for peak location. This effect is noteworthy for estimating peak location: this forecast degrades notably before becoming better – perhaps because the steepest phase of the exponential dynamics happens somewhere between $\tau = -5$ and $\tau = -1$. It is noteworthy, though, that this does not impact magnitude estimation as much.

With time series aligned with respect to their peaks as in the previous experiment, we also measured EDM performance (as captured by correlation between data and estimate) as a function of τ. This way we quantify how well our method works given that it is τ time units before or after the peak. Again, all protocols perform quite similarly in this experiment, with *pool* being notably worst than others in some cases. Figure 3**e** shows $\rho^*(h = 1; \tau)$, which starts and ends close to 0 (i.e. forecast is of poor quality further away from the peak). Performance raises up to $\rho^*(h = 1; \tau) \sim 0.5$ as the peak is approached, and remains at a similar level right after the peak before starting to decline gently. The dent at $\tau = 0$ (performance becomes factually nil) is explained because the slope of the data series changes around the peak. Unless both data and prediction are perfectly synchronized (which, Figure 3**d** proves, is not the case), this leads to an average correlation of zero at that point.

Finally, it is relevant to establish for how long a forecast remains informative. Figure 3**e** shows the EDM performance, $\rho^*(h; n_L^*, \beta^*)$, as it tries to predict h time units ahead in time. Correlation remains above 0.5 for predictions up to 7 weeks ahead of the available data, with *pool* being the preferred protocol in most cases. (Protocol *annual* becomes better around the time that correlation drops below 0.5.) We show examples of how a relatively worse (Figure 3**g**) and better (Figure 3**h**) forecast degrade as we elaborate estimates more time in advance. We see how this forecast degrades rapidly for a specific season of the Autonomous Community of Madrid, while it remains quite stable for some other season in the Community of València. This, together with the large deviations around most of the measures reported in Figure 3 (gray shadings), suggests that the right protocol might depend on the region studied, and that we might rather address this in a case by case basis. Below, we make some efforts to gain some insight about this issue.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

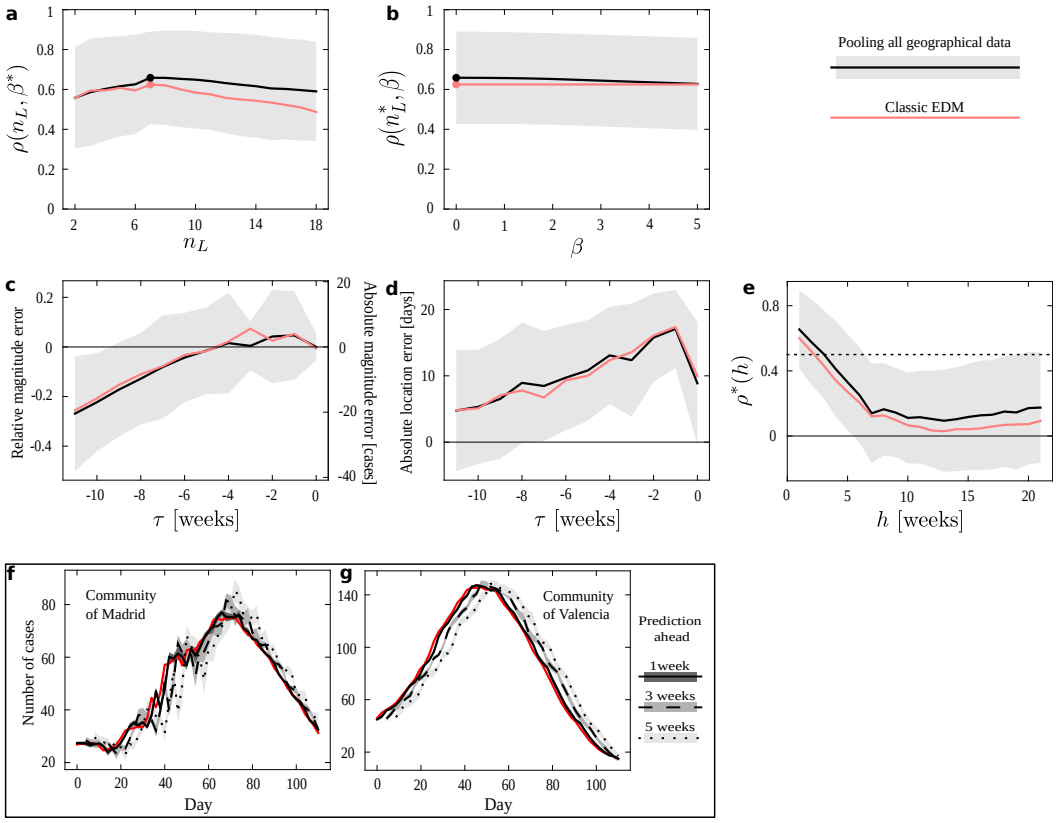

Figure 4. Pooling geographically distributed information for COVID-19 forecast. a-e Results of different numerical experiments for conditions *pool* (solid black curves, with shading indicating standard deviation), *classic* (solid red) and *annual* (dashed red). Filled circles in **a-b** mark the location of optimal meta-parameters for each protocol. The optima for *pool* are also marked by vertical solid lines. Solid horizontal lines in **c-f** mark the 0 of the vertical axis. Solid vertical lines in **c-e** mark the location of the peak in time. Dotted horizontal line in **f** marks $\rho^*(h) = 0.5$. **a** EDM performance (as measured by correlation between data and forecast) as a function of n_L with fixed, optimal β^*. **b** EDM performance as a function of β with fixed, optimal n_L. **c** Average error in forecasting the peak magnitude. **d** Average error in forecasting the peak location. **e** EDM performance as we attempt to predict more time ahead. **f-g** Examples of how forecast become worst as we attempt to predict with more anticipation. Real data (solid red curves) is compared to forecasts derived with one week (solid black), three weeks (dashed black), or five weeks (dotted black) of anticipation. The various shadings indicate standard deviation of the estimated quantity. **f** Forecasts for the Community of Madrid. **g** Forecasts for the Community of València.

3.2. Exploring EDM on COVID-19 data.

Data of the COVID-19 epidemic dynamics is affected by the various sources of unpredictability discussed above – some related to the unanticipated emergency caused by the pandemics, some others related to intrinsic properties of this malady and our social interplay with it. We have attempted to use EDM, pooling distributed geographic information from various sources, to forecast the dynamic unfolding of this crisis. Our success differed between more global (incorporating data from countries around the world) and local (as in our example from Spanish regions) attempts, and it changed over time as the pandemic changed as well. In this section we report a brief example based on the same regions as above, now studying only conditions *pool* and *classic*. While far from successful, this attempt at forecasting allows us to quantify some aspects that reveal how the new virus unfolds with dynamics very different from those of seasonal influenza. Our data series in this case give us new infections per day, instead of weeks, so some results do not translate as readily.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

Figure 4**a** shows the average EDM performance as a function of n_L with fixed optimum $\beta^* = 0$. We see an optimum $n_L^* = 7$ (days), which is much smaller than the $n_L^* = 15$ (weeks) found in the case of flu. This reveals how much more changing are the dynamics for COVID-19, and how informative patterns degrade more promptly as we attempt to compare them during longer stretches of time. This is indicative of a higher number of causal factors taking turns in dominating the dynamics – resulting in a more difficult forecast. Also, the correlation between estimates and prediction does not reach values comparable to those achieved with influenza data. Figure 4**b** shows EDM performance as a function of β with fixed, optimal $n_L^* = 7$. Again we see how the *pool* condition renders better results.

We repeated the experiments to estimate the quality of peak forecast, but in this case taking into account that COVID-19 'waves' are much more vaguely defined than seasonal peaks. Also, in some cases, EDM did not forecast the existence of a peak (suggesting, in turn, that the epidemics might grow unstopped within the time-window that we looked ahead). We report only results for cases in which a peak was predicted, and the comparison of its magnitude and location with that of an ongoing wave was possible.

Figure 4**c** shows that the error in magnitude becomes fairly small around 5 days before the epidemic peaks. By that time, EDM can produce an accurate first proxy of what the number of affected people will be. However, Figure 4**d** shows that the error in location of this peak only grows as we get closer to it. This is opposed to the results for flu, for which both magnitude and location estimates improved as the peak was approached. This indicates that EDM is consistently forecasting maxima that lie each time further away in the future of the approaching target and, in other words, suggests that COVID-19 waves do not show tell-tale signs that they are turning – thus aggravating the unpredictability of this kind of dynamics.

The window of acceptable prediction capabilities is also much smaller for COVID-19 than for flu. Figure 4**e** shows how correlation between estimate and data has dropped below 0.5 already if we attempt to predict 5 days ahead. This is an insignificant forecasting window compared to the acceptable 7 weeks that we could look ahead with a similar accuracy in the case of flu. This points, one again, to the dynamical challenge posed by the SARS-CoV-2 pandemics.

Examples of forecast for the Community of Madrid (Figure 4**f**) and Community of València (Figure 4**g**) show very small deviations from their respective averages. This is due to the very scarce data available, which at the same time reveals a poverty of dynamical patterns to draw estimates from.

3.3. EDM as a tool to characterize the epidemic unfolding

Non-parametric forecasting methods are mainly results-oriented. They are often used as black-boxes – foregoing a deeper understanding of the dynamic process as long as forecasting works. This is opposite, e.g., to compartmental modeling, in which causal relationships and meaningful parameters are inferred. With this later approach, insights can be gained about the relevant factors in the unfolding of an epidemic. But we can turn EDM on its head, using its methods not as a predictor, but as a tool for correlating and clustering the dynamics across regions and years. Then: What regions are more informative to each-other? Can we reveal a spatial structure of how the flu or COVID-19 evolved in Spain? Are there idiosyncratic regions in which the dynamics play out rather differently? How do successive influenza seasons resemble each-other?

To answer these questions we scored how often each region was within the nearest networks of each other region.

There is a question remaining which is related to the fact that we are introducing data from several regions to predict another one, so can we use EDM as a clustering tool? The answer lies on analysing how regions interact with each other -and itself- by checking how many neighbors one region takes from the others' -or its- time series. This helps us to generate a weighted and directed graph for all regions which may be useful to study

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

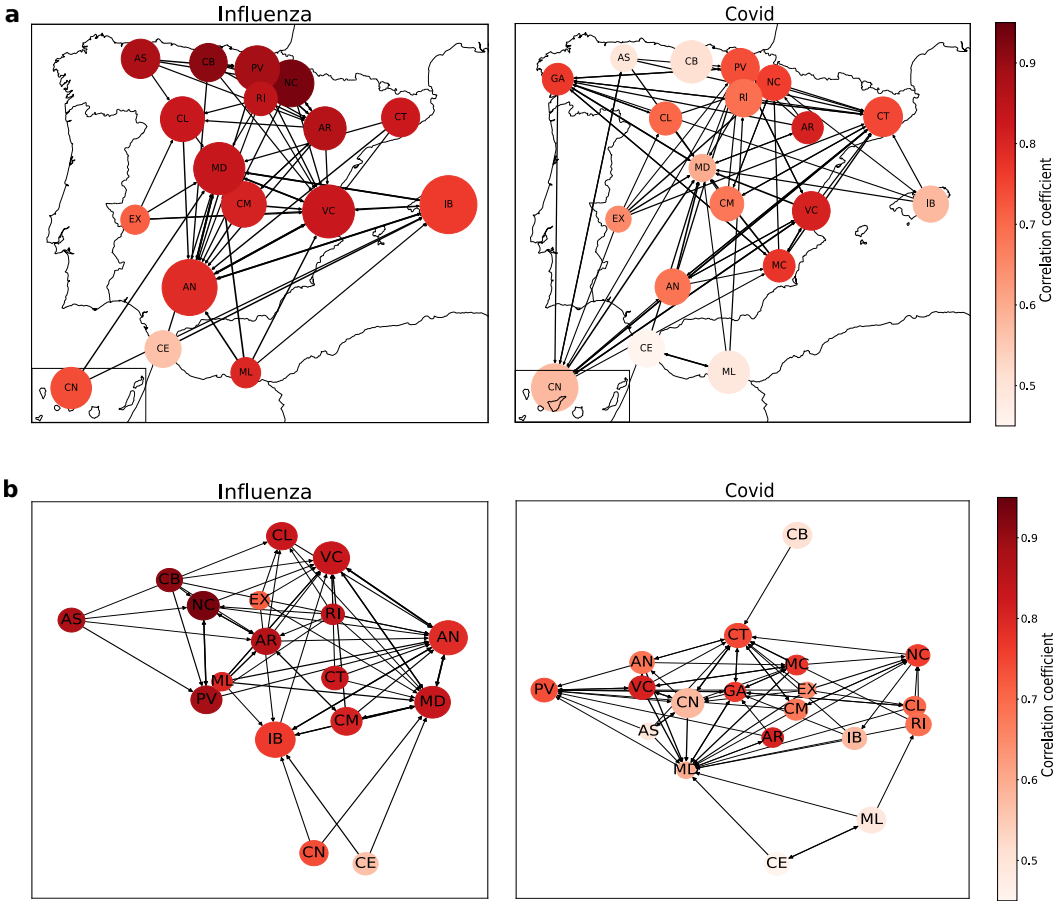

Figure 5. Influenza and COVID-19 networks. The size of the nodes is directly proportional to how many times a certain region has taken an example of itself. The darker a node is, the better it can be described -attending to the correlation coefficient ρ. Connection between generic regions A and B is plotted if the number of examples A takes from B overlaps 1.25 times the median number of examples A takes from other regions. **a** Geographical representation. **b** Graphical representation.

ongoing dynamics where there are not enough past data to make good predictions using other's regions data. We applied this technique to both diseases -influenza and COVID-19-.

As it can be seen in Figure 5**a**, northern and central autonomous regions influenza dynamics are pretty well described -with $\rho > 0.85$-, while the southern ones are one step behind -not so far- with ρ around 0.8 -Andalusia (AN), Canary Islands (CN), Extremadura (EX), Melilla (ML) and Balearic Islands (IB) as well-. Northern regions are also well connected between them and also with Andalusia, which even being at the south keeps a good relationship with other northern autonomous regions. This can be observed at Figure 5**b**, where we plot both influenza and COVID-19 networks with a random display, in order to visualize that well-described regions cluster together.

But the main fact is that none of the considered regions takes way more examples from itself than from the other ones. If we have a look to the proportional number of neighbors chosen by the EDM for one region respect from the others, it goes from 2% to 7% of the total examples, with a mean of 4.0% and a standard deviation of 1.6%[1]. This means these regions' dynamics may be similar and EDM does not notice any region to be considered "special" from the others, as all of them take examples from other autonomous regions.

The final goal of this work was to develop a non-parametric prediction method capable of estimate new dynamics when there is no historical data available, like in the case of

[1] The maximum 7% is in the confidence interval of two standard deviations, so we can assume it is just a statistical fluctuation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2022 doi:10.20944/preprints202206.0086.v1

COVID-19 pandemic -as it is a new disease with very little information at its start-. We tried to apply this method to COVID-19 data for several regions at worldwide scale -countries incidences- and at Spanish autonomous regions scale, but results were not as good as the ones obtained in influenza case. This bad performance can be explained mainly to the lack of historical data, but also because the incidence over all territories have not been the same and not even comparable, as data cannot be scaled from one region to another and examples taken by EDM might not be true to reality. In addition, the quality of the data acquired from governments have not been the best, as at the pandemic start they were running out of tests and infected people reports were not accurate enough [16].

This led us to compare how Spanish autonomous regions interacted with each other in this EDM approach considering COVID-19 dynamics from the first wave -from March 2020 to June 2020-, the second wave -from June 2020 to December 2020- and part of the third wave -from December 2020 to February 2021-, which ensures we have both ascending and descending trends so EDM will be able to choose which one is better in each analysed case. For this reason, we repeated the clustering experiment for these data and what we found out was there were many differences from the influenza epidemic network.

Having a look to the proportional number of neighbors chosen by the EDM -as we did in the section before- for one region respect from the others, it goes from 1.7% to 12.6% of the total examples, with a mean of 5.1% and a standard deviation of 3.3%.

There are some differences between the influenza and COVID-19 networks, but the most remarkable one is the fact that some regions take a large number of examples of themselves -in particular, Community of Madrid (MD), Valencian Community (VC) and Andalusia (AN)-. Correlation coefficients also reflect the bad performance of EDM predicting COVID-19 dynamics, as there are less well-described regions -with ρ over 0.85-.

In terms of connections, we have a more dispersed network, where there is no clear clustering as we had in the influenza network. The northern regions are now more connected with the southern ones, so we could think of it as an insight of a different relationship of similarity than in the influenza case -which could be related to the geographical locations and similar weather, leading to comparable incidences due to the way of life people develop-. Now we can observe that dynamics differ from the previous case studied -influenza-, probably related to the differences in autonomous regions pandemic management, as they were mainly independent from the central government and they carried out different measures to stop the propagation of this disease, while influenza has been fought for many years and this leads to more homogeneous actions. Despite this dispersion, we can observe at Figure 5**b** regions which are best described are centralized, as they are at the influenza network, denoting the potential application of this clustering method.

In summary, and taking all of this into account, there are several reasons why EDM is not able to perform well with COVID-19 pandemic data, but we can sum them up in two of them: lack of historical data and inhomogeneous disease incidences, which make regions dynamics be unpredictable from ones to the others.

4. Discussion

Among the many aspects that the COVID-19 pandemic has taught us, one clear is the need to rethink modeling approaches to predict the spread of this kind of disease in the current world. This requires a diversity of approaches, including the creation of observatories analogous to the ones of meteorologists [17]. In order to do this we need two ingredients: good data and good modeling tools.

Over the past century we have learned a lot about the dynamics that epidemics are very likely to follow. These happen to include exponential behaviors, such that the intrinsically correct models turn out to be extremely sensitive to the contingencies of real world data.

Real world data happens to have a lot of such contingencies (unknown causal factors that might be missed in the equations, errors in the collection of data, inconsistency of criteria in the recollection of data across time, etc). All of these trigger the sensitivity just

 doi:10.20944/preprints202206.0086.v1

discussed, in effect making it impossible to predict with the equations which, we know, are very likely correct.

Non-parametric modeling offers a way forward. If a global observatory is established to track this and future pandemics, we should base it on the methods introduced by Sugihara and May [14] and further studied in this paper.

Our addition to these methods, including pooling data from different regions in order to enlarge the library of patterns to look at to make the forecasts, has proven to improve the results for both the cases of influenza and COVID-19. However, the problem of predicting the epidemic peak is still challenging for the new disease, as the new peaks tend to be quite different to the older ones, independently from the region, as COVID-19 strongly depends on different political measures to fight its spread, the initial conditions and dominant strains.

The uniqueness of COVID-19 dynamics can be seen in the difference its network presents in comparison with influenza's. While neighboring regions present similar performances for the latter, they show a lot more of heterogeneity in COVID-19's network.

All analyses in this paper were based on the dynamics of the influenza and COVID-19 diseases independently. In expanding the observatory to potential epidemics in the future, we should contemplate the possibility of using the dynamics of a virus to attempt to predict the dynamics of another one (similarly to how here we use a region to predict another). This would provide very valuable information at the beginning of the pandemics. It might also helps us understand what causal agents are behind an observed contagion process – e.g. does a virus present long incubation periods, etc.

Also, such an observatory should make use of other sources of information. For example, we know that SARS-CoV-2 RNA can be located in the feces quite early after infection. Such early warning would be extremely valuable in planning to cope with the epidemics.

Author Contributions: Conceptualization, I.A., J.J.N. and J.M.; methodology, P.B., A.G. and J.M.; software, P.B. and A.G.; validation, I.A., J.J.N. and J.M.; formal analysis, P.B. and A.G.; investigation, P.B., A.G. and J.M.; resources, J.M.; data curation, P.B., A.G. and J.M.; writing—original draft preparation, P.B. and A.G.; writing—review and editing, I.A., J.J.N. and J.M.; visualization, P.B. and A.G.; supervision, I.A., J.J.N. and J.M.; project administration, J.M.; funding acquisition, J.J.N and J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Instituto de Salud Carlos III, within the Project COV20/00617 in the scope of the "Fondo COVID" of the Ministerio de Ciencia e Innovación of Spain, and by the crowdfunding program "Sumo Valor" of the University of Santiago de Compostela. Area and Nieto have been partially supported by the Agencia Estatal de Investigación (AEI) of Spain under Grant PID2020-113275GB-I00, cofinanced by the European Community fund FEDER. Mira is part of iMATUS, supported by Xunta de Galicia.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable

Acknowledgments: Data were provided by the Sistema de vigilancia de la gripe en España (SVGE), Red Nacional de Vigilancia Epidemiológica - Instituto de Salud Carlos III. We are very much indebted to Luís F. Seoane, from the Centro Nacional de Biotecnología - CSIC (Spain), for his invaluable help with this work.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202206.0086.v1

Appendix A. Data information

	Influenza	COVID-19
Andalusia (AN)	679 (00-20)	429
Aragon (AR)	615 (00-18)	421
Asturias (AS)	549 (04-20)	412
Balearic Islands (IB)	672 (00-20)	416
Basque Country (PV)	661 (00-20)	429
Canary Islands (CN)	582 (03-20)	419
Cantabria (CB)	539 (05-20)	415
Castile and León (CL)	681 (00-20)	423
Castile-La Mancha (CM)	681 (00-20)	416
Catalonia (CT)	495 (05-20)	442
Ceuta (CE)	489 (05-20)	402
Community of Madrid (MD)	661 (00-20)	457
Extremadura (EX)	582 (03-20)	416
Galicia (GA)	-	420
La Rioja (RI)	549 (04-20)	417
Melilla (ML)	356 (09-20)	405
Navarre (NC)	549 (04-20)	418
Region of Murcia (MC)	-	415
Valencian Community (VC)	679 (00-20)	429

Table A1. Length (number of data points) of each series. It spans from early 2000's to the beginning of 2020, but some series miss data from the beginning or the end. Their span is showed in brackets (beginning year - last year). Influenza data is weekly and only contains data from September to June. COVID-19 data is daily, from the beginning of the pandemic until April 19th of 2021.

References

1. A. D. Iuliano, *et al.*, Estimates of global seasonal influenza-associated respiratory mortality: a modelling study, The Lancet 391 (10127) (2018) 1285 – 1300. doi:https://doi.org/10.1016/S0140-6736(17)33293-2.
   URL http://www.sciencedirect.com/science/article/pii/S0140673617332932
2. Y. Cai, J. Li, Y. Kang, K. Wang, W. Wang, The fluctuation impact of human mobility on the influenza transmission, Journal of the Franklin Institute 357 (13) (2020) 8899–8924. doi:https://doi.org/10.1016/j.jfranklin.2020.07.002.
   URL https://www.sciencedirect.com/science/article/abs/pii/S0016003220304634
3. N. Stilianakis, A. Perelson, F. Hayden, Emergence of drug resistance during an influenza epidemic: Insights from a mathematical model, The Journal of infectious diseases 177 (1998) 863–73. doi:https://doi.org/10.1086/515246.
4. R. Casagrandi, L. Bolzoni, S. A. Levin, V. Andreasen, The sirc model and influenza a, Mathematical Biosciences 200 (2) (2006) 152 – 169. doi:https://doi.org/10.1016/j.mbs.2005.12.029.
   URL http://www.sciencedirect.com/science/article/pii/S0025556405002464
5. C. van den Dool, M. J. M. Bonten, E. Hak, J. C. M. Heijne, J. Wallinga, The effects of influenza vaccination of health care workers in nursing homes: Insights from a mathematical model, PLOS Medicine 5 (10) (2008) 1–8. doi:https://doi.org/10.1371/journal.pmed.0050200.
   URL https://doi.org/10.1371/journal.pmed.0050200
6. H. M. Dobrovolny, M. B. Reddy, M. A. Kamal, C. R. Rayner, C. A. A. Beauchemin, Assessing mathematical models of influenza infections using features of the immune response, PLOS ONE 8 (2) (2013) 1–20. doi:https://doi.org/10.1371/journal.pone.0057088.
   URL https://doi.org/10.1371/journal.pone.0057088
7. R. Soo, C. J. Chiew, S. Ma, R. Pung, V. Lee, Decreased influenza incidence under COVID-19 control measures, Singapore, Emerging infectious diseases 26 (04 2020). doi:https://doi.org/10.3201/eid2608.201229.
   URL https://wwwnc.cdc.gov/eid/article/26/8/20-1229_article
8. N. Jones, How covid-19 is changing the cold and flu season, Nature 588 (2020) 388–390. doi:https://doi.org/10.1038/d41586-020-03519-3.
   URL https://www.nature.com/articles/d41586-020-03519-3
9. S. A. Lauer, *et al.*, The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application, Annals of Internal Medicine 172 (9) (2020) 577–582, pMID: 32150748. doi:https://doi.org/10.7326/M20-0504.
   URL https://www.acpjournals.org/doi/10.7326/M20-0504

10. O. Byambasuren, O. Byambasuren, M. Cardona, K. Bell, J. Clark, M. Mclaws, P. Glasziou, Estimating the extent of asymptomatic covid-19 and its potential for community transmission: systematic review and meta-analysis, Official Journal of the Association of Medical Microbiology and Infectious Disease Canada 5 (4) (2020) 223–234. doi:https://doi.org/10.3138/jammi-2020-0030.
URL https://jammi.utpjournals.press/doi/full/10.3138/jammi-2020-0030
11. T. Frieden, C. Lee, Identifying and interrupting superspreading events-implications for control of severe acute respiratory syndrome coronavirus 2, Emerging infectious diseases 26 (03 2020). doi:https://doi.org/10.3201/eid2606.200495.
URL https://wwwnc.cdc.gov/eid/article/26/6/20-0495_article
12. E. Lorenz, Predictability: does the flap of a butterfly's wing in Brazil set off a tornado in Texas?, na, 1972.
13. M. Castro, S. Ares, J. A. Cuesta, S. Manrubia, The turning point and end of an expanding epidemic cannot be precisely forecast, Proceedings of the National Academy of Sciences 117 (42) (2020) 26190–26196. doi:https://doi.org/10.1073/pnas.2007868117.
URL https://www.pnas.org/content/117/42/26190
14. G. Sugihara, R. M. May, Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, Nature 344 (6268) (1990) 734.
15. C. Viboud, P.-Y. Boëlle, F. Carrat, A.-J. Valleron, A. Flahault, Prediction of the Spread of Influenza Epidemics by the Method of Analogues, American Journal of Epidemiology 158 (10) (2003) 996–1006. arXiv:https://academic.oup.com/aje/article-pdf/158/10/996/182648/kwg239.pdf, doi:10.1093/aje/kwg239.
URL https://doi.org/10.1093/aje/kwg239
16. Working group for the surveillance and control of COVID-19 in Spain, The first wave of the covid-19 pandemic in spain: characterisation of cases and risk factors for severe outcomes, as at 27 april 2020, Eurosurveillance 25 (50) (2020). doi:https://doi.org/10.2807/1560-7917.ES.2020.25.50.2001431.
URL https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.50.2001431
17. W. H. Press, R. C. Levin, Modeling, post covid-19, Science 370 (6520) (2021) 1015. doi:https://doi.org/10.1126/science.abf7914.
URL https://www.science.org/doi/full/10.1126/science.abf7914