Annals of Medicine and Surgery 66 (2021) 102437

Contents lists available at ScienceDirect

Annals of Medicine and Surgery

journal homepage: www.elsevier.com/locate/amsu

Experimental Research

Understanding the cycles of COVID-19 incidence: Principal Component Analysis and interaction of biological and socio-economic factors

Pablo Duarte [a], Efrain Riveros-Perez [b,c,*]

[a] Flossbach von Storch Research Institute, Germany
[b] Medical College of Georgia, Department of Anesthesiology and Perioperative Medicine, USA
[c] Outcomes Research Consortium. Cleveland Clinic, USA

ARTICLE INFO

Keywords:
COVID-19
Principal component analysis
Predictive model
Epidemiological data
Viral spread

ABSTRACT

The incidence curve of coronavirus disease 19 (COVID-19) shows cyclical patterns over time. We examine the cyclical properties of the incidence curves in various countries and use principal components analysis to shed light on the underlying dynamics that are common to all countries. We find that the cyclical series of 37 countries can be summarized in four principal components which explain over 90% of the variation. We also discuss the influence of complex interactions between biological viral natural history and socio-political reactions and measures adopted by different countries on the cyclical patterns exhibited by COVID-19 around the globe.

1. Introduction

Many infections undergo cycles and present waves of variable duration ranging from one to four years [1]. The two major contributors to the cyclic nature of respiratory viral infections are the changes in environmental parameters and human behavior [2]. The magnitude and the severe impact of COVID-19 on individual mortality, social interactions, strain on healthcare systems, and political and economic variables worldwide has led researchers to try to understand the complex interactions between the SARS CoV-2 (Severe Acute Coronavirus 2) virus, the individual host and the exposed population, and environmental factors. It is recognized that SARS CoV-2 as a coronavirus, has transmission epidemiology similar to influenza [3]. Influenza, as well as COVID-19, has followed wave patterns with a peak usually followed by a second wave a few months later [4]. Studying this cyclical pattern provides us with important insights about the nature of the cycle involving virus, host, community, and environment. Our study uses Principal Component Analysis (PCA) to model incidence patterns in different countries.

Fig. 1 shows the cyclical component of incidence series for Germany, Israel and the United States normalized such that the first observation corresponds to the peak of the first infection wave. The patterns are similar but not equal. While the cycles seem to move in a congruent way, the amplitude and length tend to vary. In this analysis, we apply frequency domain time series techniques to examine the extent to which common patterns in the cyclical components can be extracted allowing us to approximate the further movement of the series. We found that the variation of 37 incidence curves corresponding to the same number of countries can be reduced to four principal components that explain over 90% of the variation in the sample. We also show cycle predictions for countries with shorter incidence series since the peak of the first wave and a one-step ahead as a well as an out-of-sample estimation for Germany and the United States as a representation of countries with high overall incidence in two different continents. Both countries are currently in the upswing of a cycle whose turnaround point does not seem to be within the next couple of weeks.

2. Cycle extraction and commonalities

To examine the common cyclical properties of the series, we followed two general steps. First, we filtered away the trends and the high-frequency cycles (e.g. weekly fluctuations in the numbers) to solely focus on the cycles that repeat every couple of weeks and months. Second, we extracted the elements that are common to all the filtered series of a sub-sample of 37 countries using principal components analysis.

2.1. Time series filtering

Time series in general can be broken down into different constituent

* Corresponding author. 1120 15th Street BI-2144, Augusta, GA, 30912, USA.
E-mail address: eriverosperez@augusta.edu (E. Riveros-Perez).

https://doi.org/10.1016/j.amsu.2021.102437
Received 6 April 2021; Received in revised form 21 May 2021; Accepted 23 May 2021
Available online 1 June 2021

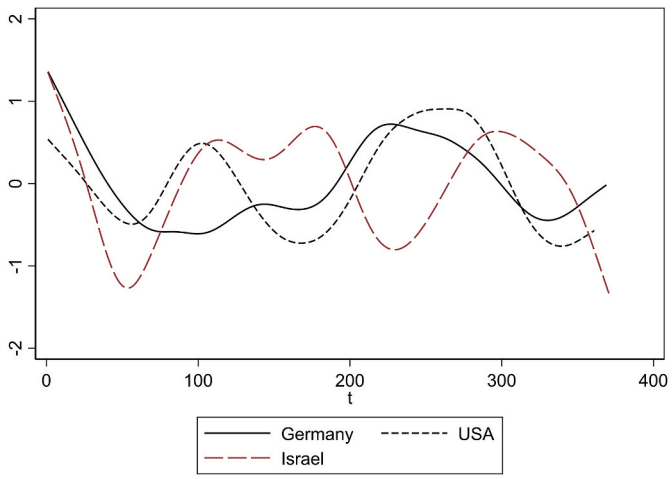

Fig. 1. New cases per 1 Million (Cyclical Component) in United States, Germany, and Israel.

factors: trend, cyclical components (weekly, monthly, and other periodicities), and an irregular component [5]. By using time series filters it is possible to separate components from each other and focus on the periodicities of interest. To filter the range of frequencies for the specific case of COVID-19 incidence series across countries, we used the popular Hodrick-Prescot filter in its double application [6].

We illustrate the procedure using the German series as an example (Fig. 2). The first application of the HP-filter takes away the cycles that repeat very often (high frequency) and leaves away everything else (curve HP-1 in Fig. 2A). The second application leaves only the very low frequencies (trend, curve HP-2 in Fig. 2A). By subtracting HP-2 from HP-1 we get the frequencies that are just between the very frequent periodicities and the trend (black dashed line in Fig. 2B).

The underlying series for the calculation of the cyclical components are the standardized logarithms of the weekly new cases per 1 Million inhabitants. The series start at a high level because we used data starting at the peak of the first wave. The reason for omitting the initial observations was that in most countries the test capacity increased at the beginning of the spread of the virus, leading to overestimation of incidence.

2.2. Common periodicities

We use the resulting filtered curves in the second step to examine common regularities at different periodicities. A straightforward way of examining commonalities consists of using a principal components analysis (PCA). The idea of PCA in a nutshell is to reduce the dimensionality of a dataset with multiple variables by identifying a smaller number of independent variables which capture the information of the dataset by summarizing the common patterns and therefore the variation of the whole dataset. In our specific case, the different variables are the cyclical components of the incidences of each country. We implemented a PCA over a sub-sample of 37 countries for which we have data corresponding to at least 358 days since the peak of the first wave. Data are more or less reliable and the testing capacity is accurate (according to the World Health Organization (WHO) criteria of 10–30 tests per confirmed case) [7]. These criteria leave countries like Cuba, Venezuela and Iran out of the sample. The countries included were Austria, Belgium, Bosnia and Herzegovina, China, Costa Rica, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Kosovo, Latvia, Macedonia, Malaysia, Netherlands, Norway, Portugal, Romania, Singapore, Slovenia, South Korea, Spain, Switzerland, Thailand, Turkey, United

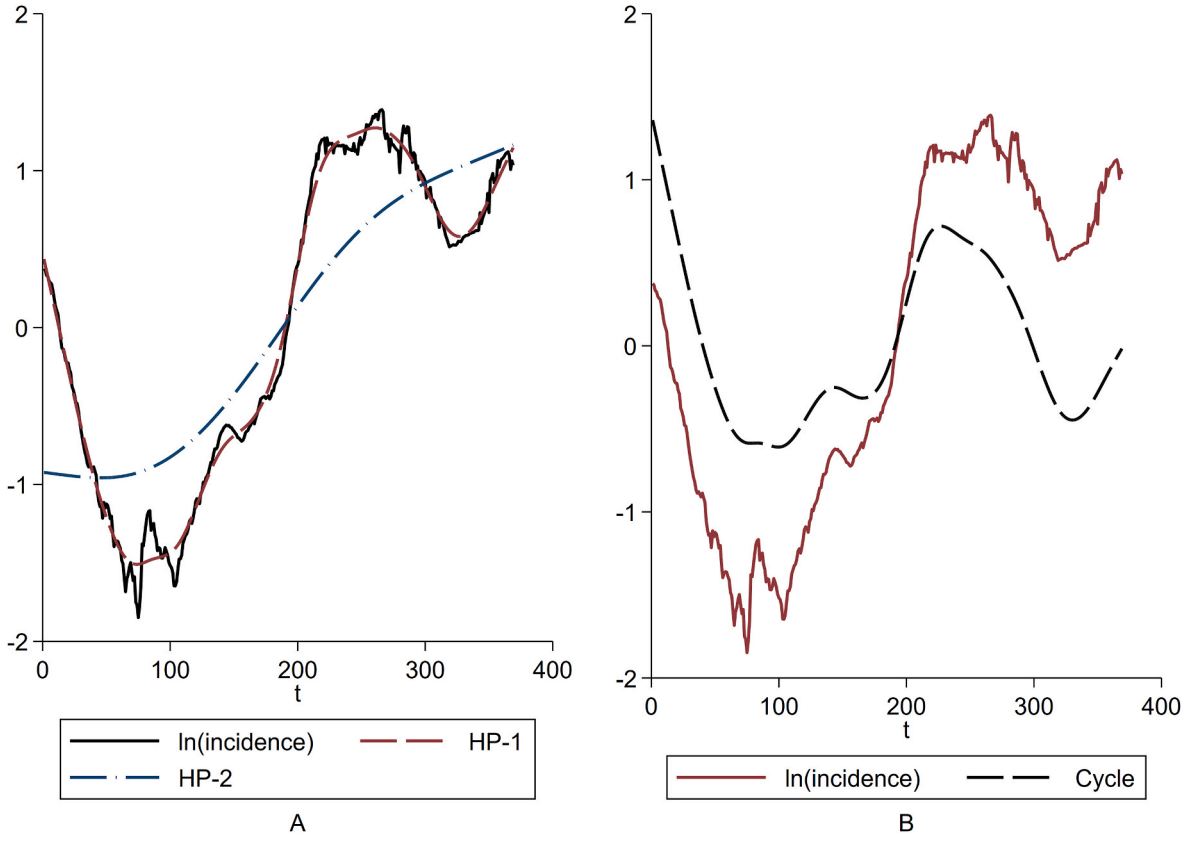

Fig. 2. Germany: Filtering the Incidence Curve. (A) First and second application of HP filter. (B) Original and resulting (filtered) incidence series. HP, Hodrick-Prescott.

States, Uruguay, and Uzbekistan. The time span goes from the peak of the first wave in each country until April 6th, 2021.

The main result from the principal components analysis is that 96.7% of the cyclical variation of 37 incidence series can be summarized in six variables or principal components. Four principal components contain over 90% of the total variation. This means that the infection dynamics as well as the social and political reactions over countries have important commonalities, such that they can be reduced to a handful of variables.

Fig. 3 shows the first 4 principal components calculated out of the 37 filtered series. Each component exhibits a different cycle length and trajectory. The first component (black line in Fig. 3), which explains 50% of the variation, shows a declining pattern until period 80 and afterwards a steady increase at a slower rate compared to the decrease until period 220. Afterwards, a new declining phase starts until it reaches a new through roughly 100 periods later. Since we do not have longer underlying time series, we cannot know with certainty for how long this cycle will expand or when a new turning point could be achieved. The second component (red line in Fig. 3) explains an additional 15% of the variation. This component suggests a lag length of approximately 60 days. The cycle length of the third component is unclear as a turnaround after period 300 is not foreseeable yet. The fourth principal component seems to have a somewhat shorter length as the second and is also entering an increasing phase, as well as the first and second principal components are.

3. Estimating cyclical series

We can use the extracted principal components, which are 358 days long, to estimate the hypothetical path of the cycle curves of countries with shorter incidence series than the ones used to calculate the principal components. Predictions were made using a simple linear regression of the cycle series on the six principal components. Fig. 4 shows the predicted cycles using the 6 principal components which explain 97% of the variation of the 37-country sample. The predicted cyclical series (red dashed line in Fig. 4) fits the official data closer in some countries than in others. The shorter the prediction horizon, such as in the UK or in Bulgaria, the more accurate the prediction was. The discrepancies are more evident when the available data end in or close to a turning point such as in Chile, India, Poland and Russia. Discrepancies can also reflect differences in the data quality as most of the countries with shorter cycles have less testing capacity (e.g. Colombia, Chile) or concerns about the transparency of their data reporting (e.g. Russia).

Fig. 5 shows one-step-ahead predictions of the cycle series for Germany and the United States starting at day 150. In other words, we wanted to answer the question: Were Germany and the US now at day 150 of the pandemic, how well would the principal components (trained model including neither of those two countries) predict the later trajectory of the cycle for both countries? The upward turning point after day 150 was well anticipated by the principal components. The trajectory towards the end of the series is overestimated for Germany and underestimated for the US and the latest turnaround towards a new increasing phase was captured rather accurately.

4. Forecasts of the cycles

Using the same approach to estimate the trajectory of the cycles for Germany and the United States for the upcoming weeks, we analyzed countries that have been at least one week ahead in the number of days since the initial peak. For Germany, Australia, Austria, China, Costa Rica, Italy, Latvia, Norway, South Korea, Thailand, and Uruguay are at least one week ahead. For the United States, the available countries are Australia, Austria, China, Costa Rica, Croatia, Czech Republic, Estonia, France, Germany, Greece, Italy, Latvia, Norway, Portugal, Slovenia, South Korea, Spain, Switzerland, Thailand, and Uruguay.

Fig. 6 shows the estimation of the trajectory one week from day 358

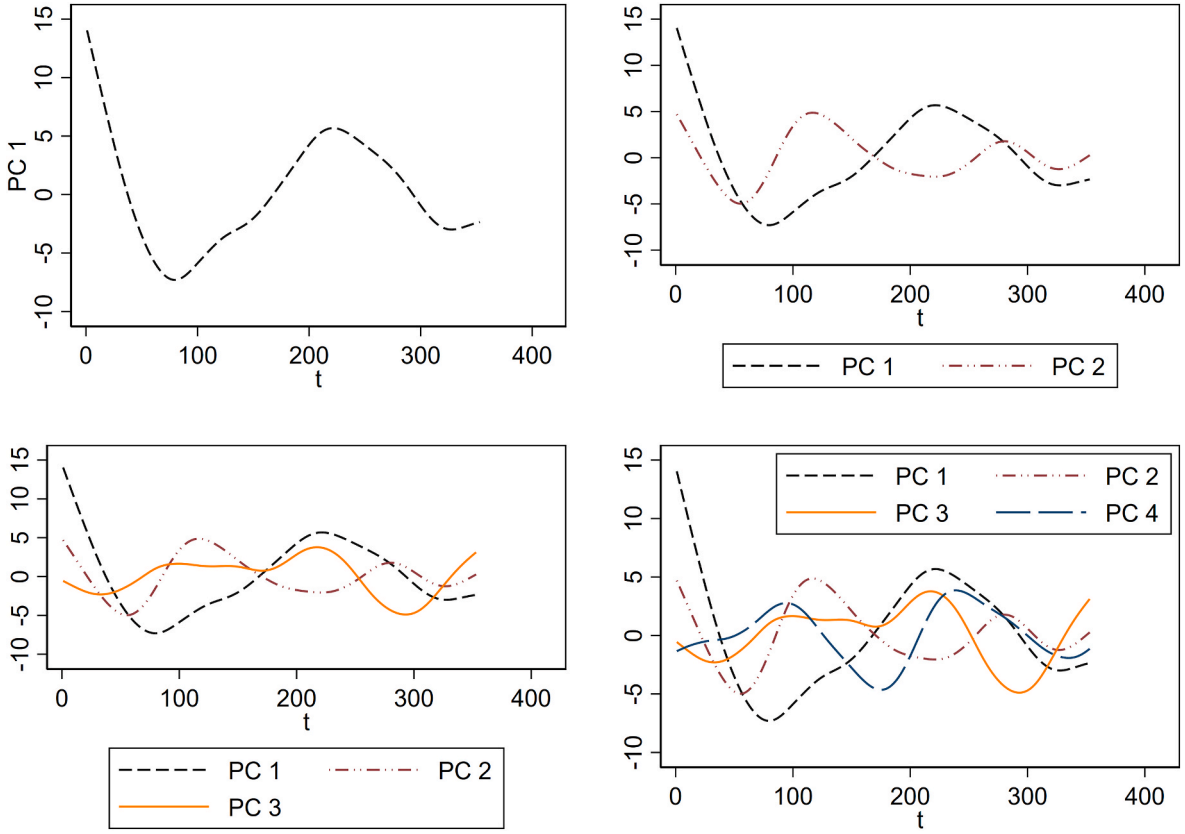

Fig. 3. Principal Components explaining 93% of variability. Each component exhibits a different cycle length and trajectory.

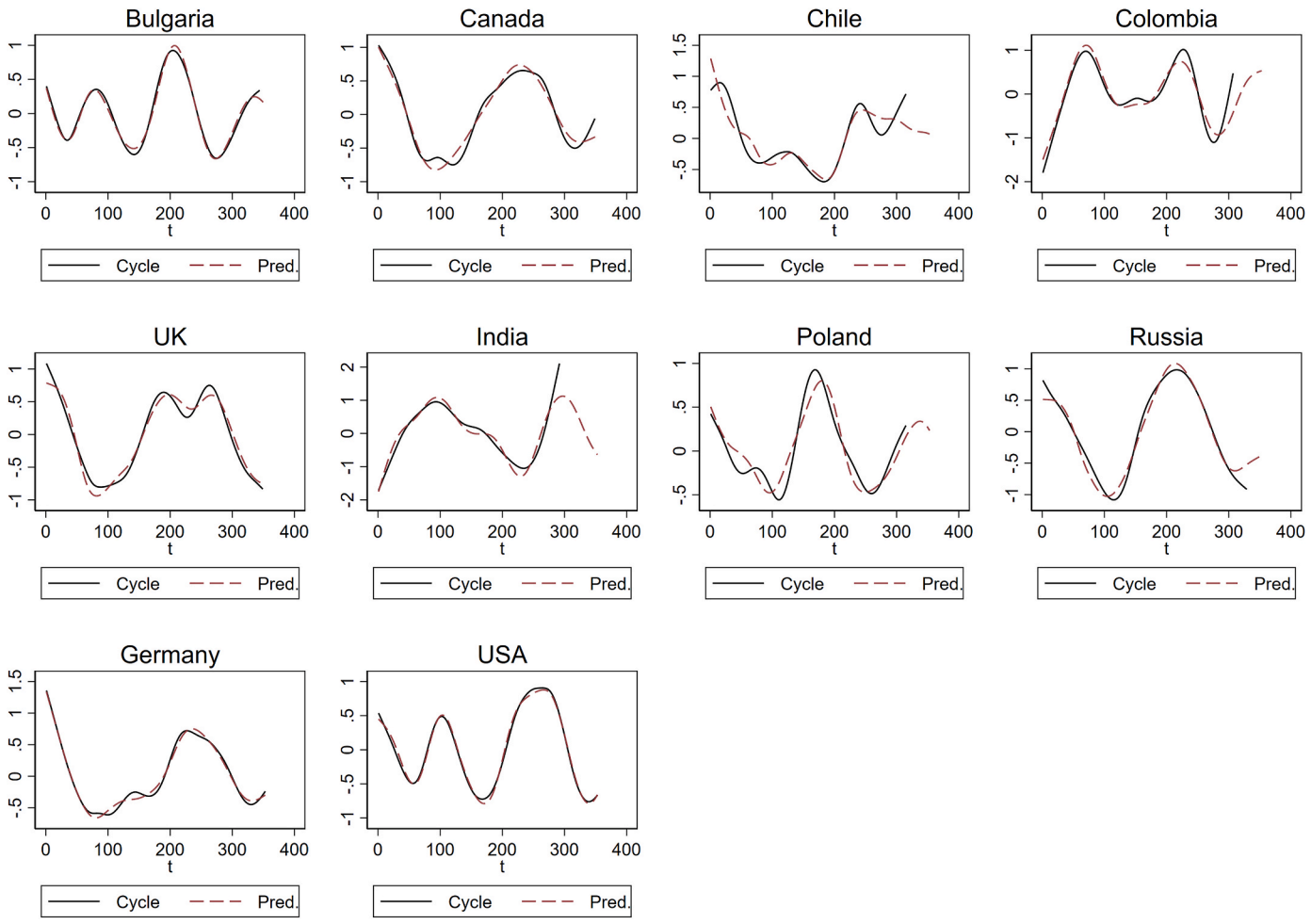

Fig. 4. Predicted Cycles for countries not used to perform the Principal Component Analysis (PCA) in relation to official data.

on. For Germany, the model projects a further upward but less steep trajectory. For the United States, even though the model is less accurate, it predicts a steepening of the cycle series which the official data is not showing at the time of the analysis.

5. Discussion

Modeling of biological phenomena is limited by the presence of randomness and noise. This randomness is the result of incomplete or insufficient knowledge of the nuances of biological variables at a smaller scale [8]. In our case, the effects of individual interactions with SARS CoV-2 are difficult to incorporate into a model based on large populations of different countries. Understanding both static complexities that do not change over time and extrinsic variations imposed over time by changes in biological aspects of the virus (e.g., new variants), the host (e.g., acquired immunity), and the community (e.g., behavior changes), is critical to comprehend patterns of viral spread. For instance, genetic differences leading to heterogeneous susceptibility to the virus, variation in viral replication from host to host, and behavioral and contact differences between individuals have been identified as important factors determining viral transmission within groups of people [9,10].

A significant body of evidence shows that possible seasonal determinants typical of respiratory viruses such as temperature, sunlight, and humidity, as well as host factors (e.g., vitamin status and behavior) contribute to the cyclical pattern of these infections [2,11–15]. Environmental conditions such as dry and unventilated air facilitates transmission of respiratory virus particles [16]. Cyclic tightening and loosening of lockdown mandates or compliance to the rule in different countries may be associated with intermittent exposure to indoor conditions that enhance transmission in patterns compatible to those displayed by the measured and predicted waves presented by our study [17]. On the other hand, despite the generalized agreement on the fact that dry environments occurring during winter season stimulate respiratory viral replication and transmission, Luo et al. challenged this notion by examining province-level variability of the basic reproductive numbers of COVID-19 in China, determining that summer conditions would not protect against viral spread [18]. Wang et al. assessed the impact of humidity and temperature on the transmission of COVID-19 taking into account socioeconomic status, mobility status, and demographics [19]. The authors conclude that changes in humidity and temperature are insufficient to reduce the reproductive viral number. Taken together, these studies underscore the complex contribution of environmental and non-environmental factors to viral spread. Our study shows that the waves are not completely seasonal, and that social and policy factors are playing a significant role in the pattern of infection across communities.

Mathematical models have been used to predict the effect of measures such as social distancing and lockdowns on COVID-19 propagation patterns [20]. However, discrepancy between predictions and actual incidence and patterns of presentation have been consistently identified [21]. Our study tries to add value to the contribution of statistical learning methods to the exploration of possibilities rather than making robust prediction about contagion dynamics in the future. We present actual, fitted, and predictive incidence data. The information contained in our numbers reflects positive testing and not mortality. In contrast with studies using compartments for exposure, we cannot discriminate

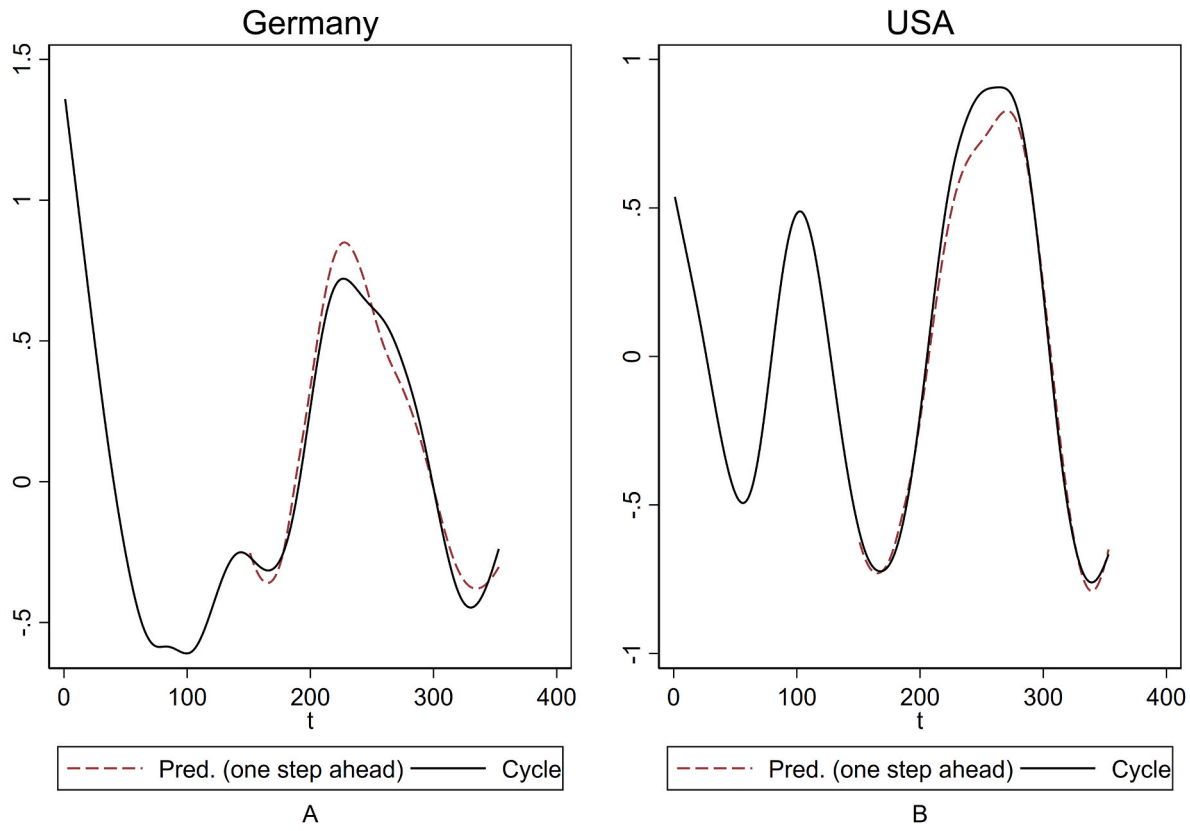

Fig. 5. One-step-ahead cycles Germany (A) and the United States (USA) (B).

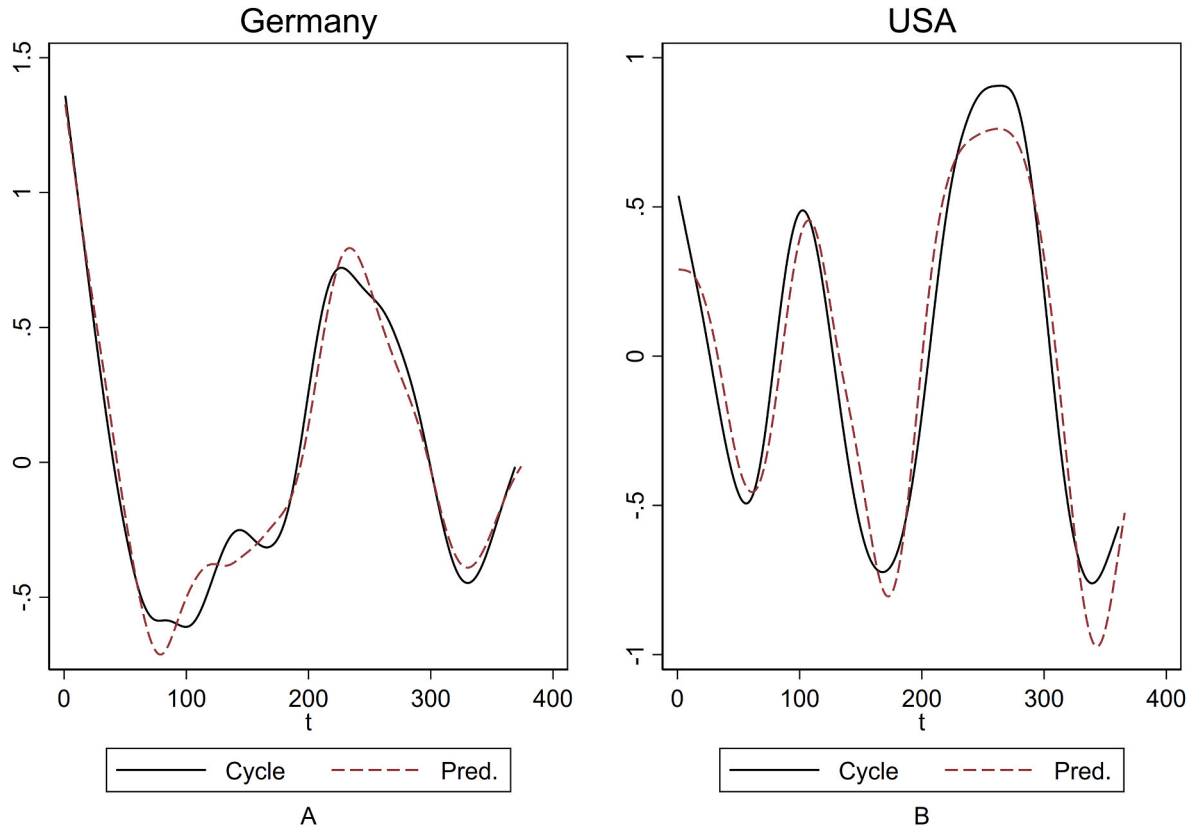

Fig. 6. Trajectory estimation Germany (A) and United States (USA) (B).

cases according to severity. We consider that simplifying information by just displaying incidence trends provides better interpretability to help decision makers incorporate data into their learning processes before coming up with policies [22]. Indeed, most governments closely follow the incidence and the reproduction number R, which is a function of the incidence, for health and economic policy making. The fact that our study includes wave patterns from countries with diverse policies facilitates this reflection process based on feedback provided by studies like ours. The introduction of vaccines is expected to change the progression and epidemiological profile of COVID-19. De Leon et al. presented a model showing the effect of the vaccination program in Israel that covered 80% of the population at the time of the study. The authors report that the shape of the outbreak as measured as new moderate and severe cases has changed, bringing the decline earlier than expected by their prediction model [23]. Fig. 1 also shows the steep decline in the cyclical component for Israel starting around day 300 after the peak of the first wave.

Our study has limitations. We did not use compartments to discriminate between asymptomatic, mild, and severe cases, and we did not analyze mortality rates. Although making predictions based on severity may show a better picture about the virulence of the virus and may help planning for increasing hospital capacity, we believe that our study provides policy makers and the general public with information that describes the overall infection spread and its relationship with control measures in different countries. As most policy makers in general also do not discriminate between cases when deciding on imposing lockdowns restrictions on say schools vs retiring homes, our study is indeed useful to describe the cyclical movements that come from social and political decision making. We argue that the comparison between countries might prove useful to isolate effective policies versus ineffective and even damaging ones. Although some control measures have been universal, regional differences are probably playing a role in differential trend patterns [24]. We also acknowledge the fact that our analysis includes a time frame that may not be representative of the total duration of the pandemic. In this regard, with the emergence of new variants, the predictions of our model may become obsolete [25]. We propose to continue training the model with new observations and reevaluate the prediction accuracy as the pandemic and the effect of new variants evolve.

Future research is necessary to continually evaluate the prediction accuracy of this and other models based on data analysis as the COVID-19 progression is fluid and rapidly changing. Experimental designs evaluating specific control measures in population groups may help elucidate the role of such measures as part of public health policy. Finally, population studies targeting vulnerable populations to characterize their unique epidemiological profile in relation to COVID-19 are warranted. Statistical learning is a powerful tool in assisting analysis of growing data in those population subgroups.

6. Conclusion

The incidence curves of the COVID-19 measured as the number of confirmed new cases per 1 million inhabitants show strong commonalities among countries. After filtering away the high-periodic elements as well as the trends from the incidence curves of 37 countries, 90% of the information in the resulting dataset can be summarized in four variables (principal components). The commonalities are not only related to the periodic nature of viral infections but also that the fact that citizens and governments have reacted to the spread of the virus in a similar fashion. The combination of viral natural history and governmental and individual behavior seem to have so much in common, that the incidence cycles of 37 countries can be reduced to a few principal components. One-step ahead forecasts for Germany and the United States show that the principal components can track the incidence cycles. How well the principal components can predict the trajectories out-of-sample will be evident in the coming weeks and months.

Ethical approval

Mathematical analysis. No IRB approval necessary.

Sources of funding

None

Author contribution

Pablo Duarte: Data collection, analysis and construction of manuscript. Efrain Riveros-Perez: Data collection and analysis. Manuscript construction and final review.

Conflicts of interest

None

Research registration unique identifying number (UIN)

N/A.

Trial registry number – ISRCTN

N/A

Guarantor

Efrain Riveros-Perez

References

[1] N. Noah, Cyclical patterns and predictability in infection, Epidemiol. Infect. 102 (2) (1989) 175–190.
[2] M. Moriyama, W.J. Hugentobler, A. Iwasaki, Seasonality of respiratory viral infections, Annual review of virology 7 (2020) 83–101.
[3] V. Grech, S. Cuschieri, COVID-19: A Global and Continental Overview of the Second Wave and its (Relatively) Attenuated Case Fatality Ratio, Early human development, 2020.
[4] P.C. Wever, L. Van Bergen, Death from 1918 pandemic influenza during the First World War: a perspective from personal and anecdotal evidence, Influenza and other respiratory viruses 8 (5) (2014) 538–546.
[5] A.C. Harvey, Forecasting, Structural Time Series Models and the Kalman Filter, 1990.
[6] R.M. De Jong, N. Sakarya, The econometrics of the Hodrick-Prescott filter, Rev. Econ. Stat. 98 (2) (2016) 310–317.
[7] W.H. Organization, Coronavirus Disease (COVID-19), 2020.
[8] F.S. Heldt, S.Y. Kupke, S. Dorl, U. Reichl, T. Frensing, Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in influenza A virus infection, Nat. Commun. 6 (1) (2015) 1–12.
[9] F. Herrerías-Azcué, T. Galla, The effects of heterogeneity on stochastic cycles in epidemics, Sci. Rep. 7 (1) (2017) 1–14.
[10] C.T. Bauch, J.O. Lloyd-Smith, M.P. Coffee, A.P. Galvani, Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future, Epidemiology (2005) 791–801.
[11] D. Fisman, Seasonality of viral infections: mechanisms and unknowns, Clin. Microbiol. Infect. 18 (10) (2012) 946–954.
[12] S.F. Dowell, Seasonal variation in host susceptibility and cycles of certain infectious diseases, Emerg. Infect. Dis. 7 (3) (2001) 369.
[13] J. Cannell, R. Vieth, J. Umhau, M. Holick, W. Grant, S. Madronich, et al., Epidemic influenza and vitamin D, Epidemiol. Infect. 134 (6) (2006) 1129–1140.
[14] C. Sloan, M.L. Moore, T. Hartert, Impact of pollution, climate, and sociodemographic factors on spatiotemporal dynamics of seasonal respiratory viruses, Clinical and translational science 4 (1) (2011) 48–54.
[15] J. Shaman, M. Kohn, Absolute humidity modulates influenza survival, transmission, and seasonality, Proc. Natl. Acad. Sci. Unit. States Am. 106 (9) (2009) 3243–3248.
[16] R. Velraj, F. Haghighat, The contribution of dry indoor built environment on the spread of Coronavirus: data from various Indian states, Sustainable cities and society 62 (2020) 102371.
[17] M.A. Capistran, A. Capella, J.A. Christen, Forecasting hospital demand in metropolitan areas during the current COVID-19 pandemic and estimates of lockdown-induced 2nd waves, PloS One 16 (1) (2021), e0245669.
[18] W. Luo, M. Majumder, D. Liu, C. Poirier, K. Mandl, M. Lipsitch, et al., The Role of Absolute Humidity on Transmission Rates of the COVID-19 Outbreak, 2020.

[19] J. Wang, K. Tang, K. Feng, W. Lv, High temperature and high humidity reduce the transmission of COVID-19, 2020. Available at SSRN 3551767.
[20] R. Sameni, Mathematical Modeling of Epidemic Diseases; a Case Study of the COVID-19 Coronavirus, 2020 arXiv preprint arXiv:200311371.
[21] J. Guan, Y. Wei, Y. Zhao, F. Chen, Modeling the transmission dynamics of COVID-19 epidemic: a systematic review, Journal of Biomedical Research 34 (6) (2020) 422.
[22] I. Holmdahl, C. Buckee, Wrong but useful—what covid-19 epidemiologic models can and cannot tell us, N. Engl. J. Med. 383 (4) (2020) 303–305.
[23] H. De-Leon, R. Calderon-Margalit, F. Pederiva, Y. Ashkenazy, D. Gazit, First Indication of the Effect of COVID-19 Vaccinations on the Course of the COVID-19 Outbreak in Israel, medRxiv, 2021.
[24] C.-C. Lai, C.-Y. Wang, Y.-H. Wang, S.-C. Hsueh, W.-C. Ko, P.-R. Hsueh, Global epidemiology of coronavirus disease 2019 (COVID-19): disease incidence, daily cumulative index, mortality, and their association with country healthcare resources and economic status, Int. J. Antimicrob. Agents 55 (4) (2020) 105946.
[25] N.G. Davies, S. Abbott, R.C. Barnard, C.I. Jarvis, A.J. Kucharski, J.D. Munday, et al., Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England, Science (2021).