Proceedings of the IISE Annual Conference & Expo 2023
K. Babski-Reeves, B. Eksioglu, D. Hampton, eds.

Monitoring and Forecasting the COVID-19 Pandemic in Peru

Abstract ID: 3622

Jacob Unterbrink[1], Charles Nicholson[1], Talayeh Razzaghi[1], Andrés D. González[1], Yuber Velazco-Paredes[2] and Brayan Alexander Lipe Huamani[2], (1) University of Oklahoma, Norman, OK, (2) Universidad Nacional de San Agustin de Arequipa, Perú

Abstract

The COVID-19 pandemic has been one of the biggest health challenges to many countries around the world in recent history. In particular, Peru has the world's highest rate of mortality over the last two years. Predicting the number of cases and deaths would assist policy makers and healthcare providers to better manage their limited resources and reduce health-related costs. In this paper, the COVID-19 case and death patterns in Peru are modeled using epidemiological and time series forecasting techniques. In particular, the results of SEIR models will be compared with time series forecasting models, and the inflection points are identified, and auto- and cross-correlation analysis are performed using Peru data.

Keywords
COVID-19, Forecasting, Pandemic modeling, SEIR, Time series analysis

1. Introduction
The COVID-19 pandemic has shaken the global economy and the lives of people everywhere. The need to understand this disease is paramount, not just for a return to normal but to save lives. However, the race for answers is not always easy. The emergence of new variants has muddled the data and made it difficult to distinguish the signal between these different strains. Efforts to unravel this issue are made more complex when attempting to account for underreported or misreported cases and COVID-related deaths. This is especially prevalent in regions whose healthcare systems have struggled to modernize – regions who need help the most – since many of their reporting methods and medical systems were overwhelmed by the volume of cases. The country of Peru provides a critical important scenario to study. By late 2021, Peru had the world's highest COVID-related death rate: 5.89 fatalities per 1,000 residents. For perspective, this is more than twice the death rate experienced in the United States, which suffered 2.38 fatalities per 1,000 residents. The COVID-19 fatality rate far exceeded the rates reported by other South American countries, e.g., Brazil (2.86), Argentina (2.54), and Colombia (2.48), during this same time-period [1]. The fatality rate across Peru varies greatly by region. The Arequipa region in southwestern Peru experienced a fatality rate that exceeded the national average.

By rigorously exploring this disease and its propagation dynamics, countries can be better prepared should a similar infection arise in the future. It may be possible to identify key signs of an emerging outbreak before it sweeps the globe, or even predict the onset of a new strain. Moreover, by fortifying struggling medical systems across the world thousands of lives can be saved, and the positive externalities associated with better treatment might serve to better the lives of thousands more.

2. Problem Description
The Arequipa region of Peru has suffered greatly from the impact of COVID-19. At the pandemic's onset, medical facilities were not equipped to cope with the surge in the demand for appropriate supplies, services, equipment, and space to effectively care for the large number of infected individuals. For example, the lack of oxygen supply and limited ICU bed capacity made the management of COVID-19 difficult. This is despite early adoption of strict preventative measures such as quarantines, social distancing, and masking (measures that are still in place to this day). Compounding this challenge is the lack of a link between university research centers and the government -- there is very little connection or collaboration between the institutions. Research generated within the academic sector is rarely considered or acted upon by the government authorities. This is in part why robust models for this region remain limited. If these models could be developed, policy makers would have an edge over COVID and future pandemics,

being able to identify, weeks in advance, when a new wave of infections might surface. Such foresight can improve decisions on resource allocation and mitigation strategies.

3. Related Research

Various research institutions have generated models to track the spread of COVID-19 over time. Some publicly available tools were developed by Stanford [2], and the US Center for Disease Control and Prevention [3]. The quantity and variety of these models continues to grow as accurate predictions prove to be a challenge. In general, these models fall into one of two categories: data-driven models or epidemiological models.

3.1. Data-Driven Models

Models falling under this framework typically outperform their epidemiological counterparts when it comes to raw predictive power. These might include agent-based simulations, or machine learning methods [4]. However, this comes at the cost of fitting propagation mechanisms directly.

3.2. Epidemiological Models

The chief epidemiological modeling framework involves compartmental models. These were first described by Kermak and McKendrick between 1920 and 1940 [5]–[7]. The idea being that a population can be divided into various compartments and flow from one compartment to another based on a set of parameters.

The most basic of these is known as the Susceptible-Infected-Recovered (SIR) model, where the population flows from the first compartment, susceptible (S), to the infected compartment (I), and finally to the recovered compartment (R). This framework can be expanded to account for any number of compartments and parameter values. These types of models are useful to understand the mechanisms underlying disease propagation, but ignore factors like geography, population heterogeneity, etc.

3.3. Modeling of COVID-19 in Diverse Geographical Locations

Various modeling approaches have been proposed to different regions around the globe due to different socio-economical, geographical, environmental, and political characteristics. For instance, deep learning methods have shown high predictive performance in the countries of Brazil, India, and Russia when forecasting cases is the primary concern [8]. While another work has demonstrated the value of incorporating a region's sociocultural context and nuance using score-driven models on data from nine Latin American countries [9]. Therefore, these studies reveal that it is beneficial to dissect countries and regions distinctly when forecasting COVID-19. This is because the propagation mechanisms can vary widely from one region to another region due to local contextual factors. For instance, in Russia, there are, on average, less than three people per household which is much less than what is observed in Pakistan (more than six people per household) [10]. This is an important factor since more populated households are more likely to encounter and spread COVID-19. Similar considerations may include governmental policies associated with mask mandates, social distancing, lockdowns, and vaccination status, as well as a population's general health, geography, and many more [11], [12], [13] .

4. Data and Methodology

4.1. Data

Regional-level data were gathered from Peruvian government agencies and then were cleaned. The data includes information on ICU beds in use/available, number of hospitalized individuals, confirmed cases, deaths, recovered cases, and negative test results. Each row of data represents an observation at a specified day starting from March 13th, 2020, through April 5th, 2022. Two columns had missing data: 15.8% from ICU beds in use, and 50.9% from beds available.

4.2. SEIR Model

To track the spread of the disease, a Susceptible-Exposed-Infected-Recovered (SEIR) model was implemented which is an extension of the SIR model that allows researchers to observe the number of susceptible (S), exposed (E), infected (I), and recovered (R) individuals in a population and also allows for recovered individuals to return to susceptibility (S). As in the SIR model, all compartments are initially empty, save for susceptible and infected. The value for susceptible is set to one minus the population total and the value for infected is set to one. The analysis of the model starts when one infected person is introduced to the remainder of the population.

Parameters need to be established to determine how many people flow from one compartment to another every day. In the SEIR model, these parameters are denoted by λ (the rate at which people move from susceptible to exposed), f (rate from exposed to infected), r (rate from infected to recovered) and w (rate from recovered to susceptible). However, the model has been updated to account for the vaccination rate (v), the natural birth/death rate (μ), and the COVID death rate (d). Vaccinated individuals move immediately from the susceptible compartment to recovered, while birthed individuals are added to the susceptible compartment. People can become deceased naturally at any point throughout the modelling process, and therefore people are removed from each compartment at rate μ (the model assumes an equal and constant birth/death rate). To compute COVID-related deaths, d is multiplied by the number of new infections at each timestamp, t.

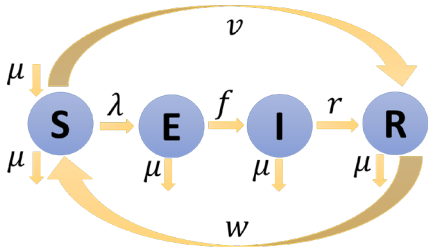

Figure 1: Visual representation of a SEIR model framework, including compartments and parameters.

The parameter λ is a function of the basic reproductive number – a disease's measure of contagion – as well as population size, duration of infection, and infected persons at time t. Other starting parameter values were estimated until they reliably fit the data. To begin the model, the number of susceptible individuals was set to 1,382,999, while the number of infected persons was set to one. Initial parameter values are as follows: 4.3569×10^{-14} for λ, 0.5000 for f, 0.1667 for r, 0.0022 for w, 0.0002 for v, 3.6530×10^{-5} for μ, and 0.0100 for d.

Most parameter values were manually updated over time. This occurred when model output deviated significantly from observed cases. Optimization of these values is a subject of future work. Root Mean Square Error (RMSE) was chosen as the error metric, and the daily RMSE was computed to track model performance over time.

4.3. Determining Inflection Points

To establish inflection points among new infections and deaths, first a seven-day rolling average was determined as a means of smoothing the curves somewhat while retaining the overall trends. From this, inflection points were identified by comparing each day with the previous days. For each day, if the value of new infections or deaths changed by more than 25 (for new infections) or two (for deaths) from the previous day, and the slope sign changed, then this day was marked as an inflection. The choice of 25 for infected cases and two for deaths is subjective, but doing so helped to ignore small bumps along the curves and isolated the more impactful points.

4.4. Time Series Modeling

Because the data were recorded at daily intervals, the features lend themselves perfectly to time series modeling. This type of approach allows for the quantitative understanding of data trends over time. Specifically, it can isolate systematic patterns such as seasonality – recurring yearly fluctuations – in the data and greatly improve forecasting abilities. Autocorrelation measures similarity between two or more observations with a time lag while cross-correlation measures the movement of two or more time series relative to one another. In this work, these two measures are computed and provide a strong starting point for pattern analysis.

5. Results

5.1. SEIR Model Analysis

One of the primary goals of the SEIR model was to forecast new infections, which was accomplished with some level of success (performance shown in Figure 2 below). While the model tends to overestimate the true number of infected individuals, it does serve to capture the overarching trends. From the daily RMSE curve, one can observe that the largest deviations were between August and December of 2020, as well as between January and June 2022. Both these periods followed an unusually large spike in COVID cases, and by visual inspection, it becomes apparent that once predicted values quickly turned upwards in value they struggle to come back down. When predicted values rise slowly (as they did between January and June 2021) they overestimate the true values by a much smaller margin. The RMSE for new daily infections was found to be 2,102.71.

The model struggled to predict deaths well. This was especially challenging for cases falling between April and August of 2021. This might partially be because the death multiplier (d) applied to new infections remained constant. Although, it does a decent job following the trend set over the first few months' worth of data and provides a baseline for future work. The RMSE for deaths was found to be 23.55.

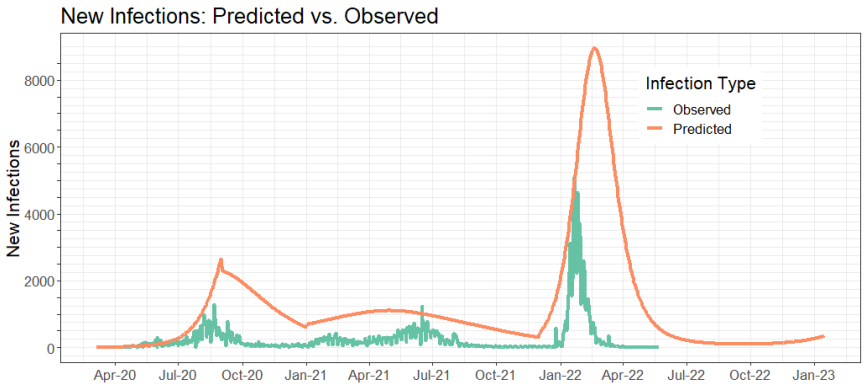

Figure 2: SEIR model results plotted against observed infection data

5.2. Inflection Points

Figure 3 shows that 20 inflection points were found for COVID-related deaths. Performing this analysis for new infections yielded 33 points. Many of these inflections were clustered around one another, giving a good indication of important date ranges to focus on. Some of the most prominent clusters resulting from new infections centered around late July, mid-June, and late December. For deaths, groups formed around late July, May, and late January.

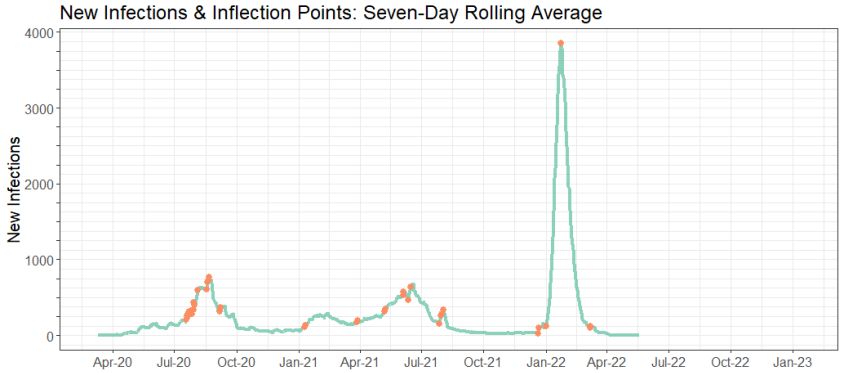

Figure 3: Inflection points for COVID-related deaths

5.3. Time Series Analysis

Autocorrelations for both cases and deaths are illustrated in Figure 4. It can immediately be seen that correlation values are far from zero, and therefore these features are not random and exhibit correlations with their lags. Moreover, both figures demonstrate seasonality and trend, oscillating and trending downward. The existence of these characteristics can aid in the building of future models, as it is now shown that predictable changes are exhibited at regular intervals.

Periodicities exist when investigating the cross-correlation between cases and deaths. Moreover, cross-correlation peaks at offsets 14. This finding is especially valuable in determining how long COVID's fatal effects take to become realized and can be used to tune models going forward.

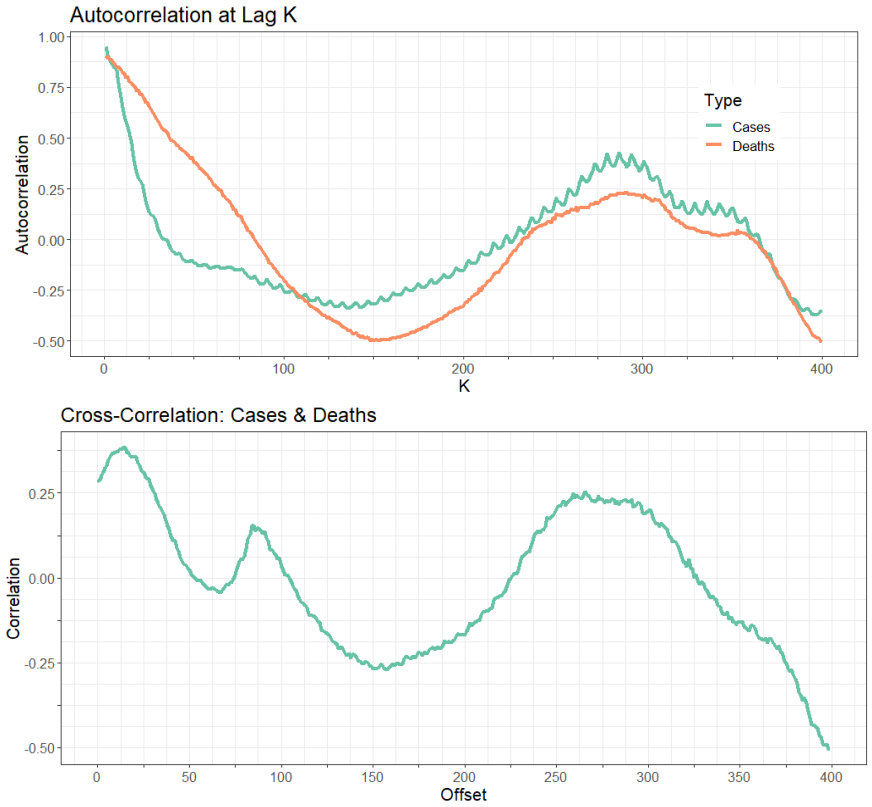

Figure 4: Autocorrelation (top) and cross-correlation (bottom) of cases and deaths

5.3.1. Time Series Forecasting

We further employed neural network autoregression (NNAR) to predict the number of COVID deaths and compared it with the existing time series models such as Exponential smoothing (ETS) with additive trend (Holt's linear trend), additive damped trend, and seasonality (Holt-Winters' additive) [12], AutoRegressive Integrated Moving Average (ARIMA) [13], mean, random walk, and drift methods [14]. All death data prior to March 15th, 2022, was used for training and the next three weeks starting from March 16th to April 5th, 2022, were selected as test set which reflects the onset of a new wave of infections. The results of forecasting models on test data are summarized in Table 1 in terms of RMSE and mean absolute error (MAE). The best results are obtained for NNAR with 4.33 RMSE and 3.36 MAE.

Table 1: Computational results of forecasting models

	NNAR	**Holt's Linear Trend**	**ARIMA**	**ETS (Damped Trend)**	**Holt-Winters' Additive**	**Mean**	**Random Walk**	**Drift**
RMSE	**4.33**	8.73	8.77	8.72	8.78	5.13	9.31	9.29
MAE	**3.36**	7.33	7.37	7.32	7.38	4.51	7.93	7.91

6. Conclusions

The purpose of this study was to develop tools and insights that would better help Arequipa policymakers curb the effects of COVID in their region. To this end, a strong foundation has been laid. A SEIR model was developed to predict new infections and deaths, and although minimal parameters were included it already captures the overarching trends found in infections and can be utilized when trying to forecast outbreaks in the future. Dates of inflection were investigated and isolated for both infections and deaths, providing a starting point for a deeper dive into the causes of outbreaks in this region moving forward. Finally, time series methods were applied to get auto- and cross-correlation among cases and deaths. This revealed COVID's seasonality and trend in the Arequipa region, which will greatly assist in future modeling attempts. Cross-correlation also allowed for the identification of highly-correlated lags between COVID cases and deaths, another discovery that can be incorporated into future work.

Moving forward, more parameters need to be accounted for and optimized to assist the SEIR model in predicting cases and deaths to improve accuracy of forecasts. Research also needs to be done to investigate critical inflection points, weather impacts, social patterns, and other factors which may influence disease propagation. Lastly, the relationship between reported cases, hospitalization, and deaths needs to be explored and understood further. Through this work, policymakers will have more tools at their disposal to make informed decisions and curb the effects of COVID or similar infectious diseases in the future.

Acknowledgements

This work was funded by the Universidad Nacional de San Agustin (UNSA), Peru, through the University of Oklahoma's IREES/LASI, and the Global Change and Human Health Institute.

References

[1] NPR (2021, November 27), "The highest COVID death rate in the world is in Peru. How did that happen? Goats and Soda." https://www.npr.org/sections/goatsandsoda/2021/11/27/1057387896/peru-has-the-worlds-highest-covid-death-rate-heres-why (accessed Feb. 18, 2023).

[2] J. O. Ferstad, Bergquist, T., Larsson, M., and Lindskog, P., "A model to forecast regional demand for COVID-19 related hospital beds," Health Informatics, preprint, Mar. 2020.

[3] Center for Disease Control and Prevention, "FluSurge 2.0, Pandemic Influenza (Flu)." https://www.cdc.gov/flu/pandemic-resources/tools/flusurge.htm (accessed Feb. 18, 2023).

[4] P. O. Siebers, C. M. Macal, J. Garnett, D. Buxton, and M. Pidd, "Discrete-event simulation is dead, long live agent-based simulation!" *Journal of Simulation*, vol. 4, no. 3, pp. 204–210, Sep. 2010.

[5] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927.

[6] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics. II.—The problem of endemicity," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 138, no. 834, pp. 55–83, 1932.

[7] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics. III.—Further studies of the problem of endemicity," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 141, no. 843, pp. 94–122, 1933.

[8] L. Xu, R. Magar, and A. B. Farimani, "Forecasting COVID-19 new cases using deep learning methods," *Computers in Biology and Medicine*, vol. 144, p. 105342, 2022.

[9] S. Contreras-Espinoza, F. Novoa-Muñoz, S. Blazsek, P. Vidal, and C. Caamaño-Carrillo, "COVID-19 Active Case Forecasts in Latin American Countries Using Score-Driven Models," *Mathematics*, vol. 11, no. 1, p. 136, 2023.

[10] United Nations, "Household size and composition around the world, 2017," https://www.un.org/en/development/desa/population/publications/pdf/ageing/household_size_and_composition_around_the_world_2017_data_booklet.pdf (accessed March 29, 2023).

[11] C. Nicholson, L. Beattie, M. Beattie, T. Razzaghi, and S. Chen, "A machine learning and clustering-based approach for county-level COVID-19 analysis," *Plos One*, vol. 17, no. 4, p. e0267558, 2022.

[12] M. C. Lucic, H. Ghazzai, C. Lipizzi, and Y. Massoud, "Integrating county-level socioeconomic data for COVID-19 forecasting in the United States," *IEEE Open Journal of Engineering Medicine and Biology*, vol. 2, pp. 235–248, 2021.

[13] O. I. Krivorot'ko, S. I. Kabanikhin, N. Y. Zyat'kov, A. Y. Prikhod'ko, N. M. Prokhoshin, and M. A. Shishlenin, "Mathematical modeling and forecasting of COVID-19 in Moscow and Novosibirsk region," *Numerical Analysis and Applications*, vol. 13, pp. 332–348, 2020.

[14] E. S. Gardner, "Exponential smoothing: The state of the art—Part II," *Int. J. Forecast.*, vol. 22, no. 4, pp. 637–666, Oct. 2006.

[15] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The **forecast** Package for *R*," *Journal of Statistical Software*, vol. 27, no. 3, 2008.

[16] J. K. Ord, R. Fildes, and N. Kourentzes, *Principles of business forecasting*, 2nd edition. New York, NY: Wessex Press, Inc, 2017.