Estimating the number of deaths due to COVID-19 in Lima and Peru during March and April 2020 using ARIMA time series and modeling

Eduardo Villarreyes (eduardo.villarreyes@unmsm.edu.pe)
Universidad Nacional Mayor de San Marcos https://orcid.org/0000-0001-7364-5342

Ana Luna
Universidad del Pacifico

Andres Soriano
Universidad Nacional Mayor de San Marcos

Research article

Keywords: Covid-19, Mortality, Peru, Inference, ARIMA

Posted Date: September 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-62317/v1

Estimating the number of deaths due to COVID-19 in Lima and Peru during March and April 2020 using ARIMA time series and modeling

Eduardo Villarreyes [1, 2], Ana Luna [2], Andrés Soriano [3]

Abstract

Background: The SARS-CoV-2 virus which causes the COVID-19 disease is a large family of viruses that cause respiratory diseases and complications in human beings. Nowadays, the pandemic is studied by the rate of infection and mortality in different countries. Since there are discrepancies regarding the number of deaths due to the pandemic, our main goal is to estimate the number of deaths in Lima and Peru due to COVID-19.

Methods: By using inference modeling techniques and statistical analysis in time series together with ARIMA-type predictions, it is possible to estimate the number of deaths caused by COVID-19. Our study took place in the city of Lima and Peru and our detailed analysis was carried out during March and April of 2020.

Results: By comparing the death toll provided by the Ministry of Health (MINSA), we have obtained approximately a difference of 325.9% regarding the number of deaths due to COVID-19 in the city of Lima and a difference of 185.9% regarding the number of deaths in Peru.

Conclusions: ARIMA time series modeling are a powerful statistical tool that predict and forecast data that are widely used in various fields of science, health, and economics. In this study, we have shown the discrepancy between the data reported by MINSA and the projection obtained in our ARIMA time series modeling. Therefore, we have a confidence level of 95% about the study that was carried out in March and April of 2020.

Keywords: Covid-19, Mortality, Peru, Inference, ARIMA.

Background

The first case of coronavirus was reported in Wuhan, China in December 2019 and was characterized as a pandemic on February 26th, 2020 in different countries such as Italy, Spain, Iran, Korea, EE. UU., etc. In Peru, the first coronavirus case was reported on March 6th, 2020 [1, 2]. After the government published in *El Peruano* newspaper, the Supreme Decree N° 044-2020-PCM on March 15th, 2020, the president declared a nationwide state of emergency to reduce rigorously the spread of the virus. On May 1st, 2020 at 0:00 a. m., Peru has 40,459 infected people and a total of 1,124 deaths [3].

Corresponding author.
E-mail addresses: eduardo.villarreyes@unmsm.edu.pe (E. Villarreyes), a.soriano@unmsm.edu.pe (A. Soriano), ae.lunaa@up.edu.pe (A. Luna).

Several studies show and analyze the differences between the official datum reported by the government and the real datum within the field of public health [4, 5, 6, 7, 8, 9]. The difference between these two data was reported through different newspapers, not only in Peru but also in other 25 countries where at least more than 87,000 people have died during the coronavirus pandemic compared to the official datum [10]. In Peru, the minister of MINSA has acknowledged the existence of underreporting cases in all diseases and not only in the epidemic [11].

The difference between the excess of deaths concerning the average of previous years and the number of deaths officially reported due to COVID-19 reaches, in some cases, a significant difference of 80% between March and April in 2020 [11, 12]. Therefore, the real picture and the data would be incomplete.

An alternative to reduce this difference is the use of an Autoregressive Integrated Moving Average (ARIMA) model presented by Box and Jenkins in 1976. One of the advantages that show, due to its benefits to adequately model the behavior of health events, is its growing and international use in the area of public health [13]. In Ref. [14] several models are studied and compared. The author concludes that ARIMA modeling is the best univariate model to predict the number of infant deaths caused by Acute Respiratory Infections (ARI). The effectiveness of ARIMA modeling was also reflected in a study whose aim was to predict cancer mortality in Spain [15]. Therefore, a correct implementation of the model allows adequate inferences to be made about unknown or unexplored phenomena in the field of biomedical science [16]. The authors of the research [17] highlight the predictive performance and the certainty in their prediction periods of ARIMA models with a seasonal component to be used as a management tool for the diverse queries.

In this study, we propose the use of official data obtained from the National Death Registry Information System (SINADEF) regarding deaths in Peru during the last three years [18] and the official information of deaths due to coronavirus by the Ministry of Health (MINSA) during March and April in 2020 [3, 19]. Consequently, we will combine the techniques of time series analysis and ARIMA modeling.

Thus far, none of the research studies in Peru have used the set of recently mentioned models regarding the predictions of the death toll due to COVID-19. This research aims to present and validate a rigorous method to estimate the number of deaths in Lima and Peru as a consequence of the pandemic.

Methods

Based on the information from the National Death Registry Information System (SINADEF), we have obtained the number of deaths in Lima and throughout Peru in the last three years [18]. Consequently, we have made several time series charts regarding the deaths that occurred between the years 2017 and 2020. In Fig. 1 and Fig. 2, we observe a notable increase number of deaths during April 2020. The trend in the city of Lima and Peru is the same. In the same atypical period, we compared the arithmetic mean of deaths reported in Lima and throughout Peru with the number obtained from previous years. During March 2020, the difference reached 222 deaths in Peru and 1006 deaths for the particular case of Lima. In April, the difference increased to 3,763 deaths in Peru and 3,202 in Lima. When calculating the standard deviation and the coefficient of variation (C.V) for the city of Lima, the values obtained were 525.38 and 20.09%, respectively for March and 1616.56 and 67.29% for April.

Thus, homogeneous data has become heterogeneous and shows that the arithmetic mean value for this last month is not reliable to be such an analysis tool (C.V > 30%).

Likewise, in the case of Peru, there is a standard deviation value and a coefficient variation of 599.10 and 6.57% for March, rising in April to 1958 and 23.28% respectively (Fig. 3 and Fig. 4).

Fig. 1. Deaths that occurred in Peru from January 2017 to April 2020. An atypical upward trend is observed in April 2020.

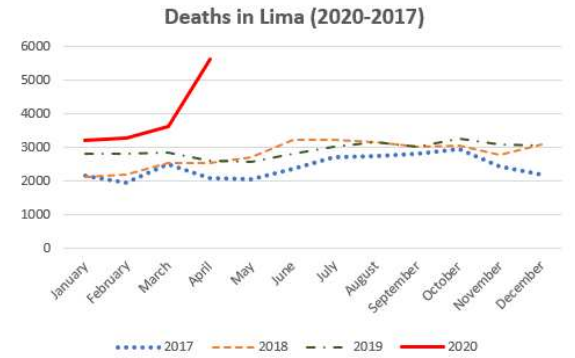

Fig. 2. Deaths in Lima from January 2017 to April 2020. The growing trend of deaths in April 2020 is very marked.

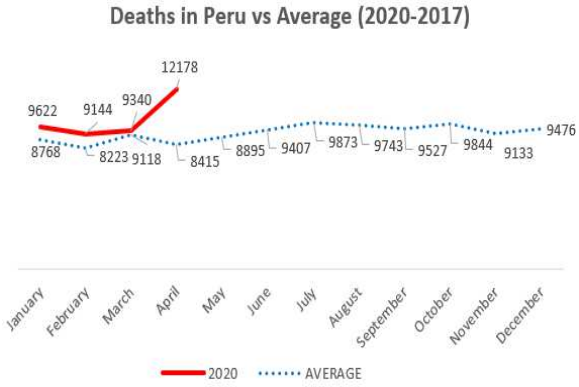

Fig. 3. The deaths that occurred in Peru in 2020 compared to the average in the years 2017-2019. There is a marked difference between the average and those registered in April 2020.

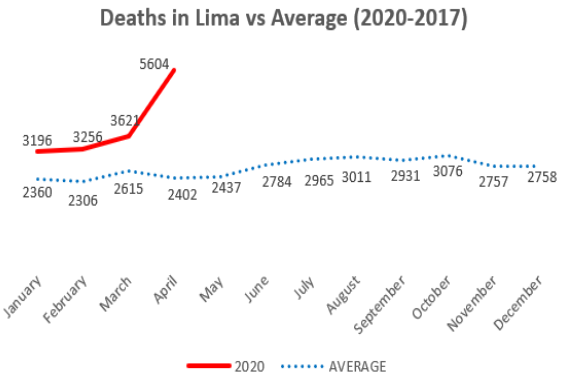

Fig. 4. Deaths that occurred in Lima in 2020 compared to the average in the years 2017-2019. There is an appreciable difference between January to April according to the average.

Based on the time series chart, the data of the total number of deaths that occurred in Lima and Peru from the beginning of January 2017 to April 2020 show the existence of stationarity. In Fig. 05, we observe a non-seasonal behavior of deaths in Lima and Peru and we also observe a very marked increase of deaths in April and similar behavior of the time series. This trend gives us a detailed analysis of deaths by using the time series from January 2017 to January 2020. Moreover, we have obtained the predictions for February, March, and April 2020.

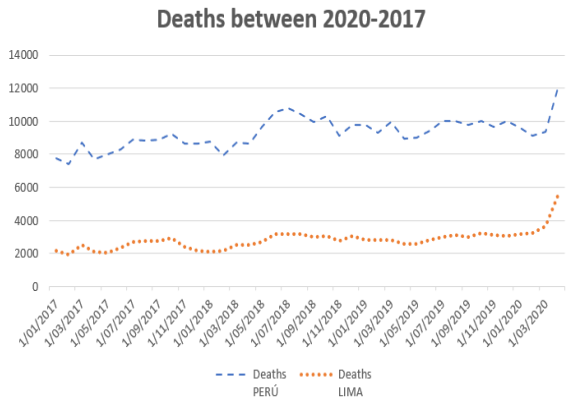

Fig. 5. Deaths that occurred in Lima and Peru from January 2017 to April 2020.

ARIMA models

Integrated Autoregressive Moving Average (ARIMA) models are a set of very powerful techniques for the analysis of time series that includes the study of people, observed groups, or varied information of a phenomenon at successive moments in time [20]. This technique allows us to study the conditional relationship of causes between different variables that change over time. It is one of the most widely used tools to make forecast inferences about the future and prediction of data. It is applied in different disciplines of knowledge such as health, econometrics, engineering, etc.

The methodology of ARIMA models consists of a time series that comprise the combination of autoregressive, differencing, and moving average term. These begin from the assumption of linear relationships where the current value of the variable of interest is expressed as a linear combination of terms such as lagged variable, current and past values of a Gaussian white noise process. [20]

In non-stationary processes, we analyze a type of lack of stationarity in the arithmetic mean which is very frequent in practice. Seasonality makes the arithmetic mean of the observations not constant; however, it evolves predictably according to a cyclical pattern. The most common case is to incorporate seasonality into the ARIMA modeling in a multiplicative way and it will result in a seasonal multiplicative ARIMA model. [21]

This model is characterized by having the expression shown in equation (1),

$$\emptyset_P(B^s)\varphi_p(B)\nabla_s^D\nabla^d z_t = \theta_q(B)\vartheta_Q(B^s)a_t \qquad (1)$$

where:

$\emptyset_P(B^s)$: is the seasonal autoregressive operator of order P.
φ_p: is the regular autoregressive operator of order p.
∇_s^D: represents seasonal differences.
∇^d: represents regular differences.
$\vartheta_Q(B^s)$: is the seasonal moving average operator of order Q.
$\theta_q(B)$: is the regular moving average operator of order q.
a_t: it is a white noise process.

Equation (1) represents correctly the seasonal series and they are written in a simplified form as the model

ARIMA (P, D, Q) × (p, d, q).

where:
P: is the order of the non-stationary autoregressive part.
D: is the number of non-stationary unit-roots (order of process integration).
Q: is the order of the non-stationary moving average part.
p: is the order of the stationary autoregressive part.
d: is the number of stationary unit-roots (order of integration of the process).
q: is the order of the stationary moving average part.

In particular, forecasting with ARIMA models make use of optimal prediction functions and they take into account those that minimize on average the squared prediction errors. The prediction of an ARIMA model has a relatively simple structure, the non-stationary operators, that is to say, the differences and the constant, if it exists, determine long-term predictions while the stationary operators, Autoregressive (AR) and Moving Average (MA) determine the short-term predictions.

The general solution of the final prediction equation of a seasonal process will have 3 components:
a. An expected trend term that will be a polynomial of degree d with coefficients that adapt over time if there is no constant in the model and a polynomial of degree d + 1 with the coefficient of the highest order term, β_{d+1} deterministic and given by $\mu/s(d+1)!$, where μ is the arithmetic mean of the stationary series.

b. A seasonal component expected to change with initial conditions.
c. A transient term of a short-term prediction that will be determined by the roots of regular and seasonal AR operators [22, 23].

Data Analysis

First, we will use the time series of the number of deaths occurred between January in 2017 to January in 2020, we will model each ARIMA time series (P, D, Q)x(p, d, q) for Lima and Peru. Then, we will use the ARIMA prediction for March and April in 2020 which will be analyzed together with the data of deaths due to COVID-19 reported by MINSA [3, 19]. We will also use the information from February in 2020 to determine the goodness of fit of our method along with the respective statistics (stationary r2 and Ljung-Box test). In the following paragraphs, we will estimate the number of deaths due to COVID-19 taking as a reference to our predictions from the ARIMA modeling and subtracting it from the data of SINADEF.

The best model for deaths in Peru is an ARIMA (9,2,2)x(0,0,2) where its non-seasonal component is an autoregressive order equal to nine, moving averages equal to two and differencing that is equal to 2. Its stationary part within the autoregressive order is zero, its part of moving averages is equal to zero and its differencing is equal to 2. It should be noted that a transformation has been made into a natural logarithm and the cyclicality of the time series from twelve months ago has also been taken into account in each observation.

In order to validate the time series, Table 01 shows the important statistics and the significance level of p = 0.135 greater than 0.05 and r^2 = 0.704 (stationary), this shows us the goodness of fit and the degree of reliability for the model. Table 02 shows the expected predictions of deaths in Lima with their respective confidence intervals.

Model fit statistics			Ljung-Box Q(18)		
R square stationary	R square	RMSE	statistics	GL	Sig.
0.704	0.545	661.023	8.420	5	0.135

Table 1. Most relevant statistics ensure the goodness of fit for the ARIMA time series regarding the death toll in Peru.

Model		Feb 2020	Mar 2020	Apr 2020
DATHSPERU-Modelo_1	Prediction	9130	9209	9096
	UCL	10139	10369	10410
	LCL	8201	8151	7913

Table 2. Predictions of death toll through the ARIMA time series in Peru.

ARIMA model closely follows the pattern of time series regarding deaths from January 2017 to January 2020 in Peru. We can observe this in Fig. 6.

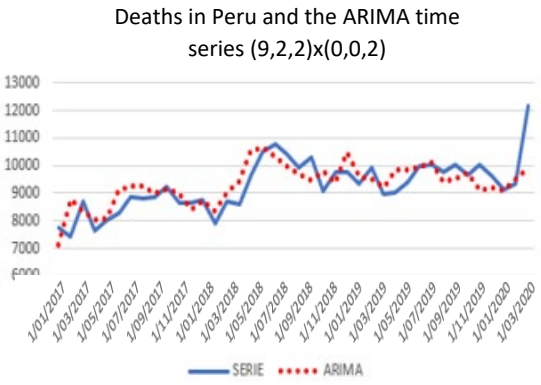

Fig. 6. The death toll in Peru from 2017 to 2020 is shown along with the best ARIMA model.

By the same token, the best model for the number of deaths in the city of Lima is an ARIMA model (5.2.8)x(1.0.1) where its non-seasonal component is an autoregressive order equal to five, moving averages equals two and whose differencing is equal to 8, its stationary part within the autoregressive order is one, its part of moving averages is equal to zero and its differencing is equal to 1. It has been carried out a transformation of the natural logarithm type.

In order to validate the time series, Table 3 shows the important statistics such as the level of significance that is p = 0.187 greater than 0.05 and r^2 = 0.503 (stationary), the goodness of fit,, and the degree of reliability. Table 04 shows the predictions of expected deaths in Peru with their respective confidence intervals.

Model fit statistics			Ljung-Box Q(18)		
R square stationary	R square	RMSE	statistics	GL	Sig.
0.503	0.502	336.293	4.798	3	0.187

Table 3. Most relevant statistics determine the goodness of fit for the ARIMA time series regarding the deaths that occurred in Lima.

Model		Feb 2020	Mar 2020	Apr 2020
DATHSLIMA-Modelo_1	Prediction	3234	3573	3561
	UCL	3978	4843	5377
	LCL	2602	2577	2260

Table 4. Predictions of the death toll in Lima using the ARIMA time series.

The ARIMA model closely follows the pattern of time series regarding the deaths that occurred from January 2017 to January 2020 in Lima. We can observe this in Fig. 7.

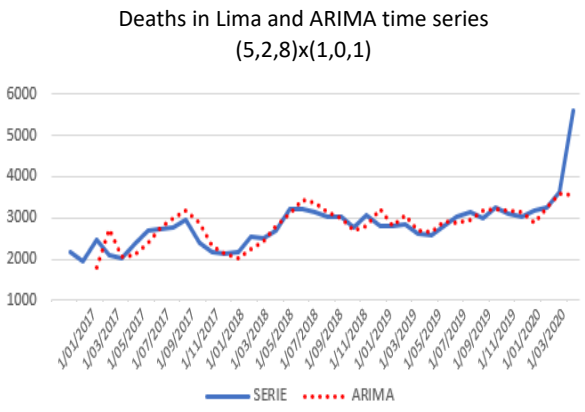

Fig. 7. The death toll in Lima from 2017 to 2020 is shown along with the best ARIMA model.

Results

Estimating the number of deaths in Lima and Peru

According to the predictions given by the ARIMA time series, in the case of Lima, a total of 3,573 deaths were expected for March. SINADEF has reported 3,621 deaths which give us a difference of 48 deaths. In the same way, for April, according to the prediction of the ARIMA time series, 3561 deaths were expected in Lima. SINADEF has reported 5,604 deaths which give us a difference of 2,043 deaths. The increased variation in deaths in April is directly related to the increase of infected people and, thereby, deaths that occurred due to Covid-19 compared to other months. We can observe this in Fig. 7.

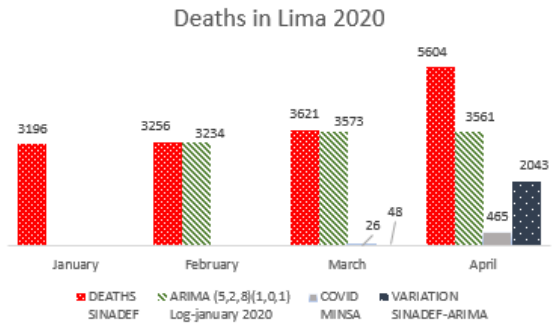

Fig. 7. Death toll registered in Lima where we observe that there was an increase in April due to COVID-19 pandemic.

According to the prediction of the ARIMA time series, for March a total of 9209 deaths were expected in Peru. SINADEF has reported 9,340 deaths, from subtracting these numbers we have a difference of 131 deaths. Likewise, for April, according to the prediction of the ARIMA time series, 9096 deaths were expected in Peru. SINADEF has registered 12178 deaths (the highest registered over all previous years) and from subtracting these numbers we obtain the difference of 3082 deaths. As well as for the case of Lima, the increase in deaths for April is directly related to the COVID-19 pandemic. We observe this in Fig. 8.

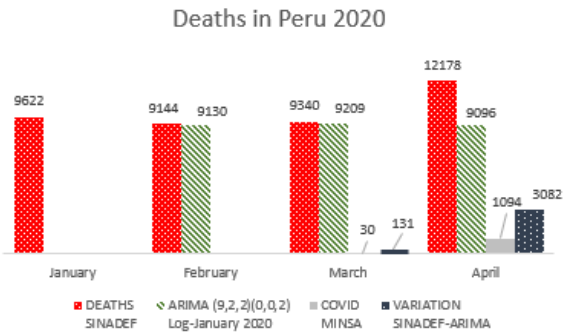

Fig. 8. Death toll registered in Peru, where we observe that for April, there was an increase due to the COVID-19 pandemic.

Conclusions

ARIMA time series model techniques are a powerful statistical tool to predict data that has a very well-known used in various areas of science, health, and economics. In the present study, we have shown the discrepancy between the data published by MINSA and the projection in our ARIMA models with a confidence level of 95% for March and April in 2020.

We have concluded that up to April 30, 2091 deaths occurred in Lima due to COVID-19. MINSA has reported 491 deaths which means that our model has detected an increase in deaths by 325.9%.

We have concluded as well that as a result of COVID-19, up to April 30, 3,213 deaths have occurred in Peru. MINSA has reported 1,124 deaths which means that our model has detected an increase of deaths by 185.9%.

The difference of death toll between SINADEF death records and ARIMA prediction model that have taken place in Lima and Peru are directly and indirectly related to the COVID-19 pandemic. For instance, we indirectly have deaths of patients with previous chronic diseases such as oncological diseases, diabetes, patients with kidney problems, neuronal problems, and many more. The situation of the patients has got worse due to the collapse of our health system during the current pandemic.

For this reason, it is important to carry out a research regarding the quality of the data as well as to the ones related to projections about the collapse of the health system, estimations, inferences, and deaths reported with ARIMAX-type, hybrid or logistic-type models in order to have new statistical tools for upcoming events.

Acknowledgements:

Not applicable.

Authors' contributions:

EV and AL conceived the study. AS and EV were responsible for the gathering of data. EV, AL and AS analysed and interpreted the data. EV and AL drafted the manuscript. All authors read and approved the final manuscript.

Funding:

Not applicable.

Availability of data and materials:

The datasets used and analysed during the current study are available.

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Competing interests:

The authors declare that they have no competing interests.

Author details:

[1] Faculty of Physics UNMSM, University City, Lima 1, Perú
[2] Academic Department of Engineering Universidad del Pacifico, Lima 11, Perú.
[3] Faculty of Mathematical Sciences, UNMSM, University City, Lima 1, Perú

BIBLIOGRAPHY

[1] Ministerio de Salud (MINSA). (2020). *Plan Nacional de Preparación y Respuesta frente al riesgo de introducción del Coronavirus 2019-nCoV.* Resolución Ministerial N 039-2020-MINSA. [Online]. Available https://cdn.www.gob.pe/uploads/document/file/505245/resolucion-ministerial-039-2020-MINSA.PDF [Accesses 02 05 2020].

[2] Agencia de noticias EFE. (2020). *El Primer caso de coronavirus en Perú está aislado en casa*. [Online]. Available https://www.efe.com/efe/america/sociedad/el-primer-caso-de-coronavirus-en-peru-esta-aislado-casa-y-trabaja-latam/20000013-4189708 [Accesses 05 05 2020].

[3]Ministerio de Salud (Minsa). (2020). *Situación actual "COVID-19" al 30 de abril del 2020.* [Online]. Available https://www.dge.gob.pe/portal/docs/tools/coronavirus/coronavirus300420.pdf [Accesses 01 05 2020].

[4] Koch E. Bravo M. et al. (2012). *Sobrestimación del aborto inducido en Colombia y otros países latinoamericanos* [Online]. Available http://handbook.usfx.bo/nueva/vicerrectorado/citas/SALUD_10/Medicina/17.pdf [Accesses 12 05 2020].

[5] Giner L. Guija J. (2014). *Número de suicidios en España: diferencias entre los datos del Instituto Nacional de Estadística y los aportados por los Institutos de Medicina Legal*. [Online]. Available https://www.elsevier.es/es-revista-revista-psiquiatria-salud-mental-286-pdf-S1888989114000056 [Accesses 05 05 2020].

[6] Ruiz M. Márquez L. Miller T. (2015). *La mortalidad materna: ¿por qué difieren las mediciones externas de las cifras de los países?* Serie Población y desarrollo. Santiago de Chile. Publicación de las Naciones Unidas.

[7] Velásquez Hurtado J. E., Kusunoki Fuero L., Paredes Quiliche T. G., Hurtado La Rosa R., Rosas Aguirre Á. M., & Vigo Valdez W. E. (2014). *Mortalidad neonatal, análisis de registros de vigilancia e historias clínicas del año 2011 en Huánuco y Ucayali, Perú*. Revista peruana de medicina experimental y salud pública, 31, 228-236.

[8] McIver D. J., and Brownstein J. S. (2014). *Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time*. *PLoS computational biology* 10.4.

[9] Eze-Nliam C., et al. (2012). *Discrepancies between the medical record and the reports of patients with acute coronary syndrome regarding important aspects of the medical history*. BMC health services research 12.1: 78.

[10] Diario The New York Time. (2020). *74,000 Missing Deaths: Tracking the True Toll of the Coronavirus Outbreak.* [Online]. Available https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html?action=click&module=RelatedLinks&pgtype=Article [Accesses 14 05 2020].

[11] Diario el comercio. (2020). *¿Exceso de muertes tiene relación directa con el COVID-19? En abril hubo 2 mil más respecto al año pasado.* [Online]. Available https://elcomercio.pe/peru/coronavirus-en-peru-la-cifra-de-muertes-durante-la-pandemia-bajo-analisis-noticia/?ref=ecr [Accesses 1 05 2020].

[12] IDL Reporteros. (2020). *Los muertos que el Gobierno no cuenta*. Consultado el 5 de mayo del 2020. URL: https://www.idl-reporteros.pe/los-muertos-que-el-gobierno-no-cuenta/

[13] Coutin Marie, Gisele. "Utilización de modelos ARIMA para la vigilancia de enfermedades transmisibles." *Revista Cubana de Salud Pública* 33.2 (2007): 0-0.

[14] Bedoya Luza, S. L. (2018). Modelamiento univariado del número de defunciones infantiles producidas por infecciones respiratorias agudas, a través de la metodología Box-Jenkins, puno 2008-2016.

[15] Ocaña-Riola, R. (2004). Eficacia del análisis de series temporales para la planificación sanitaria del cáncer en España. Atención Primaria, 34(1), 15-19.

[16] León-Álvarez, A. L., Betancur-Gómez, J. I., Jaimes-Barragán, F., & Grisales-Romero, H. (2016). Clinical and epidemiological rounds. Time series. Iatreia, 29(3), 373-381.

[17] Cárdenas, C., Sovier, C., Pérez, U., & GONZÁLEZ, C. S. (2014). Consultas de urgencia general y por causa respiratoria en la Red de establecimientos del Sistema Nacional de Servicios de Salud (SNSS): un modelo predictivo en el Servicio de Salud de Chiloé. Revista chilena de enfermedades respiratorias, 30(3), 133-141.

[18] Sistema Informático Nacional de Defunciones (SINADEF). (2020). *Tablero de control.* [Online]. Available

http://www.minsa.gob.pe/reunis/data/defunciones_registradas.asp [Accesses 18 05 2020].
[19] Ministerio de Salud (Minsa). (2020). *Situación actual "COVID-19" al 31 de marzo del 2020.* [Online]. Available https://www.dge.gob.pe/portal/docs/tools/coronavirus/coronavirus310320.pdf [Accesses 18 05 2020].
[20] Box G. E. P. Jenkins G. M. (1970). *Time series analysis: forecasting and control, 1976*. New York. EEUU. ISBN: 0-8162-1104-3.
[21] Chatfield C. (2013) *The Analysis of Time Series: An Introduction.* Chapman & Hall/CRC. Boca Raton. p.179-194.
[22] Mauricio J. (2007). *Introducción al análisis de series temporales.* Madrid: Universidad complutense de Madrid.
[23] Guirao, A. (2020). *Entender una epidemia. El coronavirus en España, situación y escenarios*. [Online]. Available https://digitum.um.es/digitum/bitstream/10201/88621/1/Entender%20una%20epidemia_Guirao2020.pdf [Accesses 15 05 2020].

Figures

Figure 1

Deaths that occurred in Peru from January 2017 to April 2020. An atypical upward trend is observed in April 2020.

Figure 2

Deaths in Lima from January 2017 to April 2020. The growing trend of deaths in April 2020 is very marked.

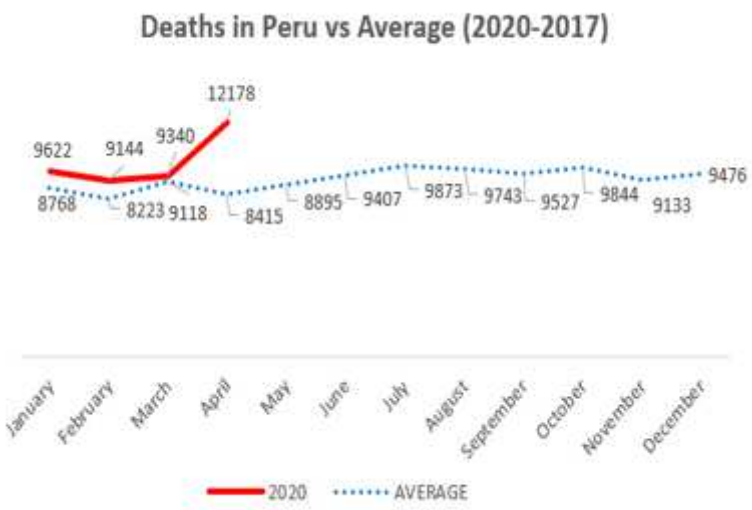

Figure 3

The deaths that occurred in Peru in 2020 compared to the average in the years 2017-2019. There is a marked difference between the average and those registered in April 2020.

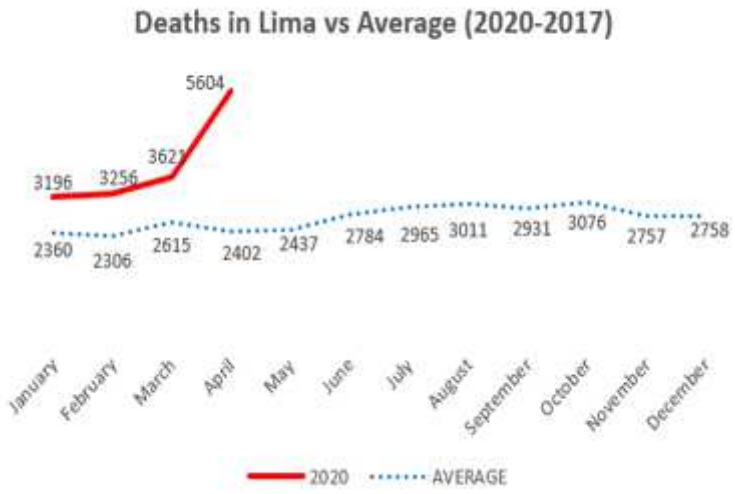

Figure 4

Deaths that occurred in Lima in 2020 compared to the average in the years 2017-2019. There is an appreciable difference between January to April according to the average.

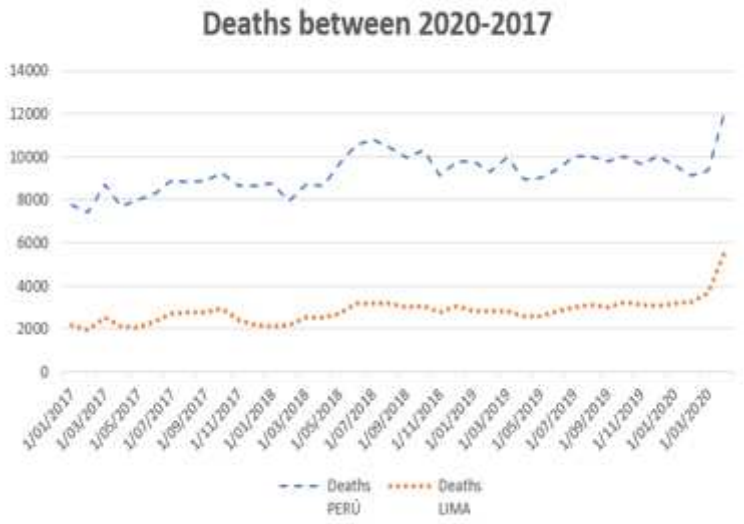

Figure 5

Deaths that occurred in Lima and Peru from January 2017 to April 2020.

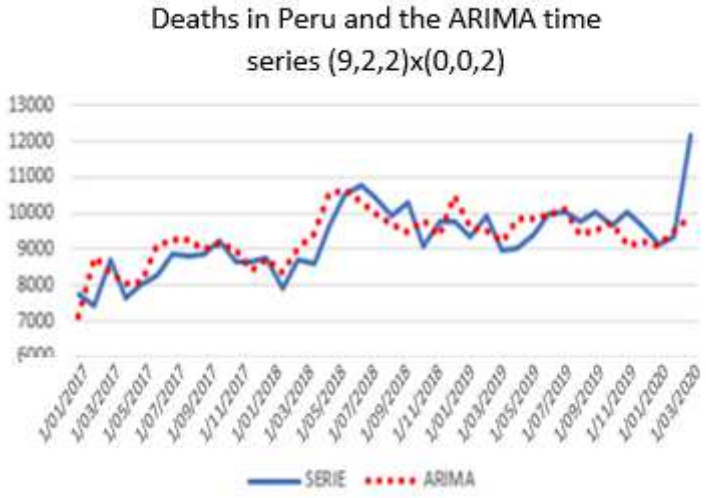

Figure 6

The death toll in Peru from 2017 to 2020 is shown along with the best ARIMA model.

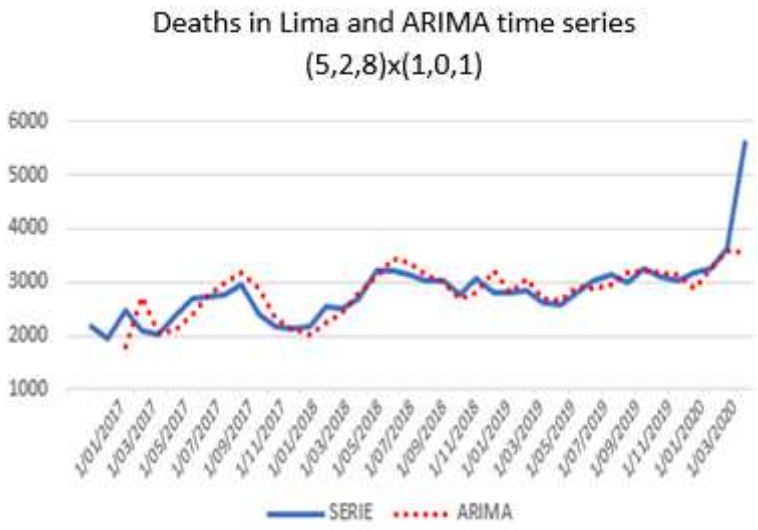

Figure 7

The death toll in Lima from 2017 to 2020 is shown along with the best ARIMA model.

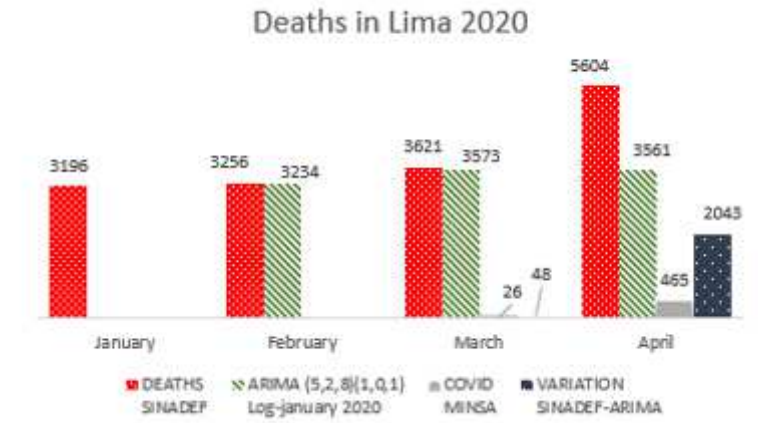

Figure 8

Death toll registered in Lima where we observe that there was an increase in April due to COVID-19 pandemic.

Figure 9

Death toll registered in Peru, where we observe that for April, there was an increase due to the COVID-19 pandemic.