Detection of Outliers in The Peruvian Fruit Production Time Series Using Arima Models

Manuel Chávez[1], *Israel Chávez*[1], *Eduardo Torres*[1], *Sandro Atoche*[1], *Stefano Palacios*[1], *Luis Trelles*[1], *Cristhian Aldana*[1], *Yesenia Saavedra*[1], *Gustavo Mendoza* [1], *Nelson Chuquihuanca*[1]

[1] Universidad Nacional de Frontera - UNF
Av. San Hilarión 101, Sullana, Piura, PERÚ

ABSTRACT

The present applied, non-experimental, descriptive and prognostic research; was aimed at detecting outliers in the agricultural production of Mangifera indica (mango), Persea americana (avocado) and Citrus x aurantifolia (lemon) at the national level, was performed by applying an ARIMA Model. To fulfill it purposes, documentary analysis was used at the National Institute of Statistics and Informatics (In Spanish, INEI). The study sample consisted of the mango, avocado and lemon production indices 2000-2020. As a result, the models were obtained arima mango (1,0,0) (2,1,2) (AIC=5448.99, BIC=5473.35 and RMSE=19067.93), arima avocado (0,1,3) (2,1,0) (AIC=4687.05, BIC=4707.91 and RMSE=4114.35) and arima lemon (1,0,1) (0,1,1) (AIC=4484.36, BIC=4501.76 and RMSE=2551.96) with a 12 months period, the diagram of boxes and whiskers was also made with which it was identified that atypical data (Outliers) abound in the periods of greatest production.

Keywords: Outliers, forecast, ARIMA, agricultural crops, times series.

INTRODUCTION

The Mangifera indica (Mango) originated in India more than 4000 years ago and has about 30 species, also describe their large proportions and each species has its own characteristic flavor; being these from the sixteenth century distributed world-wide, reaching America in the eighteenth century. We also mention that the species Mangifera indica in Peru was recorded as a yellow mango in the district of Chulucanas, cultivated to date in 22 departments of Peru having a cultivation area of approximately 30,817 ha. Its main producing regions are Piura and Lambayeque with 82% of the national harvest (Tuisima et al. 2021). The most traded varieties of Peruvian mango are Kent, Haden, Edward and Tommy Atkins, with 242,879,787 kg of mango exported in 2020 at a price of 284,101,570 USD, the main destinations being the Netherlands and the United States.

Persea americana (avocado) is a tropical fruit with nutritional benefits that con-tribute to health wellness, is consumed fresh and is of great importance for multiple industries due to the various products derived from it; for these reasons it is a very attractive product for export, expanding cultivation along the Peruvian coast. In addition, he adds that south of Lima is the Cañete valley where 1075 ha of avocado are cultivated (Collantes et al. 2017). Peru is the third country with the highest avocado exports, strengthen-ing in 2018 as one of the main agro-export items. This export growth helped agricul-tural development and created multiple jobs (Morales et al. 2020).

Citrus is the most exported product in the world, with China and Spain in first place; with 1,112,100 tons, Peru ranks fourth in South America in production and seventh in exporting 37,600 tons of citrus, with the departments of Piura, Lam-bayeque, Lima, Ica, Junín and Cusco being the main producers (Llihua et al. 2019). Likewise, in the Piura region there are 16,000 ha dedicated to the production of Citrus x aurantifolia (lemon), whose main producers are the San Lorenzo valley, Alto Piura, Medio Piura, Cieneguillo and El Chira; obtaining 60% of the production as natural consumption and 40% is applied in the manufacture of oils (Peña et al. 2018).

Materials and methods

For the development of the research, it has been chosen to use a non-experimental method and descriptive level of prognosis. Taking into account the identification of atypical data in time series and residual forecasts. Our research was developed with the free software RStudio for the elaboration of Arima models and the detection of outliers, whose estimation uses the following functions:

Autocorrelation function:

$$h_k = \frac{g_k}{g_0} = \frac{Cov(X_t, X_{t+k})}{Var(X_t)}, k = \cdots, -2, -1,0,1,2, \ldots$$

$$where\ g_0 > 0, g_k = g_{-k}, h_k = h_{-k}, h_0 = 1\ y\ |h_k| \leq 1$$

Partial autocorrelation function:

$$r_k = \frac{\sum_{t=1}^{T}(X_t - \bar{X})\,(X_{t-k} - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

Ljung – Box:

$$Q' = \frac{n(n+2)\sum_{k=1}^{m} {r_k}^2}{(n-k)} \approx X^2_{m-p-q}$$

$$\alpha \rightarrow level\ of\ significance$$

$$I\ is\ calculated$$

$$P(Q' > I) = \alpha$$

With the above, the Arima model was made and check if it was good.

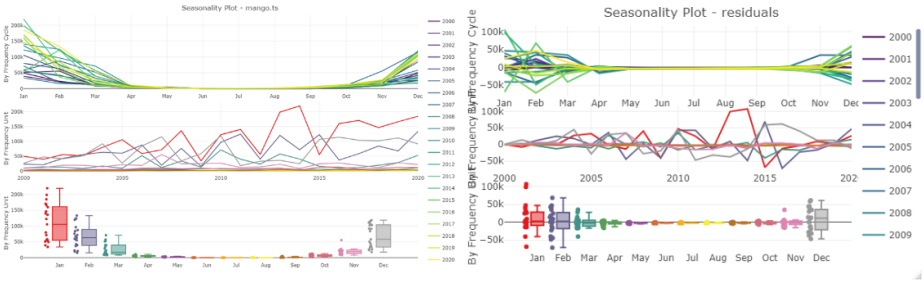
RESULTS

In the mango (see Table 1)

Table 1: Arima Model (1,0,0) (2,1,2) [12]

ARIMA MODEL	**(1,0,0) (2,1,2)**
AIC	54448.99
BIC	5473.35
ME	101.165
RMSE	19067.93
MAE	9915.798
MPE	-445.006
MAPE	458.3792
MASE	0.7390689
ACF	0.00381338

An ARIMA MODEL (1,0,0) (2,1,2) was realized with a period of 12 months. Verifying its stationarity with the Dickey – Fuller method which gave a p-value of 0.01 and to identify if it presents a good fit, the Ljung – Box test was used, which gives a p-value of 0.9514. For the detection of atypical data, the diagram of boxes and whiskers was used. It was used in the time series detecting that in the months of greatest production there is a greater amount of atypical data (January, February, March and December) and it is confirmed by making the diagram with the residual of the forecast detecting a greater amount of error in the aforementioned months (see Figure 2).



(a) Mango times series **(b)** Residual prognosis

Figure 1. Production of mango in Peru, 2000-2009 period

Avocado (see Table 2):

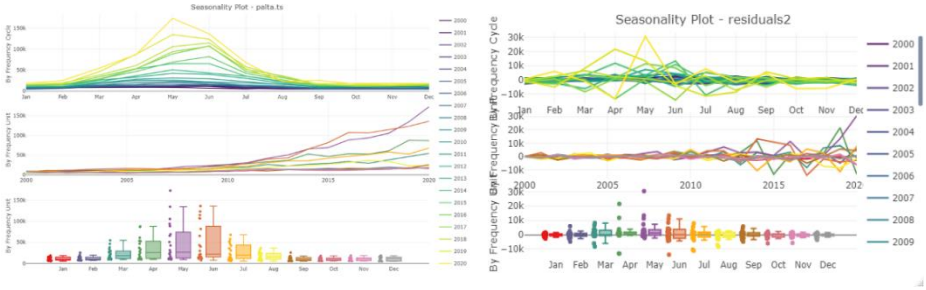
Table 2: Arima Model (0,1,3) (2,1,0) [12]

ARIMA MODEL	**(0,1,3) (2,1,0)**
AIC	4687.05
BIC	4707.91
ME	321.8878
RMSE	4114.35
MAE	2233.312
MPE	-0.8567565
MAPE	10.96324
MASE ACF	0.716675 -0.01617514

An ARIMA MODEL (0,1,3) (2,1,0) was realized with a period of 12 months. Verifying its stationary with the Dickey – Fuller method which gave a p-value of 0.01

and to identify if it presents a good fit, the Ljung – Box test was used, which gave a p-value of 0.7962 (see Table 2)

For the detection of atypical data, the diagram of boxes and whiskers was used. It was used in the time series that detecting that in the months of greatest production there is a greater amount of atypical data (April, May and June) and it is confirmed by making the diagram with the residual of the forecast detecting a greater amount of error in the aforementioned months (see Figure 2).



(a) Avocado time series **(b)** Residual prognosis

Figure 2. Production of avocado in Peru, 2000-2009 period

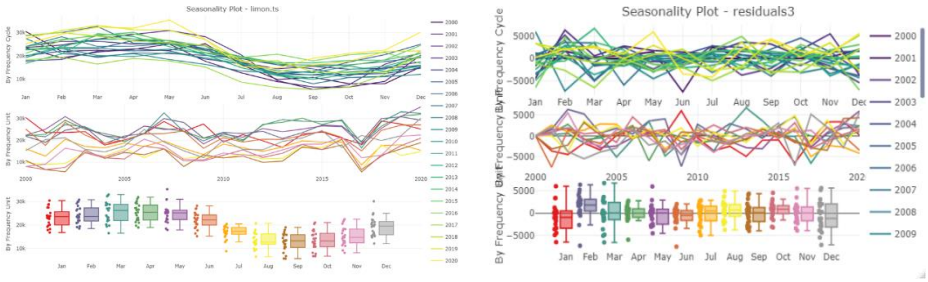
The lemon (see Table 3):

Table 3. Arima Model (1,0,1) (0,1,1) [12]

ARIMA MODEL	**(1,0,1) (0,1,1)**
AIC	4484.36
BIC	4501.76
ME	-13.12575
RMSE	2551.96
MAE	1939.362
MPE	-1.723145
MAPE	11.08032
MASE	0.5247441
ACF	-0.02334006

An ARIMA MODEL (1,0,1) (0,1,1) was made with a period of 12 months. Verifying its stationary with the Dickey – Fuller method which gave a p-value of 0.01 and to identify if it has a good fit the Ljung – Box test was used, which gave a p-value of 0.7094.

For the detection of atypical data, the diagram of boxes and whiskers was used. It was used in the time series detecting that in the months of greatest production there is a greater amount of atypical data (April, May and June) and it is confirmed making the diagram with the residual forecast detecting a greater amount of error in the aforementioned months (see Figure 3).



(a). Lemon time series **(b)**. Residual prognosis

Figure 3. Production of lemon in Peru, 2000-2009 period

CONCLUSIONS

The increase in the amount of fruit production is due to exports and the wide international market that Peru has made its way over time; finding that the months with the highest seasonal index are those that contain the highest production of fruit trees.

The behavior of the fruit trees in their selected historical data will depend of the season in which each one develops, that is to say, there will be more of the fruit trees according to the conditions that are required for their harvest.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Universidad Nacional de Frontera, Sullana, Piura, Perú.

REFERENCES

Collantes González, R., Rodríguez Berrio, A., & Canto Sáenz, M. (2017). Caracterización de fincas productoras de palto (Persea americana Mill.) y mandarina (Citrus spp.) en Cañete, Lima, Perú.

Llihua Quispe, L. J., Calderón Rodríguez, A., & Cabrera Pintado, R. M. (2019). Influencia de sacarosa y cotiledones en la microinjertación de cítricos.

Morales T., E. M., Lino N., M. D., Ortega R., E., & Castellanos S., P. L. (2020). In vitro antagonic capacity evaluation of Trichoderma spp strains against Phytophthora cinnamomi, phytopathogen of Persea americana (Palta).

Peña Castillo, R. A., & Galecio Julca, M. Á. (2018). Application of rice husk as a source of silicon and mineral fertilization of lemon (Citrus aurantifolia) in sandy soils – Piura.

Tuisima Coral, L. L., & Escobar García, H. A. (2021). Characterization of fruits of varieties of mango (Mangifera indica) conserved in Peru.