Multivariate Cuban Consumer Price Index Database, Statistic Analysis and Forecast Baseline Based on Vector Autoregressive

Reynaldo Rosado, Héctor González Diéz(✉), Orlando Grabiel Toledano-López, and Yanio Hernández Heredia

Universidad de las Ciencias Informaticas (UCI), La Habana, Cuba
{rrosado,hglez,ogtoledano,yhernandezh}@uci.cu

Abstract. The global Consumer Price Index (CPI) is a monthly multivariate time series, which allows measuring the variation of the final consumer prices of a given set of goods and services of households living in a given geographic region, city or country. The present work addresses the problem of the multivariate time series database of Cuba's CPI and a respective forecasting model based on Vector Autoregressive to establish a baseline for this dataset. An statistical analysis of the data will allow characterizing each variable of the series in terms of relevance to the multivariate problem, its causal relationships and the respective stationary analysis to evaluate the best lag to be considered in the forecasting model. The main statistics evidences of each test were reported in the paper as starting point for futures researches in the field of deep learning.

Keywords: Consumer Price Index · Multivariate Time Series Forecasting · Vector Autoregression

1 Introduction

The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. The CPI is a widely indicator of inflation, and is managed, in Cuba, by the Government of the National office of the Statistic and Information (ONEI) and by similar organizations and institutes in other countries. The CPI is systematically way taken as a reference for decision-making regarding monetary policies by governments and financial entities. It is also used for various aspects of social finance, such as retirement, unemployment and government financing [25]. The CPI is typically calculated using a Laspeyres index formula [5], which holds the basket of goods and services constant over time. However, the ONEI also calculates a multivariate CPI, which accounts for changes in the quality of goods and services over time. This is done using a hedonic regression model, which estimates the value of different product characteristics (such as Food and Drinks, Health, Houses, Transportation, among others) and adjusts prices accordingly [18,24].

Y. Hernndez Heredia et al. (Eds.): IWAIPR 2023, LNCS 14335, pp. 24–34, 2024.
https://doi.org/10.1007/978-3-031-49552-6_3

The multivariate CPI is generally considered to be a more accurate measure of inflation than the traditional CPI, as it better accounts for changes in quality. However, it is also more complex to model by the relevance of the goods and service (normally need to define the weight manually by experts). The multivariate CPI accounts for changes in the quality of goods and services over time, whereas the univariate CPI assumes that the quality of the basket of goods and services remains constant. This means that the multivariate CPI provides a more accurate measure of inflation, since it adjusts for changes in quality and captures the true cost of living. The work [15] provides an overview of identification problems in macroeconomics, including those related to constructing price indexes and the main advantage of the multivariate approach.

In the case of Cuba, the weighting reflects the data obtained in the National Survey of Household Income and Expenditures (ENIGH), which was conducted between August 2009 and February 2010. The weights of goods and services are therefore based on the consumption expenditures that households have access to at that time. The goods and services that affect Cuba's CPI are: 01 Food and non-alcoholic beverages; 02 Alcoholic beverages and tobacco; 03 Clothing; 04 Housing services; 05 Furniture and household items; 06 Health; 07 Transportation; 08 Communications; 09 Recreation and culture; 10 Education; 11 Restaurants and hotels; 12 Miscellaneous personal care goods and services [6].

The Monthly Publication of the CPI from the National Office of Statistics and Information (ONEI), allows to know the average variation experienced by the prices of a basket of goods and services, representative of the consumption of the population in a given period. Approximately 33 596 prices are collected monthly, in 8 607 establishments, located in 18 municipalities throughout Cuba, the urban area of the head municipalities of 14 provinces and 4 municipalities of Havana province, obtaining national coverage. This means that the index to be shown is only representative of the country; it does not exist at the level of regions or municipalities. The basket of goods and services includes 298 items that represent more than 90.0% of household expenditure. The data are published in the form of reports in pdf format, which makes it difficult to process and analyse them because there is no integrated view of the database [18].

Both the prices of the products and services that give origin to the CPI estimate, as well as the CPI itself, are calculated systematically, so they are time series data type. As CPI forecast helps to estimate future trends, it is key for decision making. Moreover, it allows the application of price stabilization policies to reduce the economic impact on the prices of products and services demanded by consumers. In those economies that present instability, CPI data fluctuate over time, which translates into a non-linear and non-stationary behaviour [19].

In general, several approaches in the CPI forecasting field, modelling the problem as a univariate time series, concentrating only on the study of the global indicator. Approaching it as a multivariate problem, taking into account the variation of the prices of each goods or service included in the basket, is not very well treated since the global index is composed of the weighted aggregation of the prices to each products. The most widely used statistical method for

forecasting the CPI as a univariate time series problem has been the family of the Autoregressive models [2,7,9,14,16,17]. Recently, deep learning techniques for time series forecasting have improved the performance of CPI prediction. Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) architectures have the ability to capture time dependence in data, while handling more than one output variable to estimate more than one time instant. Three examples that show good performance with simple LSTM [25] model and temporal data at different time intervals are from Mexico [11], Ecuador [20,21] and Indonesia [13]. In spite of the prominent performance of RNN models for time series forecasting, particularly for financial series, they have been studied on the basis of autoregressive models such as VAR. This is due to the limited availability of data in this scenario which is also reflected in the CPI as described in the previously works.

The aim of this paper are to propose a new multivariate time series database of Cuba's monthly CPI and a respective forecasting model based on Vector Autoregressive model as a baseline follow the statistics methodologies of analysis. A statistical analysis of the data will allow characterizing each variable of the series in terms of relevance to the multivariate problem, its causal relationships and the respective stationary analysis to evaluate the best lag to be considered in the forecasting model.

2 Multivariate Analysis

2.1 Definitions and Notation

A multivariate time series is defined as a collection of the multiples variables spatially related and individually shows a temporal relationship. Classical statistical or machine learning models need to consider the univariate or multivariate problem differently, however deep learning models can handle both indistinctly with high accuracy. Time series are usually characterized by three components: trend, seasonality and residuals [23]. In real-world time series and, in particular the CPI problem, seasonality can be affected by external agents such as the economic and financial crisis, prices of the main products in the world market, and emerging situations such as the COVID-19 pandemic.

In a more formal definition of the Multivariate Time Series we have m variables or observations, each of which has a time series. These variables are correlated in a way that the value of them at time t is related to the temporal window of size p previous values of all other variables including its own past values. We can represent each forecast in the set of variables at time t as a linear combination:

$$\begin{aligned}
\hat{y}_t^1 &= w_0^1 + w_{11}^1 y_{t-1}^1 + \ldots w_{m1}^1 y_{t-1}^m \\
&\quad + \ldots w_{1p}^1 y_{t-p}^1 + \ldots w_{mp}^1 y_{t-p}^m + \epsilon_t^1 \\
&\vdots \\
\hat{y}_t^m &= w_0^m + w_{11}^m y_{t-1}^1 + \ldots w_{m1}^m y_{t-1}^m \\
&\quad + \ldots w_{1p}^m y_{t-p}^m + \ldots w_{mp}^m y_{t-p}^m + \epsilon_t^m
\end{aligned} \tag{1}$$

Finally, the corresponding time series forecasting problem consists of the estimating a predictor $F : \mathbb{R}^{(m+1)} \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ in such a way that the expected deviation between true and predicted outputs is minimized for all possible inputs. The model associated with the Eq. (1) is to known as Vector of Auto-Regression (VAR). In the context of the CPI, this scenario makes it possible to forecast the price of goods and services that contribute to the overall or general CPI.

2.2 Statistical Analysis in Multivariate CPI.

Granger's Causality Test
The Granger causality test [8] is a statistical hypothesis test used to determine whether one time series is useful in forecasting another time series. The test is based on the idea that if a time series x "Granger-causes" another time series y, then past values of x should contain information that helps predict future values of y, beyond what can be predicted using past values of y alone.

The Granger causality test is commonly used in econometrics, finance, and other fields to investigate causal relationships between time series. Several applications in the multivariate CPI can be found in recently researches [1,10,12,22].

It is also worth noting that the Granger causality test assumes that the time series are stationary, so it is often preceded by a test for stationarity such as the Augmented Dickey-Fuller (ADF) test. Additionally, the test is sensitive to the choice of lag length and model specification, so it is important to carefully choose these parameters based on the data and the research question at hand.

Augmented Dickey-Fuller Test
The multivariate Augmented Dickey-Fuller (ADF) test is an extension of the standard ADF test that allows for multiple time series to be analysed simultaneously, taking into account possible relationships between them.The multivariate ADF test involves estimating a vector autoregressive (VAR) model for the set of time series and testing for the presence of unit roots in the model. Test examines whether the residuals of VAR model are stationary, which is equivalent to testing for stationarity of each individual time series after controlling for the other time series in the model [4].

The multivariate ADF test is useful in identifying whether a set of time series are stationary in a joint sense, which can be important for modelling and forecasting purposes. For example, if a set of economic variables are jointly non-stationary, it may be difficult to develop accurate forecasting models that account for the interrelationships between variables. It is important to note that the multivariate ADF test has some limitations and assumptions. For example, it assumes that the VAR model is correctly specified and that the residuals are normally distributed and free from serial correlation. Additionally, the test can be sensitive to the lag length of the VAR model and the number of variables included in the model [3]. Therefore, it is important to carefully select the appropriate model specification based on the data and the research question at hand.

3 Results and Discussion

3.1 Exploratory Analysis and Dataset

The Cuban Consumer Price Index database was collected from the official website National Office of the Statistic and Information ONEI [18]. This is a monthly time series from January 2010 to December 2020 with very low variability in the data as we can show in Table 1.

Table 1. Characteristic of the Cuban Consumer Price Index dataset.

	Overall	ABNA	BAT	PBC	SV	MAH	S	T	C	RC	E	RH	BSD
count	120.00	120.00	120.00	120.00	120.00	120.00	120.00	120.00	120.00	120.00	120.00	120.00	120.00
mean	103.35	109.74	101.02	91.88	101.65	102.90	107.92	105.41	81.66	99.91	100.69	104.15	105.36
std	1.37	4.35	1.33	4.07	0.96	0.87	1.47	2.35	5.08	0.86	0.95	1.95	1.20
min	100.12	99.66	92.32	77.32	99.57	100.79	100.04	100.93	72.40	95.64	99.46	100.09	99.93
25%	102.66	107.28	100.18	88.58	101.12	102.12	108.41	103.74	78.93	99.26	99.96	102.49	105.52
50%	103.15	109.58	100.32	91.52	101.70	102.79	108.41	104.49	80.03	99.59	100.53	104.80	105.68
75%	103.88	111.66	101.94	95.17	101.94	103.67	108.46	107.85	80.46	100.67	100.93	105.22	105.80
max	109.50	126.54	104.76	100.59	106.07	104.34	108.48	109.92	100.00	101.72	104.51	111.32	107.63

LEGEND

ABNA: Food and non-alcoholic beverages
BAT: Alcoholic beverages and tobacco
PBC: Clothing
SV: Housing services
MAH: Furniture and household items
S: Health
T: Transportation
C: Communications
RC: Recreation and cultur
E: Education
RH: Restaurants and hotels
BSD: Miscellaneous personal care goods and services

The values are the overall averages of the Cuban CPI for 11 category groupings and almost 298 goods and services. It is necessary to clarify that the data sets in the context of the CPI are very short series where learning models that require a lot of data are not effective in this context. Under these conditions we have modelling an appropriated problem as a time series forecasting. The Fig. 1 show the trends of the series and seasonality for each category (dashed line) respect to the overall.

3.2 Statistics Analysis

Overall, Granger's causality test is a useful tool for analysing causality between two time series, but it should be used in conjunction with other methods and careful interpretation of the results. The results of the Granger causality test involves assessing the statistical evidence for causality, determining the direction of causality, assessing the strength of causality, and considering the context and theoretical implications of the result. It is important to be cautious in interpreting Granger causality results and to consider other evidence and methods when assessing causality in time series data.

The null hypothesis is that the past values of x do not help in predicting y, while the alternative hypothesis is that the past values of x do help in predicting y. In Fig. 2 we can show that series like Transportation and Food and Non Alcohol Drinks have very low predictive power respect to another series. Also, this two series have similar causality relation respect to General CPI.

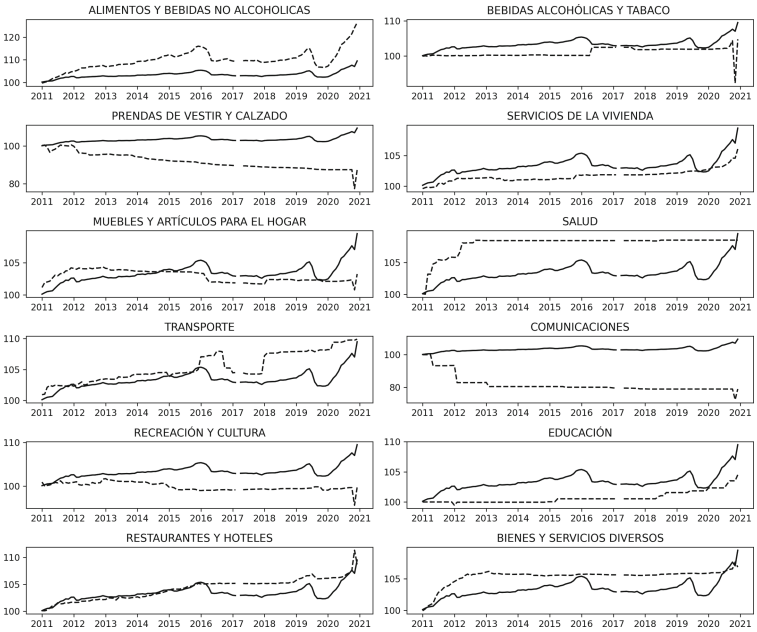

Fig. 1. Cuba CPI 2010–2021 by categories.

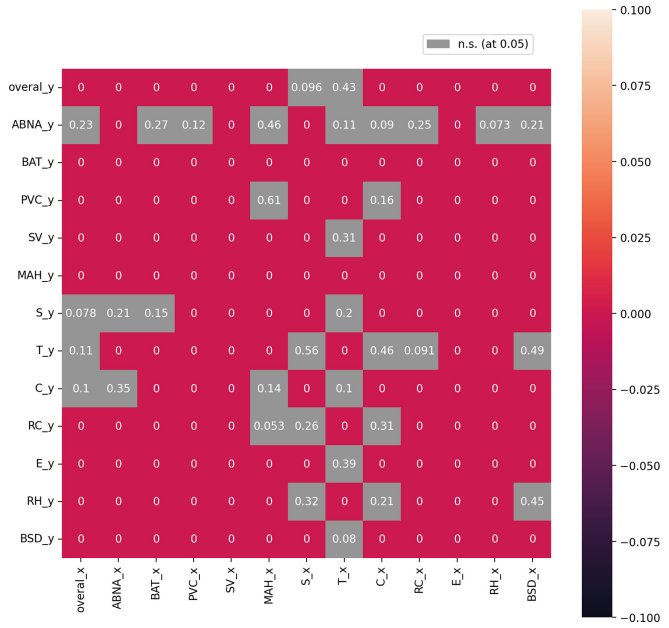

Fig. 2. Test the Granger to causality analysis in the Multivariate CPI.

The Augmented Dickey-Fuller (ADF) test is a powerful tool used to check the stationarity of the time series. This test can help to choose various parameters such as the optimal lag or the differential order to transform the multivariate series into stationary. The null hypothesis of the ADF test is that the time series is non-stationary. Therefore, if the p-value of the test is below the significance level (0.05), the null hypothesis is rejected and it follows that the time series is truly stationary. In our time series, the result of the ADF test can be find in Table 2. The test result showing that the series its non-stationary while the first differential its stationary. Also, we can report in Table 2 the best lags for each series.

Table 2. Multivariate ADF Test over original series and the first differential.

Series	Original Series				First differential of the Series			
	F-fuller	p-value	Stationarity	Lags	F-fuller	p-value	Stationarity	Lags
OVERALL	−2.867	0.049	stationary	1	−6.594	0.0	stationary	0
ABNA	−2.67	0.079	non-stationary	1	−5.402	0.0	stationary	0
BAT	−1.614	0.476	non-stationary	0	−8.457	0.0	stationary	1
PVC	−1.316	0.622	non-stationary	2	−8.123	0.0	stationary	1
SV	−2.841	0.053	non-stationary	3	−4.482	0.0	stationary	2
MAH	−1.533	0.517	non-stationary	2	−7.046	0.0	stationary	1
S	−3.631	0.005	stationary	5	−3.787	0.003	stationary	12
T	−1.531	0.518	non-stationary	0	−8.801	0.0	stationary	0
C	−7.551	0.0	stationary	11	−1.539	0.514	non-stationary	11
RC	−1.56	0.503	non-stationary	0	−7.728	0.0	stationary	1
E	1.047	0.995	non-stationary	1	−11.694	0.0	stationary	0
RH	−2.506	0.114	non-stationary	2	−8.607	0.0	stationary	1
BSD	−4.643	0.0	stationary	11	−3.783	0.003	stationary	12

Lag Selection in VAR

The choice of the best metric for lag selection in time series analysis depends on the specific modelling approach and the characteristics of the data. Akaike Information Criterion (AIC): The AIC is a measure of the relative quality of statistical models for a given set of data. It balances the goodness of fit of the model with the number of parameters used. The lower the AIC, the better the model. Bayesian Information Criterion (BIC): The BIC is similar to the AIC but places a greater penalty on the number of parameters used in the model. The BIC tends to favour simpler models with fewer parameters. Finally, Hannan-Quinn Information Criterion (HQIC): The HQIC is another model selection criterion that balances the goodness of fit with the number of parameters used. It is similar to the AIC, but it places a greater penalty on the number of parameters than the AIC [4].

In Table 3 the AIC metric drops to lowest at lag 5, then continue with instability at lag 6 and then continuously drops further.

Table 3. Lags selection in vector autoregressive.

Lag	Original Series			First differential of the Series		
	AIC	BIC	HQIC	AIC	BIC	HQIC
Lag Order = 1	−41.0	−36.11	−39.02	−40.55	−35.63	−38.56
Lag Order = 2	−44.76	−35.27	−40.93	−43.24	−33.68	−39.38
Lag Order = 3	−46.21	−32.05	−40.5	−43.78	−29.52	−38.02
Lag Order = 4	−48.29	−29.41	−40.67	−44.92	−25.91	−37.25
Lag Order = 5	−54.73	−31.05	−45.18	−51.95	−28.12	−42.34
Lag Order = 7	−726.14	−692.7	−712.66	−843.13	−809.46	−829.56
Lag Order = 8	−740.28	−701.85	−724.8	−864.16	−825.47	−848.58
Lag Order = 9	−737.67	−694.19	−720.17	−866.59	−822.81	−848.97

Forecasting Measures

As in other similar papers, we use the most common metrics for CPI time series forecasting. The Root Mean Squared Error (RMSE), Mean absolute Error(MAE), Mean absolute Percentage Error(MAPE) among others metrics were report in Table 4. It's important to note that no single metric is universally better than the others, and the choice of metric depends on the specific problem being solved and the context in which the forecasting is being applied. For example, in some cases, minimizing the overall error (as measured by MSE) may be more important than accurately predicting individual values (as measured

Table 4. Report of the forecasting metrics applied to the performance of the VAR method.

	MAPE	ME	MAE	MPE	RMSE	CORR	MinMax
Overall	0.014	−0.718	1.493	−0.007	1.922	0.655	0.014
ABNA	0.037	−1.588	4.343	−0.012	5.476	0.646	0.037
BAT	0.008	0.407	0.783	0.004	2.175	−0.092	0.008
PVC	0.012	−0.281	0.979	−0.003	1.895	0.452	0.011
SV	0.007	−0.707	0.707	−0.007	1.074	0.878	0.007
MAH	0.002	0.043	0.174	0.0	0.368	0.123	0.002
S	0.005	0.562	0.568	0.005	0.695	NAN	0.005
T	0.003	−0.114	0.371	−0.001	0.432	0.924	0.003
C	0.028	−2.046	2.223	−0.026	2.568	0.319	0.028
RC	0.004	−0.041	0.39	−0.0	0.788	0.264	0.004
E	0.005	−0.498	0.524	−0.005	0.823	0.883	0.005
RH	0.006	−0.694	0.694	−0.006	1.207	0.576	0.006
BSD	0.002	0.132	0.229	0.001	0.311	0.752	0.002

by MAE or RMSE). Conversely, in other cases, accurately predicting individual values may be more important than minimizing overall error, such as in financial forecasting.

In general the overall performance of the multivariate Cuba CPI show very good adjust in the test set (the last two years) with the mean MAPE in the very low order of the 1.4% in the general CPI. The Fig. 3 we have the forecast results over test set considered in this analysis as a baseline for futures research.

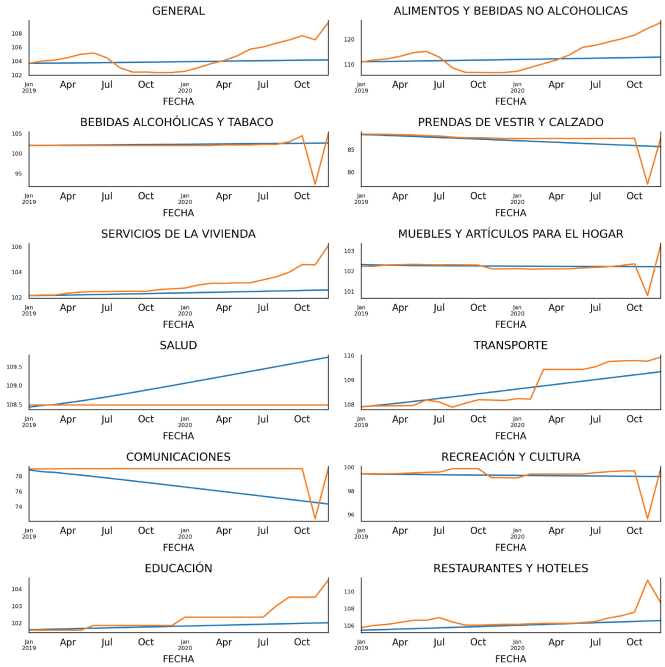

Fig. 3. Forecast over last two years of the Multivariate Cuba CPI.

4 Concluding Remarks and Further Work

A new dataset has been proposed for the CPI study in Cuba with a multivariate approach of which there are no references of previous researches in the field of forecasting. In this sense, this work has followed a standard statistical analysis methodology that has allowed establishing a baseline in terms of the VAR models, being these methods, reported in the literature as the starting point in the Multivariate CPI problem. Additionally, statistical tests for the study of the causality have been performed, showing in general strong relationships between the main components of the dataset. Likewise, the ADF test to study stationarity showed that the first differential of the series avoided stationarity with an 95% of the significance.

Future work is being planned in several directions with a view to extending this contribution. On the one hand, work is being done on an in-depth study of deep learning methods based on RNN and auto-encoder models to improve forecasting metrics and take advantage of the capabilities of the DL to handle non-linearity in the feature engineering. In another direction, variable selection should be exploited to achieve learning schemes with greater generalization.

Acknowledgement. This work has been partially funded by FONCI through project: Plataforma para el anélisis de grandes volúmenes de datos y su aplicación a sectores estratégicos.

References

1. Akin, A.C., Cevrimli, M.B., Arikan, M.S., Tekindal, M.A.: Determination of the causal relationship between beef prices and the consumer price index in turkey. Turk. J. Vet. Anim. Sci. **43**(3), 353–358 (2019)
2. Banerjee, A.: Forecasting price levels in India-an Arima framework. Acad. Mark. Stud. J. **25**(1), 1–15 (2021)
3. Cheung, Y.-W., Lai, K.S.: Lag order and critical values of the augmented dickey-fuller test. J. Bus. Econ. Stat. **13**(3), 277–280 (1995)
4. Cromwell, J.B.: Multivariate Tests for Time Series Models. Number 100. Sage (1994)
5. Diewert, W.E.: Index number issues in the consumer price index. J. Econ. Perspect. **12**(1), 47–58 (1998)
6. García Molina, J.M.: La economía cubana a inicios del siglo XXI: desafíos y oportunidades de la globalización. CEPAL (2005)
7. Ghazo, A., et al.: Applying the ARIMA model to the process of forecasting GDP and CPI in the Jordanian economy. Int. J. Financ. Res. **12**(3), 70 (2021)
8. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econom.: J. Econom. Soc. 424–438 (1969)
9. Jere, S., Banda, A., Chilyabanyama, R., Moyo, E., et al.: Modeling consumer price index in Zambia: a comparative study between multicointegration and ARIMA approach. Open J. Stat. **9**(02), 245 (2019)
10. Korkmaz, S., Abdullazade, M.: The causal relationship between unemployment and inflation in g6 countries. Adv. Econ. Bus. **8**(5), 303–309 (2020)
11. Anaya, L.M.L., Moreno, V.M.L., Aguirre, H.R.O., López, M.Q.: Predicción del ipc mexicano combinando modelos econométricos e inteligencia artificial. Rev. Mexicana Econ. Finanzas **13**(4), 603–629 (2018)
12. Mallick, L., Behera, S.R., Dash, D.P.: Does CPI granger cause WPI? Empirical evidence from threshold cointegration and spectral granger causality approach in India. J. Dev. Areas **54**(2) (2020)
13. Manik, D.P., et al.: A strategy to create daily consumer price index by using big data in statistics Indonesia. In: 2015 International Conference on Information Technology Systems and Innovation (ICITSI), pp. 1–5. IEEE (2015)
14. Mohamed, J.: Time series modeling and forecasting of Somaliland consumer price index: a comparison of ARIMA and regression with ARIMA errors. Am. J. Theor. Appl. Stat. **9**(4), 143–53 (2020)
15. Nakamura, E., Steinsson, J.: Identification in macroeconomics. J. Econ. Perspect. **32**(3), 59–86 (2018)

16. Nyoni, T.: Modeling and forecasting inflation in Kenya: Recent insights from ARIMA and GARCH analysis. Dimorian Rev. **5**(6), 16–40 (2018)
17. Nyoni, T.: ARIMA modeling and forecasting of consumer price index (CPI) in Germany (2019)
18. ONEI. Índice de precios al consumidor base diciembre 2010 (2022)
19. Qin, X., Sun, M., Dong, X., Zhang, Y.: Forecasting of china consumer price index based on EEMD and SVR method. In: 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), pp. 329–333. IEEE (2018)
20. Riofrío, J., Chang, O., Revelo-Fuelagán, E.J., Peluffo-Ordóñez, D.H.: Forecasting the consumer price index (CPI) of Ecuador: a comparative study of predictive models. Int. J. Adv. Sci. Eng. Inf. Technol. **10**(3), 1078–1084 (2020)
21. Rosado, R., Abreu, A.J., Arencibia, J.C., Gonzalez, H., Hernandez, Y.: Consumer price index forecasting based on univariate time series and a deep neural network. In: Hernández Heredia, Y., Milián Núñez, V., Ruiz Shulcloper, J. (eds.) IWAIPR 2021. LNCS, vol. 13055, pp. 33–42. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89691-1_4
22. Sünbül, E.: Linear and nonlinear relationship between real exchange rate, real interest rate and consumer price index: an empirical application for countries with different levels of development. Sci. Ann. Econ. Bus. **70**(1), 57–70 (2023)
23. Torres, J.F., Hadjout, D., Sebaa, A., Martinez-Alvarez, F., Troncoso, A.: Deep learning for time series forecasting: a survey. Big Data **9**(1), 3–21 (2021)
24. Triplett, J.: Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products (2004)
25. Zahara, S., Ilmiddaviq, M.B., et al.: Consumer price index prediction using long short term memory (LSTM) based cloud computing. J. Phys.: Conf. Ser. **1456**, 012022 (2020)