DOI: 10.18605/2175-7275/cereus.v12n4p146-159

<< Recebido em: 30/11/2020 Aceito em: 13/12/2020. >>

ARTIGO ORIGINAL

REVISTA CEREUS ISSN: 2175-7275

Basic Food Basket Monthly Price of Southern Bahia Cities: A Time Series Forecasting with Deep Learning Using a Recurrent Neural Network Approach

Preço Mensal Da Cesta Básica Em Cidades Do Sul Da Bahia: Predição da Série Temporal Utilizando Como Abordagem Uma Rede Neural Recorrente com Aprendizado Profundo

Murillo A. S. Torres[1], Mateus S. Marinho[2], Dany S. Dominguez[3], Dárcio R. Silva[4], Hélder Conceição Almeida[5]

[1] Graduando em Ciência da Computação, Universidade Estadual de Santa Cruz.

E-mail: mutorres.a@gmail.com

[2] Mestrando em Modelagem Computacional, Universidade Estadual de Santa Cruz.

[3] Doutor em Modelagem Computacional, docente da Universidade Estadual de Santa Cruz .

[4] Mestre em Modelagem Computacional, Universidade Estadual de Santa Cruz.

[5] Mestre em Modelagem Computacional, docente da Universidade Estadual de Santa Cruz.

RESUMO

O objetivo da análise de séries temporais é extrair informações não triviais de pontos organizados cronologicamente. A criação de dados de séries temporais não é uma tarefa difícil, porém o mesmo não se pode dizer de suas análises e previsões, sendo esses processos considerados em 2006 um dos dez principais desafios no campo de pesquisas em mineração de dados. A série temporal utilizada neste estudo foi uma série econômica com os preços mensais da cesta básica nas cidades de Ilhéus e Itabuna na região sul da Bahia. Ao analisar séries temporais, diferentes métodos podem ser aplicados, mas recentemente os métodos relacionados ao Aprendizado Profundo estão se tornando mais populares, sendo as redes neurais recorrentes (RNRs) as mais populares entre eles. Este trabalho utilizou uma arquitetura RNR em três experimentos distintos, prevendo em diferentes prazos os preços das séries temporais da cesta básica de 2019 e 2020. Todos os experimentos apresentaram resultado satisfatórios e sensibilidade ao comportamento das séries dos valores reais. Esses resultados sugerem que a arquitetura RNN pode generalizar bem as séries temporais para períodos ainda não vistos pela rede e que pode ter um desempenho ainda melhor em um ambiente mais abundante em dados.

Palavras-chave: Predição, Séries temporais econômicas, Aprendizado profundo, Rede neural recorrente.

ABSTRACT

The goal of time series analysis is extract non-trivial information from chronological sorted points. Time series data creation it is not a difficult task, however the same cannot be said about its analysis and predictions, with these processes being considered in 2006 one of the ten main challenges of data mining research field. The time series utilized in this research was an economic time series with the monthly prices of the basic food basket in Ilhéus and Itabuna cities. When analyzing time series, different methods can be applied, but recently the methods related to Deep Learning are becoming more popular, with recurrent neural networks (RNNs) being the most popular among them. This paper experimented an RNN architecture in three different experiments, predicting in different time terms the prices of the basic food basket time series of 2019 and 2020. All experiments presented a satisfactory prediction result and sensitivity to the real values series behavior. These results suggest that the RNN architecture can generalize well the time series to periods yet not seen by the network and that the NN can have an even better performance in a more data abundant environment.

Keywords: Forecasting, Economic time series, Deep learning, Recurrent neural network.

1. INTRODUCTION

Time series analysis is the endeavor of extracting meaningful summary and statistical information from points arranged in chronological order (NIELSEN, 2019). Most real-world data have a temporal component, whether it is measurements of natural processes (weather, sound waves) or man-made (stock market, robotics) (LÄNGKVIST; LOUTFI, 2014).

Thanks to this time component inherent to most real-world data, it is relatively easy to generate time series, however, the same cannot be said about its analysis, classification and prediction, processes which still present some challenges. The application of these processes to these series has already been considered one of the 10 challenging problems in Data Mining Research (YANG; WU, 2006).

Different areas have already detected the existent possibilities in use time series in their studies and practices, and it is not uncommon to find analyzes and predictions being performed by different sciences such as physics (Ludermir & Ferreira, 2008), medicine (Stell & Moss & Piper, 2012), astronomy (Hu et al., 2018) and economics (Wang et al., 2018). In addition to being used by different fields, the importance of time series data it is expected to grow rapidly in the coming years due to a massive production of such data from technologies like IoT (internet of things), healthcare digitalization and the rise of smart cities (NIELSEN, 2019).

A variety of methods can be applied in the analysis of these series, among those methods those that are commonly used by studies and research are: the Auto-Regressive Moving Average (ARMA) (MOURAUD, 2016), the Auto-Regressive Integrated Moving Average (ARIMA) (HAIGES et al., 2017), besides to hybrid methods that seek the combination of different methods for the development of a more robust model (SILVA; DOMINGUEZ; AMBRÓSIO, 2018).

Recently machine learning algorithms have started to become popular methods in the analysis and prediction of this data and, despite being methods that were not originally developed for time series specific data, they have proven to be useful for it (NIELSEN, 2019). Within this promising field, which is Machine Learning, there is the Deep Learning techniques and methods, which consist in the development of deep neural networks that can be understood as a (very) simplified model of human cerebral cortex, composed of a stack of layers of artificial neurons (GÉRON, 2019).

These Deep learning methods are capable of identifying structure and pattern of data such as non-linearity and complexity in time series (NAMIN; NAMIN, 2018), offering the possibility of modeling highly complex and nonlinear temporal behavior without having to guess at functional forms(NIELSEN, 2019). Today, 14 years after the rapid popularization of machine learning algorithms caused by a published paper by Hinton, Osindero and Teh (2006), there is a variety of deep learning algorithms available, with most of modern time series analysis problems being undertaken by recurrent neural networks (RNNs) (NIELSEN, 2019).

Some researchers have already applied the use of RNNs for problems of forecasting. Recent issues like energy consumption in buildings and its negative impacts was approached by Sehovac, Nesen & Grolinger (2019): their research proposes a new energy load methodology of forecasting by deep learning algorithms. In meteorological field, Jonnalagadda & Hashemi (2020) presented an interesting study of the prediction of atmospheric visibility conditions for safe transport using Auto Regressive RNN. In smart manufacturing tasks area, Shi et al. (2018) applied RNN deep learning technique for household load forecasting.

In economics field, McNally, Roche & Caton (2018) achieved good results for measuring the accuracy of forecasting the direction taken by BitCoin prices in USD using non-linear deep learning models. Wang et al. (2018) studied the application of Long Short-Term Memory (LSTM) algorithms in the optimization on stock price forecasting, in their research LSTM model reached more accurate prediction results than using backpropagation neural networks.

The time series utilized in this research was collected from the Acompanhamento do Custo da Cesta Básica project of the Universidade Estadual de Santa Cruz (State University of Santa Cruz), which has as one of its goals to record the monthly price of the basic food baskets in the region of Ilhéus and Itabuna cities, in Southern Bahia, Brazil (UESC, 2020). These data are collected and distributed to different government spheres to contribute to a more scientific public management using data to support the decision-making process.

Therefore, these time series are important resources in that region to help the regional governments identify existent social and economic needs in these cities, and also aid them on the processes related to address these needs. Using the series from these

two cities, this research aimed to develop a deep recurrent neural network capable of making predictions over these time series.

This article has the following structure: the first section gives a brief introduction about the subject addressed and presents the objectives that guided the research, the second section describes the methodology and materials used in the development of the applied deep learning method and the time series and its performed manipulations. The third section shows the results found while discuss them as they are presented, and finally the conclusions of this paper are presented at the fourth and final section.

2. MATERIALS AND METHODS

2.1 Recurrent Neural Networks (RNNs)

RNNs can be seen as very Deep networks with shared parameters at each layer when unfolded in time (LÄNGKVIST; LOUTFI, 2014). An RNN resemble in many aspects to a classic feedforward neural network, differentiating itself thanks to its structure that has reverse connections, from output to input.

Using this connection, an output of an RNN accumulate in one time step *t* the function of every previous time steps inputs, creating what can be considered a kind of memory, since part of the neural network state is preserved through time steps. (GÉRON, 2019)

Figure 1 displays the operation of a cell, also known as memory cell, present in recurrent neural networks through time. The *x(t)* represents the input on a time step *t*, *y(t)* represents the output in that same time, and *h(t)* is the preserved state of that cell, in time *t*, that will be also used as input in the next cell associating it with the *x(t+1)* input.

A deep recurrent neural network can be constructed using multiple layers of these memory cells, resulting in a deep RNN (GÉRON, 2019). These deep networks training is performed using a backpropagation through time (BPTT) approach. This approach, after the calculation of the gradient of the cost function applied to outputted sequences, propagates these values in a reverse path to the outputs, adjusting the model's parameters during this process.

The word "recurrent" is used to refer to these networks because they repeat the same task to very element in the sequence, with the characteristic of utilizing previously obtained information to predict future unseen sequential data (NAMIN; NAMIN, 2018).

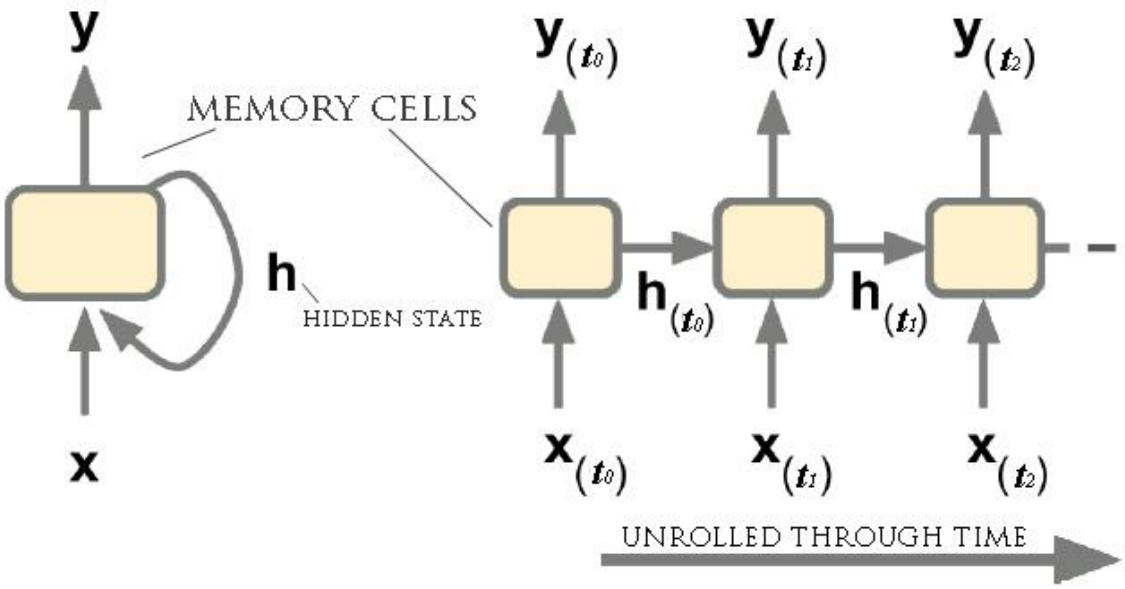

Figure 1. Operation of the Memory cells inside an RNN (Adapted from GÉRON, 2019).

2.2 The time series and data wrangling

The economic time series used in this paper are univariate and were collected by a project of the Universidade Estadual de Santa Cruz (State University of Santa Cruz) named *Acompanhamento do Custo da Cesta Básica* (UESC, 2020). These series are constituted by the price of the basic food basket each month in Ilhéus and Itabuna cities, located in Southern Bahia, Brazil. The registered values on the series represent the sum of the means of many products (UESC, 2020) which are commonly sold incorporated in a single product known as basic food basket. Figure 2 shows the behavior of both series through time.

Initially the series were loaded in two dataframes, a data structure created by Python's library Pandas. From these dataframes four training sets, out of the two time series, were created, two for each series. In each series, one of the training sets was compounded by the data points at the interval of January 2005 to December 2018, with a total of 168 data points. The other training sets of each series had data points at the interval of January 2005 to December 2019, having a total of 180 data points each.

After this initial splitting, the training sets were transformed in *k* series with 13 values each, where the values represent the months, in a way that each subsequent 13 values

series started and ended a month after the previous series until all datapoints stored in the initial training sets were utilized.

This transformation creates a set of 12 months series to be used as training set on the RNN model, and a set with the subsequent month to each of the 12 months series to be used as the training set label. At the end of the new size series creation, the training sets composed by datapoints until 2018 had 168 training instances and labels, while the other sets, with datapoints until 2019, had 180 instances.

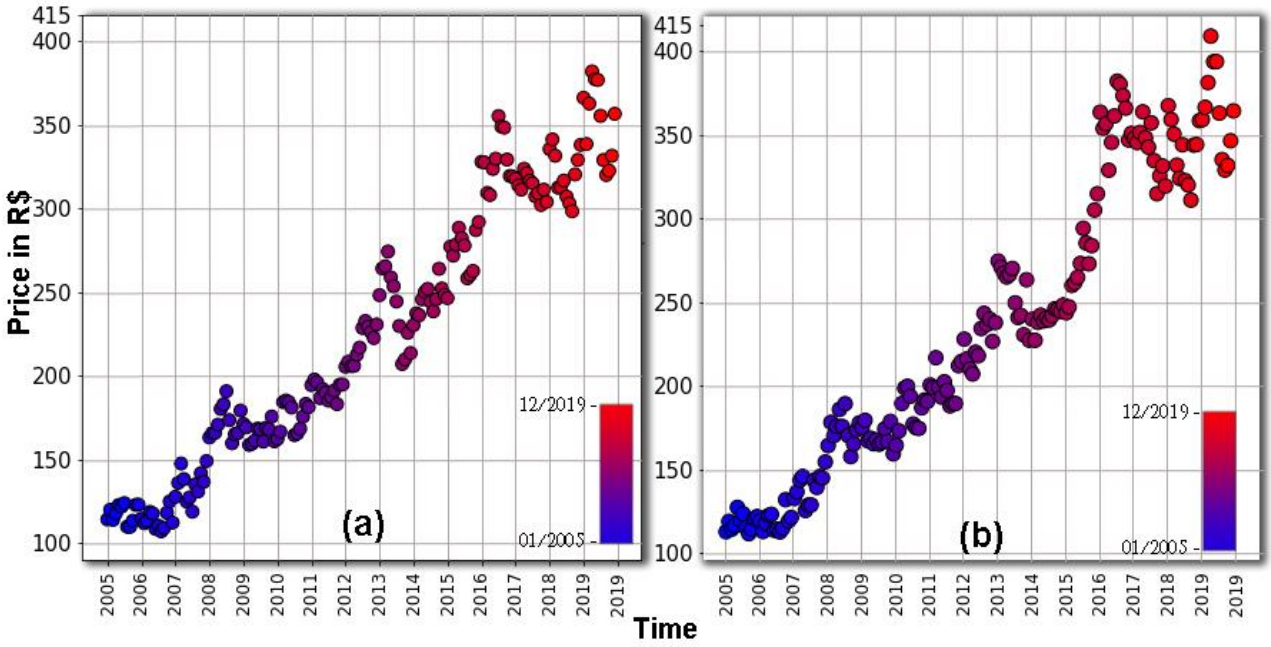

Figure 2. Itabuna's (a) and Ilhéus (b) basic food basket time series from January 2005 to December 2019.

2.3 The Recurrent Neural Network Architecture

The RNN was developed with TensorFlow library along Keras API. TensorFlow is a library developed by Google that allows training and execution of complex and large neural networks efficiently, while Keras is an API that can run on top of TensorFlow and allows the instantiation of neural networks using a model of sequential layers (GÉRON, 2019).

It was used 3 layers in the deep RNN: The top and hidden layers were declared utilizing the SimpleRNN layer available in Keras. As stated by Géron (2019), to find the number of neurons that best fit with the RNN architecture, experiments with different numbers of artificial neurons were performed. From the analysis of the neural network performance using these different numbers, it was found that 32 neurons provided the best results (Fig. 3). In order to allow the network process every sequence before resetting its

intern states, the stateful parameter was passed as true on both layers. (TENSORFLOW, 2020). Following Géron's (2019) guidance, to increase the model performance speed, maintaining basically the same precision, the output layer was declared as a Dense layer with a single neuron.

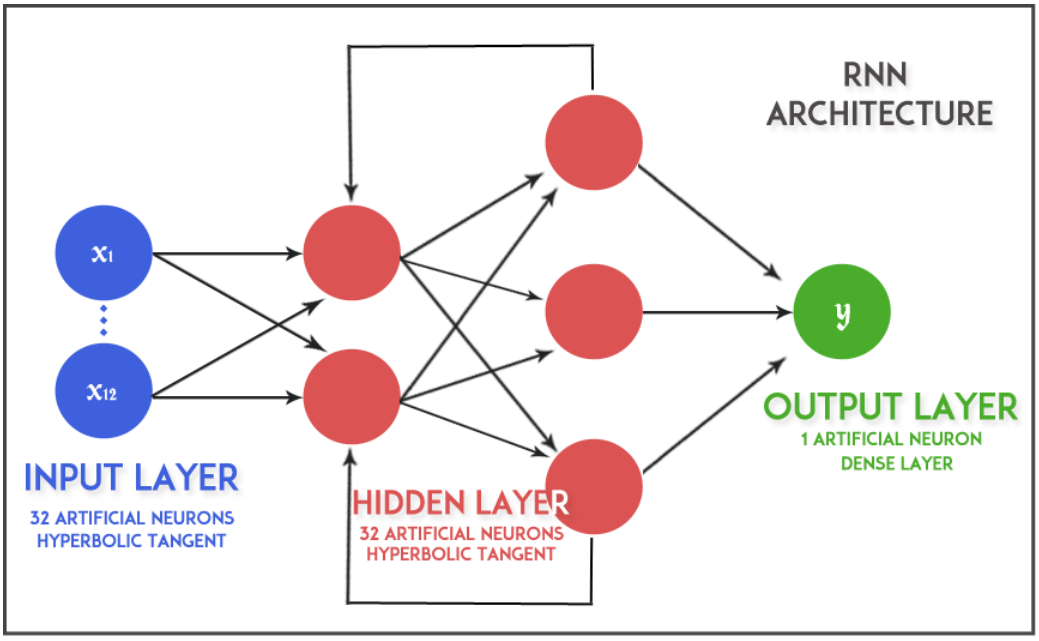

Figure 3. The proposed recurrent neural network architecture.

The activation function used was the hyperbolic tangent, standard activation function for recurrent neural networks using Keras (GÉRON, 2019). The optimizer function was the Adam's algorithm implementation, with a learning rate of 0.0003, while the cost function was the mean squared error function. On the model training, the parameter shuffle was set to false (NAMIN; NAMIN, 2018), and 20% of the training set was used to validation purposes. Beyond that, all the trainings were realized within 130 epochs.

2.4 The forecasting experiments

For each of the four training sets pre-processed, and using the previously described RNN architecture, the follow experiments were performed:

1. ***One-step forecast with model re-training using the actual values until a month before the prediction*.** This method allows assess the network

prediction capability in very short-term, allowing one month in advance planning.

2. ***Four-step forecast with mode re-training using the predicted values*.** This method allows assess the neural network prediction capacity in short-term, allowing four months in advance planning. In this experiment the model is trained until four months before the predictions, then, each month prediction is added to the model training set and then it is trained again before making the next prediction. This process is repeated until four months had been predicted. At the end of four predictions the model is then trained with the 4 actual values and then make predictions to the next four months.
3. ***12 and 8-steps forecast with retraining using only the predicted values***. First the model is trained until the last December before the predicted year. Then, it makes predictions for each month, using each prediction to train the model again before making the next prediction. The models trained with datapoints until 2018 predicted the 12 months of 2019, while the model trained with instances until 2019 predicted January to August 2020.

To measure the sensibility and accuracy of the predictions made by the recurrent neural network architecture used in all training sets and experiments, it was calculated the Root Mean Squared Error (RMSE, eq. 1), the Prediction Error (PE) of each experiment (eq. 2) and the Accuracy Mean (AM) of each experiment according to its city and year (eq. 3) using the expressions

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}{n}} \quad (1)$$

$$PE_i = Y_i - \hat{Y}_i \quad (2)$$

$$AM = \frac{\sum_{i=1}^{n} \frac{(Y_i - |PE_i|)}{Y_i}}{n} \times 100 \quad (3)$$

where $\hat{Y}_i$ represents the predicted values and Y_i the real values.

3. RESULTS AND DISCUSSION

The forecasting results for the experiments described in the previous section are offered below. Figure 4 shows the predictions performed with the training sets of Ilhéus (a) and Itabuna (b) in 2019 with the 3 proposed experiments. Figure 5 replicates the chart's structure of Figure 4 with the predictions performed with the 2020's training sets.

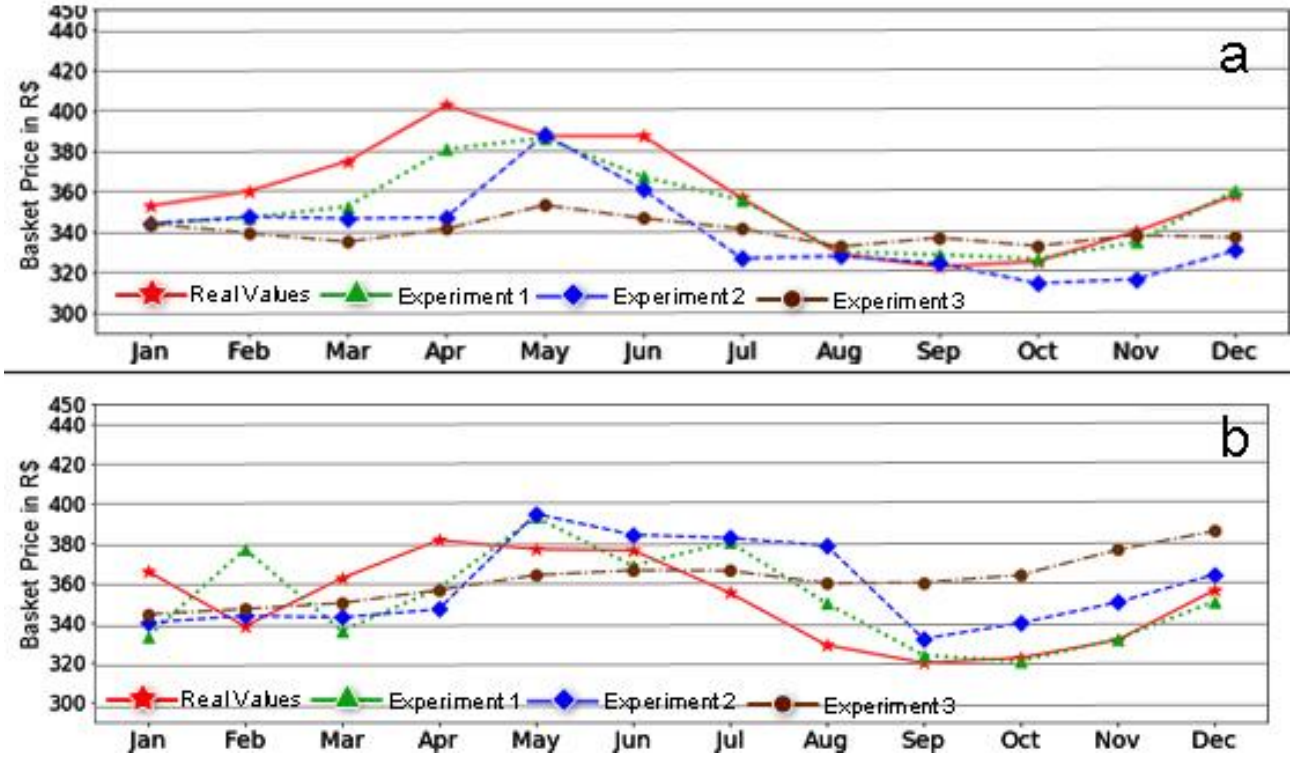

Figure 4. Ilhéus (a) and Itabuna's (b) 2019 economic time series forecast experiments results using the RNN architecture.

Except for Ilhéus 2020, experiment 1, from both figures, obtained the best predictions results for both cities. The red lines in these figures are the actual values of the basic food basket respectively in the Ilhéus (a) and Itabuna (b) cities from January to December 2019 (Fig. 4), and January to August 2020 (Fig. 5).

The lines in other colors represent the experiments according to the chart's legend. Experiment 1 was able to satisfactorily reproduce the behavior of both real values series, achieving the best accuracy averages in 2019 with 97.67% and 95.20% for Ilhéus and Itabuna, respectively. In 2020 the experiment 1 also achieved the best result for Itabuna, with an accuracy average of 97.7%, but for that same year the best results for Ilhéus was obtained by the experiment 2, with an accuracy average of 95.42% against 94.65% from the first experiment.

Although experiment 3 in Fig 5. graph b showed the lowest overall result for that year, it is possible to note that this experiment was sensitive to the trends of the real values series even though it was the only experiment that did not have any contact with the real values of that series, since it was carried out only with the values of its own predictions.

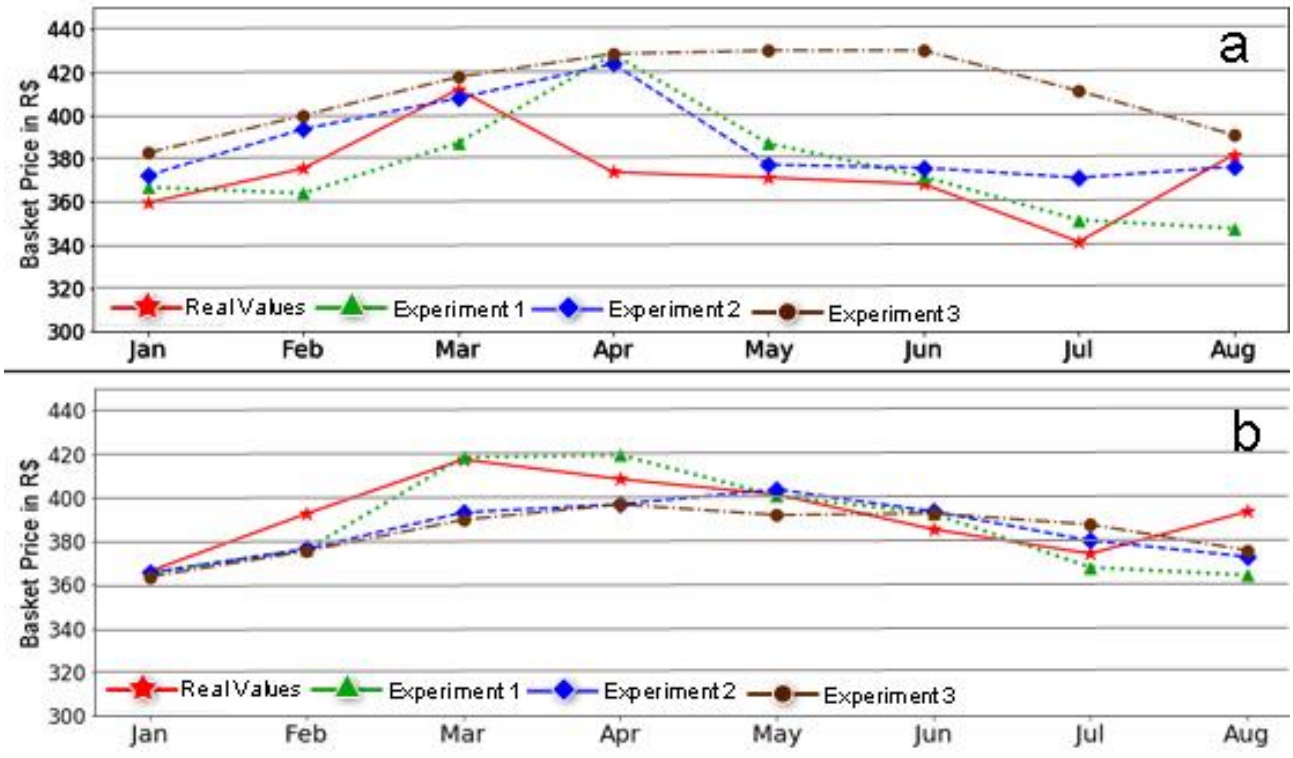

Figure 5. Ilhéus (a) and Itabuna's (b) 2020 economic time series forecast experiments results using the RNN architecture.

Table 1 shows the averages of the accuracies of all experimentations reported by this study and the results of the root mean squared error calculated for them. The 2020 experiments achieved a better accuracy mean performance than 2019, with a small difference in the RMSE average between these years. However, analyzing the cities

individually, Ilhéus 2019 had a better accuracy average performance than that of the same city in the following year. This better performance of the Ilhéus 2019 predictions is mainly related to the poor performance of the experiments 3, that showed the smallest results of all experiments, both in accuracy mean and RMSE.

The superior results of experiment 1 compared to the other experiments in both years (Table 1) were already expected since its predictions of a month ahead were carried out with the model trained with the real values until the previous month, which allows the network to abstract better the current behavior of the dataset and consequently generalize better. In relation to the other experiments, experiment 2 on average performed better than 3, behavior which, for the same reason as the superior performance of experiment 1, was also expected.

This tendency between the experiments results is replicated in the RMSE metric, as can be seen on table 1. The only exception for this tendency is the increasing on the performance of Ilhéus 2020 experiments 1 to 2. Nevertheless, this increasing is too small, less than 1% in accuracy mean and only 0.3 of difference in the RMSE, and could be caused by the random error in the prediction.

Table 1. Ilhéus and Itabuna's time series forecasting experiments accuracy means by city and year.

Year	Experiment	City	Accuracy Mean %	RMSE
2020	1	Ilhéus	94.65%	25.79
		Itabuna	97.70%	12.91
	2	Ilhéus	95.42%	25.49
		Itabuna	97.19%	13.79
	3	Ilhéus	89.42%	45.38
		Itabuna	96.64%	15.16
2019	1	Ilhéus	97.67%	11.90
		Itabuna	95.20%	21.11
	2	Ilhéus	94.88%	24.39
		Itabuna	94.15%	23.85
	3	Ilhéus	94.01%	28.31
		Itabuna	92.92%	27.44

4. CONCLUSION

The abstraction carried out by the recurrent neural network architecture proposed in this paper, was able to generalize well the reality of the economic time series, especially in very short-term tasks by predicting the values of the following month. This one-step prediction can be useful for issuing the project's monthly reports about the basic food basket price. Despite not having yet all the actual values for the last quadrimester of 2020, what prevents to assess the overall performance of this deep network this year, the results indicate that it will perform better in 2020 compared to 2019, mainly if the trends of the last four months repeat.

This possible improvement of the overall performance of the RNN in 2020 may be due to the amount of data used in the training sets, since the 2020 one also had the data points of 2019. Which may mean that this RNN architecture can achieve even better results in an environment with a greater data abundance. More research with the time series used here is underway in an attempt to develop a deep network capable of a greater overall performance, and even more sensitive to market volatility, especially in years like 2020 when global challenges, such as the current pandemic, affect and transform the economy of the countries, and particularly the undeveloped and developing economies.

REFERENCES

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow,** 2º ed. O'Reilly. Sebastopol, 2019.

HAIGES, R., WANG; Y.D. GHOSHRAY; A. ROSKILLY. A.P. Forecasting electricity generation capacity in Malaysia: An Auto Regressive Integrated Moving Average approach. **Energy Procedia,** vol. 105, p. 3471-3478, 2017.

HINTON, G.E.; OSINDERO, S.; TEH, Y-W**.** A Fast Learning Algorithm for Deep Belief Nets. **Neural Computation,** vol 18, p. 1527-1554, 2006.

HU, X.; YU, C.; LI, B.; TANG, S.; XIAO, J.; HUANG, Y. **GAIDR: An Efficient Time Series Subsets Retrieval Method for Geo-Distributed Astronomical Data**. IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), vol 1, p. 258-265, 2018.

JONNALAGADDA, J.; HASHEMI, M. **Forecasting Atmospheric Visibility Using Auto Regressive Recurrent Neural Network**. IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), vol 1, p. 209-215, 2020.

LÄNGKVIST, M.; KARLSSON, L.; LOUTFI, A. A review of unsupervised feature learning and deep learning for time-series modeling. **Pattern Recognition Letters**, vol 42, p. 11-24, 2014.

LUDERMIR, T.; FERREIRA, A**. Using Reservoir Computing for Forecasting Time Series: Brazilian Case Study**. Eighth International Conference on Hybrid Intelligent Systems, p. 602-607, 2008.

MCNALLY, S.; ROCHE, J.; CATON, S. **Predicting the Price of Bitcoin Using Machine Learning**. 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), vol 1, p. 339-343, 2018.

MOURAUD, A. Innovative Time Series Forecasting: Auto Regressive Moving Average vs Deep Networks. **Entrepreneurship and Sustainability Issues**, vol 4, p. 282-293, 2017.

NAMIN, S.S.; NAMIN, A.S. **Forecasting Economic and Financial Time Series: ARIMA vs. LSTM**. Texas Tech University. Lubbock, 2018. Retrieved from <https://arxiv.org/ftp/arxiv/papers/1803/1803.06386.pdf>.

NIELSEN, A. **Practical Time Series Analysis: Prediction with Statistics & Machine Learning**. 1º ed, O'Reilly, Sebastopol, 2019.

SEHOVAC, L.; NESEN, C.; GROLINGER, K. **Forecasting Building Energy Consumption with Deep Learning: A Sequence to Sequence Approach**. IEEE International Congress on Internet of Things (ICIOT), vol 1, p. 108-116, 2019.

SHI, H.; XU, M.; LI, R. **Deep Learning for Household Load Forecasting — A Novel Pooling Deep RNN**. IEEE TRANSACTIONS ON SMART GRID, vol 9, p. 5271-5280, 2018.

SILVA, D.R.; DOMINGUEZ, D.S.; AMBRÓSIO, P.E. **Métodos Híbridos de Redes Neurais na Previsão de Séries Temporais do Custo da Cesta Básica na Microrregião Ilhéus-Itabuna**. Anais do XX Encontro Nacional de Modelagem Computacional e VIII Encontro em Ciências e Tecnologias dos Materiais, Búzios, RJ, 2018.

STELL, A.; MOSS, L.; PIPER, I. **Knowledge-driven inference of Medical Interventions**. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, vol 1, p. 1-4, 2012.

TENSORFLOW. **Recurrent Neural Networks (RNN) with Keras**, 2020. Accessed on 17 September 2020. Retrieved from: <https://www.tensorflow.org/guide/keras/rnn>.

UNIVERSIDADE ESTADUAL DE SANTA CRUZ (UESC). **Projeto Acompanhamento do Custo da Cesta Básica (ACCB)**. Ilhéus, 2020. Accessed on 15 September 2020. Retrieved from <http://nbcgib.uesc.br/cesta/area_publica/index.php>.

WANG, Y.; LIU, Y.; WANG, M.; LIU, R. **LSTM Model Optimization on Stock Price Forecasting**. 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), vol 1, p. 173-177, 2018.

YANG, Q.; WU, X. 10 Challenging Problems in Data Mining Research. **International Journal of Information Technology & Decision Making,** vol 5, p. 597-604, 2006.