Predicting Food Prices in Iraq using Open Source Data

1st Salih Zainulabdeen
Department of Computer Science
University of Rochester
Rochester, NY, USA
szainula@u.rochester.edu

2nd Akira Sah
Department of Computer Science
University of Rochester
Rochester, NY, USA
asah@u.rochester.edu

3rd Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY, USA
jluo@cs.rochester.edu

***Abstract*—Food prices are crucial economic indicators for developing countries. However, in countries that have been affected by terrorist conflicts and corruption for the last 16 years, such price trends tend to be understudied. In this paper, we use time series techniques to study and analyze the food prices in Iraq from multiple angles. We create multiple models to predict food prices for selected commodities in selected governorates and report the results. We find that incorporating external factors, such as the number of terrorist attacks per month or the production quantity of each commodity in tonnes, which can be acquired from the released data on the web, leads to more accurate price predictions than univariate models. We hope this study would encourage more attention to countries like Iraq and help them prosper again.**

***Index Terms*—price predictions, Iraq, time-series analysis.**

I. Related Work

To enhance our prediction model and account for the external factors that might contribute to price fluctuations of commodities, we will incorporate ideas presented in previous data mining work. The first is the Flickr Background Model presented by Jin et al. (2010). The authors use the Flickr Background Model in their prediction model to account for the fluctuation of Flickr's popularity. We see a similar approach where Khatibi et al. (2019) use social media and climate data to predict touristic interest in attractions around the United States.

Moreover, the system design and implementation analysis presented by Asnhari et al. (2019) will guide our efforts into building an accurate prediction model. The paper provides valuable insight into how to combine a well-known time series model such as ARIMA (Autoregressive Integrated Moving Average) with linear regression and Fourier regression to make temporal predictions based on multiple factors. The paper concludes that when it comes to predicting the staple food prices in Indonesia with external factors like rainfall and oil price, Fourier regression provides more reliable predictions.

Finally, the article by Jin Hyun (2019) uses multiple methods for comparing two time-series data points and assessing the correlation between them. The important takeaway from this article is how to study and implement Time Lagged Cross-Correlation (TLCC); which can identify the leader-follower relationship between time series variables, find the optimal time lag (which maximizes the correlation) between them, and assess how correlated they are in that context using Pearson correlation. Moreover, TLCC is not biased by the scale of data. For example, $\vec{X} = \langle 0, 1, 2, 3, 4 \rangle$ would still be strongly correlated to $\vec{Y} = \langle 0, 0, 2, 4, 6 \rangle$ since TLCC can account for scales and time lags. Therefore, we plan to use TLCC to assess the correlation of an external factor to food prices, so that we can judge whether to include that factor in our prediction model and if so, use the optimal time lag between that factor and the food prices provided by TLCC.

II. Data Exploration

A. Datasets

1) Food Prices.: The food prices dataset was acquired from the Humanitarian Data Exchange. The dataset contains entries of monthly food prices for different commodities such as rice, wheat, etc. Those entries also contain information about the region (1 of the 18 Iraqi governorates) along with other attributes.

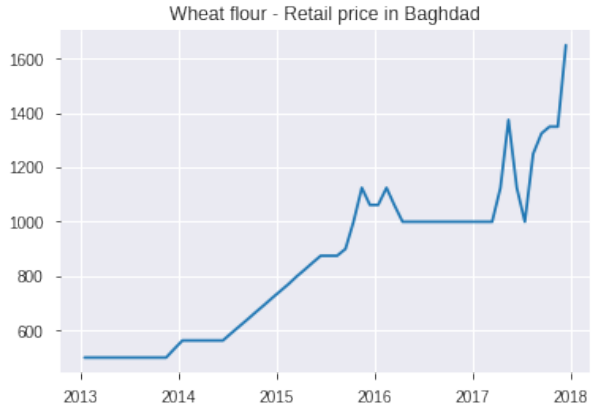


Fig. 1. Wheat flour retail prices in Baghdad from 2013-2018.

To preprocess the data, we started by analyzing missing values and inconsistencies. We found that some governantes in Iraq are missing years worth of data. For example, Nineveh governante was missing more than 50 data entries (over 4 years worth of data) due to the ISIS conflict that lasted over 3 years in Nineveh.

Another challenge that we encountered is that for some years, there are entries for over 20 commodities, and for others, there are entries for only 4 commodities. Therefore, we decided to focus on the 4 consistent commodities across the years, namely bread, rice, sugar, and wheat flour, which also happen to be the essential commodities in Iraq. Moreover, we decided to focus on the governantes with the least missing data and include Baghdad since it is the capital of Iraq. Those governantes are Baghdad, Karbala, Al-Qadisiya, and Diyala. Those governantes are all located in the central region of Iraq.

2) Conflict Data.: As we mentioned in the introduction, there are many external factors that we believe might affect food prices. One of the factors that we wanted to explore is the conflict factor. To provide context, there are many terrorist attacks and political conflicts that haven been taking place in Iraq during the last 16 years. Those terrorist attacks were recorded on the Global Terrorism Database which is maintained by the University of Maryland. Each entry represents a terrorist attack with its location (country, city, region), attack type, date, and many other attributes such as the weapons used. We limit our use of this data to the aggregate monthly number of terrorist incidents for each of the selected governorates, Baghdad, Karbala, Al-Qadisiya, and Diyala.

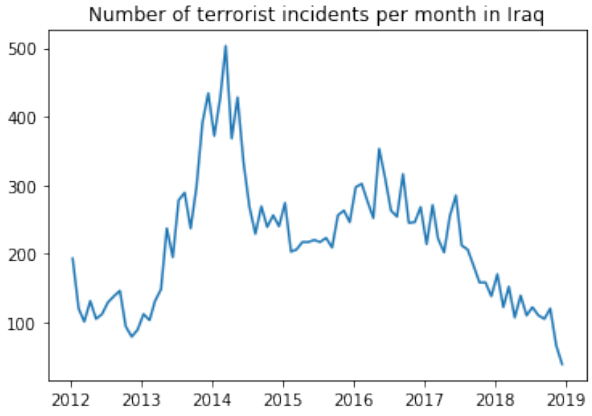


Fig. 2. Number of terrorist incidents per month in Iraq from 2012-2018.

3) Production Data.: Besides the conflict factor, we want to explore if the production quantity factor influences the food prices. For example, if the quantity of rice production (in tonnes) for 2015 was higher than in 2014, would that influence the price of rice in 2015? The Food and Agriculture Organization of the United Nations provides a dataset that contains the production quantity for each commodity in Iraq. We study the production quantity of barley (since bread is made out of barley in Iraq), rice, sugar, and wheat.

B. Correlation Analysis

Our goal is to use TLCC to evaluate the correlation between food prices and selected external factors. We define **Food Prices** = {price(c, g, t) : c ∈ **Selected Commodities** and g ∈ **Selected Governantes** and t ∈ {2013/1/15 - 2017/12/15}}. For example, price(rice, Baghdad, 2013/8/15) = 1850 Iraqi Dinar per Kilogram. We also define **Conflict Data** = {num_terrorist_incidents_per_month(g, t) : g ∈ **Selected Governante** and t ∈ {2013/1/15 - 2017/12/15}}. Lastly, we define **Production Quantity** = {production_quantity_per_month(c,t) : c ∈ **Selected Commodities** and t ∈ {2013/1/15 - 2017/12/15}}.

First, we analyze the correlation between each pair of **Conflict Data** × **Food prices**. After running this analysis on all of the pairs, only 5/16 of them portrayed significant correlation of over 0.50. Those were mostly in Baghdad, where the number of terrorist incidents per month seems to influence food prices the most. We believe this high correlation in Baghdad is due to the fact that the Baghdad has had significantly more terrorist incidents than any other governante in Iraq given it is the capital. Therefore, we only plan to use the number of terrorist incidents per month as an external factor in Baghdad only.

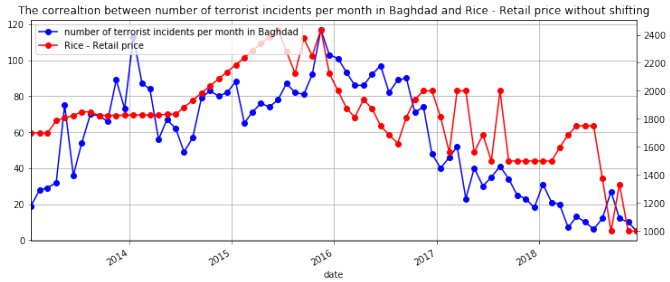


Fig. 3. The correlation between the number of terrorist attacks per month in Baghdad and rice prices.

Second, we analyze the correlation between each pair in the combinations of **Production Quantity** × **Food prices**. TLCC shows a significant negative correlation between most of the pairs, with all of the significantly correlated ones having a negative correlation between -0.6 to -0.9. Therefore, we believe food production data will contribute to the accuracy of our predictions in our multivariate models.

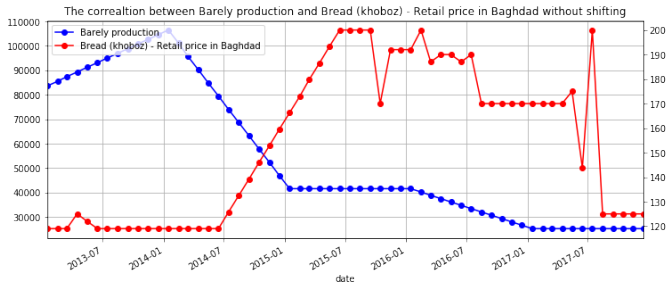


Fig. 4. The correlation between barley production and bread prices in Baghdad.

C. Seasonality Analysis

Food prices, in general, might exhibit seasonality due to the seasonal production of certain commodities and other factors. Therefore, we want to explore if the food prices for the data we have selected exhibit any seasonality. To that end, we generated graphs for each tuple in Food Prices for every year. We discovered that most of the prices do not exhibit seasonality. As a result, we will not use seasonal models such as SARIMA (Seasonal Autoregressive Integrated Moving Average).

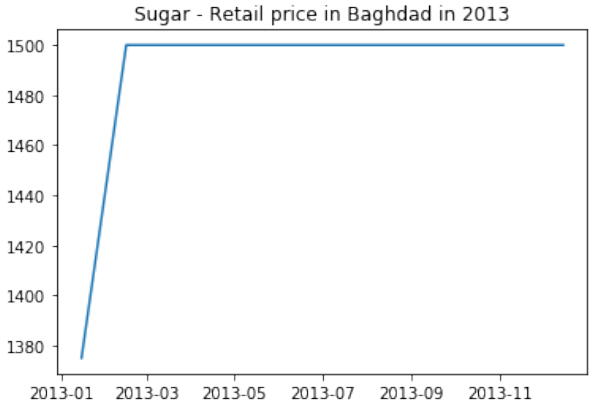


Fig. 5. Sugar - retail price in Baghdad in 2013, an example of no seasonality.

III. Methodology

In this study, we train and test 3 prediction models: ARIMA, VAR with **Conflict Data**, and VAR with **Production Quantity**. More details are included in the experiments section. We perform the following steps to train and test each model:

1) Split the data into training and testing sets, with a ratio of 2:1.
2) Train the model using the training set, pick the best parameters for the given model using respective methods.
3) Forecast the food price for the next period, then append the actual price to the training set, and re-train the model.
4) Continue forecasting in step 3's fashion until the last period in the test set.
5) Plot the actual and predicted prices for the test period.
6) Aggregate and plot the percentage errors to evaluate the accuracy of the model.

IV. Experiments and Results

We would like to note that we say a model is accurate if the majority of the predicted prices lie within 5% of the expected prices. With that on my mind, let us discuss our different models and their results.

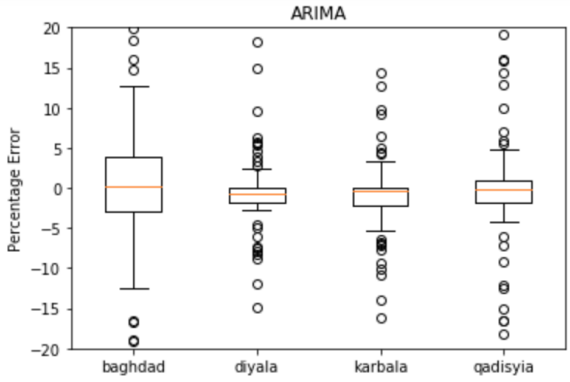
A. ARIMA

ARIMA (Autoregressive Integrated Moving Average) is one of the well-known algorithms for time-series analysis. ARIMA takes 3 parameters, **p**, number of lag orders to be included in the model, **d**, number of times differencing, and **q**, size the moving average window. We found that in our case, ARIMA predicts food prices most accurately when **d** = 1 and **q** = 0. In order to pick **p**, we had to analyze the autocorrelation plots of each tuple in **Food Prices**. Based on the autocorrelation analysis, we selected the **p** value presented in table 1.

TABLE I
ARIMA **P** values.

	Baghdad	Diyala	Karbala	Al-Qadisiya
Bread	7	11	7	4
Rice	6	3	9	5
Sugar	5	6	1	4
Wheat Flour	8	8	6	6

After selecting the parameters, we train our model using the method described above. Therefore, We discovered that our ARIMA models are especially accurate for Diyala, Karbala, and Al-Qadisiya. However, our ARIMA model was less accurate for Baghdad which indicates that we need to consider external factors in order to improve the predictions accuracy in Baghdad.



	Median |%| Error	Mean |%| Error
baghdad	3.516893	8.195354
diyala	1.578597	3.237119
karbala	1.749099	3.174152
qadisyia	1.600925	5.850724

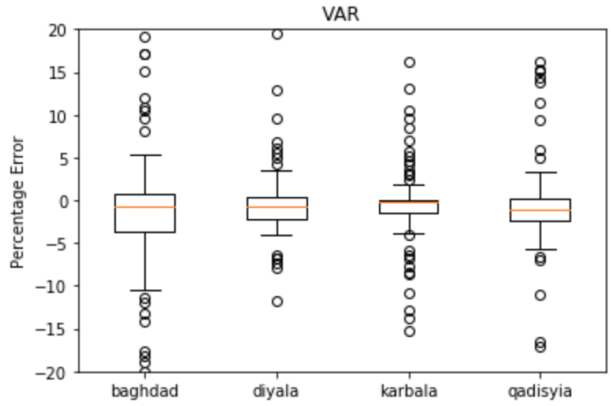
Fig. 6. ARIMA error plot.

B. VAR

Vector Autoregression (VAR) is one of the most common ways of training a multivariate prediction model. First, our training and testing data need to be preprocessed for use with this model since it only works with stationary time series. This was done by differencing each of the time series by itself. We then tested the stationarity of the data with using the Adfuller test, which indicated that all time series were stationary after one differencing with **p** less than 0.05. The one parameter of VAR in its statsmodels (Python package) implementation is **p**, which indicates the number of previous values in the input

matrix. We selected the best **p** by experimenting with a range of **p** values and selecting the best one using the AIC (Akaike's Information Criteria) metric results.

1) Using Production Quantity as an external factor with VAR: We train our VAR models with **Food Prices** × **Production Quantity** using the methodology described above. We notice that the predictions produced by VAR using **Production Quantity** as an external factor are relatively more accurate than the predictions produced by ARIMA. Figure 7 is the distribution of percentage error for every price prediction made by the models, grouped into cities. However, this model still suffers to meet our definition of accurate model as described above. This suggests that there are other external factors that are influencing food prices in Baghdad other than **Production Quantity**.



	Median |%| Error	Mean |%| Error
baghdad	2.260155	7.778216
diyala	1.852442	3.244097
karbala	1.359734	3.294857
qadisyia	1.606449	6.251480

Fig. 7. VAR error plot using **Production Quantity** as an external factor

2) Using Conflict Data as an external factor with VAR: We discussed above how there is a strong correlation between **Food Prices** and the number of terrorist incidents per month in Baghdad. As a result, we wanted to test if using the number of terrorist incidents per month in Baghdad as an external factor would improve our predictions of food prices. Indeed, our predictions in this case were improved over other models in almost all cases. Table 2 shows the improvement percentage of using **Conflict Data** in Baghdad as an external factor with VAR over other models.

V. Conclusion and Future Work

To predict food prices in Iraq, simple models such as ARIMA can provide relatively accurate predictions. However, using VAR with consideration of external factors such as **Production Quantity** will improve prediction accuracy for most of the general cases. However,in governorates like Baghdad where conflict rate is the highest in Iraq, using **Conflict Data** with VAR will noticeably improve the prediction accuracy over other models.

TABLE II
Improvement percentage of using **Conflict Data** as an external factor with VAR over other models for price prediction in Baghdad.

	ARIMA	VAR with Production Quantity
Bread	23%	1%
Rice	-4%	7%
Sugar	23%	20%
Wheat Flour	24%	10%

We plan to explore other external factors such as economic growth indicators, the presence of economic sanctions, and the food prices in neighboring countries such as Turkey. We believe exploring more related external factors will help us build more accurate prediction models for food prices in Iraq. We hope that by studying trends such as food prices, we will encourage a data science movement to study more understudied trends in countries like Iraq and help them prosper again.

References

Jin, X.; Gallagher, A.; Cao, L.; Luo, J.; and Han, J. 2010. The Wisdom of Social Multimedia: Using Flickr For Prediction and Forecast. *Proceedings of the 18th ACM International Conference on Multimedia* 1235-1244. ACM.

Khatibi, A.; Almeida, J.; Silva, A.; Belém, F.; Shasha, D.; and Gonçalves, M. 2018. Improving tourism prediction models using climate and social media data : A fine-grained approach. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* 636-639.

Asnhari, S. F.; Gunawan P. H.; and Rusmawati Y. 2019. Predicting Staple Food Materials Price Using Multivariables Factors (Regression and Fourier Models with ARIMA. *2019 7th International Conference on Information and Communication Technology (ICoICT)* 1-5.

Cheong, J. H. 2019. Four Ways to Quantify Synchrony between Time Series Data. *Towards Data Science* May 2019 archive.