Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Enhancing The ARIMAX Model By Using The Bivariate Wavelet Denoising: Application On Road Traffic Accidents Data

Nawroz M. Ahmed[1] and Qais M. Abdulqader [2]

[1]Department of Statistics College of Administration and Economic-University of Duhok, Duhok, Iraq,
[2]Department of Petroleum Geology–Technical College of Petroleum and Mineral Science-Duhok Polytechnic University, Duhok, Iraq

Article information

Article history:
Received March 2, 2023
Accepted April 25, 2023
Available online December 1, 2023

Keywords:
Traffic accidents
Excessive speed
ARIMAX
Time series
Bivariate wavelet

Correspondence:

Nawroz M. Ahmed
anawroz83@gmail.com

Abstract

The purpose of this study is to determine whether an enhanced confound model representing bivariate wavelet-autoregressive integrated moving average with exogenous variable BWARIMAX is beneficial for predicting monthly traffic accidents. A wavelet-based multiresolution analysis MRA, conducted before the ARIMAX model fitting, shows that the performance of ARIMAX models in predicting traffic accidents can be significantly improved. The method described in this study identifies the ideal wavelet function, wavelet transform, and number of decomposition levels for the MRA and consequently considerably improves forecast accuracy. The analysis of the study demonstrated the superiority of the suggested approach and revealed that utilizing the BWARIMAX method, we can extract more information from the series, which enhances the performance of the original ARIMAX model in terms of predicting. Additionally, it has been demonstrated through extensive empirical testing using a wide range of wavelet families that Daubechies and Coiflet wavelets are excellent choices for denoising data. Furthermore, the study concluded that out of the two wavelet families, the performance of the Coiflet wavelet of order 3 was better.

DOI: 10.33899/IQJOSS.2023.0181146 ,

1. Introduction

Predicting traffic accidents is an essential duty for traffic safety planners. Typically, these forecasts are useful in providing a better understanding of accident trends and current safety strategies. In other words, safety planners are interested in evaluating current policies and safety measures in order to take remedial action and analyze projected accident trends[1]. Before evaluating the series, it is crucial to remove the noise from the original data in order to better modeling and forecasting. The wavelet denoising approach, which uses a wavelet with a threshold, is a powerful mathematical tool for removing noise from the original data While preserving important features of the signal[2]. The use of wavelet analysis and time series analysis on traffic accidents data have seen a lot of implementations and proposals. Tsingotis and Vlahogranni [3]investigated the impact of meteorological data on the performance of short-term traffic forecasting models. A vector autoregressive moving average model with exogenous variables is used to examine the effects of weather and traffic mix on traffic speed prediction. The inclusion of exogenous variables increased predicting performance modestly, but vector and Bayesian estimation greatly improved the models. Junhui Zhang and Tuo Shi [4]used wavelet analysis and data from vehicle communication systems for the spatial study of traffic accidents to significantly increase the operational effectiveness of the traffic police department. Zuduo Zheng et al [5] used the Discrete Wavelet Transform DWT for the analysis of freeway traffic. The investigation demonstrates that the wavelet-based energy of individual cars may

successfully identify the sources of deceleration waves and offer information on potential causes (e.g., lane changing). In order to investigate a hybrid model for detecting traffic events, Shaura et al[6] used the wavelet transformation and logistic regression analysis. According to the findings, the method is a useful tool for detecting traffic incidents. Jabed Hossain [7] suggested a hybrid method that uses a wavelet transform to analyze a time-frequency (traffic count/hour) signal to identify acute variations in traffic flow. The improvement in accuracy for predicting long-term traffic flow was shown by the experiment results. For the purpose of forecasting short-term traffic flow, Hong Zhang et al[8]put up a brand-new hybrid technique with multivariate. In order to analyze the characteristics of traffic flow and predict the short-term state of traffic, this method combines statistical analysis with computational intelligence techniques that represent wavelet analysis and seasonal time series WSARIMA. Results of the study were encouraging. One-step ahead and ten-steps ahead forecasting accuracy improved using the newly proposed method. In order to anticipate the data on traffic accidents in Anambra State, Nigeria, Chukwutoo C. Ihueze and Uchendu O. Onwurah [9]used the ARIMA and ARIMAX models. Depending on certain statistical measures, the outcome showed that the ARIMAX model performed better than the ARIMA (1,1,1) model. To predict the price of crude oil, the author Taha Hussein Ali, and Mardin Samir Ali[10], used various linear dynamic systems, represented by ARIMA with exogenous input variables (ARIMAX models). The research's primary findings were that bivariate wavelet filtering was more effective than standard models at forecasting crude oil prices, and that using the suggested method, prices will be somewhat lower in 2020 than they were in 2019. Recently, many researchers applied the Box-Jenkins methodology to traffic accident data. In his work, Kidane Alemtsega Getahun [11]sought to apply the ARIMA approach to estimate the trend of injury, fatal, and overall traffic accidents in the Amhara region of Ethiopia from September 2013 to May 2017. It was discovered that the rate of traffic accidents in the Amhara region may be fitted by the ARIMA (2,0,0) (1,0,0) and ARIMA (2,0,0) (1,1,0)12 models. Katherina MeiBner and Julia Rieck [12]used multivariate forecasting, an extension of the ARIMA method, to predict how traffic accidents will develop over time in various geographical areas. The authors provided two additional approaches for segmenting accident data that allow the adaption of police tactics to regional features in order to identify geographical areas of interest Lunacek, Monte, et al [13] assessed different methods for traffic demand forecasting, which will help airport operations employees to accurately predict traffic and congestion. The investigation discovered that these algorithms are capable of capturing diurnal variations in surface traffic and all perform exceptionally well when anticipating the following 30 minutes of demand. Very recently, the authors Aram Nasser and Vilmos Simon[14] presented two new methods for weather-based traffic analysis and wavelet attention-based calculation. The two methodologies presented here were developed to study the temporal connections between traffic flow and exogenous meteorological elements at various frequencies and time intervals. In addition, to aid in understanding the significance of each external variable in comparison to the others. Soo-Yeon, Ji et al [15] introduced an approach for predicting future network events. Among the results of the analysis, Vector Auto- Regression with Exogenous variables VARX wavelet features can be used to analyze time series data from multivariate network traffic for the purposes of forecasting future network events and estimating their attack probability. Using ARIMA models, Vitalis Agati Ndume et al [16] examined the road traffic accidents patterns in Tanzania's selected regions. The study's main contribution to road safety was an estimation of road accident deaths by 2030. In this study, the implementation of bivariate wavelet analysis relying on discrete wavelet transform is denoted to investigate whether it can outperform bivariate ARIMA or not. for explanatory variable, we use the number of traffic accidents due to excessive speed as an indicator. Section 2 presents the methodology of bivariate time series analysis (ARIMAX model), and bivariate wavelet analysis. Section 3 deals with implementation and main results. In section 4, some conclusions are present.

2. A Brief Overview of Time Series Analysis and Wavelet Analysis

2.1 Time Series Analysis

A time series is a collection of observations, each collected at a specific time[17]. The time series data has time as the independent variable and the observed values as the dependent variables. A time series analysis that refers to any type of study employing time-series data is known as a univariate or single-series analysis. In other words, attempts to explain the behavior of data sets using only past observations on the variable of interest[18].

2.1.1 ARIMA Model

George Box and Jenkins (1976) proposed the ARIMA (p, d, q) model, which is made up of Autoregressive, Integration order, and Moving Average components. The first procedure to consider is (d), which refers to the order of the integration. To become stationary, the order of integration indicates the order of differencing to eliminate the unit root from a time series. To make a time series stationary in variance as required by the Box-Jenkins technique, we usually need to transform the series by the natural log or square root. We may refer the reader to [19][20]. The AR(p) process denotes the

Autoregressive process of order p, which defines the relationship between the target variable and the linear combination (regression) of its past values. The process can be represented as[21]:

$$y_t = C + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t \quad (1)$$

Where "ϕ" is called the autoregressive coefficient and "ε_t" is white noise. The next procedure is MA(q), which refers to a moving average process of order "q," with the variable of interest "y_t" being a function of current and multiple previous errors. The MA(q) process can be expressed as the following equation:

$$y_t = C + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (2)$$

Where θ is moving average coefficient. When these processes are combined or mixed, the ARIMA model is produced which represents the "Autoregressive Integrated Moving Average". The model is usually being shown as ARIMA (p, d, q):

$$y_t = C + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (3)$$

Backshift formulae can be used to write the ARIMA model, where B acting as a backward shift operator. It takes effect by moving the data back one period. ARIMA backshift notation (p, d, q):

$$\emptyset(\beta) y_t = C + \theta(\beta) \varepsilon_t \quad (4)$$

Where$\left(\emptyset(\beta) = 1 - \emptyset_1 \beta^1 - \cdots - \emptyset_p \beta^p\right) and, \left(\emptyset(\beta) = 1 - \theta_1 \beta^1 - \cdots - \theta_q \beta^q\right)$.

George Box and Jenkins (1970) proposed employing autocorrelation function (ACF) and partial autocorrelation function (PACF) for the initial identification of (p, q) orders (PACF). However, this strategy may only be applicable to certain patterns in time series. In this study, some specialized criteria employed, such as the Akaike's Information Criterion (AIC) for more rigorous recognition of (p, q) orders[22], and the Root Mean Square Error (RMSE) to assess the extent of forecast error in a forecast[23].

2.1.2 ARIMAX Model

Using the ARIMA model in conjunction with explanatory variables (ARIMAX) falls within the category of dynamic regressions, which includes a wide range of models, including conventional multiple regression models in which input factors have an instant effect on the output variable. In other words, the ARIMAX model represents the combination between linear regression and ARIMA process[24]. The ARIMAX model is configured in three steps: identification, parameter estimation and selection, and diagnostic checks. Identification entails deciding on data stationarity, parameter estimation and selection entails deciding on the order of AR and MA, and diagnostic check entails ensuring that no autocorrelation exists among the residuals[25].We can use formula (4) to express the ARIMAX model:

$$\emptyset(\beta) y_t = C + \beta X_t + \theta(\beta) \varepsilon_t \quad (5)$$

Here X_t is the input series or explanatory variable at time t, and y_t is the output series or dependent variable. The construction of such a model is an iterative process, though, and each step's output potentially violates the presumptions that must be made. The reference[26]has drawn the following conclusions on these presumptions:

1. The residual and the time-series used to construct the regression model both need to be stationary. If the residual is non-stationary, the original time-series must be further differentiated, and a new regression model must be created.
2. The residual of the final model conforms to the white noise hypothesis.
3. Each exogenous variable's coefficients in the final model need to be statistically significant. In some cases, some regression coefficients can lose their significance after creating the ARIMA model for the residual in the regression model. The least significant coefficient must then be eliminated, and all presumptions must be reexamined.
4. Only one-way causal relationships between exogenous and endogenous variables are possible; however, these relationships cannot exist between endogenous and exogenous variables (using Granger causality test). If the opposite causal relationship is found, the exogenous variable must be eliminated, and all presumptions must be reevaluated.
5. In the final model, the correlation coefficients between the exogenous variables and the dependent variable must have the same relationship (sign) with the regression coefficients (original timeseries).
6. In the completed model, there is no evidence of multicollinearity between exogenous variables.

2.2 Wavelet

A wavelet is indeed a small wave. A small wave grows and decomposes in a short amount of time. The opposing concept is clearly a "big wave." The sine function, which oscillates upwards and downwards on a plot of sin(u) versus $u \in (-\infty, +\infty)$, is an example of a big wave[27]. Figure1, shows the difference between sine wave and an example of Daubechies' wavelet[28].

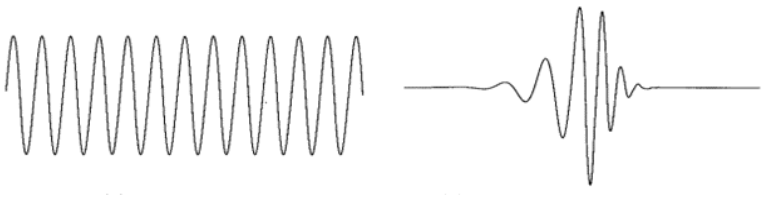

Figure 1. Sine wave from the left and Daubechies wavelet from the right

2.2.1 Wavelet Analysis

Wavelet analysis is increasingly being used to investigate localized fluctuations in power within a time series. By decomposing a time series into time-frequency space, one can determine the major patterns of variability as well as how those patterns change over time[29]. Wavelet decomposition is a redesigned short-time Fourier transform that displays decomposed signals in both the time domain and frequency domain using a time window - based function or the mother wavelet function[30]. For studying signals in the frequency domain, the Fourier transform is typically employed. However, the Fourier transform was unable to reproduce that behavior in nonlinear time-series that contain transients of short duration. A damped and long-duration vibration results from transforming a brief transient from time domain to frequency domain. A wavelet transformation is well known for its benefit in time-frequency localization. Wavelet analysis does not have this restriction, hence it works well with nonstationary time-series in contrast to Fourier transform which presumes the signal to be stationary[31].

The low frequency signal that is derived from the upper level is split into two components at each level of decomposition, the low frequency signal and the high frequency signal. The non-stationary signal can be effectively decomposed using the wavelet decomposition technique. Detail and trend components are included in the findings of the decomposition. The time-domain and frequency-domain properties of the signal are kept by expanding and translating the wavelet basis. The shape of the window is automatically flattened, and the window automatically becomes long and narrow with respect to high frequency signals[32]. Figure2 shows the process of wavelet decomposition of the original signal x.

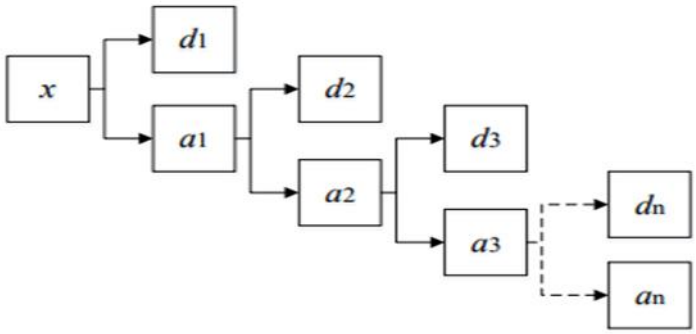

Figure 2. The algorithm of wavelet decomposition

Following wavelet decomposition, the signal x can be written as the sum of n detail parts $d_j(t)$ and one approximate part $a_n(t)$.

$$x_n = \sum_{j=1}^{n} d_j\,(t) + a_n(t) \tag{6}$$

For many practical purposes, multiple Wavelet types of transformations are employed, such as: Discrete Wavelet Transform (DWT), Continuous Wavelet Transform (CWT), Fast Wavelet Transform (FWT), Wavelet Packet Transform (WPT), and Stationary Wavelet Transform (SWT). The DWT transform is one of the most significant ones utilized in wavelets. Many different disciplines use this type of transform and its variations[33]. The DWT will be the subject of this paper because it is a fundamental tool for wavelet-based time series analysis.

The DWT differs from CWT in that it only uses a select few scales as opposed to transforming data across all scales. Again, a mother wavelet is used to create discrete wavelets, but this time, scale and shift are applied in discrete increments.

In a multiresolution analytic MRA framework, the DWT connects "filter banks" in the discrete time domain and wavelets in the continuous time domain[34].

2.2.2 Building a Hybrid Model BWARIMAX Using Wavelet Denoising

The main goal of wavelet denoising is to separate the ideal signal components from the noisy signal, which requires an estimation of the noise level. The small coefficient that is assumed to be noise is added to the estimated noise level [35]. The three main processes of signal denoising based on DWT are signal decomposition, thresholding, and reconstruction of the signal, which is then practically reduced from noise [28]. Figure3. illustrates how the process of building the BWARIMAX model works depending on wavelet-based denoising

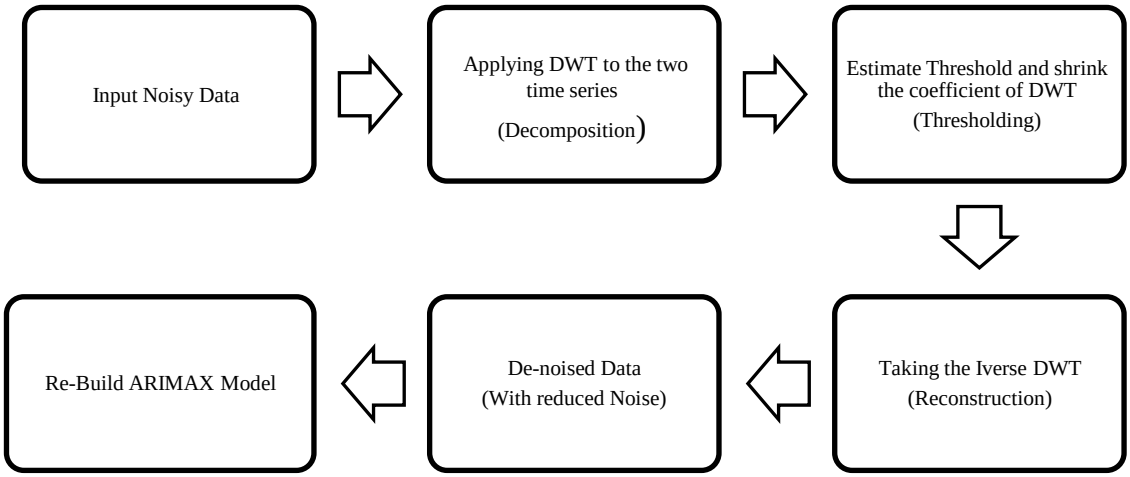

Figure3.The process of building the BWARIMAX model using wavelet-based denoising

2.2.3 Threshold Choice

The process of threshold selection is crucial since it has a direct impact on the quality of the output smoothed signal. There are several well-known techniques for estimating threshold levels. The performance of four popular criterion threshold estimating techniques—Fixed-Form, Minimax, Rigorous SHURE, and Heuristic SHURE—is examined in this research. For the stationary two datasets, three bivariate wavelet filters, including the Haar, Daubechies, and Coiflets wavelets, are applied. Additionally investigated is the effect of wavelet decomposition level. The following sources can be used by the reader to get more information about the four different threshold types and their equations[36][37]. The wavelet coefficients of a specific level would either have a hard threshold or a soft threshold depending on how the threshold of that level was estimated. Hard thresholding is the practice of setting all elements with absolute values below the threshold to zero. By first setting the elements whose absolute values are below the threshold to zero and then reducing the nonzero coefficients to zero, soft thresholding is an enlargement of hard thresholding. Discontinuities are produced by the hard technique but not by the soft procedure[38]. It was determined that the variance of the soft threshold is lower than that of the hard threshold[39]. In this study, we only concentrate on soft thresholding to improve the ARIMAX model of the data. All denoising algorithms and thresholding fundamentals are evaluated for their effectiveness in denoising the traffic accident data signals using two performance indicators, the Root Mean Squared Error (RMSE) and Akaike's Information Criterion (AIC). The following expression can be used to describe the two metrics, respectively[40].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x - x_e)^2}{n - k}} \tag{7}$$

$$AIC = Ln\sigma_e^2 + \frac{2k}{n} \tag{8}$$

Where n stands for the signal's length or sample size, x is the actual signal, x_e is the calculated signal derived from the denoised wavelet coefficients, and k is the amount of estimated model parameters.

3. Application and Main Results

3.1 Data Description

The study data includes a monthly time series of (84) observations representing the number of road accidents and the excessive speed in the Kurdistan Region of Iraq during the period starting from January 2015 to December 2021. These data have been obtained from the General Traffic Directorate of Erbil Governorate. In order of simplicity, the researcher will symbolize the variable number of accidents with the symbol (Y) as the dependent variable or the internal variable and the excessive speed variable with the symbol (X) as an independent variable or an external variable.

It is clear from table1, the statistical measures of the time series, where it is noted that the number of traffic accidents in the Kurdistan Region of Iraq during the study period ranged between (87) and (475) accidents, with a mean of (309.700) and a standard deviation of (104.329). Excessive speed ranged between (29) and (318) cases with an arithmetic average of (171.200) and a standard deviation of (68.943). Figure4 presents the time series of the study variables at their level:

Table 1. Arithmetic description of time series data

Time series	Minimum value	Maximum Value	Arithmetic mean	Standard deviation
No of accidents Y	87	475	309.700	104.329
Excessive speed X	29	318	171.200	68.943

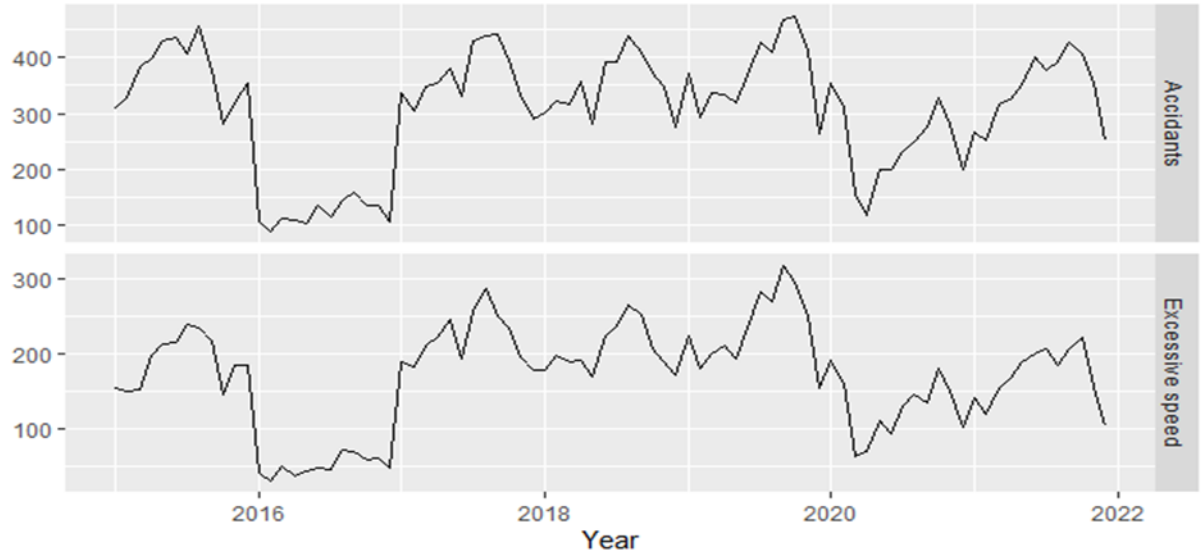

Figure 4. Monthly time series for the variables Y and X

3.2 Stationarity Test

The KPSS test (Kwait kowski-Phillips-Schmidt-Shin) is one of the special tests used to determine whether or not the time series is stationary. This test determines if the time series is stationary around an average or linear trend or whether it is not stationary because of the unit root [41]. The outcomes of the KPSS stationarity test are displayed in Table 2.

Table 2. KPSS test results

Critical values for a significant level of		10pct	5pct	2.5pct	1pct
Critical values		0.347	0.463	0.574	0.739
Time series	Value of test statistic	After the first difference			
Y	0.155	No need			
X	0.206	No need			

It is clear from table2 that the results of the KPSS test indicate that the time series of the number of road accidents is static at all levels, as the value of the test statistic is (0.155), which is less than all the given critical values. Also, from the other side, the result of the KPSS test indicates the static of the excessive speed time series, as the value of the test statistic KPSS is equal to (0.206), which is less than all the given critical values. The two series are stationary and do not need to take any differences.

3.3 Building ARIMA Model

In order to construct a suitable ARIMA model, the following steps should be followed:

1.Model Recognition Stage: For the purpose of seeking the stationarity of the time-series, the two functions representing autocorrelation ACF and partial autocorrelation PACF have been employed and relying on them to specify the order of the ARIMA models regarding each time series. as shown in figures 5,6 and table3. Selected alternatives for ARIMA models and the statistical criteria used to determine the best model. It is clear from table3, that the first stationary time series for the number of road accidents takes the model ARIMA (1,0,0) out of a total of 42 possible models, especially in comparison with the three closest alternative models. Also, the second stationary series of excessive speed takes the model ARIMA (1,0,0) out of a total of 42 possible models, especially when compared with the three nearest alternative models.

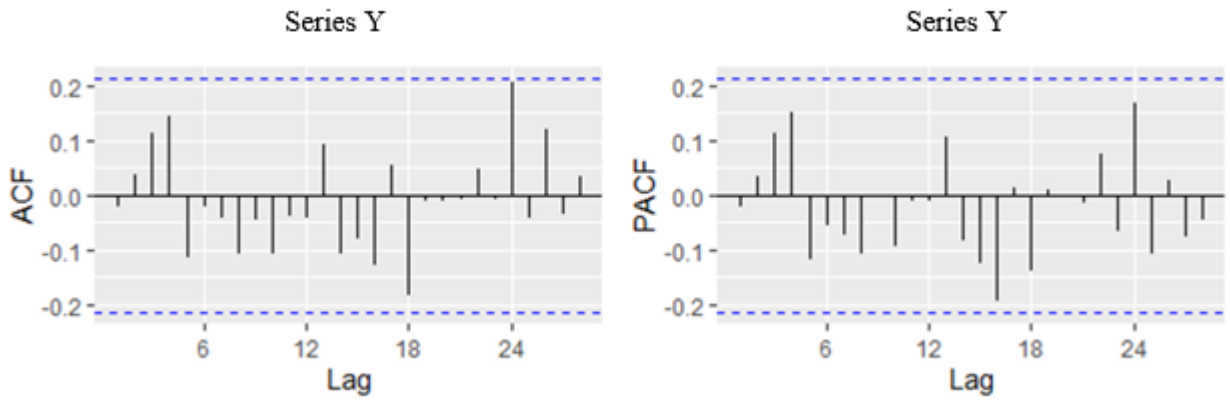

Figure 5. Time Series(Y) Interpretation of ACF and PACF Plots

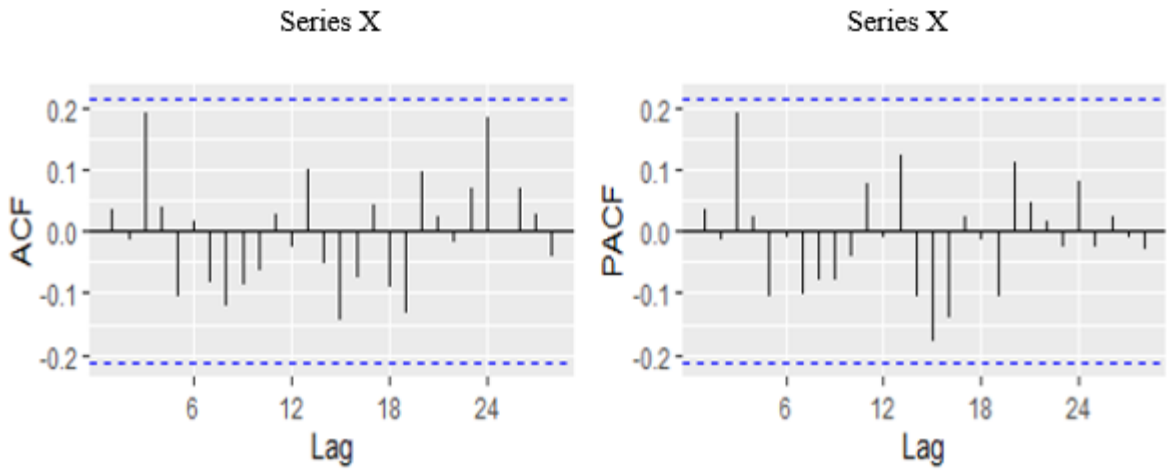

Figure 6. Time Series(X) Interpretation of ACF and PACF Plots

Table 3. presents the model chosen for the two-time series in comparison with the three closest alternative models

Series	Model	Criteria	
		AIC	RMSE
Y	**ARIMAX (1,0,0)**	**935.11**	**60.660**
	ARIMAX (0,0,1)	974.18	76.791
	ARIMAX (0,0,2)	955.58	67.809
	ARIMAX (0,0,3)	952.05	65.635
X	**ARIMAX (1,0,0)**	**855.36**	**37.707**
	ARIMAX (0,0,1)	895.79	48.105
	ARIMAX (0,0,2)	879.43	43.077
	ARIMAX (0,0,3)	871.15	40.50

2. Models Estimation Stage: The proposed ARIMA models were estimated based on the ACF and PACF functions of the stationary time series. The comparison is done for the possible alternatives of the different models for each series independently to get the best model. Table4 shows the best two models ARIMA (1,0,0) and ARIMA (1,0,0) with their estimated parameters. All parameters are significant at the 0.05 level.

Table 4. Results of estimating the best model for the variables Y and X

Time Series	Selected Model	Parameters	Standard Error	Z-Value	Pr(>|Z|)
No. road Accidents Y	ARIMA (1,0,0)	AR1=0.803	0.062	12.904	<2.2e-16***
		Mean=307.139	32.161	9.550	<2.2e-16***
Excessive Speed X	ARIMA (1,0,0)	Mean= 166.798	23.026	7.244	4.364e-13***
		AR1= 0.830	0.059	14.195	<2.2e-16***

3. Model Diagnosis Stage: By testing the residuals of the X and Y time-series and the ACF, PACF functions through figures7 and 8, it is noted that the standard residuals are randomly scattered around the zero line, and most autocorrelation

values of the ACF and PACF are within confidence limits or critical lines. Besides this, the Ljung_Box test value for the number of road accidents series is (Q*(Y) =12.139) with (p_value=0.669), and the Ljung_Box test for the series excessive speed is (Q(X)=12.132) with (p_value=0.669), and these values are greater than the significance level of 0.05. These results confirm the quality and suitability of estimated ARIMA models for time series data.

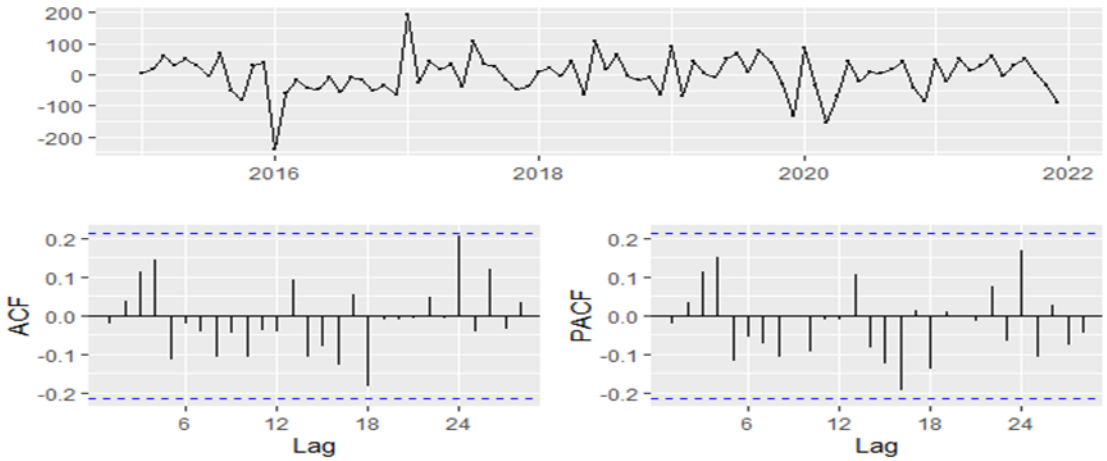

Figure 7. Standard residuals and ACF, PACF functions for the variable Y

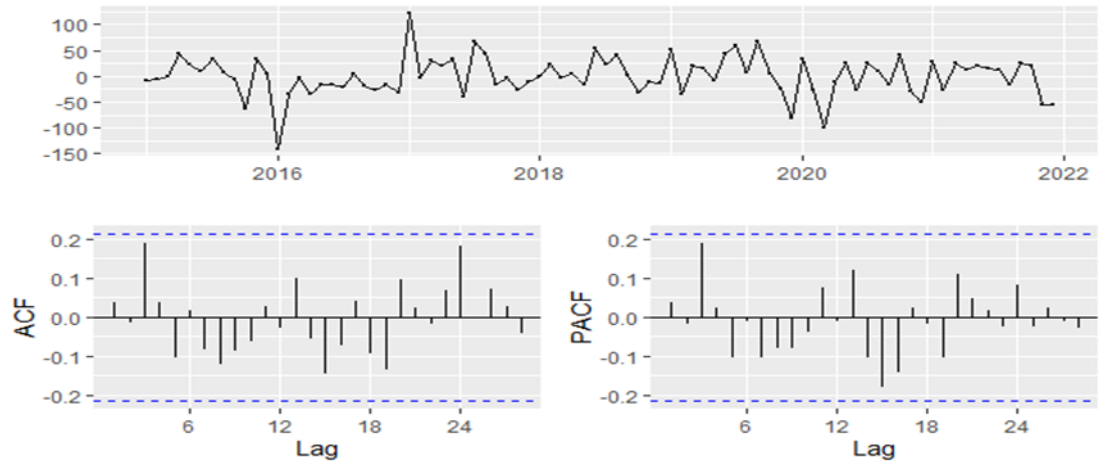

Figure 8. Standard residuals and the PACF, ACF functions for the variable X

3.4 Building ARIMAX Model

As mentioned in the theoretical section of this paper, the ARIMAX models represent an extended version of the ARIMA model that contains other independent (predictive) exogenous variables. ARIMAX models are similar to multiple regression models except that allow taking advantage of the autocorrelation that may be present in the regression residuals. This process is done to improve prediction accuracy after obtaining the ARIMA model for the number of road accidents Y, as well as the ARIMA model for the excessive speed X. Figure9 shows the errors resulting from the regression with the errors resulting from the ARIMA model.

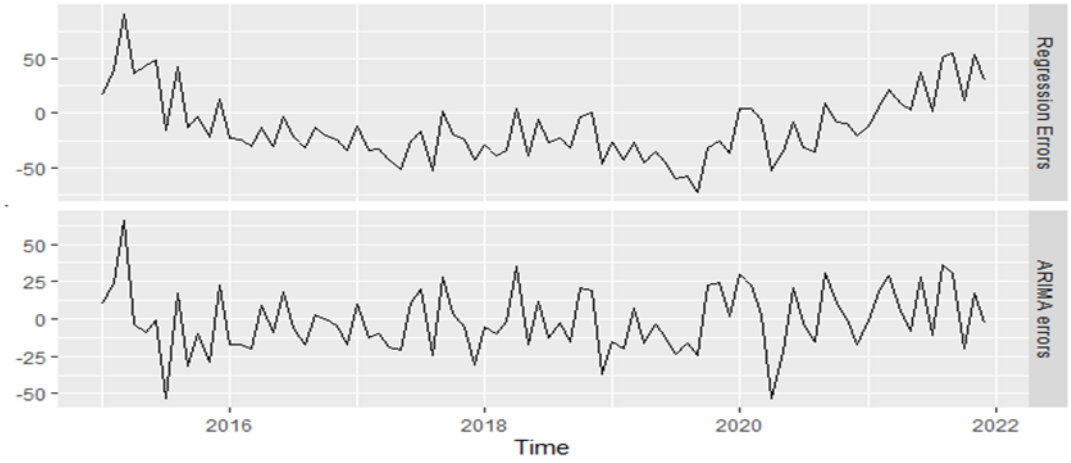

Figure 9. The combined regression errors and ARIMA model errors for the variables Y and X

Now it is required to find a suitable model to predict the number of monthly accidents with the presence of an external variable (excessive speed). The researcher used the statistical program (R) to build this model. Out of a total of (42) possible models, based on the AIC statistical criterion's lowest value, the ARIMAX (3,0,0) model was chosen. which is equal to (762.75), and the root mean square error criterion RMSE, which is equal to (20.941). Table5. shows the parameter values and their standard errors for the estimated model:

Table 5. ARIMAX (3,0,0) Model Summary

Parameter	Estimate	Standard Error	Z—Value	Pr(>|Z|)
AR (1)	0.242	0.107	2.269	0.023*
AR (2)	0.355	0.103	3.461	0.0005***
AR (3)	0.281	0.116	2.418	0.016*
Mean	68.428	19.272	3.551	0.0004***
X reg	1.482	0.050	29.616	<2.2e-16***

of the estimated Table5 clearly shows the statistical significance of parameters for both parts (ARIMA and regression) which supports the strength of this model. Figure10 shows the residuals and the location of the values of the autocorrelation coefficients.

Figure 10. Residual from regression with ARIMA (3,0,0) error

Figure10 shows that the residuals fluctuate around the zero line. Also, the all values of the autocorrelation coefficients fall within the confidence interval. Furthermore, most of the residual's values fall within the standard curve. Beside this, the value of the Ljung_Box statistical test is equal to (Q*=6.747) with a degree of freedom of (13) and the (P_value=0.874). which is another indicator of the randomness of the time series and the efficiency of this model. The equation of the ARIMAX model can be presented as the following:

$$Y_t = 1.482X_t + \eta_t \tag{9}$$

$$\eta_t = 0.242\eta_{t-1} + 0.355\eta_{t-2} + 0.281\eta_{t-3} + \varepsilon_t$$

$$\varepsilon_t \sim NID(0\ ,466.3)$$

3.5 ARIMAX Models Using Bivariate Wavelet Filter

After finding a suitable ARIMAX model for the number of monthly traffic accidents data in the presence of the exogenous variable (excessive speed) the researcher will try to use some of the wavelet filters as one of the proposed methods to reduce the impact of noise or contamination of the time series data using the program MATLAB. Forecasting will be carried out after reconstructing classically the ARIMAX model. Some wavelet filters were used such as: Haar, Daubeties, Coiflets, and at the same time some thresholding methods were used, including the Fixed Form threshold, Rigorous SURE, Heuristic SURE, and Minimax threshold with the use of soft and hard function. Figure11 shows the wavelet filter Coiftets level 3 with the Fixed Form threshold rule and soft threshold rule.

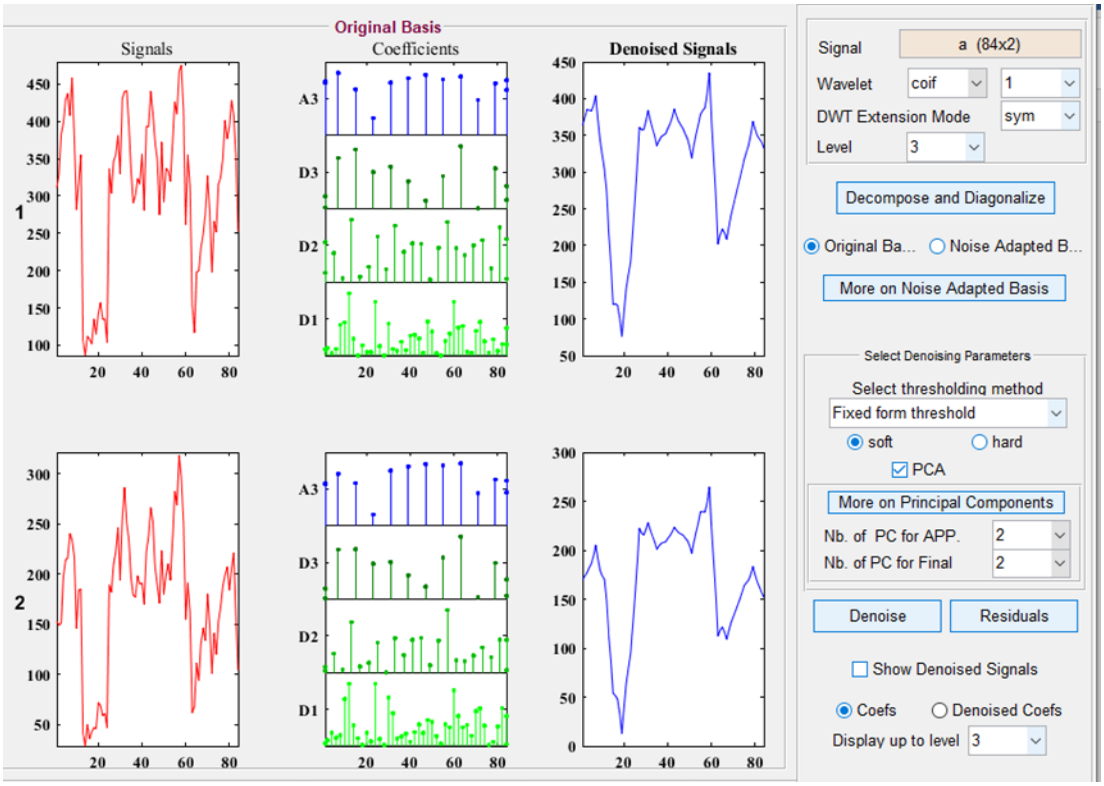

Figure 11. Coiflets level 3 bivariate wavelet with Fixed Form thresholding and soft rules for (Y, X) data

After the time series data have been converting from the time domain to the frequency domain. and using the wavelet filters, the noise reduction process was done by filtering. Then, re-converting the signal from the frequency domain to time domain, obtaining new filtered data, and reconstructing the ARIMAX model. Table6 shows the original model before filtering and a group of the best-selected models after using wavelet filters. Some thresholding methods are used according to the soft and hard threshold rule and the values of the statistical criteria are obtained.

Table 6. Comparison among the original model and the selected models after filtering for (Y, X) data.

Model	AIC	RMSE
ARIMAX (3,0,0) Original	762.75	20.941
ARIMAX (3,0,0) Using Fixed Form1 Coiflet 3_Soft	**591.38**	**7.451**
ARIMAX (3,0,0) Using Rigorous SURE Daubechies 2 Soft	654.03	10.880
ARIMAX (3,0,0) Using Heuristic SURE Coiflet 3 Soft	702.08	14.579
ARIMAX (3,0,0) Using Minimax Coiflet 3 Soft	706.63	14.988

Through table 6, we see that the values of the statistical criteria for the models that were built after filtering the data were mostly better than the original model and that the best case was through the BWARIMAX (3,0,0) model using Fixed Form thresholding method and depending on wavelet function Coiflit3 with a base soft rule, which has a minimum (AIC =591.38) and (RMSE =7.451). Table7 shows the values and std. error of the parameters for the chosen model

Table 7. BWARIMAX (3,0,0) Model Summary

Parameter	Estimate	Std Error	Z—value	Pr.(>|Z|)
AR1	1.691	0.107	15.790	<2.2e-16***
AR2	-1.175	0.189	-6.220	4.964e-10***
AR3	0.439	0.121	3.623	0.0003***
Mean	57.115	18.591	3.072	0.002**
XReg.	1.540	0.049	31.361	<2.2e-16***

The equation of the denoised ARIMAX model can be presented as the following:

$$Y_t = 1.540X_t + \eta_t \quad (10)$$
$$\eta_t = 1.691\eta_{t-1} - 1.175\eta_{t-2} + 0.439\eta_{t-3} + \varepsilon_t$$
$$\varepsilon_t \backsim NID(0\,, 59.03)$$

3.5.1 Forecasting Depending on the Denoised ARIMAX

This stage uses the estimated denoised ARIMAX model to predict new values for time series data. Table8 shows the forecasted values of the number of traffic accidents for the next twenty-four months from January 2022 to December 2023 depending on the denoised ARIMAX model.

Table 8. Forecasted values of the number of traffic accidents depending on BWARIMAX (3,0,0)

Month	Number of accidents	Month	Number of accidents
1	338	13	191
2	336	14	165
3	343	15	161
4	381	16	158
5	426	17	165
6	458	18	170
7	471	19	249
8	440	20	290
9	399	21	310
10	398	22	277
11	381	23	276
12	321	24	230

4. Conclusions

In this paper, the issue of the possibility of improving the traditional ARIMAX model based on the bivariate wavelet analysis method was mentioned. A proposed model was applied to the monthly data that represents the number of traffic accidents. Based on the results of the statistical analysis, the following can be concluded:

1- The hybrid model BWARIMAX (3,0,0) represents the optimal reliable model for predicting the number of traffic accidents.

2- The Fixed-Form threshold level was determined better than the other thresholds used in the analysis.

3- Among the wavelet family group used in the analysis, the Coiflet wavelet of order 3 gave the best result.

4- The soft rule threshold was better than the hard rule to obtain the best result.

5- Depending on the selected BWARIMAX (3,0,0) model, a slight decrease will happen in the number of traffic accidents in 2023 compared to 2022.

References

1. A. S. Al-Ghamdi, “Time Series Forecasts for Traffic Accidents, Injuries, and Fatalities in Saudi Arabia,” J. King Saud Univ. - Eng. Sci., vol. 7, no. 2, pp. 199–217, 1995, doi: 10.1016/S1018-3639(18)30627-5.
2. Q. Mustafa, “Comparing the Box - Jenkins models before . and after the wavelet filtering in terms of reducing the orders with application,” J. Concr. Appl. Math., vol. 11, pp. 190–198, 2013.
3. L. Tsirigotis, E. I. Vlahogianni, and M. G. Karlaftis, “Does Information on Weather Affect the Performance of Short-Term Traffic Forecasting Models?,” Int. J. Intell. Transp. Syst. Res., vol. 10, no. 1, pp. 1–10, 2012, doi: 10.1007/s13177-011-0037-x.
4. J. Zhang and T. Shi, “Spatial analysis of traffic accidents based on WaveCluster and vehicle communication system data,” Eurasip J. Wirel. Commun. Netw., vol. 2019, no. 1, 2019, doi: 10.1186/s13638-019-1450-0.
5. Z. Zheng, S. Ahn, D. Chen, and J. Laval, “Applications of wavelet transform for analysis of freeway traffic: Bottlenecks, transient traffic, and traffic oscillations,” Transp. Res. Part B Methodol., vol. 45, no. 2, pp. 372–384, 2011, doi: 10.1016/j.trb.2010.08.002.
6. S. Agarwal, P. Kachroo, and E. Regentova, “A hybrid model using logistic regression and wavelet transformation to detect traffic incidents,” IATSS Res., vol. 40, no. 1, pp. 56–63, 2016, doi: 10.1016/j.iatssr.2016.06.001.
7. J. Hossain, “A Hybrid Approach of Traffic Flow Prediction Using Wavelet Transform and Fuzzy Logic,” Electron. Theses Diss., pp. 1–62, 2017, [Online]. Available: https://scholar.uwindsor.ca/etd/5987.
8. H. Zhang, X. Wang, J. Cao, M. Tang, and Y. Guo, “A multivariate short-term traffic flow forecasting method based on wavelet analysis and seasonal time series,” Appl. Intell., vol. 48, no. 10, pp. 3827–3838, 2018, doi: 10.1007/s10489-018-1181-7.
9. C. C. Ihueze and U. O. Onwurah, “Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria,” Accid. Anal. Prev., vol. 112, no. November 2017, pp. 21–29, 2018, doi: 10.1016/j.aap.2017.12.016.

10. T. H. Ali and M. S. Ali, "Analysis of Some Linear Dynamic Systems with Bivariate Wavelets," Iraqi J. Stat. Sci., vol. 16, no. 3, pp. 85–108, 2019, doi: http://dx.doi.org/10.33899/iqjoss.2019.164176.
11. K. A. Getahun, "Time series modeling of road traffic accidents in Amhara Region," J. Big Data, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00493-z.
12. K. Meißner and J. Rieck, "Multivariate Forecasting of Road Accidents Based on Geographically Separated Data," Vietnam J. Comput. Sci., vol. 8, no. 3, pp. 433–454, 2021, doi: 10.1142/S2196888821500196.
13. M. Lunacek et al., "A data-driven operational model for traffic at the Dallas Fort Worth International Airport," J. Air Transp. Manag., vol. 94, no. June 2020, p. 102061, 2021, doi: 10.1016/j.jairtraman.2021.102061.
14. A. Nasser and V. Simon, "Wavelet-attention-based traffic prediction for smart cities," IET Smart Cities, vol. 4, no. 1, pp. 3–16, 2022, doi: 10.1049/smc2.12018.
15. J. Soo-Yeon, J. Bong Keun, K. Charles, L. Nandi, and H. J. Dong, "Forecasting network events to estimate attack risk: Integration of wavelet transform and vector auto regression with exogenous variables," J. Netw. Comput. Appl., vol. 203, no. 1, 2022, doi: https://doi.org/10.1016/j.jnca.2022.103392.
16. V. A. Ndume, E. C. Rutalebwa, and A. K. Runyoro, "Prediction of Road Accidents Trend in Tanzania Using ARIMA Model : The Road Safety Implication by 2021-2030," vol. 11, no. 1, pp. 1–7, 2022, doi: 10.5923/j.ijtte.20221101.01.
17. P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting - Second Edition. 2002.
18. A. PANKRATZ, Forecasting With Univariate Box- Jenkins Models: Concept and cases. John Wiley & Sons, 1983.
19. G. E. P. Box, G. M. Jenkins, R. G. C, and G. M. Ljung, Time series analysis: Forecasting and Control, Fifth edit. WILEY, 2015.
20. H. R. Makridakis S, Wheelwright SC, Forecasting methods and applications, Third. Jhon Wiley and Sons,INC., 1997.
21. G. K. Vishwakarma et al., "Forecasting steel prices using ARIMAX model: A case study of Turkey," Resources Policy, vol. 35, no. 8. pp. 297–310, 2020, [Online]. Available: https://www.statista.com/download/MTUzNDM2Mjk3MSMjMzAxOTMzIyM1MDM0OSMjMSMjcGRmIyNTdHVkeQ==%0Ahttps://doi.org/10.1016/j.eneco.2017.09.010%0Ahttps://revistas.unal.edu.co/index.php/ceconomia/article/view/73067%0Ahttps://doi.org/10.1016/j.resourpol.2019.10155.
22. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 2nd ed. Monash University, Australia, 2018.
23. R. A. Yaffee and M. McGee, Introduction to Time Series Analysis and Forecasting: with Applications of SAS and SPSS, 2nd ed. ACADEMIC PRESS, INC, 1999.
24. R. Amelia, D. Y. Dalimunthe, E. Kustiawan, and I. Sulistiana, "ARIMAX model for rainfall forecasting in Pangkalpinang, Indonesia," IOP Conf. Ser. Earth Environ. Sci., vol. 926, no. 1, 2021, doi: 10.1088/1755-1315/926/1/012034.
25. F. Islam and M. A. Imteaz, "The effectiveness of ARIMAX model for prediction of summer rainfall in northwest Western Australia," IOP Conf. Ser. Mater. Sci. Eng., vol. 1067, no. 1, p. 012037, 2021, doi: 10.1088/1757-899x/1067/1/012037.
26. B. H. Andrews, M. D. Dean, R. Swain, and C. Cole, "Building ARIMA and ARIMAX Models for Predicting Long-Term Disability Benefit Application Rates in the Public / Private Sectors Sponsored by Society of Actuaries Health Section Prepared by University of Southern Maine," Soc. Actuar., no. August, 2013.
27. D. B. Percival and A. T. Walden, "Wavelet Methods for Time SeriesAnalysis," Wavelet Methods for Time SeriesAnalysis. 2000, doi: 10.1017/cbo9780511841040.
28. C. S. Burrus, R. A. Gopinath, and H. Guo, Introduction to Wavelets and Wavelet Transforms: A Primer. 1998.
29. C. Torrence and G. P. Compo, "A Practical Guide to Wavelet Analysis," Bull. Am. Meteorol. Soc., vol. 79, no. 1, pp. 61–78, 1998, doi: 10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.
30. I. Shim, J. J. Soraghan, and W. H. Siew, "Detection of PD utilizing digital signal processing methods Part 3: open-loop noise reduction," IEEE Electr. Insul. Mag., vol. 17, no. 1, pp. 6–13, 2001, doi: 10.1109/57.901611.
31. W. Wongdhamma, "Upgrade from ARIMA to ARIMAX to Improve Forecasting Accuracy of Nonlinear Time-Series : Create Your Own Exogenous Variables Using Wavelet Analysis WAVELET TRANSFORMATION : BACKGROUND," Sas, vol. 3, no. 34, pp. 1–16, 2016.
32. Q. Xu et al., "Ultra-short-term wind speed forecast based on WD-ARIMAX-GARCH model," Proc. 2019 IEEE 2nd Int. Conf. Autom. Electron. Electr. Eng. AUTEEE 2019, pp. 219–222, 2019, doi: 10.1109/AUTEEE48671.2019.9033198.
33. G. Ramazan, S. Faruk, and W. Brandon, An introduction to wavelets and other filtering methods in finance and economics, 1st ed. Academic Press, 2002.
34. C. Charles K., An introduction to wavelets, 1st ed. Academic Press, 1992.

35. B. Ergen, “Signal and Image Denoising Using Wavelet Transform,” Adv. Wavelet Theory Their Appl. Eng. Phys. Technol., no. May, 2012, doi: 10.5772/36434.
36. I. Lo Cascio, “Wavelet Analysis and Denoising: New Tools for Economists,” no. 600, 2007.
37. P. Pallavi L, and B. Raskar, V., “Image Denoising With Wavelet Thresholding Method for Different Level of Decomposition,” Int. J. Eng. Res. Gen. Sci., vol. 3, no. 3, pp. 1092–1099, 2015.
38. M. Misiti, Y. Misiti, G. Oppenheim, and J. Poggi, Wavelet Toolbox For Use with M ATLAB. Wellesley Cambridge Press, 1996.
39. A. G. Bruce and H. Y. E. Gao, “Understanding WaveShrink: Variance and bias estimation,” Biometrika, vol. 83, no. 4, pp. 727–745, 1996, doi: 10.1093/biomet/83.4.727.
40. V. S. Chourasia and A. K. Tiwari, “Design Methodology of a New Wavelet Basis Function for Fetal Phonocardiographic Signals,” Sci. World J., vol. 2013, 2013.
41. D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, “Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root?,” J. Econom., vol. 54, no. 1–3, pp. 159–178, 1992, doi: 10.1016/0304-4076(92)90104-Y.

تحسين نموذج ARIMAX باستخدام تقليل الضوضاء المويجي ثنائي المتغير: التطبيق على بيانات حوادث المرور على الطرق

نوروز ميكائيل احمد/قسم الاقتصاد /كلية الادارة والاقتصاد/جامعة دهوك / العراق anawroz83@gmail.com

قيس مصطفى عبدالقادر / قسم جيولوجيا النفط / الكلية التقنية لعلوم النفط والمعادن/جامعة دهوك التقنية/ العراق

qais.mustafa@dpu.edu.krd

الخلاصة: تبحث هذه الدراسة في فائدة الطريقة المحسنة المدمجة والمتمثلة بنموذج (المويجة ثنائية المتغير–الإنحدار الذاتي والمتوسط المتحرك التكاملي بوجود متغير خارجي BWARIMAX) للتنبؤ بحوادث المرور الشهرية. تم إثبات زيادة أداء نماذج ARIMAX في التنبؤ بحوادث المرور على الطرق بشكل كبير من خلال التحليل متعدد التصاميم MRA القائم على المويجة. النهج الموضح في هذه الورقة يحدد أفضل نوع توليفة من تحويل المويجات ، والدوال المويجية ، وعدد مستويات التحليل المستخدمة في MRA وبالتالي زيادة دقة التنبؤ بشكل كبير. أظهر تحليل الدراسة تفوق الطريقة المقترحة وأشار إلى أنه يمكننا الحصول على مزيد من المعلومات من السلسلة عند استخدام نموذج BWARIMAX وهذا يؤدي إلى تحسين نموذج ARIMAX الأصلي في التنبؤ. إلى جانب ذلك ، بعد العديد من الاختبارات التجريبية مع العديد من العائلات المويجية ، تبين أن المويجتين Daubechies و Coiflet مناسبة جدًا عند تقليل الضوضاء من البيانات ، ومن بينهما كان أداء مويجة Coiflet من الرتبة 3 أفضل.

الكلمات المفتاحية : الحوادث المرورية ، السرعة الزائدة ، ARIMAX ، السلسلة الزمنية ، المويجة ثنائية المتغير.