Original Research Article

APPLICATION OF THE BOX AND JENKINS METHOD TO THE TIME SERIES OF SUCCESS RATES OF STATE EXAMINATIONS IN THE DEMOCRATIC REPUBLIC OF CONGO FROM 1967 TO 2021

ABSTRACT

In this article, This is about present an analysis conducted on the time series of success rates in State Examinations in the Democratic Republic of Congo (DRC) since its inception in 1967 until 2021, following the BOX and JENKINS methodology, which allows for obtaining a parsimonious ARIMA(0,1,2) Parcimony, the chosen model due to its familiarity with the presented data, has proved to be the only one to adhere to all the Box-Jenkins method steps for the series of State Exam pass rates in the DRC. At the conclusion of this study, short-term predictions calculated using the Evew12 software with first differencing applied indicate 56.1% for 2022, 61.64% in 2023, 61.61% in 2024, and 61.54% in 2025.

These forecasted results show upward trends and almost stability from 2023 to 2025.

Keywords : Time series, Box-Jenkins methodology, Success rates, ARIMA models, Parsimonious.

1. INTRODUCTION

Currently throughout the Universe, each State is organizing itself to put in place, according to its needs, the type of education desired. In the Democratic Republic of Congo, we inherited the system from our colonizer, which is none other than Belgium. The time has come for us to give our voice to the end of secondary education evaluation system in the DRC, which is the one and only way to obtain the state diploma.

The importance of education cannot be overstated. It is widely recognized that education significantly contributes to development, and without it, individuals would remain in a state of ignorance.

We have chosen for our part, an analysis based on the success rates in the so-called state exams, which we have collected since its establishment in 1967 until 2021. In total 55 observations which, taken together, form a series simple chronology.

The Box and Jenkins methodology, which enables short-term forecasts not exceeding five years (Mouhoumed Elmi, Y; 2022) is utilized. This justifies the novelty of this study and the ability to predict these rates short term over a period of four years.

The objective pursued is to analyze this chronicle and above all to predict as said above.

The approach pursued is that which consisted of collecting at the General Inspection of Primary, Secondary and Technical Education (EPST), the different success rates in state exams which form a chronicle whose size is 55 observations and which then was analyzed following the methodology of BOX and JENKINS [8-10].

The tool used is the Eviews 12 software with the primary differentiation technique

The forecasts obtained are very important for a view on the future of success rates in the State Examination which is the one and only instrument for evaluating National Education at the end of the secondary cycle in the DRC.

2. PRESENTATION OF THE BOX-JENKINS METHOD

Long before the introduction of the Box-Jenkins method, the most commonly known type of forecasting in statistical practice allowed for the acceptance of the existence of a fundamental law independent of the series, represented by this random data; the objective was to isolate the statistical characteristics of the series in order to use them for future projections (Bennai et al., 2013).

In fact, statistical analysis boiled down to attempting to fit observations using a pre-established model. In order to address this issue, professors George Box and Gwilym Jenkins jointly proposed in 1976 a methodological approach offering the possibility to judge if a variation law is satisfactory to the extent possible. This is an iterative approach that involves identifying an appropriate model capable of representing the phenomenon under study (Mouhoumed Elmi, Y; 2022 and Goureraux, C et Montfort, A; 1997). These steps are summarized by the authors into three, five, or more than five steps. In this work, these steps amount to eight, represented by the following flowchart (Aziz, A; 1986).

Flowchart 1: Flow chart showing Box-Jenkins Method

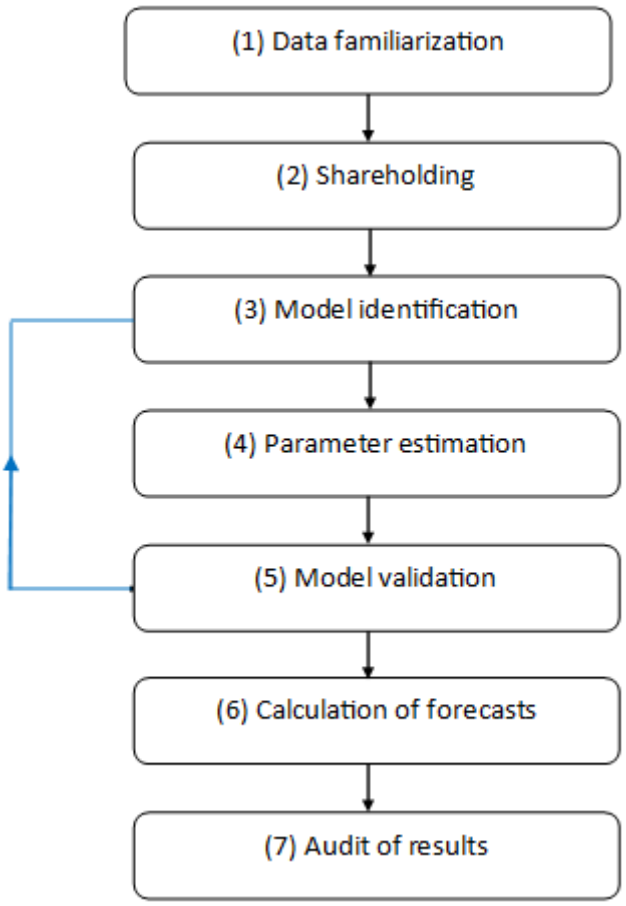


Fig 1 Box-Jenkins methodology diagram.

Source: The authors.

Comments to make on this flowchart are as follows:

(1) At the data familiarization stage, this involves:

- ✓ Represent data graphically;
- ✓ Check if there is any missing data;
- ✓ Check for outliers;
- ✓ Identify the trend;
- ✓ Check stationarity;

(2) Stationarization: stationarize the series using difference or using logarithmic transformation until obtaining the stationarized series.

(3) identification of the model: here consists of determining p and q which correspond to the series to have the number of variables at each polynomial. We used the parsimony approach, i.e. taking p and q as small as possible so as not to lose a lot of information.

(4) Parameter estimation: The best method must be used. We used the ordinary least squares method therefore:

$$\rho_k = \frac{cov(X_t, X_{t-1})}{var(X_t)} = \theta_1^k \frac{var(X_t)}{var(X_t)} = \theta_1^k \quad (1)$$

It should be noted that to verify the stationarity of the AR(1) process, it is necessary and sufficient that $|\theta_1| < 1$which corresponds to the exponential decay ($0 < \theta_1 < 1$)

Or (-1 < θ_1 < 0) oscillatory.

Process$AR(p)$

We either have a process (p), given by:

$$X_t = \theta_0 + \theta_1 X_{t-1} + \cdots + \theta_p X_{t-p} + \varepsilon_t$$

With k=1.2, 3,…

For a process $AR(1)$; we have

$$X_t = \theta_0 + \theta_1 X_{t-1} + \varepsilon_t$$

-Theoretical simple autocorrelation function of a Process$MA(1)$

Or the process $MA(1)$:

$$X_t = \varphi_1 \varepsilon_{t-1} + \varepsilon_t \qquad (2)$$

$$\rho_k = \begin{cases} \rho_0 = 1 \\ \rho_1 = \frac{\varphi_1}{1-\varphi_1^2} \\ \vdots \\ \rho_k = 0 \ \forall \ k > 1 \end{cases} \qquad (3)$$

We can see and generalize that for a process $MA(q)$;

$\rho_k \neq 0$

For$k \leq q \ et \ si \ q > q \qquad \rho_k = 0 \qquad (4)$

Identification is done by comparing the FACempirical to the FACtheoretical

- **Partial autocorrelation function for the parameter$p$ (*PACF*)**

The partial autocorrelation function is comparable to the simple autocorrelation function and also to the partial correlation between two variables $X_t \ et \ X_{t-k}$is this correlation considered after eliminating the effect of all intermediate lags i.e. the effect$X_{t-1}, X_{t-2}, \ldots, X_{t-k+1}$

From one FAC,we can find the partial autocorrelation function $FACP$.

Here is the relationship; $\theta\alpha_{11} = \rho_1$ (5)

$$et \quad \theta_{22} = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2} \qquad (6)$$

$si \ la \ série \ est \ d'ordre \ k \ noté : AR(k), nous \ pouvons \ generalisée \ en :$

$$\theta_{kk} = \frac{\rho_k - \sum_{i=1}^{k=1} \theta_{(k-1)i}\rho_{k-i}}{1 - \sum_{i=1}^{k-1} \theta_{(k-1)}\rho_i} \ pour \ k > 2 \qquad (7)$$

$$\theta_{ki} = \theta_{(k-1)i}\theta_{(k-1)(k-i)} \qquad (8)$$

The coefficients θ_{kk}coming from the estimation of the models above are the values of the simple autocorrelation function of order$k = 1, 2, 3, \ldots$

If$k = 1$

$$X_t = \theta_0 + \theta_{11} X_{t-1} + \varepsilon_t$$

If $k = 2$; SO :

$$X_t = \theta_0 + \theta_{21}X_{t-1} + \theta_{22}X_{t-2} + \varepsilon_t$$

$Si\ k = n;$

$X_t = \theta_0 + \theta_{n1}X_{t-1} + \theta_{n2}X_{t-2} + \cdots + \theta_{nn}X_{t-n} + \varepsilon_t$(9)

For a process$AR(1)$

$$X_t = \theta_0 + \theta_1 X_{t-1} + \varepsilon_t$$

With$|\theta_1| < 1$

$$\theta_{kk}\begin{cases} \neq 0\ pour\ k = 1 \\ = 0\ \forall\ k > 1 \end{cases} \qquad (10)$$

For $AR(p)$;

$X_t = \theta_0 + \theta_1 X_{t-1} + \cdots + \theta_p X_{t-p} + \varepsilon_t$

$$\theta_{kk}\begin{cases} \neq 0\ pour\ k = 1, \ldots, p \\ = 0\ \forall\ k > p \end{cases} \qquad (11)$$

NOTE :

The partial autocorrelation function at lag p (PACF) of an MA(q) process decreases exponentially or oscillatory. It is important to remember that an MA(q) process is always stationary. (Amharo, E; 2023)

The study of autocorrelation functions allows us to identify the series in a family model $ARMA$, which is the adequate model, that is to say the model which has parameters significantly different from zero. The objective here is to determine the form of the model which models the series studied.

The involvement of a few tests will allow us to deseasonalize and stationarize the series to allow the application of the so-called Box-Jenkins methodology.

(5) Model validation: We used several statistical tests as described in the text

(5) Model validation: We used several statistical tests as described in the text before validating a model among those under consideration.

(6) Forecast calculation: To make forecasts, the size of the series needs to be adjusted up to the envisaged forecast period. Therefore, the forecast formula of an ARIMA model should be used.

$$DTREUSSIT.F_t = c + \sum_{i=1}^{p+d} \theta_i X_{t-i} + \varepsilon_t + \sum_{k=j}^{q} \varphi_j \varepsilon_{t-j}$$

And

$\hat{X}_{n+h} = c + \sum_{i=1}^{p+d} \theta_i \hat{X}_{n+h-i} + \sum_{j=1}^{q} \varphi_j \hat{\varepsilon}_{n+h-j}$

If k ≥h and$\hat{\varepsilon}_{n+h-k} = \begin{cases} \varepsilon_{n+h-k}\ si\ k \geq h \\ 0\ si\ k < h \end{cases}$

If we replace the collected data in formula (13) with h=1

$$X_t = c + \sum_{k=1}^{0+1} \theta_k X_{t-k} + \varepsilon_t + \sum_{k=1}^{2} \varphi_k \varepsilon_{t-k}$$

$X_t = c + \sum_{k=1}^{0+1} \theta_k X_{t-k} + \varepsilon_t + \sum_{k=1}^{2} \varphi_k \varepsilon_{t-k}$Which brings us to;

$X_t = c + \theta_1 X_{t-1} + \varepsilon_t + \sum_{k=1}^{2} \varphi_k \varepsilon_{t-k}$And

$$\hat{X}_{55+1} = \sum_{k=1}^{0+1} \theta_k \hat{X}_{55+1-k} + c + \sum_{k=1}^{2} \varphi_k \hat{\varepsilon}_{55+1-k}$$

$$\hat{X}_{56} = \theta_1 \hat{X}_{55} + c + \sum_{k=1}^{2} \varphi_k \hat{\varepsilon}_{55-k}$$

With $h \geq q$; we obtain a relation of the form;$\hat{X}_{n+h} = c + \sum_{k=1}^{p+d} \theta_k \hat{X}_{n+h-k}$

For h=2,

$$\hat{X}_{57} = c + \sum_{k=1}^{0+1} \theta_k \hat{X}_{55+2-k}$$

$$\hat{X}_{57} = c + \theta_1 \hat{X}_{55+2-1}$$

$$\hat{X}_{57} = c + \theta_1 \hat{X}_{56}$$

Of this ; we can deduce;

For h=3,

$\hat{X}_{58} = c + \theta_1 \hat{X}_{57}$

For h=4,

$\hat{X}_{59} = c + \theta_1 \hat{X}_{58}$

3. PRESENTATION OF DATA

This work is based on the collection of data on the success rates in the state exams in the DRC, from its establishment in 1967 until 2021, presented here in Table No. 1.

Table No. 1: Different success rates in State Exams from 1967 to 2021

YEAR	RATE	YEAR	RATE	YEAR	RATE	YEAR	RATE	YEAR	RATE

1967	70	1978	18	1989	51	2000	65	2011	69
1968	65	1979	42	1990	50	2001	33	2012	61
1969	56	1980	41	1991	67	2002	41	2013	47
1970	67	1981	40	1992	46	2003	68	2014	54
1971	58	1982	30	1993	38	2004	67	2015	60
1972	68	1983	31	1994	59	2005	66	2016	61
1973	53	1984	40	1995	64	2006	63	2017	65
1974	52	1985	46	1996	56	2007	38	2018	67
1975	64	1986	51	1997	66	2008	46	2019	70
1976	70	1987	46	1998	63	2009	62	2020	72
1977	51	1988	45	1999	63	2010	66	2021	58

4. APPLICATION OF THE BOX-I NKINS METHOD

4.1. Familiarization

The study of the series we refer to as "TREUSSIT" involves an analysis of the graph and the autocorrelation function of the series, which will help detect the presence of a trend and/or seasonality using the Eviews 12 software.

4.1.1 Graph and correlogram of the TREUSSIT series

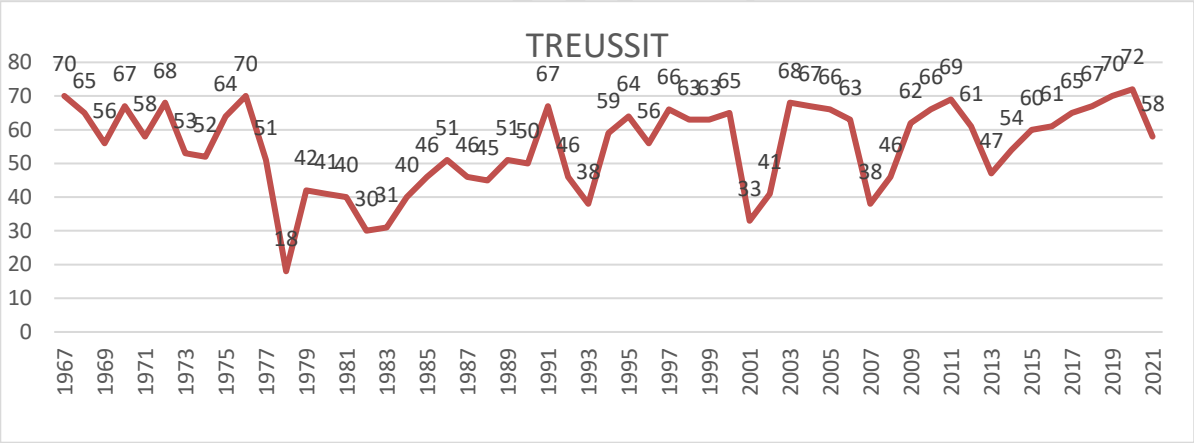


Figure 2: Graphical representation of the TREUSSIT series

Source: the authors.

We see that there is not much fluctuation around the average. Therefore, the time series is not stationary because its trend is downward.

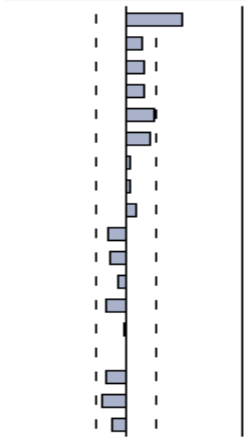
Graphical examination does not always make it possible to determine with certainty the existence of a trend. In order to remove uncertainty, we use the correlogram and appropriate tests to prove or not the stationarity of the series.

4.1.2 Correlogram

Figure 3: Correlogram of the TREUSSIT series

Source: the authors.

Date: 12/22/22 Time: 11:24
Sample (adjusted): 1967 2021
Included observations: 55 after adjustments



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.509	0.509	15.016	0.000
		2	0.148	-0.149	16.310	0.000
		3	0.158	0.204	17.816	0.000
		4	0.162	0.001	19.431	0.001
		5	0.259	0.246	23.635	0.000
		6	0.215	-0.058	26.597	0.000
		7	0.039	-0.066	26.698	0.000
		8	0.026	0.025	26.743	0.001
		9	0.082	0.028	27.203	0.001
		10	-0.163	-0.389	29.047	0.001
		11	-0.149	0.167	30.637	0.001
		12	-0.081	-0.166	31.119	0.002
		13	-0.183	-0.065	33.607	0.001
		14	-0.013	0.206	33.620	0.002
		15	0.007	-0.030	33.624	0.004
		16	-0.176	-0.131	36.109	0.003
		17	-0.222	-0.076	40.168	0.001
		18	-0.118	0.129	41.342	0.001

Reading the correlogram

Columns AC and PAC are the results of the simple autocorrelation functions and the partial autocorrelation function which are represented respectively in the form of horizontal bars on the first two columns of this table.

Their confidence interval boundaries (distaff band) are stylized by vertical dotted lines. Each term that falls outside this interval is significantly different from 0 at the 5% threshold.

Q-Stat is the statistic of Q_{BP} lag k.

Prob is the critical probability of the test on the nullity of the coefficientsρ_k

Finally the delay k is represented by the column between the PAC and AC columns

Interpretation :

From the graphical representation we see the absence of many fluctuations around the trend. We can say that this diagram is that of a non-stationary series.

From the correlogram, we see that the probabilities are all less than 0.05 Hence the series is not stationary.

The series of State Exam pass rates is therefore non-stationary according to reading the graph and the correlogram of the raw series. It is up to us to prove this thesis or not by appropriate tests.

4.1.3 Study of the stationarity of the TREUSSIT series

There are several tests to check the stationarity of a series but the most used are that of Philipe Perron and the Dickey Fuller Augmenter Test (ADF). These tests not only help to test the stationarity of the series but also to show the best way to stationarize it in case of non-stationarity.

In this article, the Augmented Dickey-Fuller test, a parametric test, was used due to the numeric nature of the data (Mantal, A; 2023). According to the examination of the autocorrelation function, all probabilities are less than 5%; therefore, this is a process that does not follow a random distribution.

We move on to hypothesis testing; we first pose the hypotheses then we apply the test at the 5% threshold.

MODEL 3

$$X_t = \alpha X_{t-1} + \beta t + c + \varepsilon_t$$

We make the following hypotheses:

$$\begin{cases} H_0 : \text{absence of trend} \\ H_1 : \text{presence of the trend} \end{cases}$$

Table 3: Model 3 of the ADF test on the TREUSSIT series

Source: the authors.

Null Hypothesis: TREUSSIT has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 1 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.497700	0.0037
Test critical values:	1% level	-4.140858	
	5% level	-3.496960	
	10% level	-3.177579	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(TREUSSIT)
Method: Least Squares
Date: 01/03/23 Time: 13:56
Sample (adjusted): 1969 2021
Included observations: 53 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TREUSSIT(-1)	-0.629654	0.139995	-4.497700	0.0000
D(TREUSSIT(-1))	0.170693	0.138336	1.233905	0.2231
C	29.51839	7.480724	3.945927	0.0003
@TREND("1967")	0.170432	0.099996	1.704386	0.0946

R-squared	0.305155	Mean dependent var	-0.132075
Adjusted R-squared	0.262613	S.D. dependent var	12.35070
S.E. of regression	10.60570	Akaike info criterion	7.633132
Sum squared resid	5511.565	Schwarz criterion	7.781834
Log likelihood	-198.2780	Hannan-Quinn criter.	7.690316
F-statistic	7.173099	Durbin-Watson stat	1.933723
Prob(F-statistic)	0.000437		

k :nknbkjbbn n

We have results found using the Eviews 12 software as follows:
the probability linked to the trend; prob=0.0946 $> 5\%$ therefore the trend is not significant and since the value of the ADF is greater than that of Mackinnon, that is to say - 4.497700 <- 3.496960 at the threshold of 5%, we have the presence of the unit root, although the series does not contain a significant trend, it is not as stationary because of the presence of the unit root. From which we accept H_0which stipulates that there is the absence of the trend .

Now let's check the presence of the constant.

MODEL 2

$$X_t = \alpha X_{t-1} + c + \varepsilon_t$$

We make the following hypotheses:

$$\begin{cases} H_0 : \text{absence of the constant} \\ H_1 : \text{presence of the constant} \end{cases}$$

Table 4: Model 2 of the ADF test on the TREUSSIT series

Source: the authors.

Null Hypothesis: TREUSSIT has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.136688	0.0019
Test critical values:	1% level	-3.557472	
	5% level	-2.916566	
	10% level	-2.596116	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(TREUSSIT)
Method: Least Squares
Date: 01/03/23 Time: 14:06
Sample (adjusted): 1968 2021
Included observations: 54 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TREUSSIT(-1)	-0.481901	0.116494	-4.136688	0.0001
C	26.26446	6.567220	3.999326	0.0002

R-squared	0.247600	Mean dependent var	-0.222222
Adjusted R-squared	0.233131	S.D. dependent var	12.25156
S.E. of regression	10.72882	Akaike info criterion	7.620078
Sum squared resid	5985.592	Schwarz criterion	7.693744
Log likelihood	-203.7421	Hannan-Quinn criter.	7.648488
F-statistic	17.11219	Durbin-Watson stat	1.845171
Prob(F-statistic)	0.000129		

We have the probability linked to the constant given by: prob =0.0002 which is less than 5%; the ADF equal to -4.13 is lower than the Mackinnon critical value at the 5% threshold given by -2.91 Hence; we accept H_1 et we reject H_0.We maintain that there is the presence of a constant.

The model to be retained must necessarily have a constant.

Let's move on to checking the presence of the unit root.

MODEL 1

$$X_t = \alpha X_{t-1} + \varepsilon_t$$

$$\begin{cases} H_0 : \text{presence of the unit root : the series is not stationary} \\ H_1 : \text{absence of the unit root : the series is stationary} \end{cases}$$

Table 5: Model 1 of the ADF test on the TREUSSIT series

Source: the authors.

Null Hypothesis: TREUSSIT has a unit root
Exogenous: None
Lag Length: 4 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-0.118200	0.6381
Test critical values:	1% level	-2.612033	
	5% level	-1.947520	
	10% level	-1.612650	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(TREUSSIT)
Method: Least Squares
Date: 01/03/23 Time: 14:16
Sample (adjusted): 1972 2021
Included observations: 50 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TREUSSIT(-1)	-0.003332	0.028193	-0.118200	0.9064
D(TREUSSIT(-1))	-0.313851	0.141883	-2.212045	0.0321
D(TREUSSIT(-2))	-0.585067	0.142515	-4.105294	0.0002
D(TREUSSIT(-3))	-0.303331	0.140699	-2.155894	0.0365
D(TREUSSIT(-4))	-0.342712	0.138753	-2.469950	0.0174

R-squared	0.304778	Mean dependent var	0.000000
Adjusted R-squared	0.242980	S.D. dependent var	12.49490
S.E. of regression	10.87142	Akaike info criterion	7.704791
Sum squared resid	5318.451	Schwarz criterion	7.895994
Log likelihood	-187.6198	Hannan-Quinn criter.	7.777602
Durbin-Watson stat	1.953591		

The series is stationary because the ADF value is greater than the Mackinnon critical value. Which means,

-0.118200 >-1.947520 , then we accept $H_0$and reject H_1, so the series is not stationary, we are facing a DS model with drift.

We know that ; To make a DS model process stationary, we will go through the differentiation method of order 1,2,… until the series becomes stationary. Here we apply the ADF test to the series in first differentiation.

4.1.4 TREUSSIT STATIONARIZATION

Hypotheses :

H_0 : The series is not stationary

H_1 : the series is stationary

The application of the ADF Test to the first differentiation series

Table 6: Application of the ADF Test to the first differentiation series

Source: the authors.

Null Hypothesis: DTREUSSIT has a unit root
Exogenous: None
Lag Length: 3 (Automatic - based on SIC, maxlag=10)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.661759	0.0000
Test critical values:	1% level	-2.612033	
	5% level	-1.947520	
	10% level	-1.612650	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DTREUSSIT)
Method: Least Squares
Date: 01/03/23 Time: 14:20
Sample (adjusted): 1972 2021
Included observations: 50 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DTREUSSIT(-1)	-2.552770	0.383198	-6.661759	0.0000
D(DTREUSSIT(-1))	1.236162	0.306610	4.031702	0.0002
D(DTREUSSIT(-2))	0.648548	0.218431	2.969115	0.0047
D(DTREUSSIT(-3))	0.343683	0.137017	2.508322	0.0157

R-squared	0.684501	Mean dependent var	-0.100000
Adjusted R-squared	0.663925	S.D. dependent var	18.55081
S.E. of regression	10.75427	Akaike info criterion	7.665102
Sum squared resid	5320.102	Schwarz criterion	7.818064
Log likelihood	-187.6275	Hannan-Quinn criter.	7.723351
Durbin-Watson stat	1.954170		

We now have a prob=o $< 5\%$; the ADF value = -6.661759 which is less than the Mackinnon critical value equal to -1.947520. We accept H_1 and we reject H_0. Hence the TREUSSIT series is stationary after the first difference.

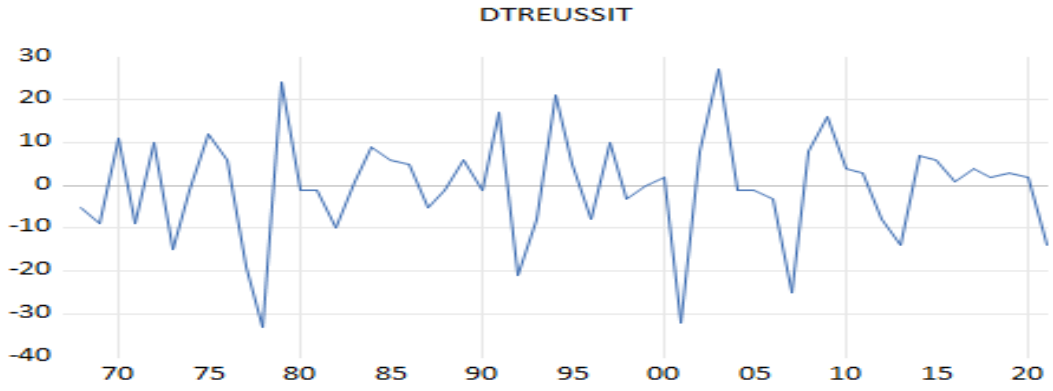


Figure 4: Graph of the stationary TREUSSIT series

Source: the authors.

We notice that there are fluctuations around the mean, which shows that our series is stationary. From where we can move on to the step of determining p and q.

4.2. Modelization

4.2.1 Determination of orders (p, q)

We have a stationary series, we now need to determine an ARMA model (p, q); To do this, we will use the correlogram from the stationary TREUSSIT series. The simple correlogram allows us to identify an MA(q) model, and the partial correlogram allows us to identify an AR(p) model.

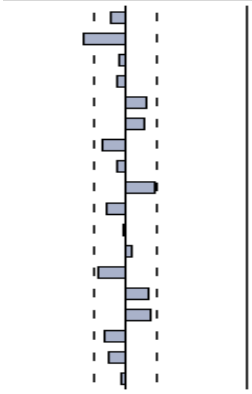
Please note that it is advisable not to exceed the first three delays so as not to lose a lot of information.

Here is the correlogram of the stationary TREUSSIT series after the first difference;

Table 7: Correlogram of the stationary TREUSSIT series

Source: the authors.

Date: 01/03/23 Time: 19:56
Sample (adjusted): 1968 2021
Included observations: 54 after adjustments



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.125	-0.125	0.8845	0.347
		2	-0.363	-0.385	8.5537	0.014
		3	-0.062	-0.206	8.7816	0.032
		4	-0.068	-0.325	9.0589	0.060
		5	0.177	-0.033	11.001	0.051
		6	0.161	0.051	12.643	0.049
		7	-0.195	-0.116	15.094	0.035
		8	-0.071	-0.035	15.424	0.051
		9	0.253	0.257	19.729	0.020
		10	-0.164	-0.114	21.574	0.017
		11	-0.009	0.069	21.580	0.028
		12	0.062	0.025	21.853	0.039
		13	-0.235	-0.230	25.930	0.017
		14	0.202	0.056	29.026	0.010
		15	0.207	0.121	32.340	0.006
		16	-0.178	0.015	34.851	0.004
		17	-0.153	-0.121	36.772	0.004
		18	-0.043	-0.137	36.928	0.005

We note that certain probabilities of the correlogram of the differentiated series are less than 5%, it is therefore not a question of a random approach, that is to say that there exists a representation of the series in the family of the ARMA model.

By differentiating our series, we removed the trend it presented in order to stationarize it. The ARIMA model is therefore a combination of this differentiation process and the classic ARMA process. Furthermore, if the time series presents both a trend and seasonality, it is possible to use the SARIMA model, which is nothing more than an ARIMA model taking into account a seasonal component. W e offer the following models with constant and order 1 differentiation:

- ARIMA (2, 1,2)
- ARIMA (0, 1,2)
- ARIMA (2, 1,0)

The models having been proposed, we can move on to the stage of estimating its parameters.

4.3. *MODEL ESTIMATION*

The model is estimated with the stationary series; we will accept the model with significant coefficients. If the model parameters are significant, then we can test these residuals, which should be white noise.

We will have to compare the values found to the ratio of Student's t statistic (for us 1.96). Using the Eviews 12 software. We have the following estimates:

4.3.1 Estimation of the ARIMA model (2, 1, 0)

Table 8: Estimation of the ARIMA model (2, 1.0)

Source: the authors.

Dependent Variable: DTREUSSIT
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 01/03/23 Time: 14:30
Sample (adjusted): 1970 2021
Included observations: 52 after adjustments
Convergence achieved after 9 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.028017	1.175798	0.023828	0.9811
AR(2)	-0.372776	0.132147	-2.820918	0.0069

R-squared	0.137300	Mean dependent var	0.038462
Adjusted R-squared	0.120046	S.D. dependent var	12.40803
S.E. of regression	11.63946	Akaike info criterion	7.784382
Sum squared resid	6773.853	Schwarz criterion	7.859430
Log likelihood	-200.3939	Hannan-Quinn criter.	7.813153
F-statistic	7.957580	Durbin-Watson stat	2.445264
Prob(F-statistic)	0.006850		

Inverted AR Roots	-.00+.61i	-.00-.61i

We have the probability linked to AR(2) equal to 0.0069 which is less than 0.05 (5%); the t-statistic value is greater than 1.96; this shows that the coefficient of AR(2) is significantly different from zero. Hence the model is good and is retained for the test of the residuals.

4.3.2 Estimation of the ARIMA model (0, 1.2)

Table 9 : Estimation of the ARIMA model (0, 1.2)

Source: the authors.

Dependent Variable: DTREUSSIT
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 01/03/23 Time: 14:32
Sample (adjusted): 1968 2021
Included observations: 54 after adjustments
Failure to improve likelihood (non-zero gradients) after 15 iterations
Coefficient covariance computed using outer product of gradients
MA Backcast: 1966 1967

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.026015	0.742054	0.035058	0.9722
MA(2)	-0.534244	0.113742	-4.696989	0.0000

R-squared	0.189632	Mean dependent var	-0.222222
Adjusted R-squared	0.174048	S.D. dependent var	12.25156
S.E. of regression	11.13445	Akaike info criterion	7.694298
Sum squared resid	6446.749	Schwarz criterion	7.767964
Log likelihood	-205.7461	Hannan-Quinn criter.	7.722708
F-statistic	12.16836	Durbin-Watson stat	2.560588
Prob(F-statistic)	0.000998		

Inverted MA Roots	.73	-.73

We have the probability linked to the MA(2) model, prob=0, less than 0.05 and the coefficient of the model significantly differs from zero because the t-statistic value is well above 1.96; we can therefore retain the model for testing the residuals.

4.3.4 Estimation of the ARIMA model (2, 1,2)

Table 10 : Estimation of the ARIMA model (2, 1,2)

Source: the authors.

Dependent Variable: DTREUSSIT
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 01/03/23 Time: 14:27
Sample (adjusted): 1970 2021
Included observations: 52 after adjustments
Failure to improve likelihood (non-zero gradients) after 26 iterations
Coefficient covariance computed using outer product of gradients
MA Backcast: 1968 1969

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.160583	0.695442	0.230908	0.8183
AR(2)	0.132592	0.236241	0.561259	0.5772
MA(2)	-0.640104	0.186311	-3.435680	0.0012

R-squared	0.227966	Mean dependent var	0.038462
Adjusted R-squared	0.196454	S.D. dependent var	12.40803
S.E. of regression	11.12265	Akaike info criterion	7.711805
Sum squared resid	6061.951	Schwarz criterion	7.824377
Log likelihood	-197.5069	Hannan-Quinn criter.	7.754962
F-statistic	7.234355	Durbin-Watson stat	2.488376
Prob(F-statistic)	0.001766		

Inverted AR Roots	.36	-.36
Inverted MA Roots	.80	-.80

On the one hand, we have the probability of AR(2) greater than 5%; that of MA(2) less than 5% and on the other side, the coefficient of the AR(2) process significantly equal to zero because its t-statistic value is less than 1.96 and that of MA(2) significantly differs from zero because its t-statistic value is much greater than 1.96. Hence the model overall is not good and should be rejected

4.4 Validation

4.4.1 Residue testing

The estimation phase allowed us to retain two models that we must test their residuals to find the model that would contain white noise residuals. Hence the residue test.

4.4.2 Test for autocorrelation of residuals

a) Testing the residuals of the ARIMA model (2, 1.0)

Residual correlogram

$$\begin{cases} H_0: \text{process is memoryless} \\ H_1: \text{process with memory} \end{cases}$$

Table 11 : Residual correlogram of the ARIMA model (2, 1.0)

Source: the authors.

Date: 01/04/23 Time: 20:51
Sample (adjusted): 1968 2021
Q-statistic probabilities adjusted for 1 ARMA term

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.240	-0.240	3.2926	
		2	-0.094	-0.161	3.8072	0.051
		3	-0.057	-0.133	3.9991	0.135
		4	-0.172	-0.266	5.7874	0.122
		5	0.125	-0.037	6.7482	0.150
		6	0.158	0.131	8.3270	0.139
		7	-0.071	-0.005	8.6534	0.194
		8	-0.095	-0.110	9.2416	0.236
		9	0.247	0.292	13.337	0.101
		10	-0.228	-0.073	16.903	0.050
		11	-0.011	-0.120	16.912	0.076
		12	0.103	0.063	17.669	0.090
		13	-0.222	-0.156	21.294	0.046
		14	0.220	0.037	24.964	0.023
		15	0.116	0.140	26.000	0.026
		16	-0.184	-0.056	28.685	0.018
		17	-0.072	-0.136	29.104	0.023
		18	-0.086	-0.190	29.729	0.028

We note that the probabilities are greater than 5% up to order k >13; from which we can say that we have a process whose residues do not follow the course of chance. Thus, we resort to the statistical test of Q_{BP} .

Consider the Box-perce statistic at delay k=18, Q_{BP}= 29.729 to compare with the value of χ^2at delay kpq-1 is equivalent to k=15; χ^2=24.99 which is lower than the statistical value Q_{BP}.

So this is not a memoryless process; Hence the model is to be rejected.

b) Test of residuals on the ARIMA model (0, 1,2)

Residual correlogram

$$\begin{cases} H_0: \text{process is memoryless} \\ H_1: \text{process with memory} \end{cases}$$

Table 12: ARIMA residual correlogram (0, 1.2)

Source: the authors.

Date: 01/04/23 Time: 14:54
Sample (adjusted): 1968 2021
Q-statistic probabilities adjusted for 1 ARMA term

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.278	-0.278	4.4142	
		2	0.009	-0.075	4.4184	0.036
		3	-0.131	-0.162	5.4297	0.066
		4	-0.016	-0.112	5.4448	0.142
		5	0.098	0.051	6.0323	0.197
		6	0.122	0.161	6.9715	0.223
		7	-0.070	0.015	7.2900	0.295
		8	-0.092	-0.083	7.8480	0.346
		9	0.227	0.252	11.314	0.185
		10	-0.185	-0.075	13.668	0.135
		11	0.000	-0.142	13.668	0.189
		12	0.053	0.068	13.874	0.240
		13	-0.189	-0.194	16.500	0.169
		14	0.163	-0.003	18.515	0.139
		15	0.120	0.185	19.629	0.142
		16	-0.174	-0.106	22.031	0.107
		17	-0.064	-0.119	22.366	0.132
		18	-0.116	-0.194	23.502	0.134

We note that the simple and partial autocorrelations are all significantly equal to zero because its values do not leave the Distaff band at the 5% threshold, with the probabilities all greater than 0.05 except at lags k=2.

To do this, we move on to the statistics $Q_{BP}$of the series which we will compare with the value of χ^2at delay 15 and we notice that $\chi^2 > Q_{BP}$, at the threshold of 5%. That is to say 24.99 >23.50 . Hence, we assimilate the process to that of chance and accept H_0because it is a process without memory, which means that it will not intervene in the forecasts.

4.4.3 Test of Heteroskedasticity

a) Arch test

$$\begin{cases} H_0: \text{the residuals are not autocorrelated (presence of homoscedasticity)} \\ H_1: \text{the residuals are autocorrelated (presence of heteroscedasticity)} \end{cases}$$

Table 13: Heteroscedasticity test (ARCH) on the ARIMA model (0, 1,2)

Source: the authors.

Heteroskedasticity Test: ARCH

F-statistic	0.330381	Prob. F(2,49)	0.7202
Obs*R-squared	0.691887	Prob. Chi-Square(2)	0.7076

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 01/03/23 Time: 14:42
Sample (adjusted): 1970 2021
Included observations: 52 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	110.5092	35.59694	3.104458	0.0032
RESID^2(-1)	0.115890	0.142793	0.811596	0.4209
RESID^2(-2)	-0.007312	0.142840	-0.051187	0.9594

R-squared	0.013306	Mean dependent var	123.8104
Adjusted R-squared	-0.026968	S.D. dependent var	191.8767
S.E. of regression	194.4467	Akaike info criterion	13.43415
Sum squared resid	1852666.	Schwarz criterion	13.54673
Log likelihood	-346.2880	Hannan-Quinn criter.	13.47731
F-statistic	0.330381	Durbin-Watson stat	2.003278
Prob(F-statistic)	0.720238		

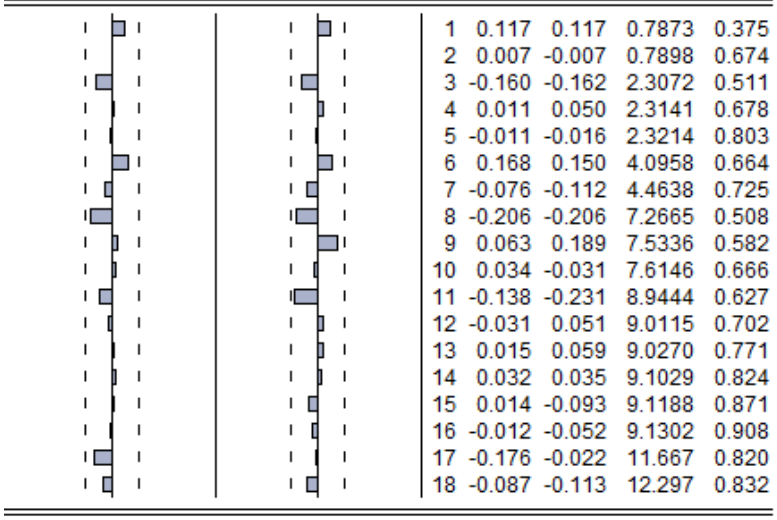
We have the test probability of ARCH =0.7 which is greater than 5%, The statistical value of Durbin-Watson (DW); DW= 2.003278 $\approx$ 2which shows the absence of autocorrelation of the residuals. Hence, we accept H_0 the absence of Heteroskedasticity, which is also Demonstrated by the correlogram below:

b) **Correlogram of the squares of the residuals**

Table 14: Correlogram of the squares of the residuals of the ARIMA model (0.1, 2)

Source: the authors.

Date: 01/04/23 Time: 16:01
Sample (adjusted): 1968 2021
Included observations: 54 after adjustments



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.117	0.117	0.7873	0.375
		2	0.007	-0.007	0.7898	0.674
		3	-0.160	-0.162	2.3072	0.511
		4	0.011	0.050	2.3141	0.678
		5	-0.011	-0.016	2.3214	0.803
		6	0.168	0.150	4.0958	0.664
		7	-0.076	-0.112	4.4638	0.725
		8	-0.206	-0.206	7.2665	0.508
		9	0.063	0.189	7.5336	0.582
		10	0.034	-0.031	7.6146	0.666
		11	-0.138	-0.231	8.9444	0.627
		12	-0.031	0.051	9.0115	0.702
		13	0.015	0.059	9.0270	0.771
		14	0.032	0.035	9.1029	0.824
		15	0.014	-0.093	9.1188	0.871
		16	-0.012	-0.052	9.1302	0.908
		17	-0.176	-0.022	11.667	0.820
		18	-0.087	-0.113	12.297	0.832

The correlogram presents autocorrelation coefficients all significantly equal to zero at the 5% threshold because none of the simple or partial autocorrelation coefficients leave the distaff band.

With probabilities greater than 5%, we can say that there is the absence of autocorrelation of the residuals which justifies the presence of homoscedasticity of the errors.

4.4.3 Normality test

$$\begin{cases} H_0: \text{the residuals are not Gaussian white noise} \\ H_1: \text{the residuals are Gaussian white noise} \end{cases}$$

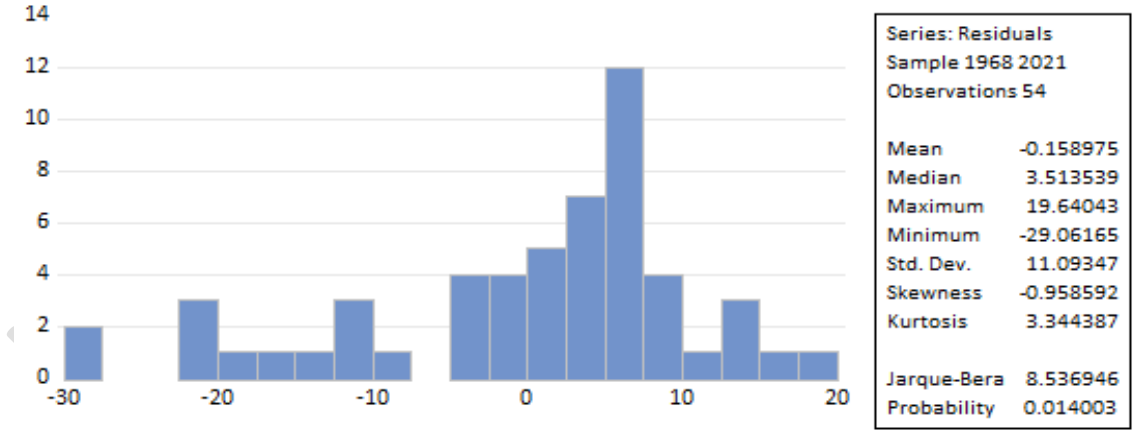


Fig 5: The histogram of the normality test of the residuals (Jacques Berat test)

Source: the authors.

We see from the histogram of the residuals that:

The skewness value being -0.958592 is less than 1.96 therefore the residuals are distributed symmetrically.

That of kurtosis is $3.344387 > 1.96$explains that the distribution of residuals is not flattened and finally that of the Jacques-Berat statistic is 8.536946 which is $>$ à $0{,}05$at the threshold value ;

Which explains why the residues follow a normal law; from which we accept H_1 and reject H_0, the residuals are therefore Gaussian white noise.

It should be noted that this process is memoryless and there is no autocorrelation of the residuals; we can say that the residuals are Gaussian white noise because the assumption of normality of the residuals is respected.

We retained the ARIMA model (0, 1,2) because it is the only one in the ARIMA process to have satisfied the conditions required to forecast with the Box-Jenkins method.

5. Forecast

To make forecasts, it is necessary to modify the size of the series, up to the forecast period envisaged. We know that the residuals are white noise; that is to say they cannot intervene in the forecast, we should therefore use the forecast formula of an ARIMA model given by:

$$DTREUSSIT.F_t = c + \sum_{i=1}^{p+d} \theta_i X_{t-i} + \varepsilon_t + \sum_{k=j}^{q} \varphi_j \varepsilon_{t-j}$$

And

$\hat{X}_{n+h} = c + \sum_{i=1}^{p+d} \theta_i \hat{X}_{n+h-i} + \sum_{j=1}^{q} \varphi_j \hat{\varepsilon}_{n+h-j}$

If k ≥h and $\hat{\varepsilon}_{n+h-k} = \begin{cases} \varepsilon_{n+h-k} \; si \; k \geq h \\ 0 \; si \; k < h \end{cases}$

If we replace the data we have in the formula, we will have;

For h=1

$$X_t = c + \sum_{k=1}^{0+1} \theta_k X_{t-k} + \varepsilon_t + \sum_{k=1}^{2} \varphi_k \varepsilon_{t-k}$$

$X_t = c + \sum_{k=1}^{0+1} \theta_k X_{t-k} + \varepsilon_t + \sum_{k=1}^{2} \varphi_k \varepsilon_{t-k}$Which brings us to;

$X_t = c + \theta_1 X_{t-1} + \varepsilon_t + \sum_{k=1}^{2} \varphi_k \varepsilon_{t-k}$And

$$\hat{X}_{55+1} = \sum_{k=1}^{0+1} \theta_k \hat{X}_{55+1-k} + c + \sum_{k=1}^{2} \varphi_k \hat{\varepsilon}_{55+1-k}$$

$$\hat{X}_{56} = \theta_1 \hat{X}_{55} + c + \sum_{k=1}^{2} \varphi_k \hat{\varepsilon}_{55-k}$$

With $h \geq q$; we obtain a relation of the form;$\hat{X}_{n+h} = c + \sum_{k=1}^{p+d} \theta_k \hat{X}_{n+h-k}$

For h=2,

$$\hat{X}_{57} = c + \sum_{k=1}^{0+1} \theta_k \hat{X}_{55+2-k}$$

$$\hat{X}_{57} = c + \theta_1 \hat{X}_{55+2-1}$$

$$\hat{X}_{57} = c + \theta_1 \hat{X}_{56}$$

Of this ; we can deduce;

For h=3,

$\hat{X}_{58} = c + \theta_1 \hat{X}_{57}$

For h=4,

$\hat{X}_{59} = c + \theta_1 \hat{X}_{58}$

We know that the results above are estimated from Dtreussit because$DTREUSSIT_t = \text{TREUSSIT}_t - \text{TREUSSIT}_{t-1}$

SO ;$\text{TREUSSITF}_t = DTREUSSITF_t + \text{TREUSSITF}_{t-1}$

If h=1, then$\text{TREUSSITF}_{t-1} = \text{TREUSSIT}_{t-1}$

We used the Eviews 12 software to have the following forecasts for the years 2022, 2023, 2024, 2025:

Table 15: Forecasts up to h=4

Source: the authors.

YEAR	TREUSSITF
2022	56.17347813585213
2023	61.6376834440594
2024	61.60563624509442
2025	61.57358904612945

These forecasts are also given by the following diagram

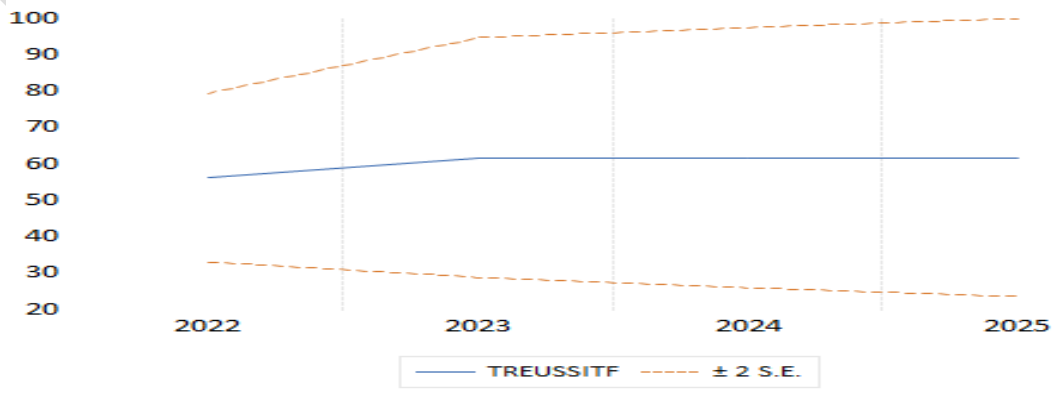


Figure 6: Graph of predicted values

Source: the authors.

The predicted values of the TREUSSIT series are obtained after determining the errors observed by the moving average of the differentiated series, it is therefore a short memory model. As a result, we can only make short-term forecasts.

The red lines represent the two limits of the confidence interval which appears to be constant from 2023 until 2025.

So, here is the graphic representation of the treussit series until the year 2025

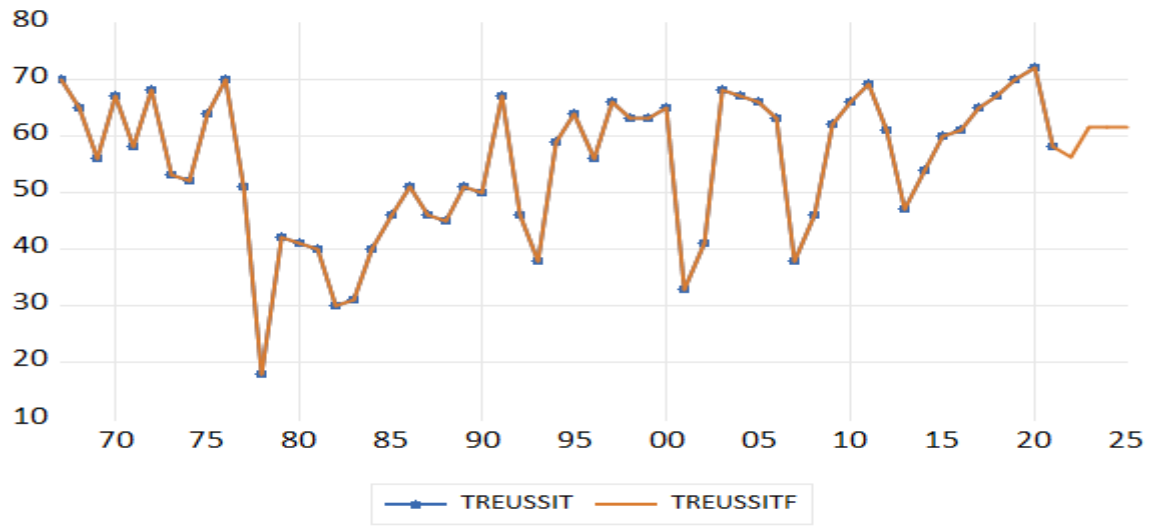


Fig 7: Graphical representation of the Treussit series predicted up to the value h=4
Source: the authors.

6. Discussion :

Our study focuses on the analysis of the variable "success rate for State Examinations in the DRC" which means studying the evolution of these rates and predicting in the short term as the model requested. It should be noted that these rates are influenced by several external variables. Let us say in passing that the Congolese education system is shaken by repeated strikes especially at the start of each school year, the mobility of students who fail before reaching the last year or even the last year; constant wars in the east of the country; the change of school programs without training of trainers and without documentation; unsatisfied motivation of trainers; unsuitable frameworks for training; the politicization of the education system; the lack of working tools on the part of both learners and trainers; The epidemics ; corruption and anti-values of all kinds; …

All these factors contribute in one way to another to vary these rates.

In developing these forecasts, we have not taken into account the effects of these factors on the pass rates for the state exam in the DRC. Our only factor is time. Which justifies the name time series.

The most parsimonious model is given after first differentiation; for our TREUSSIT series; we retained the ARIMA model (0, 1,2), which allowed the prediction.

CONCLUSION

The main goal was to predict. The model imposed on us short-term forecasts of these success rates.

To achieve this, we chose the Box-Jenkins approach, a parsimony approach. We used a computer tool which is the Eviews 12 software, which made it possible to obtain the desired results.

After the presentation of success rate data, we moved on to the graphical representation which constitutes the first step of the Box-Jenkins methodology; this graph did not completely confirm to us that the series was non-stationary; the study of the correlogram of partial and simple autocorrelation functions, the augmented Dickey Fuller test (ADF) showed the non-stationarity of this so-called TREUSSIT series and the way of stationarizing it.

Results obtained after testing the ADF for its models were presented as follows:

- Model 3, test on the constant and the trend gave the absence of the trend and the presence of the constant,
- Model 2; test on the constant indicates the presence of the constant,
- Model 1; test without constant or trend affirmed that there is the presence of a unit root.

And these results led us to a non-stationary model of the DS type with drift; the series becomes stationary after applying the first difference. Hence the TREUSSIT series becomes DTREUSSIT.

Three models are proposed ARIMA(2,1,0); ARIMA(0,1,2); ARIMA(2,1,2), among these three, only the ARIMA(0,1,2) model turned out to be the only one to have white noise and Gaussian residuals and was retained for calculating the forecasts of the series DETROIT on the horizons h.

The short-term forecasts calculated using the eviews12 software are: 56.17% for 2022, 61.64% in 2023, 61.61% in 2024 and 61.57% in 2025.

To show the veracity of our results, we invite readers to compare them with data actually recorded in the field. Already, for the year 2022, the success rate is 60%, slightly lower than that proposed by our model.

The forecasts obtained are very important for a view on the future of success rates in the State Examination which is an instrument for evaluating National Education at the end of the secondary cycle. However, we say that several factors can influence these forecasts. We cited natural disasters, pandemics, free education, anti-values,…

We believe that the Congolese State will take its results into account to improve educational results overall.

We propose the following themes for the future:

- ✓ Time series analysis of the results of each study cycle in the DRC
- ✓ Comparative study of the series of State Examination results by Province in the DRC;
- ✓ Impact of anti-values on Exeat success rates in the DRC

COMPETING INTERESTS

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References

[1] Amharo, E (2023): Study of the revenue series generated by the DGRAD in the Democratic Republic of the Congo from 2000 to 2022, Center for Research in Applied Mathematics, Kinshasa, vol. 2, 331-345.

[2] Auxois, J.Y (2017): Introduction to the study of time series; INSA/Toulouse.

[3] BOURBONNAIS Régis (2002): Econometrics, 4th edition, Dunod, Paris.

[4] BENNAI NABIL N and Adrar SOFIANES (2013): Predictive study of oil sales using the Box and Jenkins method, Abderrahmane Mira University of Bejaia, Master's thesis in economics.

[5] CHAOUCHKHOU ANE Meriem (2021): Box and Jenkins Method, Master's thesis in Statistics, Mohamed Khider University, Biskra.

[6] Mantal, A (2023): Application of the Box and Jenkins method to the production of fish caught at the Port of the City of Bandundu from 1995 to 2022, Master's thesis in Statistical Mathematics, National Pedagogical University, Kinshasa.

[7] Mouhoumed, E (2022): Modeling of France's GDP from 1949 to 2014 using the Box-Jenkins method, French Journal of Economics and Management, ISSN: 2728-0128, Vol. 3, No. 7.

8. Anderson OD. The Box-Jenkins approach to time series analysis. RAIRO-Operations Research. 1977;11(1):3-29.

9. Anderson OD. Time Series Analysis and Forecasting: Another Look at the Box-Jenkins Approach. Journal of the Royal Statistical Society: Series D (The Statistician). 1977 Dec;26(4):285-303.

10. Makridakis S, Hibon M. ARMA models and the Box–Jenkins methodology. Journal of forecasting. 1997 May;16(3):147-63.