International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Data-Driven Tax Management: Developing a Long-Run Analysis System for Niger through Outlier Detection

Moussa Khane*[1], Harouna Naroua[2], Chaibou Kadri[3], Yacouba Moumouni[4]

Submitted:10/01/2025 **Revised**: 25/02/2025 **Accepted**: 08/03/2025

Abstract: Developing countries often face challenges in conducting long-term economic analyses, which in turn affects their ability to design effective planning and policy decisions. This study applies the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to support Niger's tax administration in implementing a long-run analysis scheme. Several statistical and machine learning tools, such as boxplots, the interquartile range (IQR), the augmented Dickey-Fuller test, the Johansen test, and the vector error correction model (VECM) were employed. The dataset covers the period from January 1996 to December 2014 and reveals seven (7) outliers. Results showed that VAT, ITS, and ISB account for 93.64% of revenues in the dataset with outliers and 94.25% without outliers, confirming that cointegration tests were highly sensitive to outliers. Both datasets were non-stationary but cointegrated, with rank three (3). Tax revenue took approximately 81 days to absorb shocks with outliers, compared to 60 days without. Outliers, thus, significantly distorted the Nigerien economic planning and policy outcomes.

Keywords: *Cointegration, CRISP-DM, Data mining, Outlier detection, Tax administration, Vector error correction model.*

1. Introduction

Tax revenue is fundamental to the development and implementation of a government's economic and social agenda. To achieve this, governments worldwide require well-defined fiscal policies and sound financial planning to effectively support their programs [1]. Tax pressure, measured as the tax-to-GDP ratio, reflects a country's capacity to mobilize domestic resources. On average, tax pressure is 34.0% in OECD countries, compared to 16.0% in African nations. In developed economies such as France, it reaches 45.15%, whereas in many developing countries it remains very low; for instance, in Niger it stands at only 9.6% [2] [3].

Niger, a vast Sub-Saharan country of which a large portion (approximately 2/3) is desert, faces significant challenges in tax collection due to factors, such as limited human and material resources and the absence of direct access to the sea. The tax administration lacks a unified database, and existing tax records remain fragmented across disparate regional systems [4]. Furthermore, the current web application has not been deployed nationwide, leaving many local tax offices without real-time data access. The absence of data mining tools further prevents the administration from leveraging analytics for informed decision-making. Moreover, no automated system exists to conduct long-run analyses, making it difficult to improve data-driven tax management.

National revenue currently covers only 62% of Niger's budget, with the remaining share reliant on external financing, according to the European Union [5] . Since July 26, 2023, the closure of Niger's main maritime access route and the imposition of international sanctions have placed additional strain on the national budget. Strengthening data management, expanding digital infrastructure, and implementing advanced analytics could significantly enhance the efficiency of tax collection and improve revenue forecasting in Niger.

In the era of artificial intelligence, tax administrations worldwide are evolving toward more intelligent, data-driven systems [6]. As a result, machine learning (ML) provides advanced analytical and statistical tools that enhance the effectiveness of tax control through long-run analysis systems. These technologies enable governments to improve monitoring, oversight, and evidence-based decision-making [7].

Despite significant progress in modernizing tax administrations worldwide, Niger continues to lag behind. To date, no studies have been undertaken to develop a long-run analysis framework within the Nigerien tax administration system. To address this gap, the present study applies the CRISP-DM methodology to conduct long-run analysis (cointegration) with a focus on outlier detection

[1] *Département de Mathématiques et Informatique, Faculté des Sciences et Techniques, Université Abdou Moumouni, Niamey, Niger*

[2] *Département de Mathématiques et Informatique, Faculté des Sciences et Techniques, Université Abdou Moumouni, Niamey, Niger*

[3] *Département de Mathématiques et Informatique, Faculté des Sciences et Techniques, Université Abdou Moumouni, Niamey, Niger*

[4] *Electrical and Electronics Engineering, Higher Colleges of Technology, Ras Al Khaimah, United Arab Emirates*

** Corresponding Author Email: khane12002@yahoo.fr*

in the unified tax database.

For the long-run analysis, the study employs two stages: 1) Anomaly detection techniques, including boxplot analysis and extraction methods such as the interquartile range (IQR) and 2) Cointegration techniques, namely the i) Augmented Dickey-Fuller (ADF) test for stationarity, ii) Johansen test for cointegration, and iii) Vector Error Correction Model (VECM) to evaluate system adjustments in response to shocks.

The objective of this research is to conduct a long-run analysis of tax data through anomaly detection and assess the impact of outliers on model reliability. A comparative analysis is carried out between the original dataset, which contains anomalies, and a cleaned version from which anomalies have been removed. This approach demonstrates the importance of incorporating anomaly detection into cointegration analysis to enhance the reliability of tax data.

The remainder of this paper is structured as follows: Section 2 reviews the literature, Section 3 describes the methodology, Section 4 presents and discusses the results, and Section 5 concludes the study.

2. Literature review

Reviewing the existing literature on cointegration (long-run analysis) for forecasting, particularly in the context of tax administration, is essential for guiding research relevant to Niger's tax system. Subsection 2.1 introduces the key concepts of data mining, while Subsection 2.2 examines the theoretical foundations of cointegration and its role in long-run analysis. Finally, Subsection 2.3 discusses related studies and highlights applications of cointegration within tax administration.

2.1. Data mining

Data mining is a process of extracting and discovering patterns in large data sets (databases) involving methods at the intersection of machine learning, statistics, and database systems [8], [9] [10]. Cheng noted that the development of data mining was not uniform. Rather, it was problem oriented according to different fields[8].

In data mining, outlier detection plays a critical role, particularly in safety-critical environments, as outliers often signal abnormal operating conditions that can result in significant performance deterioration [11]. For example, in finance, outliers may indicate fraudulent transactions; in healthcare, they may point to anomalies in patient health records; and in tax administration, they may reveal irregularities or potential cases of tax evasion. Detecting and addressing such anomalies is therefore essential for ensuring system reliability, supporting accurate forecasting, and strengthening decision-making. The detection of outliers involves both the visualization and statistical analysis of extreme values within a dataset. Certain ML algorithms assign numerical scores that quantify the degree of deviation from established patterns or normal behavior. By systematically evaluating these deviations, outlier analysis enhances the performance of detection algorithms, thereby enabling more accurate and reliable identification of anomalies[12]. Hence, the approach used in this study is based on the interquartile range (IQR).

Cross Industry Standard Process for Data Mining (CRISP-DM) is a freely available, widely used methodology in the field of data mining, which is broadly adopted across various industries [13]. It offers structured guidelines for the organized and transparent execution of projects by dividing all planned tasks into six interrelated phases [14] [15].

2.2. Cointegration (Long-run analysis)

Engle and Granger defined cointegration as a statistical property of time series that enables the detection of long-term relationships among two or more non-stationary variables. The main advantage of cointegration analysis is that it allows for the estimation of dynamic relationships by decomposing the effects of long-term equilibrium forces and short-term dynamics [16].

Vector Autoregression (VAR) represents a specific form of a simultaneous equation system and can be applied when all variables are stationary. However, when the variables in the time series vector, Z_t, are non-stationary but cointegrated, the appropriate model is the Vector Error Correction Model (VECM). VECM is essentially a restricted form of VAR designed for non-stationary time series that exhibit at least one cointegrating relationship[17]. As noted by Agus et al.[18], the restriction is imposed to incorporate the cointegration structure, thereby capturing both short-term adjustments and long-term equilibrium dynamics.

According to [19], VECM is a powerful time-series framework that directly estimates the speed at which variables revert to equilibrium following a shock. It is particularly useful for distinguishing short-term effects from long-term relationships in the data. Nevertheless, as highlighted by [20], standard unit root and cointegration tests are sensitive to outliers and structural breaks, which can distort inference.

The implementation of VECM generally follows several steps. The first step is to test whether the variables are stationary. A time series is said to be integrated of order d, denoted I(d), if it becomes stationary after differencing d times. Typically, the Augmented Dickey–Fuller (ADF) test is applied to determine the presence of a unit root in an autoregressive process of order one [21]. Evidence of a unit root indicates non-stationarity.

It is important to note that not all non-stationary processes contain a unit root. Unit root processes, often referred to as difference-stationary processes, differ from trend-stationary processes despite sharing some properties. In trend-stationary processes, the mean follows a deterministic trend, and shocks have only temporary effects, with the series eventually reverting to the trend. By contrast, unit root processes exhibit permanent effects from shocks, leading to persistent deviations from the mean [22].

Following the stationarity test, the next step is to assess cointegration. Cointegration refers to a statistical property of non-stationary time series whereby two or more series are said to be cointegrated if a linear combination of them yields a stationary process. This concept is crucial because it allows researchers to identify stable long-term equilibrium relationships despite short-term fluctuations.

One of the most prominent applications of cointegration is in financial econometrics, particularly in pair trading strategies. In this context, two assets with a strong historical relationship are paired such that their residual spread—obtained through a linear combination—is stationary. This residual is effectively purged of transient variations caused by market volatility, thereby providing a reliable basis for arbitrage opportunities [23].

Several approaches have been developed to test for cointegration, including the Engle–Granger method (commonly referred to as the two-step estimation procedure), the Phillips–Ouliaris residual-based tests, and the Johansen procedure. Among these, the

Johansen maximum likelihood approach is particularly advantageous, as it overcomes the limitations of the Engle–Granger method by allowing for the simultaneous estimation and testing of multiple cointegrating vectors. This is achieved through the use of canonical correlations and the associated eigenvalue problem [24].

The Johansen framework provides two test statistics: the trace test and the maximum eigenvalue test. While both are designed to assess the number of cointegrating relationships, they differ slightly in their inferential properties and may lead to different conclusions in empirical applications [25].

Under the null hypothesis of the trace test, the number of cointegrating vectors is assumed to be equal to r, where the variable 'r' less than 'k' and k denotes the total number of endogenous variables in the system. The test evaluates this null against the alternative hypothesis that the number of cointegrating vectors is equal to k, implying full rank cointegration. The test statistic is computed as shown in Equation (1).

$$r = r^* < k \quad (1)$$

It is worthy of note to know that this computation was done against the alternative that r = k.

The test proceeds sequentially for r* = 1, 2, 3, ..., and the first non-rejection of the null is taken as an estimate of r. The null hypothesis for the "maximum eigenvalue" test is as for the trace test but the alternative is r = r* + 1 and, again, the test proceeds sequentially for r* = 1, 2, 3. …, with the first non-rejection used as an estimator of r [18].

In addition, Equation (2) shows the Johansen test, while Equation (3) represents the VECM.

$$D\big(Log(Y)\big) = b_0 + b_1 D(Log(X_1 t) + b_2 D(Log(X_2 t) + b_3(Log(Yt\text{-}1) + b_4(Log(X_1 t\text{-}1) + \cdots + ut \quad (2)$$

Where, D is the first differential operator and $D(X_t) = X_t - X_{t-1}$

$$\Delta X_t = \prod X_{t\text{-}1} + \sum_{i=1}^{p-1} \Gamma i \Delta X_{t-i} + CD_t + \varepsilon_t \quad (3)$$

Where, Δx is the first difference of x, $\prod$ is a coefficient matrix of cointegrating relationships. Γi is a coefficient matrix of Δx_{t-i} , C is a coefficient matrix of a vector of deterministic terms D_t. ε_t is an error term with mean zero and variance-covariance matrix $\sum$.

$\prod X_{t-1}$ is the first lag of linear combinations of non-stationary level variables or error correction terms (ECT) which represent long-term relationships among non-stationary level variables.

2.3. Related works on long-run analysis in Tax Administration

Authors in [26] analyzed impatcs the long-run personal income taxation determinants in Croatia. The results indicated a negative but statistically significant relationship between overall economic conditions and personal income tax revenues, alongside a positive and statistically significant relationship between both average monthly wages and the number of taxpayers with personal income tax revenues. These findings are consistent with established principles of economic and public finance theory. Income taxation has long been a central issue in economic policy, as one of the government's primary responsibilities is to design a tax system that ensures adequate revenue generation while promoting social welfare. Additionally, [27] analyzed the nonlinear relationship between economic growth and tax revenue using a hidden cointegration approach. The results demonstrated that tax revenues declined across variables and that a cointegration relationship emerged during periods of GDP growth. Similarly, [28] found evidence of both short-run and long-run relationships between economic growth and tax revenue. Moreover, the study indicated that tax revenue serves as a driver of economic growth, whereas the reverse relationship does not hold.

Despite these contributions, the existing literature has not advanced the forecasting of Niger's tax revenue through long-run analysis. Specifically, while prior studies have examined data mining techniques and the CRISP-DM methodology, as well as cointegration frameworks, these approaches have not been applied to the context of Niger's tax administration.

3. Methodology

The long-run analysis system for Niger's tax administration was developed using the CRISP-DM framework. To tailor the methodology to the specific context of tax administration, several phases of the original framework were renamed. For example, the "Business Understanding" phase was adapted to "Tax Administration Understanding," while the "Data Understanding" phase was revised to "Understanding Tax Data." These modifications ensured that the methodology accurately reflects the operational and analytical requirements of Niger's tax administration.

3.1. Tax Administration Understanding

The Niger tax administration was established on October 13, 1983. Niger, a sub-Saharan country, covers an area of 1,267,000 km². The administration faces a significant staff shortage, as retiring employees are not being replaced. It is organized into multiple departments, including the headquarters, taxpayer departments — subdivided by tax type and by taxpayer size (medium or large) — as well as audit, compliance and inquiry, and legislation departments. Additionally, the tax administration maintains regional and subregional office branches, supported by auxiliary departments such as Information Technology, Human Resources, and Planning and Forecasting.

The tax administration plays a critical role in enabling the government to implement its budgetary policies and ensure effective tax revenue collection. However, its database is neither unified nor fully integrated, as data are sourced from multiple systems. This fragmentation increases the risk of errors during centralization and retrieval, potentially compromising the accuracy of cointegration and long-run analyses. Strengthening data integration and management is therefore essential to enhance the reliability of tax revenue forecasting and policy planning.

3.2. Understanding Tax Data

The dataset used in this research comprises monthly tax records covering the period from January 1996 to December 2014. These records were obtained from the Statistics Division, the administrative unit responsible for centralizing tax data. The dataset encompasses all taxes actively administered by the Nigerien tax authority, which manages a total of 61 different tax types. Tax collection occurs at varying frequencies, viz., daily, weekly, or monthly, depending on the specific tax.

Nigerien tax system operates primarily on a declarative basis. However, certain taxes, such as the communication tax, have been discontinued following political decisions. In addition, some taxes,

particularly those associated with the mining sector, are inconsistently reported within the database. These discontinuities and inconsistencies pose challenges for maintaining comprehensive and reliable long-run tax records.

3.3. Preparation of Tax Data

The principal sources of tax revenue in Niger are the Value Added Tax (VAT), the Individual Tax on Salaries (ITS), and the Business License Tax (ISB). These three taxes represent more than 2/3 (almost equal to 64.8%) of tax revenue in Niger. [29] [30]. These taxes play a critical role in national revenue generation and form the core focus of this study. The dataset employed for the analysis spans a 228 months period, from January 1996 to December 2014, providing a comprehensive temporal scope for evaluating tax trends and cointegration. Key input variables include monthly revenue figures for VAT, ITS, and ISB, along with the total tax revenue. To improve interpretability and ensure consistency in the analysis, all monetary values were transformed into logarithmic form. This transformation aids in managing skewed distributions and stabilizing variance, thereby enhancing the robustness of statistical modeling techniques. The selected variables and preprocessing methods are aligned with best practices in fiscal data analysis.

3.4. Modeling technique

To achieve the objectives of this study, a specific modeling technique along with a set of algorithms were proposed and utilized accordingly. Hence, the proposed modeling technique was the cointegration analysis and the algorithms were the 1) Visualisation (Boxplot), 2) Analysis of extreme values, 3) augmented Dicker-Fuller test, 4) Johansen Cointegration test, and 5) Vector error correction model.

Outliers were identified and visualized using boxplots, and extreme values were analyzed to determine their appropriateness for removal. Stationarity was then assessed using the Augmented Dickey-Fuller (ADF) test, which evaluates the presence of trends and unit roots in the input variables. When a unit root was detected, cointegration tests were conducted to examine the existence of long-run relationships among the variables. Subsequently, the Vector Error Correction Model (VECM) was employed to capture both the short-term dynamics and the long-term equilibrium relationships.

The algorithms, expressed in the form of general formulas, are presented in Table 1.

Table 1. General R Code for Data Visualization, Preprocessing, and Cointegration Analysis

Procedure	General R Code
Data Visualization (Boxplot)	*Boxplot (data, ...)*
Outlier Removal	*Data [-c(x1, x2, ...),]*
Augmented Dickey-Fuller Test	*ur.df (data$x1, type = "drift", selectlags = "AIC")*
Johansen Cointegration Test	*ca.jo (data, type = c("eigen", "trace"), ecdet = c("none", "const", "trend"), K = 2, spec = c("longrun", "transitory"), season = NULL, dumvar = NULL)*
Vector Error Correction Model (VECM)	*VECM (data, lag, r = 1, include = c("const", "trend", "none", "both"), beta = NULL, estim = c("2OLS", "ML"), LRinclude = c("none", "const", "trend", "both"), exogen = NULL)*

3.5. Model assessment and deployment

This study develops and deploys a tax forecasting model for the Nigerien tax administration by applying the cointegration method to datasets both with and without outliers. By incorporating accurate long-run analysis, the model enhances forecasting precision, with key taxes assessed at the 1%, 5%, and 10% significance levels. Furthermore, the model is implemented in R-Studio, utilizing Shiny for the development of an interactive interface and R-Markdown for the generation of static reports. Finally, the system is deployed in consultation with stakeholders and under strict data security protocols to ensure confidentiality and reliability.

3.5.1. Model Assessment

The proposed model is evaluated through a systematic, step-by-step process designed to ensure consistency with the tax administration's objectives. To this end, the assessment relies on a comparative analysis of two datasets: one that retains outliers and another in which outliers have been removed. Specifically, the OLS regression algorithm, the Johansen cointegration procedure, and the vector error correction model (VECM) are applied to both datasets in order to evaluate forecasting accuracy for tax revenue. This comparative approach is particularly important, as it highlights the extent to which data quality influences predictive performance. Moreover, statistical significance levels of 1%, 5%, and 10% are employed to assess the impact of key tax variables, such as VAT, ITS, and ISB, on overall revenue generation. The results reveal that the inclusion of outliers diminishes forecasting accuracy and undermines the reliability of decision-making. Conversely, the model that appropriately manages outliers demonstrates superior performance, rendering it more suitable for fiscal planning. Consequently, rigorous outlier management emerges as a critical requirement in the development of robust forecasting tools for tax administration.

3.5.2. Model Deployment

The model is deployed using R-Studio, utilizing the Shiny package, which is an open-source tool for building interactive web applications. Input variables are imported from a Microsoft Excel database and analyzed through a user-friendly interface that includes widgets, such as text boxes, radio buttons, and drop-down menus. This interface allows stakeholders to interact with the model efficiently and visualize outputs in real time. R-Markdown is used for static report. It is worth point out that the deployment process must be carried out in alignment with business objectives and requires formal approval from the tax administration. To ensure transparency and knowledge transfer, an experience document should be prepared, compiling reports and insights from all project contributors. Security is a critical component of the deployment phase. A comprehensive security framework must be implemented, including user authentication, access control, information classification, identification protocols, and audit trails. These measures are essential to safeguard sensitive fiscal data and ensure compliance with institutional policies and data governance standards.

4. Results and discussion

Data driven tax management uses the outliers detection to make a comparative analysis between a database with outliers and the

same database where outliers are removed. Before the OLS algorthim is used to check the level of significant of the chosen data then the cointegration algorithm determines the behavior of the taxes in the databases.

4.1. Outliers visualization

To enhance the visualization of the previously identified outliers, Figure 1 presents a boxplot in which seven (07) extreme values are depicted as small circles. These values deviate significantly from the rest of the dataset and, therefore, qualify as statistical outliers. Such anomalies exert a considerable influence on both the performance and the accuracy assessment of the linear regression model. More importantly, they negatively affect the long-run cointegration analysis of tax revenue. This impact is evaluated through the application of the Ordinary Least Squares (OLS) algorithm to both datasets, with and without outliers.

Outliers in this context may arise from various sources, including data entry errors, miscalculations in tax assessment, or limitations inherent in Niger's declarative tax system. Over the period from January 1996 to December 2014, three (03) outliers were identified in ITS, one (01) in ISB, and three (03) in VAT. Determining the specific periods and amounts corresponding to these anomalies is essential, as such information enables top management to instruct the audit department to investigate their underlying causes.

Finally, the application of the OLS algorithm in RStudio demonstrates that VAT, ISB, and ITS collectively account for 93.64% of the variation in total tax revenue. These finding underscores both the critical importance of these taxes in revenue generation and the necessity of rigorous outlier detection and correction to ensure reliable econometric analysis.

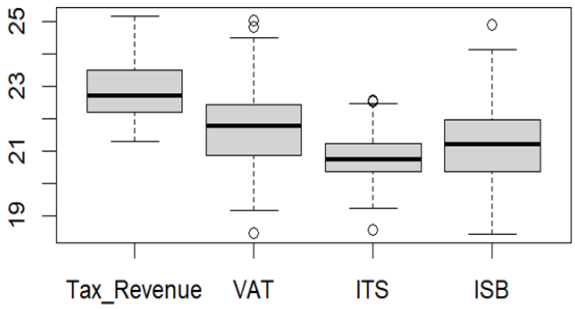

Fig. 1 Graphical visualization of the initial database

Table 2 presents the monthly distribution of outliers for each tax category, along with the specific periods in which these anomalies occur. These values were generated using the extreme value detection algorithm implemented in R-Studio, which not only identifies the magnitude of the outliers but also links them to their corresponding periods within the Nigerien tax administration dataset. It is important to note that the monthly tax amounts are expressed in logarithmic form to ensure scale comparability and improve statistical interpretation.

For VAT, three outliers are observed in January 1999, December 2012, and December 2014, with respective values of 18.47736, 25.03977, and 24.84081. Similarly, ITS displays three outliers occurring in October 2001, December 2013, and July 2014, with respective values of 18.57010, 22.55258, and 22.54660. In contrast, ISB exhibits only one outlier, identified in August 2013, with a value of 24.88954.

The timing of these outliers provides insight into their potential causes. Outliers appearing in December and January are likely attributable to delays in tax declarations, which often coincide with companies' preparation of their annual balance sheets. For ISB, the outlier in August 2013 may be explained by the provisional account period for revenue collection, during which additional controls are frequently implemented. Similarly, the July 2014 outlier for ITS could also be linked to targeted tax audits or administrative controls.

Table 2. Distribution of the outliers

Period	Taxes	Monthly amount
January 1999	VAT	18.47736
July 2014	ITS	22.54660
August 2013	ISB	24.88954
October 2001	ITS	18.57010
December 2012	VAT	25.03977
December 2014	VAT	24.84081
December 2013	ITS	22.55258

4.2. Removal of outliers

The interquartile range (IQR) method was applied in RStudio to identify and remove outliers from the dataset. Following the removal of these anomalies, the Ordinary Least Squares (OLS) analysis reveals that VAT, ISB, and ITS collectively explain 94.23% of the variation in tax revenue. Figure 2 illustrates the visualization of the cleaned dataset using the boxplot algorithm, confirming the successful elimination of extreme values.

For comparison, the OLS analysis conducted on the dataset containing outliers indicates that VAT, ISB, and ITS account for 93.64% of tax revenue variation. In contrast, the dataset without outliers demonstrates a higher explanatory power of 94.23%, reflecting improved model accuracy and statistical significance. These results clearly indicate that the presence of outliers negatively impacts the predictive performance of tax revenue models. Therefore, careful outlier detection and removal is crucial for enhancing the reliability of long-run tax revenue analysis in the Nigerien tax administration for the period from January 1996 to December 2014.

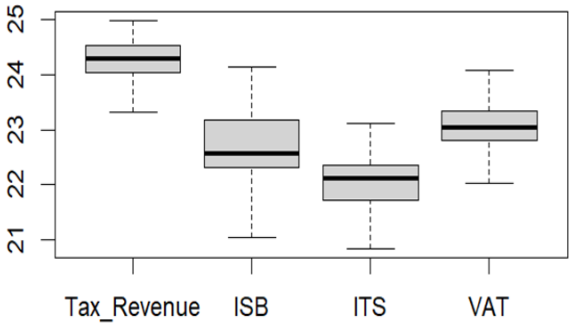

Fig. 2 Removal of outliers from database

4.3. Cointegration (A Long-run Analysis)

4.3.1. Stationary test

The stationarity test is used to determine the presence of a unit root in the time series data. The test is based on two hypotheses: 1) H_0: the series has a unit root (non-stationary), and 2) H_1: the series does not have a unit root (stationary). The decision is made by comparing the Augmented Dickey-Fuller (ADF) test statistic to critical values (CV).

Specifically, if ADF < CV, H_1 is accepted and the series is considered stationary.

Conversely, if ADF > CV, H_0 is accepted, indicating a non-stationary series.

In the initial dataset, the Total Revenue (TR) variable has an ADF value of – 2.2106, which is greater than the critical values at 1% (– 3.46), 5% (– 2.88), and 10% (– 2.57). Therefore, H_0 is accepted, confirming that TR is non-stationary. Similarly, in the dataset without outliers, TR has an ADF value of – 2.1052, which again exceeds all critical values, and H_0 is accepted. These results indicate that TR exhibits a unit root in both datasets. Figure 3 illustrates the non-stationary behavior of TR in both datasets, each displaying a clear upward trend over time.

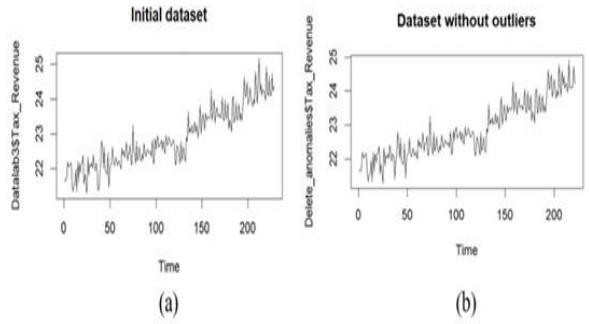

Fig3. Non stationary databases: (a) Tax revenue trend in initial database; (b) Tax revenue in database with no outlier.

4.3.2. Cointegration

The advantage of cointegration in tax administration is that it provides an analysis method for non-stationary taxes by avoiding spurious regressions put in evidence by Granger and Newbold in 1974. The process of the cointegration is based on the hypothesis and comparison of test statistic with the critical values:

H0: Non cointegration (rank of cointegration = 0)

H1: Cointegration (rank ≥ 1)

4.3.2.1. Database with outlier

In the case of dataset with outlier, when the value of test statistic is 192.50, the benchmark value at 10%; 5% and 1% are 49.65; 53.12; 60.16. When the value of test statistic is 98.40, the benchmark value at 10%; 5% and 1% are 32.00; 34.91 and 41.07. When the test statistic is 47.34, the benchmark value at 10%; 5% and 1% are 17.85; 19.96 and 24.60. At these stages the test statistic is higher than the benchmark at it implies that there is no cointegration.

But when the test statistic is 6.13, the benchmark at 10%; 5% and 1% are 7.52; 9.24 and 12.97; the value of test statistic is lower than the benchmark thus there is cointegration with a rank equals to 3 (r = 3). The variables of this dataset have shown a long-run relationship. Table 3 shows the cointegrations ranks with their test statistic and threshold for the database with outlier

Table 3: Database with outlier cointegration rank

Cointegration rank	Test statistic	10%	5%	1%
r <= 3	6.13	7.52	9.24	12.97
r <= 2	47.34	19.96	17.85	24.60
r <= 1	98.40	32.00	34.91	41.07
r = 0	192.50	49.65	53.12	60.16

4.3.2.2. Database without outlier

In the database without outlier, when the value of test statistic is 218.90, the benchmark value at 10%; 5% and 1% are 49.65; 53.12 and 60.16. When the value of test statistic is 83.93, the benchmark value at 10%; 5% and 1% are 32.00; 34.91 and 41.07. When the test statistic is 28.97, the benchmark value at 10%; 5% and 1% are 17.85; 19.96 and 24.60. At these stages the test statistic is higher than the benchmark at it implies that there is no cointegration.

But when the test statistic is 3.60, the benchmark at 10%; 5% and 1% are 7.52; 9.24 and 12.97; the value of test statistic is lower than the benchmark thus there is cointegration with a rank equals to 3 (r = 3). The variables of this dataset have long run relationship. Table 4 shows the cointegrations ranks with their test statistic and threshold in the database without outlier.

Table 4: Database without outlier cointegration rank

Cointegration rank	Test statistic	10%	5%	1%
r <= 3	3.60	7.52	9.24	12.97
r <= 2	28.97	17.85	19.96	24.60
r <= 1	83.93	32.00	34.91	41.07
r = 0	218.90	49.65	53.12	60.16

Both the two databases (initial database and database without outlier) present cointegration. Thus, the total revenue in each database is cointegrated with its dependants variables (taxes) and showing a long run relationship among taxes.

4.3.3. Vector error correction model (VECM)

Because of the error correction model, cointegration theory makes it possible to simultaneously model the long-term and short-term dynamics of taxes. VECM measures the time a disequilibrium will take before returning to its equilibrium. There are four (04) taxes that are used for the research so the vector error correction model is more appropriate.

4.3.3.1. Database with outlier

The error coefficient term is -0.3728 and is negative and it is significant at 1%. So, it is convenient to accept the specification of the type vector error correction model. Indeed, it implies that there exists an error correction mechanism in long run the disequilibrium in between Total tax revenue, VAT, ITS and ISB compensate in such that the series have the similar evolution (trend). -0.3728 represents the speed at which any disequilibrium in between the desire and effective levels of Total tax revenue is reabsorbed in a period (month) following a shock. 37.28% can be adjusted in between the desire and effective level of Total tax revenue. 37.28% effects of shock intervening in each month is reabsorbed in a month

following this shock. Thus, the shock on the Total tax revenue is corrected at 37.28% by the feedback effect. In other words, one shock found during one month is totally reabsorbed in 1/0.3728 = 2.68 months (81 days).

4.3.3.2. Database without outlier

The error coefficient term (ECT) is -0.5009, it is negative and it is significant at 1%. So, it is convenient to accept the specification of the type vector error correction model. Indeed, it implies that there exists an error correction mechanism in long run the disequilibrium in between Total tax revenue, VAT, ITS and ISB compensate in such that the series have the similar evolution (trend). -0.5009 represents the speed at which any disequilibrium in between the desire and effective levels of Total tax revenue is reabsorbed in a period (month) following a shock. 50.09% can be adjusted in between the desire and effective level of Total tax revenue. 50.09% effects of shock intervening in a given month is reabsorbed in a month following this shock. Thus, the shock on the Total tax revenue is corrected at 50.09% by the feedback effect. In other words, one shock found during one month is totally reabsorbed in 1/0.5009 = 1.99 months (60 days).

4.4. Proposed work

Figure 4 presents the flowchart of the proposed framework to ensure the model's effectiveness within the Niger tax administration. It outlines the entire process from the input of tax data to the dataset analysis and its printing out.

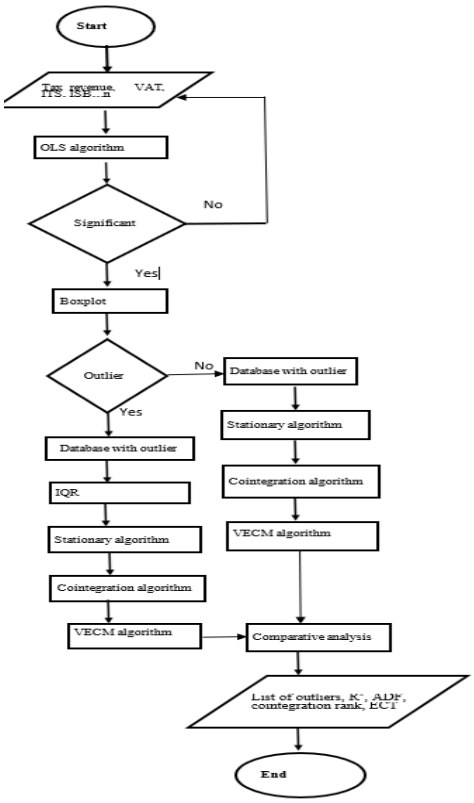

Fig 4. Flowchart of the work

5. Conclusions

The study successfully integrated disparate databases into a unique database. To improve the fiscal policy and the financial planning for the Niger tax administration a long run analysis system was developed through outlier detection. The research employed the CRISP-DM methodology alongside linear regression, boxplot analysis, the interquartile range method, augmented Dickey-Fuller test for stationary, cointegration test and the vector error correction model test.

A key finding of this study is the significant impact of accurate data processing on revenue predictions. In cleaned dataset, key taxes variables are: 1. Value added tax (VAT), 2. Individual tax on salary (ITS), and 3. Business income tax (ISB) accounted for 94.25% of variance of tax revenue. In contrast, in the dataset with outlier these variables explained 93.64% of the variance during the period from January 1996 to December 2014. The two datasets are non-stationary and are cointegrated with a cointegration rank of 3 each. But it will take eighty-one (81) days for tax revenue in dataset with outlier to reabsorb a shock found during one month whereas it will take sixty (60) days for tax revenue in database without outliers to reabsorb a shock found during one month. This underscores the importance of long run analysis through outlier detection in ensuring reliable fiscal data. Outliers in tax collection negatively impact Niger tax administration which is in line with Franses in 1998 who made research on outlier detection in cointegration analysis.

References

[1] A. T. Olufemi, O. Jayeola, A. S. Oladele, and A. O. Naimot, 'Tax revenue and economic growth in Nigeria', *Sch. Int. J. Manag. Dev.*, vol. 5, no. 7, pp. 72–85, 2018.

[2] '"statistiques-recettes-publiques-afrique-niger.pdf",' 2024. [Online]. Available: https://www.google.com/search?q=statistiques-recettes-publiques-afrique-niger.pdf

[3] F. Marro-Dauzat, 'La pression fiscale dans l'Union européenne', Touteleurope.eu. 2023. [Online]. Available: https://www.touteleurope.eu/economie-et-social/la-pression-fiscale-dans-l-union-europeenne/

[4] L. Lin, 'Application of Big Data Model in Financial Taxation Management', *Sci. Program.*, vol. 2021, no. 1, p. 7001456, 2021, doi: https://doi.org/10.1155/2021/7001456.

[5] 'L'Union Européenne et le Niger | EEAS'. 2023. [Online]. Available: https://www.eeas.europa.eu/niger/lunion-europeenne-et-le-niger_fr?s=113

[6] F. Y. Mpofu, 'Prospects, Challenges, And Implications Of Deploying Artificial Intelligence In Tax Administration In Developing Countries', *Stud. Univ. Babes Bolyai-Negot.*, vol. 69, no. 3, pp. 39–78, 2024.

[7] I. EL Yamlahi, N. B. Amine, and H. EL Ghazlani, 'La contribution de l'intelligence artificielle au développement de la direction générale des impôts au Maroc', *Rev. Econ. Kap.*, vol. 1, no. 21, 2022.

[8] J. Cheng, 'Data-Mining Research in Education', Oct. 25, 2017, *arXiv*: arXiv:1703.10117. doi: 10.48550/arXiv.1703.10117.

[9] K. Sabanci, M. F. Unlersen, and M. Koklu, 'Classification of Heuristic Information by Using Machine Learning Algorithms', *Int. J. Intell. Syst. Appl. Eng.*, vol. 4, no. Special Issue-1, pp. 252–254, Dec. 2016, doi: 10.18201/ijisae.2016SpecialIssue-146984.

[10] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2022.

[11] S. Moro, R. Laureano, and P. Cortez, 'Using data mining for bank direct marketing: an application of the CRISP-DM

methodology', EUROSIS-ETI, Oct. 2011. [Online]. Available: https://repositorium.sdum.uminho.pt/handle/1822/14838

[12] M. Goebel and L. Gruenwald, 'A survey of data mining and knowledge discovery software tools', *ACM SIGKDD Explor. Newsl.*, vol. 1, no. 1, pp. 20–33, June 1999, doi: 10.1145/846170.846172.

[13] G. C. Onwubolu and D. Davendra, *Differential Evolution: A Handbook for Global Permutation-Based Combinatorial Optimization*. Springer, 2008.

[14] T. Darmawan, 'Credit Classification Using CRISP-DM Method On Bank ABC Customers', *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 6, pp. 2375–2380, June 2020, doi: 10.30534/ijeter/2020/28862020.

[15] S. Studer *et al.*, 'Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology', *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, June 2021, doi: 10.3390/make3020020.

[16] V. Meuriot, 'The concept of cointegration: The decisive meeting between Hendry and Granger (1975)', *Pap. Polit. Econ.*, vol. 68, no. 1, pp. 91–118, 2015, doi: 10.3917/cep.068.0091.

[17] S. Washington, M. G. Karlaftis, F. Mannering, and P. Anastasopoulos, *Statistical and Econometric Methods for Transportation Data Analysis*, 3rd edn. New York: Chapman and Hall/CRC, 2020. doi: 10.1201/9780429244018.

[18] A. Suharsono, A. Aziza, and W. Pramesti, 'Comparison of vector autoregressive (VAR) and vector error correction models (VECM) for index of ASEAN stock price', *AIP Conf. Proc.*, vol. 1913, no. 1, p. 020032, Dec. 2017, doi: 10.1063/1.5016666.

[19] M. Usman, D. F. Fatin, M. Y. S. Barusman, F. A. M. Elfaki, and Widiarti, 'Application of Vector Error Correction Model (VECM) and Impulse Response Function for Analysis Data Index of Farmers' Terms of Trade', *Indian J. Sci. Technol.*, vol. 10, no. 19, May 2017, doi: 10.17485/ijst/2017/v10i19/112258.

[20] P. H. Franses and A. Lucas, 'Outlier Detection in Cointegration Analysis', *J. Bus. Econ. Stat.*, vol. 16, no. 4, pp. 459–468, Oct. 1998, doi: 10.1080/07350015.1998.10524785.

[21] I. Georgoula, D. Pournarakis, C. Bilanakos, D. N. Sotiropoulos, and G. Giaglis, 'Using Time-Series and Sentiment Analysis to Detect the Determinants of Bitcoin Prices', *MCIS 2015 Proc.*, Jan. 2015, [Online]. Available: https://aisel.aisnet.org/mcis2015/20

[22] C. Chikwira and J. I. Mohammed, 'The Impact of the Stock Market on Liquidity and Economic Growth: Evidence of Volatile Market', *Economies*, vol. 11, no. 6, p. 155, June 2023, doi: 10.3390/economies11060155.

[23] G. Vidyamurthy, *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons, 2004.

[24] F. Bilgili, 'stationarity and cointegration tests: comparison of engle -granger and johansen methodologies', *Erciyes Üniversitesi İktisadi Ve İdari Bilim. Fakültesi Derg.*, no. 13, pp. 131–141, Dec. 1998.

[25] A. Tversky and D. Kahneman, 'Rational Choice and the Framing of Decisions', in *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*, B. Karpak and S. Zionts, Eds, Springer Berlin Heidelberg, 1989, pp. 81–126.

[26] I. Palić, K. Dumičić, and B. Grofelnik, 'Analysis of Personal Income Taxation Determinants in Croatia in Long Run: Evidence from Cointegration Analysis', *Naše Gospod. Econ.*, vol. 63, no. 3, pp. 12–18, Sept. 2017, doi: 10.1515/ngoe-2017-0014.

[27] G. Çiğdem and M. Altaylar, 'Nonlinear Relationship between Economic Growth and Tax Revenue in Turkey: Hidden Cointegration Approach', *İstanbul İktisat Derg.*, vol. 71, no. 1, pp. 21–38, July 2021.

[28] W. Takumah, 'Tax Revenue and Economic Growth in Ghana: A Cointegration Approach'.2014. [Online]. Available: https://mpra.ub.uni-muenchen.de/58532/

[29] P. Vandenberghe, J. L. Carvajal, and M. Kabaka, 'Rapport d'Évaluation de la Performance', 2022, [Online]. Available : https://www.tadat.org/content/dam/tadat/en/assessments/Niger_Replublic_TADAT_%202022.pdf

[30] M. Khane, H. Naroua, C. Kadri, and Y. Moumouni, 'Enhancing Tax Administration in Niger: A Data Mining Approach to Outlier Detection', *International Journal of Intelligent Systems and Applications in Engineering*, p. 233, 2025.