IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Fraud Detection on Bulk Tax Data Using Business Intelligence Data Mining Tool: A Case of Zambia Revenue Authority

Memorie Mwanza[1] **and Jackson Phiri** [2]

Department of Electrical and Electronics, University of Zambia, Lusaka, Zambia [1]

Department of Computer Science, University of Zambia, Lusaka, Zambia [2]

Abstract: Zambia Revenue Authority (ZRA) generates large volumes of data that need complex mechanisms in order to extract useful tax information. The purpose of the study was to develop a data mining model for detection of fraud on tax and taxpayer data for ZRA. This study focused on two areas. These were (1) the baseline study that helped to establish the extent of the challenges in fraud detection for the tax payers and (2) the automation and development of the fraud detection tool using the results from the baseline study. Our baseline study showed that the current methodologies, processes, architectures, and technologies that were being used to transform raw data into meaningful and useful information were tedious and time consuming. In order to detect fraud they depended on random audits, informants and under-cover operations. A model which implements outlier algorithms for fraud detection, Continuous Monitoring of Distance Based and Distance Based Outlier Queries was then developed. We used both algorithms to analyse the domestic tax payments to detect underpayments and overpayments according to business rules. Underpayments and overpayments are marked as outliers. Results generated by our tool showed improved accuracy and takes less time in order to detect under and over payments as outliers when compared to the older methods.

Keywords: Business Intelligence, Data mining, fraud detection, outlier algorithm.

I. INTRODUCTION

Around the world today, [1] tax authorities are experiencing growing pressure to collect extra tax revenues, to discover underreporting taxpayers, and predict the irregular behavior of non-paying taxpayers. Most tax authorities require to collect tax data from a number of independent sources and perform data matching and checking with other sources to find cases of non-compliance. As a result, tax evasion detection performance has been rather limited in the absence of information technology tools.

The amount of business data that is generated has risen steadily every year [2] and more and more types of information are being stored in unstructured or semi structured formats. Traditional data mining has no power anymore to deal with the huge amount of unstructured and semi structured written materials based on natural languages. Zambia Revenue Authority is a quasi-governmental organization which is mandated to collect revenue on behalf of the Government of the Republic of Zambia. Its mandate is derived from the Zambia Revenue Authority Act 321, Section 11 (1), (2) and (3) [3]. Some of its main responsibilities include;

a) To properly assess and collect taxes and duties at the right time.
b) To ensure that all monies collected are properly accounted for and banked.

ZRA houses several critical corporate Databases used to store enormous business data commonly known as Taxes and Taxpayer Information. Information sits across different databases in different Division and Departments. Therefore, Business intelligence (BI) generally, and data mining in specific, may be effective tools for enhancing the efficiency and effectiveness of the detection of illegal activities in relation to Taxes. In this paper, a data mining model as a Business Intelligence tool for detection of fraud on tax data and analysis of bulk data for Zambia Revenue Authority is proposed. The rest of the paper is organised as follows; Section 1.0 gives the

Introduction, Section 2.0 gives the literature review, Section 3.0 is the methodology. The section that follows, Section 4.0 gives the implementation results whilst Section 5.0 gives the conclusion.

II. LITERATURE REVIEW

A. Introduction

Fraud in the financial sector which includes the tax administration sector, [4] is increasingly becoming a serious problem and as a result, this dishonest performance of taxpayers influences negatively the incomes available to public services as well as creating harm on the honest taxpayers. Therefore [5], no society can confirm that it is free from Fraud in its various form. Governments, irrespective of whether they are public or private, local or multinational, huge or small, they are affected by this reality of fraud, which seriously undermines the principles of harmony and fairness of citizens before the law and threatens business.

Currently, more and more companies are using BI tools to analyze sales and other related transactional data to detect fraud.

DOI 10.17148/IJARCCE.2016.53191

This section will therefore look at application of Business Intelligence and data mining in regions beginning with America, Asia, and will finally look at data mining in African scenarios.

B. BI and Data Mining in America

American government have incorporated the techniques of data mining and artificial intelligence into the audit planning activities mainly to detect patterns of fraud or evasion, [6] [7] which are used by tax authorities for specific purposes.

The internal revenue service (IRS), [6] [7] the institution responsible for administering taxes in the United States, has also used data mining techniques for various purposes, among which are measuring the risk of taxpayer compliance, the detection of tax evasion and criminal financial activities electronic fraud detection, detection of housing tax abuse, detection of fraud by taxpayers who receive income from tax credits and money laundering.

Further, [8] Texas and several other states, as well as tax agencies in the United Kingdom and Australia, rely on data mining to help find delinquent taxpayers and make effective resource allocation decisions.

C. BI and Data Mining in Asia

In Indian, the Chairman of the Central Board of Direct Taxes [9] confirms that Income Tax Department (ITD) has embarked on an ambitious computerization plan including implementation of a comprehensive Data Warehouse and Business Intelligence (DW & BI) to improve taxpayer services, promote voluntary compliance and deter tax evasion. The main objective of this plan when implemented is to [9] Discover non-filers with potential tax liabilities, Identify potential under-reporting taxpayers, Improving compliance of tax deductors, Identify non-compliance in service sector and implicit linkages for effective investigation

In turkey, August 2008, [10] the BI, Data Warehouse technologies were used to reduce the number of noncompliant taxpayers and ease the burden on audit units. The use of the automated system and data warehousing have resulted in a substantial decrease in the number of noncompliant taxpayers.

D. Data Mining in Developing Countries: Africa

Today, African governments and their public sector agencies everywhere have not been spared from the pressure to perform more efficiently and effectively. Previously, the traditional methods for addressing risk have served many authorities well, but there is now a need to use more advanced technologies to combat fraud, error and waste in the Tax Administration sector such as Business Intelligence, Data Warehouse and Data Mining. However, not much research has been done on these technologies in Africa as indicated by the Google, Google scholar and Science Direct Search where every research returns very little about the Data Mining and the applied algorithms in the Tax Administration sector.

Botswana, a rapidly developing country with many new organisations establishing presence every year also acknowledges the challenges to analyse data effectively and efficiently in order to gain important strategic and competitive advantage [11].

In Tanzania, Arusha region, [12] a research study indicates and suggests that Data mining is very important for healthcare as it can improve the industry as well as the well-being of the residents.

To arm themselves for this battle, [13] more and more tax authorities have turned to data mining and analytics to improve their business processes, resulting in better compliance.

The Zambian tax system [14] broadly comprises income taxes, consumption taxes, property taxes and trade taxes. These taxes are collected by the Zambia Revenue Authority (ZRA) which is the corporate body mandated to collect all taxes.

TABLE 1 TAX CATEGORIES IN ZAMBIA

	Tax category	Type of tax
1	Income taxes	Company income tax
		Pay As You Earn (PAYE)
		Withholding tax
		Mineral royalty
2	Property taxes	Property Transfer Tax
3	Consumption taxes	Import and domestic VAT
		Excise duties
4	Trade taxes	Customs duty
		Export duty

Developing countries including Zambia has not been spared from obstructions to achieving its key objectives of taxation, such as intelligent exploitation of data (data mining and predictive analysis) [15] emphasised during a presentation at a Workshop Jointly Organized by the World Bank Institute – PRMPS in South Africa.

Currently, ZRA does not have a comprehensive Business Intelligence and Data Warehouse technologies implemented. Data mining is also not comprehensively done. Fraud detection is currently handled using targeted and random Audits and Inspections.

Business Intelligence, Data Warehouse and Data Mining are therefore underlying strategic techniques for tax administrations to discover useful knowledge in support of their Tax Frauds detection, Tax Evasion and compliance enhancing agendas.

Hence, this research study focused on the development of a data mining model as a Business Intelligence tool for detection of fraud on tax data and analysis of bulk data for Zambia Revenue Authority. Further, achieving this objective was through examining the current methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information. This study however focused on the Domestic Taxes Data only.

III. RELATED WORK

Outlier detection has been studied broadly by the statistics community [16], where the objects are modelled as a distribution, and objects are marked as outliers depending

DOI 10.17148/IJARCCE.2016.53191

on their deviation from this distribution. The problem of outlier detection has also been addressed by the database and data mining communities [16], aiming at solving the problem of scalability.
The Banking Industries have not been spared from the need to keep up with their constantly changing industry to stay viable and competitive [17]. Systems previously involved manual recording of branch transactions and the generation of reports from manual ledgers according to [18] these had to be consolidated with other branches into a final report.

In higher education, BI and Data Mining is becoming an inevitable trend as seen from the fact that some universities have applied BI systems as there educational systems while others focus on the research of BI, [19]. Data mining techniques helps in increasing student's retention rate, increase educational improvement ratio, and increase student's learning outcome [20]. Thus, data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationship which help in effective decision making [20].

Today, specifically within higher education BI is viewed as a solution with much promise in regard to adding much needed efficiency on an operational level.
In health Care Industries, Information has been referred to as the life blood of healthcare as it is essential for effective clinical and administrative decision making [21]. Healthcare decision making is complex and requires access to a wide array of high-quality information [22].

University of Tennessee [23] Medical Center hospitals have introduced a new tool that takes data from the hospital's electronic health record system and clusters patients into different risk levels. It also assesses historic data to determine care strategies that have worked in the past for certain kinds of patients. In Poland, applications of Business Intelligence (BI) in different areas of the economy has been growing from year to year [24]. Recently, Poland has seen an increase in the use of the BI systems is the healthcare area.

IV. METHODOLOGY

A. Baseline Study.
A. In this study 200 Questionnaires were distributed to employees from the three (3) different business layers within ZRA namely, Domestic Taxes, Customs Services and Information Technology. Additionally, oral interviews were also used as primary sources of data. The center of this study was on providing answers to the current methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information and also designing a model which will be used to detect fraud on Taxpayer information in ZRA.

B. Business Intelligence and Data mining Model Design and Implementation.
The ZRA Fraud detection model was implemented based on a three-tier architecture design. The design segmented an application's components into three tiers of services, namely Presentation, Logic and lastly Data access.

IV.B.1 Software architecture

Presentation -The presentation tier, also called user services layer, gives a user access to the application. This layer presents data to the user. It also allows a user to setup software settings such as a desired number of payment records to be analysed for fraud, the start and stop algorithm function and also the visualisation function of the application.

Logic -The middle tier, also known as business services layer, consists of business and data rules. The two algorithms used in this design (Continuous Monitoring of Distance Based and Distance Based Outlier Queries) are implemented by this tier.

Data Access -The data tier, or data services layer, implements the data storage in this case the zra_bi database which contains the Tax Payer Profile and the payments activities.

IV.B.3 Application Use Cases, ZRA Fraud detection
User Use Case
This use cases depict the activities that a user will be able to carry out on the ZRA Fraud detection application. User will be able to;

- Start the algorithm,
- Stop the algorithm,
- Print the Screen: a user can take a print screen of the visualization results of the two algorithms,
- Export the results into a different file format: a user can export the results of the analysis to an external file and this file contains all the payments that are marked for fraud,
- View the process time: a user can view the time it takes to process the data.

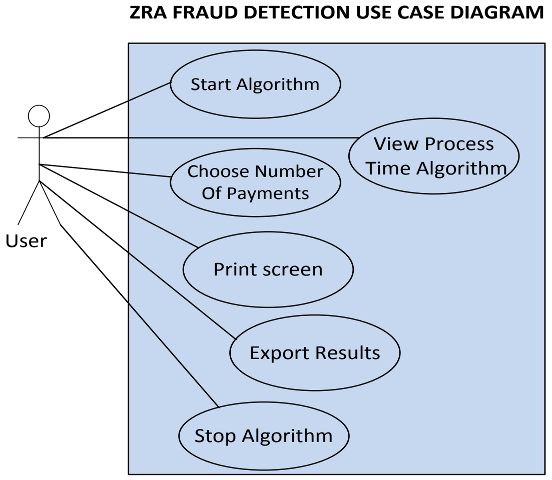
ZRA FRAUD DETECTION USE CASE DIAGRAM
Start Algorithm
View Process Time Algorithm
Choose Number Of Payments
User
Print screen
Export Results
Stop Algorithm

Fig.1. Use Case Diagram for User

Administrator Use Case
This use cases describe the activities that an administrator will be able to carry out on the ZRA Fraud detection application. The administrator will be able to;

- Manage the Business rules.
- Carry out system maintenance.

DOI 10.17148/IJARCCE.2016.53191

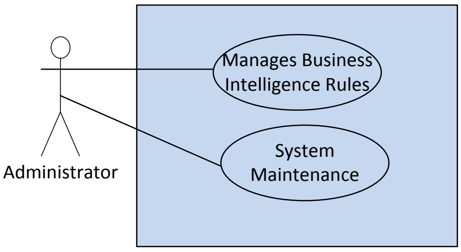

Fig.2. Use Case Diagram for Administrator

System Owner Use Case
This use cases describe the activities which the system owner will be able to carry out on the ZRA Fraud detection application.

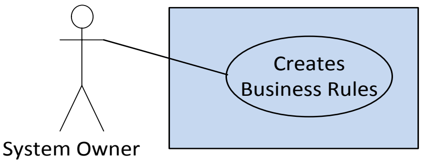

Fig.3. Use Case for System Owner

The system owner will be able to;
Create Business rules that will be used to measure against the behaviour of Taxpayers. Some of the rules that apply are [14];

- Property Transfer Tax at rate of 10% of the realisable value (Price at the time of transfer).
- Turnover Tax on business for companies and Individuals whose turnover is below per annum, the applicable rate is 3% of the turnover.
- Company Tax levied on all incorporated businesses whose turnover is above ZMK800, 000 per annum. Tax charged is, charitable 15%companies 35%, Farming 15%Mobile Telecom sector the first ZMW 250,000 profit at 35% Above ZMW 250,000 profit then 40%
- PAYE charged on income from employment including salaries and wages, overtime, bonuses is charged at the first K3000 at 0%, between 3001 to K3800 25%, K3,801 to K5,900 30% lastly above K5,900 at 35%.

IV.B.2 Classes for ZRA Fraud detection Application

- ZRAFDMain Frame -This is the main window of the ZRA Fraud Detection Application.
- Main Tab Panel -This tab holds the Outlier Algorithm User functionality
- Setup Tab -This tab allows a user to setup system options before running the two algorithms e.g. setting up the desired number of payment records to be analyzed for fraud.
- ZRAFD Controller- The controller is a class that coordinates all the three layers of the system. It gets all the user input from user interface and pass it to the logic layer for processing. It is also responsible for getting the data from the data access layer and pass it to the logic layer for processing. All processing results from the logic layer are sent back to the presentation layer by the ZRAFD Controller class. The application is implemented like this to reduce coupling among layers of the software.

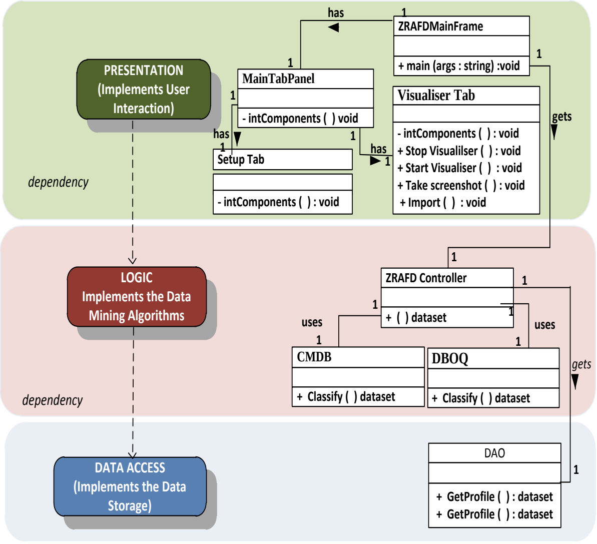
PRESENTATION (Implements User Interaction)
has
1
ZRAFDMainFrame
+ main (args : string) :void
1
MainTabPanel
- intComponents ( ) void
Visualiser Tab
- intComponents ( ) : void
+ Stop Visualiser ( ) : void
+ Start Visualiser ( ) : void
+ Take screenshot ( ) : void
+ Import ( ) : void
gets
has
Setup Tab
- intComponents ( ) : void
dependency
LOGIC Implements the Data Mining Algorithms
ZRAFD Controller
+ ( ) dataset
uses
uses
CMDB
+ Classify ( ) dataset
DBOQ
+ Classify ( ) dataset
gets
dependency
DATA ACCESS (Implements the Data Storage)
DAO
+ GetProfile ( ) : dataset
+ GetProfile ( ) : dataset

Fig. 4. ZRA Fraud detection Class Diagram

- Visualiser Tab -This tab will enable the algorithms to be run and thereafter will be able to show the visual results
- Continuous Monitoring Distance Based (CMDB)- This is the class that implements the continuous monitoring of distance based algorithm
- Distance Based Outlier Query- This class implements the distance based Outlier Queries.
- Data Access- This will return data from MySQL database

IV.B.3 Entity Relationship Diagram
The figure below is a ZRA BI Entity Relationship Diagram (ERD) comprising of the profiles and the payment entities and their attributes.

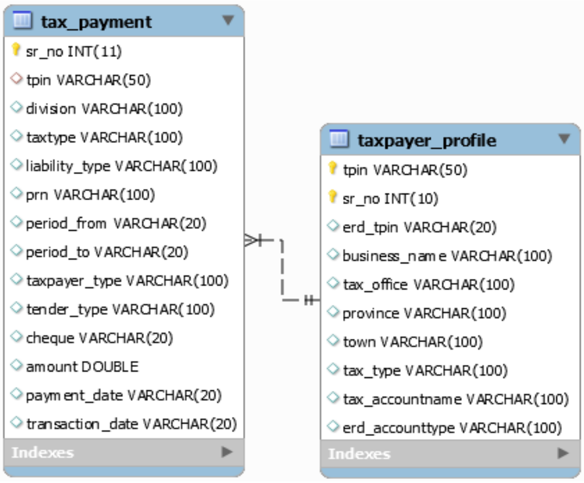

Fig.5. Entity Relationship Diagram

V. FINDINGS

The results in this section looks at the outcome of the survey based on the questionnaires and the oral interviews. The results are organised according to various and specific areas that the study set out to answer

A. Baseline Study

V.A.1 Tools used to analyses data currently in ZRA

A greater majority of the respondents, 78.2%, indicated that the data was currently analysed and turned into something that can be interpreted using Spreadsheet Analysis e.g. Excel. There were 17.9% that indicated software that allow own specific queries was used, 3.8% indicated that software that takes raw data and creates visualization e.g. pie chart software was used.

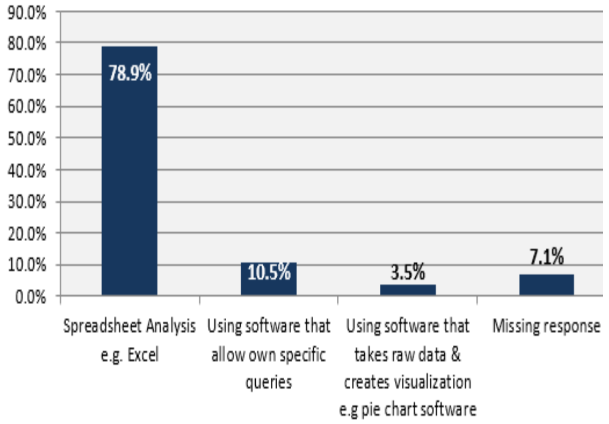

Fig.6. How Data is currently analysed

V.A.4. How fraud is currently being detected on taxes

There were 41% who indicated Targeted & random audits, only 11.5% respondents that indicated they detected fraud through data mining techniques. Using informants to detect fraud was indicated by 14.7% respondents and using under-cover operations by 32.7%.

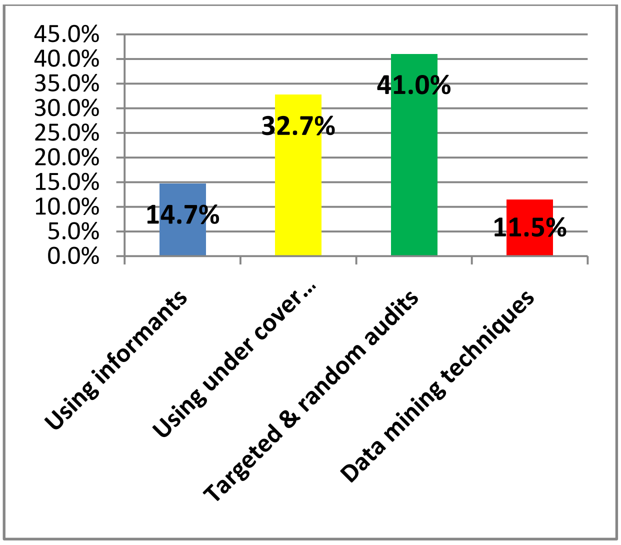

Fig.7. Current Methods of detecting Frauds

B. Model Implementation Results

V.B.2. Visualizer screen for ZRA Fraud detection

Figure 8 below shows the visualizer screen and it allows Users to run the algorithms and be able to see the visual results. Users are able to start and stop the visualization process using this window. All the points in colour red represents payments detected as outliers. Outlier payments are potential fraudulent. The left window shows results of the Continuous Monitoring of Distance Based algorithm and the right window shows Distance Based Outlier Queries algorithm.

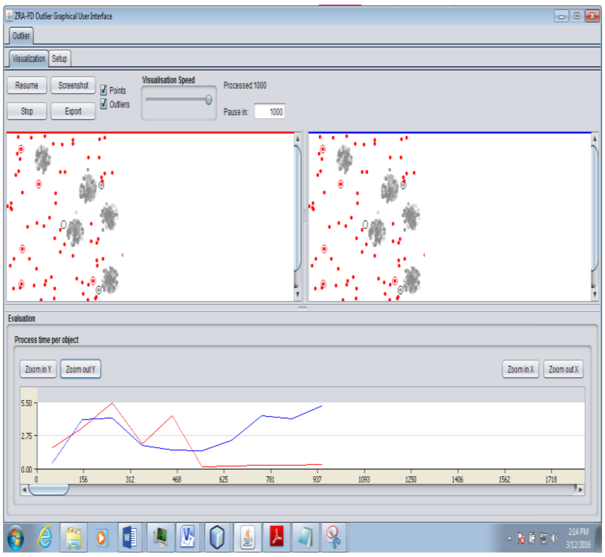

Fig.8. Visualisation Screen

V.B.4.Report of all payments

The developed tool allows us to expert the data into a spreadsheet. A Report of all payments that seem suspicious is finally produced for further analysis.

VI. DISCUSSION

This research study indicates that 78.2%, respondents confirm that data was currently analysed using Excel whilst 17.9% indicated they used software that allow to make own specific queries and 3.8% indicated that software that takes raw data and creates visualization e.g. pie chart software.

Although there is already an indication that demonstrates the power of analytics in the applications such as Tax Online and Asycuda World, there is however a growing acknowledgement of the need to implement Business Intelligence and Data mining in a comprehensive way in the day-to-day operations of ZRA in order to improve the way fraud is detected.

This study further intended to discover how fraud is currently being detected on Taxes. There was an indication by respondents on the methods used to detect fraud such as 41%, Targeted & random audits, 11.5% data mining techniques. Using informants to detect fraud was indicated by 14.7% respondents and using under-cover operations 32.7%.

Currently ZRA has moderated these challenges by putting up Units that support the functions of Data Analysis. Some of these units are Business Support System unit and Business intelligence Unit in domestic Taxes and also Data Management Intelligent Unit in Customs Services Division. These units are involved in Data Management including analysis and reporting at a higher level.

DOI 10.17148/IJARCCE.2016.53191

Feature that can further be incorporated into the application is Data Warehouse technologies. This will be able to combine data from various sources such as Asycuda World, TaxOnline, Legacy Systems such as ITAS and Asycuda++ and later consolidate plentiful variables into easy-to-interpret, actionable classifications and predictions.

The other enhancements that can be incorporated is the automated reporting with interactive dashboards feature which indicates the status of things at a specific point in time and a scorecard, on the other hand, to display progress over time towards specific goals of the organization. This therefore will mean that Tax Administrators will be better informed and their control measures and the hit rate of detecting fraud and errors which includes underpayments and underreporting, will improve.

VII. LIMITATIONS

Despite all of these positive creativities, conducting this study properly had a number of limitations some of which are; limitation of financial resources as a self-sponsored student. Limitation of time since this research study was conducted whilst working on a full time basis. There were too many procedures involved to finally have authorisation to study this topic by ZRA because of the sensitive nature of data involved such as Taxpayers data. Extraction of Data from the Data Sources (TaxOnline and Asycuda World) was so much of a challenge because of its size. It took seven days for Data to finally be ready and available for use. Data was also not normalized and was not clean. A lot of duplicates records were found.

VIII. CONCLUSION

This study confirms that Data mining is a key to many of the shortcomings of the traditional approach in combating error and fraud, and it also gives reasons to believing that data mining could meaningfully contribute to making the tax administrations fraud detection or anomaly detection more effective for tax administration in Zambia.

Using Data mining, Outlier algorithms can “learn” from the existing patterns of data for ZRA taxpayer’s payments and their profiles and are able to pick the anomalies in the data such as the values that lie outside the normal and acceptable regions based on the defined business rules. This knowledge learned using the outlier algorithms can also be used to predict the taxpayer behaviour in future cases under such similar circumstances.

The purpose of this study was to establish the extent of the challenges in fraud detection for the tax payers and also automation and development of the fraud detection tool using the results from the baseline study and data mining.

ACKNOWLEDGMENT

Mwanza M., Author, would like to thank the Zambia Revenue Authority for permitting a Research Study based on Tax Frauds.

REFERENCES

[1] O. b. H. y. L. S.-I. C. b. D. C. Y. c. Wua R. S., "Using data mining technique to enhance tax evasion detection performance," Expert Systems with Applications, An International Journal., no. 39, p. 8769–8777, 2012.

[2] M. C. Kanakalaksmil C., "A concise study on Text Mining for Business Intelligence.," International Journal of Advanced Research in Computer and Communication Engineering., vol. 4, no. 6, 2015.

[3] Resource and Planning Department, "Zambia Revenue Authority Website," ZRA, 2012 - 2013. [Online]. Available: https://www.zra.org.zm/commonHomePage.htm?viewName=organizationStructure. [Accessed 10 March 2016].

[4] M. T. Ameur F., "Taxpayers Fraudulent Behavior Modeling the Use of Datamining in Fiscal Fraud Detecting Moroccan Case," Scientifc Research, Applied Mathermatics, 2012.

[5] V. J. Castellón G. P., "Characterization and detection of taxpayers with false invoices using data mining techniques," Expert Systems with Applications, vol. 40, p. 1427–1436, 2013.

[6] V. J. González P. C., "Characterization and detection of taxpayers with false invoices using data mining techniques,," Expert Systems with Applications, International Journal, vol. 40, p. 1427–1436, 2013.

[7] Martikainen J., "Data Mining in Tax Administration - Using Analytics to enhance Tax Compliance," 2012.

[8] R. S. Micci-Barreca D., "Improving tax administration with data mining".

[9] Tewari R.K., "Data Mining and other Application in financial and Tax Crime Investigations: Experience of India," Inter-American Center of Tax Administrations - CIAT, Rio de Janeiro, 2014.

[10] Dogan U., "Data Warehouse and Data-Mining Tools for Risk Management: The Case of Turkey," The International Bank for Reconstruction and Development / the World Bank, Washington DC, 2011.

[11] K. M. A. M. D. Anderson G., "Understanding the potential of Data Mining in Botswana," Africa Journal of Computing and ICT, vol. 6, no. 1, 2013.

[12] M. S. K. D. M. D. S. A. Diwani S., "Overview Applications of Data Mining in Health Care: The case study of Arusha Region.," International Journal of Computational Engineering Research, vol. 3, no. 8, 2013.

[13] Cleary D., "irish-tax-and-customers," SAS Institute, [Online]. Available: http://www.sas.com/da_dk/customers/irish-tax-and-customers.html. [Accessed 9 March 2016].

[14] Nhekairo W., "The Taxation System in Zambia," Jesuits Centre for Theological Reflection:, Lusaka, Zambia.

[15] Msiska B., "Planning a Change Strategy for Tax Administration: A case of ZRA," [Online]. Available: http://siteresources.worldbank.org/PSGLP/Resources/5ZambiaSA.pdf. [Accessed 3 March 2016].

[16] G. A. A. N. T. K. M. Y. Kontaki M., "Continuous Monitoring of Distance-Based Outliers over Data Streams," in Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE) , Hannover, Germany, 2011.

[17] P. M. P. R. Padhy N, "The Survey of Data Mining Applications And Feature Scope," International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol. 2, no. 3, June 2012.

[18] Sahu R.K., "Application of Business Intelligence in the Banking Industry," Management Information Systems, vol. 6, no. 4, 10 July 2011.

[19] Cheng M., "Application of business Intelligence in higher Education sector," 2012.

[20] O. D. ,. Bala M., "Study of applications of Data Mining Techniques in Education," International Journal of Research in Science and Technology, vol. 1, no. 6, pp. 135 - 146, 2012.

[21] J. H. C. Maeda, "Harnessing the power of enhanced data for healthcare quality improvement:Lessons from a Minnesota hospital association pilot Project Practitioner Application.," Journal of Healthcare Management, vol. 57, no. 6, p. 406–418, 2012.

[22] K. C. Foshaya N., "Towards an implementation framework for business intelligence in health care," International Journal of Information Management, p. 20– 27, 2014.

[23] Burns E., "How predictive analytics in healthcare can lower readmissions," TechTarget, 2015.

[24] O. M. C. Batko K., "The Use of Business Intelligence Systems in Healthcare Organizations in Poland," in Proceedings of the Federated Conference on Computer Science and Information Systems, 2012.