بسم الله الرحمن الرحيم

Sudan University of Science &Technology

College of Graduated Studies

Developing a Methodology for Detecting Tax-gaps Using Data Mining Techniques

تطوير منهجية لأكتشاف الفجوة الضريبي بإستخدام تقنيات التنقيب عن البيانات

A dissertation Submitted in partial Fulfilment of the requirements for

MSc degree in computer science

By

Hana Mohammed Eltgani Mohammed

Supervisor

Dr. Mohamed Elhafiz Mustafa Musa

March 2014

Acknowledgement

I have been lucky to combine a demanding work at the VAT (Value Added Tax) Audit Unit of the Sudanese Tax Administration with studies in the Computer Science Master's Programme of the Sudan University of Science and Technology This thesis is a result of that symbiotic setting where motivation has flown in both directions.

I want to express my deepest appreciation to my family for their support and understanding and all my colleagues and especially my sister Nejood whom have contributed, either directly or indirectly, to this thesis. I believe I could not have got better ingredients for it. I hope the results will serve our common goal of enhancing tax gap detection.

Special thanks to Dr. Mohamed Elhafiz, my supervisor, for his help and guidance along the way. I am glad to have the opportunity to work with him.

Hana Mohammed

March 2014

Abstract

The tax-gap is defined as the difference between what tax-payers legally owe and what they voluntarily pay. This thesis attempts to identify new approach that can help taxation chamber to detect tax-payers whose tax returns may require auditing. The dataset used in this thesis (tax-payers dataset) contains 52568 observations, and it represents the historical data of tax-payers' monthly reports. By using the K-mean clustering algorithm a model is developed to find patterns. From the experiments it was found that the proper number of clusters is 10 clusters categorized monthly report observations into high, medium, or low tax-gap observations. It is found that the identified pattern could be used to make auditing staff focus on just 3% of tax-payers data that represents observations belonging to high and medium tax-gaps. In addition, by reducing the work load there will be an enhancement on the use of the available human, financial and technical resources.

الخلاصة

يتم تعريف الفجوة الضريبية بالفرق بين ما يجب ان يقوم دافعي الضرائب بدفعه من الناحية القانونية وما يدفعون طوعا. يقدم هذا البحث منهجية جديدة يمكن أن تساعد ديوان الضرائب علي الكشف عن دافعي الضرائب الذين قد تتطلب اقراراتهم الضريبية المراجعة. البيانات المستخدمة في هذا البحث تحوي علي 52568 ملاحظة، وهي عبارة عن البيانات المحفوظة في الاقرارات الشهرية التي يقدمها دافعي الضرائب. تم استخدمة خوارزمية K-mean لعنقدة بيانات دافعي الضرائب، و من التجارب عثر علي ان عدد العناقيد المناسب هو 10عناقيد، تم تصنيف ملاحظات الاقرارات الشهرية الي ملاحظات ذات فجوة ضريبية عالية، متوسطة، أو منخفضة. النموذج الذي تم تطويره يمكن أن يستخدم لتركيز عبء التدقيق علي 3% من ملاحظات دافعي الضرائب والتي تنتمي الي فئات الفجوة الضريبية العالية و المتوسطة وكذلك المساعدة بالإستخدام الأمثل للموارد البشرية والمالية والتقنية المتاحة.

Table of Contents

List of Figures

List of Tables

CHAPTER ONE

INTRODUCTION

1.1 Introduction

In perfect, law-abiding society, people would pay the tax they owe, and tax administrations would only facilitate necessities for citizens to carry out the responsibility. However, no such society exists. Therefore, compliance with tax laws must be cultivated, monitored and enforced in any society.

When tax-payers don’t compliance with federal tax obligations a “tax-gap” is formed. The taxations chambers try to measure the “tax-gap” or the extent to which tax-payers do not file their tax returns or pay the correct tax on time. The concept “tax-gap” is defined as the difference between what tax-payers legally owe and what they voluntarily pay. The causes of “tax-gap” can either be due to unreported return by tax-payers, underreported taxable income, or simply incomplete payment for the amount due. A key method to reduce the tax-gap is tax audit. Simply stated, tax audit is the collection of methods used to identify tax-payers not filing tax returns and paying their fair and appropriate share of taxes. [1]

Taxation is an information intensive domain that involves processing of vast amounts of data concerning a large number of tax-payers. Thus, tax audit is needed to ensure compliance with tax laws and maintain associated revenue streams. Audits indirectly drive voluntary compliance and directly generate additional tax collections, both of which help tax agencies reduce the “tax gap” between the tax owed and the amount collected. Audits, therefore, are critical to enforce tax laws and help tax agencies achieve revenue objectives, ensuring the financial health of the country.

Managing an effective auditing organization involves many decisions such as:

- What is the best audit selection strategy or combination of strategies?
- Should it be based on reported tax amounts or on the industry type?
- How should agencies allocate audit resources among different tax types?

Some tax types may yield greater per-audit adjustments. Others may be associated with a higher incidence of noncompliance. An audit is a process with many progressive stages, from audit selection and assignment to hearings, negotiation, collection, and in some cases, enforcement. Each stage involves decisions that can increase or reduce the efficiency of the overall auditing program. [2]

This chapter is divided as the following: section 1.2, discusses the research problem. Section 1.3, explains the aim of study. Section 1.4, provides an overview research methodology. Section 1.5, views the thesis organization.

1.2 Research problem

The overall research problem of this study can be stated as follows:

1. How to identify new approach that can help taxation chamber to detect tax-payers whose tax returns may require auditing?

In view of the research problem it is important to understand that the functions of a tax administration go beyond tax audits. Today's tax administrations are essentially service organisations where tax-payers are truly considered as customers. It is the tax administration's task to create and maintain an enabling environment for its customers to be able to comply with applicable tax laws and regulations with minimum effort. The research problem should thus be viewed in relation to the tax administrations' overall playground which involves both enabling and ensuring compliance.

Tax-payers with diverse abilities, behaviours, attitudes and motives pose a permanent challenge probably for all tax administrations. Different customers need different types of attention. At the same time many tax administrations face increasing efficiency and effectiveness requirements. Limited resources must be allocated where they bring best return. The above considerations must be duly addressed in connection with the research problem. [3]

1.3 Aims of the study

In spite of a seemingly promising potential, data mining has been applied in tax administrations to a relatively limited extent. The overall aim of this study is to gain understanding of how data mining applications could help tax administrations accomplish their general mission, to get the right tax at the right time, more efficiently and more effectively. This can be decomposed into the following specific aims:

- Enhance current auditing techniques for detecting tax-gaps at taxation chamber system using data mining techniques;
- Achieve equal treatment of the tax-payers;
- Best use of the available human, financial and technical resources;

1.4 Research methodology

The research methods include the following:

- Reviewing books, articles, research papers about data mining and its applications in taxation.
- Gathering data set appropriate for research problem.
- Illustrating the observations and findings of applying data mining technique in the Sudanese Taxation chamber database.

1.5 Thesis organization

The remainder of this thesis is organized as follows.

Firstly: Chapter 2 gives an overview of data mining algorithms, a review about Tax compliance management, and the use of Data mining in audit target selection.

Secondly: Chapter 3 discusses the research methodology. In addition: Chapter 4 presents the results of the experiments and analyze these results. Finally, chapter 5 provides conclusions and recommendations for future work.

CHAPTER TWO

OVERVIEW OF DATA MINING

2.1 Introduction

Data mining is a process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules [4].

According to [5] data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

The majority of data mining tasks can be broken into six general areas as follows Classification, Clustering, Association Rule Discovery, Sequential Pattern Discovery, Regression, and Deviation Detection.

This chapter presents cluster analysis and the algorithms that support it in section 2.2, a review about Tax compliance management in section 2.3, manual audit file selection strategy in section 2.4 and the use of Data mining in audit target selection in section 2.5.

2.2 Cluster analysis

The purpose of this research is to describe tax-payers whose tax returns may require auditing; therefore cluster analysis has been selected as data mining descriptive task.

Cluster analysis groups data objects based only on information found in the data which describes the objects and their similarities. [6]

Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both. It is similar to classification the difference is that when using clustering we don’t know the “answer” or clusters before the analysis. For this reason, clustering is often called unsupervised learning while classification is often called supervised learning. [6]

According to [7] clustering can be described as the following:

- **Hierarchical:** Clustering is characterised as hierarchical when the data points are not overlapping while they are hierarchical if a cluster is consisted from many clusters.
- **Exclusive, Overlapping or Fuzzy:** If the data points of a clustering belong only to one cluster, then this is an exclusive clustering, while if a data point belongs to more than one clustering equally, then it is overlapping. In fuzzy clustering each data point is assigned a probability from 0 to 1 that it belongs to a cluster. The sum of all cluster probabilities for a data points are equal to 1.
- **Complete or Partial:** Clustering is complete when it groups all the data or partial when it leaves data un-clustered.

2.2.1 Clustering algorithms

According to [7] there are three common techniques for clustering: Namely; Prototype-Based clustering, Hierarchical clustering and Density-Based clustering.

2.2.1.1 Prototype-Based Clustering:

The most common algorithm for prototype-based clustering is the K-means algorithm. This algorithm defines *K* cluster centres and iterates by assigning each data point to the closest centre value and it recalculate the centre value until the centres stop changing. The number of centre s *K* is a user defined parameter and is the actual number of clusters the specific clustering will have. The centre s in most of the cases is not in the position of a point. Although simple and vastly used, however this algorithm fails to cluster data where the points are not rounded-shaped. While the K-metroid algorithm in which the centres are represented by real points (metroids) can cluster any data and the drawback of this algorithm is its higher cost.

2.2.1.2 Hierarchical Clustering

One of the most widely used hierarchical clustering algorithms is the Agglomerative Hierarchical Clustering algorithm. In this algorithm, the clustering starts by considering each point as a cluster and gradually grouping points that are close to each other in one cluster. This is done until there is only one cluster. The grouping of points is done with the calculation of the proximity matrix. The proximity matrix is a matrix where the distance of two clusters is calculated and stored. The most common ways to calculate the proximity of two clusters is the MIN that calculates the minimum distance, MAX that calculates the maximum distance and Group Average.

2.2.1.3 Density-Based Clustering:

Density-Based clustering is mostly applied with the DB SCAN algorithm. This algorithm is based in the centre-based approach where it checks inside the radium *R* of a point. If the number of points found within the radium is greater than a point threshold *Tp*, then the point is a core point. If the number of points is smaller that *Tp* but is inside the area of a core point then the point is border point. In any other case the point is considered to be a noise point. In the evaluation of the points, the core points are the centre of the clusters, the border points are merged with the cluster and the noise points are deleted.

2.3 Tax compliance management

According to [3] the Tax compliance management is about optimising the use of resources allocated to a tax administration in order to maximise the overall level of compliance with the tax. Compliance here relates to the extent to which tax-payers meet the obligations placed on them by law is identifies four broad categories of tax-payer obligations:

- ***Registration*** in the system
- Timely ***filing*** of required taxation information
- ***Reporting*** of complete and accurate information (incorporating good record keeping)
- ***Payment*** of taxation obligations on time

A tax-payer's failure to meet any of the above obligations may be considered non-compliance. It may be due to unintentional error or intentional fraud. The compliance approach recognises that tax-payers have diverse capabilities, behaviours, attitudes and motives, and that there is a need to adjust and target the tax administration's services and intervene accordingly.

Tax-payers can be categorized into one of the following categories based on their behaviour when dealing with taxes. The behaviours are based on certain sets of values, beliefs and attitudes adopted by the person. These behaviours are usually represented by a compliance pyramid, shown in Figure 1.

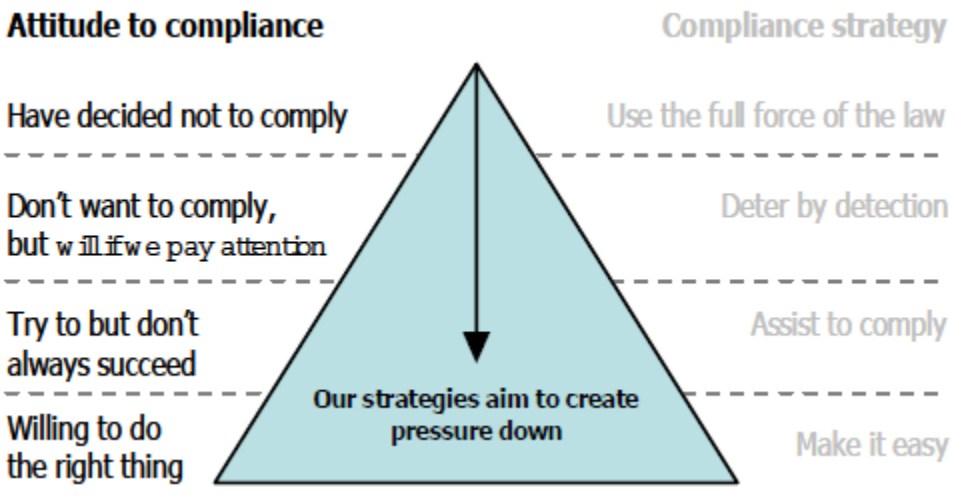

Figure : Compliance pyramid, spectrum of tax-payer attitudes to compliance [3]

With targeted activities, suited according to motivational postures, tax administrations can stimulate compliance and constrain the motivation to resist or evade compliance. It is important to note, however, that an individual tax-payer may adopt any of the attitudes in Figure 1 at different times, or adopt all of them simultaneously in relation to different issues. A tax administration's strategy should be designed so as to create an overall pressure for tax-payers to move down in the pyramid.

2.4 Manual audit file selection strategy

The procedure used currently for selecting files for auditing in the Tax Auditing Unit of the Sudanese Tax Administration tags the following as files having priorities in auditing:

i. Time priority: the file must undergo auditing before it complete five year from the last audit. Tax Administration does not have the right to request for tax gap money after more than five years (taxation dropped).

ii. Big credit balance.

iii. Empty tax-payer monthly reports.

iv. The new registered tax-payers after completing three month to make sure those tax-payers are following tax law.

2.5 Data mining in audit target selection

According to [3] The Tax Auditing Unit of the Finnish Tax Administration participated in a research project called *Titan*, studies the potential of data mining to support the selection of companies for tax audit in today's ever growing

complexity of business relationships. The way that tax auditors have identified the audit targets to date, relying on their past experience to pose queries to the database of financial reports, may not pick the best candidates for tax audit, due to the multitude of indicators for possible tax evasion. Another drawback by the current selection approach is that it focuses on only one company at a time while it would be worthwhile to view several companies simultaneously.

One objective in Titan project is to develop a general model for identifying companies that merit a tax audit. Such companies form the target group. Data mining is first used to define the profiles of companies that have been chosen for audit with a reason, that is, where the audit has yielded additional taxes. The features that differentiate these profiles from those of the companies with no need for audit are interesting.

Clustering using the self-organising map (SOM) is applied as a data mining tool to find the similarities in the audited companies of the target group. The SOM is a form of neural network frequently used in data mining tasks. The SOM algorithm projects multidimensional data onto a two-dimensional map and divides the observations into clusters. The SOM thus combines two data mining tasks; clustering and visualisation. The goal is to create a self-organising map where one cluster, called the *key cluster*, should include the majority of the target group companies.

Titan used taxation data of more than 5,000 partnership companies from year 2004 to build the model. In the data cleaning phase the data set was reduced to some 4,000 companies of which approximately 100 belonged to the target group. Three variants of the model were built and compared. At best 93 % of the auditing result could have been collected from the companies placed in the key cluster, but on the other hand all model variants were quite generous in placing also companies not in need of audit in that cluster. The model variants generally appeared to perform better in catching big evaders than distinguishing between the companies where audit is not needed and those where auditing result was low. All in all, the study concluded that the self-organising map could function as a tool to support the audit

target selection, but the application area is very complex and further research is needed.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The problem in hand contains large number of data with no prior known features that can be used for classification. Clustering the data into different groups and trying to understand the behaviour of each group is suggested as a methodology for modelling the data. The algorithm chosen for clustering the taxation data is K-mean algorithm and the tools for the implementation are R and Rattle. The following sections will present the algorithm that will be used for clustering and the tools used for implementing the solution.

3.2 K-means algorithm

K-MEANS [6] is the simplest algorithm used for clustering which is unsupervised clustering algorithm. This algorithm partitions the data set into k clusters using the cluster mean value so that the resulting clusters intra cluster similarity is high and inter cluster similarity is low. K-Means is iterative in nature it follows the following steps:

1. Arbitrarily generate *k* points (cluster centres), *k* being the number of clusters desired.
2. Calculate the distance between each of the data points to each of the centres, and assign each point to the closest centre.
3. Calculate the new cluster centre by calculating the mean value of all data points in the respective cluster.
4. With the new centres, repeat step 2. If the assignment of cluster for the data points changes, repeat step 3 else stop the process.

The distance between the data points is calculated using Euclidean distance as follows. The Euclidean distance between two points or features, $X1 = (x_{11}, x_{12...} x_{1m})$, $X2 = (x_{21}, x_{22}, ..., x_{2m})$

$Dist(X1; X2) =$

$$\sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$$

Advantages

1) Fast, robust and easier to understand.

2) Relatively efficient: O (*t k n d*), where n is objects, *k* is clusters, d is dimension of each object, and t is iterations. Normally *k*, *t* , *d* < *n*.

3) Gives best result when data set are distinct or well separated from each other.

Disadvantages

1) The learning algorithm requires apriori specification of the number of cluster centres.

2) The learning algorithm provides the local optima of the squared error function.

3) Applicable only when mean is defined i.e. fails for categorical data.

4) Unable to handle noisy data and outliers.

3.3 R and Rattle

According to [8] R is a sophisticated statistical software package, easily installed, instructional, state-of-the-art, and it is free and open source. It provides all of the common, most of the less common, and all of the new approaches to data mining. The basic modus operandi in using R is to write scripts using the R language.

Rattle is built on the statistical language R, but an understanding of R is not required in order to use it. Rattle is simple to use, quickly to deploy, and allows us to rapidly work through the data processing, modelling, and evaluation phases of a data mining project.

On the other hand, R provides a very powerful language for performing data mining well beyond the limitations that are embodied in any graphical user interface and the consequently canned approaches to data mining. When we need

to fine-tune and further develop our data mining projects, we can migrate from Rattle to R.

Rattle can save the current state of a data mining task as a Rattle project. A Rattle project can then be loaded at a later time or shared with other users. Projects can be loaded, modified, and saved, allowing check pointing and parallel explorations. Projects also retain all of the R code for transparency and repeatability. This is an important aspect of any scientific and deployed endeavour to be able to repeat our "experiments." Whilst a user of Rattle need not necessarily learn R, Rattle exposes all of the underlying R code to allow it to be directly deployed within the R Console as well as saved in R scripts for future reference. The R code can be loaded into R (outside of Rattle) to repeat any data mining task.

Rattle by itself may be sufficient for all of a user's needs, particularly in the context of introducing data mining. However, it also provides a stepping stone to more sophisticated processing and modelling in R itself. It is worth emphasizing that the user is not limited to how Rattle does things. For sophisticated and unconstrained data mining, the experienced user will progress to interacting directly with R. The typical work flow for a data mining project was introduced above.

In the context of Rattle, it can be summarized as:

1. Load a Dataset.
2. Select variables and entities for exploring and mining.
3. Explore the data to understand how it is distributed or spread.
4. Transform the data to suit our data mining purposes.
5. Build our Models.
6. Evaluate the models on other datasets.
7. Export the models for deployment.

It is important to note that at any stage the next step could well be a step to a previous stage. In Figure 2 an illustration of a typical work flow that is embodied in the Rattle inter-face

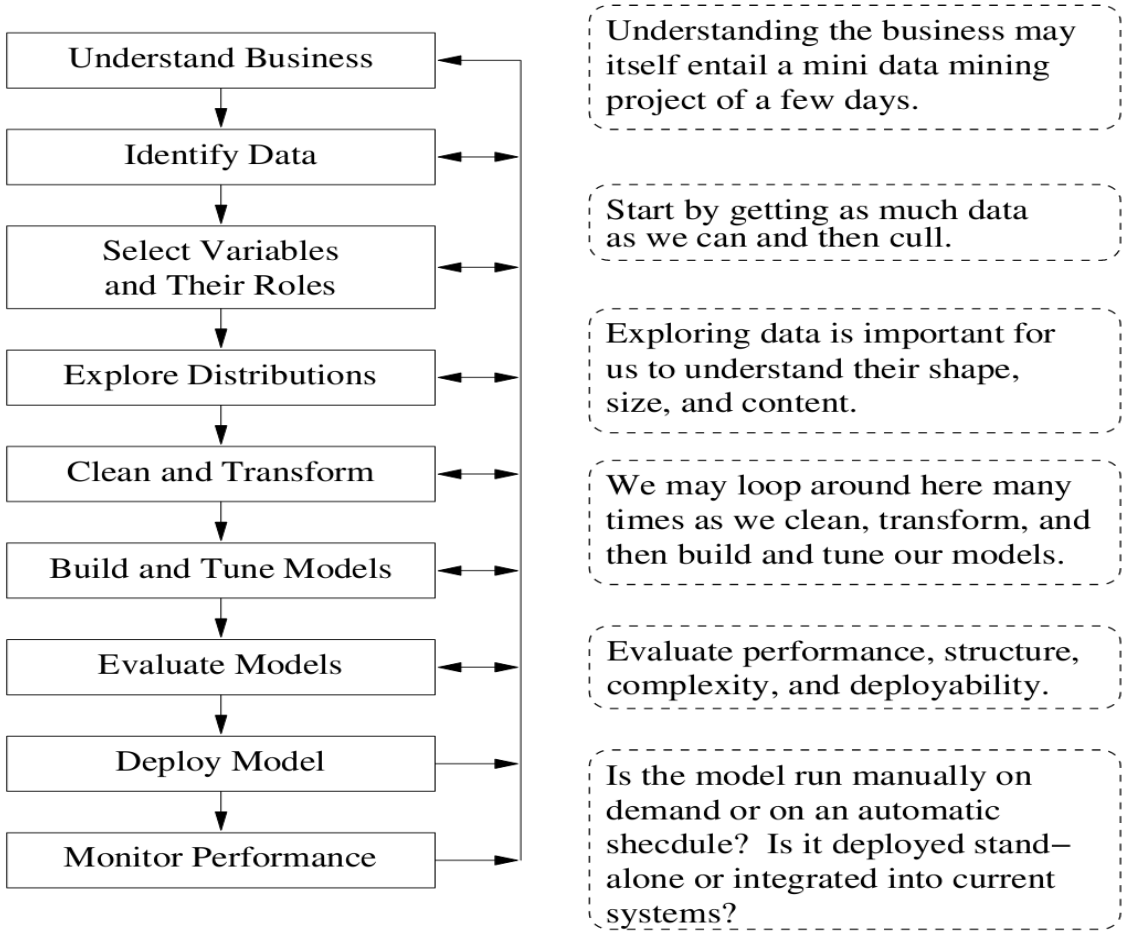

Figure : The typical work flow of a data mining project as supported by Rattle. [8]

In this thesis the typical work flow of a data mining project in Rattle is used.

3.4 Strategies for data reduction

According to [9] Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Strategies for data reduction include the following:

1. Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.

2. Attribute subset selection, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
3. Dimensionality reduction, where encoding mechanisms are used to reduce the dataset size.
4. Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.
5. Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.

From the above data reduction strategies the attribute subset selection strategy has been selected, for the step of data cleaning and transformation in Rattle typical work flow.

CHAPTER FOUR

RESULTS AND ANALYSIS

In this chapter, we present the results obtained by applying k-mean cluster analysis algorithm on dataset of 52,568 tax-payers monthly report observations. We first present in Section 4.1overview of dataset and data cleaning procedure applied on it. Sections 4.2 presents the best data reduction strategies obtained using the k-means. Sections 4.3 shows the eight clustering experiments obtained using the k-means clustering methods, provides interpretation for all cluster means that appear in the experiments and associates each cluster means with a type of tax-gap according to its characteristics. Section 4.4 discusses the addition that this thesis makes to the existing tax-gap detection technique.

4.1 Dataset overview

The source of dataset which is used in this thesis is data from taxation champers VAT system. The data consisting of tax-payers monthly reports observations and it was imported into an Excel table as an initial data preparation effort.

The database consists of 52,568 tax-payers monthly report observations from year 2001 to 2012, stored in Excel 2007 spreadsheet format. The initial database provided 23 features.

These 23 features are listed in table 1:

Table : The 23 feature in tax-payers monthly report.

Features	Data Type	Explanations
CARD_NO	Numeric	Tax-payers card number.
B_NO	Numeric	Branch number.
ORD_MON	Numeric	Month of monthly report delivered.
ORD_YY	Numeric	Year of monthly report delivered.
SAL_E	Numeric	Total sales balance.
SAL_M	Numeric	Total sales balance free from value added tax.
EXPORT	Numeric	Total balance paid for export payment.
PAY_LOC	Numeric	Total balance paid for local payment.
PAY_F	Numeric	Total balance paid for special payment.
IMPORT	Numeric	Total balance paid for import.
CREDIT	Numeric	Credit balance.
DEPT	Numeric	Debit balance.
ORD_DATE	Categoric	Date of monthly report rewarded.

IN_DATE	Categoric	Date of monthly report rewarded, entered in system.
DIS_CREDIT	Numeric	Credit balance from last monthly report.
PAY_LOC_A	Numeric	Local payment balance free from value added tax.
NILE	Numeric	Total money paid for petrol company certificates.
VAT	Numeric	Value added tax wage.
SAL_SRV	Numeric	Total money paid for other payment.
PAY_SRV	Numeric	Total money paid for other sale.
IMP_LOC	Numeric	The price of imported staff that had no VAT in it.
SEP_SERV	Numeric	Total money paid for service.
ORD_FLAG	Categoric	Order flag.

It has been realized that there are two currencies that was used in the data; “Pound” and “Dinar”. To insure consistency in the data, financial data was converted to the type of Sudanese currency that is used currently i.e. the pound instead of the dinar. One pound is equivalent to 100 dinars; the conversion has been in month of July of the year 2007. This pre-processing step was performed converting the financial data prior to the month of July of the year 2007 from pound to dinar.

4.2 Data reduction

In this thesis, the data reduction is applied to obtain a reduced representation of the dataset that is much smaller in volume, yet closely maintains the integrity of the original data.

The following section presents the results obtained using k-means clustering algorithm in dataset without data reduction.

4.2.1 K-means clustering algorithm in dataset without data reduction.

Figure 3 shows a plot of the clusters centre sum of squares as a function of the number of clusters, *k*. Clearly, the value of clusters centre sum of squares decreases as *k* increases. This is expected, because as the number of clusters increases, the algorithm can find more compact clusters for the data. In the extreme case, every point forms its own cluster and the sum of squared distances becomes zero.

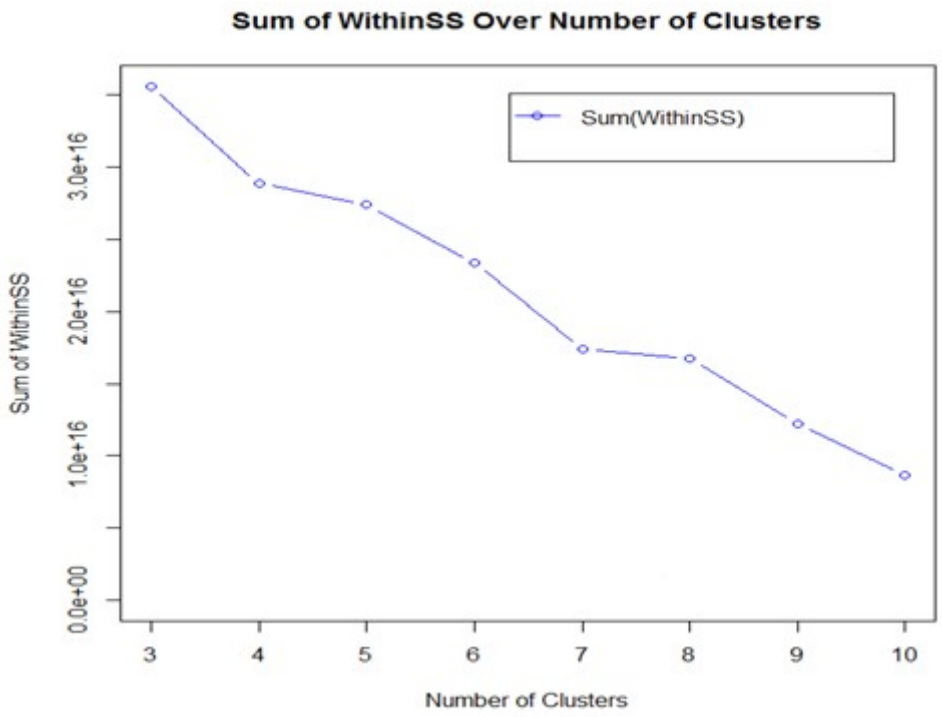

Figure : The clusters centre sum of squares a function of the number of clusters, k, for 20 features.

In this thesis, 5 clusters were found to be the suitable number of clusters as can be noticed in the above figure after applying K-Mean algorithm on the data. This is because it is found that the largest drop in the sum of the within cluster sum of squares appear when the number of clusters is 5 clusters.

For cluster analysis model in Rattle the K-mean algorithm is used to implement the model. It begins with the cluster size, which is simply a count of the number of observations within each cluster, table 2 describe the cluster size.

Table : clusters size.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
46957	5	122	10	34

From the previous table the numbers of clustered tax-payers monthly reports observations are 47128 observations instead of 52568 observations. There are 5440 observations which are not clustered. To solve this problem Data reduction techniques can be applied.

From the data reduction strategies which have been discussed in section (3.4) the attribute subset selection strategy has been selected, because the data set contains irrelevant features to our research problem.

So to remove irrelevant features the model has been reviewed to see the result of k-means clustering algorithm. The following table presents the result that consists of five vectors of the mean values for each of the variables. The cluster centres have been shown in table 3.

Table : The cluster centres of a 5 clusters.

Features	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
SAL_E	11.211	37.184.107	1.016.343	38.297.391	19.59.335
SAL_M	1.268	0.0	23.580	0.0	1.314
EXPORT	406	0.0	99	0.0	0.0
PAY_LOC	4727	25.510.460	768.316	29.377.183	415.445
PAY_F	772	0.0	0.0	0.0	7.784
IMPORT	8.027	0.0	44.106	6.853.266	63.868
CREDIT	3.469	0.0	26.245	73.023	0.0
DEPT	14.116	106.842.497	3.363.359	23.076.551	13781093
DIS_CREDIT	3.169	98.940	13.272	39.797	8.508
PAY_LOC_A	0.1	0.0	0.0	0.0	0.0
NILE	79	0.0	5.310	0.0	0.0
VAT	10	10	10	10	10

SAL_SRV	0.2	0.0	308	0.0	0.0
PAY_SRV	6	0.0	40	0.0	768
IMP_LOC	176	0.0	0.0	0.0	0.0
SEP_SERV	1	0.0	0.0	0.0	0.0

Any feature that has cluster centre having the value 0 in more than two clusters, or has fixed values in most of the clusters is removed. Thus only the most relevant features to the research will be considered. The relevant selected features for tax-gap detection using Attribute subset selection is shown in table 4.

Table : Features for tax-gap detection in tax-payer monthly report.

Variable	Explanations
SAL_E	Total sales balance.
SAL_M	Total sales balance free from value added tax.
EXPORT	Total balance paid for export payment.
PAY_LOC	Total balance paid for local payment.
PAY_F	Total balance paid for special payment.
IMPORT	Total balance paid for import.
CREDIT	Credit balance.
DEPT	Debit balance.

4.3 Experiments and Results

This section shows the results of the clustering experiments obtained using the K-means clustering methods, provides interpretation for all cluster means that appear in experiments and associates each cluster means with a type of tax-gap according to its characteristics.

Figure 4 shows the results of the eight experiments after applying K-means algorithm for *K* 3...10. Each group of nodes show cluster number (C) and the number of observations in that cluster. The arrows between nodes illustrate the splitting that happened to each clusters.

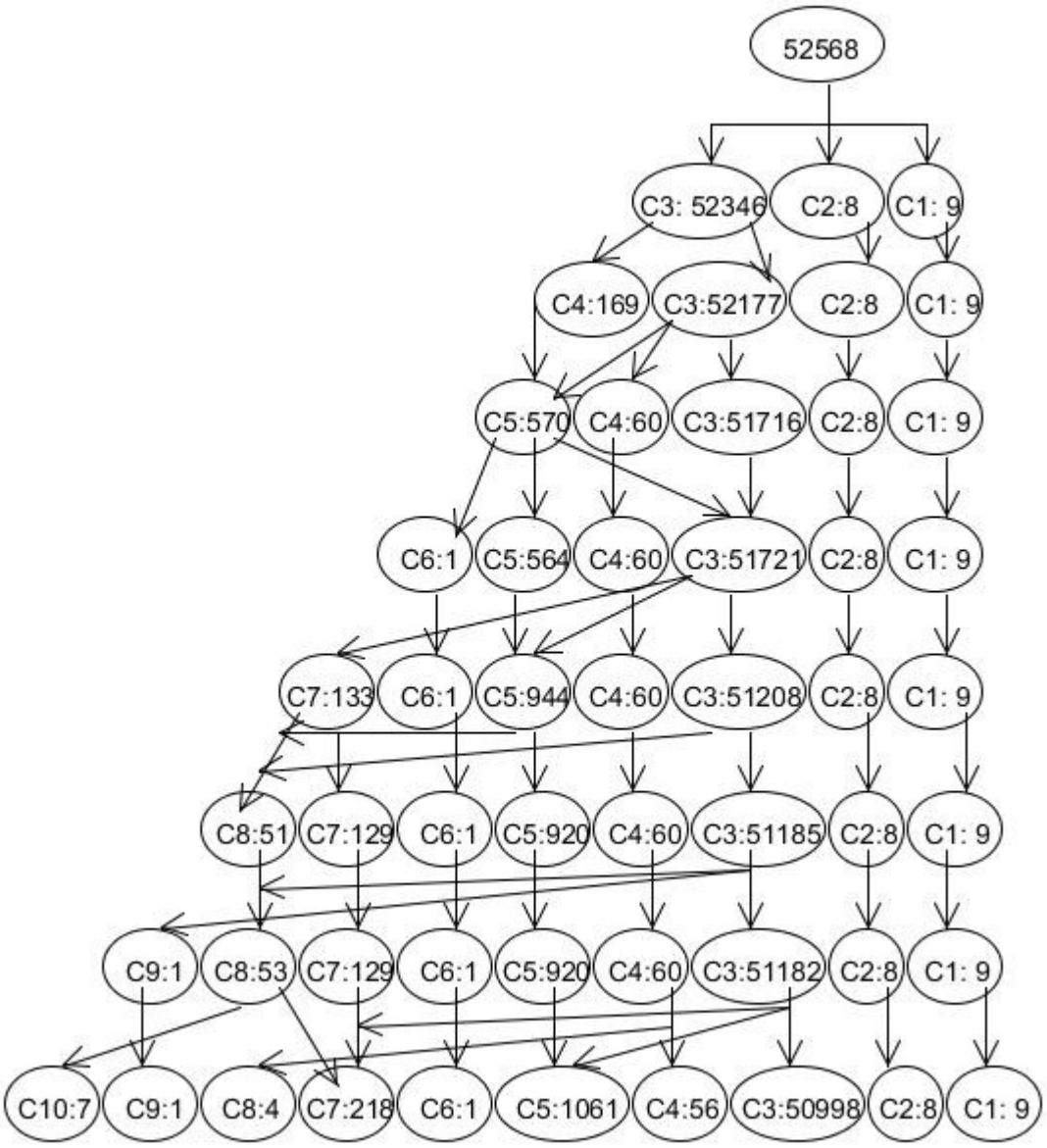

Figure : The number of observations in each cluster throughout the 8 experiments

From figure 4 it is realized that the number of clustered tax-payers monthly report observations are 52363 observations out of 52568 observations; there are 205 observations not clustered. This result is better than having 5439 observations not clustered as has been seen in previous experiment.

Also from figure 4 it could be noticed that when the number of *K* increased there are some clusters that remain fixed without splitting, while others split. In the

following sections we will continue discussing the results with the number of clusters K=10. The following paragraphs give the interpretation for that cluster. Table 5 shows the number of observations in each cluster, when K = 10.

Table : Cluster sizes for 10 clusters.

Cluster no.	Cluster size
1	9
2	8
3	50998
4	56
5	1061
6	1
7	218
8	4
9	1
10	7

Table 6 shows the cluster centre of observations in each cluster, when K = 10.

Table : Clusters' centers

K	Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
1	45.595.209	0.0	0.0	40.481.741	0.0	0.0	10.550	511.347
2	25.711.591	0.0	0.0	14.249.443	0.0	8.566.583	137.248	289.557
3	4.603	631	148	2.050	428.98	2.390	2.341	133
4	2.171.617	798	0.0	926.587	0.0	142.473	93.539	104.922
5	238.734	29.315	705.47	86.764	2.813	124.896	36.294	4.630
6	0.0	0.0	10.829.789	0.0	10.973.639	0.0	0.0	0.0
7	505.859	51.152	102	36.129	3.201	650.346	71.101	2.676
8	391.237	0.0	0.0	7.384.163	0.0	0.0	964.86	0.0

							3	
9	0.0	16.781.292	0.0	156.743	0.0	7.363	36.180	0.0
10	397.316	102.829	0.0	18.863	0.0	4.016.934	392.000	0.0

Cluster Means interpretations:

Interpretation for each of the clusters will be given in the following points:

[1] Cluster 1 (medium tax-gap**)**

Table : The cluster centres for cluster No.1.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
45.595.209	0.0	0.0	40.481.741	0.0	0.0	10.550.051	511.347

The number of observations in this cluster is 9. This cluster has been described as medium tax-gap; because dept is high that is due to the big activity size. When auditing a client who does not have documents to the local payment that will give chance for tax-gap to arise. It is found that all the 9 observations in this cluster [1] belong to the same tax-payer.

(2) Cluster 2 (high tax-gap**)**

Table : The cluster centers for cluster No.2.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
25.711.591	0.0	0.0	14.249.443	0.0	8.566.583	137.248.270	289.557

The number of observations in this cluster is 8. This cluster has been described as high tax-gap; because the imported payment is high and credit balance doesn’t match sale balance. That may denote a low profit percentage and if these cluster observations undergo auditing the profit percentage could be change to higher percentage meaning a big tax-gap would arise. All observations in this cluster (2) belong to the same tax-payer.

(3) Cluster 3 (low tax-gap**)**

Table : The cluster centers for cluster No.3.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
4.603	631	148	2.050	429	2.390	2.341	133

The number of observations in this cluster is 50998. This cluster has been described as low tax-gap; because there exists more than free from vat features in it. The purpose of auditing this group is just to ensure that tax-payers work on the right way.

The observations in this cluster are versatile, this cluster cover 97% from the dataset.

(4) Cluster 4 (high tax-gap**)**

Table : The cluster centers for cluster No.4.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
2.171.617	798	0.0	926.587	0.0	142.473	93.539	104.922

The number of observations in this cluster is 60. This cluster has been described as high tax-gap; because it has high credit balance compared to local payment, import and sales. Many advantages could be obtained from auditing this group such as discovering a tax return gap or at least removing credit balance. The observations in this cluster are versatile, and it covers .12%.

(5) Cluster 5 (high tax-gap**)**

Table : The cluster centers for cluster No.5.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
238.734	29.315	706	86.764	2.813	124.896	36.294	4.630

The number of observations in this cluster is 1061. This cluster has been described as high tax-gap because dept is low compared to the big activity size. If the cluster observations undergo auditing and no document were found to the local payment

that will give a chance for tax-gap to arise. The observations in this cluster are versatile, and it covers 2% from the dataset.

(6) Cluster 6 (low tax-gap**)**

Table : The cluster centers for cluster No.6.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
0.0	0.0	10.829.789	0.0	10.973.639	0.0	0.0	0.0

The number of observations in this cluster is 1. This cluster has been described as low tax-gap; because the given balances are suitable to each other. By auditing this cluster observation, it would be to ensure that the proper documents exist.

(7) Cluster 7 (medium tax-gap**)**

Table : The cluster centers for cluster No.7.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
505.859	51.152	102	36.129	3.201	650.346	71.101	2.676

The number of observations in this cluster is 218. This cluster has been described as medium tax-gap because credit is high that denote to the big activity size. By auditing this cluster a tax-gap might arise in case of not finding the supportive documents to the local payment or documents for import payment. The observations in this cluster are versatile, and it covers .4% from the dataset.

(8) Cluster 8 (medium tax-gap**)**

Table : The cluster centers for cluster No.8.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
391.237	0.0	0.0	7.384.163	0.0	0.0	964.863	0.0

The number of observations in this cluster is 4. This cluster has been described as medium tax-gap because credit is high that denote to the big activity size By auditing this cluster a tax-gap might arise in case of not finding the supportive documents to the local payment.

(9) Cluster 9 (low tax-gap**)**

Table : The cluster centers for cluster No.9.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
0.0	16.781.292	0.0	156.743	0.0	7.363	36.180	0.0

The number of observations in this cluster is 1. This cluster has been described as medium tax-gap because credit is high that denote to the big activity size. By auditing this cluster a tax-gap might arise in case of not finding the supportive documents to the local payment.

(10) Cluster 10 (medium tax-gap**)**

Table : The cluster centers for cluster No.10.

Sale	Sale free from vat	Export	Local payment	Special payment	Import	Credit	Dept
397.316	102.829	0.0	18.863	0.0	4.016.934	392.000	0.0

The number of observations in this cluster is 7. This cluster has been described as medium tax-gap because credit is high that denote to the big activity size. By auditing this cluster a tax-gap might arise in case of not finding the supportive documents to the local payment.

4.4 EVALUATION

For this project, cluster analysis modelling using k-mean algorithm is used to help in detecting tax-payers whose tax returns may require auditing, by describing

monthly report transaction as belong to the category of high, medium, and low risk transaction that may produce tax gap.

From the previous experiment it could be noticed that the proper number of cluster is 10 in dataset that contained 52568 observations and 8 input variables. Clustered rule were 52363 from 52568. There were 205 observations that were not clustered, and by reviewing them, it could be noticed that there is more than a missing variable necessary to cluster the remaining observations. Meaning that those variables are not set or they are mostly set to zero, this can be interpreted that the tax-payer did not exercise any activity in this period.

CHAPTER FIVE

CONCLUSIONS & FUTURE WORK

5.1 Conclusions

The following research questions were proposed:

Research problem: Is there a new approach that can help taxation chamber to detect tax-payers whom their tax returns may require auditing?

The answer to the research question is a definite yes; Cluster analysis can yield meaningful tax-payer segments that need to be effectively addressed.

These clustering techniques should help the auditor to better understand large populations of tax-payers. Data-driven methods yield “natural” groupings of tax-payers, allowing the auditors to group a large population into a series of relevant clusters and make auditors better understand the tax issues and service needs of each cluster.

The results divide the tax-payers into three broad types of auditing categories: high tax-gap, medium tax-gap, low tax-gap. These categories help auditors to select tax-payer belong to high tax-gap category for auditing.

It is found that the identified pattern could be used to make auditing staff focus on just 3% of tax-payers data that represents observations belonging to high and medium tax-gaps. In addition, by reducing the work load there will be an enhancement on the use of the available human, financial and technical resources.

5.2 Future work

The results obtained in this thesis, by applying k-mean cluster analysis algorithm on dataset of 52,567 tax-payers monthly report observations and 23 features.

When applying the algorithm on dataset the numbers of clustered tax-payers monthly reports observations are 47128 observations instead of 52568 observations. There are 5440 observations which are not clustered. The attribute subset selection strategy from data reduction techniques was applied to solve that problem.

After applying the data reduction technique it had been found from experiments that the proper number of cluster is 10 in dataset that contained 52568 observations

and 8 features. And the clustered observations were 52363 from 52568. There were 205 observations that were not clustered, and by reviewing them, it could be noticed that there is more than a missing variable necessary to cluster the remaining observations.

The k-mean cluster analysis algorithm applicable only when mean is defined i.e. fails for categorical data. As a future work comparison between the results of K-mean algorithms and other cluster analysis algorithm, by taking different number of features to see if that will give different results.

References

1. *What Is the Tax Gap?* **Toder, Eric.** s.l. : Eric Toder., 2007 . American Bar Association Conference on the Tax. p. 1.
2. **Daniele Micci-Barreca, Satheesh Ramachandran,Elite Analytics, LLC.** *Improving tax administration with data mining.* s.l. : SPSS.
3. **Martikainen, Jani.** *DATA MINING IN TAX ADMINISTRATION Using analytics to enhance tax compliance.* 2012.
4. **Michael J.A. Berry, Gordon S. Linoff.** *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second Edition.* s.l. : Wiley, 2004.
5. **Jiawei Han, Micheline Kamber.** *Data Mining:Concepts and Techniques Second Edition.* s.l. : Diane Cerra, 2006.
6. **Tan, Steinbach, Kumar.** *Introduction to Data Mining.*
7. **Manikas, Konstantinos.** *Outlier Detection in Online Gambling.* Göteborg, Sweden : Chalmers University of Technology and University of Gothenburg , 2008.
8. **Williams, Graham.** *Data Mining with Rattle and R.* s.l. : Springer.
9. **Jiawei Han, Micheline Kamber.** *Data Mining:Concepts and Techniques Second Edition.* s.l. : Diane Cerra, 2006.