THREE ESSAYS ON OPEN GOVERNMENT DATA AND DATA ANALYTICS

by

ZAMIL S. ALZAMIL

A dissertation submitted to the

Graduate School - Newark

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Professor Miklos A. Vasarhelyi

and approved by

_____

Dr. Miklos A. Vasarhelyi (Chair)

_____

Dr. Yaw M. Mensah

_____

Dr. Kevin Moffitt

_____

Dr. Deniz Appelbaum

_____

Newark, New Jersey

May, 2019

ABSTRACT OF THE DISSERTATION

Three Essays on Open Government Data and Data Analytics

By Zamil S. Alzamil

Dissertation Chairman:
Professor Miklos A. Vasarhelyi

Over the past few years, we have seen a significant grown in interest for open data, specifically open government data (OGD). This led to the availability of a large number of public sectors' datasets made available to the citizens or any other interested stakeholder. Thanks to the pressure being placed on all types of government organizations in order to release their raw data. The main motivations for publicizing access to raw materials and make it more transparent are that it can help provide higher returns from the public utilization of such data, can provide policymakers with supportive data that can assess the process of making better decisions, can generate wealth through the development and creation of new and innovative products and services or enhance the current ones, and can involve the citizens to monitor and analyze publicly available datasets to help them evaluate and assess the performance of their governments. This dissertation consists of three essays on open government data and data analytics. It explores and contributes to the literature in introducing a new model for effective and efficient open governments,

introduces application of visualizations and data analytics of governmental data and a text mining analysis for government-related financial data.

The first essay considers the process and the use of open government data (OGD) initiatives by focusing on financial reporting with the use of procurement contracts. It pursues two main arguments regarding (i) the possibility of disseminating more financial data, especially procurement tenders, and how this could transform relationships between different levels of government and citizens; (ii) and discusses the level of data transparency based on the definition of the open data model; and more importantly, the study introduces a new model for effective and efficient open government data by adding new dimensions to the open data concept when utilized by governments.

The second essay investigates the use of visualization and cluster analysis techniques in governmental, publicly available datasets. It examines the utilization of advanced data mining techniques such as hierarchical, k-means clustering and visualization in two case studies. In the first case study, we explore the literature for the use of emerging data mining techniques in auditing. Then we apply k-means and hierarchical clustering on U.S. states financial statements data. In particular, we demonstrate how cluster analysis could be applied as supportive tools for auditing governmental bodies. The second case study utilizes the Volcker Alliance's Survey data results. The survey produces extensive information about how the different U.S. states score on an annual basis on budgeting using five measures. We apply cluster analysis and visualization on the budget data. On both case studies, we demonstrate how visualization and data analytics especially cluster analysis

could be used on governmental data and to help gain more insights about financial statements and budgeting.

The third essay focuses on text mining analytics for government-related financial data. Specially, utilizing a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the two public-private agencies of the Port Authority of New York and New Jersey (PANYNJ), and the New York Metropolitan Transportation Agency (MTA). This research initiative develops a methodology to classify tweets that are related to financial bonds. We apply a frame and slot approach from the artificial intelligence literature to operationalize the FIBO ontology in a public sector/municipalities business context. FIBO is part of the Enterprise Data Management Council (EDMC) and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. One contribution of this paper is that it is the first to recognize that the FIBO structure provides a grammar of financial concepts which can be used to classify social media. We show that this grammar can be used to mine semantic meaning from unstructured textual data. The Twitter stream is monitored and analyzed with frames derived from FIBO and using keywords. The ability of the FIBO frames to detect semantic meaning in tweets is compared with naïve keyword analysis and by determining the number of false positive classified from the Twitter stream. Using FIBO frames, constituent semantic structures can be uncovered to predict reactions to policies

and programs and perform other environmental scanning more quickly than by following the feeds manually.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my dissertation chair Professor Miklos A. Vasarhelyi for the continues and unconditional support all along my Ph.D. studies, for his precious time, immense knowledge, and strong encouragement; His guidance helped me throughout all the different stages of research and the writing of this thesis. Without his thoughtful and careful supervision, this thesis would never have taken shape. I would also like to express my sincere gratitude to my dissertation committee members; Dr. Deniz Appelbaum, your continuous support and insightful advice have strengthened me in becoming a better scholar. I enjoy every moment working with you; to Dr. Kevin Moffitt, your warm encouragement and constructive feedback have deeply motivated and supported me to better advance my research; and to Dr. Yaw M. Mensah, for your detailed comments and insightful feedback. I am also grateful to Dr. Nabil R. Adam for strengthening and improving my research skills; I am thankful to know and work with Dr. Rob Nehmer, his ideas, guidance, and support have contributed to this thesis. I also want to thank Irfan Bora and Dr. Basit Shafiq.

I wish to thank my entire family: my mother and father for providing unconditional support, love and encouragement throughout the years; to my lovely wife who stood beside me and supported me all the time; and to my brothers and sister.

Lastly, I would like to express my special thanks to fellow colleagues Ussama Yaqub, Farid Razzak, Assistant Dean Goncalo Filipe, and Barbara Jensen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This dissertation incorporates three essays on open government data and data analytics. Among the five chapters of the thesis, chapter one introduces the motivation and main research issues. The three essays are included in chapters two, three and four respectively. The first essay investigates open government data (OGD) initiatives in different countries and introduces a new model for effective and efficient open governments. The second essay examines the use of visualization and cluster analysis techniques. It utilizes tools for visualizations, hierarchical and k-means clustering methods on governmental data in two case studies. The third essay focuses on text mining analytics for government-related financial data. Particularly, utilizing a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the two public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). The last chapter concludes the dissertation by providing summaries of findings and future directions.

Over the past few years, we have seen a significant grown in the interest of open data, specifically open government data (OGD). This led to the availability of a large number of public sectors' datasets made available to the citizens or any other interested stakeholder. Thanks to the pressure being placed on all types of government organizations in order to release their raw data (Janssen, Charalabidis, & Zuiderwijk, 2012). The main motivations for publicizing access to raw materials and make it more transparent are that

it can help provide higher returns from the public utilization of such data, can give policymakers supportive data that can assess the process of making better decisions, can generate wealth through the development and creation of new and innovative products and services or enhance the current ones, and can involve the citizens to monitor and analyze publicly available datasets to help them evaluate and assess the performance of their governments (Verhulst & Young, 2017; Janssen et al., 2012).

Open data is often necessary for the public sector for many reasons such as public policy development and service delivery and many other things. In this thesis, we define open data as data that is publicly available, non-confidential and non-privacy restricted and is freely available, accessed, used, re-used and redistributed without any additional cost. It also refers to the openness and diffusion of data by the utilization of information and communication technologies (ICTs), which in returns enable more accessibility of data than before (Gonzalez-Zapata & Heeks, 2015; Janssen et al., 2012; Verhulst & Young, 2017).

To accelerate the dissemination and re-use of open data, many governments around the world have opened and published their data. The publication of government data, according to Bauer (2012), should comply with the open data definition and principles (Kucera & Chlapek, 2014). The importance of open data as significant added value for governments has gained momentum after President Obama promoted transparency as a key part of his campaign. President Barack Obama enacted the United States its first open data law in May 2009. The law requires all federal agencies to publish their data in a machine-

readable format that anyone can access through USASpending.gov[1]. On January 21, 2009, President Obama signed a memorandum regarding transparency and open government data initiative which stated: *"My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government."* (The White House, 2009).

Following President Obama memorandum to make the government more transparent, Data.gov[2] was then launched in May 2009. The President said: *"information maintained by the Federal Government is a national asset."* (Obama White House, 2019). The President's main objective is to increase public participation in government and to improve their everyday lives. As of July 2017, there are approximately more than 200,000 publicly available datasets at Data.gov (Data.gov, 2019). Another example of legislation for government information in the US is the executive order of May 13, 2013, for making openness and machine readability the new default for government data (The White House, 2013), wherein section 1., General Principles stated: *"Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and the scientific discovery that improves Americans' lives and contributes significantly to job creation."* (The White House, 2013). In addition, the order also stated

[1] USASpending.gov, (https://www.usaspending.gov) is the official source of spending data for the U.S. Government. It shows how the federal government spends the money every year.
[2] Data.gov (https://www.data.gov/) is a US government website aims to bring together datasets from hundreds of sources across the federal government and from 50 non-federal sources.

that a decade ago, when the U.S. Government made both weather and Global Positioning System (GPS) data freely available and accessed, how the American entrepreneurs created products such as navigation systems, weather warning systems, location-based systems, and much more that improves the lives of millions of Americans and therefore leading to economic growth and job creation.

The first essay considers the process and the use of open government data (OGD) initiatives by focusing on financial reporting with a comparison of procurement contracts between the Federal Republic of Brazil and Saudi Arabia. It pursues two main arguments regarding (i) the possibility of disseminating more financial data in Saudi Arabia, especially procurement tenders, and how this could transform relationships between different levels of government and citizens; (ii) and discusses and suggests new dimensions for the open data definition when coupled with government data to increase transparency in governments. In addition to data availability, openness, and access, which all define open data (Group, 2017), we have added two additional dimensions, which are data analytics, and applications within governments.

The second essay examines the use of visualization and cluster analysis techniques using governmental, publicly available datasets. It investigates the use of advanced data mining techniques such as hierarchical and k-means clustering methodologies in two case studies. In the first case study, we explore the literature for the use of emerging data mining techniques in auditing. Then we apply k-means and hierarchical clustering techniques using U.S. states financial statements data. In particular, we demonstrate how cluster analysis can be applied as supportive tools for auditing. The second case study utilizes the Volcker Alliance's Survey data results. The survey produced extensive information about

how the different U.S. states score on an annual basis on budgeting using five measures or variables. We apply cluster analysis and visualization on the budget information, and we demonstrate how visualization and data clustering can be used on governmental data and to help gain more insights about budgeting.

To evaluate our results, we apply a clustering assessment methodology that uses bootstrap resampling which assesses how stable a given cluster is. This methodology is essential when using clustering algorithms such as k-means and hierarchical clustering, especially the latter, where most of the time you have to use an a priori method of choosing the optimal number of clusters. The technique was first introduced by Christian Hennig in 2007 on his published paper, "Cluster-wise assessment of cluster stability," (Hennig, 2007; Hennig, 2018). Clustering algorithms mostly produce clusters that represent the actual structure and relationships in the data, and then few clusters are miscellaneous. One common way to check if a cluster represents true structure of the data is to see how each cluster holding up under apparent variations in the dataset.

Looking at the clusterwise stability assessment results of the two clusters: k-means and hierarchical, we can notice that, overall, hierarchical clustering performs better than k-means with higher mean values with less dissolved clusters. However, assessing clustering stability is not the only important validity criterion since some clustering methods could result in stable but not valid clusters (Hennig, 2007; Hennig, 2018).

In the third essay, we use an ontology-based framework to classify social media data where we utilize a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from Twitter platform regarding financial and budget information in the public sector, namely the two public-private agencies of the

Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). We apply a frame and slot approach from the artificial intelligence literature to operationalize the FIBO ontology in a public sector/municipalities business context. FIBO is part of the Enterprise Data Management Council (EDMC) and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. One contribution of this paper is that it is the first to recognize that the FIBO structure provides a grammar of financial concepts. We show that this grammar can be used to mine semantic meaning from unstructured textual data. Twitter streams will be monitored and analyzed with frames derived from FIBO and keywords. The ability of the FIBO frames to detect semantic meaning in tweets is compared with naïve keyword analysis. Using FIBO frames, constituent semantic structures can be uncovered to predict reactions to policies and programs more quickly than by following the feeds manually.

Recently, a lot of businesses and research communities have paid attention to the utilization of social media and big data. In this last study, we decide to choose the social media platform Twitter as our source of data because of the attention, userbase and most importantly, unlike Facebook data, Twitter data is considered open, therefore, researchers, governments, and business communities could access Twitter data using Twitter Application Programming Interface (API). Thus, offers them to access Twitter data on an unprecedented scale and to analyze such data for various problems in different domains (Chae, 2015). Twitter has gained a lot of attention in recent years, making it the largest microblogging application. It is estimated that Twitter, as of the fourth quarter of 2018, has an average of 321 million monthly active users with around 500 million tweets sent per

day. Twitter has become an effective and efficient way to communicate news, express opinions and ideas (O'Regan, 2018; Statista, 2019; Yaqub, 2018; Omnicore, 2019).

CHAPTER 2: A NEW MODEL FOR EFFECTIVE AND EFFICIENT OPEN GOVERNMENT DATA

ABSTRACT: Open Government Data (OGD) is attracting stakeholders from different backgrounds. The call for OGD has been especially pronounced in the last six or seven years. Open government data demand accelerated after the launch of the United States open government data initiative portal in 2009, followed by the United Kingdom in 2010. Before that, the availability and accessibility of government data were limited to certain executives and few government employees, whereas for others, it was either partially available or completely unavailable. Publishing government data, thereby making it available to the public, could be useful in many ways such as increasing transparency and accountability in governments, increasing overall efficiency and performance, encouraging publics' engagement, and achieving trust and reputation. As an example of the role that OGD may provide, this paper compares the different financial reporting and auditing systems in the public sector between Brazil and Saudi Arabia. Also, the paper examines open government data initiatives among different countries with the focus of the Republic of Brazil and the Kingdom of Saudi Arabia's open data portals. Moreover, it assesses the level of data transparency based on the definition of the open data model, and more importantly, the paper suggests new dimensions to the open data concept when utilized by governments. In addition, it argues that the open government data in Saudi Arabia, which is an emerging initiative in a country that has centralized power, could be improved dramatically. We demonstrate that by using a sample of procurement contracts

data taken from the Council of Saudi Chambers website, which is publicly available and shows the potential of monitoring or auditing public spending.

Keywords: financial reporting; auditing; open government data; data transparency; OGD.

2.1 INTRODUCTION AND BACKGROUND

Since the time when the United States established the law of Freedom of Information Act (FOIA) in 1967, which provides the public the right to request access to any federal agency records, the interest in open government data (OGD) has increased significantly. Specifically, this interest evolved after the Obama administration decided to open federal government data and make it publicly available after the launch of data.gov (Gonzalez-Zapata & Heeks, 2015). President Obama promised during his campaign *"to create a transparent and connected democracy"*; which he then followed by signing one of his first orders in office: *"to create an unprecedented level of openness in Government"* (Coglianese, 2009). In 2010, after the 2009 President Obama's decision to open government data to the public, the United Kingdom had followed the U.S. to open its government data and to make it available to the public. After that, many governments around the world, and global organizations such as the United Nations and the World Bank have followed the U.S. and the U.K. initiatives. Achieving transparency as a goal for open government data (OGD) can be very challenging. Transparency can play a primary role in governments' decision making. By making government data available to the public, citizens, professionals, and other interested groups can access the data to help monitor

public spending and increase overall participation. Thus, enabling government officials to make better decisions (Coglianese, 2009).

In addition to better decision making, transparency can help avoid or minimize the abuse of government resources. When government officials know that their work is being monitored, they will be less likely to commit mistakes. However, too much transparency in some places may not be beneficial, as it may inhibit decision making of some officials. So, ideally, there should be a balance, where transparency to a certain level will help both officials and outsiders. Governments should also have a limit, where some confidential information should not be compromised and shared. Even FOIA itself prevents the disclosure of trade secrets, national security, and personal information (Coglianese, 2009).

Yu and Robinson (2011) identify and discuss three foundations of OGD which are: government data, open government, and open data as shown in Figure 1 below (Gonzalez-Zapata & Heeks, 2015). Government data, on the one hand, means all government-related data that is available and easily accessed. On the other hand, open government refers to allowing access for previously disclosed government information. Finally, the concept of open data refers to the openness and diffusion of any data by the utilization of information and communication technologies (ICTs), which in turn enables more accessibility of data than before (Gonzalez-Zapata & Heeks, 2015).

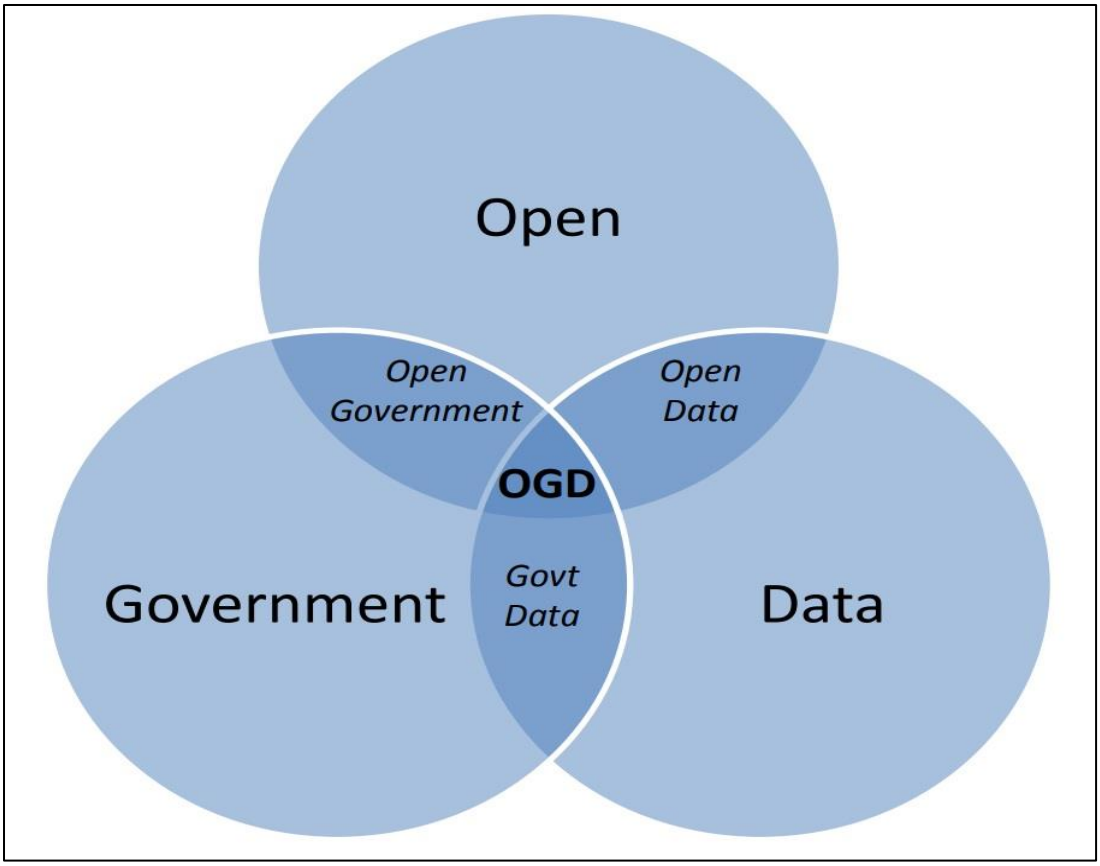


Figure 1: Open Government Data Foundations (Yu & Robinson, 2012)

The flexibility of information and communication technologies (ICTs) in e-government can play a crucial role in increasing openness and transparency in governments and in limiting or eliminating misuse of resources. A study conducted by Anderson (2009), has shown that only implementing e-government initiatives will reduce misuse of resources significantly due to the fact that e-government increases accountability and transparency, and limits contact between officials and the public. Transparency of government information is considered, nowadays, as one of the main building blocks in a modern democracy (Morrison, 2014). It is also regarded as boosting the general public's trust in different government agencies, reducing misuse of resources, helping and improving

decision making, and promoting accuracy of government information. Consequently, disseminating transparent information to citizens, or any other interested stakeholder, can play a significant role in fighting and eliminating misuse of government resources (Bertot, Jaeger, Grimes, 2010).

This paper considers the process and the use of open government data (OGD) initiatives by focusing on financial reporting with a comparison of procurement contracts of the Federal Republic of Brazil and Saudi Arabia. It pursues two main arguments regarding (i) the possibility of disseminating more financial data in Saudi Arabia, especially procurement tenders, and how this could transform relationships between different levels of government and citizens; (ii) and also discusses and suggests new dimensions to the open data definition when coupled with government data to increase transparency in governments. In addition to data availability, openness, and access, which all define open data (Group, 2017), we have added two additional dimensions, which are data analytics, and applications within governments. Data analytics can facilitate citizen's use and analysis of data, and the applications can allow citizens and any other interested stakeholder to reach out and report any problems or concerns to the government.

Our main objective for choosing the case of the Republic of Brazil's open government data initiatives and compare it to the Kingdom of Saudi Arabia's OGD is because Brazil is a founding member and part of the Open Government Partnership (OGP),[1] (dos Santos Brito, da Silva Costa, Garcia, & de Lemos Meira, 2014). Both countries have developing economies and are G20 members. Furthermore, Brazilian OGD

[1] The Open Government Partnership (OGP) is a multilateral initiative that aims to bring together governments to create action plans to promote openness, empower citizens, adopt new technologies and fight misuse of government power (https://www.opengovpartnership.org/).

initiatives are considered advanced, and thus, comparing them to Saudi Arabia's open government data initiatives would be of great interest.

The rest of the paper is organized as follows. First, we give an overview of the public sector financial reporting and auditing in Brazil and compare it to Saudi Arabia. Then, we highlight open government data (OGD) initiatives in some countries. After which we discuss the level of data transparency of government procurement contracts (GPC) between the Republic of Brazil and the Kingdom of Saudi Arabia and more importantly we introduce a new model for effective and efficient open government data. Finally, we conclude our work and provide some directions for future research.

2.2 PUBLIC SECTOR FINANCIAL REPORTING AND AUDITING: A COMPARISON BETWEEN BRAZIL AND SAUDI ARABIA

FEDERATIVE REPUBLIC OF BRAZIL

The Federative Republic of Brazil is comprised of 26 states plus the Federal District (the capital), Brasilia. The federal government has three independent branches (the division of powers) which are the following (Cardoso, 2017):

- Executive: the executive power is headed by the president and advised by Ministers. Where the president is both the head of government and the head of state.
- Legislative: formed by the Federal Senate and the House of Representatives.

- Judicial: judicial power is exercised by the Supreme Federal Court, the Superior Court of Justice, the regional federal courts and the National Justice Council.

The public sector financial reporting in Brazil uses the Treasury's Public Sector Accounting Standards (MCASP), which is converging towards the International Public Sector Accounting Standards (IPSAS). The convergence process started in 2008 (de Aquino & Batley, 2015). The Brazilian government financial reports are publicly available at:

- http://www.stn.gov.br/;
- http://www.cgu.gov.br/.

Auditing for the public sector in Brazil is performed by The Tribunal de Contas da União (TCU), or the Supreme Court of Accounts, which is the official Brazilian federal accountability office. TCU is an arm of the legislative power of the Brazilian government and is mainly responsible for the monitoring of all federal public spending. TCU has highly qualified employees to investigate and prevent any type of possible misuse of funds (Taylor & Buranelli, 2007). Auditing reports are publicly available at: (www.tcu.gov.br/).

The division of powers at the states level are the following (de Aquino & Batley, 2015):

- Executive: the governor of each state heads the executive power.
- Legislative: consists of state deputies.
- Judiciary: is made up of a Court of Justice and the Judges of law.

Financial reports for each state are publicly available at (http://www.stn.gov.br/), at each state's webpage. Also, the data is comparable; every state applies the same chart of

accounts. On the other hand, auditing is performed by States' Courts of Accounts (TCE) and each state has a TCE. Figure 2 below shows the different states in Brazil.

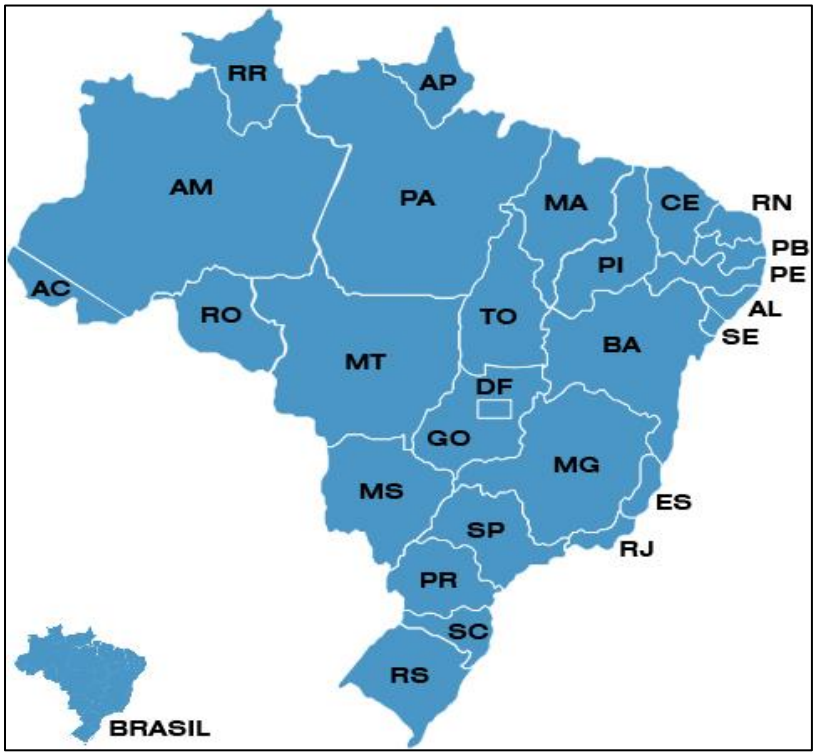


Figure 2: States of Brazil

Finally, Brazil contains 5,570 municipalities. Each municipality has its own constitutional power and legislation. However, unlike the federal government and the states, Brazilian municipalities do not have judicial power (de Aquino & Batley, 2015).

Municipalities financial reporting is in accordance to the Treasury's Public Sector Accounting Standards (MCASP), which is under the convergence towards the International Public Sector Accounting Standards (IPSAS), the same applied by the Federal government and states. Financial reports are publicly available at (http://www.stn.gov.br/). Data is comparable; every municipality applies the same chart of accounts applied by the states. In

terms of auditing, it is performed by the States' Courts of Accounts (TCE), but only two municipalities, Rio de Janeiro and Sao Paulo, have a Municipal Court of Accounts (TCM) (de Aquino & Batley, 2015).

THE KINGDOM OF SAUDI ARABIA

The Kingdom of Saudi Arabia has a monarchy system where the King combines executive, legislative, and judicial power. The basis of the country's legislation is formed from royal decrees. The king is a prime minister and is a head of a council of ministers and consultative assembly. The primary source of law is the Sharia law, which is derived from the Qur'an and Sunnah, the Prophet's traditions, Esmaeili (2009). The judges apply their personal interpretations or views of the Sharia to all civil and criminal cases. Royal decrees on the other hand supplement Sharia in other areas such as commercial, corporate, and labor law.

The administrative divisions in the Kingdom of Saudi Arabia divided into 13 regions including the capital, Riyadh. Those regions are further divided into 118 governorates. Each governorate has its own municipality headed by mayors. The governorates in Saudi Arabia are further sub-divided into sub-governorates.

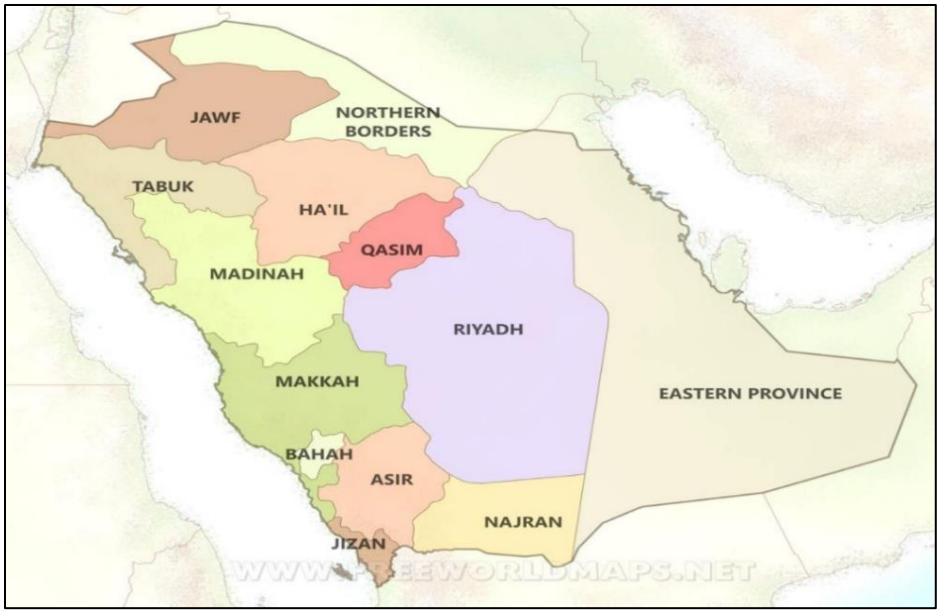


Figure 3: Regions of Saudi Arabia (source: www.freeworldmaps.net)

The public sector financial reporting in Saudi Arabia is under the control of the Ministry of Finance and supervised by the General Auditing Bureau and follows the International Public Sector Accounting Standards (IPSAS). This process began when the Ministry of Finance in the Kingdom of Saudi Arabia and the Saudi Audit Bureau initiated a study in 2008 to evaluate and analyze the government's financial reporting and began adopting the IPSAS (Dartdeloittecom, 2017). Saudi's government financial reports are publicly available at:

- The Open Data Portal in Saudi Arabia, (http://www.data.gov.sa/).
- The Ministry of Finance, (http://www.mof.gov.sa/).
- The Council of Saudi Chambers, (http://www.csc.org.sa/).

However, the data is not comprehensive and, in some cases, not complete and up to date. For instance, the Council of Saudi Chambers website (http://www.csc.org.sa/), is

a government site that publishes governments' procurement data, and most of the contracts published are missing dates, contract numbers, and details of contracts.

Auditing for the public sector in Saudi Arabia is maintained by Saudi Arabia's General Auditing Bureau (GAB). GAB is a professional, independent and credible auditing institution that enhances the efficiency of auditees, transparency, governance, and accountability (Saudi General Auditing Bureau, 2017). On March 18, 2011, the King ordered the establishment of the National Anti-Corruption Commission, Nazaha. It acts as a supportive branch to the Saudi General Auditing Bureau. The Commission's tasks include the monitoring of the implementation of orders and instructions of the public affairs. It also includes the monitoring of any misuse of resources by any government's official. All government agencies should in terms report all its approved projects, contracts, and operations to Nazaha official portal at the following: (https://www.nazaha.gov.sa/).

2.3 OVERVIEW OF OPEN GOVERNMENT DATA INITIATIVES

First, what does Open Data mean? Open Data as defined by the Open Definition, (Group, 2017): *"Open Data is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike."* To summarize, here are the main points:

- Availability and Access: means that the data must be available at no more than its reproduction cost, preferably available over the internet. It should also be available in a machine-readable format.

- Re-use and Redistribution: the data must be presented in terms and conditions that permit the re-use and redistribution of datasets.
- Universal Participation: means that there must be no discrimination on who uses or redistributes the data.

According to Maude (2012), Open Government Data (OGD) is defined as *"Public Sector Information that has been made available to the public as Open Data."*. It also denotes all governmental information available online (Matheus, Ribeiro, & Vaz, 2012). Governments around the globe are now releasing a massive amount of governmental data every day. Through all government levels, millions of records are collected and stored. Government data domain ranges from national statistics, weather forecast, water quality, procurement contracts, national maps, public sector budgeting, and performance, to all other kinds of data such as policies and inspections (AlRushaid & Saudagar, 2016). This published data could be analyzed by third parties enabling them to create innovative products and services. Open data should result in an open government where the government should act as an open system and interacts with its surrounding environment. As a result, the data should not only be published to the public, but also the government should encourage third parties for providing feedback to improve government efficiency if necessary (Janssen et al., 2012).

In 2009, the U.S. began publishing all government data to the public except personal information and national security data (Dai & Li, 2016). This initiative started when President Barack Obama, on his first full day in office as a president, in January 2009, signed a Memorandum on transparency and open government data. He announced that the new administration would start a new open government data strategy. In a note for

the head of Executive Departments and Agencies, stated, *"... We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government."* (Huijboom & Van den Broek, 2011). On December 8, 2009, The White House issued an open government directive asking all federal agencies for immediate steps to achieve a high-level of transparency, participation, and collaboration (White House, 2017). The government open data webpage (data.gov) is the main site that publishes government data.

Various countries around the world have been inspired by President Obama's initiative for open government data. In December 2009, the United Kingdom's followed the U.S. initiative encouraging different governments' agencies to publish their data for the public. David Cameron, the former U.K. prime minister, in a podcast in 2010 stated that *"extend transparency as far and as wide as possible. By bringing information out into the open, you'll be able to hold government and public services to account."*, (Kozlowski, 2016). Also, he stated regarding the importance of open data that *"If there's one thing I've noticed since doing this job, it's how all the information about government - the money it spends, where it spends it, the results it achieves - how so much of it is locked away in a vault marked sort of private for the eyes of ministers and officials only."*, he continued *"I think this is ridiculous. It's your money, your government, you should know what's going on."* (Cameron, 2010; Kozlowski, 2016). Additionally, the Australian government, in May 2010, declared an open government initiative. In July 2010, Denmark and Spain governments also started open government data initiatives and many more (Huijboom & Van den Broek, 2011).

In December 2010, the federal government of Brazil launched its first site to provide public data or OGD. The public portal can be found at (https://dados.gov.br/). The OGD project was founded by an Information Organizing Committee from the presidency which consists of the following government branches: The Office of Deputy Information for Decision Support, President's Chief of Staff, Department of Logistics and Information Technology, Institutional Relations Secretariat (SRI-PR), and two other IT companies that are owned by the federal government (Matheus et al., 2012). The data, at first, came from ministries, federal government agencies, and other public organizations. Following that, the federal government has initiated a project to create the Open Data National Infrastructure (ODNI). The aim of this project is to facilitate interoperability and standardization of all information resources (Matheus et al., 2012).

In Saudi Arabia, the government has invested heavily in its information and communication technology (ICT) infrastructures. In 2005, Yesser, an official e-government program was launched. The primary goal of the program is to encourage all government transactions to shift into digital programs. Following the implementation of e-government, specifically, in 2011, the Kingdom of Saudi Arabia launched its own open government data initiatives. Saudi's open government data initiatives aim at delivering the data through a national portal at (http://www.data.gov.sa/) as well as through the Ministry of Economy website. The government's main goal of making the data available was to increase transparency, promote citizens' participation, and encourages innovation of governments (AlRushaid & Saudagar, 2016). However, the adoption and implementation of open government data initiatives among different ministries have faced many challenges. Therefore, the level of data openness is not as high as other countries such as

the U.S., U.K. or Brazil. For instance, Table 1 below (Knowledge, 2017) shows the level of data transparency of procurement tenders presented at the Global Open Data Index (Knowledge, 2017) which measures data openness around the world. The green color on the breakdown column denotes "Yes", the red color denotes "No", and the blue color denotes "Unsure":

Place	Score	Breakdown*	Year	Location (URL)
United States	100%		2015	https://catalog.data.gov
Australia	100%		2015	http://data.gov.au
United Kingdom	90%		2015	https://www.gov.uk/contracts-finder
Spain	90%		2015	https://contrataciondelestado.es
Brazil	80%		2015	http://dados.gov.br
Denmark	45%		2015	http://envs.au.dk
Saudi Arabia	35%		2015	http://www.csc.org.sa

*Breakdown (Data-Availability).

Table 1: Procurement Tenders Evaluation (Knowledge, 2017)

As denoted from the table above, Saudi Arabia has the lowest score in its published procurement contracts among the other countries after Denmark. Brazil has a high score of 80% of data completeness and availability.

2.4 DISCUSSION

In this paper, we assess open government data (OGD) initiatives using governments' procurement contract data in Saudi Arabia and compare it with the Brazilian procurement contracts. The main objectives of comparing Saudi Arabia's tenders with Brazilian tenders are because Brazil is a founding member and part of the Open Government Partnership (OGP), the Brazilian open government data initiatives are considered more advanced compared to the Saudi's OGD. In addition, both countries have developing economies and are G20 members. Before we discuss the available procurement contracts, first, we will give examples of the kind of datasets that are publicly available in both Saudi Arabia and Brazil and its sources.

First, the Saudi Arabian government, as mentioned before, composed of ministries and those ministries are varied in terms of the data that they expose online. For instance, the Ministry of Finance is considered one of the main sources for open datasets. The data could be found on the official website of the ministry at: (https://www.mof.gov.sa/). The following table shows the datasets that are publicly available:

Category	**No. of datasets**	**Description**	**Date Published**	**File Format**	**Up to date**
Statistical Data	1	e-services user satisfactions	1438 AH (2017)	Pdf, excel, word	yes
All Cities Cost of Living Index	2	Public Investment Fund, All Cities Cost of Living Index	1435 AH (2014)	Xml, pdf, excel	No
Balance of payments estimates	1	Balance of payments	1435 AH (2014)	Xml, pdf, excel	No
Budget	4	Actual Revenues & Expenditures, budget	1435 AH (2014)	Xml, pdf, excel	No

Merchandise Trade	4	Kingdom's imports, export, import by items, export by major items	1435 AH (2014)	Xml, pdf, excel	No
National Accounts	2	Constant and current price GDP by economic activity and sector	1435 AH (2014)	Xml, pdf, excel	No
Public Credit Institutions and Programs	5	Domestic Loans Program, industrial and agricultural and real estate dev. Fund, credit and savings banks	1435 AH (2014)	Xml, pdf, excel	No
Key Performance Indicators of Saudi Economy	14	See: https://www.mof.gov.sa/	1438 AH (2017)	excel	Partially (some datasets are up to date)

Table 2: Datasets at the Saudi Arabia's Ministry of Finance (Source: https://www.mof.gov.sa/)

Other than the above datasets, there is only information about the Ministry of Finance's contracts without any explanations of the values of those contracts.

Also, the Ministry of Higher Education publishes only statistical data on its open data portal, which is available at: (https://www.moe.gov.sa/en/opendata/Pages/OpenDatasets.aspx/). The data published is in eXceL Spreadsheet (XLS), Portable Document Format (PDF) and text file formats. They have stated that their open data policies will encourage the participation of the public and increase transparency. Moreover, the Ministry of Foreign Affairs, the Ministry of Labor and Social Development, the Ministry of Health, and the Ministry of Commerce and Investment also claiming to have open data policies; however, only statistical data is available at their open data portals. For instance, the Ministry of Foreign Affairs publishes data related to visa statistics, flight clearance system statistics and Saudi affairs system

statistics. The data is published in Word documents and are not up to date. Additionally, the Ministry of Interior and the Ministry of Haj and Umrah does not hold any open data initiatives on their web portals.

Second, in Brazil, Dados[2], the official Brazilian open data portal is organized to simplify the use and search for any dataset. The portal aims to centralize the search and access to public data and to make all governmental data available for citizens. For instance, data on transportation systems, supplementary health, education indicators, public safety, electoral processes, government expenditure, are some but not all examples of the datasets available (Dados.gov.br. n.d., 2017). In all, the portal contains around 3,190 datasets.

The portal categorizes the data by different categories. For instance, it categorizes the data based on organizations, such as the Central Bank of Brazil, Brazilian Institute of Geography and Statistics, State of Alagoas, Ministry of Health, Ministry of Finance, National Telecommunication Agency, Ministry of Tourism. It also categorizes the data based on different groups, such as government and policies, health, education, public equipment and supplies, work, defense and security, industry, census, housing, sanitation and urban planning, science, information, and communication, etc. The datasets also grouped by hash-tags. Also, the datasets are available in a wide range of formats such as Comma-Separated Values (CSV), HyperText Markup Language (HTML), JavaScript Object Notation (JSON), Web Services Description Language (WSDL) and many more. Finally, the datasets are also grouped based on their download license, whether it is for open database license (OBDL), creative common attribution share license, or creative

[2] Dados: Brazilian Open Data Portal (http://dados.gov.br/).

common non-commercial. There also exists training for cataloging data that aims to provide the public with the knowledge and skill sets necessary to help promote open data initiatives at different public organizations.

2.5 DATA TRANSPARENCY ASSESSMENT: A NEW MODEL FOR EFFECTIVE AND EFFICIENT OPEN GOVERNMENT DATA (OGD)

In the following section, we will discuss the level of data transparency of government procurement contracts (GPC) between the Republic of Brazil and the Kingdom of Saudi Arabia based on a 4-condition model. This model will assess government data transparency based on the following four conditions: data availability, data openness, data analytics, and different applications within government portals. Looking at the open data definition by (Group, 2017), there are three important characteristics that define open data. They are availability and access, re-use and redistribution, and universal participation. These three characteristics are equivalent to our first 2-points of the 4-condition model which are data availability and openness. However, we believe that when open data is coupled with government, two more characteristics or dimensions should be added to the open data definition. Those are data analytics and applications. The 4-condition model, presented in Figure 4 below, for assessing data transparency in open government data environment is consisting of the following attributes:

I. Data Availability of Government's Procurement Contracts.

II. Data Openness: once the data is available, into what extent the data is open.

III. Data Analytics: if the data is available and open to an extent, is it analyzable.

IV. Applications within Government Websites.

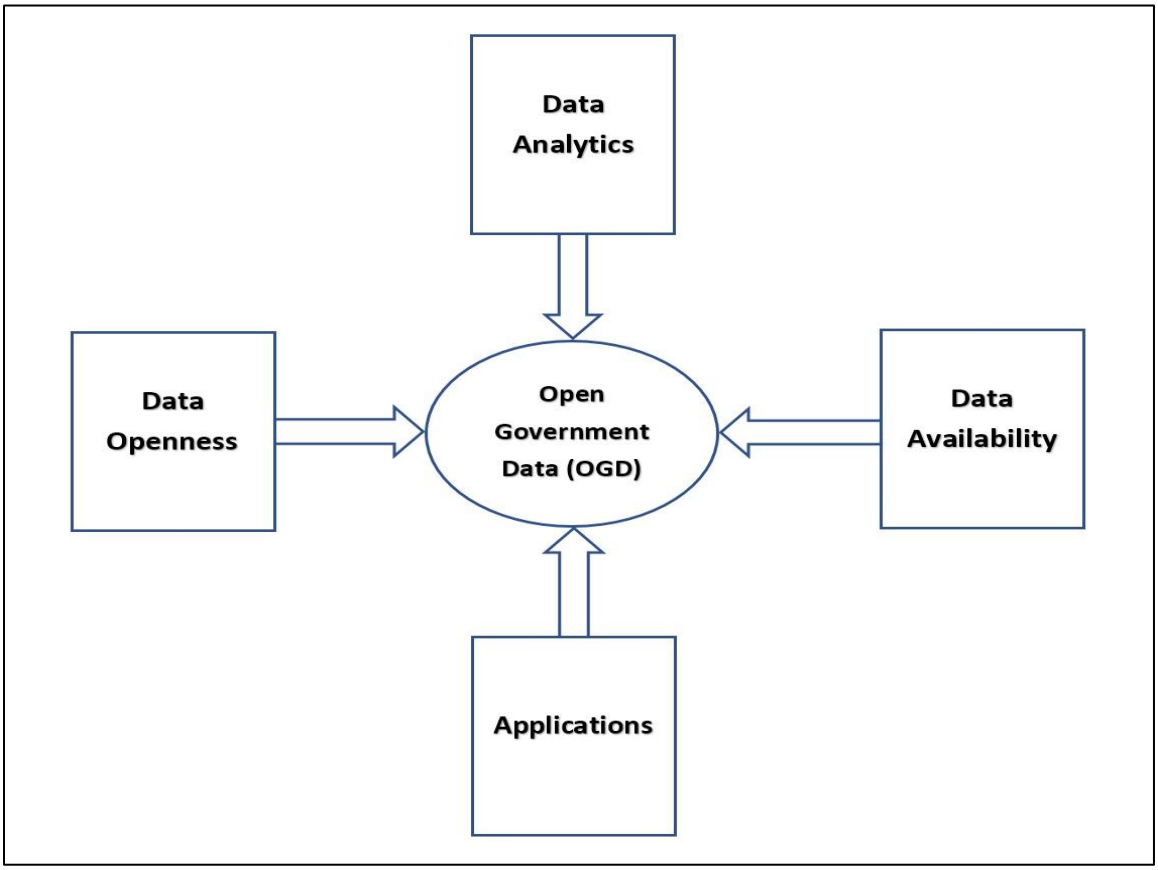


Figure 4: Proposed Model for Effective and Efficient Open Governments Data (OGD)

2.5.1 DATA AVAILABILITY

In Saudi Arabia, the government's procurement contracts are limited in availability to the public. There is no central place for government procurement contracts. Some contracts could be found at the Council of Saudi Chambers website at (http://www.csc.org.sa/), with few others at the Ministry of Finance portal at:

(https://www.mof.gov.sa/). Even though there is an electronic government procurement system, it is only for bidding; and the information about the contracts is only available for the private sector, i.e., in order to access it, one needs to be registered as part of an organization. The electronic government procurement system could be found at Etimad (http://www.saudiegp.sa/). This portal claims to support the principle of transparency among the different government agencies; but in fact, the system is designed to standardize the processes and procedures for the suppliers to apply for government procurements. Most of the government related procurement contracts are presented at the Council of Saudi Chambers website and it can be found at: (http://www.csc.org.sa/). The site has procurement contracts published in Hypertext Markup Language (HTML). In addition, the published contracts are abstracts and not comprehensive. For instance, the all-time total number of government procurement contracts presented are around 12,534. Those are broken down into around 627 pages, and in each page, there are around 20 procurement contracts.

Whereas in Brazil, they have published a centralized federal e-procurement information system called “SIASG” which stands for the Open Data of the Integrated System of Administration and General Services. They have also provided Application Programming Interfaces (APIs) to facilitate the access and download of federal procurement contracts by citizens (Dai & Li, 2016). The data could be found at the Brazilian open data portal at (http://www.dados.gov.br/). There is a total of 470,683 published procurement contracts.

2.5.2 DATA OPENNESS

Data openness means that once the data is available, to what extent or to what degree the data is open? In this section, we will demonstrate the level of openness of the data by demonstrating a sample of procurement contracts for comparison purposes between Brazil and Saudi Arabia. In Saudi Arabia, the following is an HTML sample of one contract (written in Arabic):

انشاء حدائق المرحلة الثانية بالطائف	
رقم العقد	
التصنيف الأساسي	
إسم الجهة الحكومية	أمانة محافظة الطائف
الجهة المنفذة	مصنع الأمانة والإخلاص للتجارة
قيمة العقد	9.445.000.00 ريال
تاريخ توقيع العقد	
مدة العقد	540 يوم
مكان تنفيذ العقد	محافظة الطائف
تفاصيل أخرى	تفاصيل أخري

Table 3: A procurement Contract from the Council of Saudi Chambers, Written in Arabic language (source: http://www.csc.org.sa/Arabic/)

The following table is a translation of the above contract (Table 4):

Contract Name	Garden Creation - Phase 2
Contract Number	
Class	
Government Agency	Taif Municipal
Contractor Name	Amanah and Iklas Factory
Value of Contract	9,445,000.00 SAR
Date of Signature	
Contract Duration	540 Days
Location of the Project	Taif Governorate
More Details	

Table 4: A Procurement Contract Taken from the Council of Saudi Chambers Website

As one can see from the table above, the contract number, class and signature date are unknown. The contract also missing the effective start and end dates of the project.

However, in Brazil, among the 470,683 published procurement contracts, there are 35,516 contracts that are missing supplier information. And 16,167 contracts that are missing information about bidding mode; moreover, the dates are not represented in about 1,000 contracts (Dai & Li, 2016). The following table shows an example of how procurements are presented in Brazil. It is easily accessible and there are tools for filtering the contracts:

Add Filter

Grid | Graph | Map | 1484 records | « | 1 - 100 | » | Search data ... | Go » | Filters

_id	valor_total	objeto	numero	campus	data_ini...	valor_ex...	id	data_fim
1	116689.88	CONTRATAÇÃO DE EMPRES...	142/2011	SGA	2011-06-...	103448.29	66	2012-06-.
2	22800.0	CONTRATAÇÃO DE EMPRES...	135/2011 RE	RE	2011-06-...	22740.0	77	2012-06-.
3	29400.0	CONTRATAÇÃO DE EMPRES...	97/2015	CAL	2015-09-...	24900.0	1251	2016-09-.
4	4415.02	Contratação de empresa espec...	030/2017	AP	2017-04-...	3351.98	1549	2018-03-.
5	50000.0	Serviços de publicação no Diár...	169/2012	CN	2012-08-...	48670.91	466	2013-08-.
6	17611.2	Aquisição de gêneros alimentíc...	141/2016-PR...	NC	2016-12-...	0.0	1526	2017-12-.
7	1000497...	Prestação de serviços continua...	183/2010	MC	2010-10-...	1045720...	100	2011-09-.
8	19800.0	Prestação, pela ECT, de serviç...	217/2011_S...	SC	2011-08-...	9812.08	424	2012-08-.
9	6120.0	Concessão de direito de uso d...	065/2010	CA	2010-05-...	0.0	123	2013-09-.
10	15000.0	Prestação do serviço de fornec...	89/2016	CN	2016-05-...	3853.67	1400	2018-05-.
11	489598.32	Contratação de empresa de for...	61/2016-PR...	PAAS	2016-07-...	201733.93	1404	2017-06-.
12	5279.82	Aquisição de gêneros alimentíc...	203/2014	CN	2014-12-...	5278.92	1119	2015-08-.
13	167575.47	Prestação de Serviço de Motori...	371/2013 CO...	CANG	2013-08-...	126551.97	779	2014-08-.
14	18670.45	Aquisição de Gêneros alimentí...	156/2015-PR...	SC	2016-01-...	18671.64	1277	2016-09-.
15	14599.31	Aquisição de Gêneros de Alim...	173/2016	AP	2016-12-...	2176.0	1462	2017-11-.

Table 5: Source (http://dados.gov.br/dataset/)

2.5.3 DATA ANALYTICS

Here, we ask the following questions: if the data is available and open to an extent, is it analyzable? Are there any applications or tools that facilitate the analysis of the data? Are there tutorials (manuals) explaining the different datasets and how to use them? To answer these questions, we pull a sample of procurement contracts data and examine if it is analyzable.

Using Saudi Arabia's open procurement contracts, the contracts are neither analyzable nor prepared in a ready-to-use format. It is just plain HTML thrown on hundreds of pages. Next, we are going to reorganize, clean, and extract information for demonstration purposes to show the potential of having the data in machine-readable formats. The following is a sample of procurement contracts data collected from the Council of Saudi Chambers website and put together in a machine-readable format:

Contract Name	Contract Number	Government Agency	Contractor Name	Value of Contract-SAR	Date of Signature	Contract Duration-Days	Location of the Project
Garden Creation - Phase 2		Taif Municipal	Amanah and Iklas Factory	9,445,000		540	Taif Governorate
Recunstruction - airplane lines & parking Spaces		Ministry of Defense	Almabani Company	38,208,105		266	King Khaled Airforce in Riyadh
20-feet Storage Boxes		Ministry of Interior	Bander Khaled Albahouth Corpora	338,700		120	Riyadh
Food Supplies for Soldiers		Ministry of Defense	Muhammed Alburaidi Corporatio	6,418,800		1095	Northern Region
Continuing Water Projects		Ministry of Water and Elictricity	Aljazea Company	46,927,600		720	Jazan Region
Digging 8-water Wells		Ministry of Water and Elictricity	Humod Althubaiti Corporation	2,657,600		720	Makkah Region
Project of operation and maintenance of the water network		Ministry of Water and Elictricity	Bin Sammar Corporation	7,675,300		1095	Hail Region
Digging 9-water Wells		Ministry of Water and Elictricity	Humod Althubaiti Corporation	2,986,200		540	Makkah Region
Project of operation and maintenance	5857010714	Ministry of Water and Elictricity	Osool Corporation	6,947,619		1080	Asir Region
Renovation and maintenance of Toilets	1010252975	Ministry of Islamic Affairs	Abaad Corporation	76,266		180	Riyadh
Expansion and renovation of a mosque	1010252975	Ministry of Islamic Affairs	Abaad Corporation	572,705		180	Riyadh
Playground Construction		Riyadh Municipal	Abdulaziz Alotaibi Corporation	1,839,394		450	Riyadh
Playground Construction		Riyadh Municipal	Alsalel Corporation	1,346,980		450	Riyadh
Parking and Storage Construction		Riyadh Municipal	Menaor Alshebani corporation	2,495,715		540	Riyadh
Total				127,935,984		7976	

Table 6: Saudi Procurement Contracts Sample

Putting the data in such representation will help the public with varying skills to use the data for information extraction. Thus, enabling the public to monitor government

spending. For instance, we can use the above data sample to perform data visualization. For instance, to create graphs to examine some key data patterns and metrics as follow:

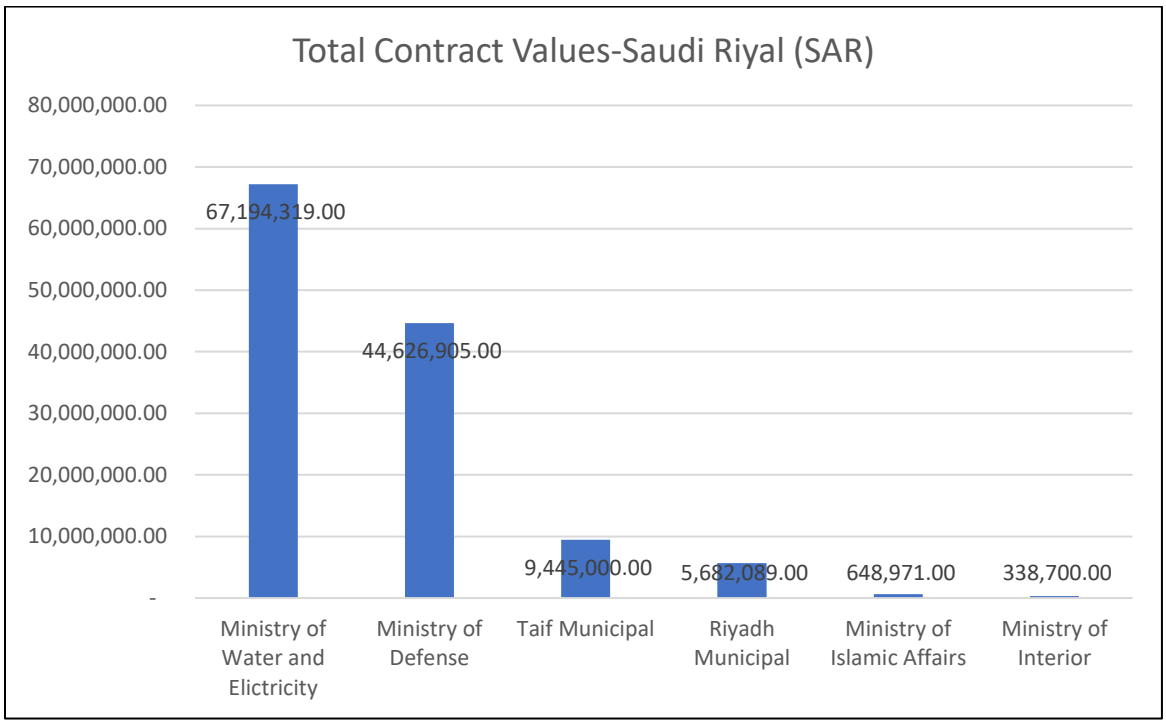


Figure 5: Total Contract Values by Government Agency

From the above figure, one can clearly see that the ministry of water and electricity takes the highest portion of total contract values. The values are in millions of Saudi Riyals (SAR).

Looking at Figure 6 below, one may see a clear downward trend going between the value of the contract and the number of days; however, there are some outliers that could draw some attention and require further analysis; for instance, notice the 38,208,105 million SAR that has a project length of 266 days, which is considered a short project length that is costly compared to other instances.

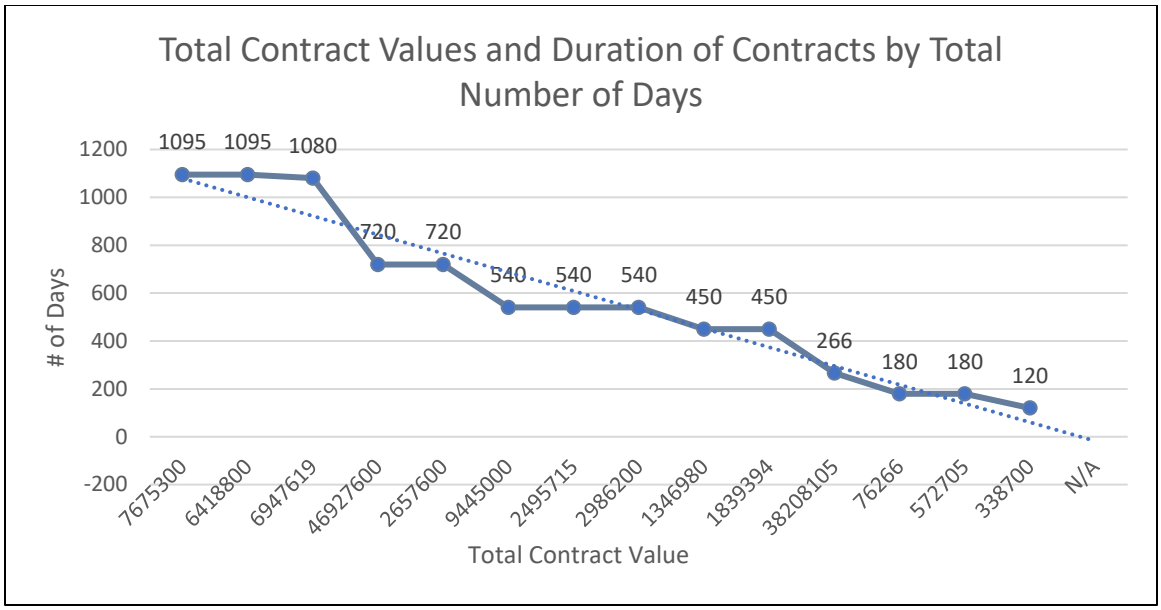


Figure 6: Total Contract Values and Duration of Contracts by Number of Days

Clearly, the Saudi open procurement contracts data as part of the open government data initiatives does not provide any graphical tools, or manuals, or rich and different format of datasets that could facilitate the data analytics attribute. The previous discussion is just for illustration purposes of the potential of presenting the data in a machine-readable format; we have extracted, cleaned, and put in the procurement contracts in a machine-readable format and show the usefulness of such processes.

In Brazil, the open data website (http://www.dados.gov.br/) provides some ready-to-use visualization analysis tools and applications that could help the citizens or any other interested stakeholder to use and analyze the available data. For instance, Figure 7 below shows procurement contracts presented in an easy to interpret point-graph with selected features from the right side, this means that you could choose the features that you want to visualize in an interactive web-based graph. In addition, the data is presented in machine-

readable and ready-to-use rich formats, thus, making it easily analyzable by anyone. Also, each dataset is supported by a description and a series of additional metadata. The datasets also organized such that the data from the same resource or have common metadata, are grouped together to help facilitate searching and understanding of their content.

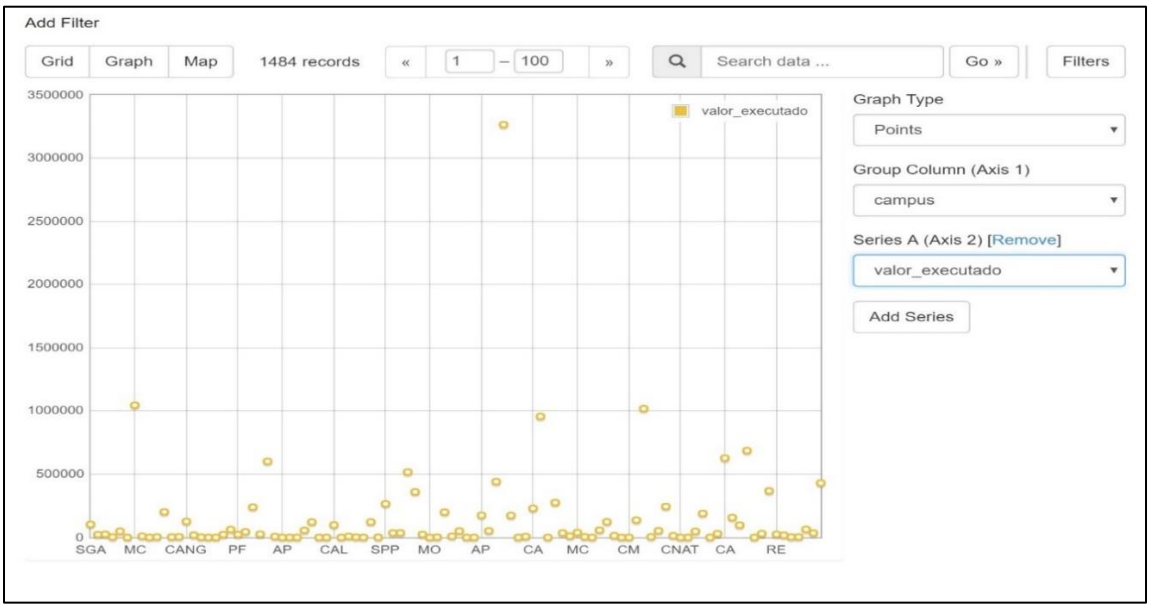


Figure 7: An Illustration of One of The Graphical Tools Available Within the Brazilian OGD Website (source: http://dados.gov.br/dataset/)

2.5.4 APPLICATIONS WITHIN GOVERNMENT WEBSITES

On this final attribute of our 4-condition model for a transparent, open government data, we ask the following questions: Are there applications or unique places within the different open government data portals that allow the citizens to report any problem? Is there a gap between data analytics (third attribute in our model) and applications?

Government 2.0 as defined by Gartner Research (DiMaio, 2009) stated: *"the use of Web 2.0 technologies, both internally and externally, to increase collaboration and transparency and potentially transform the way government agencies relate to citizens and operate."* In Saudi Arabia, there is no evidence of any applications within the different government portals that could be used or facilitate data reporting except the traditional "Contact Us" links. However, there has been an increase in the use of emerging technologies during the past three to four years, especially of social media platforms such as Twitter. Citizens are attempting to use the Twitter platform to interact with different levels of government officials and to report any issues or suggestions as their major communication channel.

Even in Brazil, there is no existence of any type for applications or centralized place by the government that would help the public to report any issues, concerns, or suggestions. However, there is a system to request new datasets called the Electronic Citizen Information Service (e-SIC), but it has nothing to do with reporting.

Clearly, there is a big gap between data analytics and applications for reporting and interacting with public officials. We believe that open government data is not fully open or transparent unless there exist some applications that could help centralize the reporting process by citizens. Therefore, once a centralized place is put in place and accessed by anyone, actionable results have greater potential to occur for better and more transparent governments.

2.6 CONCLUSIONS AND FUTURE RESEARCH

This paper discusses open government data among different countries, focusing on the cases of the Republic of Brazil and Saudi Arabia open government data initiatives. It first compares the different financial reporting and auditing systems in Brazil and Saudi Arabia. Second, the paper gives an overview of OGD initiatives in different countries. In addition, it assesses the level of data transparency based on the definition of open data, and more importantly, this study introduces a new model for effective and efficient open government data by adding new dimensions to the open data concept when utilized by governments which expands the open data definition to include its potential to encourage possible better decision making by governments. The assessment of the proposed attributes for data transparency has been conducted using procurement contracts available at the Republic of Brazil and Saudi Arabia's open data portals. We find that open government data initiatives in Saudi Arabia lack appropriate datasets, formats, analytical tools, and applications, and this could be improved dramatically to enhance government efficiency; whereas the Brazilian OGD initiatives lack applications. We also find that there is a gap between data analytics and applications for reporting and interacting with public officials. Open government data should be analyzable, and the subsequent results should be actionable, for the evolution of better governments.

Saudi Arabia in one side is a developing country but in another, it is developed such that Information and Communication Technologies (ICTs) are highly utilized; but more work is needed on openness and transparency. Brazil, on the other hand, is following and implementing a transparent and open system when it comes to government data. The paper also suggests that public audit (armchair audit) could be of great use in Saudi Arabia where it has not been utilized yet. There exist some limitations such that more detailed theoretical

corpus around the proposed dimensions' proposal is needed. Our long-term goal of this research is to develop a unified framework that applies our proposed 4-condition model for better government data transparency. Also, we need to address how the structure of Brazil and Saudi Arabia's governments, its public sector financial reporting and auditing systems could affect their open data initiative processes. Finally, more data and evidence are needed to support our proposed model and also by assessing other countries' open government data initiatives.

REFERENCES

AlRushaid, M. W., & Saudagar, A. K. J. (2016). Measuring the Data Openness for the Open Data in Saudi Arabia e-Government-A Case Study. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 7(12), 113-122.

Andersen, T. B. (2009). E-Government as an anti-corruption strategy. Information Economics and Policy, 21(3), 201-210.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. Government information quarterly, 27(3), 264-271.

Cameron, D. (2010). PM's Podcast on Transparency. Available at: https://www.gov.uk/government/news/pms-podcast-on-transparency.

Cardoso, Ricardo L. (2017). Public sector financial reporting and auditing in Brazil.

Coglianese, C. (2009). The transparency president? The Obama administration and open government. Governance, 22(4), 529-544.

Dados.gov.br. (n.d.). Retrieved September 24, 2017, from http://dados.gov.br/.

Dai, J., & Li, Q. (2016). Designing Audit Apps for Armchair Auditors to Analyze Government Procurement Contracts. Journal of Emerging Technologies in Accounting, 13(2), 71-88.

Dartdeloittecom (2017). Available at: https://dart.deloitte.com/resource/1/dbbdf810-488c-11e6-8970-3bb2b71b01a5/. Accessed September 27, 2017.

Data.gov., Retrieved February 16, 2019, from https://obamawhitehouse.archives.gov/21stcenturygov /tools/data-gov.

de Aquino, A. C. B., & Batley, R. A. (2015). Accounting and Accountability: The Political Effects of Technical Reforms in Brazil.

DiMaio, A. (2009). Government 2.0: A gartner definition. Retrieved July, 1, 2011.

dos Santos Brito, K., da Silva Costa, M. A., Garcia, V. C., & de Lemos Meira, S. R. (2014, June). Brazilian government open data: implementation, challenges, and potential opportunities. In Proceedings of the 15th Annual International Conference on Digital Government Research (pp. 11-16). ACM.

Esmaeili, H. (2009). On a Slow Boat towards the Rule of Law: The Nature of Law in the Saudi Arabia Legal System. Ariz. J. Int'l & Comp. L., 26, 1.

FLORIAN. BAUER. (2012). Linked open data: the essentials. EDITION MONO.

Gonzalez-Zapata, F., & Heeks, R. (2015). The multiple meanings of open government data: Understanding different stakeholders and their perspectives. Government Information Quarterly, 32(4), 441-452.

Group, O. K. (n.d.). The Open Definition. Retrieved September 24, 2017, from http://opendefinition.org/.

Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. European journal of ePractice, 12(1), 4-16.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.

Knowledge, O. (n.d.). Datasets / Procurement tenders. Retrieved September 05, 2017, from http://global.census.okfn.org/dataset/procurement.

Kozlowski, S. (2016). A Vision of an ENHanced ANalytic Constituent Environment: ENHANCE (Doctoral dissertation, Rutgers University-Graduate School-Newark).

Kucera, J., & Chlapek, D. (2014). Benefits and risks of open government data. Journal of Systems Integration, 5(1), 30-41.

Matheus, R., Ribeiro, M. M., & Vaz, J. C. (2012, October). New perspectives for electronic government in Brazil: the adoption of open government data in national and subnational governments of Brazil. In Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance (pp. 22-29). ACM.

Maude, F. (2012). Open Data White Paper-Unleashing the potential. The Stationary Office Limited on behalf of HM Government, Cabinet Office, London, United Kingdom.

Morrison, J. (2014). Armchair auditing and the great town hall transparency swindle.

Obama White House, Data.gov. Retrieved February 16, 2019, from https://obamawhitehouse.archives.gov/21stcenturygov/tools/data-gov.

Saudi General Auditing Bureau, Gab.gov.sa. Retrieved 27 September 2017, from http://www.gab.gov.sa.

Taylor, M. M., & Buranelli, V. C. (2007). Ending up in pizza: accountability as a problem of institutional arrangement in Brazil. Latin American Politics and Society, 49(1), 59-87.

The White House, Office of the Press Secretary. (2013, May 9). Executive Order -- Making Open and Machine Readable the New Default for Government Information. Retrieved February 17, 2019, from https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government.

The White House. (2009, February 24). MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES. Retrieved February 16, 2019, from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2009/m09-12.pdf.

Verhulst, S., & Young, A. (2017). Open Data in Developing Economies: Toward Building an Evidence Base on What Works and How.

White House. 2017. Open Government Initiative. Available at: http://whitehouse.gov/open/.

Yu, H., & Robinson, D. G. (2011). The new ambiguity of open government. UCLA L. Rev. Discourse, 59, 178.

CHAPTER 3: APPLICATIONS OF DATA ANALYTICS: VISUALIZATION AND CLUSTER ANALYSIS OF GOVERNMENTAL DATA

ABSTRACT: Since data analytics is one way to explore the data and to help uncover hidden relationships, in this study we plan to explore the literature for the use of emerging data mining techniques in auditing. In particular, visualization and unsupervised learning methods such as cluster analysis as supportive tools to help gain more insights into governmental data, thus bring more opportunities for practitioners and auditors. Especially, governmental financial statements and budgeting. We have conducted two case studies; the first one uses the U.S. states financial statements data for the years 2000 to 2016. The selected attributes for our study consist of nine numeric variables and were suggested by government data experts. From the original data tables, we obtained per capita for each variable for examination. The second case study utilizes the Volcker Alliance's Survey data results. The survey produces extensive information about how the different U.S. states scores on an annual basis on budgeting using five measures or variables. On both case studies, we demonstrate how visualization and data analytics especially cluster analysis could be used on governmental data and to help gain more insights about financial statements and budgeting. Our main objective of using cluster analysis is for grouping and ranking the states. The statistical open-source software R is used in the analysis of both studies. To evaluate our results, we use clustering assessment methodology that uses bootstrap resampling which assesses how stable a given cluster is. This methodology is

important when using clustering algorithms such as k-means and hierarchical clustering, especially the latter, where most of the time you have to use an a priori method of choosing the optimal number of clusters. The method was first introduced by Christian Hennig in 2007 (Hennig, 2007). This study contributes to the literature in two ways. First, the two case studies bring some advanced visualization and data mining techniques into the governmental domain, especially using financial statements and budgetary data. Second, the visualization and clustering results bring more opportunities for auditors and practitioners and to help get more insights into the data.

3.1 INTRODUCTION AND BACKGROUND

Data mining is the process of gaining insights and identifying interesting patterns and trends from data stored in large databases in such a way that the insights, patterns, and trends are previously unknown, statistically reliable, and actionable, meaning that some decisions could be taken to exploit the knowledge (Elkan, 2001). Data mining is also defined as *"a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gaining knowledge from a large database,"* (Turban, Sharda, and Delen, 2011, p 21).

Cluster analysis as a data mining approach can help find similar objects in data. Kaufman & Rousseeuw (2009) have defined cluster analysis as *"the art of finding groups in data."* To illustrate, Figure 8 below shows eight objects and two variables. The X-axis represents the weight, and the Y-axis represents the height. Clearly, two distinct groups of objects could be found which are the following: {LEO, PET, LIE, JAC}, and {KIM, ILA,

TAL, TIN}. These two groups are called clusters and the way to discover them is called cluster analysis. Objects in the same group are similar to each other, and objects in different groups are as dissimilar as possible (Kaufman & Rousseeuw, 2009).

Cluster analysis is a type of unsupervised learning where there are no predefined classes required. Cluster analysis could be used alone to get insight into data distribution, or it could be used as a preprocessing step where the results of the clustering algorithm used as an input for other algorithms.

Cluster analysis is also considered as an Exploratory Data Analysis (EDA) tool which is used for pattern recognition and knowledge discovery. EDA is well-established statistical techniques, mostly graphical that can assist in uncovering underlying structure, explore important variables, test hypothesis, detect anomalies and outliers, and many more. Eventually, the discovered insights from data could be used to support various decision making, test and evaluate current and past decisions, or just to conduct different experiments on a specific domain (Hossain, 2012).

Classifying similar objects is an essential human activity; it is part of our everyday life. The daily process of learning to distinguish between cats and dogs, cars and motorcycles, green and yellow colors, since we were kids, improve our subconscious classification schemes. This is one of the reasons that cluster analysis is considered as part of artificial intelligence and pattern recognition (Kaufman & Rousseeuw, 2009).

Clustering algorithms usually work with two different input structures. The first structure represents the objects *n* based on their *m* attributes, represented as an *n-by-m* matrix, where the rows correspond to the objects or points, and the columns correspond to

the attributes or features (Kaufman & Rousseeuw, 2009). The second structure could be represented by proximities between objects or *n*-by-*n*, which is the distances between every two objects whether using similarity (how close two objects to each other) or dissimilarity (how far two objects from each other) measures.

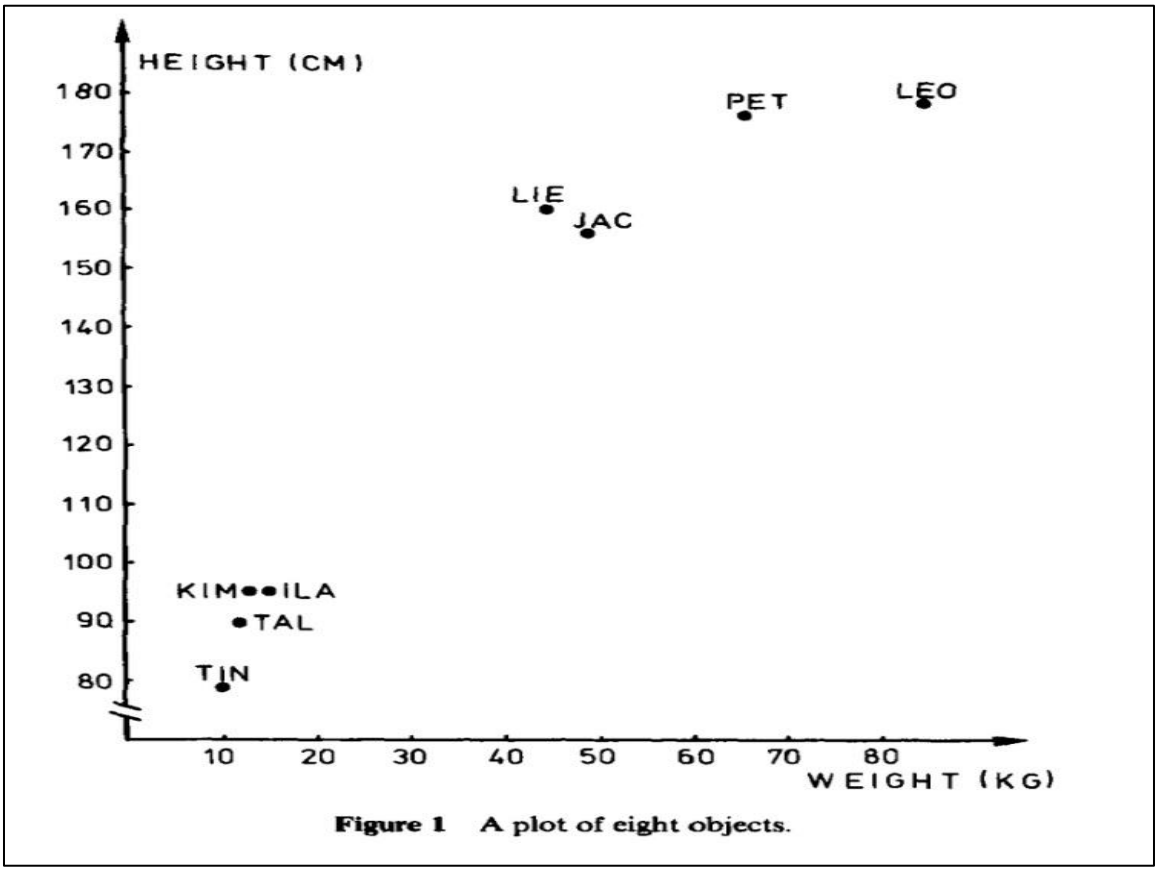


Figure 8: Cluster Analysis (source: Kaufman & Rousseeuw, 2009)

3.2 MOTIVATION: LITERATURE REVIEW

There exist many studies in different domains discussing cluster analysis. In accounting research, many studies are addressing the possibility of using unsupervised learning or cluster-based analysis to identify some interesting patterns. In risk analysis, for

instance, cluster analysis is applied to identify groups of accounts, or customers who encounter unusual behavior. Williams and Huang (1997) combine k-means clustering and the decision tree (i.e., C4.5) supervised algorithm for insurance risk analysis in order to find customer groups that highly affect the insurance portfolio. They use an approach named the hot spot methodology where they use k-means clustering to group the data and then apply the decision tree algorithm to get sets of rules, and then evaluate the results in order to identify the high risks patients. The Australian government's Medicare healthcare datasets are used. The datasets maintain all payment information made to doctors and patients. This method can help better facilitate the exploration of different key groups or clusters to reassess the insurance premiums for different patients.

In anomaly detection, Thiprungsri & Vasarhelyi (2011) use cluster analysis to detect anomalies. The purpose of their paper is to examine the use of cluster analysis techniques to automate fraud filtering during the auditing process and to support auditors when evaluating group life insurance claims. By using a simple k-means clustering approach, they are able to detect possible fraudulent claims resulted from the anomalous clusters. Thus, facilitating the investigation processes for internal auditors.

Cluster analysis is also widely used in many other domains such as market research, particularly in market segmentation (Thiprungsri & Vasarhelyi, 2011). For instance, divide customers into segments or clusters based on demographic information; therefore different clusters could be targeted with specific services. Kim and Ahn (2008) apply a hybrid k-means clustering algorithm to segment an online shopping market for customers. Specifically, a Korean diet shopping mall site. They use the personal information of their

customers to direct-market their products by sending recommendations of new and existing products in their online shopping mall to their customers.

Cluster analysis is also used in fraud detection. Outlier detection is the normal technique used to detect fraud. For example, fraud in banking could be categorized into two categories: behavioral and applicational fraud (Bolton & Hand, 2001). Behavioral fraud is if you have your credit card data, for instance, compromised. On the other hand, applicational fraud which is when customers apply for new credit cards. So, clustering methodologies could identify high-risk customers from others.

Srivastava, Kundu, Sural, and Majumdar (2008) apply a k-means clustering algorithm to identify fraudulent credit card transactions using the amount of a transaction attribute. Then they categorize customers or profiles based on their spending. For instance, high spenders will be in the high-spending group, low-spenders will be assigned to the low-spending groups, and medium-spenders will be assigned to the medium-spending group. Thus, by using this profiling approach, they are able to identify possible fraudulent transactions.

Cluster analysis is also used in detecting fraud in telecommunication domain. Hilas and Mastorocostas (2008) use both supervised and unsupervised machine learning approaches to investigate the usefulness of applying those methods in telecommunication's fraud detection. They use cluster analysis for their unsupervised learning approach; in particular, they use the hierarchical agglomerative clustering technique. Five profiles of telecommunication users' behavior have been identified and used as a user classification method to discriminate between fraudulent and legitimate usage in the telecommunications environment. The method successfully groups a distinct cluster of outliers.

In healthcare domain, cluster analysis is applied to identify or group patients with similar types of cancer (Thiprungsri & Vasarhelyi, 2011). It is also being used to identify adverse drug reactions (ADRs). Harpaz et al. (2011) present a new pharmacovigilance data mining method based on the Biclustering model, which aims to identify drug groups that share a common set of ADRs. Biclustering technique differs from the typical clustering methodology in which clustering can be applied to only rows or columns of a matrix separately; while Biclustering performs clustering in both rows and columns simultaneously. In their paper, the authors are using the spontaneous reporting system (SRS) taken from the U.S. food and drug administration (FDA), specifically the Adverse Event Reporting System database (AERS) for the year of 2008. The performed statistical tests show that it is unlikely that the Biclusters' structure, their contents, could have arisen by chance. The experimental results show that many cases for drugs on the same class were associated with a specific adverse reaction; this suggests that any member of the class could result in similar adverse events.

Clustering is mainly used for data exploration and pattern recognition using visualizations and other techniques. As the size of a data set increases, more methods and tools are required to facilitate interpretation and understanding the data. Cluster analysis is one way that helps understand the data structure of large datasets. Fua, Ward, and Rundensteiner (1999) apply hierarchical clustering methodology for the visualization of large multivariate datasets. Their research mainly focusses on the interactive visualization of large multidimensional datasets using a parallel display technique. They propose a hierarchical clustering approach using a software called XmdvTool[1] that displays

[1] http://davis.wpi.edu/xmdv/

multiresolution views of the data with the ability to navigate and filter the large data sets to facilitate the discovery of hidden trends and patterns.

Eisen, Spellman, and Brown (1998) use cluster analysis for genome pattern exploration. They utilize a hierarchical clustering method in order to support biologists in the study of gene expression datasets. Pairwise average linkage is used to calculate the distances between observations. They use the hierarchical tree of genes to represent different degrees of similarities and relationships between genes. Thus, ordering the different genes so that the genes with similar expressions are adjacent to each other. The ordered data is then displayed graphically. The quantitative values represented using a naturalistic color scale rather than numbers. They find that this alternative way of representing a large volume of quantitative data is easier for brains to interpret and capture.

Advanced data analytics techniques bring more challenges to researchers and practitioners in the field of accounting and especially auditing. Challenges such as the ability to adapt to new and sophisticated data analytics techniques. Another challenge would be how to decide which data analytics technique is suitable to be used on a certain task. At the same time, the modern data analytics techniques provide more opportunities for researchers and practitioners in the accounting domain to be explored and utilized (Schneider, Dai, Janvrin, & Ajayi, 2015; Liu & Vasarhelyi, 2014).

Data analytics tools along with the digitization and dissemination of governmental data (e.g., financial statements) to the public, allow auditors and any other interested party to monitor and analyze government performance (Liu & Vasarhelyi, 2014). Various data analytics techniques such as clustering methods can be applied for exploratory purposes,

thus help auditors get more insights into the data and automate the discovery of previously unknown patterns and trends (Liu & Vasarhelyi, 2014).

However, few studies have been published in the extent of using sophisticated data analytics techniques in the government domain using government-related financial statements data and other budgeting related information. Bloch and Fall (2015) discuss government debt, finances, and sustainability across countries. They also assess factors such as financial liabilities, assets, pension liabilities, national accounts, non-financial assets and debt composition for analyzing general government debt.

Arisandi (2016) discusses applying an exploratory data analytics (EDA) technique such as data clustering using the U.S. government and non-profit sector data. Ten years period of data is used (2004-2013) of the comprehensive annual financial report (CAFR)[2] of the fifty-one states' level data in the United States and a non-financial data such as crime statistics, transparency awards data, leadership change and public employee corruption convection from many publicly available sources. Ratio analysis and k-means cluster analysis are applied. The clusters place the states with similar characteristics based on the selected variables together into groups. The author is able to detect trends and potential anomalies by using cluster analysis; thus, applying such techniques could support users understanding of information.

In these two case studies, we examine the usability of applying modern data analytics techniques, especially data clustering and visualization to get more insights into

[2] A Comprehensive Annual Financial Report (CAFR): is a set of financial statements for a government body such as a state, municipality, or any other government entity that comply with the accounting the requirements of the Government Accounting Standards (GASB). It also should be audited by an independent auditor using generally accepted government auditing standards (GASB, 2010; Hegar, 2019).

data and to extract meaningful knowledge that can assess a broad range of government officials especially to auditors for better decisions.

3.3 CLUSTER ANALYSIS OVERVIEW

Excessive amounts of data are generated and collected every day from various domains. From satellite images, biomedical, businesses, marketing, security, internet searches, and more importantly government-related data. Extracting knowledge from these "big" data exceeds human abilities. Cluster analysis is one way of mining these data and is one of the important data mining techniques for discovering previously unknown patterns and trends (Kassambara, 2017). In the literature, it is often referred to as "unsupervised learning" because we do not have an a priori knowledge of the data and in which variables belong in which cluster and the algorithm will learn how to cluster the data (Kassambara, 2017).

Since our focus on these two case studies is in visualization and cluster analysis techniques using k-means and hierarchical clustering, the following two subsections attempt to give an overview of the two clustering techniques used in this study.

3.3.1 K-MEANS CLUSTERING

K-means clustering algorithm is a popular clustering method due to its simplicity and efficiency. It uses unsupervised learning and is one of the most common data mining methods that is used in data clustering (grouping) and pattern recognition. K-means was first introduced by J. MacQueen (MacQueen, 1967). It partitions a given dataset (i.e.,

unlabeled) into a set of *k* clusters (groups), where *k* denotes the number of groups pre-selected by the analyst. It uses centroids to form clusters by optimizing the within clusters' squared errors. Therefore, it automatically groups a dataset into *k* partitions known as clusters. K-means clustering algorithm divides *m* points in *n* dimensions into *k* clusters so that the within-cluster sum of squares is minimized, meaning that each observation (a record in a table) belongs to the cluster with the nearest mean. Therefore, objects within the same cluster are as similar to each other as possible (i.e., intra-class similarity), whereas objects in different clusters are as dissimilar as possible (i.e., inter-class similarity) from other clusters (Hartigan & Wong, 1979; Jain, 2010; Kassambara, 2017).

In k-means clustering algorithm, each cluster is formed by its center "centroid" which represents the mean of the total points assigned to the cluster. K-means proceeds by manually choosing the number of clusters (*k* initial clusters) and then iteratively apply the following steps (Wagstaff, Cardie, Rogers, & Schrödl, 2001; Kassambara, 2017):

I. Specify the number of clusters "*k*".
II. Randomly select *k* objects from the dataset as the initial cluster centers.
III. Scan through the list of *m* observations, then assign each observation to its nearest cluster's center using the Euclidean distance.
IV. Each cluster's center is then updated to be the average of the new observations assigned.
V. Repeat Steps III and IV iteratively until there are no more reassignments or the maximum iterations number is reached.

The fundamental idea of the k-means clustering is forming clusters of data so that the total intra-class variation is as low as possible. There exist many k-means algorithms;

the standard one is the Hartigan-Wong algorithm which uses the Euclidean distance (Kassambara, 2017). The main advantages of k-means algorithm are its speed and simplicity; however, there exist some disadvantages such as that the algorithm does not yield the same results with each run since *k* is assigned randomly at the beginning (Singh, Malik, & Sharma, 2011).

3.3.2 HIERARCHICAL CLUSTERING

Hierarchical clustering or hierarchical cluster analysis (HCA) is an unsupervised data mining technique which aims to build clusters hierarchy (Hierarchical clustering, 2017). Hierarchical clustering results in a tree called dendrogram. It is of great interest in a broad range of domains. Hierarchical trees provide views of the data at different abstraction levels, therefore making it ideal for interactive visualization and exploration (Zhao, Karypis, & Fayyad, 2005). In hierarchal clustering, the data are not partitioned into clusters in one step, it takes multiple partitioning steps which may move from all objects in one cluster to *k* clusters each containing one object. Hierarchical clustering technique is appropriate for tables with a few hundred records. Also, you can choose the desired number of clusters after the tree is built by cutting it at a certain height (Bobric, Cartina, & Grigoras, 2009).

Hierarchical clustering is subdivided into two methods: agglomerative and divisive. On the one hand, the agglomerative method (or bottom-up) were each observation forms its own cluster; each pair of clusters are then merged as the hierarchy moves up. On the other hand, divisive method (or top-down) is the opposite of agglomerative clustering were

all observations form one big cluster; then they split recursively as one moves down the hierarchy (Hierarchical clustering, 2017).

The merging in the agglomerative hierarchical clustering method uses heuristic criteria, e.g., sum of squares, single linkage (i.e., nearest neighbor), complete linkage (i.e., farthest neighbor), Ward's method (i.e., decreases in variance between the two clusters that are being merged) or average linkage (i.e., average distances between objects). The splitting in the divisive clustering methods uses the following heuristic. It starts with all objects as one cluster of size *m*. Then, at each step, clusters are partitioned into pairs of clusters with maximizing the distance between the two clusters. Finally, the algorithm stops when objects are partitioned into *m* clusters of size 1. The hierarchical clustering resulted by either agglomerative or divisive is usually presented in a two-dimensional diagram called a dendrogram (Everitt, Landau, Leese & Stahl, 2011).

3.4 Data Analytics and Visualization of States Financial Statements: A Case Study

3.4.1 INTRODUCTION

In this case study, we use unsupervised learning methods, in particular, cluster analysis. We investigated two clustering techniques: k-means and hierarchical clustering methodologies. The statistical open-source software R is used in the analysis. The data used is the U.S. states financial statements for the years 2000-2016. The selected variables for examination consisted of nine numeric variables and were suggested by government data experts. From the original data table, we obtained per capita for each variable. The

nine variables are total general fund revenues, excess (deficiency) of revenues over expenditures, total operating expenses, education expenses, the net change in fund balance, general fund of total other financing sources, general fund transfers to other funds, general fund transfers from other funds and pension expense.

Cluster analysis is used in this study as a supportive method to get more insight into government data, specifically U.S. States' financial statements data. Our main objective of using cluster analysis is for grouping and ranking of states. Unlike linear discriminant analysis (LDA), cluster analysis does not require prior knowledge of data (Roy, Kar, & Das, 2015). Two different clustering techniques were used, k-means and hierarchical clustering. Finally, we compare the results of the clusters and how they are different from each other.

In this study, we used government data, particularly, states' financial statements for clustering. The dataset contains nine numeric variables. These variables are based on the average per capita for the years (2000-2016) and are consist of the following:

1. Per capita total general fund revenues.
2. Per capita excess (deficiency) of revenues over expenditures.
3. Per capita total operating expenses.
4. Per capita education expenses.
5. Per capita net change in fund balance.
6. Per capita general fund total other financing sources.
7. Per capita general fund transfers to other funds.
8. Per capita general fund transfers from other funds.
9. Per capita pension expense.

The following figure shows a two-dimensional biplot. This shows how the original variables behave relative to our principal components (only in the directions of the first two principal components).

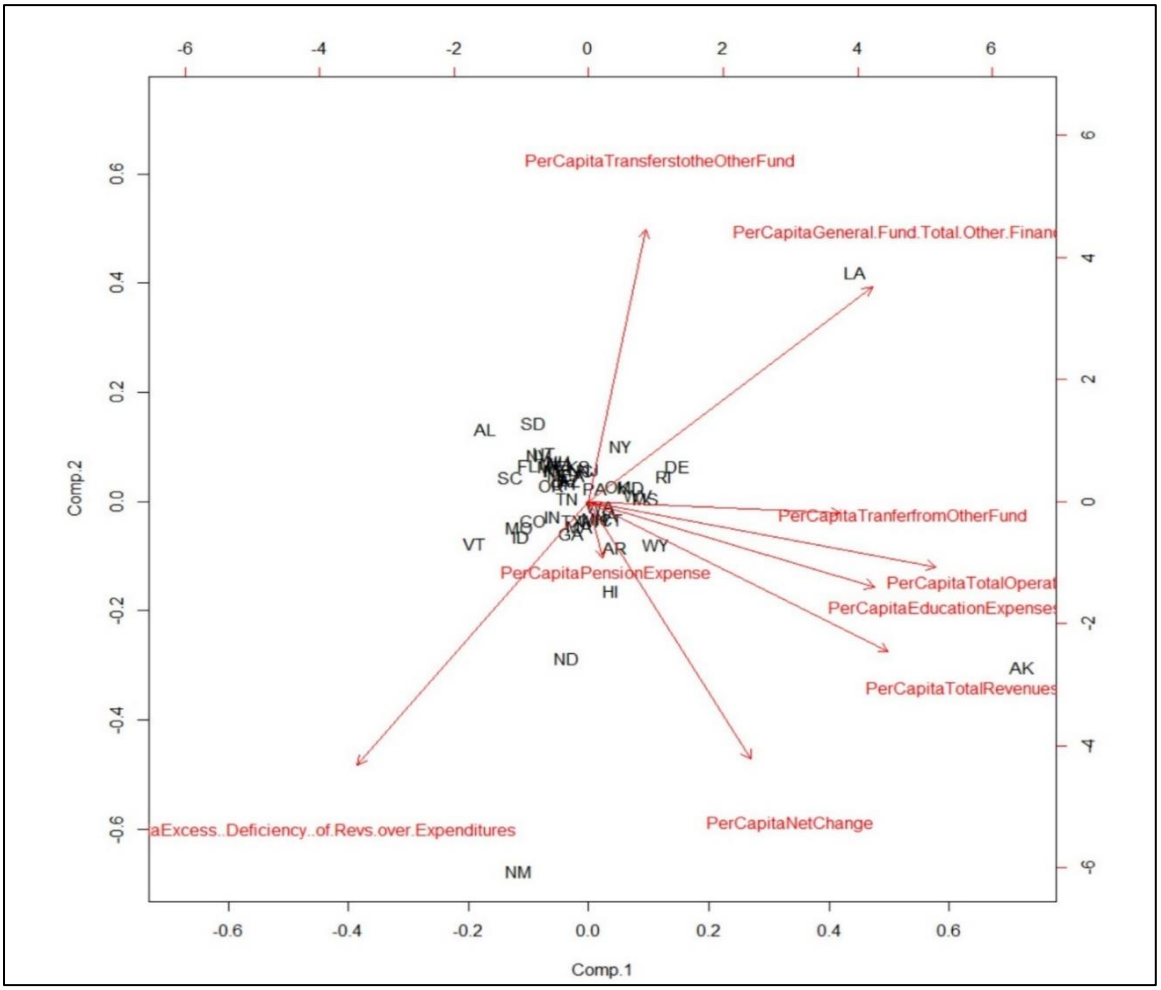


Figure 9: A Two-dimensional Biplot of the Variables

3.4.2 K-MEANS CLUSTERING

In using the k-means clustering algorithm, we must first initiate the number of clusters that is *k*. One of the fundamental problems when dealing with k-means is to

determine the number of clusters. Choosing the right *k* value has a significant impact on the performance of the algorithm; and the question is which *k* should be selected when there is no prior knowledge on the data (Bholowalia & Kumar, 2014). There exist many methods that help determine a "good" number of clusters to proceed with. Some widely used methods are the elbow, silhouette and gap statistics (Kassambara, 2017). Since this is beyond the scope of our work, we decide to choose a direct and widely used method which is the elbow method.

The elbow method is a way of interpretation and validation of consistency within clusters by computing the within clusters sum of squares that is designed to help finding the appropriate number of clusters in a dataset. It is one of the oldest methods used for determining a "true" number of clusters. It looks at the percentage of the explained variance as a function of the number of clusters *k*. It basically comes from the idea that when adding additional clusters, the modeling of the data will not improve. Therefore, the percentage of the explained variance by the clusters is plotted against the total number of clusters. Thus, the first clusters will add the most information or marginal gain, and at a certain point, the gain will drop significantly. This significant drop means that adding an additional cluster would not gain incremental meaningful insights. If one were to graph these findings, the optimal point resides in the curve of a bent arm, the elbow crease (Bholowalia & Kumar, 2014; Kodinariya & Makwana, 2013).

The true *k* "number of clusters" is chosen at the elbow crease point or the "elbow criterion". The idea is to start with k = 2, and then keep increasing the number by 1, calculating your clusters, and at a certain value for *k*, the cost will drop significantly, hence reaching a plateau when you increase *k* further. This is the "right" *k*-value you want to

proceed with. The reason is that after this, you increment the clusters, but the new clusters are near the existing ones (Bholowalia & Kumar, 2014). To summarize, the elbow methods algorithm depends on the following: it first initializes k = 1, and then measure the cost of the optimal quality solution, and if at a certain point, the cost of the solution drops significantly then that is the "right" *k* value.

Figure 10 below shows a scree plot where the x-axis represents the number of clusters, and the y-axis shows the sum of squared distances of the clusters' means to the global mean (Bholowalia & Kumar, 2014). The figure below shows that six clusters would be a good fit.

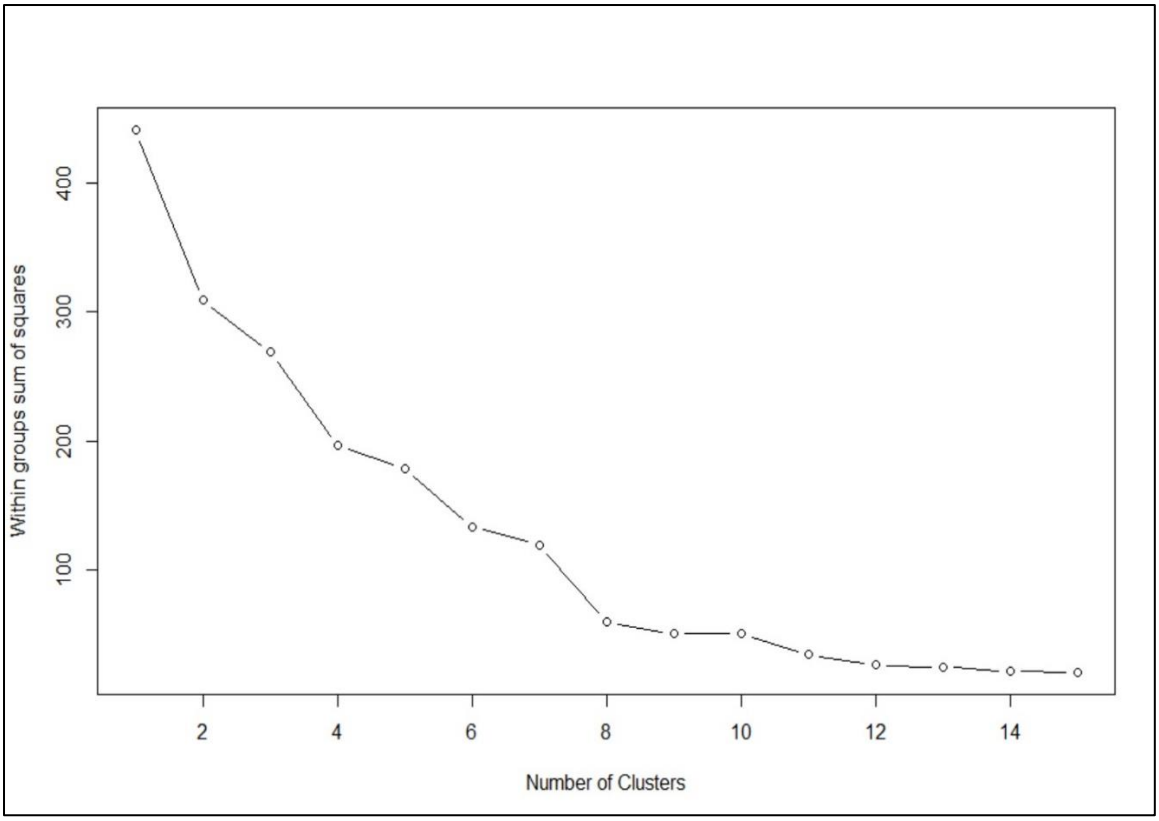


Figure 10: A Scree Plot that Shows the Number of Clusters and Within Group Sum of Squares (Elbow Method)

The next plot shows a two-dimensional representation of the clusters for the average per capita scores using six clusters. We use the first-two principal components (PCs), and they account for 67.15% of the data variability. The 1st principal component explains around 43% of the overall variability; the 2nd principal component accounts for 24% of the data variability.

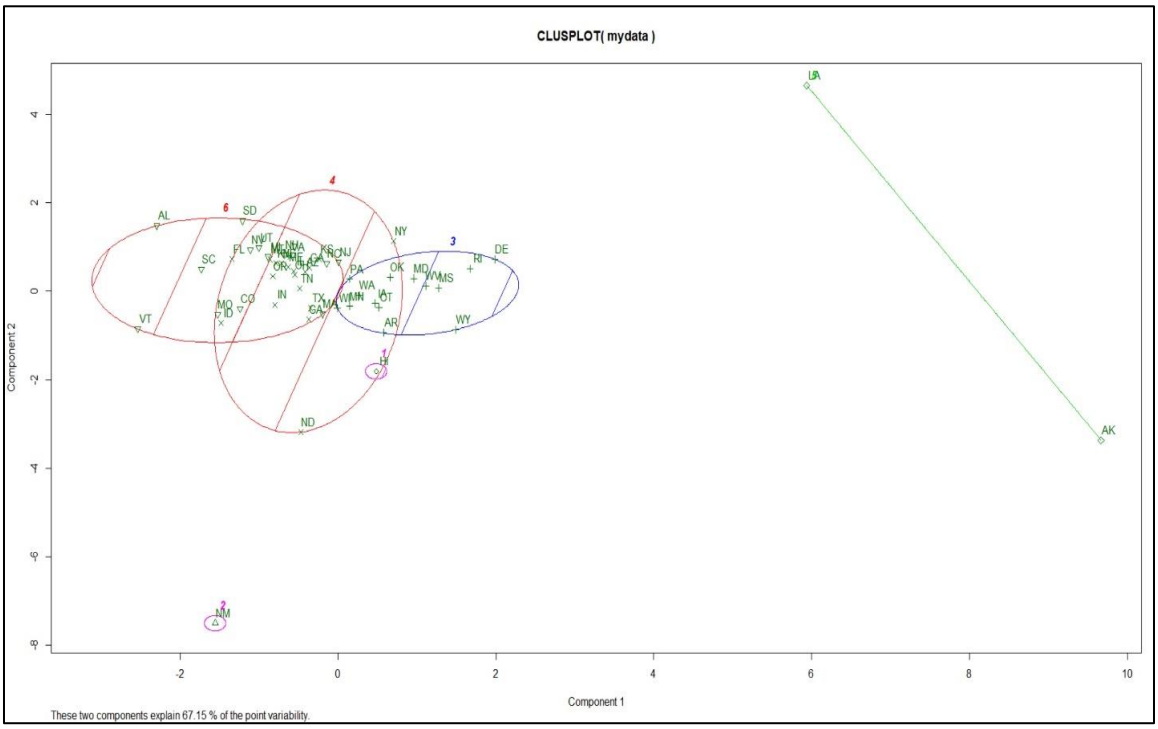


Figure 11: Two-dimensional Representation of K-means Results - Six Clusters

As shown in Figure 11 above, the states on each cluster are more similar to the ones in their group than they are to the others. The table below shows the states that belong to a certain cluster:

Cluster	Members
#1	HI
#2	NM
#3	AR, CT, DE, IA, MD, MN, MS, OK, PA, RI, WA, WV, WI, WY

#4	AZ, CA, FL, GA, ID, IL, IN, KS, KY, ME, MT, NE, NY, ND, OH, OR, TN, TX, VA
#5	AK, LA
#6	AL, CO, MA, MI, MO, NV, NH, NJ, NC, SC, SD, UT, VT

Table 7: U.S. States Distribution on Each Cluster

3.4.3 HIERARCHICAL CLUSTERING

Here, we use the agglomerative hierarchical clustering method, in particular, Ward.D, with Euclidean as a measure of distance. Ward.D is a method that finds a minimum variance by minimizing the total within-cluster variance (Everitt et al., 2011). The following dendrogram shows the clusters' hierarchy.

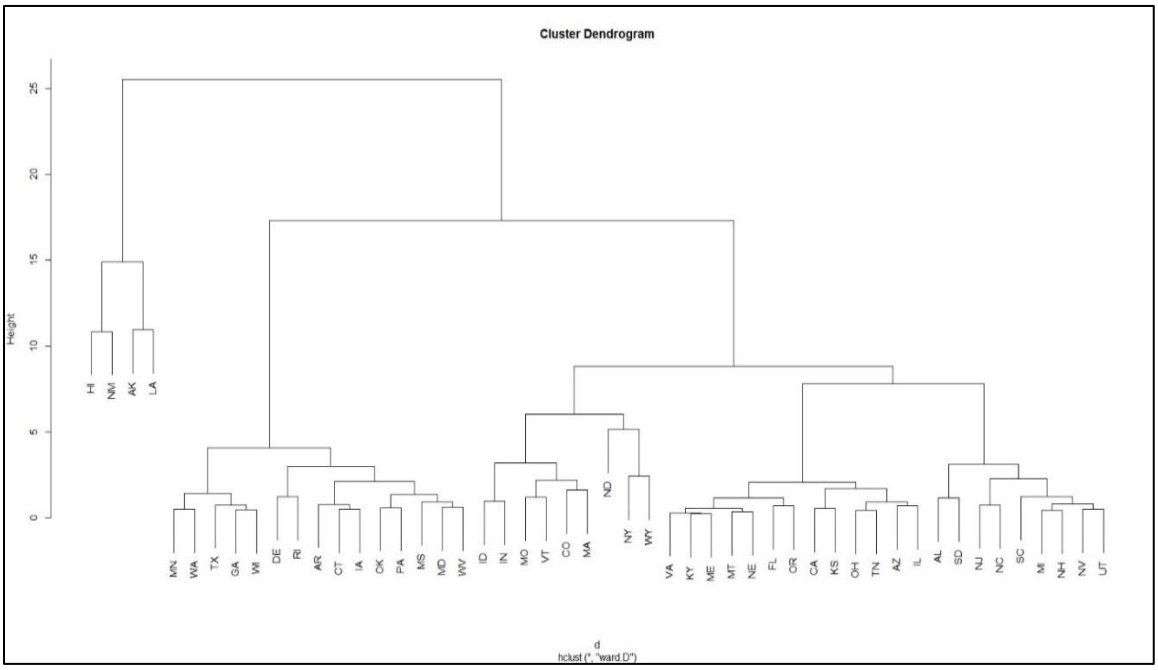


Figure 12: A Dendrogram Representation of the Hierarchy

Furthermore, Figure 13 below shows four clusters, and the red lines separate the four clusters.

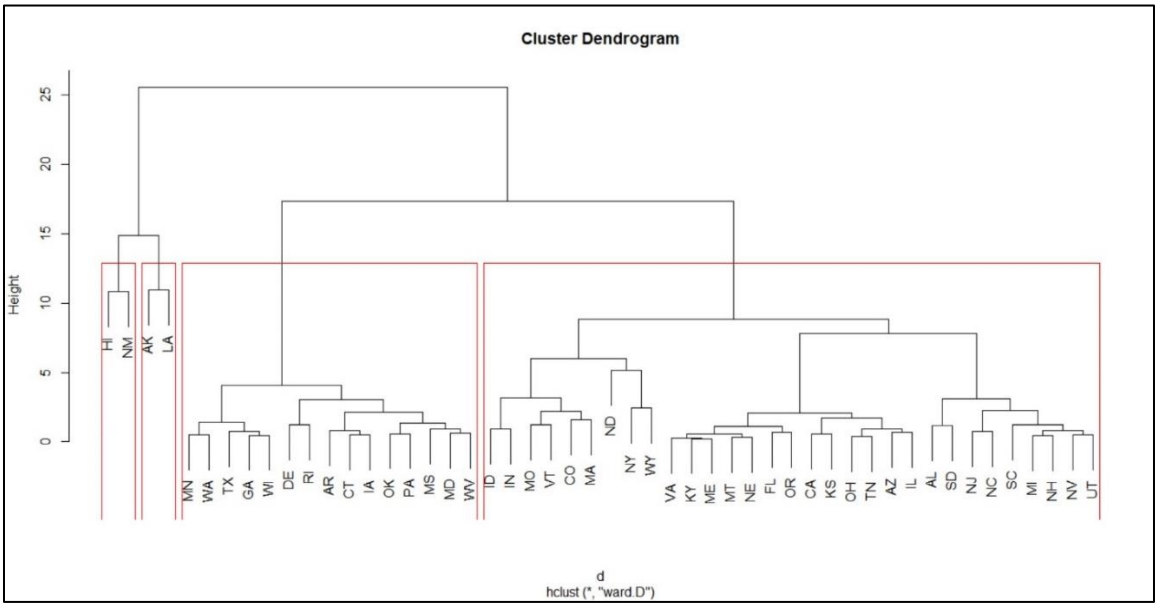


Figure 13: A Dendrogram Representation of the Hierarchy - Four Clusters

The table below shows the states that belong to a specific cluster.

Cluster	Members
#1	HI, NM
#2	AK, LA
#3	MN, WA, TX, GA, WI, DE, RI, AR, CT, IA, OK, PA, MS, MD, WV
#4	ID, IN, MO, VT, CO, MA, ND, NY, WY, VA, KY, ME, MT, NE, FL, OR, CA, KS, OH, TN, AZ, IL, AL, SD, NJ, NC, SC, MI, NH, NV, UT

Table 8: U.S. States Distribution on Each Cluster

3.4.4 DISCUSSION

In this case study, we experiment using the U.S. states financial statements for the years 2000 through 2016. Visualization such as the two-dimensional biplot of the variables shows us how the original variables behave relative to our first-two principal components, thus exploring those variables in length could help us examine how each feature could affect external variables such as credit ratings of the states (e.g., Moody's Ratings), municipal bonds, or net population change in an area. Then we use unsupervised learning approach such as k-means and hierarchical clustering methods on our data to further investigate and get more insights about the states and compare, for instance, the states in the same group (i.e., cluster) with others.

The states that form a cluster are more similar to each other based on the examined variables. By comparing the two unsupervised methodologies, k-means and hierarchical, we can confirm the quality of the clusters. For instance, we have Alaska and Louisiana in the same cluster in both methods, thus confirming that the two states have similarities and thus gives us an idea that additional investigation should be conducted in order to get more insights into the states. In addition, the third clusters in both the hierarchical and the k-means methods are almost the same except for four states.

This case study explores the potential for applying some data mining methodologies to get more insights into governmental data especially states financial statements. It appears to be one of the few studies that use visualization and cluster analysis in the governmental domain. Those methods could bring more opportunities for auditors and practitioners or any interested government official and help get more insights into data. Cluster analysis is mainly being used for ranking and grouping the states. Further analysis is required for each cluster's results; also, by comparing and relating states' financial statements with external

variables. Since we have almost use the same methodologies in these two case studies, we will get into more details in the second case study.

3.5 Data Analytics and Visualization of States Budgeting: A Case Study

3.5.1 INTRODUCTION

In this case study, we utilize the Volcker Alliance's[3] survey data results. The survey produces extensive information about how the U.S. states score on an annual basis on budgeting using five measures: Budget Forecasting, Budget Maneuvers, Legacy Costs, Reserve Funds, and Transparency. However, readers may have difficulties to fully understand such rich data by just looking at the tables. Data analytics is one way to explore the data and to help uncover hidden relationships. We explored data visualization and data mining techniques as part of our experiment. This case study contributes to the literature by showing how certain data mining techniques could be of good use when applied to the government domain.

First, we show how applying variables correlation coefficient among the five variables could draw some patterns between two variables. Also, we utilize Tableau

[3] Volcker Alliance was launched in 2013 and it is a non-partisan Alliance that partner with organizations such as academic, public and private sectors that conducts research on governments with the objectives to improve overall governments' performance and increase accountability and transparency at the federal, state and local levels (https://www.volckeralliance.org).

software[4] for data visualization part where we use the map feature to illustrate Moody's[5] different ratings for the States. We also show a map view of the Volcker Alliance's scores. Next, we apply cluster analysis techniques on The Volcker Alliance's survey scores. That is, how the states are similar to one another regarding their budgetary practices, and may we find previously unknown relationships and patterns with non-judgmental cluster analysis.

3.5.2 ABOUT THE DATASET

In 2016, The Volcker Alliance began a series of studies exploring budgetary and financial reporting practices of all fifty states in the U.S. The first report study is called Truth and Integrity in State Budgeting: What is the Reality? And it was first published in November 2017 (The Volcker Alliance, 2017). It proposes sets of best practices for policymakers. The goal of the published study is to contribute to the main mission of the Volcker Alliance to improve the effectiveness of different government administrations at all levels and to make processes such as state budgeting more transparent to various stakeholders especially citizens. Following their first published report regarding budgetary and financial reporting practices, they have published two more reports in the following year. The first one called Truth and Integrity in state Budgeting: Preventing the Next Fiscal

[4] Tableau is an interactive data visualization tool that help people discover and understand their data better (https://www.tableau.com).
[5] Moody's provides credit ratings and research on area such as debt instruments and securities as well as analytics in financial risk management (https://www.moodys.com/).

Crisis, and the second report is Truth and Integrity in State Budgeting: What is the Reality? Fifty State Report Cards (The Volcker Alliance (1), 2018; The Volcker Alliance (2), 2018).

The Volcker's reports contain grades ranging from (A to D-minus) related to the state's budgetary practices for the fiscal years of 2015 through 2017. Each one of the states receives a grade in five main categories; a grade is given based on the state's commitment to best practices in several budgeting indicators. Each grade reflects the average of the three years for the state's budgeting area. The five categories are as follow (The Volcker Alliance, 2017):

a. Budget forecasting: how states estimate expenditures and revenues for the upcoming fiscal year and for the long term;
b. Budget maneuvers: how much the states make a one-time action to make up recurring expenditures;
c. Legacy costs: measures how well the states are funding commitments to the public employees to cover costs such as retirement including pensions and healthcare;
d. Reserve funds: measures how well the states manage both the general fund reserves and the rainy-day funds;
e. Budget transparency: measures how well the states are disclosing budget information, including tax expenditures, debts, and the estimate costs of deferment infrastructure maintenance.

The grades are derived from the survey that the Volcker's made for each of the five critical categories for the states. They have assigned a value of zero or one for each indicator of the five categories. For instance, in the budget forecasting category, they measure the following indicators: revenue growth projections, multiyear revenue forecasts,

multiyear expenditure forecasts, and consensus revenue forecasts. For budget maneuvers, they measure indicators such as deferring recurring expenditures, revenue and cost shifting, funding recurring expenditures with debt and using asset sales and up-front revenues. For legacy costs, they measure the following: public employee OPEB funding and public employee pension funding. For reserve funds, they measure the following indicators: positive reserve or general fund balance, reserve funds disbursement policy and reserves tied to revenue volatility. And finally, for budget transparency, they measure the following indicators: consolidated budget website, provided debt tables, disclosed deferral infrastructure replacement costs and disclosed tax expenditures.

The datasets used are publicly available and freely accessed through the Data Lab of The Volcker Alliance's website[6]. The Alliance also provides online tools for data exploratory with interactive visualizations.

3.5.3 DATA VISUALIZATION

First, we use correlation coefficient analysis which measures the strength of association or relationship between two variables (McGraw & Wong, 1996). This analysis could assist in understanding and validating the survey design as well as establishing hints for future budget behaviors. Correlation analysis of the five variables shows a moderate positive correlation between legacy costs and budget maneuvers at ~0.512 which indicates

[6] The Volcker Alliance data can be found at: https://www.volckeralliance.org. Publicly available and free of cost.

that the scores in these two categories are interrelated with each other as shown in Figures 14 and 15 below.

	Budget.Forecasting	Budget.Maneuvers	Legacy.Costs	Reserve.Funds	Transparency
Budget.Forecasting	1.000000000	-0.007919089	-0.03613848	0.25110021	0.18377649
Budget.Maneuvers	-0.007919089	1.000000000	0.51272449	0.22466741	-0.11578494
Legacy.Costs	-0.036138475	0.512724489	1.00000000	0.02784838	0.04485754
Reserve.Funds	0.251100213	0.224667415	0.02784838	1.00000000	0.09371242
Transparency	0.183776490	-0.115784941	0.04485754	0.09371242	1.00000000

Figure 14: Variables Correlation Coefficient

This variables correlation coefficient (Figure 14) and the pairwise scatterplot matrix (Figure 15) analysis could assist in understanding and validating the survey design as well as establishing hints of future budget behaviors, also could assist in selecting appropriate variables to build future models.

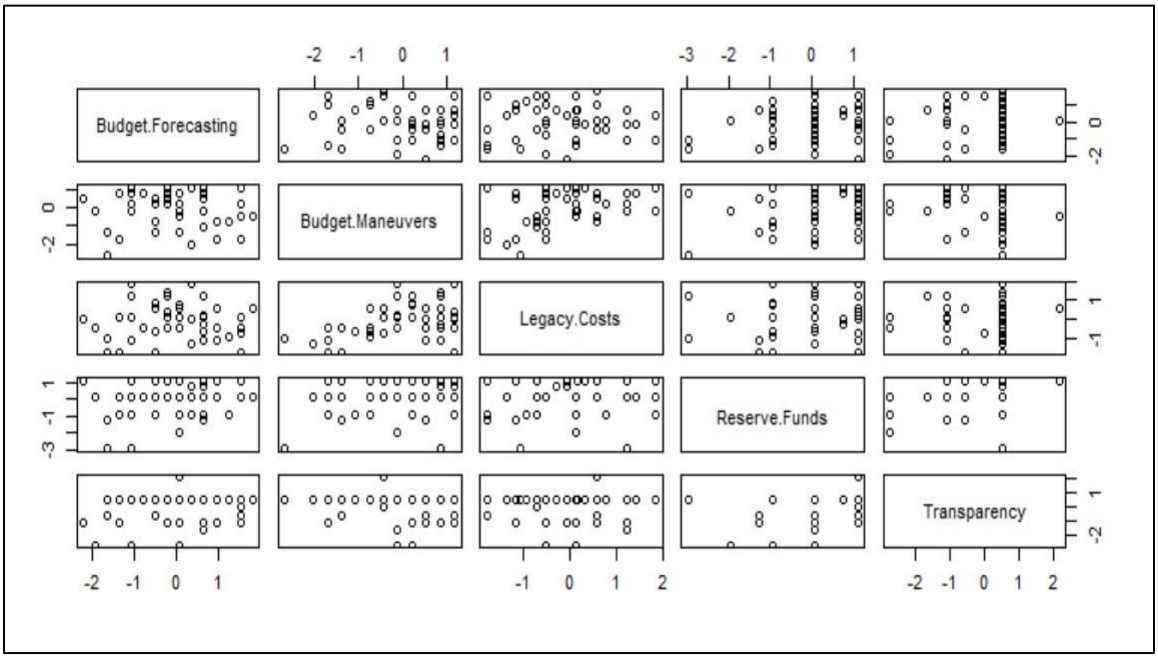


Figure 15: A Pairwise Scatter Plot Matrix for Our Five Variables

Figure 16 below shows a map view of the different grades of the Volcker Alliance's scores (i.e., A, B, C, D, D-) of the five variables for the states utilizing Tableau software. As you can see below the vertical axis is representing the grades (A to D-), and the horizontal axis represents the budget categories which are: budget forecasting, budget maneuvers, legacy costs, reserve funds, and transparency. This map view of the scores allows the readers to directly compare the scores of different categories of the different states. In the following section, we will present more visualization after getting the clusters' results.

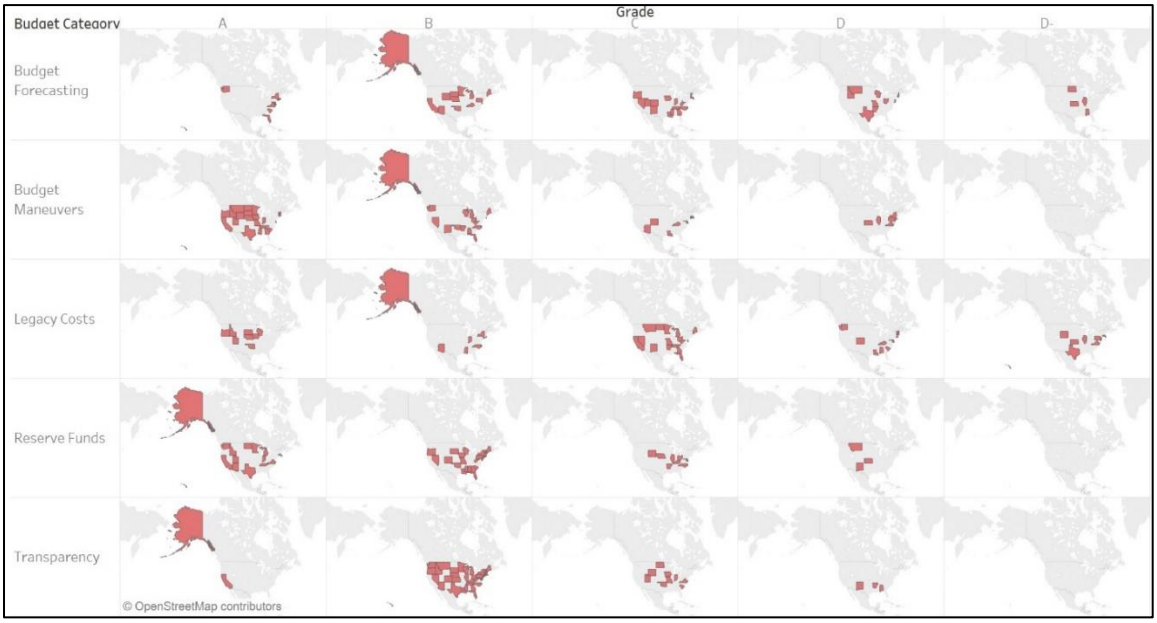


Figure 16: A Map View of Volcker Alliance's Scores for the States

3.5.4 K-MEANS CLUSTERING

Second, we apply k-means clustering - basically, the concept of *"birds of a feather flock together."*, McPherson, Smith-Lovin, & Cook (2001). K-means clustering algorithm divides *m* points in *n* dimensions into *k* clusters so that the within-cluster sum of squares is

minimized, which means that each observation (record in a table) belongs to the cluster with the nearest mean (Hartigan, & Wong, 1979; Jain, 2010).

We use five variables in our cluster analysis. The variables are taking from the survey data scores of Volcker's research results:

1. Budget Forecasting.
2. Budget Maneuvers.
3. Legacy Costs.
4. Reserve Funds.
5. Transparency.

In addition, we take the average for the data (i.e., three years), this includes the average for the years 2015, 2016 and 2017. To determine a good cut for the number of clusters, we use the elbow method. As mentioned earlier, the elbow method calculates the within clusters sum of squared errors (SSE) for each cluster *k* to determine the optimal number of clusters. Then plot a line of the SSE for each *k*. Therefore, if the line looks like an arm, then the elbow on the arm represents a good value "cut" for *k* (Bholowalia & Kumar, 2014). Looking at the graph below, using the elbow method, we find that seven clusters would be a good cut. The figure below shows a scree plot where the x-axis represents the number of clusters, and the y-axis shows the sum of squared distances of the clusters' means to the global mean.

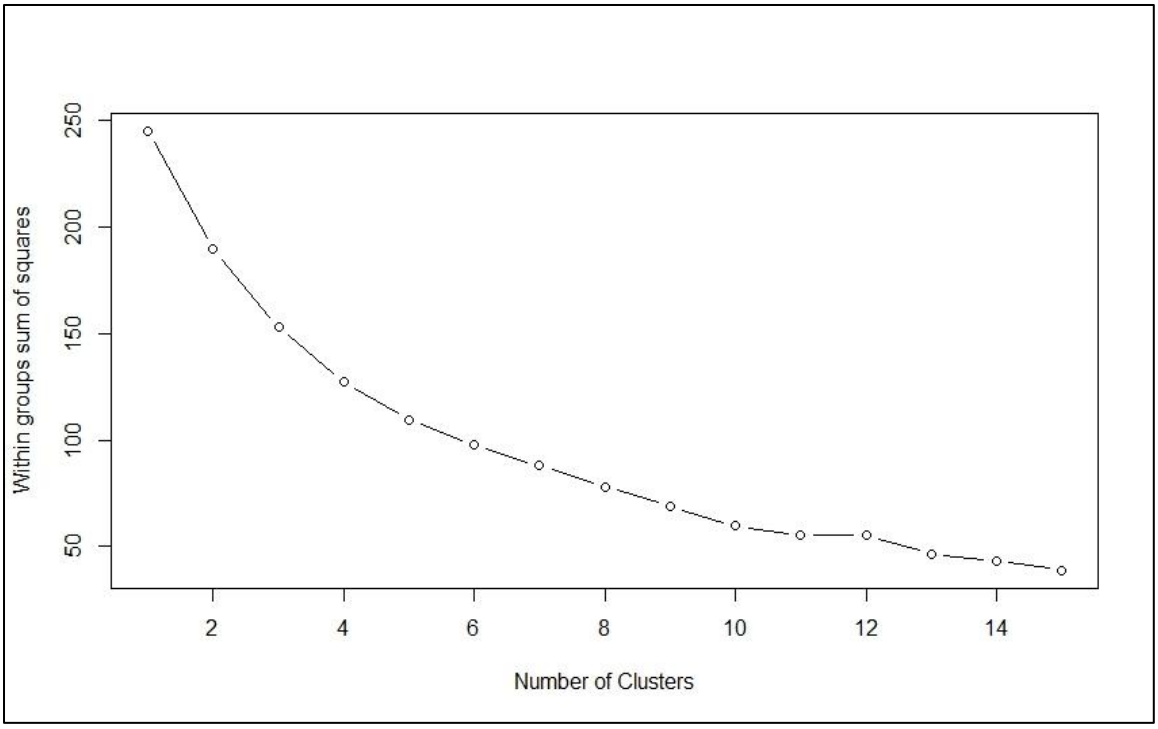


Figure 17: Within Clusters Sum of Squares (Elbow Method)

The next cluster plot, Figure 18, shows a two-dimensional representation of the seven clusters for the average scores using the five variables, which are budget forecasting, budget maneuvers, legacy costs, reserve funds, and transparency. Before plotting the clusters, we use the principal components analysis (PCA). The PCA is a statistical technique, and it is widely used by statistician and data scientists to find trends and patterns in the data (Smith, 2002). PCA attempts to find a combination of attributes that lead to maximum separation of data points. By applying PCA on our dataset where we have five numerical features, we obtain five principal components (PC 1-5). Each one explains a proportion of the total variation in the dataset. That is to say: PC 1 explains 32% of the total variance, which means around one-third of the information in the dataset. PC 2 explains 27% of the variance. So, by knowing PC 1 and PC 2, you can have an accurate

view or variability in your data, as just PC 1 and PC 2 can explain around 58% of the variance. The following table shows the importance or variability of each one of the four components.

Importance of components	PC 1	PC 2	PC 3	PC 4	PC 5
Standard deviation	1.2551352	1.1599979	0.9719874	0.8466411	0.64612677
Proportion of variance	0.3150729	0.2691190	0.1889519	0.1433602	0.08349596
Cumulative proportion	0.3150729	0.5841919	0.7731438	0.9165040	1.00000000

Table 9: The Principal Components Distribution

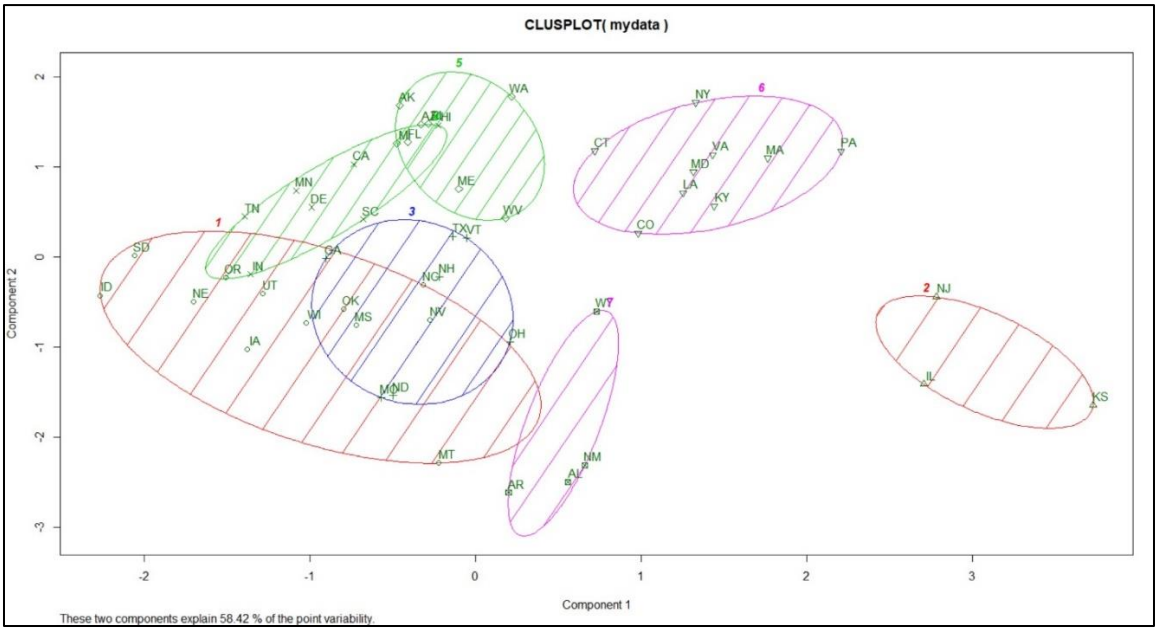


Figure 18: A Visualization of Our k-means Clusters

As shown from the previous figure, the states in the same cluster are more similar to the ones in their group than they are to the states in other clusters. The table below shows each state that belongs to a certain cluster.

Cluster	**Members**
Cluster 1	ID, SD, NE, IA, UT, OR, WI, OK, MS, NV, NC, MT
Cluster 2	NJ, IL, KS
Cluster 3	TX, VT, GA, MO, ND, OH, NH
Cluster 4	TN, MN, DE, CA, HI, SC, IN
Cluster 5	AK, WA, AZ, FL, ME, WV, MI, RI
Cluster 6	CT, NY, PA, MA, VA, MD, LA, KY, CO
Cluster 7	NM, AL, AR, WY

Table 10: State Distributions Within Each Cluster

3.5.5 HIERARCHICAL CLUSTERING

After clustering the data using k-means algorithm in the previous section, here we illustrate our results using hierarchical clustering algorithm. As in the first case study, we also use Ward.D method with Euclidean as a measure of distance for building the hierarchy. The figure below shows a dendrogram representation of the hierarchical tree.

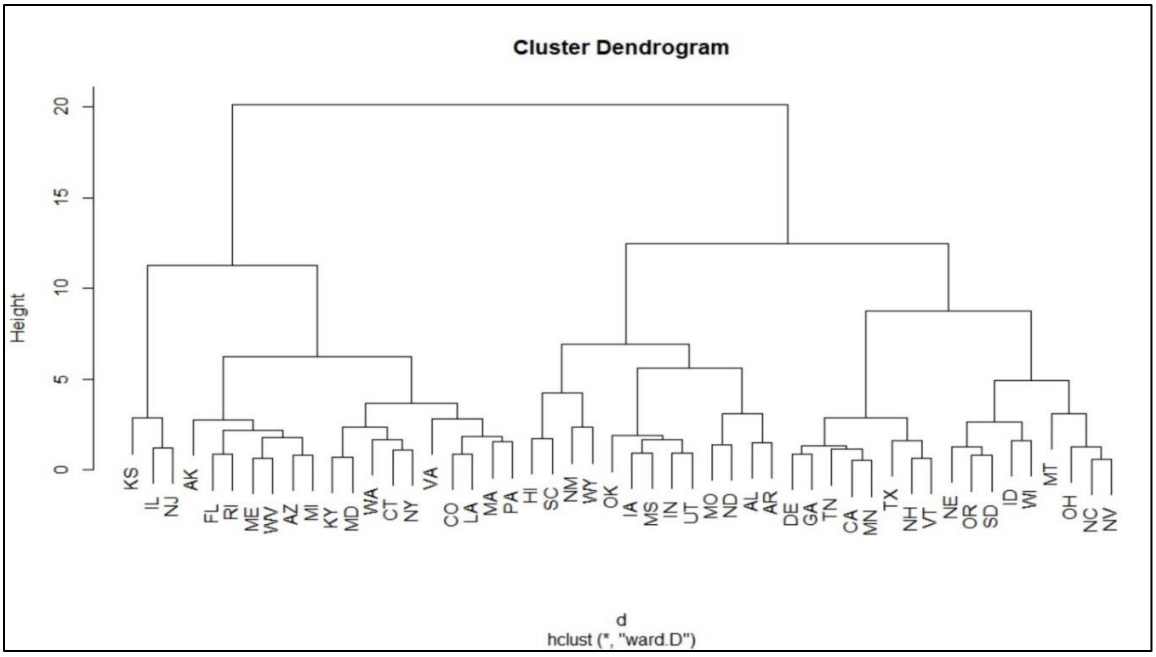


Figure 19: A Visualization of Our Hierarchical Clustering Dendrogram

In the following figure, we cut the hierarchical tree at a certain height where we get seven clusters. The height of the dendrogram cutting point controls the number of clusters obtained. Similar to the *k* in k-means clustering. We use the cutree function in R to cut the hierarchical tree. The resulted clusters are separated with a borderline around each cluster.

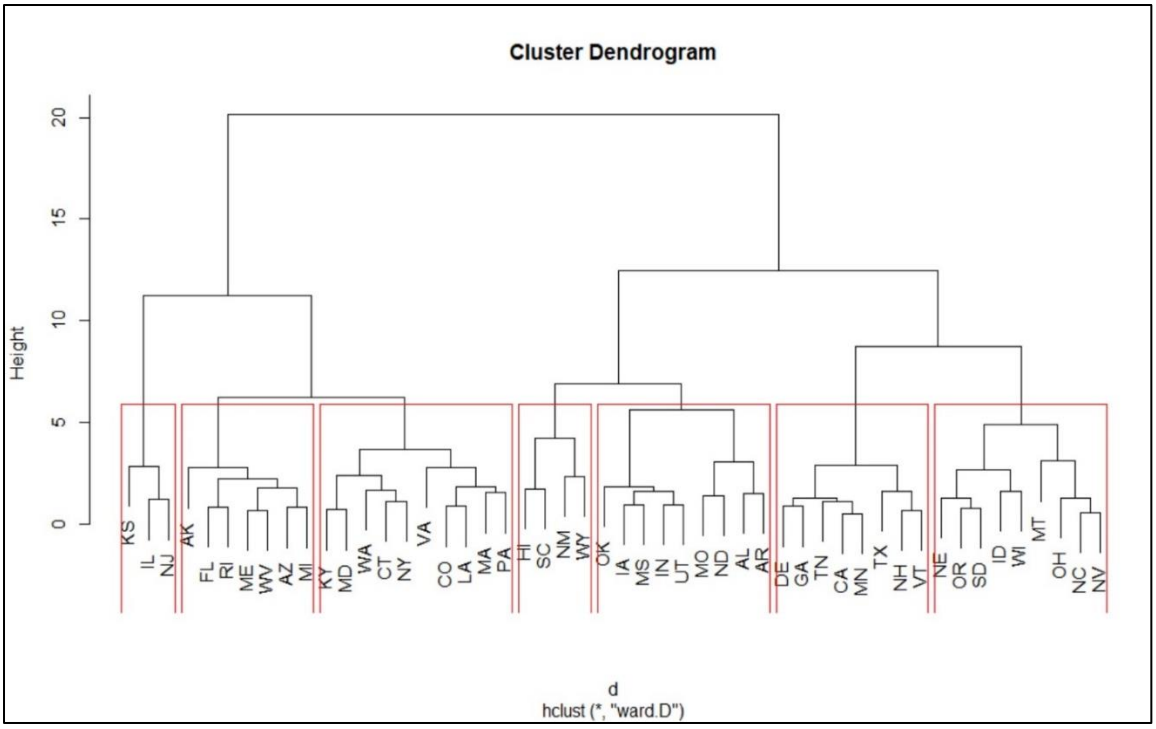


Figure 20: A Visualization of Our Hierarchical Clustering Dendrogram

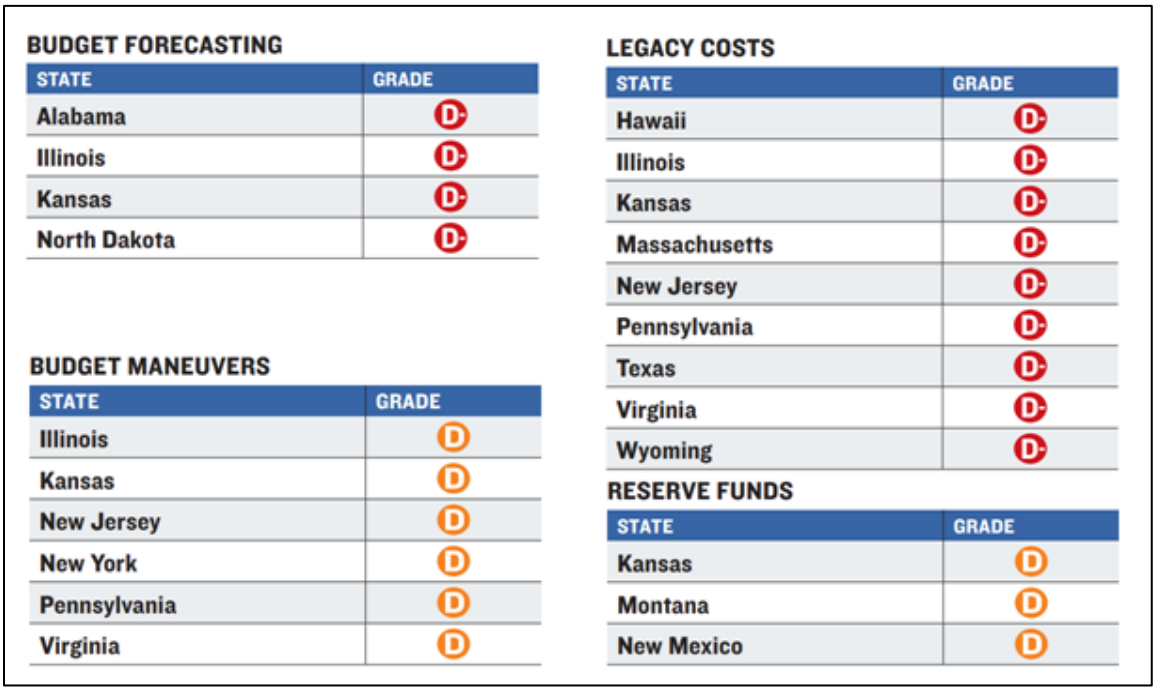
As shown from the previous figure, the states are clustered as follow.

Cluster	**Members**
Cluster 1	NJ, IL, KS
Cluster 2	AK, FL, RI, ME, WV, AZ, MI
Cluster 3	KY, MD, WA, CT, NY, VA, CO, LA, MA, PA
Cluster 4	HI, SC, NM, WY
Cluster 5	OK, IA, MS, IN, UT, MO, ND, AL, AR
Cluster 6	DE, GA, TN, CA, MN, TX, NH, VT
Cluster 7	NE, OR, SD, ID, WI, MT, OH, NC, NV

Table 11: State Distributions Within Each Cluster

The following table (Table 12) compares the two-clusters' results (k-means and hierarchical clusters). The states that populate each cluster of the hierarchical method are

moderately different from the clusters obtained from k-means with the exception of NJ, IL, and KS. These are identical in both methods. Their similarities affirm that the clusters for both methods are well distributed. Many readers may be surprised to see that Kansas belongs with New Jersey and Illinois. After all, there is much press about the budgetary woes of New Jersey and Illinois, not to mention the findings of the Volcker survey, but little publicity in general about Kansas state. However, upon closer examination of the survey results, we find that Kansas scores poorly in areas of legacy costs and budget maneuvers which are the two main contributing factors to the clusters. In fact, Kansas scores poorly in four out of five of the Volcker's survey variables. Although the reader might notice these survey results in their report (Figures 21) below, with cluster analysis this condition is highlighted immediately (The Volcker Alliance, 2017).



BUDGET FORECASTING

STATE	GRADE
Alabama	D-
Illinois	D-
Kansas	D-
North Dakota	D-

BUDGET MANEUVERS

STATE	GRADE
Illinois	D
Kansas	D
New Jersey	D
New York	D
Pennsylvania	D
Virginia	D

LEGACY COSTS

STATE	GRADE
Hawaii	D-
Illinois	D-
Kansas	D-
Massachusetts	D-
New Jersey	D-
Pennsylvania	D-
Texas	D-
Virginia	D-
Wyoming	D-

RESERVE FUNDS

STATE	GRADE
Kansas	D
Montana	D
New Mexico	D

Figure 21: Budget Forecasting, Budget Maneuvers, Legacy Costs, Reserve Funds - The Lowest Graded States (The Volcker Alliance, 2017)

Cluster	K-means	Hierarchical
Cluster 1	ID, SD, NE, IA, UT, OR, WI, OK, MS, NV, NC, MT	NJ, IL, KS
Cluster 2	NJ, IL, KS	AK, FL, RI, ME, WV, AZ, MI
Cluster 3	TX, VT, GA, MO, ND, OH, NH	KY, MD, WA, CT, NY, VA, CO, LA, MA, PA
Cluster 4	TN, MN, DE, CA, HI, SC, IN	HI, SC, NM, WY
Cluster 5	AK, WA, AZ, FL, ME, WV, MI, RI	OK, IA, MS, IN, UT, MO, ND, AL, AR
Cluster 6	CT, NY, PA, MA, VA, MD, LA, KY, CO	DE, GA, TN, CA, MN, TX, NH, VT
Cluster 7	NM, AL, AR, WY	NE, OR, SD, ID, WI, MT, OH, NC, NV

Table 12: K-means and Hierarchical Clusters Results

Next, after getting the clustering results, we demonstrate additional data visualization using Tableau software. The following figure shows Moody's U.S. States' Bond Ratings.

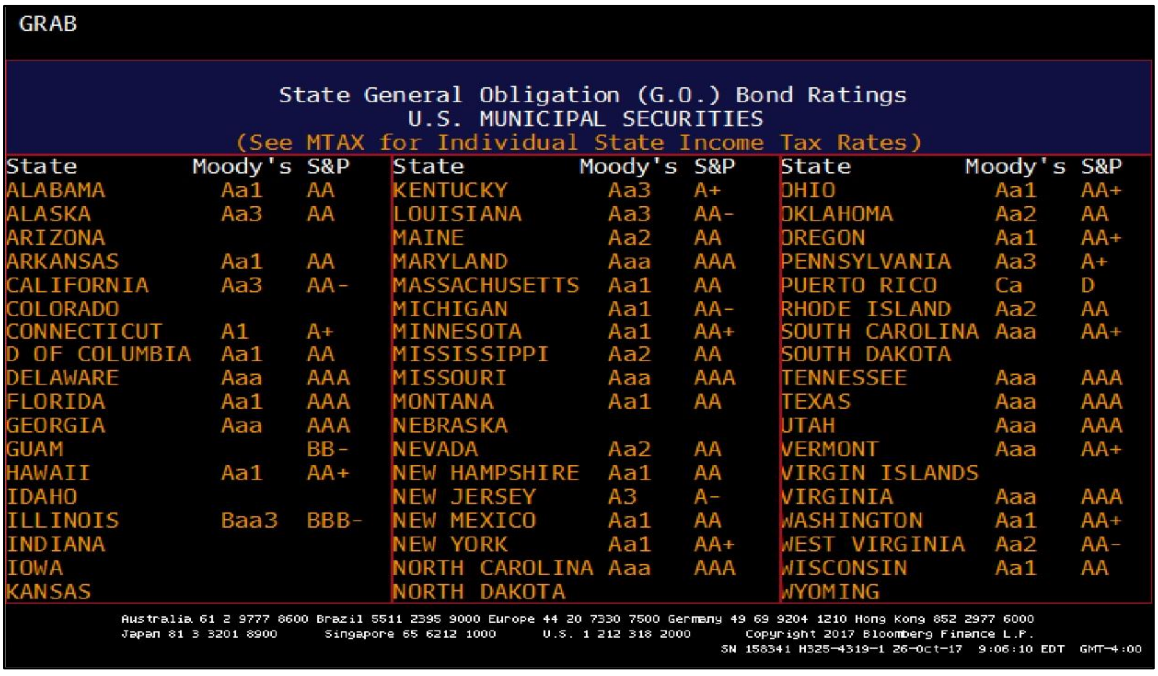


Figure 22: Moody's U.S. States' Bond Ratings

Next, the following figure shows a map view of Moody's ratings of the U.S. states using Tableau.

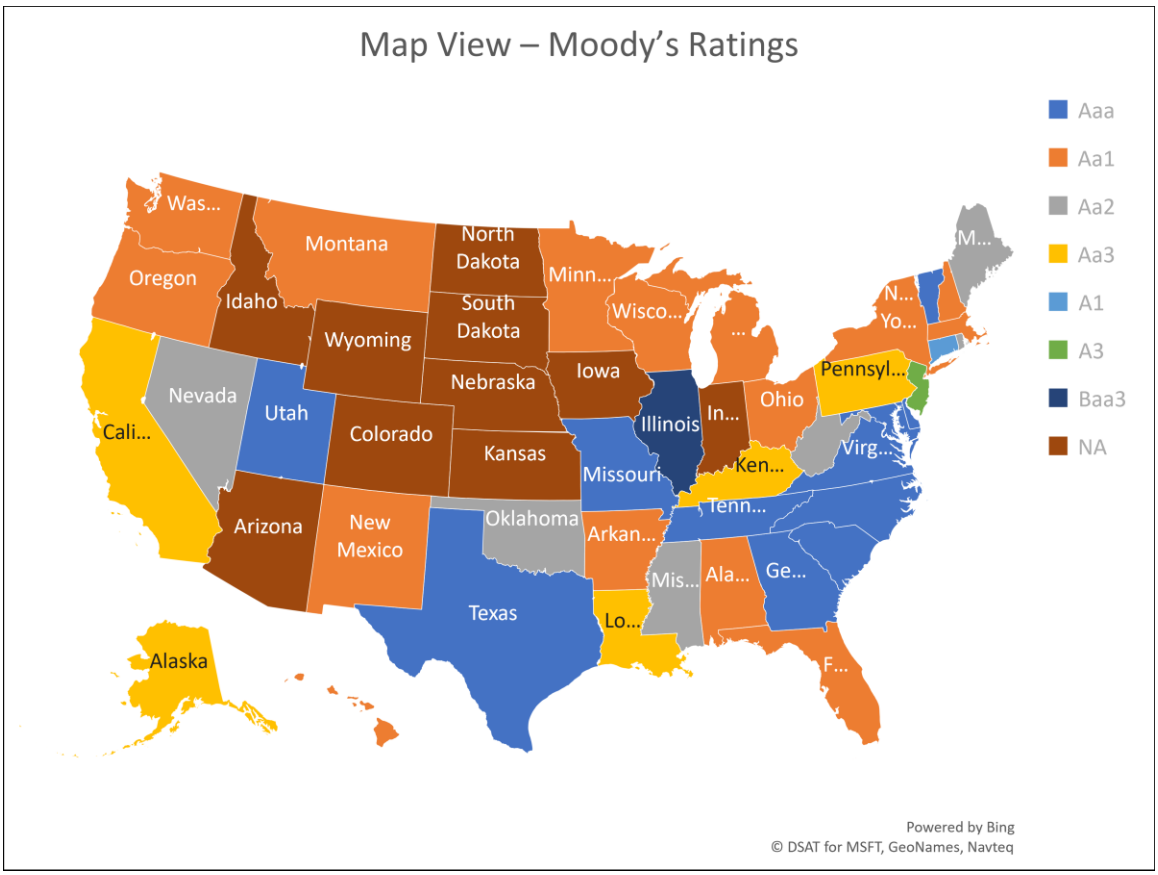


Figure 23: Map View of Moody's Ratings of the States

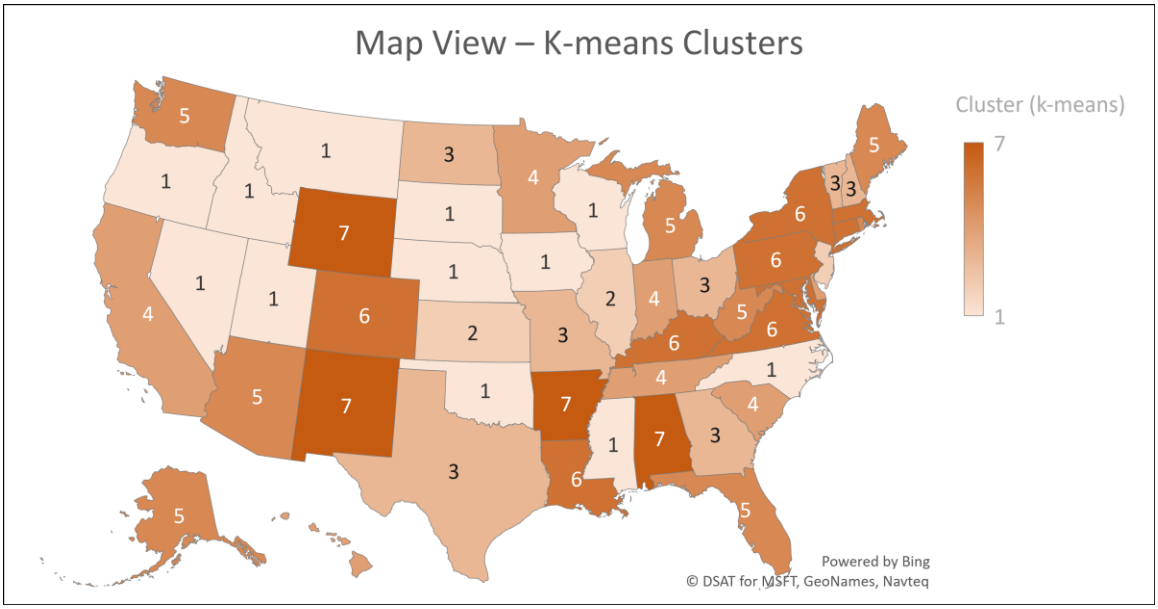


Figure 24: Map View of the K-means Clustering Results

The figure above shows a map representation of the k-means results.

From this data visualizations of the clusters' results and the Moody's U.S. States' Bond Ratings, we can notice some relationship between the states and the states' ratings such that Illinois was graded Baa3, and New Jersey graded as A3. However, Kansas was not graded by Moody's. Both ratings considered low or moderate credit risk which is not high enough.

3.5.6 ASSESSMENT OF THE RESULTS

To evaluate our results, we plan to apply clustering assessment methodology that uses bootstrap resampling which assesses how stable a given cluster is. This methodology is important when using clustering algorithms such as k-means and hierarchical clustering,

especially the latter, where most of the time you have to use an a priori method of choosing the optimal number of clusters. The method was first introduced by Christian Hennig in 2007 on his published paper, "Cluster-wise assessment of cluster stability," (Hennig, 2007; Hennig, 2018).

Clustering algorithms mostly produce clusters that represent the actual structure and relationships in the data, and then there are few clusters that are miscellaneous. One common way to check if a cluster represents true structure of the data is to see how each cluster holding up under apparent variations in the dataset. We use the function clusterboot() that is available in the "fpc[7]" package using the statistical software R. The clusterboot() function uses a bootstrap resampling to evaluate the stability of each cluster (Hennig, 2007). Clusterboot() function can perform clustering and also evaluate the resulted clusters. It can run many clustering algorithms including k-means and hierarchical clustering (Zumel, 2014, 2015).

Running the clusterwise cluster stability assessment algorithm by resampling using clusterboot function in our dataset for our hierarchical clustering method, results in the following (Table 13):

Cluster (HC)	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Mean value	0.63043	0.69833	0.62334	0.50435	0.68573	0.60433	0.56000
Dissolve	7	5	8	13	5	10	13

Table 13: The Mean and The Number of Times a Cluster is Being Dissolved (Hierarchical Clusters)

[7] Fpc is a package in R statistical software which has the clusterboot method among others that runs the clusterwise stability assessment for clustering (Hennig, 2007; Hennig, 2018).

The above results show the clusterwise means for each cluster. Meaning the average values for the 20-bootstrap iterations. Values close to 1 indicate stable clusters. Any mean value that is less than or equal to 0.50 indicates a dissolve cluster. Mean values between 0.6 and 0.75 considered clusters that clearly indicating patterns in the data. Mean values equal to 0.85 or above are considered highly stable clusters (Hennig, 2007). It also shows the number of times each cluster has been dissolved. As you can see, the second cluster has dissolved only five times. Thus, results in the highest mean value of 0.69833 which indicates the most stable one, followed by the fifth cluster with a mean value of 0.68573 and then the first cluster with a value of 0.63043. With the fourth cluster being the lowest that has a mean value of 0.50435 meaning we cannot trust it. The low value indicates that it dissolved the most among the other clusters.

Next, by running clusterwise cluster stability assessment using the clusterboot function for our k-means clustering method, we get the following results (note that we are using 20-resampling runs for each cluster on both the hierarchical and the k-means clustering methods).

Cluster (K)	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Mean value	0.52234	0.57123	0.58506	0.54961	0.86000	0.66667	0.57313
Dissolve	9	9	10	11	1	8	10

Table 14: The Mean and The Number of Times a Cluster is Being Dissolved (K-means Clusters)

Notice that cluster number five in Table 14 above has a high mean value that is 0.86000 for which indicates high stability. This cluster has been dissolved only once.

Looking at the clusterwise stability assessment results of the two clusters: k-means and hierarchical, we can notice that, overall, hierarchical clustering performs better than k-means with higher mean values with less dissolved clusters. However, assessing clusters stability is not the only important validity criterion since there are some clustering methods that could result in stable but not valid clusters (Hennig, 2007).

3.6 CONCLUSIONS AND FUTURE RESEARCH

In this study, we explore the literature for the use of emerging data mining techniques in auditing. In particular, cluster analysis techniques and visualization as supportive tools for auditors and practitioners. We apply two different clustering techniques, k-means, and hierarchical clustering. We found that cluster analysis along with the use of data visualization could help get more insights into the data.

On the first case study, we use the U.S. states financial statements data. The second case study utilizes the Volcker Alliance's Survey data results. The survey produces extensive information about how the different U.S. states score on an annual basis on budgeting using five measures. On both case studies, we show how visualization and data clustering could be used on governmental data and to help gain more information about financial statements and budgeting. The cluster results also show that there are some similarities between the two methods, k-means and hierarchical, and this could give us an idea about our data quality. We have also used an assessment methodology to evaluate the stability level of the clusters' results. In addition, we have now clear and unusual patterns and relationships to explore in greater depth.

The contribution of the second essay is two-fold. First, the two case studies presented in the second essay bring some advanced visualization and data mining techniques into the governmental domain, especially using financial statements and budgetary data. Second, the visualization and clustering results bring more opportunities for auditors and practitioners and to help get more insights into the data. Finally, we have used two clustering methodologies, k-means and hierarchical clustering; However, there exist some limitations in the study such that we need to more evaluate the results of the first case study such that examining the clusters with external variables; and also more data is needed for the second case study, we have only used three years of data.

Future research could apply and compare additional data clustering techniques such as partition around medoids (PAM) clustering method. Also, by comparing the results of cluster analysis using external variables such as net population change and gross domestic product (GDP) growth. In addition, the states that are grouped together on the same cluster could be further analyzed.

REFERENCES

Arisandi, D. (2016). The implementation of data analytics in the governmental and not-for-profit sector (Doctoral dissertation, Rutgers University-Graduate School-Newark).

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

Bloch, D., & Fall, F. (2015). Government debt indicators: Understanding the data.

Bobric, E. C., Cartina, G., & Grigoras, G. (2009). Clustering techniques in load profile analysis for distribution stations. Advances in electrical and computer engineering, 9(1), 63-66.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control VII, 235-255.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863-14868.

Elkan, C. (2001, August). Magical thinking in data mining: lessons from CoIL challenge 2000. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 426-431). ACM.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical clustering. Cluster Analysis, 5th Edition, 71-110.

F. V. Jensen, An introduction to Bayesian networks. UCL press London, 1996, vol. 210.

Fua, Y. H., Ward, M. O., & Rundensteiner, E. A. (1999, October). Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of the conference on Visualization'99: celebrating ten years (pp. 43-50). IEEE Computer Society Press.

GASB. (2010, December). The User's Perspective. Retrieved February 11, 2019, from https://gasb.org/cs/ContentServer?c=GASBContent_C&cid=1176158164302&d=&pagename=GASB/GASBContent_C/UsersArticlePage&pf=true.

Harpaz, R., Perez, H., Chase, H. S., Rabadan, R., Hripcsak, G., & Friedman, C. (2011). Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clinical Pharmacology & Therapeutics, 89(2), 243-250.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

Hegar, G, (Texas Comptroller of Public Accounts). GUIDE TO UNDERSTANDING COMPREHENSIVE ANNUAL FINANCIAL REPORTS (CAFRS). Retrieved February 11, 2019, from https://comptroller.texas.gov/transparency/budget/cafr-faq.php.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis, 52(1), 258-271.

Hennig, C. (2018, July 20). Fpc: Flexible Procedures for Clustering. Retrieved February 12, 2019, from https://cran.r-project.org/web/packages/fpc/index.html.

Hierarchical clustering. (2017, October 16). Retrieved October 24, 2017, from https://en.wikipedia.org/wiki/Hierarchical_clustering.

Hilas, C. S., & Mastorocostas, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. Knowledge-Based Systems, 21(7), 721-726.

Hossain, M. S. (2012). Exploratory data analysis using clusters and stories. Virginia Polytechnic Institute and State University.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Kassambara, A. (2017). Practical guide to cluster analysis in R: unsupervised machine learning (Vol. 1). STHDA.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

Kim, K. J., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. Expert systems with applications, 34(2), 1200-1209.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.

Liu, Q., & Vasarhelyi, M. A. (2014). Big questions in AIS research: Measurement, information processing, data analysis, and reporting. Journal of information systems, 28(1), 1-17.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological methods, 1(1), 30.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1), 415-444.

Roy, K., Kar, S., & Das, R. N. (2015). Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic press.

Schneider, G. P., Dai, J., Janvrin, D. J., Ajayi, K., & Raschke, R. L. (2015). Infer, predict, and assure: Accounting opportunities in data analytics. Accounting Horizons, 29(3), 719-742.

Singh, K., Malik, D., & Sharma, N. (2011). Evolving limitations in K-means algorithm in data mining and their removal. International Journal of Computational Engineering & Management, 12, 105-109.

Smith, L. I. (2002). A tutorial on principal components analysis.

Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. IEEE Transactions on dependable and secure computing, 5(1), 37-48.

The Volcker Alliance (1), Truth and Integrity in State Budgeting: Truth and Integrity in State Budgeting: Preventing the Next Fiscal Crisis , December 2018, https://www.volckeralliance.org/sites/default/files/TruthAndIntegrityInStateBudgetingWhatIsTheRealityFiftyStateReportCards.pdf.

The Volcker Alliance (2), Truth and Integrity in State Budgeting: What is the Reality? Fifty state Report Cards, 2018, https://www.volckeralliance.org/sites/default/files/TruthAndIntegrityInStateBudgetingWhatIsTheRealityFiftyStateReportCards.pdf.

The Volcker Alliance, Truth and Integrity in State Budgeting: What is the Reality?, November 2017, https://www.volckeralliance.org/sites/default/files/TruthAndIntegrityInStateBudgetingWhatIsTheReality_0.pdf.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. International Journal of Digital Accounting Research, 11.

Turban, E., Sharda, R., & Delen, D. (2011). Decision support and business intelligence systems. Pearson Education India.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In ICML (Vol. 1, pp. 577-584).

Williams, G. J., & Huang, Z. (1997, November). Mining the knowledge mine. In Australian Joint Conference on Artificial Intelligence (pp. 340-348). Springer, Berlin, Heidelberg.

Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. Data mining and knowledge discovery, 10(2), 141-168.

Zumel, N. (2015, September 04). Bootstrap Evaluation of Clusters. Retrieved February 12, 2019, from https://www.r-bloggers.com/bootstrap-evaluation-of-clusters/.

Zumel, N., Mount, J., & Porzak, J. (2014). Practical data science with R (pp. 101-104). Greenwich, CT: Manning.

CHAPTER 4: AN ONTOLOGY-BASED FRAMEWORK FOR CLASSIFYING SOCIAL MEDIA: TEXT MINING ANALYSIS FOR FINANCIAL DATA

ABSTRACT: In this paper we utilize a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the two public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). We apply a frame and slot approach (Frame-based system) from the artificial intelligence literature to operationalize the FIBO ontology in a public sector/municipalities business context. FIBO is part of the Enterprise Data Management Council (EDMC) and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. One contribution of this paper is that it is the first to recognize that the FIBO structure provides a grammar of financial concepts. We show that this grammar can be used to mine semantic meaning from unstructured textual data. Twitter streams will be monitored and analyzed with frames derived from FIBO and keywords. The ability of the FIBO frames to detect semantic meaning in tweets is compared with naïve keyword analysis. Using FIBO frames, constituent semantic structures can be uncovered to predict reactions to policies and programs more quickly than by following the feeds manually.

4.1 INTRODUCTION

In this paper, we utilize a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the collective public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). We apply a frame and slot methodology from the artificial intelligence literature to operationalize the FIBO ontology in the public sector/municipalities business context. Frames and slots are part of the knowledge representation and reasoning (KR) which is a field of artificial intelligence (AI) which is a way of representing information about anything in the world in a form that a computer system can understand (Lassila & McGuinness, 2001). Examples of knowledge representation include semantic networks, logical representation, production rules, and frame representation.

FIBO is part of the Enterprise Data Management Council (EDMC) and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. One contribution of this paper is that it is the first to recognize that the FIBO structure provides a grammar of financial concepts. We show that this grammar can be used to mine semantic meaning from unstructured textual data. Twitter streams will be monitored and analyzed using frames derived from FIBO and keywords. The ability of the FIBO frames to detect semantic meaning in tweets is compared with naïve keyword analysis. Using FIBO frames, constituent semantic structures can be uncovered to predict reactions to policies and programs more quickly than by following the feeds manually.

Collectively, PANYNJ and the MTA are responsible for nearly all of New York and northern New Jersey's transportation infrastructure – subways, buses, commuter rail, bridges, tunnels, airports, and ports. Individually they are both considered to be the two largest transportation agencies in the world (Polycarpou, 2014) and together they cover over 25 square miles of a densely populated urban metropolis (Campbell et al., 2017). Both agencies are considered to be "public benefit corporations" in that they operate as quasi-private corporations that serve the public good. As such, public commentary about these entities could provide valuable insights about perceived successes and issues. Both agencies are chartered by the states that they serve (NY, NJ, and a small portion of CT) but are funded by self-issued debt (bonds) and the tolls that they collect. Initially, the intent was not to rely on any local state funding, but chronic huge operating deficits have forced both agencies to seek state subsidies and institute frequent fare increases.

Some attribute this seemingly perpetual deficit to the confusing shared control of the PANYNJ between the two states (Polycarpou 2014). There seems to be a considerable misalignment between who uses, who pays, and who controls the agencies. MTA is primarily used by NYC residents but is run by the Governor of NY. About 80% of the daily transit riders are NYC residents, yet 25% of the subsidies are provided from the suburbs. If the fares were to be adjusted to address this gap, the increase could amount to $0.73/ride.

This study contributes to accounting academic research in that it appears to be the first that applies a formal business ontology to unstructured social feeds, such as Twitter. Using such ontologies facilitates the identification and understanding of the nuanced threads in Twitter that pertain to such issues as probable and/or pending municipal bond releases. Twitter data has been found to be relevant for predictive sentiment analysis (Pak

and Paroubek 2010). During the time period of this study, PANYNJ bond series were rated as stable Aa3 by Moody's and the Port Authority Board approved the application to raise funds to upgrade one of its airports (Mid-Hudson News Network, 2018). It is anticipated that these announcements would evoke a public reaction, and this study applies a formal accounting ontology to the processing of these social media extractions to facilitate and organize understanding. Such understanding would be of interest to bond issuers, regulators, analysts, investors, and academics.

4.2 LITERATURE REVIEW

Twitter is an online news, and social networking service where users post and interact with SMS-like posts called "tweets." Tweets are publicly visible by default or can be restricted by the poster to only be sent to his/her "followers" (Stutzman 2007). Twitter is regarded as an unfiltered source of current information and news (Syed, Gillela, & Venugopal, 2013). As of 2016, Twitter had 319 million active users. Most of the "tweets" are of news and social networking value (Syed et al., 2013).

Twitter data is usually classified as unstructured big data (Warren, Moffitt, & Byrnes, 2015). In fact, about 90% of big data is unstructured and is comprised of emails, social media posts such as Facebook and Twitter, phone calls, audio files and video streams (Syed et al., 2013). Unstructured textual data has shown to be of increasing importance to firms that want to differentiate themselves (Warren et al., 2015). Because of the volume of data, ideally, automated text mining techniques should be applied to Twitter texts to extract high-quality information. This text mining process begins with structuring the tweets by

parsing, possibly adding and/or removing certain linguistic features, and adding them to a database, this stage called pre-processing the data. The resulting structured texts are then analyzed for patterns and interpreted using text categorization, text clustering, extractions, sentiment analysis, summarization, and modeling. However, due to restrictions on the maximum number of characters (now 280 as of November 7, 2017), many tweets contain shortcuts and abbreviations which may challenge this analysis process.

Despite these challenges, Twitter feeds have been studied by researchers in a broad range of domains. Kavanaugh et al. (2012) conducted a six-month exploratory study on how social media data could be leveraged to improve the government's communications and services with citizens. They collected data using Twitter, Facebook, and Flicker platforms relevant to Arlington County in the State of Virginia. More specifically, they analyzed the collected feeds by monitoring public opinion before and after public events and identify important community topics over time and location. They collected area-specific social media data mostly from Twitter, Facebook and comments from YouTube videos and conducted group interviews with 24 county officials. They applied semantic analysis to obtain popular topics and frequency analysis from social media data. Their findings could be summarized as follows: first, local government agencies do not have a full understanding of the use of social media, nor realize its benefits or potential - who is the actual audience, who should oversee managing governments' different social media accounts and what is the actual effect of their tweets. Secondly, new tools should be made available for both citizens and governments to help gather and get insights into the massive amount of data generated. Thirdly, there exists a need for digital libraries to archive generated content related to crises and other big events.

On crime data analytics, Chen et al. (2004) suggested a general framework for crime data mining techniques. They accessed the Tucson Police Department (TPD) crime dataset and studied the existing literature. The dataset contains around 1.3 million criminal records from 1970 onward. On this basis, they developed a general data mining framework that shows the relationships between the type of crime and the data mining techniques applied. For instance, the framework shows that cluster analysis could be used in crime association and prediction. Social network analysis, on the other hand, could help on crime association and pattern recognition. To illustrate their framework applicability, they applied three use cases. These are name entity extraction, deceptive identity detection, and criminal network analysis. For name entity extraction, they extracted names from police narrative reports from the TPD. They used an entity extraction system which follows a three-step process to extract names, address, vehicle type, narcotic name and individual characteristics for each. The system first extracts noun phrases; then it calculates a feature score based on pattern and lexicon lookup. Finally, it applies a feedforward neural network to predict which category each noun phrase belongs to. The system extractor performs well in identifying these five entity types. This case study shows the potential value of using entity text extraction, one dimension of their general framework, to crime data. For deceptive identity detection, their team manually collected a sample of deceptive records. Based on this taxonomy, they identified name, birth date, SSN, and address to represent a criminal identity. Then they employed string comparators to each record with its corresponding one to calculate similarities. Finally, for criminal network analysis use case, they collected 272 incident report summaries from TPD database. They then created a network of potential suspects by extracting criminal relations from incident reports. Co-

occurrences measure the likelihood that a two-criminals are linked by computing how frequently they appear on the same incidents. However, one limitation of this approach is that it creates static networks, and criminal networks usually are dynamic in nature.

In the business domain, there exists a wide adoption of social media and text mining for various purposes. For instance, He, Zha, & Li, (2013) study the feasibility of using text mining techniques to increase competitive advantages of businesses. They analyzed and monitored not only businesses' social media content, but also businesses competitors' content. They conducted a case study using text mining to analyze unstructured content from Facebook and Twitter of three large pizza chains: Domino's Pizza, Pizza Hut and Papa John's. Their approach contains three steps which are text collection and pre-processing, processing and analyzing, and actionable results. In the last step, they extracted knowledge by identifying patterns, issues, trends, etc., and by providing recommendations for further actions. They utilized two software packages for text mining: SPSS Clementine and NVivo 9[1]. SPSS was used for linguistic methods such as grouping, extracting and indexing, while NVivo 9 was used for queries. The first step entailed obtaining insights into quantitative data for each pizza chain such as number of tweets, number of followers, shares, replies, comments. Then they combined the content of the three chains to extract five popular themes such as ordering and delivering, quality, marketing tweets, etc. Then they analyzed each different theme among the three competitors. Findings in this study suggest that social media and competitive text mining analysis could contribute significantly to gaining more insights for better organization-level decision making.

[1] SPSS Clementine is which is called now SPSS Modeler is a text mining tool from IBM (https://www.ibm.com/products/spss-modeler); NVivo 9 is another text mining tool for qualitative data analytics from QSR International.

In research on consumer brand sentiment in Twitter, Mostafa (2013) analyzed 3516 tweets to evaluate polarity of consumer sentiments of brands such as IBM, T-Mobile, KLM, Nokia and, DHL. He used a predefined lexicon including around 6800 seed objectives in his text analysis. The author findings show a general consumer positive sentiment towards many well-known brands. He suggested that the study contributes to the literature on consumers' general sentiment over international brands.

In stock market movement prediction, Bollen, Mao, & Zeng, 2011 conducted a sentiment text analysis of Twitter streams which showed that Twitter feeds could provide useful sentiment information or mood related to the stock market. Those moods could improve the accuracy of stock movement prediction. They investigated the public sentiment of Twitter and related that to movement in the Dow Jones Industrial Average (DJIA) by extracting a time series of the DJIA daily closing values. They found a significant correlation between moods and the DJIA closing values.

However, the process of “structuring” unstructured data or tweets to obtain high-quality information about accounting and financial information is challenging as this type of big data is unfamiliar to the profession. Accounting and finance typically analyze numbers and have only recently expanded into the textual analysis of financial statement footnotes and text, in addition to management conference calls (Warren et al., 2015). To provide structure to the textual data describing concepts of budgeting, accounting, and finance, formalization of taxonomies and hierarchies must be expanded (Moffitt & Vasarhelyi, 2013). Standardized semantic understanding and natural language processing are required to differentiate words and phrases.

4.3 ONTOLOGY BASED ACCOUNTING INFORMATION SYSTEM RESEARCH APPLIED TO TWITTER

4.3.1 SLOTS AND FRAMES STRUCTURE

The Financial Industry Business Ontology (FIBO) developed under the Enterprise Data Management Council (EDMC) and the Object Management Group (OMG). It is a standard representation of the ontology of the financial industry. The public side of FIBO is an operational ontology published in Research Description Framework RDF OWL[2] and other formats that are freely available with attribution. Behind the published FIBO is a conceptual ontology referenced and developed by the content teams which create FIBO. FIBO is a rich knowledge representation of the financial industry. The content teams consist of subject matter experts (SMEs) in various subareas of finance such as business entities and derivatives. This research uses only a small subset of the knowledge available in FIBO, the representation of municipal bonds. While it is possible to use FIBO as a collection of terms in which to get keywords for analyzing a twitter stream, doing so would lose the knowledge base aspect of the FIBO ontology. As such, frames are used here to maintain the knowledge embedded in the FIBO ontology in order to help filter twitter feeds to find threads referencing municipal bonds.

Frames are used for knowledge representation. Frames represent related knowledge about a narrow topic. A frame is similar to a record in a database, attributes and values of a record are the slots and slot fillers of the frame. Semantic nets, on the other hand, are a two-dimensional representation of knowledge, frames add a third dimension by allowing

[2] OWL is a Web Ontology Language which is a group of knowledge representation languages.

different nodes to have structures. Frames are good option to represent physical or conceptual structures, such as cars and financial instruments. In contrast, semantic nets are usually used in broad knowledge representations. A frame structure is defined in Table 15 below:

A Typical Frame with Frame Object "Car"	
Slot	Filler
Maker	Ford
Model	F-150
Year	2013
Engine	Gasoline
Tires	Goodyear
Color	Blue

Table 15: A Frame General Structure

Since Twitter feeds tend to contain minimal detailed information, they are unlikely to conform to highly complex frame structures. This being the case, we reduced the government issued bonds portion of FIBO into a frame structure using the concepts as the slots except in the case of the lowest level concepts where these became slot fillers. A portion of the FIBO diagram for municipal bonds is shown in Figure 25 below.

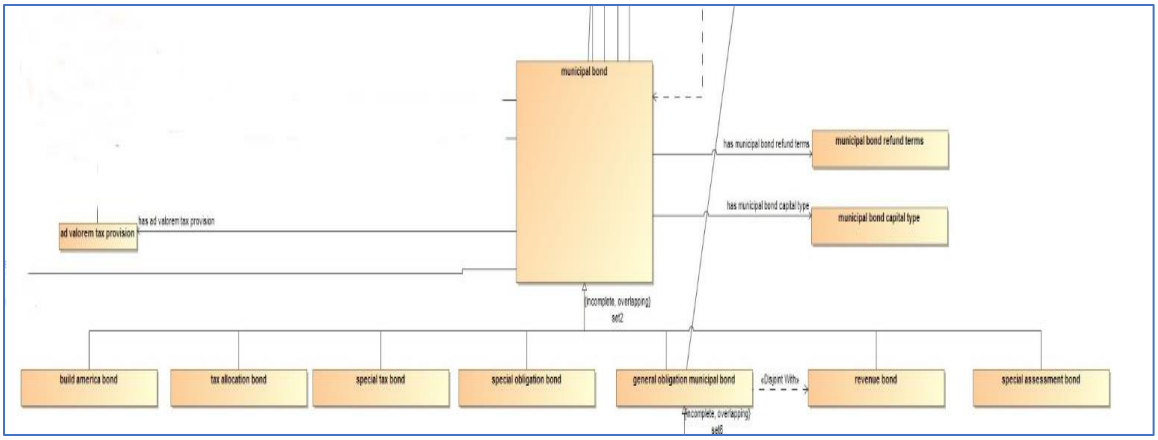


Figure 25: FIBO Municipal Bonds

The resulting frame for Government Issued Bonds is then as follows (Table 16).

Slot	**Filler**
Government Issued Bond	
Municipal Security	
Municipal Debt Issuer	
Municipal Bond	
Debt Obligor	
Funds usage	
Municipal Bond Capital Type	
Municipal Bond Refund Terms	
Municipal Trustee	
Ad valorem tax provision	
Municipal Bond Type	(Build America, Tax Allocation, Special Tax, Special Obligation, General Obligation, Revenue, Special Assessment, Consolidated Bond)

Table 16: Government Issued Bond Frame from FIBO

4.3.2 DEFINE THE RESEARCH OBJECTIVES

Although previous research discusses data standards for analysis of the softer qualitative data in financial statements (Warren et al., 2015), research has not been found that discusses formalizing financial textual information about municipal bonds in social media sources such as Twitter. This paper applies a frame and slot methodology from the artificial intelligence literature to operationalize the FIBO ontology in a public sector/municipalities business context. FIBO is part of the Enterprise Data Management (EDM) Council and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. FIBO concepts are vetted by subject matter experts (SMEs), so they should reflect high-quality financial concepts. One contribution of this paper is that it is the first to recognize that the FIBO structure provides a grammar of financial concepts. We show that this grammar can be used to mine semantic meaning from unstructured textual data. We compare this to an approach using naïve keywords. Twitter streams will be monitored and analyzed with frames derived from FIBO based on this framework. With such analysis, constituent semantic structures can be uncovered to predict reactions to policies and programs more quickly than by following the feeds manually.

4.4 DEMONSTRATION OF THE SOLUTION

4.4.1 IMPLEMENTATION OF THE TWITTER FEED SYSTEM

Based on the approaches of previous Twitter studies and of accounting ontology applications, the components of our proposed initial methodology are as follow (Figure 26):

1. Target a population of Tweets.
2. Develop an API for the targeted Tweets.
3. Collect and assemble the data.
4. Data aggregation and pre-processing.
5. The Financial Industry Business Ontology (FIBO) term search.
6. Construct validity test (FIBO versus naive keywords).
7. Design of the descriptive statistics algorithm and generic classifiers.
8. Tuning the FIBO generic classifiers.
9. Using the tuned classifiers to classify other twitter feeds.
10. Evaluation of the final results.

The sequential workflows of our methodology are outlined in Figure 26 below.

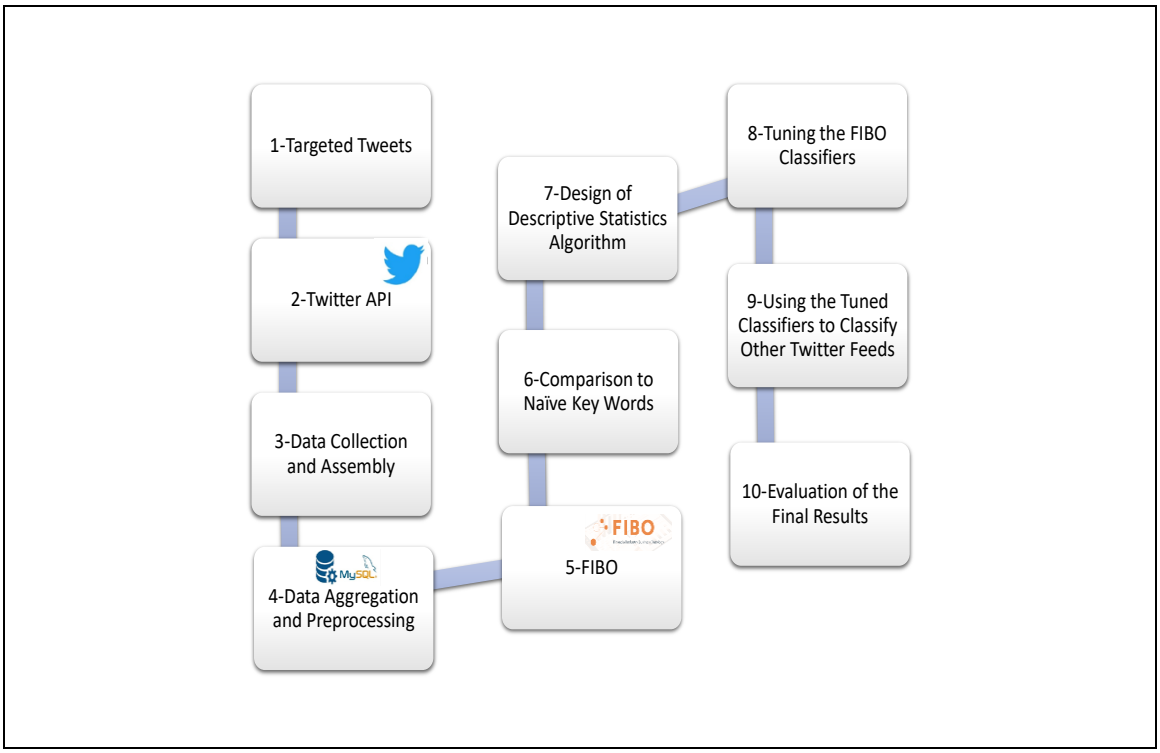


Figure 26: Proposed FIBO-Twitter Framework

Each of the steps in the methodology is described more fully here:

1. Targeted Tweets: The first step of our framework. Initially, we have chosen two major transportation agencies: The Port Authority of New York and New Jersey (PANYNJ), and The Metropolitan Transportation Authority (MTA). Also, one major division of PANYNJ is the PATH subway system. One reason for choosing PANYNJ and the MTA systems is because they are responsible for nearly all of New York and northern New Jersey's transportation infrastructure – subways, buses, commuter rail, bridges, tunnels, airports, and ports. In order to expand our search, we add as many words (keys) as possible, targeting all possible public tweets that mention those transportation hubs. For instance, to fetch tweets mentioning PANYNJ, we include the following keys (PANYNJ,

PORTAUTHORITY, PORT AUTHORITY, PABusTerminal, PORTNYNJ, Port Authority NY&NJ, the Port authority, Port authority of NY & NJ, Port authority of NY and NJ, the Port of New York and New Jersey). For tweets that mention MTA, the following keys were included (NYCT Subway, #MTA, #MTATransparency, NYCTSubway, NYCTBus, @MTA, LIRR, NYC Subway, #nycsubway). Finally, for tweets mentioning the PATH, the following keys were included (Path train, Path service, Pathtrain, #pathtrain).

2. Twitter API: The second step of our framework. After deciding on what will be the focus of the study, the Twitter Micro-blogging social media platform was chosen as the source of the data (https://twitter.com/). Twitter enables millions of users to share their feelings, opinions regarding any issue, any day. As a result, Twitter is considered as a rich source of data for opinion mining and sentiment analysis, Pak & Paroubek (2010).
3. Data Collection and Assembly: In this step, by accessing the Twitter Application Programming Interface (API), we wrote a Python code using Python 2.7 (https://www.python.org/) to fetch all Twitter stream that contains at least one of the targeted keys mentioned in Step One. Many Python libraries have been used such as Tweepy which allow us to access the Twitter API and StreamListener method that allows us to stream real-time messages (tweets) and route them to a storage space. We run the code for the three sessions at the same time to stream real-time data, one for each agency (PANYNJ, PATH, MTA). So far, the data has been collected from January 29th until October 19th, 2018.

The following Figure is a screenshot of the live-stream from Twitter API running in the background:

Figure 27: PANYNJ Twitter Live-stream (Python 2.7)

4. Data Aggregation and Preprocessing: After collecting the raw data, we aggregate six fields from each tweet. The fields are the following: date and time of the tweet, the original message (tweet), user identification number, number of followers, number of likes, number of posts that a certain user has. Then, we perform a preprocessing step, first by removing the commas from the text body, then uploading the data into a database. We use a Structured Query Language Platform, MySQL. MySQL is an open-source relational database management system (https://www.mysql.com/). We load all the fetched data to MySQL using 'Load Data' statement which reads rows from text or comma separated files (CSV) into SQL tables. The following Figure shows a screenshot of part of the PATH table from the MySQL database management system:

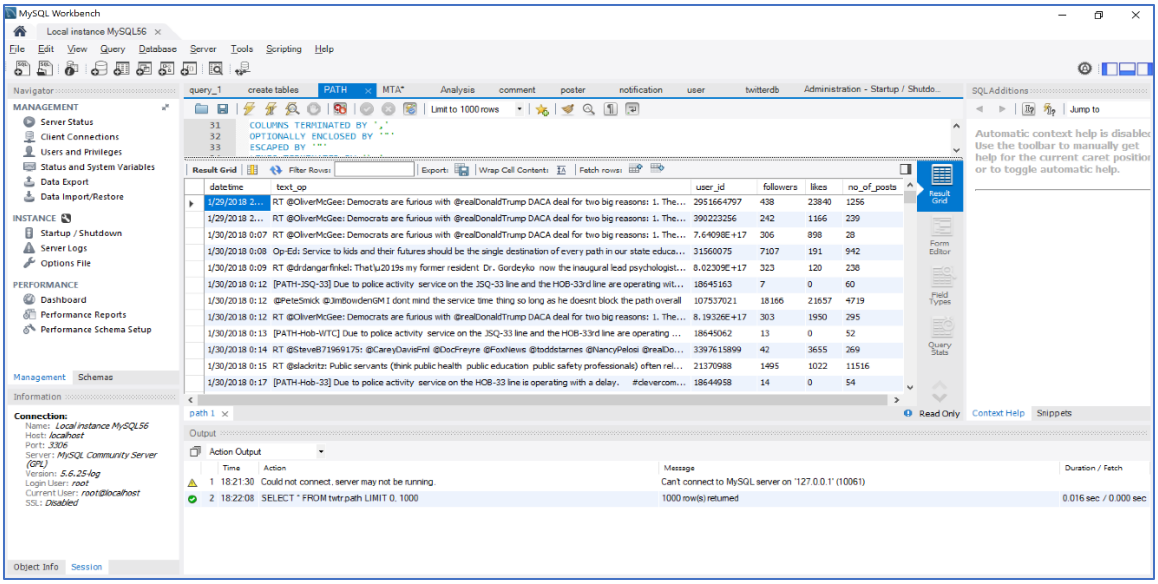


Figure 28: The PATH Table Stored in MySQL Workbench

Figure 29 below shows the three tables stored in our database management system (DBMS):

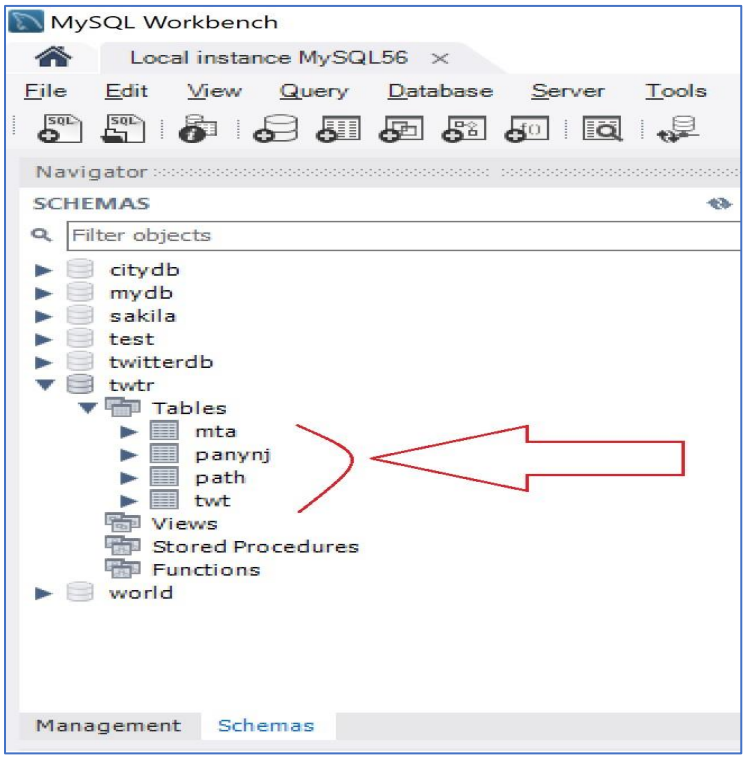


Figure 29: The PANYNJ, MTA, PATH Tables Stored in Our Database

After the data aggregation and preprocessing during the period from 1/29/2018 through 10/19/2018, the intermediate datasets consist of the following (Table 17 below):

Table Name	**# of Tweets**	**Aggregation Period**
PANYNJ	101,634	1/29/2018 – 10/19/2018
MTA	541,542	1/29/2018 – 10/19/2018
PATH	106,222	1/29/2018 – 10/19/2018
Total	**749,398**	1/29/2018 – 10/19/2018

Table 17: The PANYNJ, MTA, and PATH Tables

The total number of records means that the number of tweets or retweets that include at least one of the key terms mentioned previously for each of the agencies.

5. The Financial Industry Business Ontology (FIBO) term search: After data collection, aggregation and preprocessing, we search the databases for data structures which fill the slots of the frame for government bonds developed from the FIBO ontology. We use the framework derived in the next section to perform the search. We search both on individual tweets and threads.
6. Construct validity test: After collecting all the tweets related to the bond information of the two agencies, we compare the results to a naïve keyword search of the database. This indicates whether the semantics provided by the FIBO SMEs is superior to using simple keyword searches and is a form of model validation for our methodology. Our metric is the number of false positives of each classification method. This serves as a test of the construct validity of using FIBO terms as opposed to naïve keywords.

7. Design of the Descriptive Statistics Algorithm: In this step, we collect the matches for all the terms in the FIBO frame, including synonyms and near synonyms, for the MTA and PANYNJ twitter streams. This allows us to see which slots and terms are relevant for this application. This step results in the creation of two generic classifiers: one derived from the MTA stream and the other from the PANYNJ stream. We then "tune" these generic classifiers to see if we could improve their performance in terms of correct classification and the false positives metric.
8. Tuning the FIBO classifiers: In this step, we use each of the generic classifiers to see whether we could improve its results. We did this by testing whether classifiers built from combinations of several frames perform better than the generic classifiers. Since individual tweets are very short, we do not expect to improve performance. Our expectations end up being correct.
9. Using the tuned classifiers to classify other twitter feeds: In our final test, we use the tuned MTA classifier to classify the PANYNJ twitter stream and the tuned PANYNJ classifier to classify the MTA twitter stream. We then use each of the tuned classifiers to classify the PATH twitter stream.
10. Evaluation of Final Results: In this final step, we evaluate results of the MTA and PANYNJ classifiers. With such evaluation of final results, we provide a proof of concept for using FIBO frames in text mining as constituent semantic structures.

Steps 1 through 5 above are followed to produce the three twitter streams (MTA, PANYNJ and PATH). The remaining part of the paper discusses steps 6 through 10 in

additional detail and then concludes with parting thoughts and suggestions for future research.

4.4.2 INITIAL TESTING: CONSTRUCT THE VALIDITY TEST

The following tables show an illustration of the findings between FIBO terms search and the naïve terms search for the three Twitter streams: MTA, PANYNJ, and PATH. The results only show the application of FIBO concepts, not the synonyms and near synonyms. Adding the latter would only improve the results. False positives are tweets which were initially classified as being relevant but are not. An example would be if someone tweets on the MTA feed about her dog named "Revenue."

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
280	138	125	20	16%

Table 18: MTA FIBO Terms Search

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP	Ratio of # of Tweets of Naïve to FIBO
85	36	41	18	43.9%	30.4%

Table 19: MTA Naïve Terms Search

The Port Authority of New York and New Jersey (PANYNJ) Tweets:

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
117	61	54	1	1.9%

Table 20: PANYNJ FIBO Terms Search

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP	Ratio of # of Tweets of Naïve to FIBO
31	7	22	4	18.2%	26.5%

Table 21: PANYNJ Naïve Terms Search

The Port Authority Trans-Hudson (PATH) Tweets:

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
48	14	34	21	61.8%

Table 22: PATH FIBO Terms Search

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP	Ratio of # of Tweets of Naïve to FIBO
2	1	1	1	100%	4.2%

Table 23: PATH Naïve Terms Search

In our initial validation above (Tables 18-23) we compare the results between the FIBO terms search which is present in table 16 (excludes the synonyms) and the naïve terms search. One can clearly notice that the FIBO terms search is returning more records compared to naïve terms search in all datasets. For instance, looking at the MTA dataset,

one can notice a total of 280 records retrieved using FIBO terms search compared to 85 records using naïve terms search. Also, the percentage of false positives is significantly higher in naïve terms search compared to FIBO search with a total of 43.9% compared to 16% respectively. Similarly, looking at PANYNJ dataset, we find a 26.5% total ratio of the number of tweets of naïve to FIBO terms searches which is significantly lower. In addition, we also find that there is a low percentage of false positives, 1.9%, using FIBO compared to 18.2% of false positives using naïve terms search. In the PATH tables (tables 22-23), we also notice a higher number of records retrieved using FIBO terms search compared to naïve terms search with less false positives; however, 61.8% of false positives using the FIBO search is still considered a high number but it is still lower than the naïve terms search (100% FP).

So, from completing the initial validation we can clearly see that by using FIBO terms search, we can retrieve more records compared to naïve terms search. This is an indication that the SMEs' expertise has been captured in the FIBO ontology and was successfully transferred into our methodology. Additionally, in two out of the three cases, we observe a lower percentage of false positives using the FIBO terms search. In the next section, we will further classify the tweets to see if we can obtain more accurate results by getting the synonyms and apply the frame and slots methodology. Please note that the percentage of false positives is calculated based on the unique number of tweets.

4.4.3 DESIGN OF THE IMPLEMENTED SYSTEM ON A REAL TWITTER FEED

In our methodology, we derive the following frame and slots terms and their synonyms from the FIBO ontology Municipal Bonds Full (Table 24 below):

Term	**Synonym**	**Near / Broader Synonym Or Role-performing Item**
Government Issued Bond	Sovereign Bond, Treasury Bond	Government Debt
Municipal Security	Municipal Debt Instrument	Muni
Municipal Debt Issuer	Muni Issuer,	Issuer, Municipality
Municipal Bond	Muni Bond, Muni	
Debt Obligor	Owing Party, Borrower	Obligor
Funds usage	Funds Purpose, Disbursement Purpose	Loan Purpose, Credit Facility Purpose, Credit Purpose
Municipal Bond Capital Type	Muni Capital Type	Capital Type
Municipal Bond Refund Terms	Muni Refund Terms,	Refund Terms
Municipal Trustee	Muni Trustee	Trustee,
Ad valorem tax provision		Property Tax Provision, Real Property Tax Provision, Sales Tax Provision
Municipal Bond Type	N/A	
Build America	Build America Bond	
Tax Allocation	Tax Allocation Bond	
Special Tax	Special Tax Bond	
Special Obligation	Special Obligation Bond	
General Obligation	General Obligation Bond	
Revenue	Revenue Bond	
Special Assessment	Special Assessment Bond	

Table 24: Frame and Slots Terms and Their Synonyms from the FIBO Ontology Municipal Bonds Full

For most terms, we have added synonyms that refer to broader topics, since these words may be used in natural language (where the context is implicit) to refer to these concepts. These are synonyms and near synonyms in the FIBO specification. Since the FIBO terms, synonyms and near synonyms are all developed by subject matter experts in the financial industry; we expect that searches which include a broader selection of terms will find more tweets which discuss municipal bonds. We anticipate that very few tweets would discuss municipal bonds explicitly and that the public will tweet more often about topics relating to bonds and their FIBO concepts and synonyms/near synonyms.

4.5 CONSTRUCTION OF THE FRAME-BASED SYSTEM

A frame consists of a set of slots which are filled by values, procedures, or links to other frames. We need to formalize the municipal bonds-type frames taken from FIBO as shown in Table 24 above.

4.5.1 DESIGN OF THE DESCRIPTIVE STATISTICS ALGORITHM

For our design methodology, we perform multiple tests aiming at finding what best constitutes a frame, that is, to tune the results given which slots and synonyms are best at classifying the Twitter feed. The frame and slots terms and their synonyms and near synonyms are taken from Table 24 above which is derived from the FIBO Ontology Municipal Bonds Full. We assume the slots are represented as they appear in Table 25

(note the partial representation of Table 24 for illustration purposes on how to pick the slots):

Concept	**Synonym**	**Near / Broader Synonym Or Role-performing Item**
$S_{1,1}$ = Government Issued Bond	$S_{1,2}$ = Sovereign Bond, Treasury Bond	$S_{1,3}$ = Government Debt
$S_{2,1}$ = Municipal Security	$S_{2,2}$ = Municipal Debt Instrument	$S_{2,3}$ = Muni
$S_{3,1}$ = Municipal Debt Issuer	$S_{3,2}$ = Muni Issuer	$S_{3,3}$ = Issuer, Municipality
...	...	...
$S_{18,1}$ = Special Assessment	$S_{18,2}$ = Special Assessment Bond	

Table 25: Illustration of The Slots of The Frame and The Synonyms from FIBO Ontology Municipal Bonds Full

Definition of the Text Terms:
The set of text terms would follow similar logic but by adding additional dimension to include terms from Column 1 + Column 2 + Column 3 of table 12 as follows:

$T = S_{1,1} \vee S_{1,2} \vee S_{1,3} \vee S_{2,1} \vee S_{2,2} \vee S_{2,3} \vee S_{3,1} \vee S_{3,2} \vee S_{3,3} \vee S_{4,1} \vee S_{4,2} \vee S_{4,3} \ldots S_{18,1} \vee S_{18,2} \vee S_{18,3}$

SQL Query:

```
SELECT * FROM TWTR.(PANYNJ, PATH, MTA) WHERE text_op LIKE '% S1,1 %'
OR text_op LIKE '% S1,2 %'
OR text_op LIKE '% S1,3 %'
OR text_op LIKE '% S2,1 %'
OR text_op LIKE '% S2,2 %'
OR text_op LIKE '% S2,3 %'
OR text_op LIKE '% S3,1 %'
OR text_op LIKE '% S3,2 %'
OR text_op LIKE '% S3,3 %'
```

OR text_op LIKE '% $S_{4,1}$ %'
OR text_op LIKE '% $S_{4,2}$ %'
OR text_op LIKE '% $S_{4,3}$ %'
OR ...
OR text_op LIKE '% $S_{18,1}$ %'
OR text_op LIKE '% $S_{18,2}$ %';
OR text_op LIKE '% $S_{18,3}$ %'

The definitions above are part of our design methodology for frame construction. In these definitions, we use the logical condition ‘OR’ or ‘V’ for the slots. So, the question is what constitutes or best constitutes a frame? In order to explore whether or not all slots need to be included in the search, we developed descriptive statistics in regard to each term on Table 24’s frequency of occurrence. We also developed statistics regarding the frequency of the slots being filled, excluding retweets. Based on these descriptive statistics, we proceeded by developing an efficient method for extracting frames from within the Twitter stream. In this case, efficiency is defined by the number of slots needed to best represent a frame.

4.5.2 DESCRIPTIVE STATISTICS

In this section, we perform the descriptive statistics on our population. We use the three tables which are the MTA, PANYNJ, and PATH. The following tables show the frequency of the Frames and Slots Terms and their Synonyms from the FIBO Ontology Municipal Bonds Full for the system on our population (real Twitter feed) for the MTA, the PANYNJ, and the PATH datasets:

FIBO Concepts	# of Tweets	FIBO Synonyms	# of Tweets	Contextualized Synonym Or Role-performing Item	# of Tweets	Frequency of tweets row-wise
Government Issued Bond	0	Sovereign Bond Treasury Bond	0 0	Government Debt	0	0
Municipal Security	0	Municipal Debt Instrument	0	N/A	N/A	0
Municipal Debt Issuer municipal dept	0 1	Muni Issuer	0	Issuer Municipality	0 3	0 0
Municipal Bond	0	Muni Bond Muni	3 10	N/A	N/A	0 0
Debt Obligor	0	Owing Party Borrower	0 0	Obligor	0	0 0
Funds usage	0	Funds Purpose Disbursement Purpose	0 0	Loan Purpose Credit Facility Purpose Credit Purpose	0 0 0	0 0 0
Municipal Bond Capital Type	0	Muni Capital Type	0	Capital Type	0	0
Municipal Bond Refund Terms	0	Muni Refund Terms	0	Refund Terms	0	0
Municipal Trustee	0	Muni Trustee	0	Trustee	5	0
Ad valorem tax provision	0	N/A	N/A	Property Tax Provision Real Property Tax Provision Sales Tax Provision	144 0 21	0 0 0
Municipal Bond Type	0	N/A	N/A	N/A	N/A	
Build America	0	Build America Bond	0	N/A	N/A	0
Tax Allocation	0	Tax Allocation Bond	0	N/A	N/A	0
Special Tax	1	Special Tax Bond	0	N/A	N/A	0
Special Obligation	0	Special Obligation Bond	0	N/A	N/A	0
General Obligation	0	General Obligation Bond	0	N/A	N/A	0
Revenue	253	Revenue Bond	2	N/A	N/A	2
Special Assessment	0	Special Assessment Bond	0	N/A	N/A	0

Table 26: Frequency of the Frames and Slots Terms and their Synonyms from the FIBO Ontology Municipal Bonds Full for The MTA Dataset

FIBO Concepts	# of Tweets	FIBO Synonyms	# of Tweets	Contextualized Synonym Or Role-performing Item	# of Tweets	Frequency of tweets row-wise
Government Issued Bond	0	Sovereign Bond Treasury Bond	0 0	Government Debt	0	0
Municipal Security	0	Municipal Debt Instrument	0	N/A	N/A	0
Municipal Debt Issuer municipal debt	0 0	Muni Issuer	0	Issuer Municipality	0 8	0 0
Municipal Bond	2	Muni Bond Muni	0 0	N/A	N/A	0
Debt Obligor	0	Owing Party Borrower	0 0	Obligor	0	0 0
Funds usage	0	Funds Purpose, Disbursement Purpose	0 0	Loan Purpose Credit Facility Purpose Credit Purpose	0 0 0	0 0 0
Municipal Bond Capital Type	0	Muni Capital Type	0	Capital Type	0	0
Municipal Bond Refund Terms	0	Muni Refund Terms	0	Refund Terms	0	0
Municipal Trustee	0	Muni Trustee	0	Trustee	8	0
Ad valorem tax provision	0	N/A	N/A	Property Tax Provision Real Property Tax Provision Sales Tax Provision	18 0 0	0 0 0
Municipal Bond Type	2	N/A	N/A	N/A	N/A	2
Build America	0	Build America Bond	0	N/A	N/A	0
Tax Allocation	0	Tax Allocation Bond	0	N/A	N/A	0
Special Tax	3	Special Tax Bond	0	N/A	N/A	0
Special Obligation	0	Special Obligation Bond	0	N/A	N/A	0
General Obligation	0	General Obligation Bond	0	N/A	N/A	0
Revenue	111	Revenue Bond	7	N/A	N/A	7

Special Assessment	0	Special Assessment Bond	0	N/A	N/A	0

Table 27: Frequency of the Frames and Slots Terms and their Synonyms from the FIBO Ontology Municipal Bonds Full for The PANYNJ Dataset

FIBO Concepts	**# of Tweets**	**FIBO Synonyms**	**# of Tweets**	**Contextualized Synonym Or Role-performing Item**	**# of Tweets**	**Frequency of tweets row-wise**
Government Issued Bond	0	Sovereign Bond	0	Government Debt	2	0
		Treasury Bond	0			0
Municipal Security	0	Municipal Debt Instrument	0	N/A	N/A	0
Municipal Debt Issuer	0	Muni Issuer	0	Issuer	0	0
municipal dept	0			Municipality	2	0
Municipal Bond	0	Muni Bond	0	N/A	N/A	0
		Muni	1			0
Debt Obligor	0	Owing Party	0	Obligor	0	0
		Borrower	0			0
Funds usage	0	Funds Purpose	0	Loan Purpose	0	0
				Credit Facility Purpose	0	0
		Disbursement Purpose	0	Credit Purpose	0	0
Municipal Bond Capital Type	0	Muni Capital Type	0	Capital Type	0	0
Municipal Bond Refund Terms	0	Muni Refund Terms	0	Refund Terms	0	0
Municipal Trustee	0	Muni Trustee	0	Trustee	8	0
Ad valorem tax provision	0	N/A	N/A	Property Tax Provision	1	0
				Real Property Tax Provision	0	0
				Sales Tax Provision	0	0
Municipal Bond Type	0	N/A	N/A	N/A	N/A	0
Build America	0	Build America Bond	0	N/A	N/A	0
Tax Allocation	0	Tax Allocation Bond	0	N/A	N/A	

Special Tax	0	Special Tax Bond	0	N/A	N/A	0
Special Obligation	0	Special Obligation Bond	0	N/A	N/A	0
General Obligation	0	General Obligation Bond	0	N/A	N/A	0
Revenue	48	Revenue Bond	0	N/A	N/A	0
Special Assessment	0	Special Assessment Bond	0	N/A	N/A	

Table 28: Frequency of the Frames and Slots Terms and their Synonyms from the FIBO Ontology Municipal Bonds Full for The PATH Dataset

4.5.3 CONSTRUCTION OF OUR FIBO CLASSIFIERS

After performing the descriptive statistics on the MTA, PANYNJ and the PATH datasets, we find that of the eighteen FIBO slots we have, only six slots are filled by at least one value. From this, we construct generic classifiers which only use FIBO terms for which we found results in the descriptive analysis. From the results in Tables 26 and 27, the two generic classifiers are: MTA: Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond; PANYNJ: Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond. Please note that we decide to discard the PATH table going forward since it returns only a few values of the frames and slots.

We now take the slots with more than one value and see if we can get better results (fewer false positives) when we tune the system by requiring that multiple slots are used to classify a tweet as being more specifically about bonds. We construct classifiers from the descriptive statistics by testing pairwise combinations of the slots with their synonyms and

near synonyms for slots having more than one match. Since an AND condition can at most only include the already tested OR conditions, we only test the pairwise comparisons. If our results had been different, we would have continued to test triples, etc. Given the twitter streams in this study, this proves unnecessary. The two sets of tuned classifiers based on the descriptive statistics above are as follows.

MTA

1. (Property Tax OR Sales Tax Provision) AND Trustee
2. (Property Tax OR Sales Tax Provision) AND Municipality
3. (Property Tax OR Sales Tax Provision) AND (Muni Bond OR Muni)
4. (Property Tax OR Sales Tax Provision) AND (Revenue OR Revenue Bond)
5. (Revenue OR Revenue Bond) AND Trustee
6. (Revenue OR Revenue Bond) AND Municipality
7. (Revenue OR Revenue Bond) AND (Muni Bond OR Muni)

PANYNJ

1. (Revenue OR Revenue Bond) AND Municipality
2. (Revenue OR Revenue Bond) AND Municipal Bond
3. (Revenue OR Revenue Bond) AND Trustee
4. (Revenue OR Revenue Bond) AND Property Tax
5. (Revenue OR Revenue Bond) AND Municipal Bond Type
6. (Revenue OR Revenue Bond) AND Special Tax

Tuning the Results

We found the following classification results during tuning:

MTA

1. (Property Tax OR Sales Tax Provision) AND Trustee = 0
2. (Property Tax OR Sales Tax Provision) AND Municipality = 0
3. (Property Tax OR Sales Tax Provision) AND (Muni Bond OR Muni) = 0
4. (Property Tax OR Sales Tax Provision) AND (Revenue OR Revenue Bond) = 12, Zero false positive. 100% true positive.
5. (Revenue OR Revenue Bond) AND Trustee = 0
6. (Revenue OR Revenue Bond) AND Municipality = 0
7. (Revenue OR Revenue Bond) AND (Muni Bond OR Muni) = 0

PANYNJ

1. (Revenue OR Revenue Bond) AND Municipality = 0
2. (Revenue OR Revenue Bond) AND Municipal Bond = 0
3. (Revenue OR Revenue Bond) AND Trustee = 0
4. (Revenue OR Revenue Bond) AND Property Tax = 0
5. (Revenue OR Revenue Bond) AND Municipal Bond Type = 0
6. (Revenue OR Revenue Bond) AND Special Tax = 0

Therefore, we can derive the best classifiers as follows:

MTA:

In Table 29 below we show Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond (the generic classifier from Table 26).

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
423	256	167	20	12%

Table 29: Results from the Best MTA Classifier

Compare this with the only other candidate for MTA: (Property Tax OR Sales Tax Provision) AND (Revenue OR Revenue Bond) = 12, FP 0%. If we choose this, #4, we lose 147 (true positives) – 12 = 135 Tweets which are false negatives for the #4 classifier. This is too much loss of information for a small gain in precision.

PANYNJ:

We demonstrate in Table 30 Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond (the generic classifier from Table 27).

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
148	75	73	13	17.8%

Table 30: Results from the Best PANYNJ Classifier

This has 148 classifications with a 17.8% FP rate. There are no other candidates here, so we choose the OR classifier above.

Using the Tuned Classifiers to Classify Other Twitter Feeds

In Table 31 below, we use the best MTA classifier to classify the whole PANYNJ population which is the following:

“Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond”.

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
148	76	69	3	4.3%

Table 31: Results of Classifying PANYNJ Data with the MTA Best Classifier

Then we use the best PANYNJ classifier to classify the whole MTA population which is the following:

“Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond”.

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
418	268	150	7	4.7%

Table 32: Results of Classifying MTA Data with the Best PANYNJ Classifier

4.6 EVALUATION OF THE FINAL RESULTS AND THE METHODOLOGY

Since our objective is to be able to scan any type of big data source and apply a scientific methodology which would enable us to better understand and extract some knowledge from messy, unstructured data, we find that by using the publicly available FIBO ontology as the basis for our frames and slots, we are able to actually extract related meaningful knowledge. Our data results show that by organizing the unstructured data in this more structured way, we can extract more related tweets from Twitter with fewer false

positives. The formal concepts and synonyms of FIBO provide structure and organization to the messy Twitter media, and this organization allows it to be joined to structured financial data that share the same concepts.

By applying different classifiers and tuning the search results on the last section, we discover that when using the best MTA classifier to classify the PANYNJ dataset, the results show the same number of tweets of 148 but a significant increase in the precision - 4.3% using the best MTA classifier compared to 17.8% false positives. Similarly, by using the best PANYNJ classifier to classify the MTA dataset, we get almost the same number of tweets of slightly above 400 with an increase in precision of 4.7% false positives compared to 12%. Based on our current tests to date, it appears that both of the "best" classifiers perform similarly on both data sets. However, the MTA classifier demonstrates the more drastic improvement of false positive reduction when applied to the PANYNJ twitter feeds (13.5%) versus that of PANYNJ classifiers applied to MTA data (7.3%). It would be interesting to see if these results hold over time or if we might find a degradation or improvement of classifier performance.

Generally, there exist some limitations. For instance, in such environmental scanning, we need more rich and voluminous datasets. We collected ten months of data, but subsequently, feel that additional tests may require a longer time frame. Like any big data project, the longer the period of data collection, the more instances we could extract, analyze, and subsequently achieve better accuracy. Furthermore, regarding the PATH results above, many of the Tweets in the original feed created for PATH pick up the term "path" in the Tweet itself. This leads to phrases such as "revenue path" being classified

and, consequently, often a false positive. We need to consider this in any future refinements of the methodology.

4.7 TESTING OUR METHODOLOGY ON A NEW POPULATION

On this section, we test our methodology using new data collected from October 22, 2018 to March 10, 2019. The table below shows the number of records each table contain:

Tables	Raw data	Cleaned data
MTA	323,413	322,110
PANYNJ	6,956	6,904

Table 33: Testing Data - The MTA and PANYNJ Tables

4.7.1 FIBO VS. NAÏVE TERMS SEARCH

Methodology	# of Tweets	Retweets	Unique # of Tweets
FIBO Terms Search	448	230	218
Naïve Terms Search	77	36	41

Table 34: MTA Table - FIBO & Naïve Terms Search

Methodology	# of Tweets	Retweets	Unique # of Tweets
FIBO Terms Search	31	27	4
Naïve Terms Search	2	0	2

Table 35: PANYNJ Table - FIBO & Naïve Terms Search

From the tables above (MTA, PANYNJ), clearly, we see that using the FIBO ontology would fetch more records than by using the naïve key-word search.

4.7.2 TESTING OUR TUNED CLASSIFIERS USING THE NEW DATASET

MTA

In the following table, we use the previously introduced best MTA classifier to classify the new MTA population. The classifier is as follows:
"Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond".

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
527	272	255	5	1.9%

Table 36: Results of Classifying MTA Test Data Using the MTA Best Classifier

Furthermore, we use the best PANYNJ classifier to classify the new MTA population which is the following:

"Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond".

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
506	270	236	2	<1% (0.8%)

Table 37: Results of Classifying MTA Test Data Using the PANYNJ Best Classifier

The results above show that by testing our tuned classifiers (i.e., best classifiers) in a new population we get similar results, or even more accurate output. Using the best MTA classifier to classify the new MTA dataset, we see an increase in precision of 1.9% compared to 12% which is the percentage of false positives. Moreover, the total number of records captured is 527 compared to 423 records. Similarly, by applying the best PANYNJ classifier on the new population (i.e., the new MTA dataset), we see an improvement in precision that is 0.8% compared to 4.7% of false positives. We can also notice the increased number of records fetched using our classifiers compared to the training phase. The total number of records went up to slightly above 500 in both classifiers compared to around 400 records even though the duration of data collection is almost half the time in the testing-data-collection period compared to the initial datasets (10-months compared to 5-months). We think this is because of the controversial issues that occur during the testing phase such as Amazon's Co. backing up from opening its new HQ2 in Long Island City, NY, and how this could cause a huge loss in tax revenue that could be used to support NYC infrastructure (e.g., MTA). Also, the MTA board approved fare and toll increases on subways, railroads, buses and tolled crossings on Feb 27, 2019.

These testing results prove the ability and accuracy of our proposed methodology to be generalized and used in other FIBO ontologies and their concepts or even to be extended and applied using different data outlets. During the testing period, there exist some debatable subjects in NYC related to the context of new sources of revenue for the city, fixing the infrastructure, etc.

PANYNJ

The following table shows the results of using the best MTA classifier to classify the new PANYNJ population. The classifier is as follows:

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
31	27	4	0	0%

Table 38: Results of Classifying PANYNJ Test Data Using the MTA Best Classifier

Furthermore, we use the best PANYNJ classifier to classify the new PANYNJ population which is the following:

“Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond”.

# of Tweets	Retweets	Unique # of Tweets	False Positives	% of FP
31	27	4	0	0%

Table 39: Results of Classifying PANYNJ Test Data Using the PANYNJ Best Classifier

The results above show similar findings compared to our previous tests. That is using the tuned classifiers (i.e., best MTA and PANYNJ classifiers) to classify other twitter feeds such that in Tables 38 & 39 (The PANYNJ new population). We find that we almost receive similar percentages of false positives compared to the testing of our initial population.

4.8 CONCLUSIONS

This study was structured to capture tweets about past and pending economic events regarding PANYNJ and the MTA using a methodology developed from a slot and frame structure and the FIBO ontology. During the time period of this study, PANYNJ bond series were rated as stable Aa3 by Moody's, and the Port Authority Board approved the application to raise funds to upgrade one of its airports (Mid-Hudson News Network, 2018). We anticipated that these economic events and other activities would provide sufficient Twitter data to demonstrate this type of methodology for environmental scanning.

Additionally, we anticipate that utilizing a methodology developed from a slot and frame structure and the FIBO ontology could facilitate the joining of "messy" unstructured data, such as that found in Twitter, to more structured financial data sharing the same FIBO concepts. The methodology of using the FIBO ontology in this manner should create a fuzzy match between data types that are ordinarily challenging and time-consuming to combine.

Although this study has been carefully grounded in accounting ontology theory, we should mention several assumptions made in our study that are typical for most examinations of social media. First, there is a mistaken assumption that Twitter feeds represent the true population. However, Twitter feeds only represent the tweets of the population that choose to actively interact on Twitter. Secondly, there are many subscribers of Twitter who do not actively post but may retweet and or just read tweets. Basically, most Twitter studies do not capture the tweets of the broad population, but only that of active Twitter participants. Thirdly, another assumption is that tweets represent the participant's actual meanings (semantic state). Twitter posts only display what participants elect to post, and as such could be abbreviated and/or modified. Some participants may feel comfortable posting in a manner similar to an unstructured "stream of consciousness" while others might post in a more measured and constrained manner. Or other tweets may be more commercial or promotional in purpose. These issues point to the potential benefits of using a structured ontology such as that of FIBO for understanding a broad range of tweets that are of a more financial nature.

Furthermore, Twitter might not be the medium with which users typically would choose to explicitly discuss municipal bond matters. They actually might tweet more frequently about concepts related to municipal bonds, such as "revenue". It is this conceptual context which allows the FIBO Conceptual Ontology method to shine. Using the FIBO Conceptual Ontology, it is possible to capture both the directly and indirectly related tweets. Many may argue that a simple keyword search of "municipal bond revenue" or other terms would suffice. However, such a search would not classify all the utterances and concepts in the Twitter data that are relevant indirectly and that are captured with the

FIBO ontology frames and slots. In fact, many users tweet about revenue in the context of their MTA/PANYNJ experiences and observations and are not thinking about bonds at all, yet the tweets relate to the financial health of the bond-issuing agency. Here are a few of such conceptually related tweets:

- @nyc311 Why are @MTA buses incapable of accepting dollar bills? Seems to me you are losing out on easy revenue!
- And of course there are not ticket collectors in sight. Riders suffer revenue suffers @NYGovCuomo @SenSchumer https:\/\/t.co\/S1Jpsd8CE3

So even though the initial FIBO concept "revenue" does not relate exclusively to bonds (in the tweets captured it broadly relates to either the MTA, PANYNJ, NJPATH), it still relates indirectly to bonds in that revenues determine the overall financial condition of the bond-releasing agency and its bond performance and could potentially influence bond buying and selling behavior. But it is not expected that users would often tweet about municipal bonds unless this topic is currently in the news. As mentioned earlier in the paper, the authors anticipated that the recent bond release would provide more instances for bond-related tweets than would normally be expected. Absent the recent bond release; there might have been zero explicit tweets about municipal bonds.

Additionally, there may be room for other FIBO ontologies to be applied to these Twitter data streams. Since many of the relevant tweets, particularly at the broad concept level of "revenue" could relate to municipal budgets or other financial contexts, it is hoped that interested parties could apply the methodology demonstrated here using other FIBO ontologies pertaining to other firms/municipalities. The method demonstrated here in this

research should be generalizable across other FIBO ontologies and their concepts. This research demonstrates the potential for a FIBO ontology classifier methodology, when applied to messy social media feeds, to facilitate its fuzzy match to the relevant numbers in financial reports that share the same FIBO concepts. To illustrate, a bond analyst examining the revenue streams of the MTA to arrive at a bond-buying decision may want to include a "real-world" perspective of its revenues, such as that supplied by Twitter, as one of many sources of insightful information for an environmental scan. This analyst can then relate all "revenue" relevant tweets as identified by this FIBO ontology method to the revenue number in the financial report. In this case, the tweets and the number would share the same FIBO classifying concepts.

Secondly, this methodology could be applied to other sources of social media, such as Facebook postings and Instagram. It is anticipated that these data sources may be similarly challenging as Twitter to extract, classify, and match. Other more structured forms of social media such as blogs, articles, and reports may also provide rich resources for FIBO ontology and conceptually based classification and extraction.

To summarize, in this paper we utilize a text mining implementation of Financial Industry Business Ontology (FIBO) to extract insights from Twitter regarding financial and budget information in the public sector, namely the collective public-private agencies of PANYNJ and the MTA. This research initiative is approached by developing a methodology to classify tweets as being about financial bonds or not related to financial bonds. We apply a frame and slot methodology from the artificial intelligence literature to operationalize the methodology using the FIBO ontology in the public sector/municipalities business context. FIBO provides standards for defining the facts,

terms, and relationships associated with financial concepts. One contribution of this paper is that it is the first to recognize that the FIBO structure provides a grammar of financial concepts which can be used to classify social media. We show that this grammar can be used to mine semantic meaning from unstructured textual data. Twitter streams are monitored and analyzed with frames derived from FIBO concepts. Using FIBO frames, constituent semantic structures can be uncovered to predict reactions to policies and programs and to perform other environmental scans more quickly than by following the same feeds manually. Using FIBO frames, we can capture more nuanced tweets that relate conceptually to the relevant financial data, a fuzzy match which might otherwise be overlooked in a less structured scanning approach.

This study contributes to accounting academic research in that it appears to be the first that applies a formal business ontology to unstructured social feeds, such as Twitter. This allows the tweets in the feed to be classified as either relevant to an environmental scan or not. We envision a situation where either firm management or their auditors are looking to scan the firm's environment for potential risk or additional insights. Social media provides a rich potential source of environmental signals but also carries a lot of noise. The FIBO ontology was built by consulting with domain experts whose financial expertise has been captured in the ontology. Therefore, using such ontologies facilitates the identification and understanding of the nuanced threads in Twitter that pertain to such financial issues as probable and/or pending municipal bond releases. During the time period of this study, PANYNJ bond series were rated as stable Aa3 by Moody's and the Port Authority Board approved the application to raise funds to upgrade one of its airports (Mid-Hudson News Network, 2018). It is anticipated that these announcements would

evoke a public reaction, and this study applies a formal accounting ontology to the processing of these social media extractions to facilitate and organize understanding. The authors anticipate that this methodology would be of interest to bond issuers, regulators, analysts, investors, and academics.

REFERENCES

Accounting Information Systems Elements," Journal of Information Systems (2017), vol. 31, no. 3, pp. 45 – 61.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

Campbell, G., C. Godian, E. Golden, A. Khoa, D. Peterson, E. Rankin, M. Reed, K. Rosenberg, R. Stoffers, and R. Todd. (2017)."Reforming the Port Authority of New York and New Jersey."https://wws.princeton.edu/sites/default/files/content/WWS%20591a%20Port%20Authority%202017.pdf, accessed on February 10, 2018.

Chae, B. K. (2015). Insights from hashtag# supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. International Journal of Production Economics, 165, 247-259.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. computer, 37(4), 50-56.

Financial Industry Business Ontology (FIBO), https://www.edmcouncil.org/financialbusiness.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464-472.

Hevner, A., March, S. T., Park, J. and Ram, S. 2004. Design Science in Information Systems Research. MIS Quarterly, 28(1), 75 – 105.

Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., ... & Xie, L. (2012). Social media use by government: From the routine to the critical. Government Information Quarterly, 29(4), 480-491.

Lassila, O., & McGuinness, D. (2001). The role of frame-based representation on the semantic web. Linköping Electronic Articles in Computer and Information Science, 6(5), 2001.

Mattessich, Richard, "Accounting Representation and the Onion Model of Reality: A Comparison of Baudrillard's Orders of Simulacra and his Hyperreality," Accounting, Organizations and Society (2003), vol. 28, no. 5, pp. 443 - 470.

Moffitt, K. and M.A. Vasarhelyi. (2013). AIS in an Age of Big Data. *Journal of Information Systems* Vol. 27, No. 2 pp. 1-19.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications, 40(10), 4241-4251.

Murthy, Uday and Guido L. Geets, "An REA Ontology-Based Model for Mapping Big Data to

O'Regan, G. (2018). The smartphone and social media. In World of Computing (pp. 257-265). Springer, Cham.

Omnicore (2019, January 06). Twitter by the Numbers: Stats, Demographics & Fun Facts. Retrieved February 20, 2019, from https://www.omnicoreagency.com/twitter-statistics/.

Pfeffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45 – 77.

Polycarpou, L. (2014). Study: Reforming the Port Authority and the MTA. .(http://blogs.ei.columbia.edu/2014/06/23/study-reforming-the-port-authority-and-the-mta/) accessed on February 10, 2018.

Smith, S.J. (2014) "It's Time to Kill the Port Authority of New York and New Jersey." (https://nextcity.org/daily/entry/its-time-to-kill-the-port-authority-of-new-york-and-new-jersey) accessed on February 10, 2018.

Statista. (2019). Twitter: Number of active users 2010-2018. Retrieved February 20, 2019, from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

Stutzman, Fred (April 11, 2007). "The 12-Minute Definitive Guide to Twitter" (https://web.archive.org/web/20080704074026/http://dev.aol.com/article/2007/04/definitive-guide-to-twitter) accessed on February 10, 2018.

Syed. A., K. Gillela, and C. Venugopal. (2013). The future revolution on Big Data. *International Journal of Advanced Research in Computer and Communication Engineering* 2 (6): 2446-2451.

Warren, J.D. Jr., Moffitt, K. C., and Byrnes, P. (2015). How Big Data Will Change Accounting. *Accounting Horizons*, Vol. 29, No. 2. Pp. 397-407.

Yaqub, U. (2018). Citizen centric stakeholder theory: sentiment and behavior analyses in social media (Doctoral dissertation, Rutgers University-Graduate School-Newark).

CHAPTER 5: CONCLUSION AND FUTURE RESEARCH

In this thesis, we have discussed three essays on open government data and data analytics. The first essay introduces a new model for effective and efficient open government data; the second study examines the use of visualization and cluster analysis techniques. It utilizes tools for visualizations, hierarchical and k-means clustering methods on governmental data in two case studies. The third essay focuses on text mining analytics of government-related financial data. Particularly, utilizing a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the two public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). The following summarizes the major findings and future research of the three essays.

During the past decade, the availability and accessibility of a large number of public datasets have increased significantly. This led to a great return in various domains. However, a lot of today's focus is on data dissemination, where data can be found almost everywhere. Whereas in order to achieve the success of open data as a model in the government domain, the focus should be shifted to the use cases and the quality of the published data. One of the major arguments nowadays regarding open data initiatives is that data is supply driven (Janssen et al., 2012). Most governments' open data initiatives from different countries lack standardizations in data, and are not providing application

and tools that facilitate the utilization of such data, and are not providing opportunities for users to give feedback or to interact with governments' officials.

The first essay compares the different financial reporting and auditing systems in the public sector between Brazil and Saudi Arabia. Also, the paper examines open government data initiatives among different countries with the focus of the Republic of Brazil and the Kingdom of Saudi Arabia open data portals. Moreover, it assesses the level of data transparency based on the definition of the open data model, and more importantly, the paper introduces a new model for effective and efficient open government data by adding additional dimensions to the open data concept when utilized by governments, which expands the open data definition to include its potential to encourage possible better decision making by governments.

The assessment of the proposed attributes for data transparency has been conducted using a sample of procurement contracts data available at the Republic of Brazil and Saudi Arabia's open data portals. We find that open government data initiatives in Saudi Arabia lack appropriate datasets, formats, analytical tools, and applications, and this could be improved dramatically to enhance government efficiency; whereas the Brazilian OGD initiatives lack applications. We also find that there is a gap between data analytics and applications for reporting and interacting with public officials. Open government data should be analyzable, and thus, the subsequent results should be actionable, for the evolution of better governments.

Saudi Arabia in one side is a developing country, but in another, it is developed such that Information and Communication Technologies (ICTs) are highly utilized; but more work is needed on openness and transparency. Brazil, on the other hand, is following

and implementing a transparent and open system when it comes to government data. The paper also suggests that public audit (armchair audit) could be of great use in Saudi Arabia where it has not been utilized yet. There exist some limitations such that more detailed theoretical corpus around the proposed dimensions' proposal is needed. Our long-term goal of this research is to develop a unified framework that applies our proposed 4-condition model for better government data transparency. Also, we need to address how the structure of Brazil and Saudi Arabia's governments, its public sector financial reporting and auditing systems could affect their open data initiative processes. Finally, more data and evidence are needed to support our proposed model and also by assessing other countries' open government data initiatives. Future research could be developing a unified framework that applies our proposed four-condition model for better governments' data transparency.

The second essay explores the literature for the use of emerging data mining techniques in auditing. In particular, cluster analysis techniques and visualization as supportive tools for auditors and practitioners. We have conducted two case studies. The first case study uses the U.S. states financial statements data. The second case study is utilizing the Volcker Alliance's Survey data results. The survey produced extensive information about how the different U.S. states scores on an annual basis on budgeting using five measures. On both case studies, we show how visualization and data clustering techniques could be used on governmental data and to help gain more information about financial statements budgeting. The cluster results also show that there are some similarities between the two methods, k-means and hierarchical, and this could give us an idea about our data quality. We have also used an assessment methodology to evaluate the

stability level of the clusters' results. In addition, we have now clear and unusual patterns and relationships to explore in greater depth.

The contribution of the second essay is two-fold. First, the two case studies presented in the second essay bring some advanced visualization and data mining techniques into the governmental domain, especially using financial statements and budgetary data. Second, the visualization and clustering results bring more opportunities for auditors and practitioners and to help get more insights into the data. Finally, we have used two clustering methodologies, k-means, and hierarchical clustering. However, there exist some limitations in the study such that we need to more evaluate the results of the first case study such that examining the clusters with external variables, and also more data is needed for the second case study, we have only used three years of data.

Future research could apply and compare additional data clustering techniques such as partition around medoids (PAM) clustering method. Also, by comparing the results of cluster analysis using external variables such as net population change and gross domestic product (GDP) growth. Also, the states that are grouped together on the same cluster could be further analyzed.

The third essay utilizes a text mining implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the two public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). We apply a frame and slot approach from the artificial intelligence literature to operationalize the FIBO ontology in a public sector/municipalities business context. FIBO is part of the Enterprise Data Management

Council (EDMC) and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts.

This study contributes to accounting academic research in that it appears to be the first that applies a formal business ontology to unstructured social media feeds, such as Twitter. This allows the tweets to be classified as either relevant to an environmental scan or not. We envision a situation where either firm management or their auditors are looking to scan the firm's environment for potential risk or additional insights. Social media provides a rich potential source of environmental signals but also carries a lot of noise. The FIBO ontology was built by consulting with domain experts whose financial expertise has been captured in the ontology. Therefore, using such ontologies facilitates the identification and understanding of the nuanced threads in Twitter that pertain to such financial issues as probable and/or pending municipal bond releases. During the time period of this study, PANYNJ bond series were rated as stable Aa3 by Moody's and the Port Authority Board approved the application to raise funds to upgrade one of its airports (Mid-Hudson News Network, 2018). It is anticipated that these announcements would evoke a public reaction, and this study applies a formal accounting ontology to the processing of these social media extractions to facilitate and organize understanding. We anticipate that this methodology would be of interest to bond issuers, regulators, analysts, investors, and academics.

In future research, this methodology could be applied to other sources of social media, such as Facebook postings, or Instagram. It is anticipated that these data sources may be similarly challenging as Twitter to extract, classify, and match. Other more structured forms of social media such as blogs, articles, and reports may also provide rich

resources for FIBO ontology and conceptually based classification and extraction. Also, future research could be in developing more complex frames (e.g., apply inheritance).