Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 12, Iss. 1, pp. 38-52, 2014

Using Clustering and Text Mining to Create a Reference Price Database*

Rommel Carvalho, Eduardo de Paiva, Henrique da Rocha, and Gilson Mendes
Department of Strategic Information (DIE)
Brazilian Office of the Comptroller General (CGU)
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro
Brasília – Distrito Federal – Brazil
{rommel.carvalho,eduardo.paiva,henrique.rocha,liborio}@cgu.gov.br

Resumo – Desde 2004, a Controladoria-Geral da União (CGU) tem publicado diversos dados relacionados a gastos do governo brasileiro no Portal da Transparência. Em 2010, a CGU começou a publicar diariamente as despesas diretas do Governo Federal. No entanto, inconsistências que prejudicam a transparência e prestação de contas foram encontradas nessa base de dados. Esse artigo apresenta como a CGU usa técnicas de agrupamento e mineração de texto para recuperar informações essenciais para a melhoria da prestação de contas do governo, incluindo o que foi comprado, o preço pago por item, o preço de referência por produto, etc. Essa análise permitiu que a CGU chegasse a algumas conclusões preliminares que são apresentadas nesse trabalho como forma de ilustrar os resultados da pesquisa. Finalmente, essa informação será eventualmente disponibilizada no Portal da Transparência, pertmitindo que todo cidadão possa ter acesso a informação de quanto o governo está realmente pagando, de uma forma geral, pelos produtos comprados. Dessa forma, é possível melhorar a prestação de contas e a transparência dos gastos públicos não apenas dentro do âmbito da CGU, como órgão de controle interno, mas também para todos os cidadãos brasileiros, que, no final, pagam a conta.

Palavras-chave – preço de referência, agrupamento, mineração de texto, gasto público, prestação de conta, transparência, empenho e compra governamental.

Abstract – Since 2004, Brazil's Office of the Comptroller General (CGU) has been publishing several data related to government expenditures in the Transparency Portal. In 2010, CGU started publishing daily every financial statement produced by the Federal Government. Nevertheless, inconsistencies which hinder accountability have been found in this data base. This paper presents how CGU uses clustering and text mining techniques to retrieve essential information for a good accountability, which includes what was bought, the price paid per item, a price reference per product, etc. This analysis has allowed CGU to draw some preliminary conclusions which are presented as a means to illustrate the research results. Finally, this information will eventually be incorporated in the Transparency Portal, allowing every citizen to understand how much the Government is really paying, in general, for products. Thus, improving social control and providing a solid accountability not only to CGU, as an internal control agency, but also to Brazil's citizens who, in the end, are the ones paying the bill.

Keywords – reference price, cluster, text mining, public expenditure, accountability, note of purchase, and government purchase.

1 Introduction

Economy and good management of public resources have always been priority in the Brazilian Government, as in any democracy, as well as the wish of its citizens. This wish has been reinforced by the public manifestations in 2013 born in the social media in Brazil.

However, in order to make good use of public resources, it is necessary to have a realistic budget planning. Unfortunately, even though all purchases are registered in a centralized system (SIASG[1], in Portuguese), there are problems with both quality of data entered by the user and the product classification currently available.

For instance, it is not rare to verify that fields like quantity and price per unity is incorrectly filled. Although the total price of the purchase is correctly informed by the user, he/she sometimes incorrectly enters 1 into the quantity, which makes the price per unit be the same as the total price of the purchase.

Moreover, some of the classifications available in the system are too general to specify a single product. For example, the code 113026 refers to the basic food basket. However, in some cases it is sometimes used to refer to purchases of the basket, as expected, and in others it refers to specific products from the basket, like rice, beans, etc.

*This paper is an extension version of the paper "Methodology for Creating the Brazilian Government Reference Price Database" presented in the X National Meeting on Artificial and Computational Intelligence (ENIAC 2013).

[1]The Integrated Administration and General Services System (SIASG) enables automated control actions and management of government procurement, which provided a significant improvement in the management of spending on cost and streamlining of procedures [1].

Therefore, the government manager is not able to answer simple questions like: How much does the government usually pays for a liter of regular gasoline? How much does a packet of A4 white paper cost? We need new chairs, what is the mean price paid by the government? The Reference Price Database discussed in this paper provides the answer to these and many other questions. Besides that, it allows the manager to compare its purchase performance with other agencies.

This paper addresses the methodology for creating a database of average price paid per product by the Brazilian Federal Government. The main goal is to have a price reference so that auditors as well as regular citizens can assess whether purchases made by the Government are overpriced or not. This is an essential tool to providing both accountability as well as transparency of the daily purchases made by the Brazilian Government.

The information used in order to compute the average price per product comes from the note of purchase commitment, which are public available as Open Government Data (OGD)[2] at the Brazilian Transparency Portal[3]. The note of purchase commitment, first stage in executing a public purchase, has a list of items that specify every product or service that is being committed. Every item has its description, price, quantity, among other information. This information was chosen because it is the most detailed information about a purchase available in the Brazilian Government database and also because every purchase made by the Federal Government must have this data entered into the SIASG system in order to be able to get the money to pay the supplier.

Figure 1 presents a note of purchase for water[4]. The most important part of the note of purchase used for computing the reference price in this research is highlighted in red just below the Detail of Expenditure (*Detalhamento do Gasto*, in Portuguese). Some of the problems encountered during this research can be seen in this note of purchase. For instance, the same category code 9873 (see the number in the end of each description presented just after the 'ITEM DE MATERIAL:') is used for both 500 ml water bottle and 20 l water container. Moreover, there is a purchase of 17,460.15625 20 l water container. However, it is not possible to buy 0.15625 of a container. Thus, we can suspect that this number might have been incorrectly informed. These problems are discussed in more detail and addressed in Section 2.

The main challenge in defining a reference price per product is how to identify which product is being described in the note of purchase commitment. The problem is that most of the relevant information is in the description text field. Fortunately, the description is semi-structured, which allows the extraction of a few extra information. The most relevant information which can be extracted is the category code (*código do material* in Portuguese).

The category code is a code used by the SIASG system to identify the product or service which is being acquired. Although important, this information is not precise enough. For instance, the code 21806 is used to identify simple batteries. However, there is no structured information which describes what type of battery it is (AA, AAA, C, D, etc).

One of the main contributions of this research is showing how to apply well known Data Mining and Text Mining techniques to tackle the problems previously described in order to allow the extraction of useful information from a dataset that is subject to errors that arise from common user mistakes (*e.g.*, entering incorrect values for numeric fields in the system). These techniques can be used to extract useful information from other governmental systems, for instance.

The paper is structured as follows. Section 2 describes the proposed methodology which relies on Data Mining techniques to identify the products in each purchase in order to obtain a reliable and precise reference price. Section 3 presents some statistical results obtained from the products analyzed, including confidence intervals of price average. Section 4 describes some patterns and preliminary conclusions found during the process of constructing the product reference price database. Finally, Section 5 draws some conclusions, describes the deployment plan of the constructed database, and presents some future work.

2 Methodology

As explained in Section 1 the main challenge when trying to define a reference price per product is to be able to correctly classify the products. Although a category code is available, in most cases, it is too broad thus not enough to pinpoint a specific product.

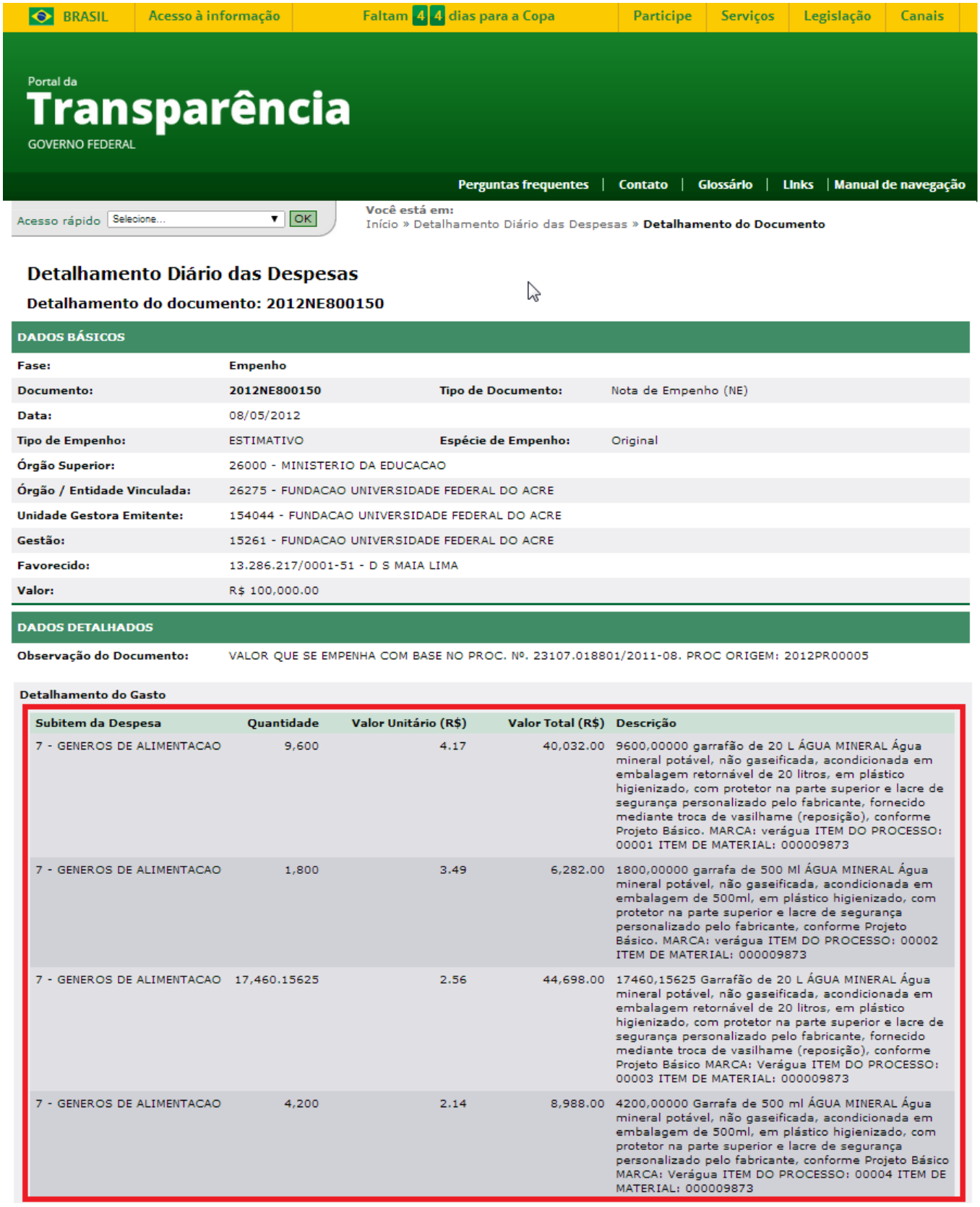
For instance, Figure 2 presents the result of searching for category codes for paper of size A4 and that has 75 g/m^2 (keywords 'papel A4 75' in Portuguese) in the system of category code (CATMAT[5], in Portuguese). As shown, there are 9 different codes for that type of paper. Some codes are more general, like the 347498 which refers to recycled

[2]Brazil has made a commitment via its action plan with the Open Government Partnership (OGP) as one of its founding governments. The OGP is a global effort to make governments better by providing more transparent, effective and accountable governments [2]. OGD plays a big part on making this possible, as it can be seen in most action plans, which focus their efforts on OGD initiatives [3].

[3]The Transparency Portal was created in November 2004 for the purpose of making it possible for public managers and citizens at large to follow up on the financial execution of all programs and actions of the Federal Government more easily. The Portal shows all data on the SIAFI's (Federal Government Integrated System for Financial Management) financial execution among other information [4]. The Portal can be accessed at http://www.portaldatransparencia.gov.br/.

[4]This note of purchase is available in the Transparency Portal at `http://www.portaldatransparencia.gov.br/despesasdiarias/empenho?documento=154044152612012NE800150`.

[5]This search was made at the website `http://www.comprasnet.gov.br/Livre/Catmat/conitemmat1.asp`.



BRASIL | Acesso à informação | Faltam 44 dias para a Copa | Participe | Serviços | Legislação | Canais

Portal da Transparência
GOVERNO FEDERAL

Perguntas frequentes | Contato | Glossário | Links | Manual de navegação

Acesso rápido Selecione... OK

Você está em:
Início » Detalhamento Diário das Despesas » **Detalhamento do Documento**

Detalhamento Diário das Despesas

Detalhamento do documento: 2012NE800150

DADOS BÁSICOS

Fase:	**Empenho**		
Documento:	**2012NE800150**	**Tipo de Documento:**	Nota de Empenho (NE)
Data:	08/05/2012		
Tipo de Empenho:	ESTIMATIVO	**Espécie de Empenho:**	Original
Órgão Superior:	26000 - MINISTERIO DA EDUCACAO		
Órgão / Entidade Vinculada:	26275 - FUNDACAO UNIVERSIDADE FEDERAL DO ACRE		
Unidade Gestora Emitente:	154044 - FUNDACAO UNIVERSIDADE FEDERAL DO ACRE		
Gestão:	15261 - FUNDACAO UNIVERSIDADE FEDERAL DO ACRE		
Favorecido:	13.286.217/0001-51 - D S MAIA LIMA		
Valor:	R$ 100,000.00		

DADOS DETALHADOS

Observação do Documento: VALOR QUE SE EMPENHA COM BASE NO PROC. Nº. 23107.018801/2011-08. PROC ORIGEM: 2012PR00005

Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	9,600	4.17	40,032.00	9600,00000 garrafão de 20 L ÁGUA MINERAL Água mineral potável, não gaseificada, acondicionada em embalagem retornável de 20 litros, em plástico higienizado, com protetor na parte superior e lacre de segurança personalizado pelo fabricante, fornecido mediante troca de vasilhame (reposição), conforme Projeto Básico. MARCA: verágua ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000009873
7 - GENEROS DE ALIMENTACAO	1,800	3.49	6,282.00	1800,00000 garrafa de 500 Ml ÁGUA MINERAL Água mineral potável, não gaseificada, acondicionada em embalagem de 500ml, em plástico higienizado, com protetor na parte superior e lacre de segurança personalizado pelo fabricante, conforme Projeto Básico. MARCA: verágua ITEM DO PROCESSO: 00002 ITEM DE MATERIAL: 000009873
7 - GENEROS DE ALIMENTACAO	17,460.15625	2.56	44,698.00	17460,15625 Garrafão de 20 L ÁGUA MINERAL Água mineral potável, não gaseificada, acondicionada em embalagem retornável de 20 litros, em plástico higienizado, com protetor na parte superior e lacre de segurança personalizado pelo fabricante, fornecido mediante troca de vasilhame (reposição), conforme Projeto Básico MARCA: Verágua ITEM DO PROCESSO: 00003 ITEM DE MATERIAL: 000009873
7 - GENEROS DE ALIMENTACAO	4,200	2.14	8,988.00	4200,00000 Garrafa de 500 ml ÁGUA MINERAL Água mineral potável, não gaseificada, acondicionada em embalagem de 500ml, em plástico higienizado, com protetor na parte superior e lacre de segurança personalizado pelo fabricante, conforme Projeto Básico MARCA: Verágua ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000009873

Figure 1: Some of the information of a note of purchase available in the Transparency Portal for water.

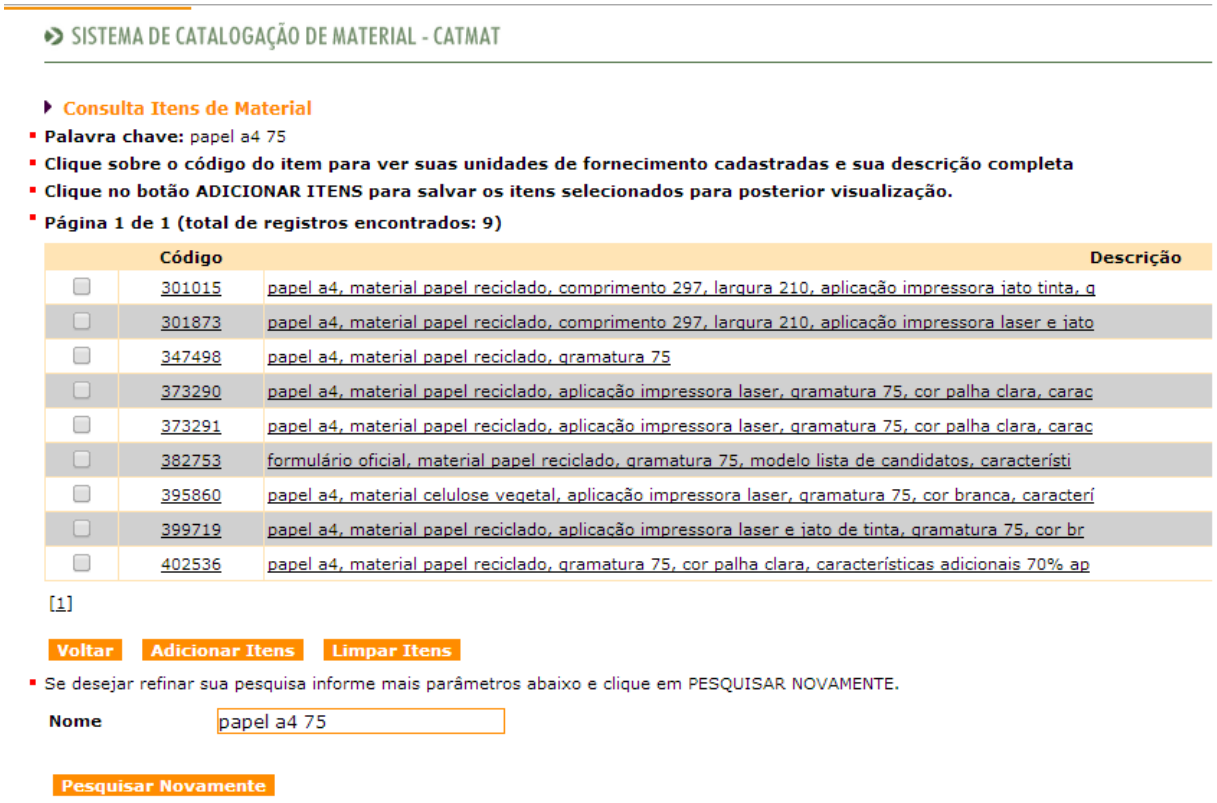


Figure 2: Result of searching for category codes used for papers of size A4 and that has 75 g/m^2.

papers of size A4 and that has 75 g/m^2. Other codes are more specific, differentiating papers used for laser and ink jet printers, for instance. However, it is common to have purchases of more specific types of papers being classified in a more general category code, even though there exists a more specific code for the type of paper purchased. There are also some cases where purchases are simply incorrectly classified. Thus, we cannot simply rely on the category code to group purchases of the same type in order to compute its reference price.

Although the category code defines both products and services, the only codes which will be considered when constructing the price database are of products. The reason for ignoring services is that the same service (*e.g.*, software development) can vary too much in its details, which will reflect in totally different prices (*e.g.*, a simple agency web page when compared to a system for allowing electronic votes for presidency). Therefore, it is not reasonable nor useful to compute averages in these situations.

Before using any advanced techniques, experts from CGU analyzed a simple and intuitive methodology for computing the reference prices. This methodology includes 6 major steps:

1. Retrieve the notes of purchase commitment for a given period from the Brazilian Transparency Portal database.
2. For every note of purchase commitment, retrieve the category code from the SIASG's database.
3. Filter the resulting dataset by category code to retrieve only the notes of purchase commitment of a given product (*e.g.*, code 21806 which refers to simple batteries).
4. Filter the resulting dataset by keywords in order to pinpoint a specific product (*e.g.*, aaa).
5. Filter the resulting dataset by price range.
6. Finally, compute the reference price for the product.

Steps 1 and 2 are performed by an Extract, Transform and Load (ETL) process[6]. Step 3 is a simple and straightforward Structured Query Language (SQL) query. In step 4, the experts define keywords that should be present in the description field of the note (*e.g.*, aaa) but also keywords that should not be present in the description (*e.g.*, car, if we want to make sure a car battery will not be in our result). The goal of step 4 is to pinpoint the specific product we want to compute the reference price for. Even after pinpointing a specific product, the experts realized that the price still had a large variance, which had a huge impact when computing the reference price, which resulted in an

[6]ETL refers to a process in databases and in data warehousing that involves: extracting data from outside sources; transforming it to fit operational needs; loading it into the desired target, which can be a database, a data warehouse, among others [5, 6].

unrealistic price when compared to standard prices found in commerce. The reason for such difference was that these purchases had not only some outliers, but also some other subtleties which were not easily captured by the expert when first looking at the data. For instance, there could be purchases of pairs of batteries as well as boxes of batteries. Since these subtleties were hard to catch by hand, the experts came up with a price range which could be thought as reasonable for the type of product in the unit of measure they were thinking about plus a large margin of error in order to account for unexpected values (*e.g.*, overprice and errors), which is the reason for step 5. Finally, step 6 was, at first, just the computation of the price average. Nevertheless, once the average was computed, the experts realized that the result was still not quite correct (too far from what it seemed to be the right one). The problem was that defining these price range was a though task and some datapoints (which could be outliers) had a huge impact in the average. Therefore, they decided to use the median as the reference price.

As it can be seen, in this initial methodology, a lot of questions arise, especially in steps 4 and 5. How can we find out which keywords to use? Do we have to have experts manually searching for them? But there are hundreds of thousands of records, how many experts will be needed? How can we define the price range? Is this just an expert feeling? What if there is no expert available that knows that type of product (*e.g.*, some specific medicine)? How can we trust the expert chose a reliable price range?

These questions led to the development of the proposed methodology in this paper, which uses well known Data Mining techniques to address these problems. Section 2.1 describes how to find similar products by clustering them based on the purchase price, which is a solid justification for price ranges used in step 5. Finally, Section 2.2 uses text mining techniques in order to classify the clusters found by using keywords from the description text field, which is an automatic way of doing step 4.

2.1 Using clustering to define reliable price range

As previously explained, one major challenge is to find group of purchases that describe the exact same product (*e.g.*, water) and in the same unit (*e.g.*, box of 12 or box of 24). Intuitively, experts used price range for such tasks, however, choosing the correct range and justifying the reason for that was not trivial. In this Section we will use the notes of purchase commitments for water bottles (500 ml) from 2011 and 2012 to show how clustering can reliably find these price ranges.

Table 1 presents a sample from the water bottle (500 ml) product dataset from years 2011 and 2012.

Table 1: Sample from water bottle (500 ml) dataset from years 2011 and 2012

	Description	Price (R$)	Quantity	Total Price (R$)
1	0000000001,00000 garrafa agua mineral garrafa de 500 ml de agua mineral sem gas, marca agua da pedra. marca: agua da pedra item do processo: 00003 item de material: 000009873	75.36	1.00	75.36
2	10,00000 cx agua mineral agua mineral com gas - pvc, com 24 unid de 500 ml marca: mil item do processo: 00215 item de material: 000009873	29.36	10.00	293.60
586	500,00000 garrafa agua mineral agua mineral c/ gas, garrafa c/ 500 ml marca: veragua item do processo: 00005 item de material: 000009873	0.32	500.00	160.00
587	300,00000 garrafa agua mineral agua mineral c/ gas, garrafa c/ 500 ml marca: veragua item do processo: 00005 item de material: 000009873	0.32	300.00	96.00

The main assumption that must hold when performing this clustering analysis is that most of the purchases of the same product in the same unit (*e.g.*, box of 12 water bottles) have similar price while purchases of the same product in different unit (*e.g.*, box of 24 water bottles) have significantly different price. The same must be true if different products (*e.g.*, aa vs D batteries) are present in the same dataset which is being clustered. This is a reasonable assumption, since error and/or fraud is usually the exception, not the rule.

This is exactly what cluster analysis do. Clustering has the objective of grouping a set of objects in order to maximize their similarity inside the same group (called cluster) and minimize the similarity they have to objects in other groups (clusters) [7].

There are several clustering algorithms. One of the most common is the k-means [8, 9], which represents each cluster by a single mean vector.

Every analysis made on the note of purchase commitment dataset was done using the R software[7], including the clustering analysis.

[7]http://www.r-project.org/

Although the k-means algorithm was used during our first analysis, we switched to fixed point clusters (FPC) [10,11] algorithm available as an R package [12]. The main reason for using FPC is that it is not necessary to define the number of clusters in advance (before running the clustering algorithm) as in k-means clustering algorithm. FPC computes the number of clusters automatically via bootstrap where several times 2 bootstrap samples are drawn from the data and the number of clusters is chosen by optimizing an instability estimation from these pairs, as explained in [13]. This automatic process is crucial, since we want to analyze several products and not every expert would be able to identify the correct number of clusters without some additional training.

Figure 3 shows the result of clustering the water 500ml product dataset from years 2011 and 2012. At this point, all we can say is that there are price ranges which seem to group similar products. But, what do these groups mean? Do they really represent similar products? Just by looking at their value, we are not able to validate these cluster. We need to understand what they have in common, *i.e.*, which type of product they represent. Section 2.2 presents how we use text mining techniques in the description text field in order to classify each cluster.

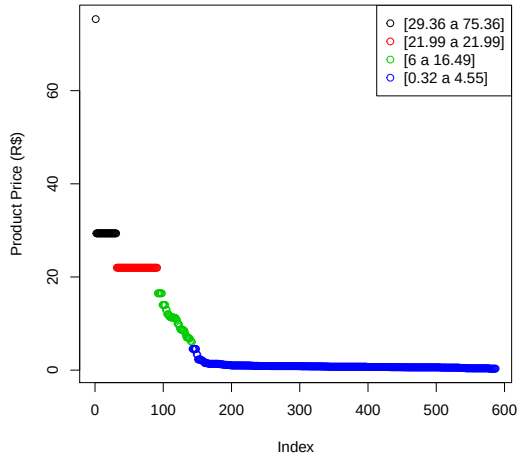


Figure 3: Result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012

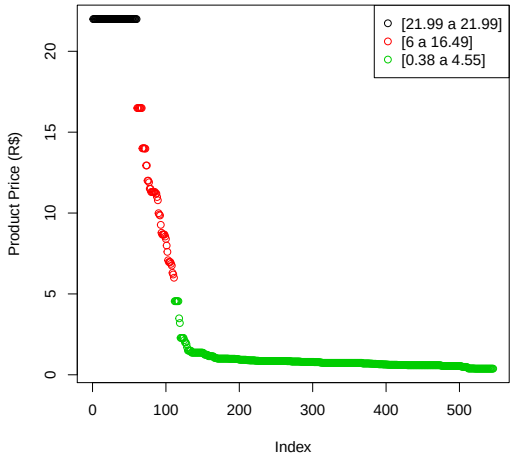


Figure 4: Result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012 without outliers

2.2 Using text mining for classifying product clusters

Text mining refers to the process of discovering useful information from text. This useful information is usually found via statistical pattern learning. Text mining usually involves the process of structuring the text, finding patterns using the structured data found, and finally analyzing its result [14, 15].

There are several tasks involved in text mining: text categorization and text clustering [16, 17], concept/entity extraction [18], production of granular taxonomies [19], sentiment analysis [20, 21], document summarization [22–24], and entity relation modeling [25, 26]. The overall goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) [27] and analytical methods.

Section 2.1 presented how we use clustering to define groups of purchases that fall in the same price range. However, in order to make sure these purchases are indeed talking about the same product, we need to extract its semantics from the description text field. Therefore, our main goal in this Section is to use text mining for classifying each cluster to be able to make such assessment.

After applying some text mining techniques, like creating the corpus, which represents a collection of text documents, preprocessing (*e.g.*, stripping white spaces, removing stopwords) the corpus, and creating the term-document matrix, we were able to find the most frequent words in each cluster. Finally, with these most frequent words per cluster, we were able to compute the words that better define the cluster, which we interpret as the cluster classification. These words are the set of most frequent words minus the intersection of the sets of most frequent words from all clusters but the one being analyzed as defined in Equation 1. All text mining analysis was done using the Text Mining Infrastructure in R (TM) package [28–30].

$$words.define.cluster_i = frequent.words_i \setminus \bigcap_{c \setminus i} (frequent.words_c) \quad (1)$$

Table 2 presents the text mining classification result for each cluster found for the water bottle (500 ml) product dataset from years 2011 and 2012. The first column presents the number of the cluster, the second presents the most frequent words in each cluster, and the third column presents the words that better define the cluster. As previously

explained this third column is computed as defined in Equation 1. In this case, we have, for instance, that the words that define cluster 1 are derived as presented in Equation 2. The same reasoning applies to the other three clusters.

$$\begin{aligned} words.define.cluster_1 = \{&000009873, 00215, 24, 500, agua, cx, gas, item, marca:, material:, mil, mineral, ml, \\ &processo:, pvc, unid\} \setminus \bigcap_{c=2,3,4} (frequent.words_c) \\ words.define.cluster_1 = \{&000009873, 00215, 24, 500, agua, cx, gas, item, marca:, material:, mil, mineral, ml, \\ &processo:, pvc, unid\} \setminus \{000009873, 500, agua, item, marca:, material:, mineral, processo:\} \\ words.define.cluster_1 = \{&00215, 24, cx, gas, mil, ml, pvc, unid\} \end{aligned} \quad (2)$$

As it can be seen in Table 2, the first and second clusters represent purchases of boxes of 24 bottles (from the keyword *cx*, which is short for *caixa* - box in Portuguese - and keyword 24). The third cluster represents purchases of boxes of 12 bottles (from the keyword 12). Finally, the fourth cluster represents purchases of single bottles (from the keyword *garrafa*, which is bottle in Portuguese).

Table 2: Result of the text mining classification of each cluster found in the water bottle (500 ml) product dataset

	Most Frequent Words in the Cluster	Words that Better Define the Cluster
1	000009873 / 00215 / 24 / 500 / agua / cx / gas / item / marca: / material: / mil / mineral / ml / processo: / pvc, / unid	00215 / 24 / cx / gas / mil / ml / pvc, / unid
2	000009873 / 00214 / 24 / 500 / agua / cx / item / marca: / material: / mil / mineral / ml / processo: / pvc, / unid	00214 / 24 / cx / mil / ml / pvc, / unid
3	000009873 / 12 / 500 / agua / gas, / item / marca: / material: / mineral / processo:	12 / gas,
4	000009873 / 500 / agua / garrafa / gas, / item / marca: / material: / mineral / ml / processo:	garrafa / gas, / ml

Although a human being has to verify the sets of keywords in order to better classify each cluster, it is a lot faster, easier, and less error-prone than going through every single purchase manually. This product alone, which is one with the products with fewer purchases, has over 500 purchases just in 2011 and 2012. It would be unrealistic to ask an expert to look at all these purchases for hundreds of products. However, with this methodology, this is actually feasible. In fact, it is quite fast. It took an expert, in average, less than 10 minutes to generate the report with this classification information and assess what each cluster means in a more human understandable way. This is why this methodology is so useful.

One question that may arise from the classification found by the text mining algorithm is why the first two groups were not identified as just one. Since the words that better define both groups are the same.

The problem is that the highest datapoint is way higher then all the other values (see Figure 3). In fact, this is a clear outlier. When finding the clusters for this product, this outlier ends up causing the creation of a new cluster.

After some exploratory analysis, mostly using box plots and scatter plots, we were able to identify that for most of the products we analyzed (most of the 72 products currently available in our database), at least 1% of the datapoints are outliers (a more detailed discussion will be presented in Section 4). Therefore, for the purporse of better identifying the clusters, we may ignore these outliers by filtering the 1% of the datapoints with the highest and lowest values. In other words, we remove everything below the 0.5th percentile and we remove everything above the 99.5th percentile.

Figure 4 shows the result of clustering the water bottle (500 ml) product dataset from years 2011 and 2012, but now removing the supposed outliers. Notice that now the number of clusters found was 3, as expected, not 4. As a result, our text mining now is able to correctly classify all three clusters.

During the process of removing outliers, we might end up removing data that was not necessarily an outlier. Nevertheless, even by removing some of these data, we still have enough datapoints to compute with a reasonable confidence the average price per product. After all, we still have 99% of the datapoints to make that estimate. Furthermore, as we will see in Section 3, we just compute confidence intervals for large datasets (more than 30 datapoints). For clusters with less then 30 datapoints, we simply do not add to our price database, since we are not confident in its accuracy.

Finally, we also generate an SQL query in order to allow the recovery of all datapoints that contain the words that better describe each cluster. This way, we might get some purchases that were classified as part of one cluster (*e.g.*, box of 24 bottles), when in fact they could be part of a different cluster (*e.g.*, single bottle). In fact, this is the case for the purchase with the highest price. Although it was classified as a purchase of a box of 24 bottles, it is in fact a purchase of a single bottle (as it can be seen in the purchase description in Table 1).

Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 12, Iss. 1, pp. 38-52, 2014

3 Results

This Section presents the last step in our methodology, which is computing the statistical results per type of product (*i.e.*, each cluster defined previously) in order to define the reference price of the product which will be added to our reference price database.

Although we have analyzed the overall fit of the clusters found only manually and for a random sample of the purchases, we believe that it classified most of the purchases in the correct category (*e.g.*, bottle, box with 12 bottles, and box with 24 bottles). However, a more thorough performance analysis is needed in order to confirm our initial analysis/conclusion[8].

Nevertheless, we also compared the final mean price found for each of the products and categories we evaluated with the price found in regular stores and they were similar. Since finding reference price (means) is our main goal, we believe our approach presents at least a satisfactory result. If we had too many products being classified in the wrong category, the final mean price would be much higher or lower than expected. For instance, on the one hand, if too many purchases of single bottles were misclassified as purchases of boxes of 12 bottles, we would end up having a mean price for this category much lower than what is expected. On the other hand, if too many purchases of boxes of 12 bottles were misclassified as purchases of single bottles, we would end up having a mean price for this category much higher than expected. Since all categories presented a price similar to those found in regular stores, we believe they were correctly classified in our approach.

One problem we were able to identify with our approach is in cases where a product A shares the same price range with a different product B when they both share the same category code. If this happens (and it does happen in some cases), we are not able to differentiate between the two using our approach (which uses only price to differentiate the two). However, we were able to select several category codes that did not have this problem and we focused our analysis on these products. In future work we plan on incorporating other features in order to allow the differentiation in these cases.

Table 3 presents some of the summary statistics computed for the water bottle (500 ml) product. These summaries were computed for each cluster and also for the entire dataset (if there subtypes of products were not identified) for comparison. Furthermore, these summaries were computed per purchase, but also per product. Each purchase has a quantity associated with it, which is used to compute the summaries per product. The reason to compute these summaries per product is that the price paid in a purchase of 100 thousand bottles should have a higher weight than the price paid in a purchase of a single bottle.

Table 3: Statistics summary per cluster for the price of water bottles (500 ml) in 2011 and 2012

Cluster	Range [Min,Max]	Quart.[1st,3rd]	Mean 95% Conf. Int.	Mean	Median	Size
All - per Purchase	[0.38, 21.99]	[0.62, 1.37]	[3.53, 4.71]	4.12	0.82	546
All - per Product	[0.38, 21.99]	[0.54, 0.9]	[2.66, 2.72]	2.69	0.60	158,567
1 - per Purchase	[21.99, 21.99]	[21.99, 21.99]	[21.99, 21.99]	21.99	21.99	60
1 - per Product	[21.99, 21.99]	[21.99, 21.99]	[21.99, 21.99]	21.99	21.99	11,352
2 - per Purchase	[6, 16.49]	[8.61, 12.94]	[10.06, 11.80]	10.93	11.20	51
2 - per Product	[6, 16.49]	[8.67, 11.3]	[9.64, 9.76]	9.70	8.67	7,158
3 - per Purchase	[0.38, 4.55]	[0.59, 0.89]	[0.80, 0.91]	0.86	0.74	435
3 - per Product	[0.38, 4.55]	[0.39, 0.78]	[0.76, 0.77]	0.77	0.59	140,057

The first thing we notice is that the number of purchases of single bottles is by far larger than the purchase of boxes of bottles (about 80% of the total purchases). Therefore, the reference price for single bottles should be considered more reliable, or closer to the "real" fair price, than the reference price for boxes. This is actually reflected in the mean 95% confidence interval. The confidence interval range for single bottle purchases is a much narrower ([0.80, 0.91]) than box purchases ([10.06, 11.80] for the box with 12 bottles). This is not always true, as it can be seen for the box with 24 bottles, since all 60 purchases had the same price associated with them (21.99).

The same principle holds when comparing the mean confidence intervals of purchases and products, since the number of products bought is much higher than the number of purchases made.

It is also important to validate if these values correspond to reality. A simple way to validate the price is to compare to the price a regular citizen would pay, since this product is common and easy to find. For instance, a water bottle (500 ml) can be bought in the supermarket for a little less than R$1.00 here in Brazil. So, if the Government pays, in average, R$0.77 per bottle, than we known we have a good price for reference. Furthermore, we also have some good news in the sense that the Government is paying less than a regular citizen who is buying just a few bottles.

Nevertheless, it does not sound reasonable to expect every single purchase to be made for that computed average price every time, since the price may vary during the year, it varies from state to state, etc. Therefore, it is more reasonable to define a reference interval instead of single price per product. This is where the ranges shown in Table 3

[8]This is already being done, however, due to the size of the database, and the fact that we have to manually classify each purchase we want to validate, this is taking a long time and we were not able to finish it at this point.

play a key role. As it can be seen, 50% of all purchases of single bottles paid between R$0.59 and R$0.89 and 50% of all bottles purchased cost between R$0.39 and R$0.78. A more interesting interval is shown in the mean 95% confidence interval. Although the mean of purchases of single bottles was R$0.86, its 95% confidence interval goes from R$0.80 to R$0.91. On the other hand, while the mean of the price paid per bottle was R$0.77, its 95% confidence interval goes from R$0.76 to R$0.77.

Even though it is hard to decide which interval to use for the reference price. After some discussion with the experts at CGU, it was decided that the reference price will be the interval from the minimum value between both confidence intervals (per purchase and per product) to the maximum value between both confidence intervals. Thus, the reference price for water bottle (500 ml) will be from R$0.76 to R$0.91.

It is unquestionable that, independent of which interval is chosen, having these parameters will be of great value to those responsible for the procurements, for auditors, and also for citizens.

The department responsible for all purchases at CGU (*i.e.*, responsible for the procurements) has already shown interest in our reference price database and has asked us to compute reference price for many other products.

The department at CGU responsible for auditing and inspecting all Federal expenses has also shown interest in our database. They use the information in our database to verify if the procurements they have to analyze do not have overprice. For instance, for this water product we can identify a purchase of R$1.8 thousand bottles at the price of R$3.49, a price 453% higher than the average price (R$0.77). This purchase alone incurred in a loss of almost R$5 thousand. Notice that we are describing a purchase of a product that is worth less than R$1.00. In fact, the potential loss with purchases of water bottles (just the purchases of single bottles, not boxes) is a bit over R$13 thousand. This value was computed as the sum over all purchases above the maximum reference price (R$0.91) of the purchase price minus R$0.91 times the quantity of bottles bought. Imagine the losses that would be discovered for products with higher average price. This is why this reference price database is so valuable and important for a better management of public resources.

Today, we have already created a fist price reference database for 72 products[9]. These data are already being used by internal auditors from CGU in its current activities.

Finally, citizens will be able to verify how their money is being spent which will result in a much greater transparency and accountability. At this stage, it is being discussed at CGU how to make this database available as OGD for all citizens in our Transparency Portal. A consequence of this discussion was the inclusion of the Reference Price Database in Brazil's action plan in the OGP partnership. According to the new plan[10], CGU has to implement until October 2014 the following commitment:

> **Development of a Database of the Federal Public Administration Purchases Prices:** to develop a database containing reference prices for the most purchased items by the Federal Government, from data published on the Transparency Portal. The interface shall provide for the identification of items average prices, thus constituting an efficient strategy for formulating budgets and procurements, disseminating best practices in public purchases, as well as for supporting actions aimed at fighting corruption, especially in circumstances where overprice purchases are identified.

4 Purchase patterns and other preliminary conclusions

In this Section we describe a few patterns found in the dataset as well as some preliminary conclusions, which will be an initial step of a further and future analysis made by experts from CGU.

The first thing we realized when using the methodology proposed is that it does not work for every product. In fact, this was already expected, as we have defined a major assumption that must hold (similar products have similar prices). Table 4 presents a sample from the diesel product dataset from 2012, which is a product that breaks this assumption, since it has a wide range of prices for the same product. We will demonstrate what happens when we try to apply the proposed methodology when this assumption does not hold.

Figure 5 presents the result of the clustering analysis. The clustering algorithm has found two different clusters in this dataset, but do they actually represent different products?

As expected, the available information in the description of the purchases cannot explain the clusters found, as it can be seen in Table 5. This can be easily verified by looking at the descriptions on Table 4. Although the price varies drastically, the descriptions is basically the same, which is the liter of diesel.

In other words, all we learned when applying the proposed methodology was that although the dataset indicates that there are two different products, because of its large difference in price, it is in fact the same product. Although this is not the result we expected, it is still useful. It shows that there is, at least, something strange going on with purchases of this product. Either there is not enough information to pinpoint the exact product being bought (see discussion on patterns of uniform and nonuniform products below) or there are too many errors and/or frauds

[9]The same type of product in a different unit is considered a different product. For instance, a 350 ml coke can is a different product than a 2 l coke bottle.

[10]See latest action plan at http://www.opengovpartnership.org/country/brazil.

Table 4: Sample from the diesel product dataset from 2012

	Description	Price (R$)	Quantity	Total Price (R$)
57	0000000001,00000 litro oleo diesel oleo diesel. marca: valecard item do processo: 00002 item de material: 000016993	20756.27	1.00	20756.27
58	0000000001,00000 lt oleo diesel oleo diesel, nome oleo diesel marca: item do processo: 00003 item de material: 000016993	20166.69	1.00	20166.69
5163	0000120000,00000 ano/lts oleo diesel oleo diesel, nome oleo diesel, para um periodo de 12(doze)meses marca: petrobras item do processo: 00002 item de material: 000016993	0.01	120000.00	1998.00
5164	56400,00000 litro oleo diesel oleo diesel marca: trivale item do processo: 00003 item de material: 000016993	0.01	56400.00	1000.00

Table 5: Result of the text mining classification of each cluster found in the diesel product dataset

	Most Frequent Words in the Cluster	Words that Better Define the Cluster
1	0000000001,00000 / 000016993 / diesel / item / marca: / material: / oleo / processo:	0000000001,00000
2	000016993 / diesel / item / marca: / material: / oleo / processo:	

happening. For instance, we see in Figure 5 that there are several purchases paying more than R$5,000.00 for a liter of diesel when its standard price in December 2012 was less than R$3.00 per liter[11].

Another interesting discovery during the use of our proposed methodology was the pattern of uniform and nonuniform products. Figure 6 presents the purchases price for products considered uniform accross different purchases, while Figure 7 presents purchases price for products considered nonuniform accross different purchases. Here, uniform means that if two purchases have similar descriptions then they will receive similar products. This is not always the case. For instance, when describing an office chair, even if two different purchases describe it as blue chair with wheels, they might end up getting two totally different products, with respect to quality, shape, among other characteristics.

In both Figure 6 and Figure 7 the prices in red are the purchases greater than two times the median of the overall purchases for that product. The prices in blue are the ones lower than half the median of the overall purchases for that product. The values in black are the most common prices for that product. Notice that just by looking at the plot of each product, it is possible to assess if the product is uniform or not, even without actually knowing what the product is. On the one hand, uniform products usually have an slope close to zero for the purchases with common price, *i.e.*, almost all purchases had the same (or very similar) price. On the other hand, nonuniform products usually have an slope closer to negative one for the purchases with common price, *i.e.*, almost no purchase (or very few) had the same (or very similar) price.

This pattern is interesting because nonuniform products could present a greater challenge when trying to define a reference price. Therefore, it might be better to not even define a reference price for these products or at least make it clear to the end user that there is a er uncertainty about this reference price when compared to uniform products.

Moreover, another pattern discovered is shown in Figure 6 and Figure 7. After an exploratory analysis in different products it was identified that the non-black values can be usually justified by difference in units (*e.g.*, pack of 12 bottles vs pack of 24 bottles), overprice (red values), economy (blue values), or by a simple error while inputting the data into the system.

This pattern can be useful for various activities. The overpriced purchases can be further investigated to punish those responsible. Those responsible for buying for a better price could be contacted in order to identify and document a best practice when buying certain types of products for the Government. Finally, those responsible for inputting incorrect data into the system could be contacted to receive an special training in order to avoid future mistakes.

Another surprising discovery is that although it is common sense that the more you buy the cheaper it gets, this is not the case in most of the purchases analyzed. Figure 8 presents the correlation between the price and quantity of the purchases made in 2012 for the uniform products discussed previously. Although the expected would be a number close to -1, it can be seen that the correlations are all close to 0.

Finally, another similar discovery is that purchases of boxes, bags, etc, in average, paid more than purchases of single products. This was verified for various different products. As it can be seen in Table 3 from Section 3, a box of 24 water bottles was bought in average for R$21.99, while buying 24 water bottles would cost only $R\$0.77 * 24 = R\$18.48$. The same thing happens with a box of 12 water bottles. It costs in average R$9.70, while buying 12 water bottles would cost slightly less, $R\$0.77 * 12 = R\$9.24$.

[11] http://www.anp.gov.br/preco/

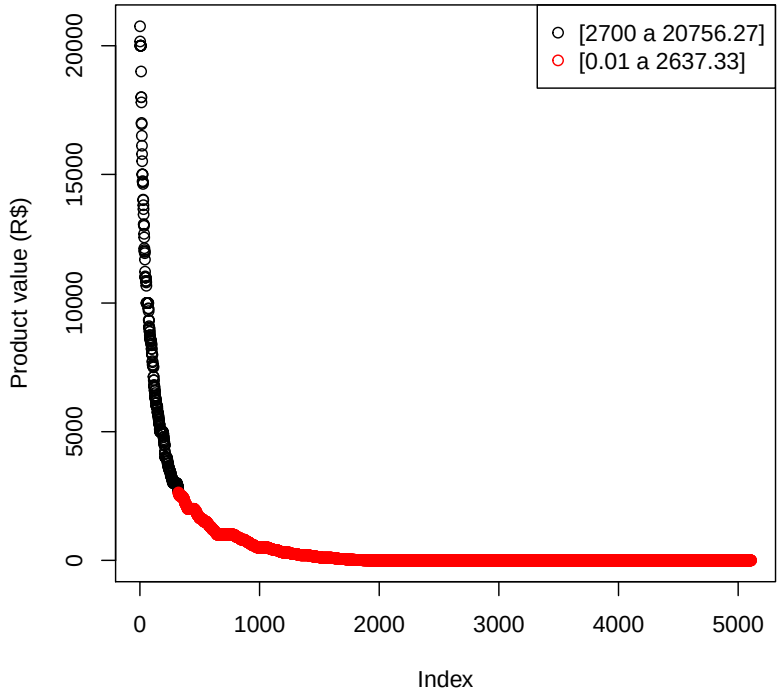


Figure 5: Result of clustering the diesel dataset from 2012

These discoveries contradict what is expected in a regular commercial transaction. Therefore, it invites further study and investigation, which is also being considered as a future project at CGU.

5 Conclusions

This paper presented a major problem that prohibits the creation of a reference price database in the Brazilian Government, which is being unable to categorize Government purchases precisely. The main challenge comes from the fact that the information available is not structured and the only classification available is too broad to allow the definition of a reference price.

In order to tackle this problem, CGU developed a methodology, described in Section 2, which uses Data Mining techniques to overcome the challenges presented. Section 3 demonstrated the methodology works by showing how this methodology was used to compute the reference price of water bottles (500 ml).

As described in Section 3, many different stakeholders are interested in the price reference database, including the department responsible for the procurements at CGU, the department at CGU responsible for auditing and inspecting Federal expenses, as well as citizens. Thus, CGU has made an effort to create an initial version of the database with 72 products. Moreover, since transparency and accountability are major goals of the institution, it is being discussed how to make this database available for all citizens in our Transparency Portal. One of the consequences of this discussion was the commitment in Brazil's second action plan for the OGP partnership to create the Reference Price Database.

Although the paper only presents how the methodology was applied to only one product (water bottle) due to simplicity and length restrictions, we were able to successfully apply the same methodology to several other products. Table 6 presents some of these products, their unit of measure, and their respective reference price. The reference price for these products were validated by the department at CGU responsible for its purchases. Furthermore, the prices were compared to the values one can find at a regular store or supermarket and they were similar, which was also considered another way of validating them.

Besides allowing simple queries in the database, tools for improving auditing and inspection are being designed. The focus will be in allowing the identification of who is spending over the reference price, which agencies are being able to buy for less, how much the Government is spending over the reference price, among other features.

Finally, Section 4 described what happens when the methodology is used in purchases that do not conform to the assumption of similar products having similar prices. Furthermore, some initial conclusions and insights from the exploratory analysis were presented. They proved useful in understanding the overall domain of Government purchases and promoting transparency and accountability.

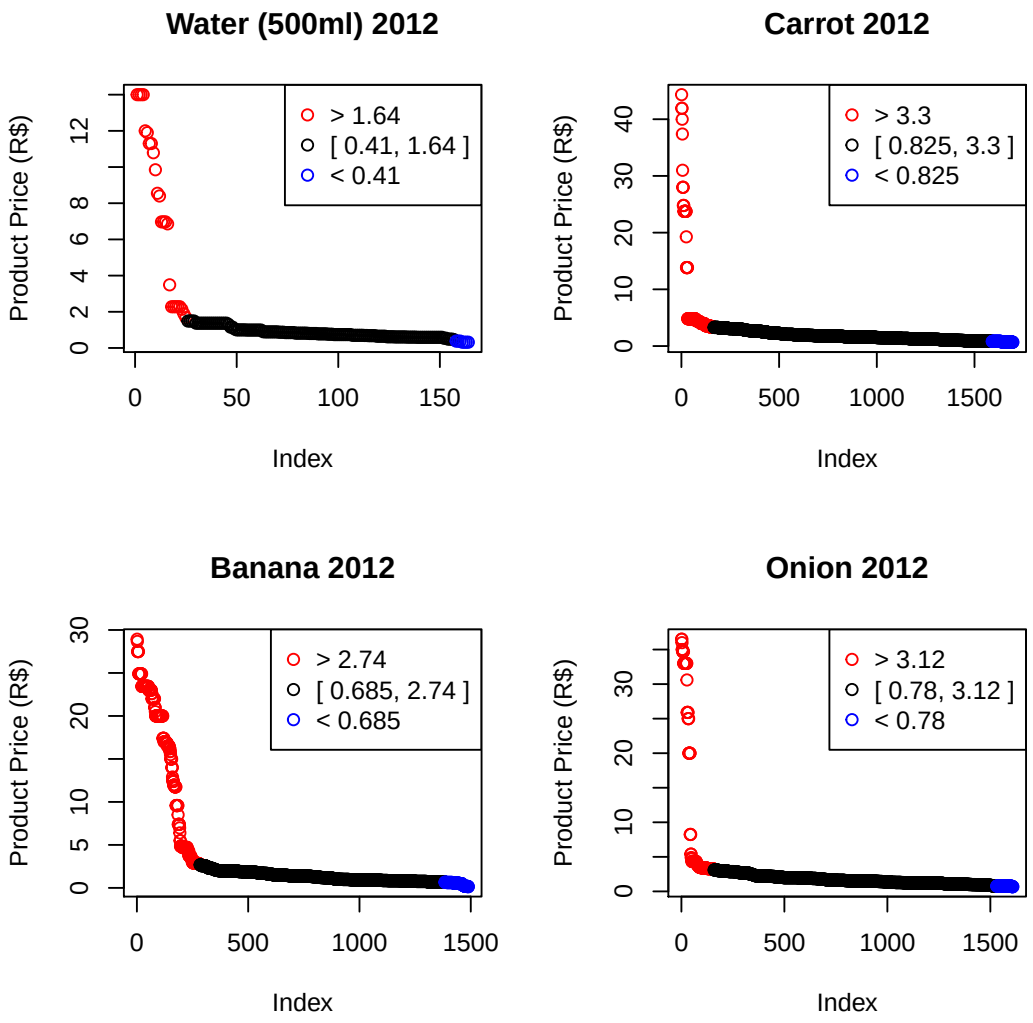


Figure 6: Result of clustering the diesel dataset from 2012

One of the major benefits of the methodology and resulting reference price database presented in this paper is in improving the management of public resources. As an example, recently, the Court of Auditors from the Federal District (DF in Portuguese) did a similar task of finding reference prices but for a specific procurement. The procurement of the DF Health Secretariat, which was initially estimated in almost R$86 millions to buy different medications, was only allowed to continue after it was reestimated. For everybody astonishment, the reestimated price fell to almost R$13 millions. The use of reference prices was responsible for savings over 85%, about R$73 millions[12].

6 Acknowledgments

The authors would like to thank CGU for the support in this project and for letting the authors publish these results.

REFERENCES

[1] C. H. d. A. Moreira. "Implementation of E-Procurement System in Brazil". In *Proceedings: Towards Frontiers in Public Procurement*, 2010.

[2] "About | Open Government Partnership". http://www.opengovpartnership.org/about.

[3] "Country Commitments | Open Government Partnership". http://www.opengovpartnership.org/countries.

[12] http://www.tc.df.gov.br/web/tcdf1/noticias/-/asset_publisher/a5YM/content/valor-de-licitacao-para-compra-de-remedios-tem-reducao-de-85

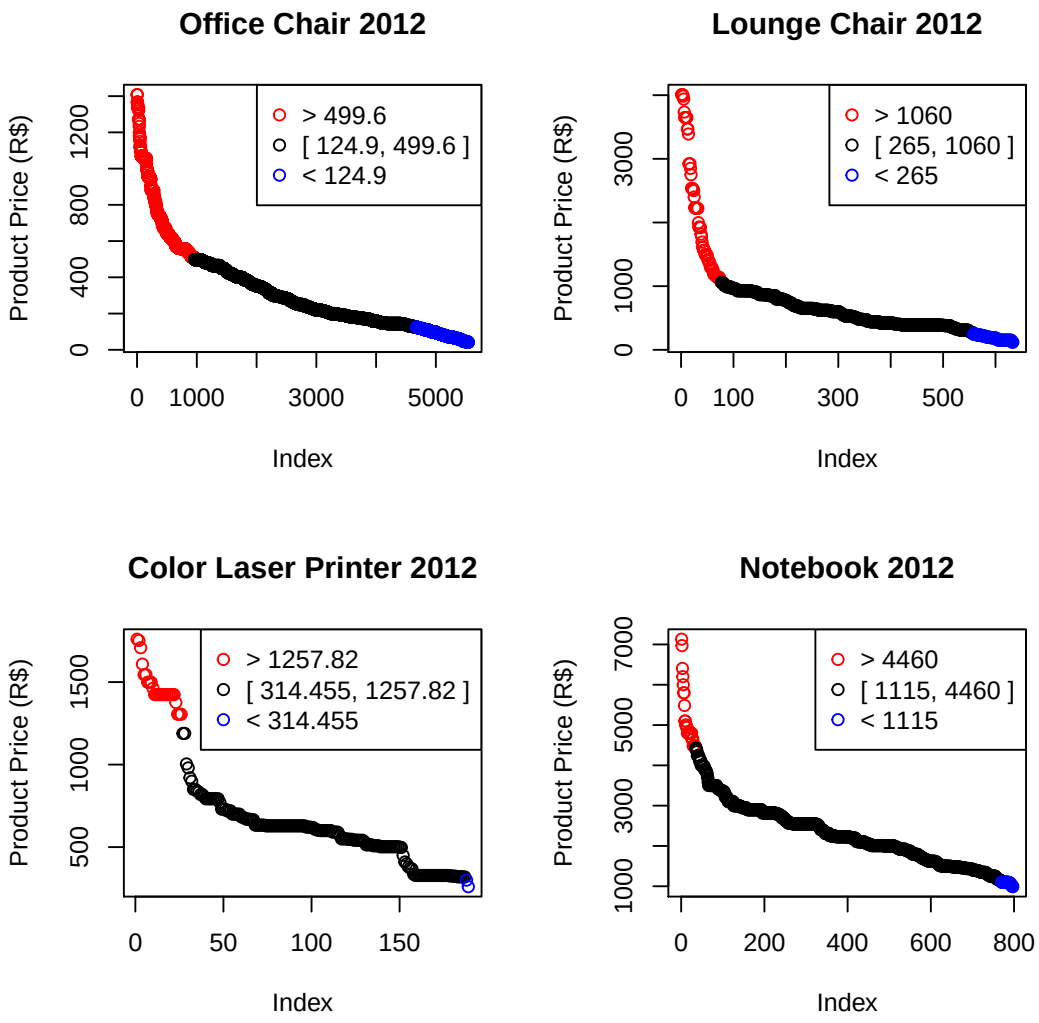


Figure 7: Result of clustering the diesel dataset from 2012

[4] "CGU -Transparency Portal". `http://www.cgu.gov.br/english/AreaPrevencaoCorrupcao/AreasAtuacao/IncrementoPortal.asp`, 2013.

[5] "Extract, transform, load". `http://en.wikipedia.org/w/index.php?title=Extract,_transform,_load&oldid=535476953`, February 2013. Page Version ID: 535476953.

[6] R. Kimball and J. Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin*. Wiley, first edition, September 2004.

[7] "Cluster analysis". `http://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=537295570`, February 2013. Page Version ID: 537295570.

[8] "k-means clustering". `http://en.wikipedia.org/w/index.php?title=K-means_clustering&oldid=536835500`, February 2013. Page Version ID: 536835500.

[9] J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm". *Applied Statistics*, vol. 28, no. 1, pp. 100, 1979.

[10] C. Hennig. "Fixed Point Clusters for Linear Regression: Computation and Comparison". *Journal of Classification*, vol. 19, no. 2, pp. 249–276, December 2002.

[11] C. Hennig. "Clusters, outliers, and regression: fixed point clusters". *J. Multivar. Anal.*, vol. 86, no. 1, pp. 183–212, July 2003.

[12] "R: Linear Regression Fixed Point Clusters". `http://rss.acs.unt.edu/Rdoc/library/fpc/html/fixreg.html`, 2013.

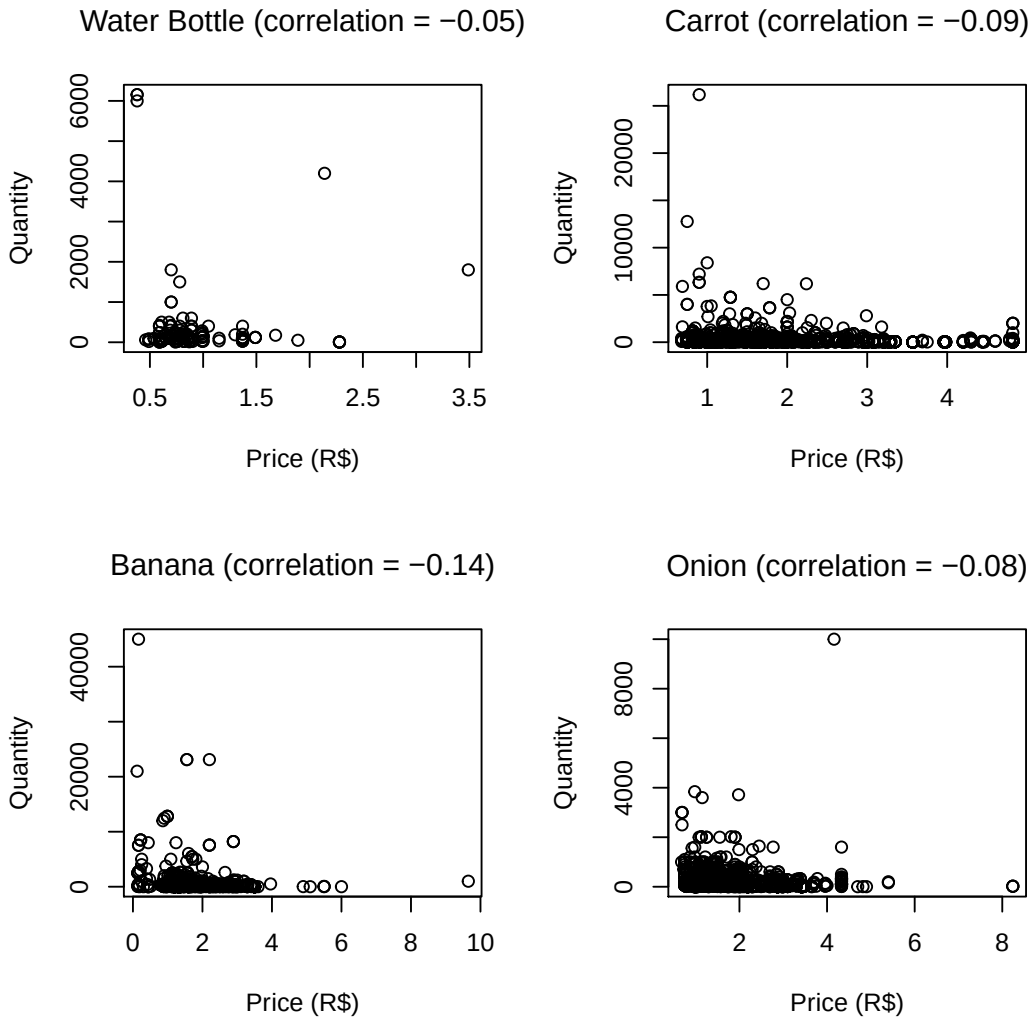


Figure 8: Correlation between price and quantity of different products

[13] Y. Fang and J. Wang. "Selection of the number of clusters via the bootstrap method". *Comput. Stat. Data Anal.*, vol. 56, no. 3, pp. 468–477, March 2012.

[14] "Text mining". `http://en.wikipedia.org/w/index.php?title=Text_mining&oldid=532797054`, January 2013. Page Version ID: 532797054.

[15] P.-N. Tan, M. Steinbach and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, first edition, May 2005.

[16] A. Srivastava and M. Sahami, editors. *Text Mining: Classification, Clustering, and Applications*. Chapman and Hall/CRC, first edition, June 2009.

[17] F. Sebastiani. "Machine learning in automated text categorization". *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, March 2002.

[18] H. Al Fawareh, S. Jusoh and W. Osman. "Ambiguity in text mining". In *International Conference on Computer and Communication Engineering, 2008. ICCCE 2008*, pp. 1172 –1176, May 2008.

[19] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler and O. Zamir. "Text mining at the term level". In *Principles of Data Mining and Knowledge Discovery*, edited by J. Zytkow and M. Quafafou, volume 1510 of *Lecture Notes in Computer Science*, pp. 65–73. Springer Berlin / Heidelberg, 1998.

[20] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, July 2008.

[21] M. Gamon, A. Aue, S. Corston-Oliver and E. Ringger. "Pulse: Mining Customer Opinions from Free Text". In *Advances in Intelligent Data Analysis VI*, volume 3646 of *Lecture Notes in Computer Science*, pp. 741–741. Springer Berlin / Heidelberg, 2005.

Table 6: Reference price for products using the methodology described in this paper.

Product	Unity of Measure	Reference Price (min.)	Reference Price (max.)
Mineral Water	20 liters	R$ 5.56	R$ 5.85
Banana	kg	R$ 1.97	R$ 2.04
Potato	kg	R$ 2.27	R$ 2.38
Roasted Coffee	kg	R$ 10.41	R$ 10.86
Roasted Coffee	250 g	R$ 2.73	R$ 2.82
Roasted Coffee	500 g	R$ 6.25	R$ 6.79
Onion	kg	R$ 2.30	R$ 2.38
Carot	kg	R$ 2.04	R$ 2.12
Orange	kg	R$ 1.16	R$ 1.21
Sausage	kg	R$ 8.76	R$ 9.00
Ham	kg	R$ 11.78	R$ 12.21
Alcohol Fuel	liter	R$ 2.13	R$ 2.19
Gasoline Fuel	liter	R$ 2.88	R$ 2.90
Diesel Fuel	liter	R$ 2.34	R$ 2.36

[22] J. Larocca Neto, A. D. Santos, C. A. Kaestner and A. A. Freitas. "Document Clustering and Text Summarization". Postgraduate thesis, Pontificia Universidade Catolica do Parana.

[23] B. Larsen and C. Aone. "Fast and effective text mining using linear-time document clustering". In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pp. 16–22, New York, NY, USA, 1999. ACM.

[24] M. Hu and B. Liu. "Mining and summarizing customer reviews". In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pp. 168–177, New York, NY, USA, 2004. ACM.

[25] A. M. Cohen and W. R. Hersh. "A survey of current work in biomedical text mining". *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 57–71, March 2005.

[26] A. Kao and S. R. Poteet. *Natural Language Processing and Text Mining*. Springer, January 2007.

[27] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. BARNES & NOBLE, 1999.

[28] "tm - Text Mining Package". http://tm.r-forge.r-project.org/.

[29] I. Feinerer, K. Hornik and D. Meyer. "Text Mining Infrastructure in R". *Journal of Statistical Software*, vol. 25, no. 5, pp. 1–54, 2008.

[30] I. Feinerer. "Introduction to the tm Package Text Mining in R". Technical report, January 2013.