TRIBUNAL DE CONTAS DA UNIÃO

INSTITUTO SERZEDELLO CORRÊA
ESCOLA SUPERIOR DO TRIBUNAL DE CONTAS DA UNIÃO

EVILÁSIO VILAR SILVA

ANÁLISE GEORREFERENCIADA EM PAGAMENTOS DA UNIÃO:
impacto de longas distâncias de fornecimento nos preços de materiais

Brasília
2019

EVILÁSIO VILAR SILVA

ANÁLISE GEORREFERENCIADA EM PAGAMENTOS DA UNIÃO: impacto de longas distâncias de fornecimento nos preços de materiais

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista em Análise de Dados.

Orientador: Prof. Luiz Felipe Carvalho Silva

Brasília
2019

REFERÊNCIA BIBLIOGRÁFICA

SILVA, Evilásio Vilar. **Análise Georreferenciada em Pagamentos da União**: impacto de longas distâncias de fornecimento nos preços de materiais. 2019. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF. 62f.

CESSÃO DE DIREITOS

NOME DO AUTOR: Evilásio Vilar Silva
TÍTULO: Análise Georreferenciada em Pagamentos da União**:** impacto de longas distâncias de fornecimento nos preços de materiais.
GRAU/ANO: Especialista/2019

_____

Evilásio Vilar Silva
evilar@gmail.com

Silva, Evilásio Vilar
Análise Georreferenciada em Pagamentos da União: impacto de longas distâncias de fornecimento nos preços de materiais / Evilásio Vilar Silva.
Brasília, 2019.
62f.
Orientador: Luiz Felipe Carvalho Silva.
TCC (Pós-Graduação - Especialização em Análise de Dados para o Controle) - Escola Superior do Tribunal de Contas da União, 2019.
1. Análise de dados. 2. Tesouro Gerencial. 3. Georreferenciamento. 4. Sobrepreço. 5. Execução financeira. I. Silva, Luiz Felipe Carvalho. II. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle. III. Título.

EVILÁSIO VILAR SILVA

ANÁLISE GEORREFERENCIADA EM PAGAMENTOS DA UNIÃO:
impacto de longas distâncias de fornecimento nos preços de materiais

Trabalho de Conclusão do Curso de Pós-graduação *lato sensu* em Análise de Dados para o Controle, realizado pela Escola Superior do Tribunal de Contas da União, como requisito para a obtenção do título de especialista em Análise de Dados.

Brasília, 13 de dezembro de 2019.

Banca Examinadora:

_____
Prof. Luiz Felipe Carvalho Silva, Esp.

_____
Prof. Felipe Leitão Valadares Roquete, MSc.

RESUMO

No ano de 2015, o governo federal lançou o Tesouro Gerencial, um sistema capaz de fornecer informações gerenciais referentes à execução orçamentária, financeira, contábil e patrimonial da União, em substituição ao SIAFI Gerencial, também utilizado para esta finalidade desde 1995. A atual plataforma aumentou o nível de transparência das informações e possibilitou gerar melhores relatórios para embasar análises e apoiar o processo de decisão do governo em diversos aspectos, inclusive para melhoria de políticas públicas. Com base nas informações advindas dessa plataforma, este artigo busca documentar uma análise geográfico-financeira envolvendo as duas partes fundamentais de um contrato administrativo: contratante e contratada. Na execução financeira da União, estas entidades estão materializadas nas figuras da Unidade Gestora e do Favorecido da Ordem Bancária, respectivamente. Deste modo, levando-se em consideração as citadas figuras, e aplicando-se a metodologia exploratória e explicativa, por meio de uma abordagem do problema quantitativa e de pesquisa experimental, dividiu-se o experimento em duas partes. Primeiramente, será feita uma análise mais geral e exploratória, utilizando toda a massa de dados de pagamento do Tesouro Gerencial, desde o ano de 2015. Posteriormente, sobre uma amostra, será validada uma hipótese criada durante essa exploração inicial: se longas distâncias de fornecimento poderiam impactar o preço final do produto e, consequentemente, prejudicar o alcance da finalidade pública. O modelo de referência CRISP-DM foi utilizado como metodologia para mineração de dados. Casos de interesse encontrados durante a análise exploratória inicial foram relatados. A partir de uma análise sobre a amostra de validação concluiu-se que, apesar de distâncias e preços possuírem alguma correlação, ela é devida a um motivo que difere da pressuposição que embasou a proposta original de pesquisa: tipos de produtos de maior valor agregado costumam ser fornecidos a maiores distâncias. Contudo, não foi constatada clara influência da distância de fornecimento no preço final do produto dentro da maioria das classes de materiais.

Palavras-chave: Análise de dados. Tesouro Gerencial. Georreferenciamento. Distância. Sobrepreço. Materiais. Fiscalização financeira. Pagamento. Ordem bancária. Análise de dados governamentais. SIAFI.

ABSTRACT

In 2015, the Brazil's federal government launched the platform called "Tesouro Gerencial", a system capable of providing management information regarding the Union's budget, financial, accounting, and asset execution, in place of the "SIAFI Gerencial", which has been used for this purpose since 1995. This new platform increased the level of transparency of information and made it possible to generate better reports to support analysis and to support government decision-making in various aspects, including to improve public policies. Based on the information from these reports, this paper seeks to document a geo-financial analysis involving the two fundamental parts of an administrative contract: hirer and hired. In the financial execution of the Brazil's Union, these entities are represented by the Management Unit and the Banking Order Beneficiary, respectively. Thus, taking into consideration the mentioned figures, and applying the exploratory and explanatory methodology, through a quantitative problem approach and experimental research, the experiment was divided into two parts. Firstly, a more general and exploratory analysis will be made using the full mass of Management Treasury payment data since 2015. Subsequently, on a sample, a hypothesis created during this initial exploration will be validated: if long distances of supply could impact the final price of the product, weakening the public welfare. The CRISP-DM reference model was used as a methodology for data mining. Cases of interest found during the initial exploratory analysis were reported. From an analysis of the validation sample it was found that, although distances and prices have some correlation, it is due to a reason that differs from the assumption underlying the original research proposal: higher value-added products types tend to be supplied over longer distances. However, no clear influence of supply distance on final product price was found inside most material classes.

Key-words: Data analysis. Tesouro Gerencial. Georeferencing. Distance. Overpriced. Products. Financial supervision. Payment. Bank Order. Analysis of government data. SIAFI.

LISTA DE FIGURAS

Figura 1 – Fluxograma da execução orçamentária e financeira 15
Figura 2 – Estágios da Despesa Pública 16
Figura 3 – Tela de consulta a ordens bancárias (CONOB) no SIAFI 18
Figura 4 – Marcos do reordenamento das finanças públicas 19
Figura 5 – Tela inicial do Tesouro Gerencial 20
Figura 6 – Processo de carga do DWTG 21
Figura 7 – Acesso ao DWTG via cliente JDBC 22
Figura 8 – Fases do modelo de referência CRISP-DM 24
Figura 9 – Números totais relacionados aos pagamentos no DWTG 26
Figura 10 – Painel de informações elaborado para exploração de dados 27
Figura 11 – Trecho de código Python utilizando expressões regulares 30
Figura 12 – Informações contidas em um gráfico *boxplot* 31
Figura 13 – Comparativo entre diferentes algoritmos de agrupamento 34
Figura 14 – Principais finalidades dos pagamentos federais em 2018 37
Figura 15 – Dispersão: Valores x Distâncias por CNAE 38
Figura 16 – *Boxplot*: *Outliers* de distância 39
Figura 17 – Fornecedor EPP com distribuição em todo o Brasil 40
Figura 18 – Empresa sediada em PE com 94% dos fornecimentos para outras UFs 41
Figura 19 – Dispersão: Distância x Preço 44
Figura 20 – Dispersão: Distância x Preço por tipo de produto 45
Figura 21 – Dispersão: Distância x Preço para cada tipo de produto 47
Figura 22 – Agrupamento dos pontos da amostra 48

LISTA DE TABELAS

Tabela 1 – Descrição do conjunto de dados 28
Tabela 2 – Principais tarefas de mineração de dados..... 33
Tabela 3 – Distribuição de produtos na amostra..... 42
Tabela 4 – Análise descritiva geral..... 43
Tabela 5 – Teste de correlação entre distância e preço para cada tipo de produto.. 45
Tabela 6 – Percentual de sobrepreço para cada tipo de produto..... 49

LISTA DE ABREVIATURAS E SIGLAS

Art.	Artigo
BI	Business Inteligence
CGU	Controladoria-Geral da União
CNAE	Código Nacional de Atividades Econômicas
CNPJ	Cadastro Nacional de Pessoas Jurídicas
CPF	Cadastro de Pessoas Físicas
CRISP-DM	Cross Industry Standard Process for Data Mining
DWTG	Data Warehouse do Tesouro Gerencial
EPP	Empresa de Pequeno Porte
HC	Habeas Corpus
IBGE	Instituto Brasileiro de Geografia e Estatística
ISC	Instituto Serzedello Corrêa
JDBC	Java Database Connectivity
LRF	Lei de Responsabilidade Fiscal
ME	Microempresa
MRE	Ministério das Relações Exteriores
NF-e	Nota Fiscal Eletrônica
OB	Ordem Bancária
ρ	Coeficiente de Correlação de Spearman
R	Linguagem de programação R
SQL	Structured Query Language
SERPRO	Serviço Federal de Processamento de Dados
SIAFI	Sistema Integrado de Administração Financeira do Governo Federal
SIASG	Sistema Integrado de Administração de Serviços Gerais
STJ	Superior Tribunal de Justiça
STN	Secretaria do Tesouro Nacional
TCU	Tribunal de Contas da União
UA	Unidade Administrativa
UF	Unidade Federativa
UG	Unidade Gestora
Un	Unidade
UO	Unidade Orçamentária

SUMÁRIO

1 INTRODUÇÃO 11
1.1 OBJETIVOS 13
1.1.1 Objetivo geral 13
1.1.2 Objetivos específicos 13

2 REFERENCIAL TEÓRICO 14
2.1 EXECUÇÃO ORÇAMENTÁRIA E FINANCEIRA 14
2.1.1 Execução Financeira da União 16
2.1.2 Estágios da Despesa Pública 16
2.1.2.1 Pagamento 17
2.2 SISTEMA INTEGRADO DE ADMINISTRAÇÃO FINANCEIRA DO GOVERNO FEDERAL 17
2.3 TESOURO GERENCIAL 19
2.3.1 Data Warehouse do Tesouro Gerencial 20

3 METODOLOGIA APLICADA 23
3.1 CRISP-DM 23
3.1.1 Entendimento do negócio 24
3.1.1.1 Inventário e escolha das fontes de dados 25
3.1.2 Entendimento dos Dados 25
3.1.2.1 Coleta de dados inicial 26
3.1.2.2 Exploração dos dados 26
3.1.3 Preparação dos dados 27
3.1.3.1 Descrição do conjunto de dados 27
3.1.3.2 Criação de atributos derivados 28
3.1.3.3 Remoção de outliers 30
3.1.4 Modelagem 32
3.1.4.1 Mineração de dados 32
3.1.4.2 Correlação de Spearman 34
3.1.5 Avaliação 35
3.1.6 Implantação 36

4 RESULTADOS OBTIDOS 37
4.1 ANÁLISE EXPLORATÓRIA 37
4.1.1 Fornecimento de produtos e serviços comuns a grandes distâncias 38
4.1.2 Fornecedores de pequeno porte com abrangência nacional 39
4.1.3 Empresas e órgãos que contratam principalmente fora de sua UF 40
4.2 VALIDAÇÃO DA HIPÓTESE DA PESQUISA 42
4.2.1 Estatística descritiva 43
4.2.2 Testes de correlação 44
4.2.3 Análise de sobrepreços 48

5 CONCLUSÕES E TRABALHOS FUTUROS 50

REFERÊNCIAS **52**

TERMINOLOGIA **55**

APÊNDICE A – ANÁLISE DESCRITIVA DAS VARIÁVEIS POR TIPO DE PRODUTO **58**

1 INTRODUÇÃO

O termo georreferenciamento vem do latim geo, que significa terra, e referenciar, tomar como ponto de referência, localizar. Georreferenciar, portanto, significa localizar algum ponto da Terra.

Por outro lado, os sujeitos de um contrato administrativo, segundo a Lei nº. 8.666/93, art. 6º, são: o contratante, o órgão ou a entidade signatária do instrumento contratual; e o contratado, a pessoa física ou jurídica signatária de contrato com Administração Pública.

Essas entidades são partes fundamentais na execução do objeto do contrato, e a localização dessas pode ser crucial para o sucesso da execução contratual. A partir dessas localizações, representadas por coordenadas geográficas, pode-se chegar à distância entre esses sujeitos, que é um importante fator a ser analisado, como será mostrado a seguir.

Kano, Kano e Takechi (2015, p. 1) concluíram em seu estudo que “o efeito da distância é significativamente grande, sugerindo que o preço das barreiras geográficas para o transporte regional é alto”. Entretanto, “os produtores não repassam totalmente o aumento dos custos de transporte”.

Em um país com dimensões continentais como o Brasil, longas distâncias podem acarretar em aumento do tempo de transporte, além de pagamento de taxas e impostos adicionais.

No caso de fornecimento de bens ou serviços comuns ou de baixo valor agregado, longas distâncias podem acarretar em acréscimo significativo no preço final, por causa do custo do frete. Se for perecível, o produto ainda poderá ter sua qualidade deteriorada.

Além dos problemas elencados acima, existem objetos licitados cuja localização geográfica do fornecedor é indispensável para a execução satisfatória do contrato. Exemplo clássico é a contratação de empresa para o fornecimento de combustível. Nesse caso, a localização do posto para o abastecimento é essencial para a eficácia do fornecimento.

É oneroso para a Administração contratar uma empresa distante da sua sede, tendo que se deslocar até lá para proceder com os abastecimentos dos veículos. Tal expediente acarretaria em acréscimo no consumo de combustível e de

tempo gasto para realizar a atividade de abastecimento. Assim sendo, no exemplo apresentado, a consideração da localização geográfica é imprescindível.

Sobre tal aspecto, o Superior Tribunal de Justiça se manifestou:

> […] 3. Conforme a decisão emitida pela Corte de Contas Estadual, não há o que censurar na compra dos combustíveis, quanto há um único posto de abastecimento na cidade; não poderia a Administração concordar que os veículos do Município se deslocassem a longas distâncias para efetuar o abastecimento, com visíveis prejuízos ao Erário […] (SUPERIOR TRIBUNAL DE JUSTIÇA, 2008).

Nesse mesmo sentido, o TCU já foi provocado por uma representação que versava sobre limitação imposta por um edital que objetivava a contratação de empresa especializada para a prestação de serviços de manutenção preventiva, corretiva e assistência técnica para os veículos oficiais pertencentes à frota do Tribunal Regional do Trabalho da 2ª Região. Esse edital exigia que somente poderiam participar empresas sediadas a um raio de 12 km da sede do TRT-2.

Acerca dessa situação, o Tribunal de Contas da União apresentou o seguinte parecer:

> 10. Em tese, a limitação geográfica tem potencial de restringir a participação de empresas, mas pode ser necessária. Caso contrário, a Administração será obrigada a levar seus veículos a oficinas localizadas a distância considerável.
> 11.Cabe considerar que isso demanda não só combustível, como no exemplo hipotético da ora representante, mas também tempo de mão de obra, considerando o motorista que busca e leva o veículo na oficina, ainda mais se for considerado o trânsito caótico das grandes cidades, como é o caso de São Paulo. O custo desse motorista é bastante superior ao mero custo do combustível empregado no deslocamento (TRIBUNAL DE CONTAS DA UNIÃO, 2015).

Em alguns casos, longas distâncias de fornecimento podem tornar a execução contratual mais difícil ou até mesmo impossível, levantando indícios de não haver uma contraprestação por parte do fornecedor.

Tendo em vista o panorama de cunho geográfico apresentado, o presente estudo se depara com o seguinte problema de pesquisa: longas distâncias entre as partes de um contrato administrativo prejudicam o alcance da finalidade pública?

1.1 OBJETIVOS

1.1.1 Objetivo geral

O trabalho tem como objetivo auxiliar o acompanhamento da execução financeira da União quanto aos impactos das distâncias geográficas entre o fornecedor e o consumidor do material, em especial no tocante aos preços contratados.

1.1.2 Objetivos específicos

Tendo como foco alcançar respostas ao problema da pesquisa se têm estipulados os seguintes objetivos específicos:

1. Extrair, explorar e examinar as ordens bancárias emitidas pela União nos últimos cinco anos;
2. Identificar casos de especial interesse, em que a distância seja, realmente, um fator relevante;
3. Encontrar outras dimensões envolvidas no processo;
4. Processar linguagem natural para extração de atributos importantes no trabalho;
5. Utilizar métodos estatísticos e algoritmos de aprendizagem de máquina durante a preparação dos dados e para geração da conclusão do experimento.

2 REFERENCIAL TEÓRICO

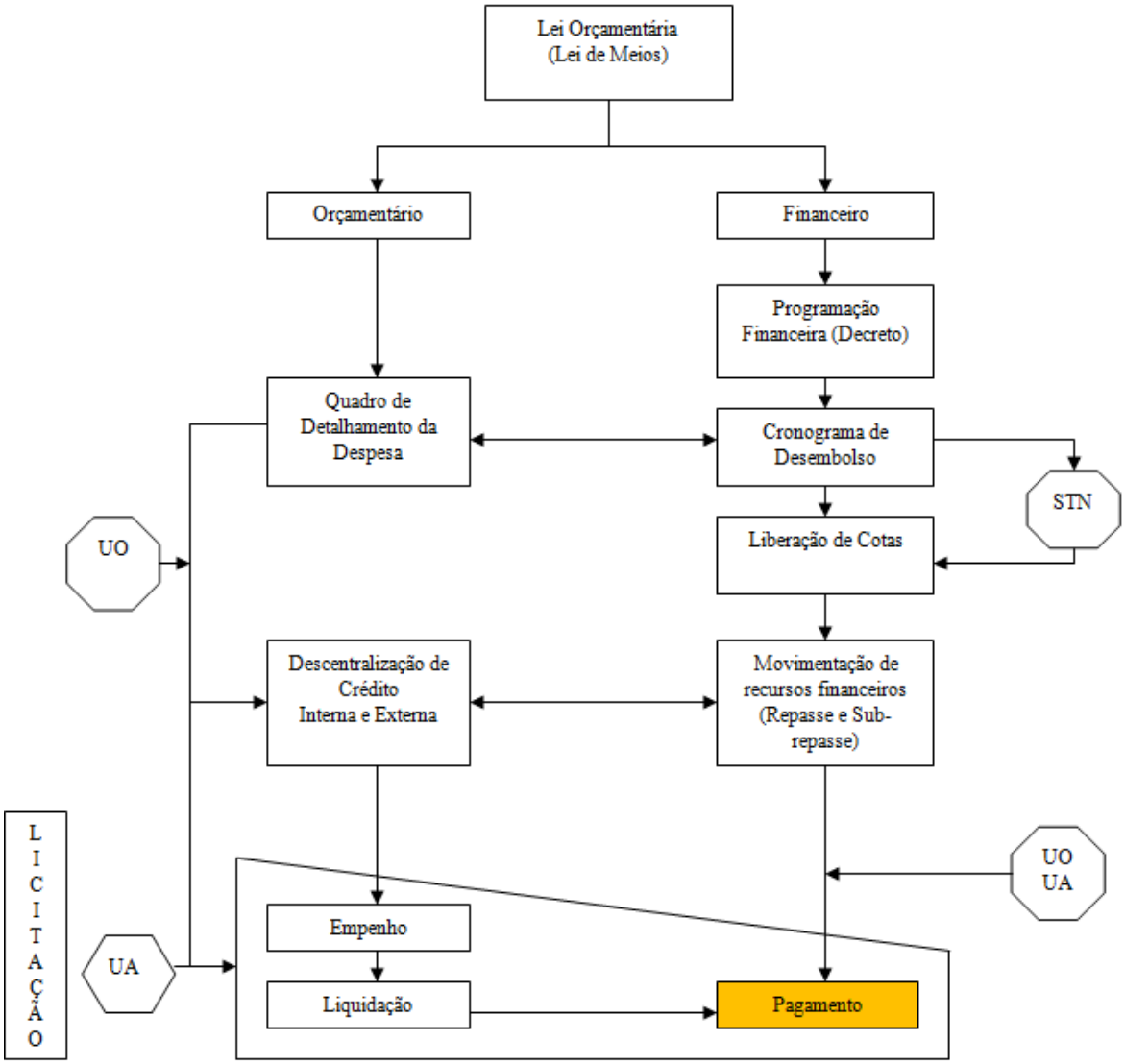
Neste capítulo serão apresentadas fundamentações para o problema e para os objetivos elencados no capítulo anterior, por meio de diversas fontes de pesquisa e de estudos que abordam o assunto, sejam esses digitais ou impressos, de forma a contribuir com as conclusões do experimento.

2.1 EXECUÇÃO ORÇAMENTÁRIA E FINANCEIRA

Segundo Crepaldi e Crepaldi (2013), normalmente há interpretações equivocadas do que venha a ser execução orçamentária e financeira. É perfeitamente compreensível esse equívoco, pois a execução orçamentária e a financeira ocorrem concomitantemente. Essa afirmativa tem como sustentação o fato de que tanto a execução orçamentária quanto a execução financeira estão atreladas uma à outra. Havendo orçamento e não existindo o financeiro, não poderá ocorrer a despesa. Por outro lado, pode haver recurso financeiro, mas não se poderá gastá-lo, se não houver a disponibilidade orçamentária.

Nesse sentido, pode-se definir execução orçamentária como sendo a utilização dos créditos consignados no Orçamento ou Lei Orçamentária Anual - LOA. Já a execução financeira, por sua vez, representa a utilização de recursos financeiros, visando atender à realização dos projetos e/ou atividades atribuídas às Unidades Orçamentárias pelo Orçamento.

Figura 1 – Fluxograma da execução orçamentária e financeira



Fonte: Tribunal de Contas da União (2019).

Da figura acima se verifica que a execução financeira e orçamentária é um processo longo, que vai desde a dotação orçamentária até o pagamento. Esse processo envolve vários agentes, desde legisladores até a unidade executora. É, portanto, sujeito a várias alterações durante o seu transcorrer.

Executar o orçamento e as finanças públicas é realizar as despesas públicas nesse previstas, seguindo à risca os três estágios da execução das despesas que serão apresentados no item 2.1.2.

Vale ressaltar que o foco desse estudo será na etapa derradeira da execução do gasto público: o pagamento, que faz parte da execução financeira.

2.1.1 Execução Financeira da União

A execução financeira representa o fluxo de recursos financeiros necessários à realização efetiva dos gastos dos recursos públicos para a realização dos programas de trabalho definidos. Recurso é dinheiro ou saldo de disponibilidade bancária (enfoque na execução financeira) e crédito é dotação ou autorização de gasto ou sua descentralização (enfoque na execução orçamentária).

De acordo com Jund (2008), o exercício financeiro no Brasil é o espaço de tempo compreendido entre 1º de janeiro e 31 de dezembro de cada ano, no qual a Administração promove a execução orçamentária e demais fatos relacionados com as variações qualitativas e quantitativas, que tocam os elementos patrimoniais da entidade ou órgão público.

2.1.2 Estágios da Despesa Pública

Conforme a Lei n.º 4.320/1964, a despesa pública é executada em três estágios: empenho, liquidação e pagamento e esses serão brevemente explicados a seguir.

Figura 2 – Estágios da Despesa Pública

Fonte: Controladoria Geral da União (2019)

O empenho é uma etapa na qual o governo faz a reserva do recurso para o pagamento de um bem ou de um serviço que se pagará ao termino do mesmo, sendo essa fase a que propicia organizar os gastos de diversas áreas do Governo, evitando que se gaste mais do que foi planejado.

A fase denominada de liquidação é quando se verifica que o governo recebeu aquilo que comprou. Ou seja, quando se confere que o bem foi entregue corretamente ou que a etapa da obra foi concluída como acordado, por exemplo.

Por fim, se estiver tudo em conformidade com as fases anteriores, o Governo pode fazer o pagamento, repassando o valor ao vendedor ou prestador de serviço contratado.

A seguir serão apresentados mais detalhes acerca desta última etapa, foco deste experimento, o pagamento.

2.1.2.1 Pagamento

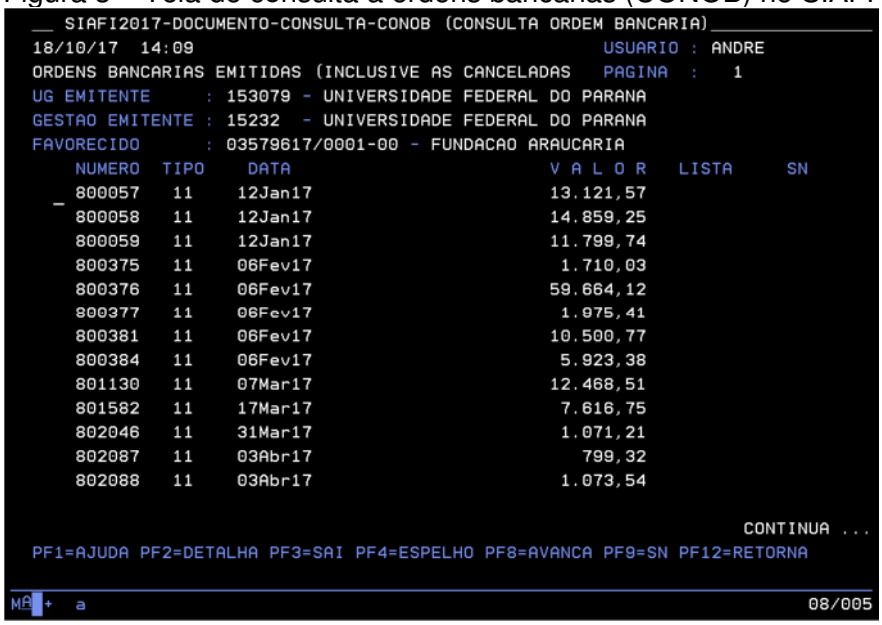
O pagamento corresponde ao último estágio da execução de despesa, envolvendo o dispêndio de recursos financeiros oriundos do Orçamento Geral da União e se faz exclusivamente por meio de Ordem Bancária (OB) e da Conta Única do Governo Federal. Ele se destina ao pagamento de compromissos, bem como a transferência de recursos entre as Unidades Gestoras, tais como liberação de recursos para fins de adiantamento, suprimento de fundos, cota, repasse, sub-repasse e afins. A OB é, portanto, o único documento de transferência de recursos financeiros (BIDERMAN, 2013).

2.2 SISTEMA INTEGRADO DE ADMINISTRAÇÃO FINANCEIRA DO GOVERNO FEDERAL

De acordo com exposição de Silva, Mota e Pinto (2008), o Sistema Integrado de Administração Financeira do Governo Federal, conhecido como SIAFI, é um sistema contábil informatizado que processa e controla a execução orçamentária, financeira, patrimonial e contábil da União. Ele foi implantado em 1987, tornando-se, desde então, importante instrumento para o acompanhamento e controle da execução orçamentária, financeira, patrimonial e contábil do Governo Federal. Atualmente, se apresenta como um dos mais abrangentes instrumentos de administração das finanças públicas, dentre os seus congêneres conhecidos no mundo.

.

Figura 3 – Tela de consulta a ordens bancárias (CONOB) no SIAFI



Fonte: SIAFI, 2015

Explicam Silva, Palmeira e Quintana (2007) que o Governo Federal até o ano de 1986 administrava os gastos com base na posição do caixa, controlando as contas bancárias por meio dos registros no Banco do Brasil, de forma que os gestores não tinham segurança real dos gastos da Administração Pública, sendo urgente naquela época uma solução para a falta de informações acerca das finanças do Governo, que também tinha como agravante a condição de taxas inflacionárias que transformavam os orçamentos em peças que nem sempre se conseguiam executar.

Também contribuía para a falta de uniformidade o fato de que apenas um órgão arrecadava, mas vários faziam execução, cada qual com seu critério sobre o mesmo assunto.

O SIAFI representou, à época, uma tentativa bem-sucedida de controle dos recursos governamentais através das TICs, pois conseguiu unificar e organizar as contas governamentais por meio da Conta Única do Tesouro Nacional, assim como padronizar diversos tipos de lançamentos contábeis (STN, 2019).

Figura 4 – Marcos do reordenamento das finanças públicas

Fonte: Silva, Mota e Pinto (2008).

O Decreto nº 92.452 de 1986 constituiu o Sistema de Administração Financeira Federal e o Sistema de Contabilidade Federal. Por meio dessa e de outras normas foi possível alcançar maior transparência em gastos públicos, propiciando, com isso, a entrega de *accountability* para toda sociedade brasileira.

2.3 TESOURO GERENCIAL

A Secretaria do Tesouro Nacional (STN) passou a partir de 2015 a disponibilizar uma nova plataforma para ofertar informações gerenciais relacionadas à execução orçamentária, financeira, contábil e patrimonial da União, sendo tal sistema uma forma de substituição para o SIAFI Gerencial, que era utilizado com tal finalidade desde 1995, e que já possuía diversas limitações decorrentes da própria evolução da tecnologia.

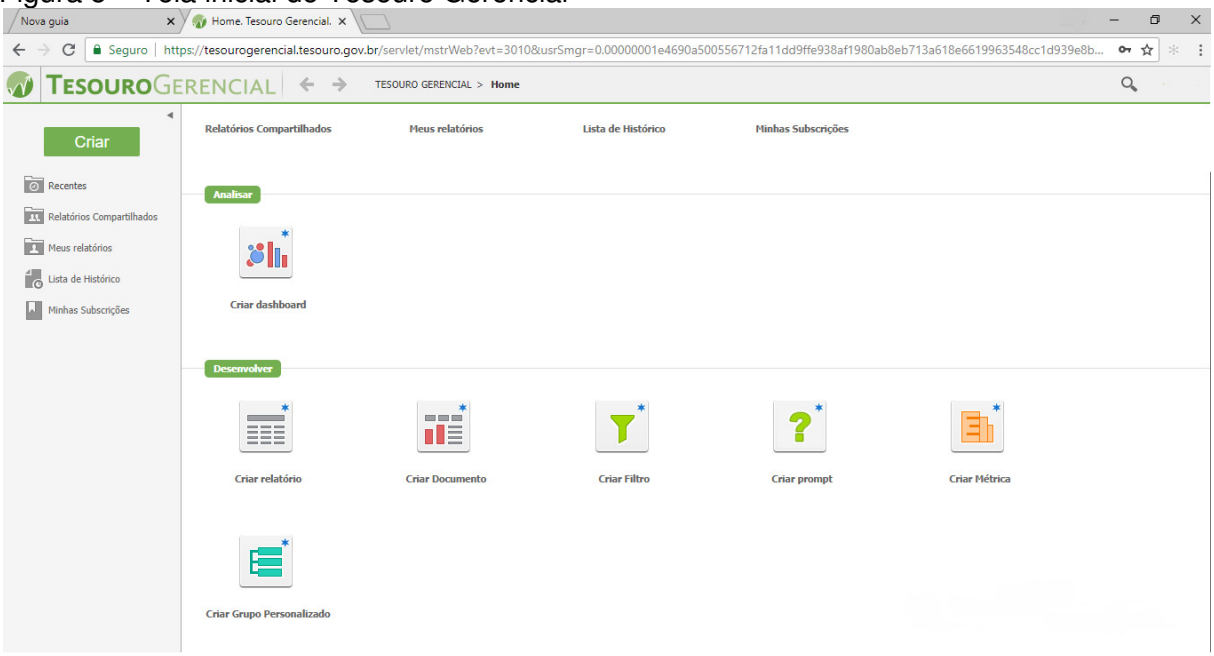
O Tesouro Gerencial surgiu como ferramenta para unificar dados provenientes de vários sistemas estruturantes do governo, tendo sido desenvolvido pelo Serpro, com tecnologia baseada na plataforma de BI proprietária da MicroStrategy.

Esse sistema ampliou a gama de informações que estavam disponíveis aos usuários do SIAFI Gerencial até aquele momento e permitiu análises mais consistentes no apoio de decisões do governo em diversos aspectos, inclusive para melhoria de políticas públicas. A partir de então, um gestor público utilizando o Tesouro Gerencial poderia consultar todos os lançamentos contábeis efetuados no SIAFI no dia anterior, no seu maior nível de detalhe.

De acordo com explicação de Souza (2015), entre as possibilidades que se abriram com a solução, está a construção de painéis de indicadores com atualização diária e automática e visualização por meio de dispositivos móveis (SOUZA, 2015).

O acesso ao portal do Tesouro Gerencial é disponibilizado aos servidores pelos próprios órgãos da Administração Pública a que esses estão vinculados.

Figura 5 – Tela inicial do Tesouro Gerencial



Fonte: Tesouro Gerencial (2019).

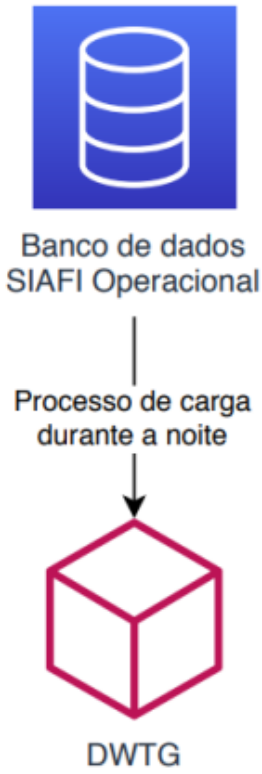
Outro aspecto importante do sistema é que ele é capaz de gerar automaticamente consultas SQL para que se faça execução direta no Data Warehouse, sendo esse componente o objeto de estudo da próxima seção.

2.3.1 Data Warehouse do Tesouro Gerencial

O Data Warehouse do Tesouro Gerencial (DWTG) é uma base de dados Teradata desnormalizada, que tem suas informações provenientes do SIAFI.

O processo de carga do SIAFI para o DWTG (Teradata) tem periodicidade diária e está ilustrado no diagrama abaixo.

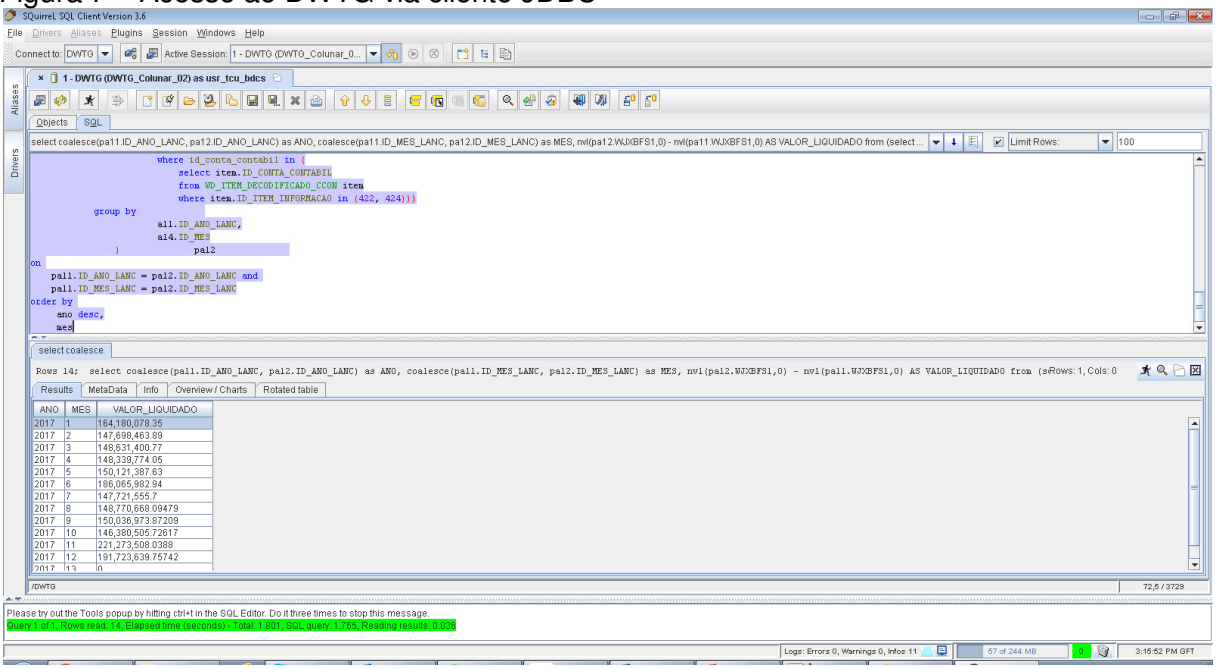
Figura 6 – Processo de carga do DWTG



Fonte: Elaboração própria.

O acesso ao DWTG pode ocorrer por meio de um cliente JDBC, que seja adequadamente configurado. Através dessa interface, mostrada a seguir, é possível enviar consultas SQL personalizadas para suprir qualquer necessidade de informação constante nessa base.

Figura 7 – Acesso ao DWTG via cliente JDBC



Fonte: Elaboração própria.

O Tribunal de Contas da União (TCU) disponibiliza no ambiente LabContas um subconjunto do DWTG, que pode ser acessado por seus servidores e colaboradores. A partir dessa fonte se construiu o experimento apresentado neste trabalho.

3 METODOLOGIA APLICADA

A pesquisa desenvolvida pode ser classificada conforme explicação de Gil (2008) como exploratória e explicativa em relação aos objetivos estabelecidos. Exploratória, porque consiste na realização de estudo com foco em familiarização do pesquisador com o objetivo investigado; explicativa, pois busca identificar os fatores que determinam ou que contribuem para a ocorrência dos fenômenos.

Sobre os procedimentos metodológicos, a pesquisa se classifica como experimental em que se determina um objeto de estudo, seleciona-se as variáveis que seriam capazes de influenciá-lo e define-se as formas de controle e de observação dos efeitos que a variável produz no objeto.

O problema tem como abordagem a percepção quantitativa, e o tipo de pesquisa pode ser classificado como de campo, empírica, uma vez que a investigação não é restrita apenas a aspectos teóricos, já que a ênfase decorre da análise de dados concretos que são extraídos por meio de observação de fatos.

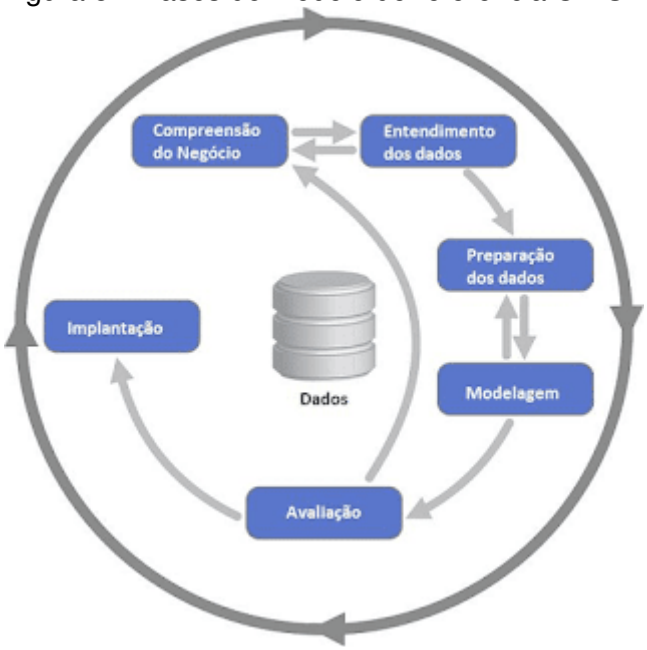
Foi adotada como metodologia de mineração de dados o modelo de referência CRISP-DM (do inglês, Cross-Industry Standard Process for Data Mining), sendo esse modelo detalhado a seguir, da mesma forma que as etapas de execução do experimento.

3.1 CRISP-DM

Segundo explica Shearer (2000), em 1986 foi criado o Grupo de Trabalho CRISP-DM, cujo foco de atuação se centrou em padronização de conceitos e de técnicas que propiciam informações especificas para tomada de decisões. Esse grupo propôs uma metodologia como o mesmo nome, destinada a auxiliar administradores e responsáveis no processo geral de planejar e executar a mineração de dados, englobando desde a especificação do processo até a apresentação dos resultados.

Essa metodologia é constituída por seis etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. Essas fases são executadas de forma iterativa e cíclica, podendo ser compreendidas como um ciclo de vida de mineração de dados, mostrado na figura abaixo.

Figura 8 – Fases do modelo de referência CRISP-DM



Fonte: Traduzido de Chapman (2000).

3.1.1 Entendimento do negócio

Conforme Wirth e Hipp (2000), essa fase inicial se concentra no entendimento dos objetivos e requisitos do projeto sob uma perspectiva de negócio e, em seguida, na conversão desse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar do projeto desenvolvido para atingir os objetivos.

Neste caso concreto, durante a fase de entendimento do negócio foram executadas as seguintes tarefas:

1. Estudo da legislação aplicável para identificação dos dispositivos relacionados ao tema;
2. Discussão com o orientador sobre as questões de negócio envolvidas no trabalho;
3. Identificação das situações de destaque no tocante à distância geográfica entre as partes de um contrato;
4. Análise de inventário e escolha de fontes de dados disponíveis para o experimento.

A seguir será apresentada uma importante tarefa executada durante essa fase, o inventário de recursos. Alguns detalhes de negócio resultantes dessa etapa do CRISP-DM já foram melhor detalhados nos capítulos anteriores deste documento

3.1.1.1 Inventário e escolha das fontes de dados

Como ponto de partida deste experimento havia diversas alternativas de fonte de dados, entre esses ComprasNet, SIASG, NF-e e Tesouro Gerencial. Após uma análise minuciosa, o Tesouro Gerencial foi escolhido em virtude das seguintes características em relação aos demais:

- Maior abrangência (engloba todos os pagamentos da União, independentemente de valores ou objeto contratado);
- Maior qualidade de dados (como se trata do dado que realmente foi desembolsado pela União é, portanto, menos passível de posteriores correções ou modificações);
- Maior disponibilidade de dados (os dados estão em completa disponibilidade no TCU, ao contrário da base de NF-e, por exemplo).

Cabe ressaltar que essas fontes não são mutuamente excludentes, mas complementares, e podem ou devem ser utilizadas em conjunto. Porém, em virtude da limitação na capacidade de processamento dos dados causada pelo *hardware* e *software* utilizado no experimento, optou-se por não utilizar outras fontes além do DWTG. Essa grande demanda de recursos se deveu ao amplo escopo da análise.

Adicionalmente ao DWTG, foram utilizadas como coadjuvantes as bases do IBGE e da Receita Federal do Brasil, para a obtenção de informações sobre municípios e empresas incluídas na análise, respectivamente.

3.1.2 Entendimento dos Dados

A etapa de entendimento de dados tem seu início com uma coleta preliminar de dados e prossegue com atividades para familiarização do pesquisador com os dados. Nessa fase também é possível identificar problemas de qualidade dos dados, descobrir as primeiras ideias sobre os dados ou detectar subconjuntos interessantes para formar hipóteses.

Existe um vínculo estreito entre o entendimento de negócio e o entendimento de dados. A formulação do problema de mineração de dados e o plano do projeto requerem pelo menos alguma compreensão dos dados disponíveis (WIRTH; HIPP, 2000).

A seguir serão mostrados os passos executados durante essa fase.

3.1.2.1 Coleta de dados inicial

As informações estão disponíveis em um servidor de banco de dados SQL Server, no LabContas. Nesse ambiente está disponível uma massa enorme de dados de pagamentos desde o ano de 2015: 48.787.425 registros de pagamentos para 2.735.762 favorecidos distintos, totalizando o montante de R$ 9.954.654.088.180,39. Na época da extração, a base estava atualizada até o dia 16/08/2019.

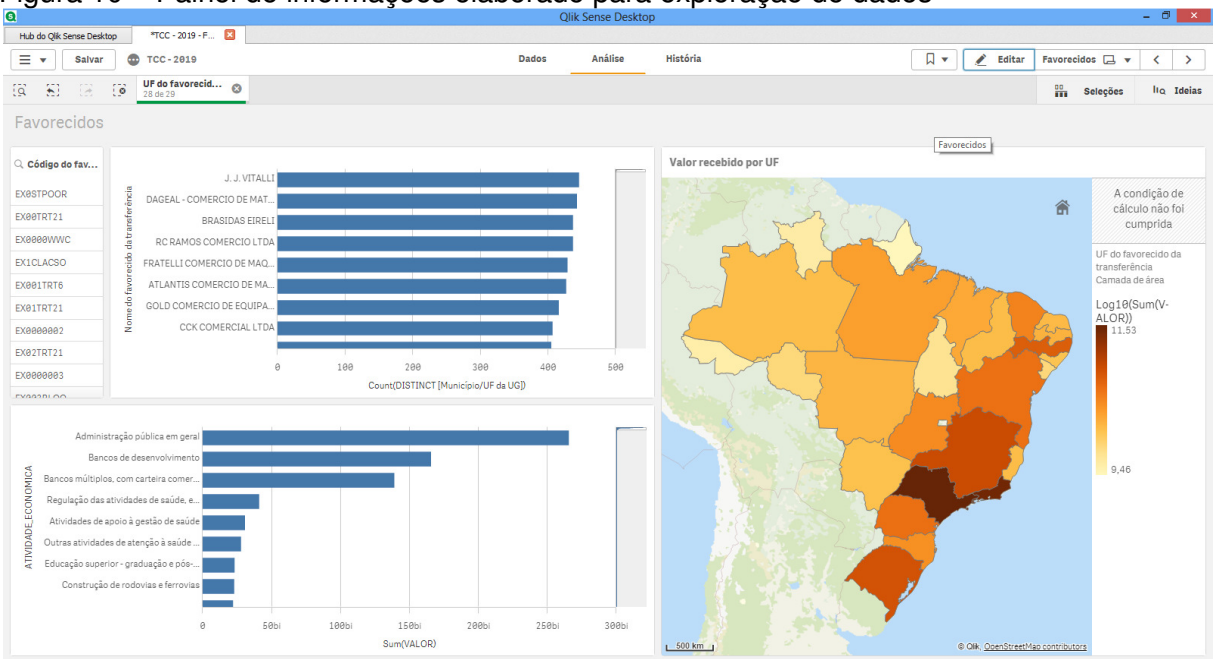
Figura 9 – Números totais relacionados aos pagamentos no DWTG

Fonte: Elaboração própria.

3.1.2.2 Exploração dos dados

A exploração dos dados foi feita com auxílio do *software* Qlik Sense. Essa plataforma de BI foi escolhida em virtude de seu desempenho, capacidade de customização e praticidade na atividade exploratória e na criação de mapas.

Figura 10 – Painel de informações elaborado para exploração de dados



Fonte: Elaboração própria.

Adicionalmente, foi utilizada a biblioteca QVD Map Library para a geração de mapas como o da figura acima. Essa biblioteca para Qlik Sense disponibiliza *shapes* de polígonos para todas as UFs e municípios brasileiros.

Os *insights* produzidos e os resultados das atividades executadas nessa etapa com o auxílio do painel mostrado acima serão documentados detalhadamente no capítulo 4.

3.1.3 Preparação dos dados

A denominada fase de preparação de dados envolve as atividades que são aplicadas para construir o conjunto de dados final utilizado na modelagem, conforme explicam Wirth e Hipp (2000). Assim, as tarefas de preparação de dados geralmente são executadas por diversas vezes e não em uma ordem fixa. Dentre essas tarefas, pode-se citar a seleção de tabela, de registros e de atributos, limpeza de dados, construção de atributos derivados e transformação de dados.

3.1.3.1 Descrição do conjunto de dados

Na tabela abaixo estão relacionados os principais atributos relacionados ao pagamento disponível no BD_DWTG e que apoiarão nesta análise. Esses estão

classificados em dois tipos: a) dados qualitativos, que geralmente são expressos por atributos categóricos; e b) dados quantitativos, expressos em quantidades e valores.

Tabela 1 – Descrição do conjunto de dados

Atributo	Tipo	Descrição
Número da OB	Qualitativo	Número que identifica unicamente a Ordem Bancária (chave primária).
Número da NE	Qualitativo	Número que identifica unicamente a Nota de Empenho relacionada à OB acima (chave estrangeira).
Data de emissão da OB	Qualitativo	Data em que a OB foi emitida.
Observação da OB	Qualitativo	Campo de texto livre com informações importantes sobre a OB. Geralmente informa o fim a que se destina o pagamento e a nota fiscal relacionada a ele.
Lista credor da OB	Qualitativo	Número com a com o código da lista de credores relacionados à OB (chave estrangeira).
Tipo de favorecido da transferência	Qualitativo	Informa se o favorecido é pessoa física ou jurídica.
Código do favorecido da transferência	Qualitativo	Identifica o recebedor do pagamento. Geralmente é um CPF ou CNPJ, mas pode assumir códigos de inscrições genéricas, no caso de entidades domiciliadas no exterior, pessoas físicas ou reembolso de despesas.
Nome do favorecido da transferência	Qualitativo	Nome do recebedor do pagamento.
Código do município do favorecido da transferência	Qualitativo	Código do município do recebedor do pagamento na base de dados do IBGE.
Código do município da UG	Qualitativo	Código IBGE do município onde está sediada a unidade gestora responsável pelo pagamento.
Valor da transferência	Quantitativo	Valor do pagamento, em R$.

Fonte: Elaboração própria.

3.1.3.2 Criação de atributos derivados

Durante a etapa de preparação dos dados, atributos são transformados de modo a criar um conjunto de informações que agregue mais valor à solução proposta. A seguir serão apresentados dois importantes atributos, criados a partir dos dados brutos do DWTG.

3.1.3.2.1 Distância entre duas coordenadas

A distância entre o contratante e o contratado surge como informação fundamental para este trabalho, sendo este atributo derivado dos municípios onde estão sediados a UG e o favorecido do pagamento. Ao se fazer uma simples consulta à base de dados do IBGE, é possível descobrir as coordenadas geográficas (pares de latitude e longitude) onde estão localizadas essas cidades.

Para o cálculo da menor distância em quilômetros, entre duas coordenadas, foi utilizada a seguinte equação, derivada da lei dos cossenos para triângulos esféricos (PEREIRA; MOREY, 2017):

$$d = \text{acos}(\sin \varphi 1 \cdot \sin \varphi 2 + \cos \varphi 1 \cdot \cos \varphi 2 \cdot \cos \Delta\lambda) \cdot R$$

Em que φ representa latitude, λ representa longitude e R representa a constante de raio da Terra, de aproximadamente 6.371 quilômetros.

3.1.3.2.2 Preço unitário do produto

Outro importante atributo criado durante esta etapa foi o referente ao preço unitário, pois este é derivado da divisão do valor total do pagamento pela quantidade de itens que foram efetivamente pagos, mas essa última informação não é trivial de ser obtida.

Em um primeiro momento, foi avaliada a possibilidade de extrair a informação sobre a quantidade da lista de itens da nota de empenho. Porém se verificou que os números lá presentes não eram precisos, pois nem tudo que é empenhado é efetivamente pago. Foram identificados diversos casos em que o valor do empenho foi reduzido, mas a lista de itens não foi atualizada pelo gestor do SIAFI, no momento do pagamento.

Portanto, a solução adotada foi extrair apenas a descrição do item e sua respectiva unidade de medida de um campo textual chamado descrição do item, na lista de itens da nota de empenho.

Por questão de praticidade, foram consideradas apenas OBs que pagavam empenhos que possuíam apenas um tipo de produto na lista de itens. Serviços contratados também foram excluídos durante essa etapa, por serem objetos cuja mensuração é mais complexa. Somente em virtude desta limitação de escopo, o

número de registros citados na seção 3.1.2.1 foi reduzido para aproximadamente 1,2 milhões (um milhão e duzentos mil) antes do processamento textual.

A partir desses dois atributos (descrição do item e unidade de medida) utilizou-se uma biblioteca Python para o tratamento de expressões regulares no campo de observação da OB e foi extraída a informação a respeito da quantidade de itens efetivamente paga.

Figura 11 – Trecho de código Python utilizando expressões regulares

```
%%time
import re

df['quantidade_item'] = 0
for index, row in df.iterrows():
    lista = re.findall(r" ([\d+\.]*\d+) " + row['unidade_medida'].split()[0], row['observacao_ob'])
    if (row['unidade_medida'] == 'UN'):
        lista.extend(re.findall(r" ([\d+\.]*\d+) " + row['nome_item'].split()[0], row['observacao_ob']))
    df.at[index,'quantidade_item'] = sum(list(map(int, [l.replace('.', '') for l in lista])))
```

```
CPU times: user 7min 21s, sys: 186 ms, total: 7min 21s
Wall time: 7min 22s
```

Fonte: Elaboração própria.

Como o campo de observação é um texto de livre criação pelo usuário, só foi possível obter a informação sobre quantidade para cerca de 1% dos registros, mas como será necessária apenas uma amostra para validar a hipótese desse estudo, tal limitação não se transformou em impedimento.

3.1.3.3 Remoção de outliers

O termo *outlier* dentro da estatística representa um valor atípico, um registro que possui afastamento dos demais registros da série ou que seja inconsistente. A existência de registros nessa condição tipicamente causa prejuízos de interpretação dos resultados dos testes estatísticos que se aplicam para as amostras coletadas.

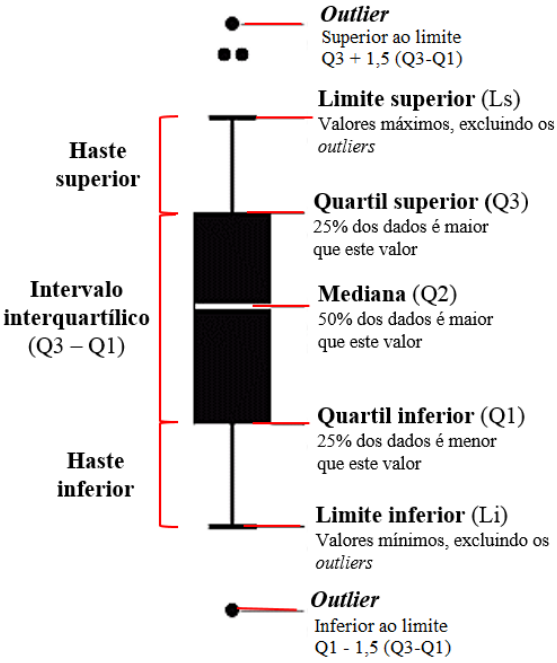
Existem vários métodos para identificação de *outliers* e entre esses podem ser citados: gráfico de *boxplot*, modelos de discordância, teste de Dixon, teste de Grubbs e Z-scores. Cada um desses possui suas características próprias e se adapta melhor em certas situações.

O método que teve o melhor comportamento para a detecção de *outliers* de preços para esta análise foi o *boxplot*.

De acordo com explicação de Valladares Neto et al. (2017), o boxplot foi empregado pelo matemático americano John Turkey (1915-2000) no ano de 1970, e teve maior divulgação a partir de 1977.

O boxplot é um recurso visual que resume os dados e consegue exibir, em uma só figura, a mediana, os quartis e os valores pontuais máximos e mínimos. Portanto, apresenta valores de tendência central, dispersão e simetria dos dados agrupados.

Figura 12 – Informações contidas em um gráfico *boxplot*



Fonte: Adaptado de Valladares Neto *et al*. (2017).

No caso em estudo, como os dados sobre quantidades de produtos foram derivados de um campo textual livre, estão bastante sujeitos a erros. Para corrigir essas distorções foi utilizada a linguagem Python para excluir da amostra de validação essas observações consideradas inconsistentes, consideradas *outliers.*

Após diversas experimentações, foi estabelecido o tamanho amostral mínimo de 17 observações para cada tipo de produto antes da aplicação dessa técnica, de forma com que as medianas e limites ficassem mais bem estabelecidos para cada tipo de produto. Em virtude dessa decisão, o tamanho amostral foi reduzido para menos da metade antes de remoção de *outliers.*

Após a remoção de todos os valores atípicos, o tamanho da amostra de validação ficou limitado a apenas 1.075 registros. A partir dessa amostra será validada a hipótese de pesquisa.

3.1.4 Modelagem

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ideais. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas requerem formatos de dados específicos.

Existe um vínculo estreito entre a preparação de dados e a modelagem. Muitas vezes são percebidos problemas de dados durante a modelagem ou se obtêm ideias para construir novos dados (WIRTH; HIPP, 2000).

3.1.4.1 Mineração de dados

Segundo explicam Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de descoberta de conhecimento decorre da mineração de dados e a correspondente análise dos mesmos com aplicação de algoritmos de descoberta, que envolvem limitações computacionais, bem como produzem padrões de dados em agrupamentos.

Assim, a mineração de dados tem como objetivo a aplicação de uma ou mais tarefas, sendo as mais populares a classificação, a regressão, a associação, o agrupamento e a sumarização de dados, que são expostos na tabela resumo que segue com as tarefas mais comuns, descrições e exemplos de aplicação.

Tabela 2 – Principais tarefas de mineração de dados

Tarefa	Descrição	Exemplos
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes.	• Classificar pedidos de crédito; • Esclarecer pedidos de seguros fraudulentos; • Identificar a melhor forma de tratamento de um paciente.
Regressão	Usada para definir um valor para alguma variável contínua desconhecida.	• Estimar o número de filhos ou a renda total de uma família; • Estimar o valor em tempo de vida de um cliente; • Prever a demanda de um consumidor para um novo produto.
Associação	Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação.	• Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado.
Agrupamento	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos.	• Agrupar clientes por região do país; • Agrupar clientes com comportamento de compra similar; • Agrupar seções de usuários Web para prever comportamento futuro de usuário.
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados.	• Tabular o significado e desvios padrão para todos os itens de dados; • Derivar regras de síntese

Fonte: Dias (2001).

Neste experimento foi utilizado agrupamento para separar distâncias (numéricas) em três categorias: pequena, média e grande.

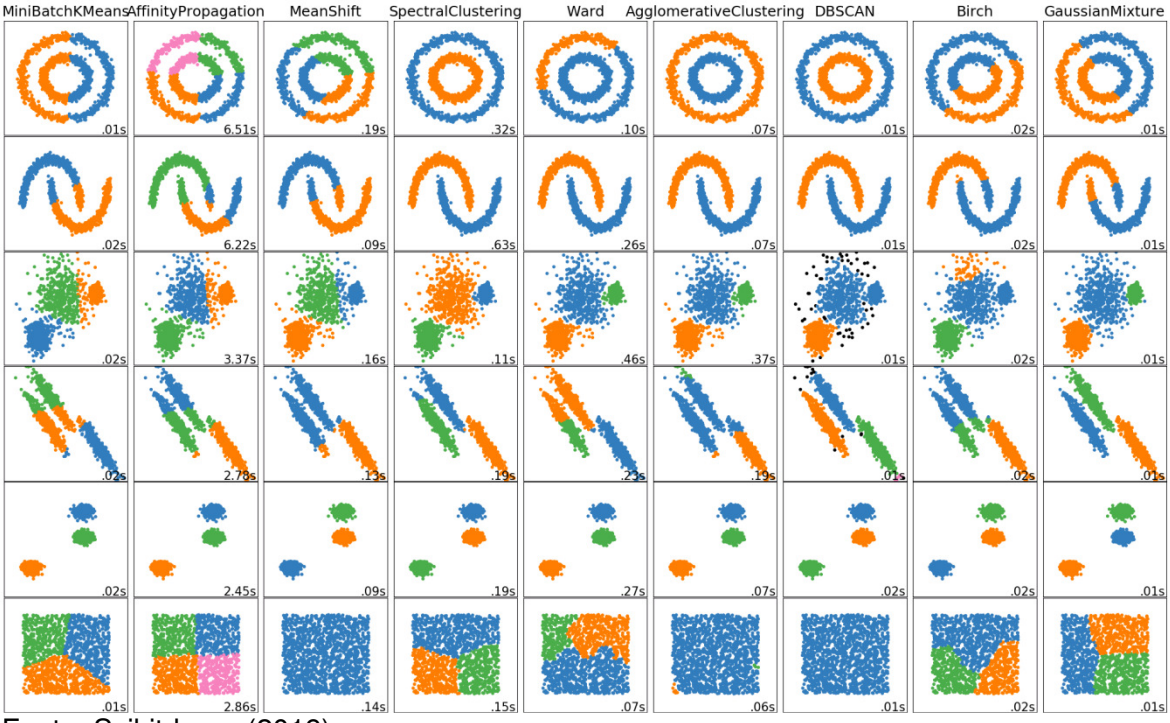
3.1.4.1.1 Agrupamento

A tarefa de agrupamento, segundo explicam Camilo e Silva (2009), tem como foco a identificação de registros similares, sendo um cluster ou grupo uma coleção de registros que podem ser vistos como similares entre si, mas diferentes de outros registros nos demais grupos.

Essa tarefa difere da classificação, pois não necessita que os registros sejam previamente categorizados. Por isso é chamada de aprendizado não supervisionado. Essa tarefa também não tem a pretensão de classificar, estimar ou predizer o valor de uma variável. Ela apenas identifica os grupos de dados similares.

A biblioteca Python utilizada no experimento, chamada *scikit-learn,* disponibiliza vários métodos de agrupamentos diferentes, que devem ser utilizados conforme a distribuição dos dados e as características dos grupos que se deseja criar. A figura abaixo ilustra as combinações dessas variáveis.

Figura 13 – Comparativo entre diferentes algoritmos de agrupamento



Fonte: Scikit-learn (2019).

Por meio do experimento, o melhor resultado identificado para o objetivo envolveu o agrupamento k-Means.

De acordo com explicação de Camilo e Silva (2009), o algoritmo k-Means utiliza o conceito de centroide e, inicialmente, faz a seleção aleatória de k registros, cada um representando um agrupamento. Para cada registro restante, é calculada a similaridade entre o registro analisado e o centro de cada agrupamento. O objeto é inserido no agrupamento com a menor distância, ou seja, maior similaridade. O centro do cluster é recalculado a cada novo elemento inserido.

3.1.4.2 Correlação de Spearman

O coeficiente de correlação de Spearman (p) mede a intensidade da relação entre duas variáveis. A correlação de Spearman entre duas variáveis é similar à correlação de Pearson entre os valores de postos daquelas duas variáveis.

Enquanto a correlação de Pearson avalia relações lineares, a correlação de Spearman avalia relações monótonas, sejam elas lineares ou não, explicam Kendall e Gibbons (1990).

Os valores para esse coeficiente variam de -1 a 1. O sinal indica a direção positiva ou negativa do relacionamento e o valor representa a intensidade da relação entre as variáveis. Uma correlação de valor zero indica que não há relação entre as variáveis. Ou seja, elas são estatisticamente independentes.

3.1.5 Avaliação

De acordo com Wirth e Hipp (2000), nesta fase do projeto são construídos um ou mais modelos que parecem ter alta qualidade, sob uma perspectiva de análise de dados. Antes de prosseguir para a implantação final do modelo, é importante avaliar mais detalhadamente o modelo e revisar as etapas executadas para construir o mesmo, para garantir que esse atinja adequadamente os objetivos de negócios. Um objetivo principal é determinar se há algum problema de negócio importante que não foi suficientemente considerado. Ao final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.

Após uma análise descritiva geral dos dados foi realizado um teste de correlação de Spearman para verificar se há uma correlação entre o preço unitário e a distância de fornecimento. O teste de Spearman foi escolhido como o mais adequado após a realização dos testes de Shapiro-Wilk e Kolmogorov-Smirnov, que apresentaram p-valor inferior a 0,05, indicando que nenhuma das variáveis analisadas apresenta uma distribuição normal. Portanto, os testes não paramétricos são os mais adequados.

Para todos os testes realizados, no presente trabalho, o nível de significância adotado foi de 5% e como medida de decisão foi adotado o p-valor, ou seja, quando o p-valor encontrado for inferior a 0,05, esse indica que o resultado tem significância estatística.

As tarefas de avaliação dos resultados e determinação dos próximos passos, incluídas nessa fase do modelo CRISP-DM, serão apresentadas nos dois próximos capítulos.

3.1.6 Implantação

De acordo com exposição de Wirth e Hipp (2000), a construção do modelo não implica finalização do projeto, pois há necessidade de organização e de sistematização das informações para que se possa utilizá-las. Dependendo dos requisitos, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível.

Em muitos casos, será o usuário quem executará as etapas de implantação e não o analista, sendo relevante entender quais ações devem ser realizadas, para então aplicar os modelos que foram criados.

Os resultados obtidos nesse experimento serão detalhados no próximo capítulo. Os arquivos fontes (scripts de carga SQL para o painel de informações e Jupyter notebook utilizado para validação) estão no repositório[1] criado para este projeto.

No TCU, o produto deste trabalho poderá ser integrado com soluções já desenvolvidas no Tribunal, como o painel de Despesas da Administração Pública Federal, por exemplo.

[1] Disponível em https://gitlab.com/evilar/distancias-dwtg/tree/master.

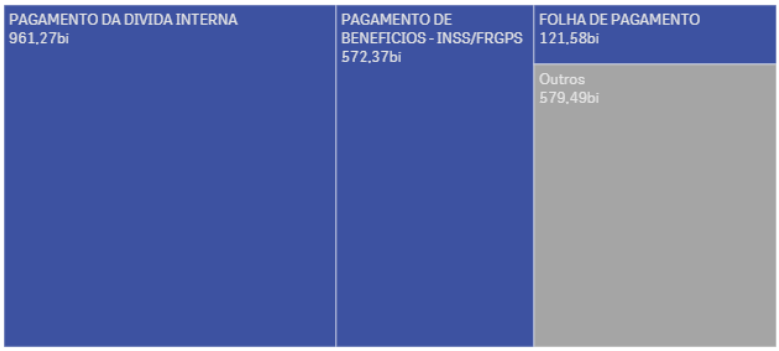
4 RESULTADOS OBTIDOS

Os resultados serão apresentados em duas partes: primeiramente, serão mostradas situações de especial interesse descobertas durante a análise exploratória inicial, utilizando toda a massa de dados de pagamento do Tesouro Gerencial, desde o ano de 2015. Posteriormente, sobre uma amostra, será validada uma hipótese criada durante essa exploração inicial: se longas distâncias de fornecimento podem impactar o preço final do produto e, consequentemente, prejudicar o alcance da finalidade pública.

4.1 ANÁLISE EXPLORATÓRIA

Conforme citado na Seção 3.1.2.1, a massa de dados inicialmente analisada reúne 48.787.425 registros de pagamento para 2.735.762 favorecidos distintos, totalizando o montante de aproximadamente R$ 10 trilhões. Essa exploração foi realizada com auxílio do *software* Qlik Sense para criação de um painel de informações.

Figura 14 – Principais finalidades dos pagamentos federais em 2018



Fonte: DWTG (2019).

Da figura acima se depreende que a maior parte dos pagamentos da União, cerca de 74%, é destinada para despesas de caráter obrigatório e continuado. Essa situação provoca uma redução da flexibilidade orçamentária, resultando em menor capacidade de manejo dos recursos públicos por parte dos agentes políticos.

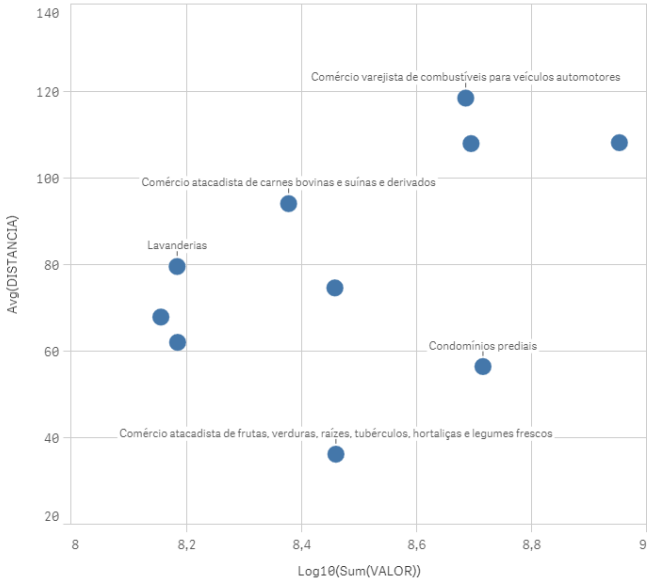
Durante essa fase inicial da pesquisa surgiram casos que, apesar de não serem flagrantemente ilegais, podem levantar suspeitas e merecem ser destacados. Algumas dessas situações serão mencionadas a seguir.

4.1.1 Fornecimento de produtos e serviços comuns a grandes distâncias

No caso de fornecimento de bens ou serviços comuns ou de baixo valor agregado, longas distâncias podem acarretar em acréscimo significativo no preço final, por causa do custo do frete ou do deslocamento. Se for perecível, o produto ainda poderá ter sua qualidade deteriorada.

A figura a seguir mostra situações em que a distância tem caráter fundamental na relação de prestação de bens ou serviços.

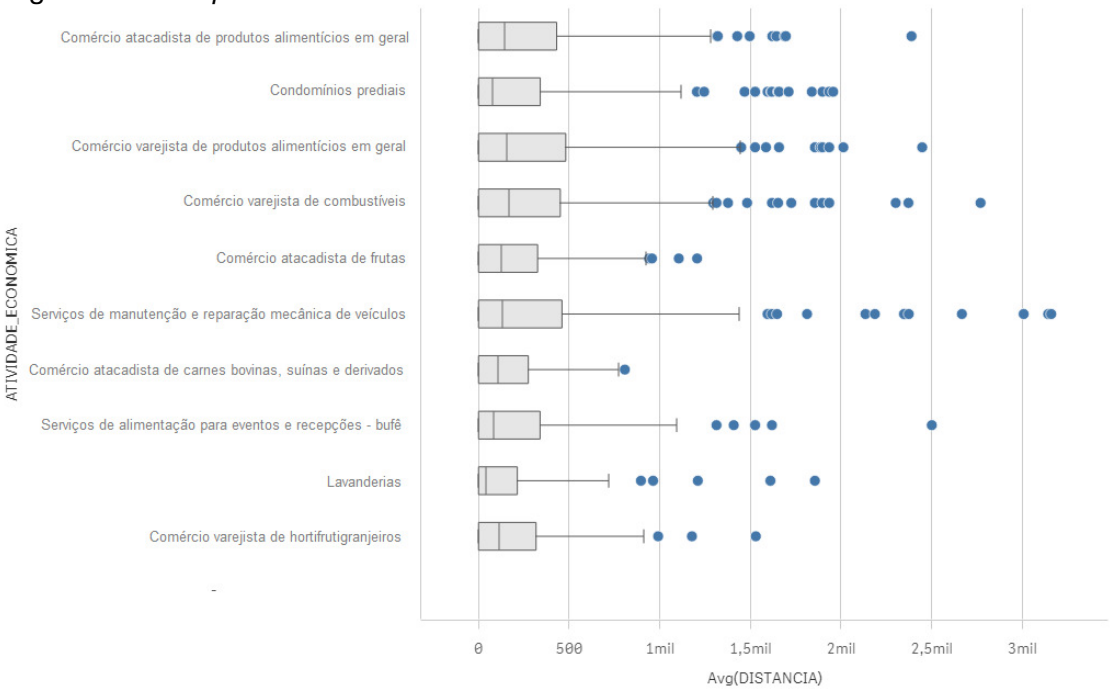
Figura 15 – Dispersão: Valores x Distâncias por CNAE



Fonte: DWTG (2019).

Geralmente, nesses casos as distâncias são curtas, em virtude do caráter mencionado acima. Porém, existem situações de *outliers* de distância que podem, a princípio, levantar suspeitas.

Figura 16 – *Boxplot*: *Outliers* de distância



Fonte: DWTG (2019).

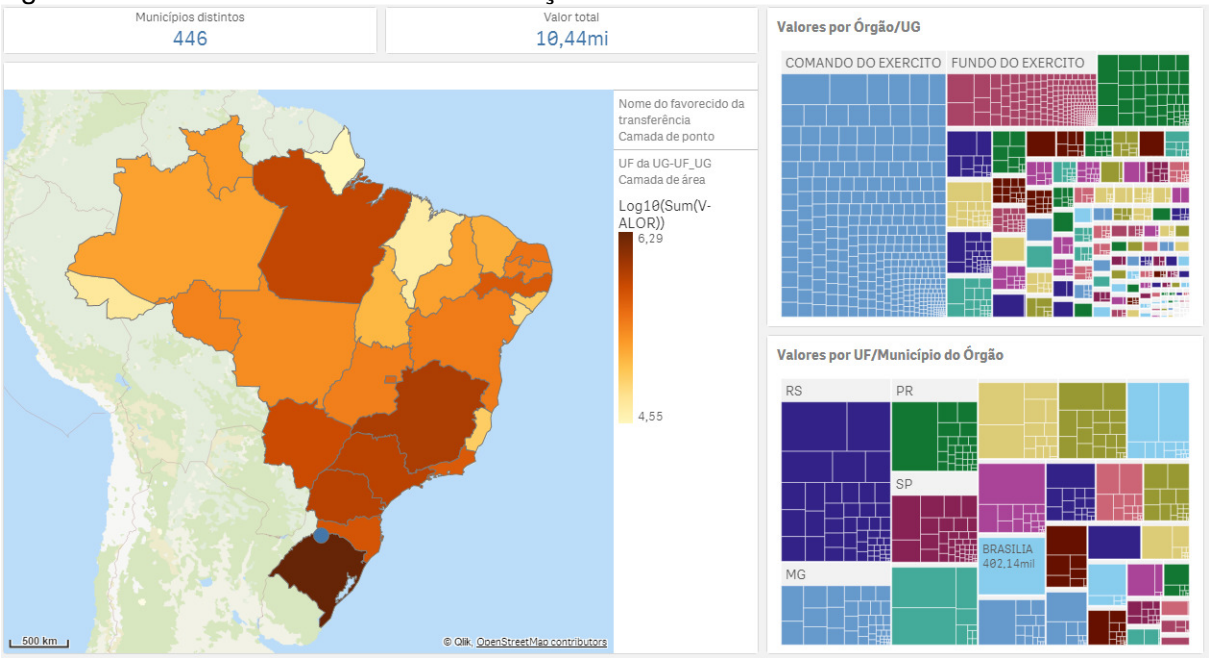
A ilustração acima mostra fornecedores de produtos ou serviços comuns que, geralmente, contratam com a Administração Pública em longas distâncias. Essas situações merecem uma análise mais detalhada, posteriormente.

4.1.2 Fornecedores de pequeno porte com abrangência nacional

Durante a análise exploratória foram encontrados diversos casos de pequenas empresas (MEs e EPPs), com pequeno capital social, que conseguem fornecer para órgãos sediados em todo o território nacional. Geralmente, essas situações ocorrem em função da adesão dos contratantes ao mecanismo de ata de registro de preços.

Vale ressaltar que essa situação pode ser cumulativa com a anterior, levantando um maior grau de suspeição.

Figura 17 – Fornecedor EPP com distribuição em todo o Brasil



Fonte: DWTG (2019).

A figura acima mostra a situação de uma empresa de pequeno porte situada no município gaúcho de Frederico Westphalen, que consegue dominar grande parte dos fornecimentos de materiais de construção para todo o Brasil. As cores mais escuras no mapa representam as UFs das UGs que efetuaram mais pagamentos para a referida varejista.

Desde 2015, esta companhia forneceu produtos desse gênero para diversas repartições públicas sediadas em 446 municípios diferentes de todo o território nacional, perfazendo um total de mais de dez milhões de reais. Os produtos fornecidos são bens comuns, como materiais hidráulicos (torneiras, válvulas, etc.) e de construção em geral (lixas, luvas, etc.).

Foram encontrados na análise outros casos similares envolvendo outras empresas neste mesmo ramo e região do país.

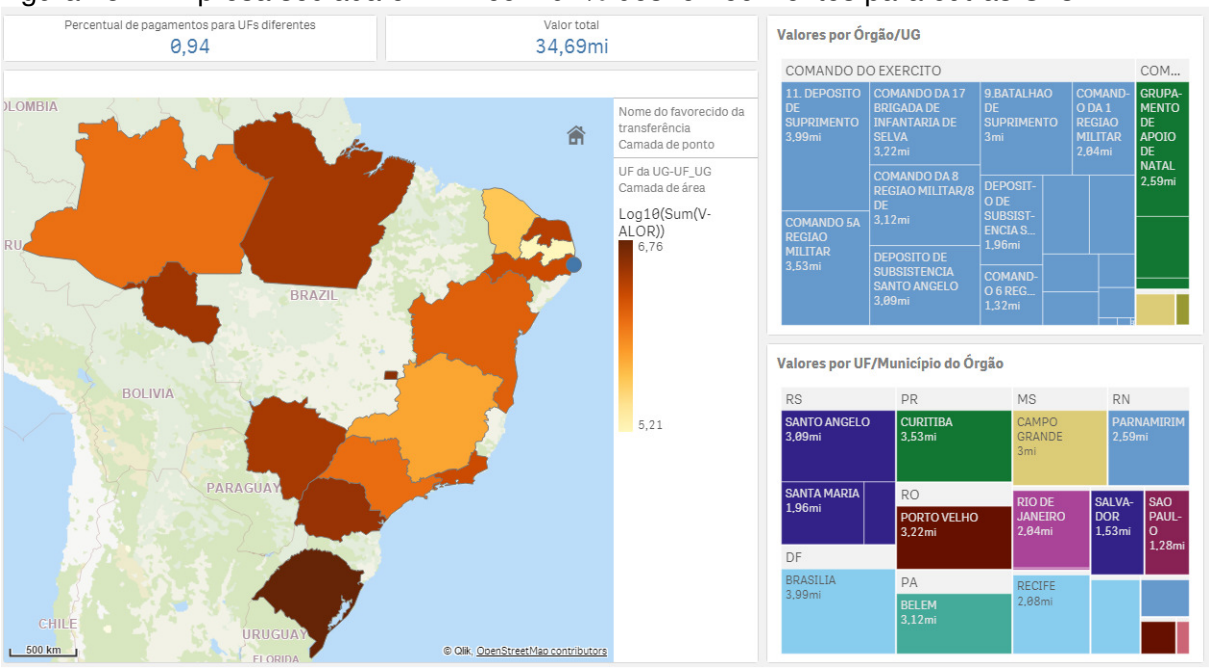
4.1.3 Empresas e órgãos que contratam principalmente fora de sua UF

Em um país com fortes subdivisões territoriais como o Brasil, principalmente no tocante à legislação tributária específica de cada ente federativo, a transposição de barreiras interestaduais pode acarretar em aumento do tempo de transporte, além de pagamento de taxas e de impostos adicionais.

De acordo com o tipo de fornecimento, esta situação pode tornar a execução contratual mais difícil ou até mesmo impossível, levantando indícios de não haver uma contraprestação por parte do fornecedor.

Porém, ao contrário dessa lógica, existem empresas e repartições públicas que contratam mais fora do que dentro do estado no qual se situa a sede. Como exemplo, é mostrada a situação abaixo de uma empresa sediada no município pernambucano de Jaboatão dos Guararapes, cuja principal atividade econômica é o comércio atacadista de carnes. Cerca de 94% do total fornecido por essa empresa, de aproximadamente trinta e cinco milhões de reais, são para repartições públicas situadas em outros Estados da Federação.

Figura 18 – Empresa sediada em PE com 94% dos fornecimentos para outras UFs



Fonte: DWTG (2019).

O mapa acima mostra que essa atacadista forneceu mais carne para UGs sediadas no Estado do Rio Grande do Sul, que possui um dos maiores rebanhos bovinos e suínos do Brasil, do que para repartições públicas localizadas em sua própria UF.

4.2 VALIDAÇÃO DA HIPÓTESE DA PESQUISA

Durante esta etapa foi validada uma hipótese criada durante a exploração inicial: se longas distâncias de fornecimento podem impactar o preço final do produto, prejudicando, assim, o alcance da finalidade pública.

Para tornar viável esse experimento, foi criada uma pequena amostra contendo atributos de preço unitário, distância e tipo do produto para 1.075 (mil e setenta e cinco) pagamentos do Governo Federal, dos anos de 2015 a 2019. O procedimento para geração da amostra de validação foi descrito na Seção 3.1.3.2.2. Adicionalmente, para esta análise, foram excluídos os registros com distância zero e de tipos de produtos raros, com baixa frequência.

A seguir é mostrada a distribuição de produtos e seus respectivos percentuais de representatividade na amostra de validação.

Tabela 3 – Distribuição de produtos na amostra

Produto	Frequência	Percentual (%)
TOALHA DE PAPEL (PACOTE)	10	0,93
GASOLINA COMUM (LITRO)	12	1,12
REFRIGERADOR DUPLEX (UN)	13	1,21
BATERIA NÃO RECARREGÁVEL (UN)	13	1,21
TELA PROJEÇÃO (UN)	14	1,30
PAPEL (RESMA)	15	1,40
SUPORTE FIXAÇÃO PROJETOR (UN)	15	1,40
FREEZER (UN)	18	1,67
TELEFONE SEM FIO (UN)	20	1,86
ÓLEO DIESEL (LITRO)	22	2,05
COPO DESCARTÁVEL (PACOTE)	25	2,33
MOUSE PAD (UN)	25	2,33
PAPEL SULFITE (RESMA)	25	2,33
ESTANTE METÁLICA (UN)	27	2,51
REFRIGERADOR DOMÉSTICO (UN)	31	2,88
COPO DESCARTÁVEL (CAIXA)	33	3,07
FRIGOBAR (UN)	37	3,44
FILTRO LINHA (UN)	39	3,63
MONITOR VIDEO (UN)	43	4,00
LÂMPADA LED (UN)	46	4,28

MICROCOMPUTADOR (UN)	67	6,23
BEBEDOURO ÁGUA GARRAFÃO (UN)	69	6,42
FORNO MICRO-ONDAS (UN)	75	6,98
COMPUTADOR (UN)	100	9,30
PAPEL A4 (RESMA)	137	12,74
PROJETOR MULTIMÍDIA (UN)	144	13,40

Fonte: DWTG (2019).

4.2.1 Estatística descritiva

A análise descritiva da amostra de validação está apresentada logo abaixo, em formato de tabela. Foram extraídos os principais indicadores estatísticos: média, mediana, desvio padrão, entre outros. Essas medidas foram importantes na seleção dos melhores métodos estatísticos para atender ao objetivo da pesquisa.

Inicialmente, a análise descritiva traz uma abordagem geral das informações, incluindo todos os grupos de mercadorias. Após essa avaliação, os valores foram segregados por tipos de produto, para que houvesse um conhecimento mais específico acerca dos dados.

Tabela 4 – Análise descritiva geral

Variável	Média	Mediana	Desvio Padrão	Coef. de Variação	Min.	Máx.
Valor total	39.451,33	4.594,60	236.022,65	598,26	24,51	6.334.110,72
Valor unitário	1.152,69	426,66	1.496,21	129,80	1,10	6.392,79
Distância	974,46	814,00	769,06	78,92	6,00	3.411,00
Quantidade de itens	439,90	15,00	1.521,27	345,82	1,00	15.000,00

Fonte: DWTG (2019).

Na tabela acima é possível verificar que a variável “valor total” apresenta uma média de 39.451,33 e um coeficiente de variação elevado, isso é causado pela diversidade das informações e de produtos que compõem a amostra.

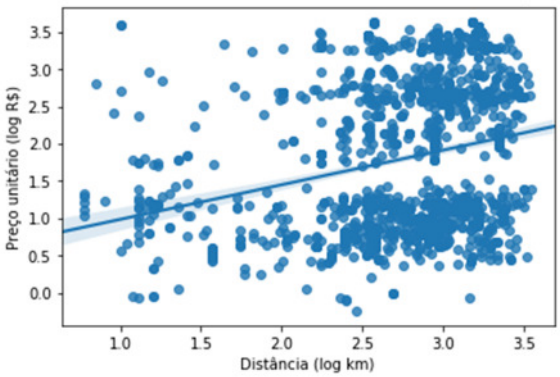
Os principais indicadores estatísticos para cada classe de produtos estão listados no Apêndice A.

4.2.2 Testes de correlação

Inicialmente, foi realizado um teste de correlação de Spearman sobre toda a amostra de validação para verificar se há uma correlação entre o preço unitário e a distância de fornecimento. Os dados da amostra estão plotados no gráfico a seguir.

Preliminarmente, observou-se uma pequena correlação positiva (ρ = 0,2565) entre preço do produto e distância de fornecimento. Isso significa que conforme a distância aumenta, o preço do produto também aumenta. Como o p-valor foi inferior a 0,05, a hipótese nula de que não há correlação estatisticamente significante foi rejeitada.

Figura 19 – Dispersão: Distância x Preço



Fonte: DWTG (2019).

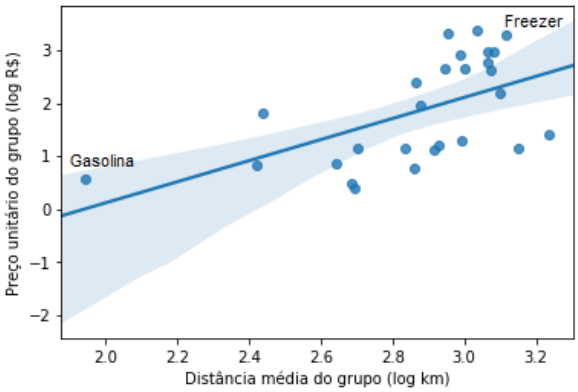
Porém, após uma investigação mais cuidadosa, constatou-se que essa correlação se dava principalmente pelo fato de que classes de produtos com maior valor agregado, como eletrodomésticos ou veículos, tendem a ser fornecidos a maiores distâncias frente a tipos de itens menos elaborados.

Com base nessa situação, Fleury (2004) explica que em regra geral se pode verificar que tendo menor valor agregado o produto gera uma maior participação de despesas de transporte no faturamento da empresa, sendo tal percentual em média de 3,5%. No caso da indústria farmacêutica, com maior nível de agregação, esse percentual é de apenas 0,8%, podendo chegar a 7,1% na indústria de papel e de celulose, que fabrica produtos de menor valor agregado.

Deduz-se, então, que o acréscimo no custo de transporte provocado por longas distâncias pode tornar inviável economicamente o fornecimento de alguns produtos de baixo valor agregado. Porém, alguns desses casos apareceram por

diversas vezes na amostra analisada, conforme a Figura 19, e podem, ao menos, ser considerados incomuns.

Figura 20 – Dispersão: Distância x Preço por tipo de produto



Fonte: DWTG (2019).

Na distribuição de pontos da figura acima, após o agrupamento por tipo de produto, o coeficiente de correlação (ρ) foi de 0,52492, com significância (p) de 0,0002, portanto se trata de uma correlação moderada.

Quando a influência da classe do produto foi eliminada do experimento, e as amostras foram analisadas dentro do seu respectivo grupo (tipo de material), as correlações, inclusive as estatisticamente significantes, tomaram direções e proporções diversas, conforme mostra a tabela a seguir.

Tabela 5 – Teste de correlação entre distância e preço para cada tipo de produto

Produto	Correlação	P-valor
TOALHA DE PAPEL (PACOTE)	0,4587	0,1824
GASOLINA COMUM (LITRO)	-0,04113	0,899
REFRIGERADOR DUPLEX (UN)	-0,0563	0,8549
BATERIA NÃO RECARREGÁVEL (UN)	0,1217	0,692
TELA PROJEÇÃO (UN)	-0,1392	0,635
PAPEL (RESMA)	0,2663	0,3374
SUPORTE FIXAÇÃO PROJETOR (UN)	0,0076	0,9787
FREEZER (UN)	0,3686	0,1323
TELEFONE SEM FIO (UN)	-0,1197	0,6152
ÓLEO DIESEL (LITRO)	-0,7247	0,7489
COPO DESCARTÁVEL (PACOTE)	0,5885*	0,0019
MOUSE PAD (UN)	-0,2353	0,2575
PAPEL SULFITE (RESMA)	0,4163*	0,03844
ESTANTE METÁLICA (UN)	0,1554	0,4388

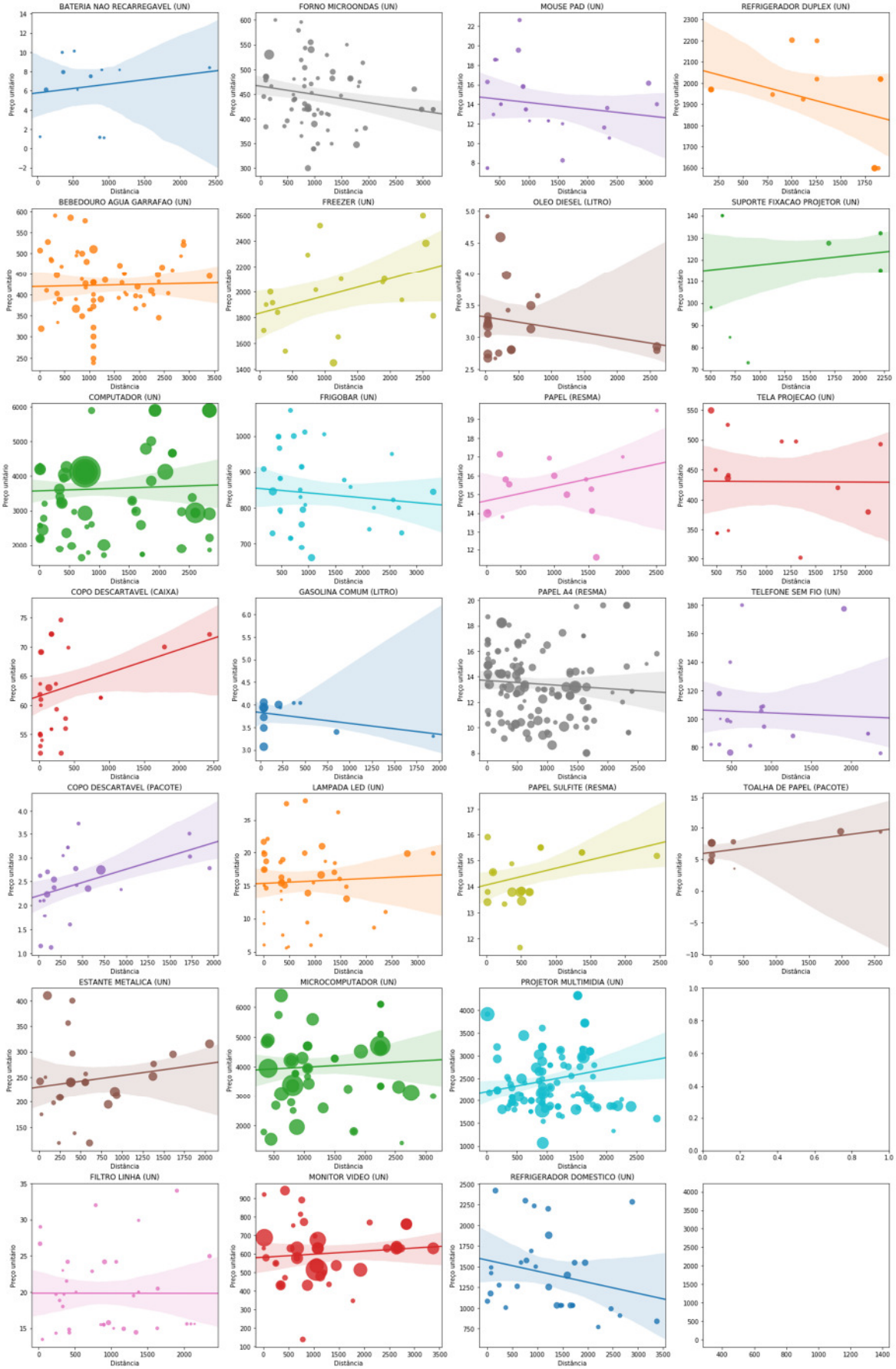
REFRIGERADOR DOMÉSTICO (UN)	-0,3643*	0,0439
COPO DESCARTÁVEL (CAIXA)	0,3167	0,07248
FRIGOBAR (UN)	-0,1021	0,5476
FILTRO LINHA (UN)	0,0534	0,7467
MONITOR VIDEO (UN)	0,0805	0,6079
LÂMPADA LED (UN)	0,0122	0,9359
MICROCOMPUTADOR (UN)	0,0297	0,812
BEBEDOURO ÁGUA GARRAFÃO (UN)	-0,0135	0,9126
FORNO MICRO-ONDAS (UN)	-0,2803*	0,01483
COMPUTADOR (UN)	0,0019	0,9849
PAPEL A4 (RESMA)	-0,1808*	0,0345
PROJETOR MULTIMÍDIA (UN)	0,1516	0,06954

* = Resultado com significância estatística ($p < 0,05$).

Fonte: DWTG (2019).

A seguir são mostrados os gráficos de dispersão para cada classe de itens que foi objeto da análise.

Figura 21 – Dispersão: Distância x Preço para cada tipo de produto



Fonte: DWTG (2019).

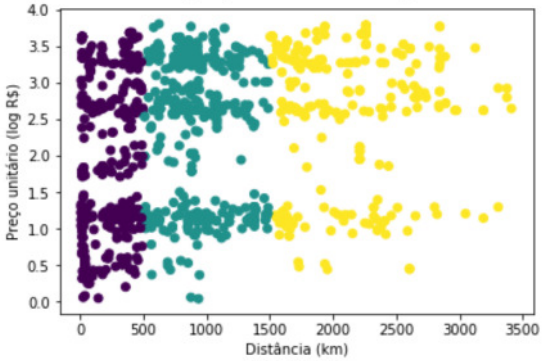
Como pode ser visualizado nos gráficos anteriores, não há um comportamento generalizado de variação do preço em função da distância dentro das classes de produtos. Portanto, não se pode inferir que a distância provoca aumento do preço final de forma generalizada.

4.2.3 Análise de sobrepreços

Nessa fase foi verificado se é mais caro, e qual a média de sobrepreço que um produto fornecido em longas distâncias possui em relação a um produto similar fornecido em curtas distâncias.

Como etapa intermediária, no processo de descoberta, foi realizado um agrupamento k-Means na amostra, visando transformar o atributo numérico de distância em categórico ordinal. Ficaram bem evidentes três categorias de distâncias, com cortes na região de 500 e 1500 quilômetros, aproximadamente.

Figura 22 – Agrupamento dos pontos da amostra



Fonte: DWTG (2019).

Verifica-se, da parte inferior do gráfico, que produtos com baixo preço unitário costumam ser fornecidos em curtas distâncias (menores que 500 km). Conforme a distância aumenta, essa classe de produtos vai ficando mais rara. Tal observação corrobora com os resultados apresentados na seção anterior.

As razões entre o preço unitário médio dos dois grupos (mais próximos e mais distantes) para cada tipo de produto também assumiram comportamentos distintos para as diversas classes de produtos e não se mostraram conclusivas, conforme a tabela abaixo.

Tabela 6 – Percentual de sobrepreço para cada tipo de produto

Produto	Sobrepreço dos mais distantes
BATERIA NÃO RECARREGÁVEL (UN)	+46%
BEBEDOURO ÁGUA GARRAFÃO (UN)	+2%
COMPUTADOR (UN)	+8%
COPO DESCARTÁVEL (CAIXA)	+14%
COPO DESCARTÁVEL (PACOTE)	+39%
ESTANTE METÁLICA (UN)	+20%
FILTRO LINHA (UN)	+9%
FORNO MICRO-ONDAS (UN)	-16%
FREEZER (UN)	+20%
FRIGOBAR (UN)	-5%
GASOLINA COMUM (LITRO)	-11%
LÂMPADA LED (UN)	-10%
MICROCOMPUTADOR (UN)	+16%
MONITOR VÍDEO (UN)	-4%
MOUSE PAD (UN)	-6%
ÓLEO DIESEL (LITRO)	-12%
PAPEL (RESMA)	-12%
PAPEL A4 (RESMA)	-12%
PAPEL SULFITE (RESMA)	+9%
PROJETOR MULTIMÍDIA (UN)	-4%
REFRIGERADOR DOMESTICO (UN)	-8%
REFRIGERADOR DUPLEX (UN)	-12%
TELA PROJEÇÃO (UN)	-24%
TELEFONE SEM FIO (UN)	-5%
TOALHA DE PAPEL (PACOTE)	+43%

Fonte: DWTG (2019).

5 CONCLUSÕES E TRABALHOS FUTUROS

Esta pesquisa propõe uma análise exploratória dos pagamentos do Governo Federal. Para isso, foi utilizada uma grande massa de dados do Tesouro Gerencial com todos os pagamentos desde o ano de 2015. Adicionalmente, este estudo se propõe a responder a seguinte questão: se longas distâncias entre as partes de um contrato administrativo prejudicam o alcance da finalidade pública.

Como se trata de uma pesquisa de abordagem preponderantemente quantitativa, e também visando uma maior objetividade, concentrou-se em verificar o prejuízo para a finalidade pública em termos de preço final de materiais. Porém, cabe ressaltar que a análise do alcance da finalidade pública é muito mais ampla e não deve se ater somente ao aspecto financeiro. Alguns desses outros aspectos foram introduzidos no Capítulo 1.

Diante disso, infere-se que o cunho geográfico deve sempre respeitar o princípio da proporcionalidade, e cada caso deve ser analisado especificamente, não se podendo afirmar categoricamente que feriu os princípios da Administração Pública.

Para validar experimentalmente a hipótese de que longas distâncias de fornecimento podem impactar o preço unitário final, foi extraída uma amostra de produtos efetivamente pagos e feita uma análise de correlação.

A partir dessa análise se chegou a uma conclusão um pouco diferente da hipótese original: tipos de produtos de maior valor agregado costumam ser fornecidos sob maiores distâncias de fornecimento, enquanto os de menor preço costumam ser providos mais proximamente. Porém, não foi constatada clara influência da distância de fornecimento no preço final do produto dentro da maioria das classes de materiais.

Entre os principais obstáculos encontrados durante a análise está a falta de detalhamento sobre as unidades executoras ou subunidades no Tesouro Gerencial. Na versão disponível no TCU para essa base, o maior nível de detalhamento do ente público contratante é o de Unidade Gestora, porém pode haver subdivisões desse. Como exemplo, a Segunda Comissão Brasileira Demarcadora de Limites (MRE), com sede no Rio de Janeiro e subsedes em Santana do Livramento/RS, Ponta Porã/MS e Corumbá/MS. Grande parte dos gastos está concentrada fora do Rio de

Janeiro, porém o detalhamento da UG Executora ou subsedes da UG principal é inexistente. Casos semelhantes ocorrem em UGs da Embrapa e de outros órgãos.

Outro problema que acontece no SIAFI é a ocorrência de inscrições genéricas ou de recebimento sem a respectiva identificação dos recebedores. Essa situação ocorre nos pagamentos de dívida interna, benefícios do RGPS e folha de pagamento, entre outros.

Como possibilidade de melhoria do presente trabalho se propõe buscar informações sobre os itens em notas fiscais eletrônicas. Essa informação é muito mais abundante do que a descrição dos itens nas observações das OBs e está presente em quase todas as aquisições de materiais (cerca de 90%, ante 2% sobre quantidades). Dessa forma, seria obtido um conjunto de dados muito maior sobre quantidade e preços unitários dos itens efetivamente pagos.

REFERÊNCIAS

BIDERMAN, Maria Tereza Camargo. **Dicionário de termos financeiros e bancários**. São Paulo: DISAL, 2013.

BRASIL. Decreto n.º 92.452, de 10 de março de 1986. Cria, no Ministério da Fazenda, a Secretaria do Tesouro Nacional (STN), extingue a Secretaria Central de Controle Interno (SECIN), e dá outras providências. **Diário Oficial da União**. Brasília, 11 março 1986.

BRASIL. Constituição da República Federativa do Brasil de 1988. Emendas Constitucionais de Revisão. **Diário Oficial da União**. Brasília, 05 out. 1988.

BRASIL. Lei n.º 4.320, de 17 de março de 1964. Estatui Normas Gerais de Direito Financeiro para elaboração e controle dos orçamentos e balanços da União, dos Estados, dos Municípios e do Distrito Federal. **Diário Oficial da União**. Brasília, 04 maio 1964.

BRASIL. Lei n.º 8.666, de 21 de junho de 1993. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. **Diário Oficial da União**. Brasília, 21 jun. 1993.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados**: Conceitos, Tarefas, Métodos e Ferramentas. Goiânia: UFG, 2009.

CASTRO, Domingos Poubel de; GARCIA, Leice Maria. **Contabilidade Pública no Governo Federal**: Guia para reformulação do ensino e implantação da Lógica do Siafi. São Paulo: Atlas, 2004.

CASTRO, Marcus Vinícius Borela de. **Mineração de dados com rastro**: Boas práticas para documentação de processos e sua aplicação em um projeto de classificação textual. Brasília, 2019, 107f. Monografia (*lato sensu* em Análise de Dados) – Escola Superior do Tribunal de Contas da União. Instituto Serzedello Corrêa, Brasília, 2019.

CHAPMAN, PETE; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ, Thomas; SHEARER, Colin; WIRTH, Rüdiger. **CRISP-DM 1.0**: Step-by-step data mining guide. Copenhagen: SPSS, 2000, v. 16.

CONTROLADORIA GERAL DA UNIÃO. Portal da Transparência. **Execução da despesa pública**. Disponível em: http://www.portaltransparencia.gov.br/entenda-a-gestao-publica/execucao-despesa-publica. Acesso em: 02 out. 2019.

CREPALDI, Guilherme Simões; CREPALDI, Sílvio Aparecido. **Orçamento Público**: Planejamento, elaboração e controle. São Paulo: Saraiva, 2013.

DIAS, Maria Madalena. **Um Modelo de Formalização do Processo de Desenvolvimento de Sistemas de Descoberta de Conhecimento em Banco de Dados**. Florianópolis, 212 f. Tese (Doutorado em Engenharia da Produção) - Universidade Federal de Santa Catarina, Florianópolis, 2001.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

FLEURY, Paulo Fernando. Gestão estratégica do transporte. **Revista da Madeira**, n. 81, jun. 2004. Disponível em: http://www.remade.com.br/br/revistadamadeira_materia.php?num=558&subject=Transporte&title=Gest%E3o%20estrat%E9gica%20do%20transporte. Acesso em: 29 set. 2019.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2008.

IBM. **IBM SPSS Modeler CRISP-DM Guide**, 2016. Disponível em: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf. Acesso em: 29 set. 2019.

JUND, Sergio. **Administração financeira e orçamentária**. 3. ed. Rio de Janeiro: Elsevier, 2008.

KANO, Kazuko; KANO, Takashi; TAKECHI, Kazutaka. **The Price of Distance**: Pricing to market, producer heterogeneity, and geographic barriers. The Research Institute of Economy. Tokyo/JP: Trade and Industry, 2015.

KENDALL, Maurice George; GIBBONS, Jean Dickinson. **Rank correlation methods**. 5. ed. London: Griffin, 1990.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MEIRELLES, Helly Lopes. **Direito Administrativo Brasileiro**. 36. ed. São Paulo: Malheiros, 2010.

PEREIRA, Ana Carolina Costa; MOREY, Bernadete Barbosa. Revisitando a lei dos cossenos para triângulos esféricos: um aporte histórico do século XV. **História da Ciência e Ensino: Construindo interfaces**, v. 15, p. 81-95, 2017.

SCIKIT-LEARN. **2.3. Clustering**. Disponível em: https://scikit-learn.org/stable/modules/clustering.html. Acesso em: 10 out. 2019.\

SENADO FEDERAL. **Orçamento Federal**: Glossário. Disponível em: https://www12.senado.leg.br/orcamento/glossario. Acesso em: 29 set. 2019.

SHEARER, Colin. The CRISP-DM model: The new blueprint for data mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13-22, 2000.

SILVA, Gerson Luiz Cardoso da; PALMEIRA, Eduardo Mauch; QUINTANA, Alexandre Costa. Sistema Integrado de Administração Financeira do Governo Federal-SIAFI-Necessidade Criação e Evolução. **Observatório de la Economia Latinoamericana**, v. 86, p. 1-16, 2007.

SILVA, Paulo Henrique Feijó da; MOTA, Francisco Glauber Lima; PINTO, Liane Ferreira. **Curso de Siafi**: Uma abordagem prática da Execução Orçamentária e Financeira. 2. ed. Brasília: Ed. do autor, 2008.

SOUZA, Leandro. **Tesouro Nacional**: BI com Serpro e MicroStrategy. Baguete, 23 nov. 2015. Disponível em: https://www.baguete.com.br/noticias/23/11/2015/tesouro-nacional-bi-com-serpro-e-microstrategy. Acesso em: 02 out. 2019.

STN. **Conheça o SIAFI**, 2019. Disponível em: http://www.tesouro.fazenda.gov.br/conheca-o-siafi. Acesso em 01 out. 2019

SUPERIOR TRIBUNAL DE JUSTIÇA. **STJ - HC 88.370 RS 2007/0181783-1**. Relator: Ministro Napoleão Nunes Maia Filho. Data de Julgamento: 07/10/2008. T5 - Quinta Turma. Data de Publicação: DJe 28/10/2008.

TESOURO NACIONAL. **Ordem Bancária**: Manual Simplificado. 08 fev. 2019. Disponível em: https://sisweb.tesouro.gov.br/apex/cosis/thot/obtem_arquivo/29110:1683431:inline. Acesso em: 29 set. 2019.

TRIBUNAL DE CONTAS DA UNIÃO. Secretaria de Orçamento, Finanças e Contabilidade. **Fluxograma da Execução Orçamentária e Financeira**. Disponível em: https://portal.tcu.gov.br/lumis/portal/file//fileDownload.jsp?fileId=8A8182A14D110A73014D1EFE3BA91199. Acesso em: 29 set. 2019.

TRIBUNAL DE CONTAS DA UNIÃO. **TCU – Acórdão N.º 520/2015**. Processo nº TC 000.548/2015-4. 2ª Câmara. Relator: Ministro Vital do Rêgo. Disponível em: http://www.tcu.gov.br/Consultas/Juris/Docs/judoc/Acord/20150305/AC_0520_04_15_2.doc. Acesso em: 30 set. 2019.

VALLADARES NETO, José; SANTOS, Cristiane Barbosa dos; TORRES, Érica Miranda; ESTRELA, Carlos. Boxplot: Um recurso gráfico para a análise e interpretação de dados quantitativos. **Revista Odontológica Brasil-Central**, v. 26, n. 76, p. 1-6, 2017.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a standard process model for data mining. In: **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. Manchester/UK: Citeseer, 2000, p. 29-39.

TERMINOLOGIA

Accountability	Termo da língua inglesa que pode ser traduzido para o português como responsabilidade com ética e remete à obrigação, à transparência, de membros de um órgão administrativo ou representativo de prestar contas a instâncias controladoras ou a seus representados.
CNAE	Classificação Nacional de Atividades Econômicas, que tem como objetivo categorizar os ramos de atuação de pessoas jurídicas em todo o país.
ComprasNet	Portal de Compras do Governo Federal é um site web, instituído pelo Ministério do Planejamento, Orçamento e Gestão, para disponibilizar a sociedade, informações referentes às licitações e contratações promovidas pelo Governo Federal, bem como permitir a realização de processos eletrônicos de Aquisição.
CRISP-DM	Abreviação de CRoss Industry Standard Process for Data Mining, que pode ser traduzido como Processo Padrão Inter-Indústrias para Mineração de Dados. É um modelo de processo de mineração de dados que descreve abordagens comumente usadas por especialistas em mineração de dados para atacar problemas.
Georreferenciamento	Tornar as coordenadas de um local conhecidas num dado sistema de referência.
ISC	Instituto Serzedello Corrêa, Escola Superior do Tribunal de Contas da União.
DWTG	Data Warehouse do Tesouro Gerencial, construído a partir de dados do SIAFI.
JDBC	Java Database Connectivity é um conjunto de classes e interfaces (API) escritas em Java que fazem o envio de instruções SQL para qualquer banco de dados relacional.
Jupyter notebook	Um aplicativo Open Source da web que permite criar e compartilhar documentos com códigos Python, equações,

	visualizações e textos explicativos.
LabContas	O "Laboratório de Informações de Controle" é um ambiente de conhecimento, informação e tecnologia com a finalidade de agregar valor às bases de dados e torná-las úteis e disponíveis aos processos de trabalho do TCU, propiciando, também, o desenvolvimento e aplicação de estratégia digital de controle.
Python	Linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte.
Qlik Sense	Plataforma de BI que permite criar visualizações de dados, relatórios e painéis interativos e personalizados.
re	Biblioteca da linguagem Python que fornece operações de correspondência de expressões regulares.
Scikit-learn	Biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python.
SIASG	Sistema utilizado para facilitar e agilizar os processos de compra e aquisição de materiais e serviços do Governo Federal.
SQL	Structured Query Language (Linguagem de Consulta Estruturada) é a linguagem de pesquisa declarativa padrão para banco de dados relacional. Muitas das características originais do SQL foram inspiradas na álgebra relacional.
SQL Server	Sistema gerenciador de Banco de dados relacional desenvolvido pela Microsoft.
SIAFI	Sistema Integrado de Administração Financeira do Governo Federal, que é um sistema contábil que tem por finalidade realizar todo o processamento, controle e execução financeira, patrimonial e contábil da União.
Teradata	Teradata Corporation é uma empresa de Tecnologia da Informação que desenvolve e vende soluções para Data Warehouse.
UG	Unidade Gestora do SIAFI, que é responsável por

	administrar dotações orçamentárias e financeiras próprias ou descentralizadas.
UO	Unidade Orçamentária, entidade da administração direta, inclusive fundo ou órgão autônomo, da administração indireta (autarquia, fundação ou empresa estatal) em cujo nome a lei orçamentária ou crédito adicional consigna, expressamente, dotações com vistas à sua manutenção e à realização de um determinado programa de trabalho.
OB	Ordem Bancária, meio de pagamento destinado a repasse de recursos do Governo Federal.

APÊNDICE A – ANÁLISE DESCRITIVA DAS VARIÁVEIS POR TIPO DE PRODUTO

A seguir estão elencados os principais indicadores estatísticos para as variáveis em cada classe de produtos.

Produto	Media	Mediana	Desv. Padrão	Coef. de Variação	Min	Max
TOALHA DE PAPEL (PACOTE)						
Valor Total Pago	7456,575	6102,3	6890,1565	92,403771	35,21	18975
Valor Unitário	6,801188	7,59	1,9405094	28,531918	3,521	9,41
Quantidade	1084	850	945,11587	87,187811	3,521	9,41
Distância	539,2	23	893,94707	165,79137	11	2590
GASOLINA COMUM (LITRO)						
Valor Total Pago	14195,524	18022,25	12140,624	85,524312	366,5	39425
Valor Unitário	3,7424531	3,9271954	0,3266337	8,7277963	3,07	4,0529
Quantidade	3839,9167	5000	3305,9718	86,094884	3,07	4,0529
Distância	354,41667	120	532,25518	150,17781	37	1935
REFRIGERADOR DUPLEX (UN)						
Valor Total Pago	3747,8738	3938	1775,7767	47,380908	1597,54	7987,7
Valor Unitário	1924,8431	1969	197,24053	10,247096	1597,54	2202,16
Quantidade	2	2	1,1094004	55,47002	1597,54	2202,16
Distância	1182,5385	1258	660,9367	55,891349	175	1910
BATERIA NÃO RECARREGÁVEL (UN)						
Valor Total Pago	418,71538	202,5	506,01029	120,84827	24,51	1769
Valor Unitário	6,4711282	7,94	3,1187463	48,194785	1,1	10,125
Quantidade	94,846154	30	100,44658	105,90474	1,1	10,125
Distância	759,92308	747	578,35994	76,107696	36	2417
TELA PROJEÇÃO (UN)						
Valor Total Pago	2196,9979	1126,98	2339,2873	106,47654	344	7688,94
Valor Unitário	430,34245	438,40279	74,40923	17,290702	300,99	549,21
Quantidade	4,9285714	2,5	4,8765369	98,944227	300,99	549,21
Distância	1010,9286	622	584,00324	57,768991	448	2154
PAPEL (RESMA)						
Valor Total Pago	6192,3333	5649	5851,1462	94,490169	510	24482,5
Valor Unitário	15,435347	15,54	1,7861153	11,571592	11,58045	19,47
Quantidade	422,53333	400	422,18383	99,917283	11,58045	19,47
Distância	992,8	998	759,80403	76,531429	13	2510
SUPORTE FIXAÇÃO PROJETOR (UN)						
Valor Total Pago	582,05467	279,98	691,59817	118,82014	98	2677,92
Valor Unitário	118,198	127,52	22,007715	18,619363	73	139,99
Quantidade	4,8666667	2	5,4877642	112,76228	73	139,99
Distância	1116,8	622	715,60494	64,076373	509	2211
FREEZER (UN)						
Valor Total Pago	4798,9878	3539,92	3677,0541	76,621451	1900	14499,1
Valor Unitário	1990,3925	1968,37	306,60528	15,404262	1449,91	2597,59
Quantidade	2,5	2	2,1666667	86,666667	1449,91	2597,59

Distância	1166,7222	1029	883,05611	75,686919	65	2659
TELEFONE SEM FIO (UN)						
Valor Total Pago	1511,9345	858	1763,6001	116,64527	100	7630
Valor Unitário	104,158	96,275	28,966679	27,810325	75,8	180
Quantidade	15,3	10	20,911958	136,67947	75,8	180
Distância	1017,15	806,5	718,18461	70,607541	248	2367
ÓLEO DIESEL (LITRO)						
Valor Total Pago	22529,953	22329	15166,313	67,31622	666,63	55046
Valor Unitário	3,2409286	3,15025	0,5865788	18,099098	2,66652	4,9146
Quantidade	7053,5	7500	4463,9923	63,28762	2,66652	4,9146
Distância	460,77273	215	716,66741	155,53599	31	2605
COPO DESCARTÁVEL (PACOTE)						
Valor Total Pago	3381,2796	960	8063,7049	238,48087	35,6	41100
Valor Unitário	2,449796	2,42	0,6465544	26,392172	1,12	3,71
Quantidade	1373,48	350	2971,9888	216,38384	1,12	3,71
Distância	448,16	279	551,66582	123,09573	16	1955
MOUSE PAD (UN)						
Valor Total Pago	1009,5856	745	825,86903	81,802774	185,5	3228
Valor Unitário	14,0538	13,6	3,6562071	26,01579	7,45	22,6
Quantidade	72,24	50	54,230825	75,070356	7,45	22,6
Distância	1161,4	900	847,6471	72,98494	290	3190
PAPEL SULFITE (RESMA)						
Valor Total Pago	12858,458	7650	12990,858	101,02967	309,8	41350,68
Valor Unitário	14,423257	14,546143	0,975278	6,7618427	11,66	15,89
Quantidade	919,36	500	953,77305	103,74315	11,66	15,89
Distância	570,6	507	534,72049	93,711968	13	2460
ESTANTE METÁLICA (UN)						
Valor Total Pago	15440,846	11900	14270,491	92,420396	500	54587
Valor Unitário	243,93115	240	69,903434	28,657035	119	410
Quantidade	64,518519	50	61,366157	95,11402	119	410
Distância	580,44444	401	497,63167	85,732869	13	2053
REFRIGERADOR DOMÉSTICO (UN)						
Valor Total Pago	4423,6497	3244,98	3388,3867	76,597085	1421,66	15036
Valor Unitário	1411,9874	1275,775	464,3379	32,885414	769,4	2418,78
Quantidade	3,2903226	2	2,3853525	72,496006	769,4	2418,78
Distância	1243,1613	1226	865,91883	69,654585	7	3376
COPO DESCARTÁVEL (CAIXA)						
Valor Total Pago	2097,3221	1779	2168,4422	103,391	540	12598
Valor Unitário	62,883646	61,28	7,2485827	11,526976	51,78	74,6
Quantidade	33,30303	30	33,97636	102,02183	51,78	74,6
Distância	331,27273	175	519,64383	156,86285	12	2439
FRIGOBAR (UN)						
Valor Total Pago	3549,6603	1754,66	4017,9474	113,19245	714,89	21141
Valor Unitário	840,03668	830	106,96977	12,733941	657,167	1071,57
Quantidade	4,3783784	2	5,1745288	118,18368	657,167	1071,57
Distância	1104,1892	854	800,86716	72,529886	175	3307

FILTRO LINHA (UN)						
Valor Total Pago	687,71538	471	642,98395	93,495647	78	2838,6
Valor Unitário	19,832311	19,68	5,2150061	26,295504	13,46	33,99
Quantidade	36,179487	25	36,748537	101,57285	13,46	33,99
Distância	964,23077	898	631,34161	65,476194	22	2353
MONITOR VÍDEO (UN)						
Valor Total Pago	106229,36	22675,09	234494,03	220,74314	429,99	1441116,8
Valor Unitário	600,59154	629,86348	155,29405	25,856849	137,51286	941,5
Quantidade	184,5814	36	440,11031	238,43698	137,51286	941,5
Distância	1149,4884	875	917,2109	79,792969	7	3376
LÂMPADA LED (UN)						
Valor Total Pago	3357,5765	1480	3922,9914	116,83997	110	16579,98
Valor Unitário	15,581661	15,665	5,4383567	34,902291	5,58	27,94
Quantidade	204,26087	100	240,04822	117,52042	5,58	27,94
Distância	758,08696	447,5	780,30267	102,9305	6	3307
MICROCOMPUTADOR (UN)						
Valor Total Pago	111603,56	22219,4	229330,5	205,48673	1412,25	1281371,1
Valor Unitário	4032,2272	3954,3	1169,5371	29,004743	1412,25	6392,785
Quantidade	31,074627	5	64,21171	206,63711	1412,25	6392,785
Distância	1387,194	1058	738,52059	53,238449	337	3125
BEBEDOURO ÁGUA GARRAFÃO (UN)						
Valor Total Pago	4097,7632	2899,98	4278,5963	104,41297	333,82	25500
Valor Unitário	422,59817	426,66	70,491167	16,680424	238,32778	590
Quantidade	9,9565217	7	10,287793	103,32718	238,32778	590
Distância	1276,6522	1081	803,34077	62,925579	10	3411
FORNO MICRO-ONDAS (UN)						
Valor Total Pago	2986,0023	1539,96	5991,6904	200,65927	376,6	50349,05
Valor Unitário	449,49052	446,87	61,301387	13,637971	299,89	599,99
Quantidade	6,6	3	11,734848	177,80073	299,89	599,99
Distância	985,52	875	743,32795	75,424948	59	3190
COMPUTADOR (UN)						
Valor Total Pago	222787,44	44379,96	699708,25	314,06988	1750	6334110,7
Valor Unitário	3631,2163	4123,77	1014,2789	27,932209	1602,9333	5899,2507
Quantidade	56,69	15	171,14139	301,8899	1602,9333	5899,2507
Distância	1030,84	766	830,65352	80,580257	15	2846
PAPEL A4 (RESMA)						
Valor Total Pago	15019,078	7920	19043,814	126,79749	172	144991
Valor Unitário	13,424427	13,181	2,7099863	20,18698	7,995	19,6
Quantidade	1178,438	600	1498,8443	127,18907	7,995	19,6
Distância	854,9927	621	714,1272	83,524362	6	2823
PROJETOR MULTIMIDIA (UN)						
Valor Total Pago	21604,202	8917,71	39461,562	182,65688	1329	289506,96
Valor Unitário	2488,3197	2169,0225	719,04543	28,896826	1067,0735	4322,4267
Quantidade	9,3680556	4	19,308442	206,10939	1067,0735	4322,4267
Distância	1095,8681	1058	580,61446	52,982151	10	2840