Relatório de Visualização de Dados (IN1171)

Victor Félix Pimenta*
CIn - UFPE

RESUMO

O controle de gastos e a gestão dos recursos públicos são tarefas importantes em diversas partes do mundo e essa é uma questão que a aprendizagem de máquina tem tentado ajudar a atenuar. No Brasil há incentivos para projetos que ajudem a identificar corrupção e incentivar a fiscalização, porém há poucos trabalhos científicos nessa área. O objetivo dessa pesquisa é estabelecer um modelo capaz de extrair conhecimento de uma base de dados de gastos públicos, identificando padrões e encontrando na massa de dados um panorama geral do destino das verbas direcionadas a servidores públicos, permitindo a participação dos cidadãos nessa atividade.

Palavras-chave: Visualização de dados—Gastos Públicos—Séries Temporais—Clusterização

1 INTRODUÇÃO

Apesar do tema de controle de gastos e corrupção ser de interesse de vários países, o desenvolvimento de soluções tende a ser localizado, empregando regras manuais para identificar possíveis problemas. A consequência disso é que há poucos trabalhos científicos utilizando técnicas de inteligência artificial para minerar os dados disponíveis em busca de problemas e inconsistências. Além disso, tratar de corrupção demanda conhecimento profundo das regras políticas e envolve evidências externas. Em face disso, este trabalho foca na transparência e prestação de contas, elaborando uma ferramenta que auxilie do processo de conscientização social que deve ocorrer para que haja uma melhoria na gestão pública [1].

A fonte de dados escolhida foi a Cota para Exercício de Atividade Parlamentar (CEAP), disponível no portal da câmara de deputados que fornece dados sobre os gastos públicos dos deputados do Brasil, compondo um conjunto de séries temporais contendo informações sobre o uso do dinheiro [6]. Nesse contexto foram utilizadas técnicas de previsão de séries temporais e clusterização para traçar um perfil dos gastos da CEAP dos políticos do Brasil e, assim, atuar como uma ferramenta para auxiliar no monitoramento do panorama nacional para melhor julgar, cobrar e votar em políticos. Também esperamos que o modelo elaborado possa ser suficientemente robusto para permitir sua utilização com outras fontes de dados de outros políticos e até outros países.

O restante desse artigo está dividido da seguinte maneira: na seção seguinte apresentaremos uma breve revisão dos principais trabalhos na área, bem como o estado da arte no que diz respeito a clustering e séries temporais e o modelo escolhido; na seção 3 detalharemos a estrutura dos dados utilizados para a pesquisa, seus problemas e desafios; na seção 4 descreveremos os procedimentos realizados no pré-processamento dos dados; em seguida descreveremos os experimentos realizados; na seção 6 traremos os resultados dos testes; finalmente na seção 7 faremos uma conclusão e levantaremos ideias para trabalhos futuros.

*e-mail: vfp@cin.ufpe.br

2 REVISÃO DA LITERATURA

Em sistemas democráticos, os gastos públicos e a sua fiscalização são de grande importância, pois são um fator determinante no crescimento e desenvolvimento social [2]. Porém é necessário que esses gastos sejam aplicados de maneira correta, caso contrário eles passam a caracterizar desperdício de verba pública. Várias trabalhos na área tentam avaliar a distribuição dos investimentos e ações realizadas por prefeituras e governos [7] [8], mas há pouco estudo sobre os gastos pessoais de membros do governo e a fiscalização dos mesmos. No Brasil a CEAP reembolsa os deputados com o objetivo de viabilizar o trabalho parlamentar, mas essa verba pode ser empregada de maneira indevida.

O contexto do estudo realizado foi focado nos conceitos de séries temporais e clusterização, áreas com uma literatura forte, o que permitiu identificar boas referências durante a pesquisa em face da ausência de trabalhos com o mesmo foco. Um ponto de interesse é o estudo dos *outliers*, onde diversos trabalhos foram publicados, podendo ser divididos quanto a técnica empregada em modelos estatísticos, redes neurais, aprendizagem de máquina ou sistemas híbridos [4]. *Outliers* são interessantes dentro do contexto de corrupção e gastos públicos pois permitem identificar entre o conjunto de informações os pontos que mais se afastam da média. Assumindo que o comportamento padrão é de gastos justos e de honestidade, encontrar situações suficientemente afastadas da média são um forte indicativo de problemas ou consumo malicioso de dinheiro do contribuinte. Contudo, em se tratando de dados temporais, a identificação de outliers se torna um tanto diferente, pois cada medição está correlacionada com as vizinhas e esse tipo de situação não é perfeitamente incorporada por modelos onde cada dimensão do vetor de entrada representa uma informação de natureza diferente. Nesse contexto existem algumas alternativas para melhor agrupar esse tipo de dado, podendo medir *outliers* como um ponto em uma série, uma série num conjunto de dados, além de poder modelar dados que são recebidos iterativamente, como um *data stream* [3].

Por conta deste trabalho tratar-se de uma pesquisa com um objetivo original, o trabalho de clusterização de gastos perde robustez no que diz respeito a medidas de *benchmark* e métricas de comparação da acurácia. Além disso, o projeto foi desenvolvido sem o auxílio de um especialista na área, o que torna o levantamento de hipóteses e a validação das mesmas um processo árduo, impedindo que as conclusões sejam mais assertivas, optando-se por uma perspectiva mais geral.

3 DESCRIÇÃO DOS DADOS

Os dados utilizados para esse trabalho foram obtidos do portal da câmara dos deputados do Brasil, utilizando as ferramentas disponíveis pelo *framework* desenvolvido pelo projeto Operação Serenata de Amor. Esses dados correspondem a gastos realizados por deputados federais e estaduais referentes a CEAP. Cada instância da base de dados representa as informações de uma compra realizada por um deputado, composta por atributos relacionados à compra (e.g. valor da compra, número da nota fiscal, CNPJ da empresa vendedora) e ao deputado (e.g. nome, partido, estado de atuação). Além dessa informações também foram utilizados a sequência de legislaturas a qual o deputado participou como forma de determinar janelas temporais de interesse, e a situação do mandato, identificando se o parlamentar trabalhou durante todo o período ou se foi

um mandato atípico.

Por tratarem-se de gastos reais e que descrevem uma situação complexa, os dados utilizados possuíam uma série de problemas e pontos a serem considerados antes que pudessem ser aplicados ao modelo. Na próxima seção descreveremos os procedimentos aplicados e decisões tomadas para lidar com os problemas nesses dados.

4 Avaliação

Um ponto crucial para a validação deste trabalho é medir a qualidade do modelo desenvolvido. Para isso é necessário garantir que a divisão dos dados encontrada é suficientemente sólida e saber justificar quais características estão presentes em cada conjunto.

Para responder a primeira questão foram utilizados medidas de estabilidade dos clusters. A métrica inicialmente aplicada foi a silhueta, que mede a separação de cada cluster, aplicando um score para cada amostra de acordo com a distância entre ela e os demais pontos no cluster comparado aos fora dele. Outra medida aplicada foi a elaboração de um classificador treinado utilizando uma amostra com suas etiquetas obtidas na clusterização. A acurácia desse classificador aplicada a um conjunto de teste determina a qualidade da divisão dos grupos [5]. A técnica de silhueta traz um benefício visual de fácil interpretação para o modelo e o classificador traz um score adicional que pode ser comparado a fim dar mais solidez às conclusões.

A resposta à segunda pergunta é mais complexa de ser respondida e é crucial para determinar a validade do projeto. Algumas alternativas foram consideradas com o objetivo de responder essa pergunta, incluindo testes práticos e técnicas de visualização de dados. Essas alternativas e os problemas relacionados estão descritos na seção seguinte.

5 Experimentos

Com o modelo definido e os dados previamente limpos pudemos iniciar os experimentos sobre os dados. Inicialmente o objetivo era encontrar a melhor divisão dos dados seguindo os critérios mencionados na seção anterior. Porém, conforme os experimentos foram realizados, notamos que a diferença entre os clusters não era suficientemente grande para que fosse possível uma determinada configuração de parâmetros pudesse dominar sobre as demais. Nas configurações em que tivemos uma média significativamente alta para as silhuetas o modelo identificou apenas 2 clusters, tornando o processo de justificativa difícil, ao passo que aumentar a sensibilidade do modelo diminuiu drasticamente a qualidade da silhueta como pode ser visto nas figuras 1 e 2. O classificador treinado também não apresentou o desempenho esperado. A alta granularidade do problema acabou por impactar na qualidade da acurácia do classificador, cuja taxa de acerto só foi relevante para o cluster majoritário.

Algumas alternativas foram consideradas a fim de justificar a divisão dos dados inicialmente sem muito sucesso. Estas estão descritas a seguir:

- Para cada cluster foi extraída a série temporal média e a variância de cada instância. A hipótese considerada nessa abordagem era encontrar em cada cluster um conjunto de instâncias com variância maior, o que levaria à conclusão de que aqueles atributos seriam os representantes daquele cluster. Porém as séries de cada cluster se mostraram bastante similares.

- Foi treinado um classificador Random Forest para cada cluster de maneira binária, onde os demais clusters foram marcados como 'outros'. Este procedimento teve o objetivo de extrair do classificador os atributos mais importantes que pudessem vir a determinar o conteúdo do grupo. Entretanto o classificador empregado acabou apresentando uma variabilidade quanto aos scores dos atributos a depender da execução.

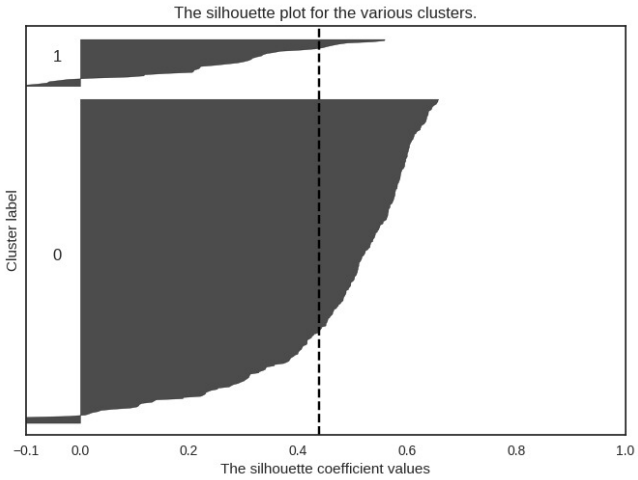

Figura 1: Silhueta obtida utilizando distância cosseno e k=3.

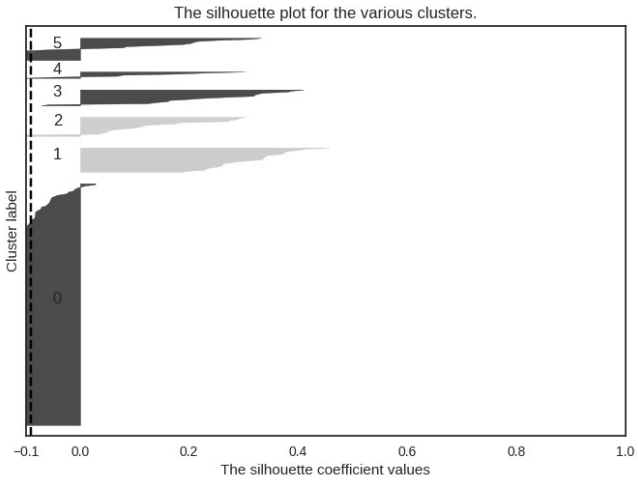

Figura 2: Silhueta obtida utilizando distância cosseno e k=2

- A partir dos testes feitos com os gastos normalizados viu-se que as silhuetas mostraram um resultado inferior ao com os dados absolutos. O único modelo que manteve um resultado bom para a silhueta foi no grupo de controle onde o método MST-KNN não foi utilizado, somente a divisão obtida pelo Kmeans. Foi feita a tentativa de utilizar a distância euclidiana no método MST-KNN, porém a aplicação deste diminuiu a qualidade dos resultados.

A dificuldade em justificar os clusters no contexto de gastos públicos já era esperada e fica evidenciada pelas complicações descritas acima. Em face dessa situação optamos por utilizar técnicas de visualização de dados como forma de encontrar relações nos dados que não estejam necessariamente relacionadas ao valor dos gastos, utilizando outros atributos, com o objetivo de utilizar os clusters como guia para explorar os dados. O que ficou claro era que mesmo que uma configuração do modelo fosse escolhida como a melhor, seria difícil justificar essa decisão por tratar-se de um problema não-supervisionado. A solução escolhida foi realizar o processo inverso: etiquetar um conjunto de indivíduos quanto à forma das suas séries temporais e utilizar essa intuição como base para o modelo, auxiliando na tarefa de determinar os melhores parâmetros.

O cluster escolhido era composto por 11 deputados cujas séries de gastos possuíam um comportamento em comum de possuirem picos

de gastos no fim de cada ano. Utilizando essa divisão como base, foi empregada uma métrica F1 para determinar qual das clusterizações obtidas pelo modelo melhor capturava esse comportamento. O parâmetros obtidos foram a distância de Jensen-Shanon e 3 como número de vizinhos e foram aplicados sobre a subquota de publicidade. O resultado foi uma clusterização com alta granularidade, gerando inúmeros elementos isolados, porém os poucos clusters grandes gerados apresentaram uma forma bem diferenciada para as séries temporais.

Como o objetivo nesse ponto era observar características visuais presentes nos clusters, foi desenvolvida uma aplicação web para inspecionar os dados e encontrar comportamentos escondidos nos números. Para que isso fosse possível foi necessário largar mão do método de normalização dos dados para que os dados pudessem ser melhor visualizados. Nesse ponto o tamanho dos clusters tem um papel fundamental na avaliação dos mesmos, pois conjuntos pequenos podem acabar sendo demasiadamente específicos, ao passo que clusters grandes podem ter uma forma mais geral, de difícil avaliação. Outro ponto crucial é o tipo de gasto, ou subquota, selecionado para visualizar, pois cada uma delas possuem uma forma diferente e uma única escolha de parâmetros seria incapaz de dividir corretamente os dados para todas as subquotas. No caso da divisão mencionada anteriormente, onde os deputados pertencentes ao conjunto possuem picos de gasto no fim do ano, essa relação está diretamente ligada aos gastos com publicidade.

A ferramenta possui duas partes principais, onde é possível realizar uma série de inspeções nos dados. A primeira parte possui duas visões que projetam os gastos dividindo-os em subregiões, sendo a primeira um gráfico de barras empilhado (figura 3) e a segunda um gráfico de pizza (figura 4). Ambas as visões podem ser segmentadas quanto a três critérios: estado, partido e subquota, além de poderem ser filtradas quanto aos três critérios e também por deputado. A segunda parte da ferramenta é focada nos clusters, tendo uma visão com um grafo relacional dos deputados (figura 5) e outra visão com as séries temporais dos gastos de cada cluster (figura 6).

Munidos da ferramenta de visualização desenvolvida iniciamos um processo de levantamento de hipóteses e avaliação dos clusters. É importante ressaltar que cada uma das divisões aplicáveis aos dados é potencialmente capaz de trazer informações úteis para as conclusões desse projeto. As medidas de qualidade definidas na seção 4 e o foco no tamanho dos clusters definido aqui servem como guia para melhor explorar o espaço de possibilidades, porém não definem uma solução categórica de qual é divisão correta dos dados. Na seção seguinte traremos as hipóteses levantadas e os resultados obtidos.

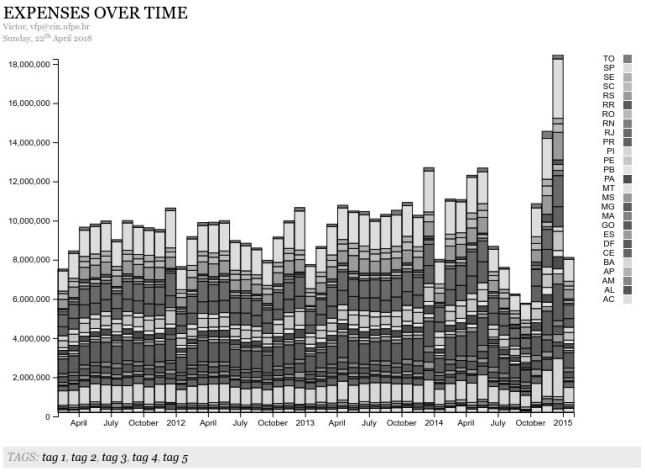

Figura 3: Projeção mensal dos dados divididos por estado.

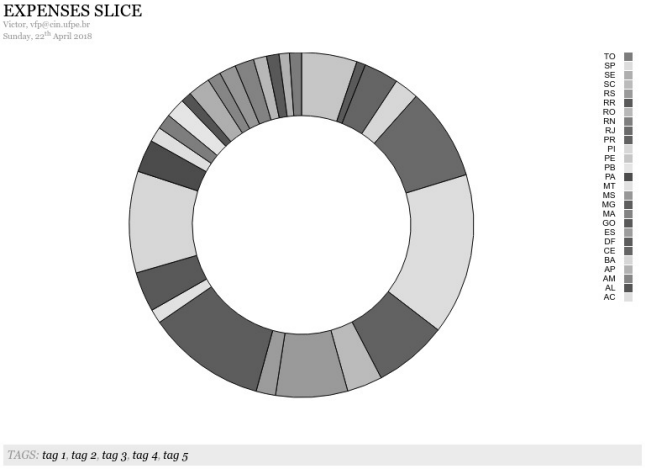

Figura 4: Projeção absoluta dos dados divididos por estado.

6 RESULTADOS

Usando a página é possível extrair diversas conclusões sobre o conjunto de dados de entrada. Como os dados são muito complexos, pode-se explorá-los de diferentes ângulos e chegar a conclusões diferentes. A ferramenta interativa é muito flexível e, como conseqüência, depende do conhecimento prévio do usuário sobre o assunto para alcançar os melhores resultados. Esta seção apresenta uma série de comportamentos gerais extraídos dos dados para a legislatura 54 para servir como exemplos de seu potencial.

6.1 Cenário geral

O cenário geral da câmara dos deputados tem alguns atributos típicos. A maioria dos parlamentares é das regiões nordeste e sudeste, as mais populosas, com forte presença dos estados de São Paulo, Minas Gerais, Rio de Janeiro e Bahia. Em relação ao tipo de gasto, emissão de passagens aéreas é a subquota mais proeminente, seguida de publicidade, telecomunicações e manutenção de escritórios.

Observando para a série temporal dos dados é possível perceber uma mudança no tipo de transporte usado pelo deputado em toda a legislatura. No início, há um foco em veículos terrestres e embarcações marítimas, que depois são alterados para terrestre e aeronave no final de 2014. A origem desse comportamento é desconhecida.

6.2 Polarização

Olhando para atributos específicos, há alguns sinais de polarização de alguns aspectos. Alguns deles são esperados como o tipo de transporte através do país, onde embarcações marítimas são mais frequentes na região norte onde existem mais rios, também os táxis e os serviços postais são mais utilizados na região sudeste com foco no estado de São Paulo, onde está localizada a maior e mais populosa metrópole. Outra situação esperada é o número mais alto de congressistas do espectro esquerdo no nordeste e da direito no sudeste, o que pode ser sentido nas eleições.

Algumas outras concentrações são mais difíceis de entender e requerem mais estudos, como altos gastos com segurança pela esquerda, mais especificamente pelo PSD. Outros partidos também dominam certos campos como o PT no transporte terrestre e fluvial e PR com o aluguel de embarcações.

6.3 Perfis

A configuração selecionada para os testes foi utilizando a distância *robust*, que é uma medida que une o coeficiente de correlação de Pearson e o coeficiente de correlação de postos de Spearman, e 2 como

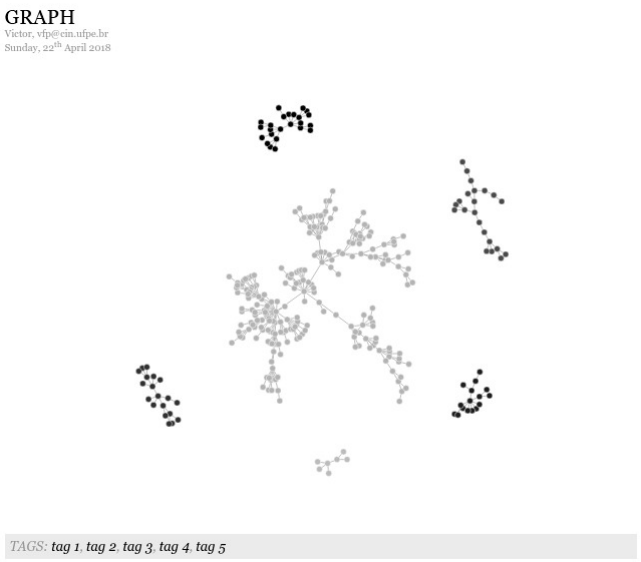

Figura 5: Grafo obtido utilizando distância cosseno e k=2.

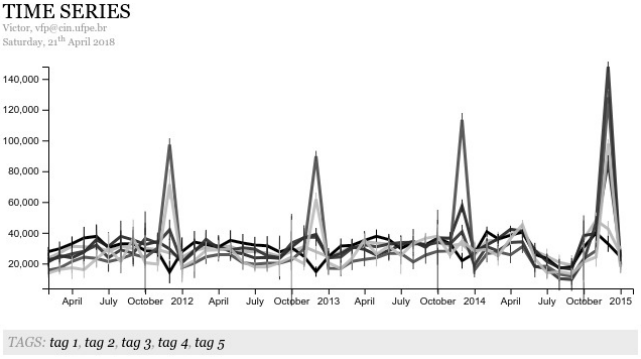

Figura 6: Série temporal referente ao grafo anterior.

número k de vizinhos para o K-NN. Essa configuração foi escolhida pois apresentou o melhor score de silhueta entre as demais e também porque os clusters construídos possuíam um tamanho ideal, nem muito pequenos, nem muito grandes. Seguindo essa configuração encontramos 6 clusters, dos quais tentamos identificar quais eram as características de destaque. Para se destacar esperávamos que as características dentro do cluster fossem significativamente diferentes do padrão da população inteira. A seguir descreveremos os clusters e seus principais pontos.

1. O primeiro cluster é o maior deles, contendo 125 deputados. Diferente dos demais, esse grupo possui picos no fim de cada ano causado por um alto gasto em publicidade. Por conta desse ponto o gasto total com publicidade neste cluster se iguala ao gasto com emissão de bilhetes aéreos, que é o maior gasto geral.

2. O segundo cluster possui 92 deputados, porém suas características são bastante similares do padrão geral, sem nenhuma propriedade que se destaque entre os demais clusters.

3. O terceiro cluster é composto por 64 deputados, concentrados no estado de Minas Gerais, superando o estado majoritário de São Paulo. Estes estão distribuídos principalmente nos partidos PT e PP, diminuindo a fatia tipicamente grande do PSDB. Algumas subquotas também são afetadas, com uma maior proporção de aluguel de veículos aquáticos e manutenção de escritório.

4. O quarto cluster tem 25 deputados, cuja maioria são do partido PR, o que não acontece no quadro geral. Os gastos desse grupo também possuem uma predominância em emissão de bilhetes aéreos maior do que o normal.

5. O penúltimo cluster possui apenas 8 deputados, contemplando poucos partidos e estados e marcados por um único pico no fim de 2014 onde os gastos com publicidade e consultoria dominam.

6. O último cluster possui apenas 4 deputados, pequeno demais para possuir qualquer característica destacável.

7 Conclusão e trabalhos futuros

Neste trabalho aplicamos uma técnica de clusterização para dividir o espaço de dados disponíveis no portal da câmara federal e isolar grupos de interesse para a população. Nesse contexto escolhemos um caso para demonstrar a validade do sistema proposto e que tipos de hipóteses podem ser levantadas. A análise realizada destaca alguns comportamentos dentro do vasto conjunto de dados inicial. É possível apontar alguns perfis interessantes e, usando a ferramenta, navegar dentro desses conjuntos de forma a obter mais detalhes sobre como a verba pública está sendo direcionada pelos deputados. Por exemplo, é possível isolar algum partido ou estado e visualizar como eles empregam a sua quota durante o ano, filtrando os pontos de interesse.

Sem dúvidas há mais informações e conclusões para serem extraídas desses dados, porém esperamos que esse trabalho possa demonstrar o potencial da ferramenta que deverá ser continuamente aprimorada e como ela pode ser útil para auxiliar a população para impulsionar a fiscalização social desejada. Para isso disponibilizamos uma versão de demonstração da ferramenta no link https://vfpimenta.github.io/.

Referências

[1] V. da Silva Figueiredo and W. J. L. dos Santos. Transparência e controle social na administração pública. *Temas de Administração Pública*, 8(1), 2013.

[2] S. Fan, B. Yu, and A. Saurkar. Public spending in developing countries: trends, determination, and impact. *Public expenditures, growth, and poverty*, pp. 20–55, 2008.

[3] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267, 2014.

[4] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[5] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323, 2004.

[6] R. Paranhos, W. Nascimento, D. Silva, J. A. da Silva Júnior, and D. B. Figueiredo Filho. Cota para o exercício de atividade parlamentar (ceap) dos deputados do nordeste (2011-2014). *Revista Estudos Legislativos*, (10), 2017.

[7] Y. Psycharis. Public spending patterns. In *Regional Analysis and Policy*, pp. 41–71. Springer, 2008.

[8] T. Zhang and H.-f. Zou. Fiscal decentralization, public spending, and economic growth in china. *Journal of public economics*, 67(2):221–240, 1998.