Análise de votações nominais do legislativo brasileiro utilizando componentes principais

Resumo

A literatura sobre análise quantitativa de votações nominais em casas legislativas é bem extensa e relativamente antiga. No entanto, poucos são os trabalhos que se debruçam sobre a aplicação de tais métodos nas casas legislativas brasileiras. Este artigo trás uma revisão de métodos de análise de votações legislativas nominais e discute a utilização da análise de componentes principais (ACP) como um método simples e eficaz para analisar votações nominais de casas legislativas. Apresentaremos também nossa abordagem para o tratamento de abstenções e sobre a realização de análises agregadas por partido, temas pouco explorados na literatura devido a pouca relevância no contexto de casas legislativas dos Estados Unidos. A avaliação do modelo é apresentada com medições comparadas ao W-NOMINATE.

1 Introdução

Modelos espaciais para análise de votações no âmbito legislativo existem pelo menos desde 1957, com Downs [1], e se tornaram mais numerosos e mais utilizados a partir da década de 1980, com o aumento da disponibilidade e redução de custo de processamento computacional e com a proposição em 1985 do famoso algoritmo NOMINATE por Poole e Rosenthal [2], até hoje o mais conhecido e utilizado.

O objetivo desses modelos de escalonamento dimensional é representar os parlamentares em um espaço geométrico com algumas poucas dimensões (frequentemente uma ou duas) de tal forma que o comportamento de cada um nas votações seja em grande parte explicado por sua posição nesse espaço. Essa posição do legislador é também chamada de "ponto ideal" e é estimada a partir dos votos observados nas votações nominais.

Existe uma literatura relativamente ampla sobre o assunto focando no Congresso Americano ou outras entidades estadounidenses como a suprema

corte, inclusive comparando o desempenho de diferentes modelos [3, 4], mas são poucos os estudos de votações em entidades brasileiras. Um exemplo é o estudo de Leoni, que analisou as votações da Câmara dos Deputados entre os anos 1991 e 1998 utilizando W-NOMINATE [5].

A Análise de Componentes Principais (ACP) é o método estatístico mais popular para redução dimensional de grandes conjuntos de dados [6], estando amplamente disponível em softwares e bibliotecas de matemática e estatística. Com redução dimensional queremos dizer que algoritmos são aplicados a informações n-dimensionais (ex: atuação de vários parlamentares em várias votações) sendo simplificadas em informações com menos dimensões (duas no caso dos mapas de votações por nós construídos). A redução de dimensões funciona como se enxergássemos apenas a "sombra" de determinado fenômeno. Embora seja uma informação simplificada e não completa, boa parte das vezes pode ser suficiente para melhorar o nosso entendimento sobre o fenômeno estudado, principalmente ao considerar que a informação original, n-dimensional, é de difícil assimilação devido a sua complexidade inerente.

A ACP é utilizada pelo aplicativo Radar Parlamentar[1] para a elaboração de mapas espaciais de votações do legislativo brasileiro. Mas para que essa aplicação seja possível, não basta simplesmente "usar a ACP". Uma série de decisões devem ser tomadas, tais como a modelagem dos dados que servirá de entrada ao algoritmo da ACP, o que fazer com abstenções, como considerar não só parlamentares mas também partidos, como comparar mapas de períodos diferentes, como tratar a mudança de partido de um legislador etc.

Dessa forma, o objetivo deste artigo é *apresentar e avaliar uma abordagem completa para a elaboração de mapas de votações espaciais para o legislativo brasileiro*. A abordagem aqui apresentada é basicamente a mesma utilizada pelo software Radar Parlamentar, porém com alguns ajustes menores para a apresentação impressa, sem os recursos interativos da web.

São contribuições deste artigo: 1) apresentação didática e detalhada dos principais trabalhos da literatura sobre modelos espaciais de votações parlamentares; 2) apresentação da nossa abordagem utilizando a ACP aplicada ao legislativo brasileiro; 3) apresentação de medidas de avaliação de nosso modelo em comparação com o W-NOMINATE. Além disso, disponibilizamos o banco de dados completo do Radar Parlamentar para que outros pesquisadores possam avançar em pesquisas futuras sobre análise quantitativa de votações parlamentares.

Este artigo está organizado da seguinte forma: na Seção 2 apresentamos em detalhes os principais trabalhos da literatura sobre modelos espaciais de

[1] `http://radarparlamentar.polignu.org/`

votações parlamentares, dando destaque aos trabalhos de Keith Poole; na seção 3 apresentamos nossa abordagem para a elaboração de mapas espaciais de votação, levando em conta questões ligadas a realidade do legislativo brasileiro; na Seção 4 explicamos como é realizada a avaliação do modelo, que é apresentada na Seção 5, juntamente com os mapas espaciais produzidos, e discutida na Seção 6. Por fim, nossas conclusões são apresentadas na Seção 7.

2 Revisão da literatura

Nesta seção fazemos uma revisão da literatura com o objetivo de contextualizar o leitor com o histórico de pesquisas na elaboração de mapas espaciais de votações. Optamos aqui por uma abordagem mais detalhada e didática dos conceitos apresentados pelos principais trabalhos da área, em detrimento de uma revisão mais completa da literatura em termos de trabalhos abordados. Esperamos dessa forma que este artigo seja útil também para que novos pesquisadores desvendem os complicados conceitos envolvendo a elaboração de mapas espaciais de votações.

Métodos de escalamento multidimensional analisam similaridades entre elementos de um conjunto, de forma a obter distâncias entre esses elementos em um espaço geométrico, possibilitando uma análise visual sobre os dados mais amigável que uma grande listagem de números [7]. Trabalhos como os de Pearson [8] e Hotelling [9] já consideram desde o início do século XX a aplicação dessas técnicas de escalamento multidimensional em áreas como economia, psicologia e educação.

A aplicação de modelos multidimensionais em economia está relacionada com a elaboração de modelos de escolhas racionais, nos quais consumidores manifestam preferências sobre produtos. Esses mesmos princípios passaram a ser usados nas ciências políticas. Já na década de 70, Davis et al. [10] modelam matematicamente o processo no qual eleitores escolhem candidatos. Para cada candidato o cidadão avalia as posições dos candidatos em diferentes assuntos, o que fornece uma localização dos candidatos em um espaço multidimensional. A escolha feita pelo cidadão é função da localização dos candidatos e dele próprio nesse espaço.

Seguimos agora com a descrição de modelos multidimensionais voltados mais especificamente à elaboração de mapas espaciais de votações no âmbito de casas legislativas.

Keith Poole: optimal classification e NOMINATE

Os trabalhos mais proeminentes publicados sobre a construção de mapas espaciais de votações são os de Keith Poole. Apresentaremos na sequência conceitos básicos e procedimentos utilizados por Poole na construção dos mapas espaciais. Todos esses conceitos e procedimentos estão descritos detalhadamente no livro de Poole sobre modelos espaciais de votações [11].

O primeiro conceito é do que seria uma *votação perfeita*. Considere um mapa espacial de votações onde cada ponto representa um parlamentar e cada linha representa uma votação, de forma que pontos de um lado da linha representam parlamentares que votaram SIM, enquanto que os pontos do outro lado da linha representam parlamentares que votaram NÃO. Para uma pequena quantidade de votações pode ser possível uma construção perfeita de tal mapa, como, por exemplo, podemos observar na Figura 1.

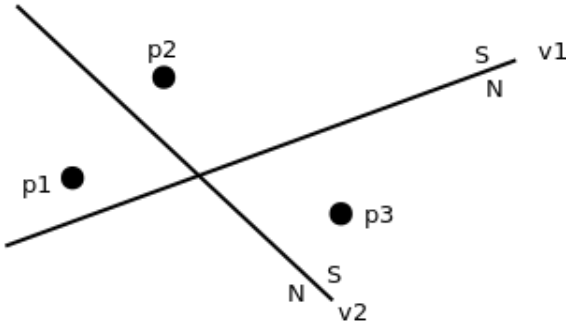

Figura 1: Mapa espacial de votação com classificação perfeita.

No exemplo fornecido, podemos ver pelo mapa que o parlamentar $p1$ votou SIM na votação $v1$ e NÃO para a votação $v2$. Já $p2$ votou SIM para $v1$ e SIM para $v2$. Por fim, $p3$ votou NÃO para $v1$ e SIM para $v2$. Como a partir do mapa descrevemos perfeitamente o comportamento dos parlamentares, dizemos que trata-se de uma votação perfeita.

Mas conforme o número de votações e de parlamentares cresce, percebe-se que é impossível posicionar perfeitamente todos os pontos em relação a todas as retas. Por isso é importante entender que o mapa de votações não é necessariamente construído de forma a descrever perfeitamente o comportamento dos parlamentares. Todo mapa gera distorções, e todo tipo de mapa procura amenizar algum tipo de distorção. Em mapas de votações, uma "distorção" a ser amenizada é a quantidade de classificações incorretas presentes no mapa.

Na concepção dos trabalhos de Poole, considera-se que em um mapa de votações cada parlamentar possui seu *ponto ideal* no espaço. Nesse mesmo espaço, uma votação também possui pontos associados às suas possíveis

opções (SIM e NÃO). As retas que representam as votações na Figura 1 são construídas em função dos pontos que representam suas respectivas votações. Dessa forma, a estimativa do comportamento de um parlamentar numa dada votação é função de relações entre seu ponto ideal e a representação espacial da votação.

Dados os pontos ideias associados a parlamentares e votações, uma primeira abordagem simplista para determinar o voto do parlamentar em uma votação seria dizer que o parlamentar vota deterministicamente na opção mais próxima de seu ponto ideal. Mas em vez disso, Poole utiliza o conceito de *função utilidade*, que atribui a cada ponto no espaço um valor. Quanto mais alto esse valor, maior é a *probabilidade* de que o parlamentar vote na opção associada a esse ponto do espaço.

Duas premissas importantes são usualmente aplicadas às funções utilidade: 1) as funções são de pico único (i.e., possuem apenas um ponto de valor máximo), sendo esse pico localizado no ponto ideal do parlamentar; 2) a função é simétrica, ou seja, o parlamentar é indiferente a duas opções igualmente distantes de seu ponto ideal. Uma função utilidade com essas características é ilustrada na Figura 2.

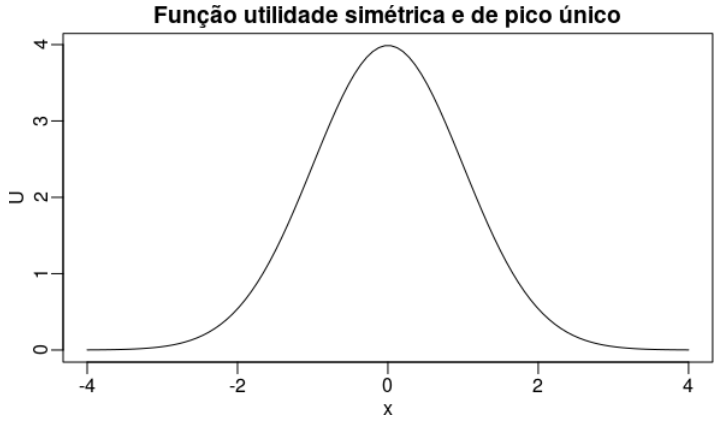

Figura 2: Exemplo de função utilidade simétrica e de pico único. O eixo x representa um espaço unidimensional de preferências políticas, enquanto que o eixo y representa o valor da função utilidade correspondente a um dado x.

A função utilidade possui uma parcela determinística e também uma parcela estocástica, que possibilita a modelagem de *erros de votação*. Um erro de votação seria a ideia de que o parlamentar não votou de acordo com suas preferências políticas. Um erro pode ter acontecido no sentido de que o parlamentar pode ter avaliado erroneamente a localização espacial das opções de uma dada votação. Mas o erro pode refletir também o fato de que fatores

subjacentes não captados pelo modelo foram decisivos na determinação da opção escolhida.

Dados os conceitos básicos apresentados (pontos ideias, funções utilidade e erros de votação) vamos descrever agora os principais métodos de construção de mapas espaciais de votações elaborados por Poole. São eles o *Optimal Classification* e o NOMINATE.

O *Optimal Classification* (OC) consiste em um processo iterativo[2] que procura maximizar a proporção de classificações corretas em um mapa espacial de votações. Dada uma configuração inicial de um mapa de votações, primeiramente aplica-se um algoritmo que maximiza a classificação correta fixando os pontos e movendo as linhas. Em um segundo passo, fixa-se as linhas e move-se os pontos para maximizar a classificação correta. Esses dois passos são repetidos várias vezes até que o erro (proporção de classificações incorretas) estabilize. Os algoritmos empregados garantem que a cada passo o erro nunca aumenta.

O OC não define uma posição exata dos parlamentares no mapa de votações, assim como não define uma distância exata entre dois dados parlamentares. O que o OC fornece são regiões do espaço nas quais os parlamentares podem ser posicionados. Essas regiões são denominadas de politopos e representam padrões de opções escolhidas nas votações. Voltemos à Figura 1: note que se alterarmos ligeiramente a posição de um parlamentar, digamos $p1$, seu padrão de opções escolhidas não se altera. Esse padrão se mantêm enquanto o ponto ideal não atravessar uma das retas que representam as votações. Dessa forma, dizemos que essa região do espaço delimitada pelas retas que mantêm o padrão de escolhas de $p1$ é o seu politopo.

Após a construção do mapa podemos observar que normalmente alguns parlamentares caem do lado errado de algumas retas. Isso significa que se o leitor do mapa fosse reconstituir os votos dados pelos parlamentares em cada votação, ele se enganaria em alguns casos. Esses erros representam uma imperfeição do mapa construído. A literatura costuma apresentar esses erros como uma incapacidade do modelo obtido em *predizer* corretamente os resultados de algumas votações. É preciso ficar atento com o uso do termo *predição*, pois os mapas de votações ou funções utilidades obtidos não serão utilizados na tentativa de predizer o resultado de votações futuras, ou mesmo votações passadas que não foram utilizadas para a produção do mapa de votações.

Para a utilização do OC, não é preciso premissas sobre a distribuição

[2]Um processo iterativo é aquele no qual um algoritmo é repetido várias vezes, sendo que depois de uma certa quantidade de repetições há uma *convergência*, ou seja, o resultado não mais se altera com mais repetições do processo.

da função utilidade. Apenas considerar que ela é simétrica e de pico-único. Embora o conceito de função utilidade não apareça diretamente na aplicação do algoritmo, ele é importante para explicar os erros de classificação, no sentido de que há uma certa probabilidade de que o legislador vote na opção contrária do que o mapa de votações indica.

Como exemplo de aplicação, Poole [11] mostra os resultados do OC quando aplicado à votação de revogação das leis do milho na Casa dos Comuns do parlamento inglês em 1846. Nessa situação, o algoritmo apresentou uma taxa de classificação correta de 95,2% para os 430 parlamentares que votaram nessa matéria. Já em outro estudo [3], analisando da 80ª à 104ª legislatura do senado dos EUA, Poole e Rosenthal obtiveram taxas de classificação correta para duas dimensões que vão de 85,8% a 91,3%.

O outro método de construção de mapas espaciais de votações consagrado por Poole é o NOMINATE, que constitui na verdade uma família de algoritmos com ligeiras variações entre si. Diferentemente do OC, o NOMINATE produz posições exatas dos pontos ideais de parlamentares e votações, podendo-se assim atribuir distâncias entre esses pontos.

No NOMINATE temos algumas premissas a mais sobre as funções utilidade. A mais importante é a de que a função utilidade é gaussiana (também chamada de "exponencial"). Outra opção utilizada em outros trabalhos [4] são funções quadráticas. Essa diferença diz respeito a como o parlamentar vai se comportar em relação a opções cada vez mais longes de seu ponto ideal. Essa diferença pode ser visualizada nos gráficos da Figura 3.

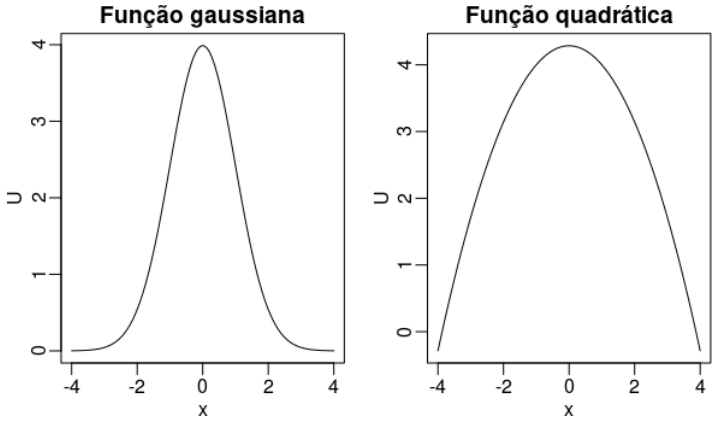

Figura 3: Ilustração gráfica da diferença entre uma função gaussiana e uma função quadrática.

Com funções gaussianas, as opções cada vez mais longes do ponto ideal se tornam cada vez mais indistinguíveis. Por exemplo: numa reforma do congresso, um parlamentar pode achar que uma casa legislativa com mais de

100 membros torne o debate inviável. Dessa forma, propostas de casas com 800 ou 1400 membros são igualmente ruins para o parlamentar, pois ambas já passaram do ponto aceitável.

Já para funções quadráticas, conforme as opções se distanciam do ponto ideal, elas se tornam cada vez pior avaliadas. Por exemplo, um parlamentar que é contra o aumento de impostos vai achar que um aumento de 100% é inaceitável, mas mesmo assim achará um aumento de 200% ainda pior.

Agora vamos detalhar a função utilidade utilizada no NOMINATE [2]. Considere a localização x de um parlamentar e a localização o de uma opção de uma votação. Essas localizações são pontos em um espaço multidimensional, onde cada dimensão representa preferências sobre um determinado tema político. Para uma votação temos o_y, a localização da opção SIM, e o_n, a localização da opção NÃO.

A função utilidade utilizada pelo NOMINATE [2] é

$$U(x, o) = \beta * e^{\frac{-w^2 * d^2}{2}} + \varepsilon,$$

onde d representa a distância entre os pontos x e o. Quanto mais perto x está de o, maior o valor de U, sendo U simétrica e de pico único. β e w são parâmetros da função utilidade, sendo β o fator de ruído, que determina o peso da parcela determinística da função utilidade. ε é a parcela estocástica, que representa erros distribuídos independentemente conforme a distribuição logística. Poole afirma que teoricamente a distribuição normal para ε seria mais adequada. No entanto, devido às limitações computacionais da época, Poole optou pela distribuição logística, que é mais simples do ponto de vista computacional e razoavelmente similar à distribuição normal.

No NOMINATE temos então três grupos de parâmetros: pontos ideias dos parlamentares (um x para cada parlamentar), pontos ideais das opções das votações envolvidas (um o_n e um o_y para cada votação) e os parâmetros da função utilidade (β e w). O objetivo do NOMINATE é encontrar os valores de todos esses parâmetros para que se possa desenhar o mapa espacial de votações.

Dada uma configuração inicial de pontos ideias de legisladores e opções de votações, o NOMINATE aplica sucessivamente os três seguintes passos: 1) estima-se os parâmetros da função utilidade com base nos parâmetros dos legisladores e das votações; 2) estima-se os parâmetros dos legisladores com base nos parâmetros das votações e da função utilidade; 3) estima-se os parâmetros das votações com base nos parâmetros dos legisladores e da função utilidade. Cada um desses passos possui seu próprio algoritmo com suas complexidades. Os três passos são repetidos até a convergência, que é quando o refinamento para de ter efeito e os mesmos valores são produzidos.

Tanto o OC quanto o NOMINATE partem de uma *configuração inicial* do mapa de votações a ser refinada. Poole define essa configuração inicial por meio da análise de componentes principais.

Analisando 172 votações realizadas por 440 parlamentares na 85ª legislatura do congresso dos EUA, Poole e Rosenthal [2] obtêm uma taxa de classificação correta de 78,9% utilizando o NOMINATE com apenas uma dimensão.

Formas de se avaliar os resultados produzidos

Para avaliar os resultados de um modelo espacial de votações podemos usar a taxa de classificação correta, que nos informa a porcentagem de acertos e erros do modelo. Sendo U a função utilidade, se os parâmetros obtidos pelo modelo definem que $U(\text{SIM}) > U(\text{NÃO})$ pra um determinado parlamentar em uma determinada votação, temos um acerto caso o parlamentar realmente tenha votado SIM naquela votação ou um erro se na realidade o parlamentar votou NÃO.

No entanto, o problema é que modelos extremamente ingênuos podem obter boas taxas de classificação correta. Exemplo: o modelo pode prever que todos os parlamentares votam na opção vencedora. Dessa forma, numa votação em que 90% dos legisladores votaram com a maioria, o modelo ingênuo teria apenas 10% de erro. Por isso, utiliza-se também a PRE (redução proporcional do erro) [5], que mede em quantos porcentos o erro foi reduzido do modelo ingênuo para o modelo avaliado. Ou seja,

$$PRE = \frac{\text{votos da minoria} - \text{erros do modelo}}{\text{votos da minoria}}.$$

A fórmula da PRE considera apenas uma votação. Para se avaliar um conjunto de votações, se utiliza a APRE (redução proporcional do erro agregado) [5], onde a quantidade de votos na minoria e a quantidade de erros são consideradas para todas as votações.

Aplicações no congresso brasileiro

Aplicando o W-NOMINATE, uma das variações do NOMINATE, à Câmara dos Deputados, Leoni [5] encontra taxas de classificação correta de 86,4% e 90,4% para as 49ª e 50ª legislaturas respectivamente. Já em termos de APRE, as taxas encontradas são de 52,3% e 64,8% para as mesmas legislaturas.

Izumi [12] elabora mapas espaciais de votações para o Senado brasileiro. No entanto, ele argumenta que o NOMINATE pode não ser adequado por

causa dos pressupostos envolvendo a função utilidade. O primeiro pressuposto questionado é a simetria da função utilidade, uma vez que, por exemplo, a redução em 5% dos impostos pode ser algo muito mais importante para um parlamentar do que um aumento da mesma magnitude.

O segundo pressuposto questionado da função utilidade é o de que os erros são independentes e identicamente distribuídos entre os legisladores e as votações. Argumentos contrários a esse pressuposto: 1) existem partidos mais coesos que outros (o PT, por exemplo, costuma ser bem mais coeso que outros partidos); 2) migrações de parlamentares entre partidos e migrações de partidos para dentro ou fora da base governista podem alterar a variação de erros ao longo do tempo; 3) em determinados contextos o voto estratégico pode prevalecer sobre o voto sincero. Interessante notar aqui como os argumentos 1) e 2) são características que diferenciam o estudo do legislativo brasileiro do legislativo norte-americano.

Embasado por esses questionamentos, Izumi prefere adotar o Optimal Classification, pois este não se sustenta sobre pressupostos da função utilidade ou da distribuição de erros. Dessa forma, aplicando o OC a seis legislaturas do Senado (da 48ª à 53ª), Izumi encontra taxas de classificação correta entre 90,7% e 98,5%. Já em termos de APRE, as taxas encontradas são entre 50% e 94%. Embora o trabalho de Izumi se aplique ao Senado, cabe notar aqui que todos seus argumentos também se aplicariam à Câmara dos Deputados.

Cabe notar que parte das críticas de Izumi ao NOMINATE se aplicam também ao OC. Como disse Poole [13], as únicas premissas para a aplicação do OC são 1) o espaço de escolha é euclideano; e 2) preferências são simétricas e de pico-único. Como o próprio Izumi disse, a simetria é uma premissa que pode ser questionada na prática, e essa não somente para o caso brasileiro.

Modelos lineares de Heckman e Snyder

Muitos trabalhos consideram modelos lineares para a análise de votações nominais menos adequados do que os modelos não lineares, como o NOMINATE de Poole. Heckman e Snyder [14], porém, demonstram rigorosamente a equivalência de modelos lineares com os resultados obtidos por Poole, tendo como principal vantagem a simplicidade e eficiência computacional dos métodos lineares.

Para Heckman e Snyder, a decisão de se votar SIM ou NÃO em uma votação é modelada como o resultado de um processo de escolha racional no qual os legisladores usam suas preferências para ponderar sobre as características da votação. Assim, dizemos que uma opção é localizada em um espaço no qual cada dimensão seria uma característica da opção. Nesse caso,

a função de utilidade recebe um vetor no espaço de características e retorna um número real. Cada legislador teria sua própria função utilidade e escolheria a opção que resultasse no maior valor da função utilidade. Diferentemente do NOMINATE, temos aqui uma função utilidade quadrática.

A função utilidade de Heckman e Snyder, assim como a do NOMINATE, é incrementada com uma parcela de erros aleatórios. Considera-se que uma das fontes de erro seja a dificuldade para o parlamentar estimar o valor de cada característica.

Os autores utilizam diferentes métodos algébricos para estimar as preferências dos legisladores. Um dos métodos mais simples utilizados é justamente a Análise de Componentes Principais (ACP). Leo ►*Elaborar relação entre o PCA e a função utilidade. No que o nosso uso aparentemente mais simples da PCA é diferente do uso aparentemente mais elaborado da PCA como feito por Heckman?*◄

Dimensionalidade dos mapas de votações

Um debate existente na literatura é sobre a quantidade de dimensões necessárias para representar um mapa espacial de votações. De acordo com McCarty [15], essa é uma polêmica sem conclusão definitiva.

Segundo McCarty [15], a utilização de uma segunda dimensão no D-NOMINATE[3] acrescenta um poder explicativo de apenas 4% sobre o poder de explicação de 87% já fornecido pela primeira dimensão. No entanto, de 1945 até a década de 60, questões raciais e de direitos civis fazem com que a segunda dimensão seja mais significativa.

A primeira dimensão presente nos mapas de votações do legislativo americano estaria ligada ao espectro político-ideológico presente nesse país, possuindo uma escala que vai do extremo liberal ao extremo conservador. No período em que a segunda dimensão se torna significativa, evidencia-se também um padrão de votação que diferencia parlamentares do norte e do sul dos EUA.

Já para Heckman e Snyder [14], existem pelo menos cinco dimensões significativas. Essas dimensões a mais podem não fazer tanta diferença na taxa global de sucesso de classificação, mas são decisivas em algumas votações específicas, por representarem substantivamente assuntos específicos. Exemplos de tais assuntos: direitos civis e eleitorais, agricultura, ajuda internacional, gasto militar, teto da dívida, água, aborto e reforma do congresso.

Utilizando o W-NOMINATE, Leoni [5] chegou à conclusão de que uma dimensão explica a maior parte das votações na Câmara dos Deputados,

[3] O D-NOMINATE, ou *dynamic NOMINATE*, é uma variação do NOMINATE para a comparação intertemporal dos pontos ideais nos mapas de votações.

pois dimensões adicionais não melhoram significativamente a capacidade explicativa do modelo em termos da taxa de classificação correta e da APRE. Izumi [12], utilizando os mesmos critérios de Leoni, mas aplicados ao OC, também conclui pela unidimensionalidade do espaço legislativo do Senado brasileiro.

Discussão

Leo ►*Oi Saulo, principalmente essa parte que eu queria ver oq vc acha... se oq tá escrito aí não é besteira =S*◄

Saulo ►*Leo, acho que você teve a sensação de que poderia ter um erro conceitual ao dizer nessa seção que a abordagem não tem uma teoria por trás. De fato, ao fazermos a PCA estamos assumindo de forma implícita as mesmas coisas basicamente que Heckman e Snyder. Nosso modelo é praticamente o mesmo que o deles, na verdade eu não sei se há alguma diferença além da tendência deles de dar mais ênfase a interpretação do significado das dimensões, enquanto nós nos preocupamos mais com a interpretação mais pragmática de estimar as distâncias entre parlamentares. Me parece que a hipótese de função de utilidade quadrática (de Heckman-Snyder) está relacionada ao fato de que a PCA minimiza distâncias euclidianas, ou seja, quadráticas. Ao fazermos uma "fotografia puramente algébrica" adotamos um algoritmo para fazer isso; o algoritmo que escolhemos não é o único possível. Diz-se que esta é a forma que mais minimiza a perda de informação, porém ao dizer isso pressupõe-se uma métrica; se usássemos outra métrica o resultado seria outro.*◄

Como visto, os trabalhos em ciências políticas normalmente não buscam somente a elaboração de mapas espaciais de votações, mas também a elaboração de teorias e modelos que embasem esses mapas. Porém, toda teoria parte de algumas suposições, que no caso concreto podem ou não serem verdadeiras. Davis et al. [10], por exemplo, pressupõe que todos os eleitores avaliam os candidatos sob o prisma dos mesmo assuntos. Poole, tanto no OC quanto no NOMINATE, pressupõe funções utilidade simétricas e de pico único. Já no NOMINATE, temos ainda que considerar que erros são independentes e identicamente distribuídos. Mesmo Heckman e Snyder [14], que utilizam modelos lineares, se apoiam fortemente em modelos de escolha discretas, considerando que o legislador modela suas preferências de forma multidimensional e aplica uma função utilidade quadrática sobre esses valores para tomar sua decisão.

Embora o uso de suposições para a elaboração de teorias seja comum e necessário, podemos aqui ponderar sobre uma vantagem da utilização da ACP para a elaboração dos mapas de votações. Como trata-se de uma abordagem puramente algébrica, sem uma teoria por trás, podemos entender o gráfico gerado pela ACP como uma fotografia do comportamento dos parlamentares

em um determinado período. Essa fotografia não é perfeita, há uma perda de informação, mas que pode ser medida como veremos mais a frente. Essa fotografia não traz consigo explicações sobre o processo de decisão ou do porquê dos resultados, mas possibilita a um analista político que faça considerações sobre a conjuntura política de uma determinada casa legislativa em um determinado período, ajudando o analista a chegar a possíveis conclusões.

Ou seja, mesmo sem o embasamento de uma teoria, o mapa de votações produzido pela ACP é útil. E justamente por não estar baseado em uma teoria forte, pode-se dizer que esse mapa não tem a validade limitada por suposições que podem não ser verdadeiras no caso concreto. Saulo ►*Em suma, entendo o que você quer dizer aqui, mas acho que não está correto dizer dessa forma. Existe uma teoria por trás sim. Nossas suposições são similares as de Heckman e Snyder. Na minha opinião a força deste tipo de modelo é que as suposições são simples, e o modelo é fácil de interpretar, fácil de reproduzir, e dá resultados parecidos com os outros modelos, então se o objetivo é informar o público mais amplo este modelo é suficiente, e até, pela simplicidade, preferível.*◄

3 A ACP para análise de votações nominais

Nesta seção apresentaremos nossa abordagem completa de análise quantitativa de votações nominais para a elaboração de mapas espaciais de votação. Algumas das decisões de modelagem apresentadas levam em conta características do legislativo brasileiro, o que diferencia nossa análise de outros trabalhos da literatura que usualmente focam no legislativo dos EUA.

Considera-se uma casa legislativa com M membros (parlamentares) e N votações nominais de interesse. O voto x_{ij} de um parlamentar j em uma votação i será modelado por um valor numérico como segue:

$$x_{ij} = \begin{cases} 1 & \text{, se parlamentar votou } sim \\ -1 & \text{, se parlamentar votou } não \\ 0 & \text{, em qualquer outro caso} \end{cases}$$

Os outros casos além do sim e do não podem consistir em abstenção, obstrução ou ausência do parlamentar, ou situação em que este não esteja exercendo o mandato na data em que a votação ocorreu. Todos esses casos representam uma impossibilidade de verificar a opinião do parlamentar sobre a votação, e por isso são modelados por um valor euclidianamente equidistante das duas opções.

Normalmente a análise de componentes principais não é adequada para variáveis categóricas, porém neste caso as categorias podem ser claramente representadas em um eixo cartesiano com dois extremos: SIM e NÃO. O

valor de x_{ij} pode ser interpretado como um estimador para um ponto de utilidade máxima ξ_{ij} do legislador j face à decisão i situado em uma escala contínua de valores deste eixo, tal que quando $\xi_{ij} > 0$ o legislador tende a preferir o SIM, e com mais convicção ou maior importância dada à questão quanto mais distante do zero, e analogamente para $\xi_{ij} < 0$ e a opção NÃO. Ora, o comportamento observado que é o voto, por sua natureza categórica, não permite dizer o grau de importância dada ou a convicção com que o parlamentar decidiu por uma ou outra opção, mas é razoável supor que os x_{ij} tal como definidos acima forneçam um estimador para os ξ_{ij}.

Fica definida a matriz de votações $\mathbf{X}$:

$$\mathbf{X} = \begin{array}{c} \\ \text{votações} \downarrow \end{array} \begin{array}{c} \overrightarrow{\text{membros}} \\ \begin{array}{ccccc} 1 & & j & & M \end{array} \end{array} \begin{array}{c} \\ \begin{array}{c} 1 \\ \\ i \\ \\ N \end{array} \end{array} \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NM} & \dots & x_{NM} \end{bmatrix}$$

Por definição esta matriz contém apenas os valores -1, 0 e 1. Para realizar a análise de componentes principais, define-se a matriz centralizada $\mathbf{X}^*$, subtraindo de cada entrada a média da linha:

$$x^*_{ij} = x_{ij} - \langle x_{ij} \rangle_j \tag{1}$$

onde $\langle \cdot \rangle_j = \frac{1}{M} \sum_{j=1}^{M} \cdot$ denota a média nos j.

Define-se a matriz de centralização $\mathbf{C}$ por:

$$c_{ij} = \langle x_{ij} \rangle_j \qquad i = 1..N;\ j = 1..M$$

de forma que:

$$\mathbf{X}^* = \mathbf{X} - \mathbf{C}$$

A variância (amostral) $\text{var}(i)$ de cada votação, ou dimensão é:

$$\text{var}(i) = \frac{\sum_{j=1}^{M} \left(x_{ij} - \langle x_{ij} \rangle_j \right)^2}{M-1} = \frac{M}{M-1} \left(\langle {x_{ij}}^2 \rangle_j - \langle x_{ij} \rangle_j^2 \right)$$

$$\text{var}(i) = \frac{M}{M-1} \langle {x^*_{ij}}^2 \rangle_j \tag{2}$$

A análise de componentes principais consiste em uma rotação de base $\mathbf{R}$ deste espaço vetorial tal que os dados (centralizados) transformados $\mathbf{\Gamma} =$

$\mathbf{R} \cdot \mathbf{X}^*$ concentram a máxima variância possível na primeira dimensão, a segunda dimensão possui a máxima variância possível sob a restrição de ser ortogonal à primeira, e assim sucessivamente. A cada vetor da nova base é dado o nome de *componente principal*, os valores de $\mathbf{R}$ são chamados *pesos* (ou *loadings*) e as coordenadas obtidas em $\mathbf{\Gamma}$ são chamadas de *scores*.

Como a matriz de rotação $\mathbf{R}$ é ortonormal, sua inversa é igual à transposta $\mathbf{R}^t$, e tem-se $\mathbf{X}^* = \mathbf{R}^t \cdot \mathbf{\Gamma}$.

Se forem mantidos apenas os $d \leq N$ primeiros componentes principais, a parte relevante da matriz de rotação, que chamaremos de $\mathbf{R}_{(d)}$, e da matriz de scores, $\mathbf{\Gamma}_{(d)}$, terão apenas d linhas, e $\mathbf{R}^t_{(d)} \cdot \mathbf{\Gamma}_{(d)}$ será a melhor aproximação de $\mathbf{X}^*$ que pode ser obtida com um modelo linear deste tipo com d dimensões, onde "a melhor aproximação" se refere à minimização da soma dos quadrados das diferenças das entradas[4].

Utilizando uma nomenclatura usual em análise de votações legislativas, as coordenadas de cada parlamentar j retidas em $\mathbf{\Gamma_{(d)}}$ podem ser entendidas como o *ponto ideal* do parlamentar no espaço d-dimensional de preferências políticas.

Exemplificando para o caso comum em que $d = 2$, a equação $\mathbf{X}^* \approx \mathbf{R}^t_{(2)} \cdot \mathbf{\Gamma}_{(2)}$ foi reescrita abaixo:

$$\underset{\text{vot.}}{\begin{bmatrix} x^*_{11} & \cdots & \overset{\text{membros}}{x^*_{1M}} \\ \vdots & \ddots & \vdots \\ x^*_{N1} & \cdots & x^*_{NM} \end{bmatrix}} \approx \underset{\text{vot.}}{\begin{bmatrix} R_{11} & \overset{\text{C.P.}}{R_{21}} \\ \vdots & \vdots \\ R_{1N} & R_{2N} \end{bmatrix}} \cdot \underset{\text{C.P.}}{\begin{bmatrix} \gamma_{11} & \cdots & \overset{\text{membros}}{\gamma_{1M}} \\ \gamma_{21} & \cdots & \gamma_{2M} \end{bmatrix}}$$

Centralização e Normalização

Em diversos contextos em que se aplica a ACP é comum realizar a *centralização* (subtraindo de cada entrada o valor médio da linha) e a *normalização* (multiplicando cada entrada por um fator de escala igual ao inverso da variância da linha, de forma a obter variância unitária para todas as direções da base original) de $\mathbf{X}$ antes de proceder à análise.

O algoritmo de determinação das componentes por SVD não é baseado na variância em si, e sim na soma dos quadrados. Para variáveis centralizadas as duas quantidades são proporcionais (vide equação 2), por isso a centralização é recomendável para variáveis que não possam ser supostas de média zero. No caso de votações legislativas a centralização introduz N parâmetros ao modelo (através dos valores L.I. da matriz $\mathbf{C}$), que podem ser interpretados como sendo relacionados aos tamanhos da maioria e minoria de cada votação.

[4]Em outras palavras, o modelo minimiza a norma de Frobenius da matriz de votações.

Já a normalização é em geral recomendável quando as componentes originais possuem unidades de medida distintas, para evitar que dimensões com variâncias numericamente grandes predominem artificialmente. Como todas as votações possuem a mesma "escala", não se faz necessária a normalização. De fato, para o caso de uma votação quase unânime o fator de escala (1/variância) seria muito alto, pois a variância de uma votação quase unânime é baixa, e esta votação receberia um peso maior na composição das componentes principais apenas por ter sido menos acirrada.

Estas considerações sugerem a adoção da centralização, mas não da normalização, na análise de votações utilizando ACP.

Escolha do Número de Dimensões *d*

O modelo será tanto mais preciso na classificação correta das votações quanto maior for o número de dimensões retidas $d \leq N$. Porém está claro que um modelo simples é mais útil: analisar cada uma do total de N dimensões seria tão trabalhoso quanto analisar individualmente cada uma das N votações (e tão completo quanto). O objetivo é simplificar, retendo o essencial da informação.

Uma forma de quantificar a informação retida (ou perdida) ao considerar apenas d dimensões é observar qual é a fração $\nu_d \leq 1$ da variância total explicada:

$$\nu_d = \frac{\sum_{i=1}^{d} \frac{M}{M-1} \langle {\gamma_{ij}}^2 \rangle_j}{\sum_{i=1}^{N} \mathrm{var}(i)}$$

onde o numerador é a soma da variância das d primeiras componentes principais, e o denominador é a variância total da matriz de votações.

Quanto maior for ν_d mais preciso será o modelo. Uma prática comum é adotar d tal que se fosse adotado $d+1$ o ganho em ν_d seria pequeno. Dito isso, o critério é arbitrário, e deve depender do objetivo da análise. Para uma visualização do aspecto geral de distribuição dos parlamentares é prático utilizar $d = 2$, já que assim a visualização no plano é muito mais simples. Seja qual for, a escolha deve vir acompanhada do valor de ν_d correspondente, afim de que se possa ter uma idéia de quanta informação está sendo desconsiderada.

Análise por partido

No modelo apresentado, nada impede que os valores de $\mathbf{X}$ possuam valores reais, situados por exemplo no intervalo [-1;1], em vez de apenas os valores discretos {-1;0;1}. Esta observação permite uma extensão direta do modelo

para analisar os parlamentares agregados por partido em vez de considerá-los individualmente, bastando considerar o voto médio do partido em cada votação antes de iniciar a análise.

O voto médio do partido k na votação i é definido por:

$$x_{ik} = \frac{1}{|k|} \sum_{j \in k} x_{ij} \tag{3}$$

onde $j \in k$ denota que o parlamentar j pertence ao partido k, e $|k|$ é o número de parlamentares do partido k considerados.

Esta análise é útil para analisar afinidades partidárias e coalizões em ambientes com vários partidos, como é tipicamente o caso das casas legislativas no Brasil.

No Radar Parlamentar, porém, não aplicamos essa técnica, pois para esse software optamos pela coexistência de partidos e parlamentares individuais no mesmo mapa especial. O método proposto acima é válido, mas seu resultado não possui relação direta com o resultado da análise por parlamentar. Assim sendo, no Radar Parlamentar os partidos são posicionados no centroide das posições ocupadas por seus parlamentares. Ou seja, a posição de um partido P com k parlamentares $p_1, ..., p_k$ se dá por:

$$x_P = \frac{\sum_{i=1}^{k} x_{p_i}}{k} \tag{4}$$

No Radar Parlamentar também optamos por diferenciar os partidos em função do tamanho da bancada de cada partido no período analisado. Para isso, cada circunferência representando um partido possui o raio proporcional à raiz quadrada do tamanho da bancada no período.

A Figura 4 permite visualizar o resultado do posicionamento do partido pelo centroide de seus parlamentares, assim como observar o efeito do tamanho das circunferências representando os partidos.

Tratamento de valores faltantes

Todos os métodos de análise de votações legislativas encontrados na literatura revisada descartam ausências e abstenções antes de iniciar a análise, considerando explicita- ou implicitamente que tais atitudes não trazem informação acerca das preferências políticas do legislador, e notando que tais situações representam a minoria dos casos. Por exemplo, Heckman e Snyder notam que as abstenções representam menos de 1% dos votos para câmara e senado estadounidenses, e as assumem aleatórias em relação a resultados das votações e a preferências dos parlamentares [14, p.40].

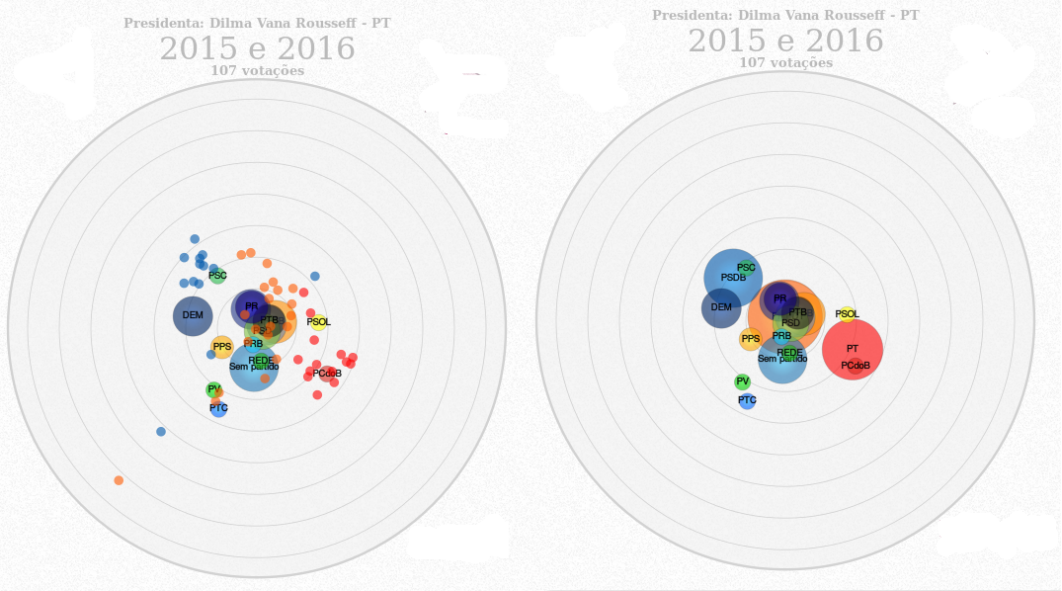

Figura 4: Nos mapas espaciais justapostos é possível observar o resultado do posicionamento de três partidos em função de seus centroides: PT (vermelho), PSDB (azul) e PMDB (laranja). No período analisado, PT, PSDB e PMDB contaram com 14, 13 e 21 senadores respectivamente, o que afeta o tamanho das circunferências representando os partidos.

Já na legislatura da Câmara dos Deputados brasileira no período de 2011 a 2014, temos cerca de 5% entre abstenções e obstruções dentro do universo de votos dados, enquanto que ausências são cerca de 36% dentre o total possível de votos a serem dados. Esses valores foram estimados com base nos dados abertos da Câmara. Propõe-se então que o comportamento observado de ausentar-se ou abster-se de uma votação traz sim informação acerca das preferências do parlamentar que se deseja estimar, e por isso essa informação não deve ser descartada na análise.

No modelo proposto, ausência, abstenção e obstrução são modeladas através do valor 0. Em relação à alternativa de descartar estas situações, que serão referidas genericamente como votos "nulos", esta modelagem introduz um viés no sentido oposto ao voto médio dos que realmente votaram. Ou seja, supondo sem perda de generalidade que o voto da maioria é sempre SIM, o voto médio dos que votaram será sempre maior que zero, e se o parlamentar faz voto nulo sua preferência nesta votação será modelada como sendo ligeiramente oposta ao SIM (pois seu voto é numericamente menor do que a média), mas não tão oposta quanto se o parlamentar tivesse efetivamente votado NÃO.

Este viés pode parecer arbitrário, porém esta abordagem é consistente tanto com a ideia de que um voto nulo representaria uma indiferença do parlamentar quanto aos resultados SIM e NÃO (o voto nulo é euclidianamente equidistante das duas alternativas) quanto da ideia de que ao não votar o parlamentar pode ter uma preferência contrária àquela que se imagina que será aprovada na votação, como em um "boicote" pessoal (ou em grupo) à votação. Em outras palavras, um parlamentar teria maior tendência de comparecer e não se abster nem obstruir a votação em propostas nas quais ele esteja inclinado a votar com a maioria. Se estas hipóteses são arbitrárias, pode-se dizer que são pelo menos tão arbitrárias quanto a alternativa de considerar que um voto nulo equivale a um parlamentar com preferência igual à preferência média da casa. Nossos resultados sugerem que de fato a forma proposta de modelagem melhora os índices de classificação correta.

No caso da análise por partidos, ao excluir votos nulos do cálculo da média na equação 3 estaria-se buscando considerar que a opinião "do partido" é composta apenas pela opinião daqueles que votaram ou SIM ou NÃO. Outra opção é excluir apenas as ausências, se os dados permitirem discriminar esta opção. Os resultados aqui apresentados não excluem estes votos, para que a análise reflita o fato de que uma abstenção ou mesmo uma ausência não são equivalentes a concordar com a opinião geral do partido. Além disso a análise fica mais simples, já que não há necessidade de tratamento especial de partidos que tenham estado por exemplo cem porcento ausentes em uma dada votação.

Tratamento de votações unânimes

E tb de parlamentares que faltam bastante.

Acho que no caso, simplesmente não consideramos esses problemas e usamos todo mundo.

Mas pelo menos sobre votações unânimes isso tem a parte ruim de deixar o mapa mais embolado e difícil de ler. Por outro lado, ajuda a mostrar a realidade de que na prática, apesar das diferenças, eles não são tão diferentes assim.

NOMINATE faz uns cortes pra não pegar todo mundo.

Lidando com migração partidária

Nossa abordagem: quando o parlamentar troca de partido, se torna um novo parlamentar. Desvantagem: parlamentar estava no partido A no período 1 e foi para o partido B no período 2. Na matriz de votação que engloba os períodos 1 e 2, a linha do parlamentar no partido A estará cheias de direita

à esquerda e a linha do parlamentar no partido B estará cheias de zeros à esquerda.

ToDo ►*Citar Izumi*◄

Análise temporal

Para o analista político é interessante visualizar como a conjuntura de alianças parlamentares se altera ao longo do tempo. Daí a necessidade de algoritmos que possibilitem a comparação entre mapas de votações de períodos distintos. Por conta dessa necessidade, desenvolvemos um algoritmo próprio de orientação e escalonamento de um conjunto de mapas de votações.

O objetivo de nosso algoritmo para análise temporal não é possibilitar a comparação entre as posições de um parlamentar ao longo do tempo, pois os eixos e posições se alteram em seus significados de um período para o outro. O que pretendemos é que se possa comparar as distâncias relativas dos parlamentares entre si ao longo do tempo. Dessa forma é possível verificar nos mapas de votação, por exemplo, o afastamento e aproximação entre partidos ao longo do tempo. Na Figura 5 temos um exemplo de alteração de posições relativas com profundo significado político.

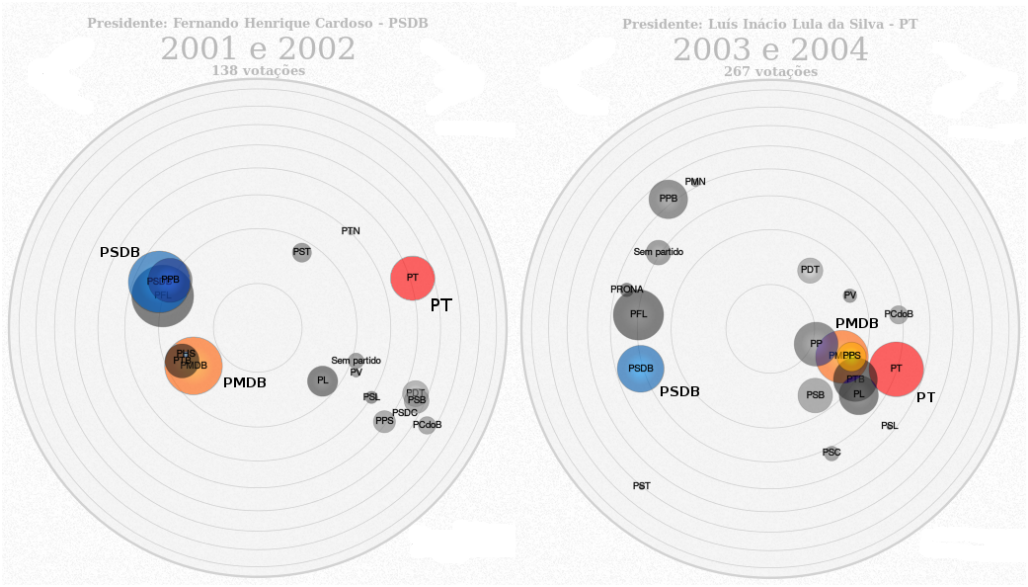

Figura 5: Os dois mapas de votações destacam a mudança de posição relativa do PMDB, indo da proximidade com o PSDB no fim do governo FHC para a proximidade com o PT no primeiro governo de Lula.

O problema está no fato de que mesmo que pouco se altere entre as posições relativas dos parlamentares de um período para o outro (i.e. parlamentares continuaram votando com as mesmas alianças), um mesmo parlamentar pode ocupar posições totalmente diferentes nos diferentes mapas produzidos pela ACP. Até mesmo para exatamente o mesmo conjunto de votações, detalhes de precisão de cálculos devido a utilização de diferentes máquinas podem resultar em mapas rotacionados um em relação ao outro.

Para melhor visualizar as mudanças nestas posições relativas é conveniente rotacionar o resultado de uma das análises em torno da origem, de tal forma que seja minimizada alguma medida das "distâncias"percorridas pelos partidos entre os períodos. A Figura 6 ilustra essa ideia, em que o resultado da ACP de 2011 foi rotacionado em 160º.

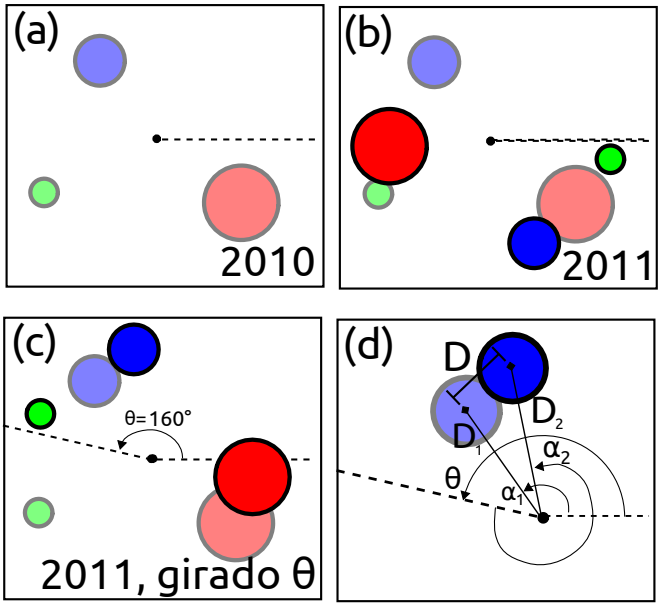

Figura 6: (a) Um resultado hipotético da análise de componentes principais para o ano de 2010 envolvendo três partidos. (b) Superposto ao primeiro, um resultado para o ano seguinte. (c) O mesmo resultado anterior, porém rotacionado de 160° para minimizar a movimentação dos partidos na animação entre um ano e outro. (d) Detalhe para um dos partidos.

Temos aqui um problema de otimização, em que buscamos o ângulo θ que minimiza uma função objetivo $E(\theta)$ que forneça uma medida das distâncias percorridas pelos parlamentares no plano. A função objetivo escolhida é a

soma das distâncias quadradas percorridas pelos parlamentares. Ao utilizar o quadrado das distâncias, de forma a penalizar mais as distâncias muito grandes. Temos então:

$$E(\theta) = \sum_{k \in \{parlamentares\}} D_k^2 \tag{5}$$

Caso a intenção seja trabalhar o mapa de votações de partidos, recomendamos a ponderação da distância percorrida pelo partido utilizando a quantidade de parlamentares no partido, conforme a Equação 6. Nota-se nesse caso a analogia desta função objetivo com a definição de energia: para um conjunto de bolas k massa M_k que iriam do ponto (p_{k1}) ao ponto (p_{k2}) em um tempo Δt (o tempo da animação), estamos buscando o ângulo θ que minimiza a energia envolvida neste movimento.

$$E(\theta) = \sum_{k \in \{partidos\}} M_k \cdot D_k^2 \tag{6}$$

Em vez da analogia com energia, poderíamos utilizar a analogia com a quantidade de movimento, para isso usando as distâncias não quadráticas. Porém, essa outra abordagem tem mais chance de resultar em partidos "caminhando longas distâncias".

Da definição de $E(\theta)$ para partidos, chegamos à seguinte fórmula para θ[5]

$$\theta = \arctan \frac{\sum_k M_k D_{k1} D_{k2} \sin \alpha_k}{\sum_k M_k D_{k1} D_{k2} \cos \alpha_k} + C\pi \tag{7}$$

onde $C = 0 \,\text{ou}\, 1$, um dos casos correspondendo ao mínimo e o outro ao máximo. Para o mapa de parlamentares, basta considerar $M_k = 1$.

Para encontrar a solução basta calcular $E(\theta)$ para os θ que satisfazem (7) e verificar qual delas corresponde ao mínimo. Se durante o cálculo o denominador do argumento do arco-tangente for nulo, então o mínimo está em $\frac{\pi}{2}$ ou $\frac{3\pi}{2}$.

O mesmo procedimento deve ser repetido espelhando-se um dos eixos (por exemplo multiplicando todas as coordenadas x por -1), já que o sentido dos eixos que resulta da ACP é arbitrário. Assim o problema se resume à avaliação de $E(\theta)$ em 4 casos (espelhado e não-espelhado, cada um com dois valores de θ), e escolha do caso que resultar no mínimo.

O resultado desse algoritmo pode ser exemplificado pela Figura 5, com a ressalva de que o algoritmo de rotação é aplicado ao mapa de parlamnetares e

[5]Para uma demonstração completa, ver `https://github.com/radar-parlamentar/radar/raw/master/doc/algoritmo_rotacao.pdf`.

que os partidos visualizados são meramente os centroides dos parlamentares que os compõe.

%%%%%%%

Questões de escala entre períodos...? Efeito de extremistas na visualização temporal: um extremista de um período afeta a visualização de outros períodos (deixando parlamentares muito concentrados). Procede ???

Leo ►*Comparar em linhas gerais nosso método de análise temporal com o D-NOMINATE (dynamic nominate).*◄

Heckman encontra uma alta correção entre as coordenadas de um legislador ao longo do tempo (e não são apenas para as duas primeiras coordenadas... chega-se a até 6 ou 8 fatores com correlação ao longo do tempo). Mas duvido que no caso brasileiro essas correlações seriam altas.

Questões de escala

Leo ►*Falar do recuro de zoom do radar*◄

4 Medidas de Avaliação do Modelo

As principais medidas de avaliação de modelos de construção de mapas espaciais de votações são a taxa de classificação correta e a APRE. Conforme já foi explicado em detalhes na Seção 2, a taxa de classificação correta fornece a proporção de votos corretamente previstos pelo modelo. Já a APRE mede a redução no erro que o modelo proposto traz em relação ao modelo ingênuo que prevê que todo parlamentar vota com a maioria.

No entanto, para a aplicar essas medidas a ACP, precisamos definir um *preditor*. A ACP por si só não define um "valor esperado" para o voto de um dado parlamentar em uma dada votação. Por isso, a definição do preditor utilizado e aqui apresentado é de nossa autoria.

Seja a matriz $\widehat{\mathbf{X}}$, tal que:

$$\widehat{\mathbf{X}} = \mathbf{R}^{t}_{(d)} \cdot \mathbf{\Gamma}_{(d)} + \mathbf{C} \tag{8}$$

$\widehat{\mathbf{X}}$ possui valores em $\mathbb{R}$ que se aproximam dos valores discretos da matriz de votos original $\mathbf{X}$.

Para $\widehat{x}_{ij} > 0$ o preditor prevê que o parlamentar j vota SIM na votação i; para $\widehat{x}_{ij} < 0$ o modelo prevê voto NÃO; e para $\widehat{x}_{ij} = 0$ o modelo prevê um voto arbitrário (para facilitar a reprodutibilidade dos resultados foi adotado SIM nestes casos). Este modelo prevê apenas votos SIM ou NÃO, ou seja, não prevê a possibilidade de abstenções, obstruções ou ausências.

É claro que a definição de tal preditor depende de limiares arbitrários que mapeiam os valores reais de $\widehat{x}_{ij}$ para os valores discretos que representam opções em uma votação (SIM ou NÃO). Então um problema dessa avaliação é que a escolha de outros limiares ou outros mapeamentos possíveis poderiam produzir resultados bem diferentes. Um exemplo seria associar uma faixa de valores de $\widehat{x}_{ij}$ à abstenção.

Por isso apresentaremos também como medida de avaliação do modelo a porcentagem da variância retida pelas dimensões da ACP. Quanto maior é a variância retida em uma dimensão, mais essa dimensão é explicativa do fenômeno estudado. Uma desvantagem dessa medida é não ser diretamente comparável com outros métodos.

Leo ►*Saulo, talvez você queria explicar melhor o que é a proporção da variância retida pelas componentes principais.*◄

5 Resultados

Nesta seção apresentaremos os resultados que a ACP fornece quando aplicada a casas legislativas brasileiras. Os dados utilizados são os dados abertos provenientes das respectivas casas legislativas. Realizamos análises sobre votações das casas que, ao nosso conhecimento, disponibilizam abertamente os dados de votações nominais. São essas casas: Câmara dos Deputados, Senado e Câmara Municipal de São Paulo[6].

Seguindo o mesmo procedimento adotado por Izumi [12] e Leoni [5], adotamos a legislatura completa como unidade de análise. Assim, para a Câmara dos Deputados e para o Senado escolhemos as legislaturas 2007-2010 e 2011-2014. Dessa forma iremos mostrar o uso da ACP em diferentes casas e em diferentes momentos do tempo. Para a Câmara Municipal de São Paulo, devido a restrição de dados fornecidos, escolhemos a legislatura 2013-2016.

A relação de partidos presentes nos mapas espaciais produzidos é exibido na Figura 7. A grande quantidade de partidos no Brasil causa uma dificuldade para diferenciá-los no mapa. Por isso, além de cores utilizamos também formas geométricas para ajudar a distinguir os partidos. Embora algumas cores tenham relação com as cores oficial dos respectivos partidos, isso não é um padrão aplicável a todos os partidos, justamente pela grande quantidade de partidos.

O recurso de utilizar formas geométricas para distinguir partidos foi utilizado neste artigo, mas não no Radar Parlamentar. Como numa aplicação web dispomos de recursos interativos, optamos no Radar Parlamentar por

[6]O banco de dados completo com as votações sobre as quais trabalhamos pode ser acessado em `http://radarparlamentar.polignu.org/dados/downloads/`.

Figura 7: Legenda com as cores dos partidos presentes nos gráficos ACP.

destacar membros de um partido definido pelo usuário quando o usuário interage com o nome do partido.

Os mapas espaciais de votações produzidos pela ACP são exibidos nas Figuras 8, 9, 10, 11 e 12.

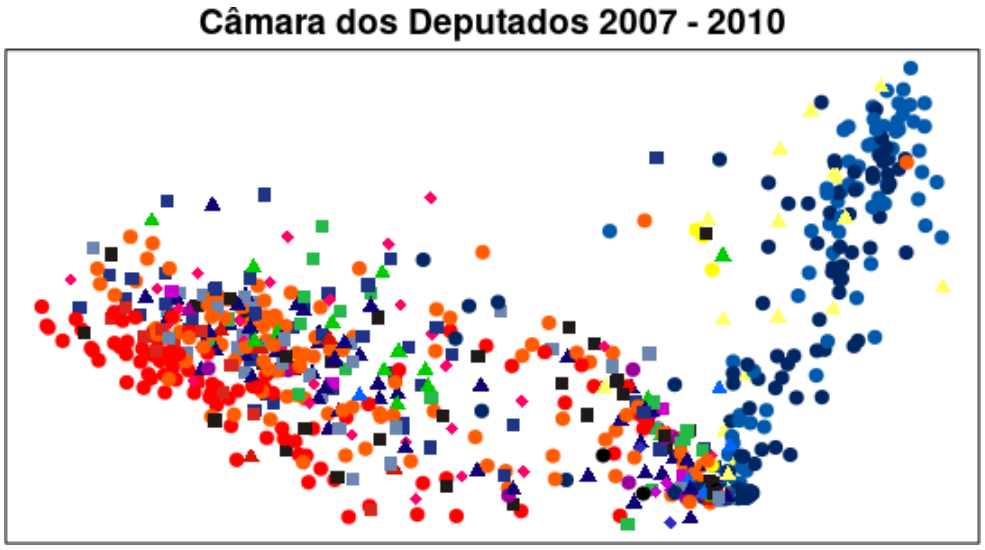

Figura 8: Mapa espacial de votações gerado com a ACP para votações da Câmara dos Deputados entre 2007 e 2010.

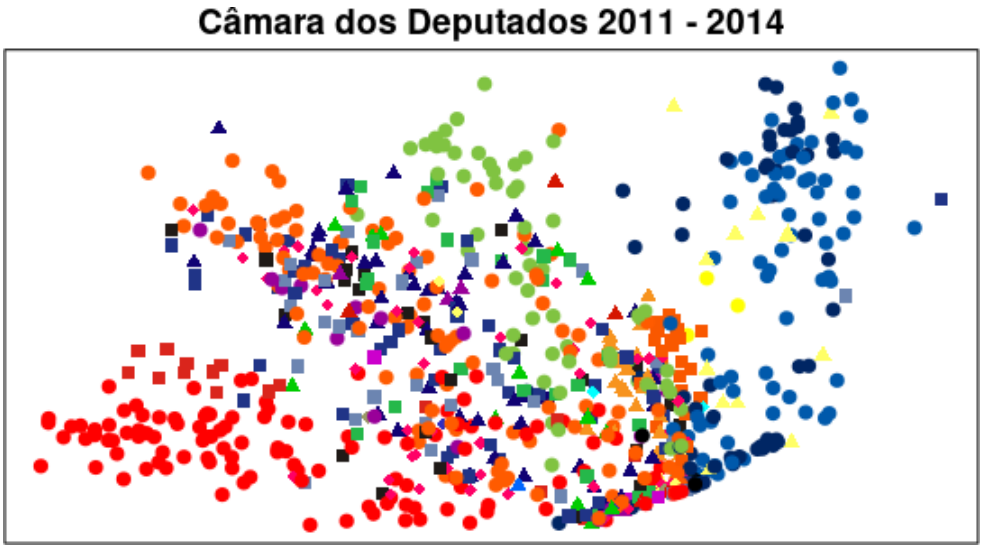

Figura 9: Mapa espacial de votações gerado com a ACP para votações da Câmara dos Deputados entre 2011 e 2014.

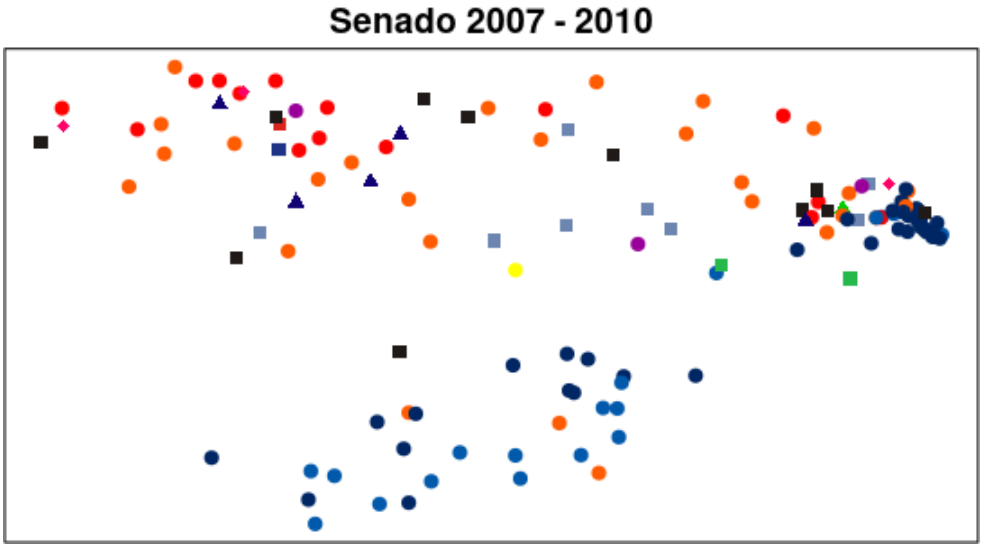

Figura 10: Mapa espacial de votações gerado com a ACP para votações do Senado entre 2007 e 2010.

ToDo ►*Atualizar dados cmsp. O recesso parlamentar da cmsp se inicia em 15 de dezembro. Então talvez valha a pena esperar essa data, assim garantimos que temos os dados da legislatura completa.*◄

As medidas de adequação produzidas pela ACP são exibias na Tabela 1. Nessa mesma tabela pode-se comparar os resultados da ACP com os resul-

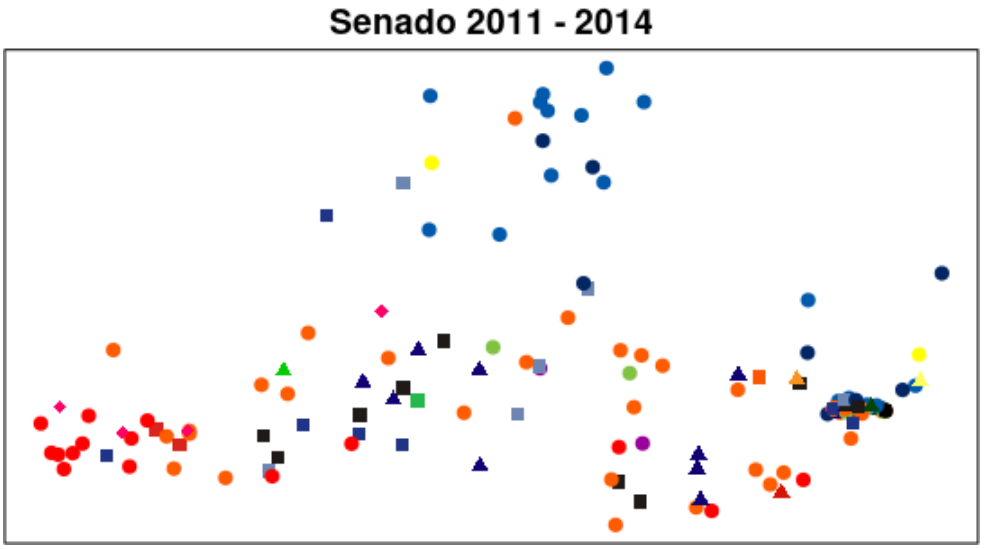

Figura 11: Mapa espacial de votações gerado com a ACP para votações do Senado entre 2011 e 2014.

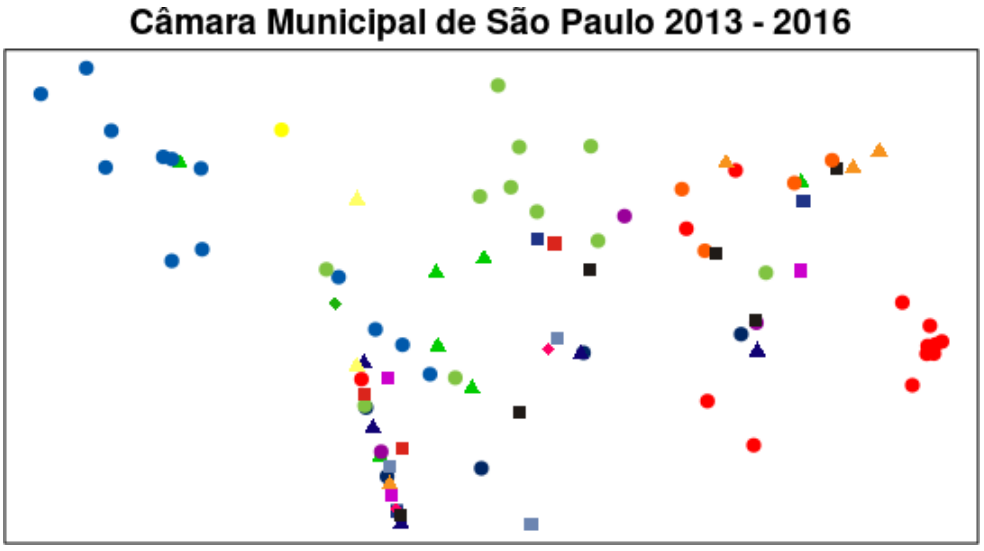

Figura 12: Mapa espacial de votações gerado com a ACP para votações da Câmara Municipal de São Paulo entre 2013 e 2016.

tados produzidos pelo W-NOMINATE. Já a Tabela 2 mostra os tempos de execução para ambos os métodos. Por fim, na Tabela 3 apresentamos a porcentagem de variância retida por cada uma das dimensões principais da ACP. Nessas tabelas utilizamos as seguintes abreviações: *cdep* para Câmara dos Deputados, *sen* para Senado e *cmsp* para Câmara Municipal de São Paulo.

Análise	*Classificação Correta (%)*				*APRE (%)*			
	ACP		*W-NOMINATE*		*ACP*		*W-NOMINATE*	
	1D	*2D*	*1D*	*2D*	*1D*	*2D*	*1D*	*2D*
cdep2007-2010	93	93	91	92	45	49	55	62
cdep2011-2014	88	90	87	89	27	37	42	52
sen2007-2010	89	96	92	93	24	62	67	74
sen2011-2014	88	93	87	90	25	40	46	56
cmsp2013-2016	94	96	96	97	58	70	77	83

Tabela 1: Medidas de adequação produzidas pela ACP e pelo W-NOMINATE nas casas legislativas e períodos analisados.

Análise	*ACP*	*W-NOMINATE*
cdep2007-2010	1,26 ± 0,02	70,21 ± 1,67
cdep2011-2014	0,67 ± 0,03	54,18 ± 1,87
sen2007-2010	0,03 ± 0,01	2,22 ± 0,08
sen2011-2014	0,03 ± 0,00	3,27 ± 0,09
cmsp2013-2016	0,02 ± 0,00	10,35 ± 0,88

Tabela 2: Tempos de execução, apresentando médias e desvios padrão em segundos, calculados com base em 10 execuções.

Análise	*1D*	*2D*	*3D*	*4D*	*5D*	*1D + 2D*
cdep2007-2010	33	6	3	2	2	39
cdep2011-2014	18	8	5	3	3	26
sen2007-2010	27	17	6	4	3	44
sen2011-2014	24	13	7	3	3	37
cmsp2013-2016	41	8	6	4	3	49

Tabela 3: Porcentagem da variância retida pelas dimensões da ACP.

Este artigo não objetiva realizar uma comparação direta entre a ACP e W-NOMINATE. Os valores produzidos pelo W-NOMINATE foram colocados apenas para que o leitor tenha alguma referência para avaliar os resultados que encontramos para a ACP. Dessa forma, o W-NOMINATE foi escolhido meramente por sua proeminência entre os métodos de produção de mapas espaciais de votações, mesmo critério utilizado por Leoni [5].

Para a execução do W-NOMINATE utilizamos as opções padrões de minvotes=20 e lop=0.025. Isso significa que o algoritmo descarta parlamentares ausentes em pelo menos 20 votações e descarta as votações nas quais menos de 2,5% dos deputados votou com a minoria[7]. Já na execução da ACP não descartamos nenhum voto para a análise. Uma discussão sobre a escolha desses valores é realizada na próxima seção.

As análises apresentadas nesta seção foram feitas no software de estatística **R**, e o código que produziu esses resultados está disponível em `https://github.com/radar-parlamentar/pesquisa/tree/master/R`. Para os cálculos envolvendo o W-NOMINATE, utilizamos o pacote wnominate[8] disponibilizado pelo próprio Keith Poole. Já para o cálculo da ACP, utilizamos a função prcomp, nativamente disponível no R.

Esses scripts foram executados em um laptop Dell Vostro 5480, com 8GB de memória RAM e com a 5ª geração do Processador Intel® CoreTM i7-5500U. O sistema operacional utilizado foi o ArchLinux com kernel Linux versão "4.7.2-1-ARCH #1 SMP PREEMPT Sat Aug 20 23:02:56 CEST 2016" para arquitetura x86_64.

O registro do hardware e sistema operacional é importante, pois para um mesmo conjunto de dados, mapas diferentes podem ser produzidos pela ACP em computadores diferentes, principalmente no que diz respeito à orientação do mapa. Essas diferenças acontecem principalmente devido a diferenças no tratamento dado a números de ponto flutuante em arquiteturas diferentes.

6 Discussão

Heckman e Snyder [14] encontram padrões políticos nos mapas espaciais por eles produzidos que são compatíveis com análises já feitas por cientistas políticos utilizando outros métodos. Segundo eles, isso ilustra a "razoabilidade" dos mapas produzidos pelo modelo proposto.

De forma similar, a "razoabilidade" de nossos resultados pode ser defendida por certos padrões políticos identificados, sendo o principal a polarização

[7] "lop" é abreviação de *lopsided*.

[8] `https://cran.r-project.org/web/packages/wnominate/`

minvotes	lop	*Classificação Correta (%)*		*APRE (%)*	
		1D	*2D*	*1D*	*2D*
0	-1	88	90	27	37
0	0	87	89	28	40
20	0.025	85	88	33	45

Tabela 4: Resultados obtidos utilizando-se a ACP para a legislatura 2011-2014 da Câmara dos Deputados. lop=0 indica que apenas as votações unânimes foram descartadas, enquanto que lop=-1 não descarta nenhuma votação.

PT-PSDB, presente nos cinco mapas. Da mesma forma, o considerável espalhamento do PMDB nas casas federais poderia ser esperado.

Mas ao mesmo tempo em que os mapas confirmam padrões já esperados, também evidenciam outros comportamentos interessantes, mas não tão óbvios. Um deles é o relativo afastamento que se pode perceber do PMDB em relação ao PT na Câmara dos Deputados, quando se compara as legislaturas 2007-2010 e 2011-2014. Isso parece ilustrar um processo de enfraquecimento da base parlamentar do governo do PT ao longo do tempo, o que parece fazer sentido considerando o impeachment de Dilma Rousseff.

Pelos valores apresentados na Tabela 1, podemos observar a tendência de que a classificação correta na ACP seja ligeiramente melhor que a mesma taxa no W-NOMINATE, enquanto que a APRE seja melhor no W-NOMINATE. Contudo, embora os algoritmos tenham importância nos resultados obtidos, observamos que esses indicadores também são influenciados pelos parâmetros minvotes e lop. Pela Tabela 4, que mostra resultados para a legislatura 2011-2014 da Câmara dos Deputados utilizando a ACP, observamos que quanto mais votações excluímos da análise, seja por minvotes ou lop, mais a taxa de classificação correta piora e mais a APRE melhora.

Uma justificativa parcial para as escolhas de minvotes e lop são: 1) Aparentemente o W-NOMINATE não converge com minvotes=0 e lop=0.0. Em nossas execuções, meia-hora não foi suficiente para as análises das legislaturas do senado (sen2007-2010 e sen2011-2014). 2) Utilizar minvotes=20 e lop=0.025 para a ACP resultou em mapas espaciais sem significado político aparente, causando a impressão de um mapa construído aleatoriamente, conforme pode-se verificar na Figura 13. Nesse mapa, nem mesmo a polarização PT-PSDB é possível de se identificar.

Para o mapa apresentado na Figura 13, os parâmetros utilizados resultam em 87 de 808 (10%) legisladores descartados e 115 de 382 (30%) votações excluídas. Esses valores provavelmente são bem maiores do que cientistas políticos poderiam esperar em análises sobre o congresso americano, conside-

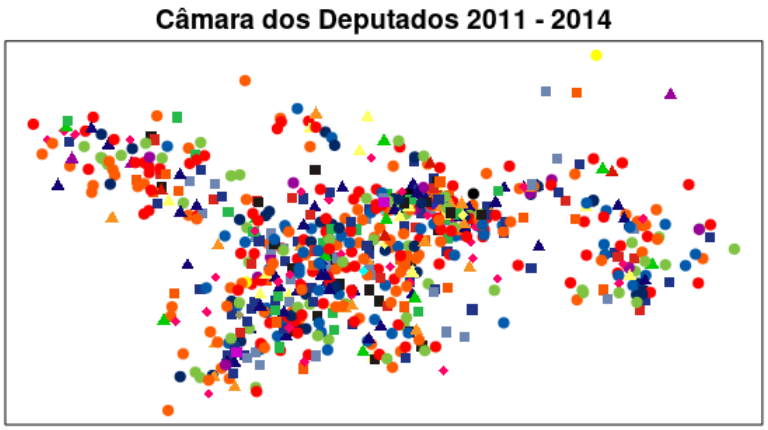

Figura 13: Mapa espacial de votações produzido para a Câmara dos Deputados na legislatura 2011-2014 utilizando a ACP com parâmetros minvotes=20 e lop=0.025. Não há aparente significado político nesse mapa.

rando que, utilizando também lop=0.025, Poole descartou 10% das votações da 85ª legislatura do congresso americano [2]. Além disso, segundo um levantamento realizado por Roller e Stamm [16], em média um senador americano falta a 2,5% das votações, enquanto que apenas dois senadores faltaram mais do que 10% das vezes, sendo um desses devido a um derrame.

Sobre os tempos de execução, mostrados na Tabela 2, confirma-se a tese defendida por Heckman e Snyder [14]: a de que modelos lineares são bem mais eficientes. Em nosso caso, encontramos que a ACP foi até 517 vezes mais rápida que o W-NOMINATE, o que ocorreu para a Câmara Municipal de São Paulo. É interessante também notar na Tabela 2 que os desvios padrão são relativamente pequenos, tanto para a ACP quanto para o W-NOMINATE, o que evidencia uma boa previsibilidade quanto ao tempo de execução para ambos os algoritmos.

Leo ►*Aqui talvez o Saulo queira encerrar com mais algumas palavras sobre a ACP, como sobre a simplicidade e transparência da ACP e a fácil interpretação do mapa produzido.*◄

Análise intertemporal

Observa-se pelas Figuras 8 e 9 que os dois mapas possuem a mesma orientação, ou seja, a polarização entre os principais partidos se mantém na mesma distribuição espacial: PT na região inferior esquerda e PSDB na região direita.

No entanto, a ACP não garante essa orientação entre mapas de períodos diferentes. Pode-se observar nas Figuras 10 e 11 que uma melhor comparação entre os mapas é possível ao se espelhar verticalmente a segunda figura, como podemos ver na Figura 14. É esse processo de espelhamento, ou rotação, que ocorre quando da aplicação do algoritmo de análise intertemporal apresentado na Seção 3.

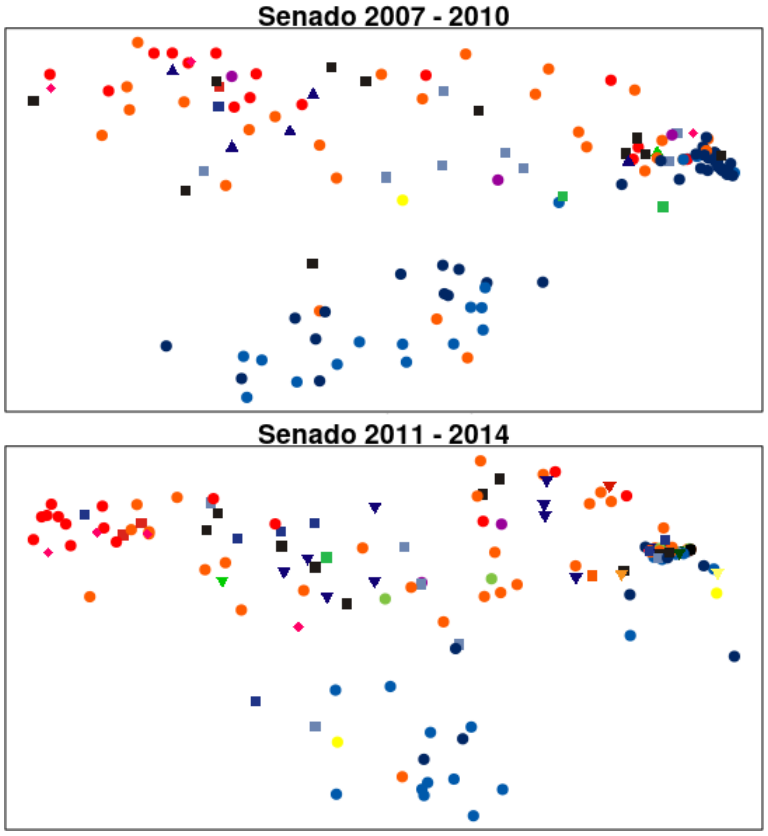

Figura 14: Mapas espaciais de votações gerados com a ACP para duas legislaturas do Senado, sendo o segundo mapa espelhado verticalmente para se aumentar a semelhança entre os dois mapas e assim facilitar a análise intertemporal.

Análise de dimensionalidade

Pelos números da Tabela 3, observamos que a explicação acrescentada pela segunda dimensão não é desprezível, principalmente nas legislaturas do senado. Também observamos que em alguns casos, a explicação acrescentada

pela terceira dimensão também possui poder explicativo quando comparado com a explicação fornecida pelas outras dimensões, caso no qual se destaca a legislatura cdep2011-2014. Temos ainda que mesmo que a partir da terceira dimensão pouco se acrescente, a utilização das duas primeiras dimensões possui poder explicativo modesto, com valores encontrados indo de 26% a 49%.

Essa análise sobre os números apresentados pela Tabela 3 levam a um entendimento de que o legislativo brasileiro opera em uma realidade multidimensional. Ou pelo menos de que idealmente os mapas espaciais de votações deveriam ser construídos com mais dimensões para que diferenças significativas entre os parlamentares fossem visualizadas.

Leoni [5] e Izumi [12] não chegam a mesma conclusão. Diferentemente, eles concluem pela unidimensionalidade do espaço legislativo da Câmara dos Deputados e do Senado. Essa diferença ocorre pois eles utilizam a taxa de classificação correta e a APRE como critérios para determinar a dimensionalidade do espaço legislativo.

ToDo ►*Tentar mostrar um mapa em 3 dimensões??!!!*◄

7 Conclusões

resultados coisas pra melhorar / investigar ex: modelagem de partidos q não votam em alguma votação; suplentes; troca de parlamentares análise de sensibilidade

Referências

[1] A. Downs, "An economic theory of political action in a democracy," *The Journal of Political Economy*, vol. 65, no. 2, pp. 135–150, 1957.

[2] K. T. Poole and H. Rosenthal, "A spatial model for legislative roll call analysis," *American Journal of Political Science*, vol. 29(2), pp. 357–384, 1985.

[3] K. T. Poole and H. Rosenthal, *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press, USA, Nov. 2000.

[4] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 02, pp. 355–370, 2004.

[5] E. Leoni, "Ideologia, democracia e comportamento parlamentar: a Câmara dos Deputados (1991-1998)," *Dados*, vol. 45, pp. 361 – 386, 2002.

[6] M. Kantardzic, "Section 3.4 Principal Componente Analysis," in *Data Mining, Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, INC., 2003.

[7] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[8] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine (6)*, vol. 23, pp. 559–572, 1901.

[9] H. Hotelling, "Analysis of a complex of statistical variables into principal components," 1933.

[10] O. A. Davis, M. J. Hinich, and P. C. Ordeshook, "An expository development of a mathematical model of the electoral process," *The American Political Science Review*, vol. 64, no. 2, pp. 426–448, 1970.

[11] K. T. Poole, *Spatial models of parliamentary voting*. Cambridge University Press, 2005.

[12] M. Y. Izumi, "Governo e oposição no senado brasileiro," *DADOS – Revista de Ciências Sociais*, vol. 59, pp. 91–138, 2016.

[13] K. T. Poole, "Nonparametric unfolding of binary choice data," *Political Analysis*, vol. 8, no. 3, pp. 211–237, 2000.

[14] J. J. Heckman and J. M. Snyder, "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators," *The RAND Journal of Economics*, vol. 28, pp. S142–S189, 1997.

[15] N. McCarty, "Measuring legislative preferences," in *The Oxford handbook of the American Congress*, pp. 66–94, Oxford University Press Oxford, 2011.

[16] E. Roller and S. Stamm, "The best and worst attendance records in the senate," 2014. http://www.theatlantic.com/politics/archive/2014/05/the-best-and-worst-attendance-records-in-the-senate/455949/.