



SPG 31 · METODOLOGIA E MÉTODOS DE PESQUISA EM CIÊNCIAS SOCIAIS

Quantos votos meu partido vai fazer?

Um tutorial de validação preditiva com dados abertos (2016-2024)

Felipe Lamarca

IESP-UERJ

Tomás Paixão Borges

IESP-UERJ

30 de setembro de 2026

Duas perguntas

SUBSTANTIVA

É possível prever, com boa aproximação, o desempenho dos partidos nas eleições para vereador usando apenas informação disponível antes do pleito?

METODOLÓGICA

Como se conduz, na prática, um exercício preditivo rigoroso, dos dados brutos à validação dos resultados?

Respondemos com um tutorial reprodutível: os 92 municípios fluminenses em três eleições para vereador (2016, 2020 e 2024).

Descrever, explicar, prever

DESCREVER

Resumir padrões e regularidades, sem pretensão causal ou preditiva.

EXPLICAR

Dizer por que uma variável afeta outra, com um modelo teórico e uma forma funcional interpretável.

PREVER

Antecipar casos ainda não observados. Pesa a informação disponível antes do evento.

Explicar e prever pedem procedimentos diferentes ([Shmueli, 2010](#)), mas se complementam ([Mullainathan; Spiess, 2017](#); [Hofman *et al.*, 2021](#); [Verhagen, 2022](#)).

O protocolo é o mesmo qualquer que seja o algoritmo: definir o alvo, separar treino e teste sem contaminação, comparar contra um benchmark e medir fora da amostra. Por isso o estimador aqui é o mais simples, a regressão linear.

Um tutorial em oito passos

PARTE I · PREPARAR

1. O desfecho e a base
2. Treino, teste e validação
3. Vazamento de dados
4. Associações iniciais

PARTE II · MODELAR

5. Benchmarks e modelos
6. Ajuste

PARTE III · AVALIAR

7. Desempenho contra os benchmarks
8. Onde o modelo falha

Uma tarefa preditiva se define por quatro elementos: o alvo, o conjunto de informação (o que se sabe e quando), a função de perda e o critério de sucesso.

Partido × município × ano

4.785

observações: 2.107 em 2016, 1.360 em 2020 e 1.318 em 2024

92

municípios do estado do Rio de Janeiro

9

variáveis, todas do TSE via Base dos Dados

- **Desfecho:** proporção de votos do partido na eleição para vereador.
- **2012** entra só como passado: para prever 2016 usamos 2012; para 2020, 2016; para 2024, 2020.
- **De-para de partidos:** fusões e mudanças de nome preservam o histórico. O desempenho anterior do UNIÃO em 2024 é a soma de DEM e PSL em 2020.

Separar pelo tempo

ALEATÓRIO

Sortear linhas (70/30, por convenção).

POR MUNICÍPIO

Sortear municípios inteiros: o modelo generaliza para lugares que nunca viu?

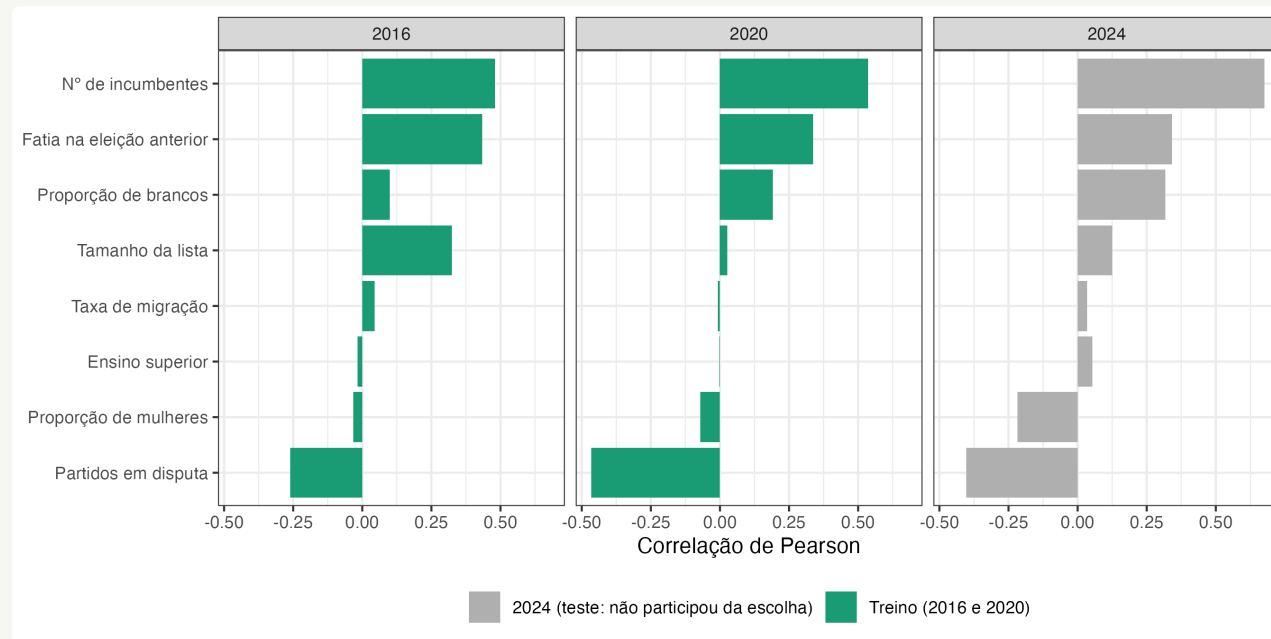
FORA DO TEMPO · ADOTADO

Treinar no passado e prever a eleição seguinte: a situação real de uma previsão.

- **Treino:** 2016 e 2020 (3.467 linhas). **Teste:** 2024 (1.318 linhas), não observadas no ajuste.
- **Validação:** escolhas de forma funcional com treino em 2016 e validação em 2020.
- O erro é condicional ao partido lançar lista: 26,6% das combinações família-município de 2020 não disputam em 2024.

O que se sabe em agosto

- **Financiamento de campanha fica de fora:** a prestação de contas completa só sai depois da eleição.
- **Vazamento por seleção:** escolher preditores olhando o teste. Por isso 2024 aparece em cinza.
- **Vazamento por pré-processamento:** padronizar ou imputar na base inteira antes de separar.



Correlação de Pearson entre a proporção de votos e cada preditor, por ano.

Benchmarks antes de modelos

Benchmarks

- Média do desfecho no treino
- **m0**: repete a votação anterior
- Média histórica do partido
- Proporção de candidatos do partido no município

Modelos (regressão linear)

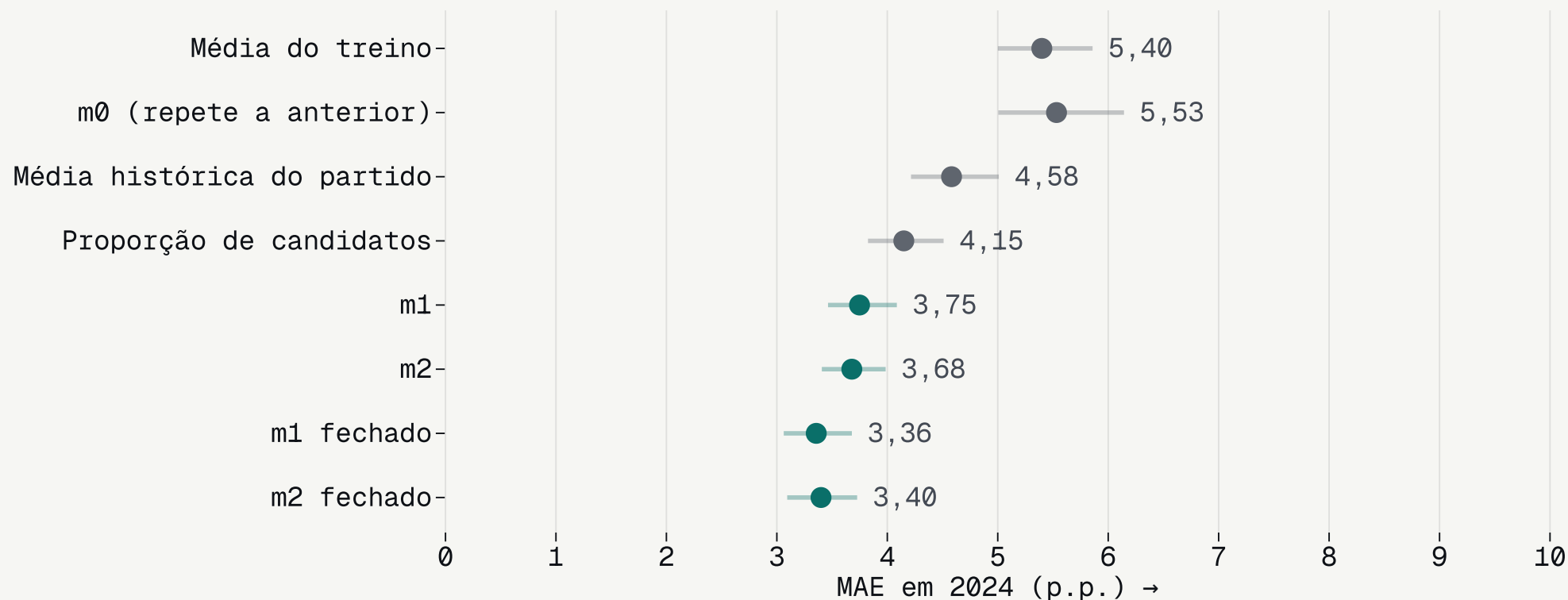
- **m1**: votação anterior + nº de incumbentes na lista
- **m2**: m1 + tamanho da lista, partidos em disputa, proporção de brancos e de mulheres

Métricas em pontos percentuais: **MAE** (erro típico) e **RMSE** (penaliza erros grandes). O **R² fora da amostra** compara o modelo com a média do ano previsto e pode ser negativo.

Os modelos superam os benchmarks

métrica (p.p.)

● MAE ○ RMSE



Barras: intervalo de 95% por bootstrap dos 92 municípios. Fechado: as previsões de cada município são divididas pela soma, para somar 100%. O m2 fechado tem MAE 3,40, RMSE 4,51 e R^2 fora da amostra 0,58.

Mais variáveis, pouco ganho fora da amostra

Erro dentro e fora do treino (p.p.)

Modelo	Preditores	MAE no treino	MAE em 2024
m1	2	3,38	3,75
m2	6	2,94	3,68

O m2 ajusta o passado bem melhor, mas quase não melhora o futuro.

As duas transições fora do tempo

m2 fechado	MAE	RMSE	R²
2016 → 2020	3,39	4,50	0,573
2016 + 2020 → 2024	3,40	4,51	0,584

A mesma receita entrega o mesmo desempenho nas duas janelas.

Para a pergunta do título

40 de 92

municípios com o partido mais votado certo (m0: 13)

0,73

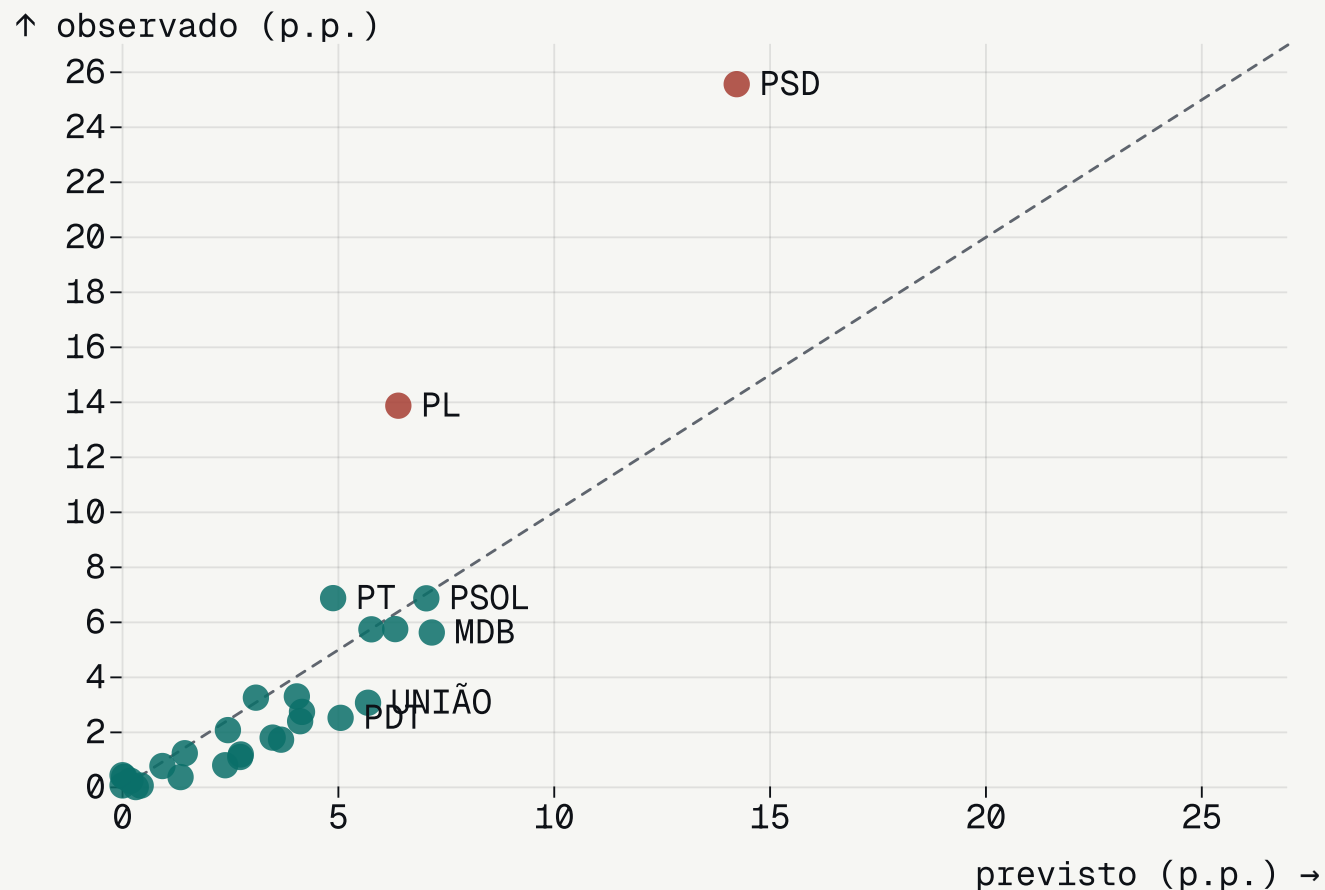
correlação de Spearman média dentro do município (m0: 0,33)

78,5%

das previsões a menos de 5 p.p. do observado

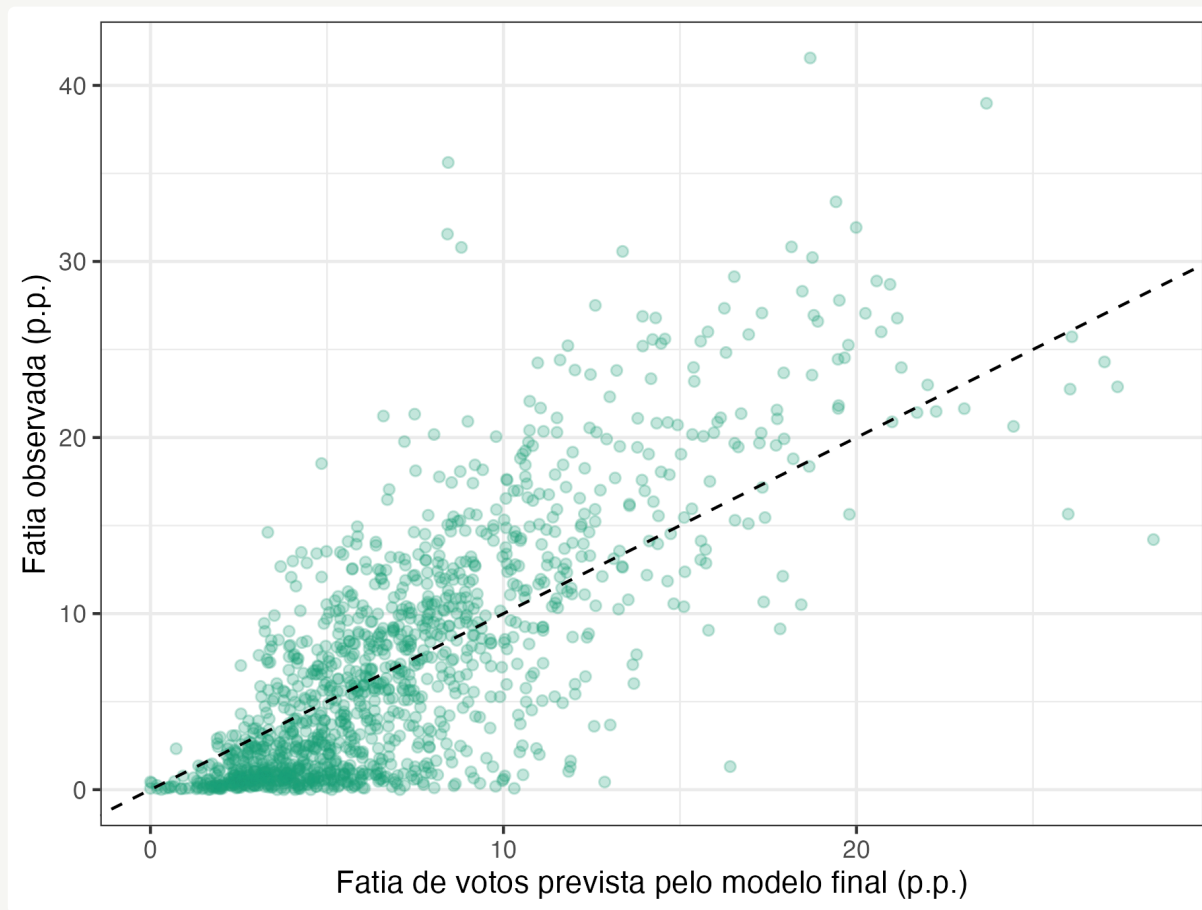
O modelo acerta bem a ordem e o nível dos partidos médios, e deve ser usado com desconfiança para os maiores.

Rio de Janeiro, 2024



- 28 partidos disputaram; as previsões somam 100%.
- **PSD** subestimado em 11,3 p.p. e **PL** em 7,5 p.p.
- REPUBLICANOS, PSOL, PODE, PP e SOLIDARIEDADE ficam a menos de um ponto.

Calibração em 2024



A nuvem acompanha a diagonal até uns 10 p.p. e se achata acima disso.
O modelo comprime as previsões em direção à média.

Proporção prevista contra proporção observada. A diagonal tracejada é a previsão perfeita.

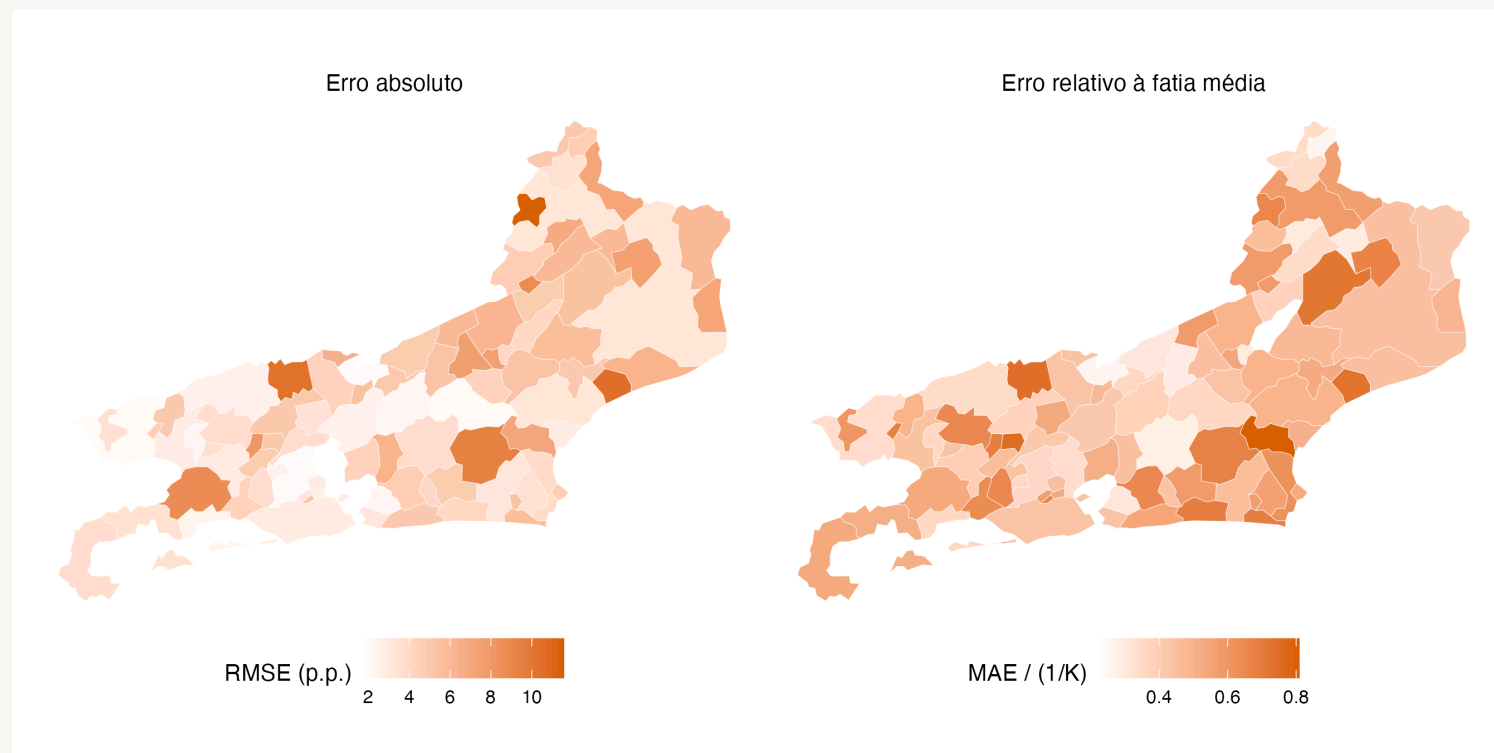
2024 é território conhecido?



Observações de 2024 no plano das duas primeiras componentes principais do treino, com o contorno convexo.

- **99,7%** de 2024 cai dentro do contorno convexo.
- Benchmark sem deslocamento: metade retida do treino, **99,34%**; 2024, **99,33%**.
- Mas o nº de partidos em disputa cai **0,80 DP** entre treino e teste: o contorno não vê o centro.
- Fora ficam listas de MDB, PRD e SOLIDARIEDADE em câmaras de 5 a 8 partidos ([King; Zeng, 2006](#)).

Onde o erro se concentra



Erro por município em 2024: absoluto (RMSE, à esquerda) e relativo à proporção média do município (à direita).

Em termos relativos, o modelo é pior nas cidades grandes e competitivas. Por partido, 22 de 29 têm resíduo mediano negativo, e o resíduo acompanha o tamanho do partido (correlação de 0,92).

Um passo além: modelos mais flexíveis

Modelo (previsões fechadas)	MAE 2020	MAE 2024	RMSE 2024	R ² 2024
m2 fechado	3,39	3,40	4,51	0,584
m2 + efeito de partido	3,38	3,30	4,39	0,606
m2 + efeito de partido e de município	3,35	3,14	4,23	0,634
<i>Random forest</i>	3,20	2,95	4,10	0,657

- **Multinível** ([Gelman; Hill, 2006](#)): ganho real e modesto, com efeitos interpretáveis (PP +1,80; MDB +1,72; PT -1,32 p.p.).
- **Floresta aleatória**: menor erro, mas sem coeficientes; acerta o mais votado em 39 municípios, contra 40 do linear.
- O protocolo não muda quando o estimador muda.

Cinco decisões

- 01 Definir o que se sabe e em que momento
- 02 Separar treino e teste como na situação real de uso
- 03 Montar benchmarks antes dos modelos
- 04 Medir fora da amostra
- 05 Verificar se o resultado sobrevive a outra janela

Sim, dá para prever. Seis preditores conhecidos em agosto e a restrição de soma 1: 3,4 p.p. de erro em 2024, o mesmo na transição de 2016 para 2020.

Uma divisão proporcional ao tamanho das listas, sem parâmetros, erra 4,15 p.p. e supera a média histórica de cada legenda, que custa 36.

Referências

- GELMAN, Andrew; HILL, Jennifer. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. New York, NY: Cambridge University Press, 2006.
- HOFMAN, Jake M. et al. Integrating Explanation and Prediction in Computational Social Science. **Nature**, [s. l.], v. 595, p. 181–188, 2021. DOI [10.1038/s41586-021-03659-0](https://doi.org/10.1038/s41586-021-03659-0).
- KING, Gary; ZENG, Langche. The Dangers of Extreme Counterfactuals. **Political Analysis**, [s. l.], v. 14, n. 2, p. 131–159, 2006. DOI [10.1093/pan/mpj004](https://doi.org/10.1093/pan/mpj004).
- MULLAINATHAN, Sendhil; SPIESS, Jann. Machine Learning: An Applied Econometric Approach. **Journal of Economic Perspectives**, [s. l.], v. 31, n. 2, p. 87–106, 2017. DOI [10.1257/jep.31.2.87](https://doi.org/10.1257/jep.31.2.87).
- SHMUELI, Galit. To Explain or to Predict? **Statistical Science**, [s. l.], v. 25, n. 3, p. 289–310, 2010. DOI [10.1214/10-STS330](https://doi.org/10.1214/10-STS330).
- VERHAGEN, Mark D. A Pragmatist's Guide to Using Prediction in the Social Sciences. **Socius**, [s. l.], v. 8, 2022. DOI [10.1177/23780231221081702](https://doi.org/10.1177/23780231221081702).

Obrigado!

Felipe Lamarca

IESP-UERJ

felipelamarca.com

Tomás Paixão Borges

IESP-UERJ