Análise de métodos de Inferência Ecológica

em dados de redes sociais

Gustavo Penha[12], Thiago N. C. Cardoso [2], Ana Paula Couto da Silva [1], Mirella M. Moro[1]

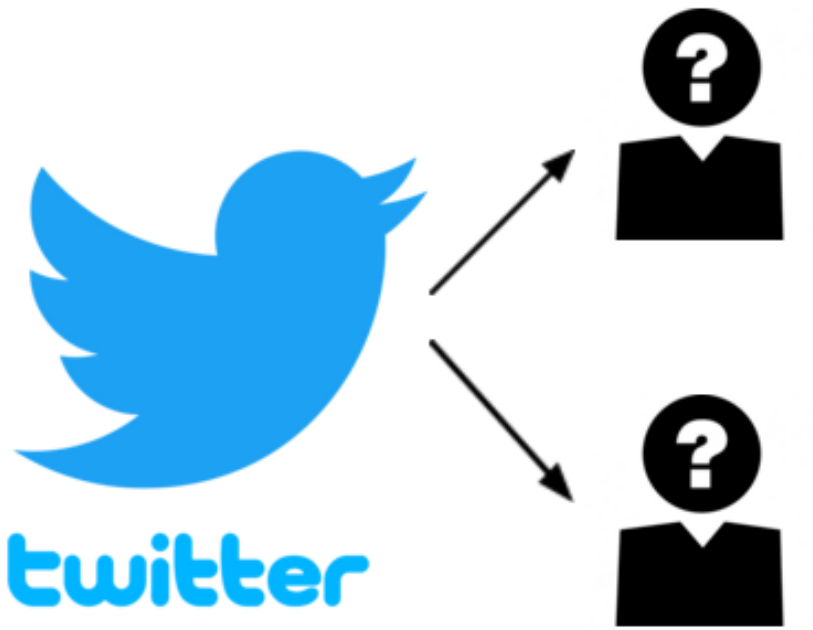



HEKIMA [2]

DCC DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO UFMG [1]

Motivação

- Redes Sociais Online se tornaram extremamente populares e geram um grande volume de dados espontâneo.
- Conhecer os atributos demográficos dos usuários pode ser útil, por exemplo, para o direcionamento de campanhas de marketing.
- Propomos então, a utilização de métodos de **Inferência Ecológica para inferir características demográficas** para grupos de usuários.



Definição do problema

- Dado um conjunto de variáveis **agregadas** observadas para uma população, inferir variáveis **desagregadas** para essa mesma população.
- Dessa forma, os métodos de **Inferência Ecológica** tem como objetivo extrair tais pistas sobre o comportamento individual (desagregados) a partir de informações agregadas.

Tabela: Ilustração do problema de IE no caso 2x2. Tabela de porcentagem de votos para os candidatos Dilma e Aécio e a distribuição de homens e mulheres em uma seção eleitoral *i*.

	Dilma	**Aécio**	
Homem	?	?	52%
Mulher	?	?	48%
	65%	35%	

Questões de Pesquisa

- **RQ1**: Existem padrões na base de dados social que influenciam o resultado dos métodos de Inferência Ecológica?
- **RQ2**: Qual modelo de Inferência Ecológica apresenta os melhores resultados na base de dados social?
- **RQ3**: Os erros dos métodos de Inferência Ecológica na base de dados social são estatisticamente diferentes dos erros de uma base de dados eleitoral?

Configuração de Experimentos

- **Base de dados**: 122 dias de coleta de posts geolocalizados sobre assuntos relacionados à Dilma Rousseff no Twitter utilizando o Zahpee Monitor. Cerca de 150 mil usuários com a sua opinião em relação à Dilma e os seus atributos de gênero e idade. O censo utilizado é o IBGE de 2010.

Tabela: IE para os dados de gênero em uma cidade *i*.

	Favorável à Dilma	**Não favorável a Dilma**	
Homem	?	?	52%
Mulher	?	?	48%
	18%	82%	

Tabela: IE para os dados de idade em uma cidade *i*.

	Favorável à Dilma	**Não favorável a Dilma**	
Menos de 40 anos	?	?	74%
Mais de 40 anos	?	?	26%
	23%	78%	

Configuração de Experimentos

- **Modelos de Inferência Ecológica**
  - (KING,1997)
  - (WAKEFIELD,2004)
  - (IMAI; LU; STRAUSS, 2008)
- **Métricas de avaliação**
  - $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c}_i)^2}$ e $MAE = \frac{1}{N}\sum_{i=1}^{N}|c_i - \bar{c}_i|$
- **Procedimento de avaliação**
  - Otimização dos hiperparâmetros: Projeto fatorial 2^k e *grid-search* nos hiperparâmetros que explicam maior variação do MAE.
  - Comparação dos métodos: Intervalos de confiança e testes pareados.
  - Comparação entre diferentes bases: Testes não pareados.

Análises e Resultados - RQ1

- Existem poucas cidades com muitos usuários que fazem posts geolocalizados com o tema Dilma e muitas cidades com poucos usuários.

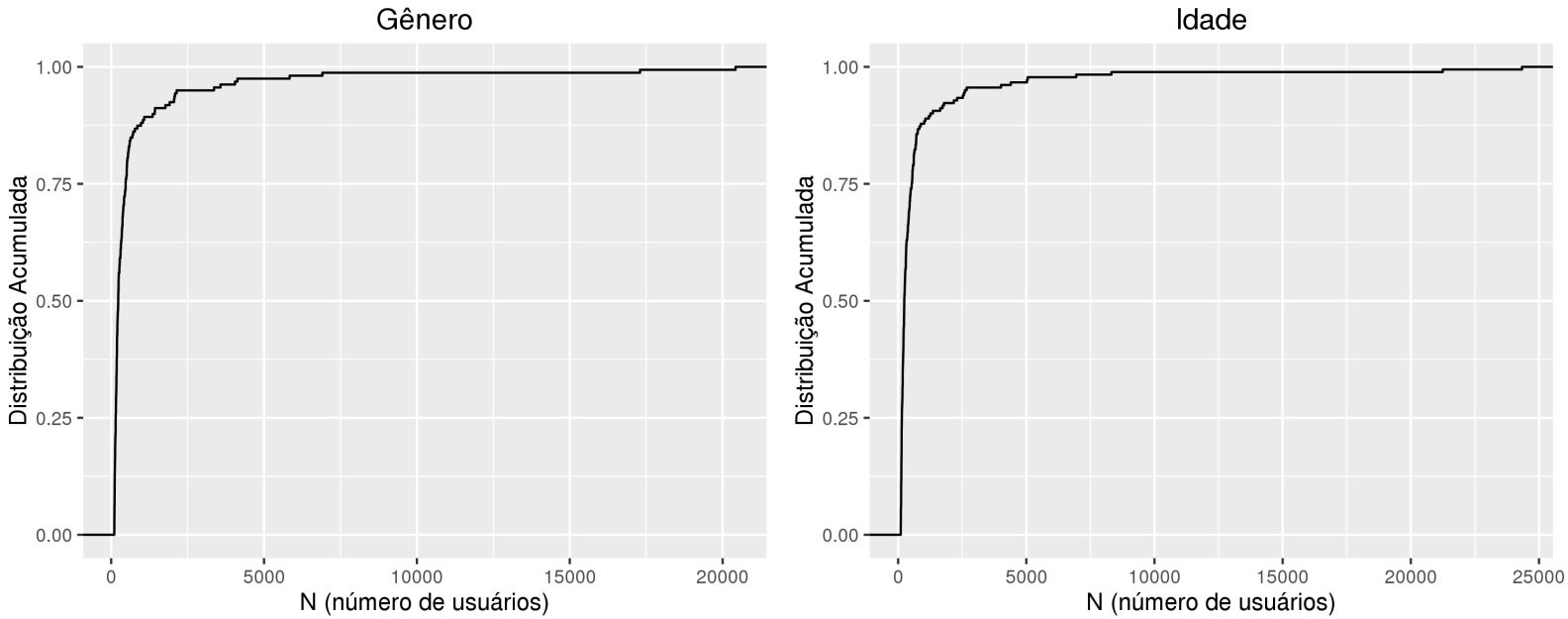


Figura: ECDF do número de usuários por cidade (N_i).

Análises e Resultados - RQ1

- **RQ1:** Essa característica influencia o resultado dos modelos de Inferênfcia Ecológica: cidades com maior número de usuários apresenta erros menores para os métodos.

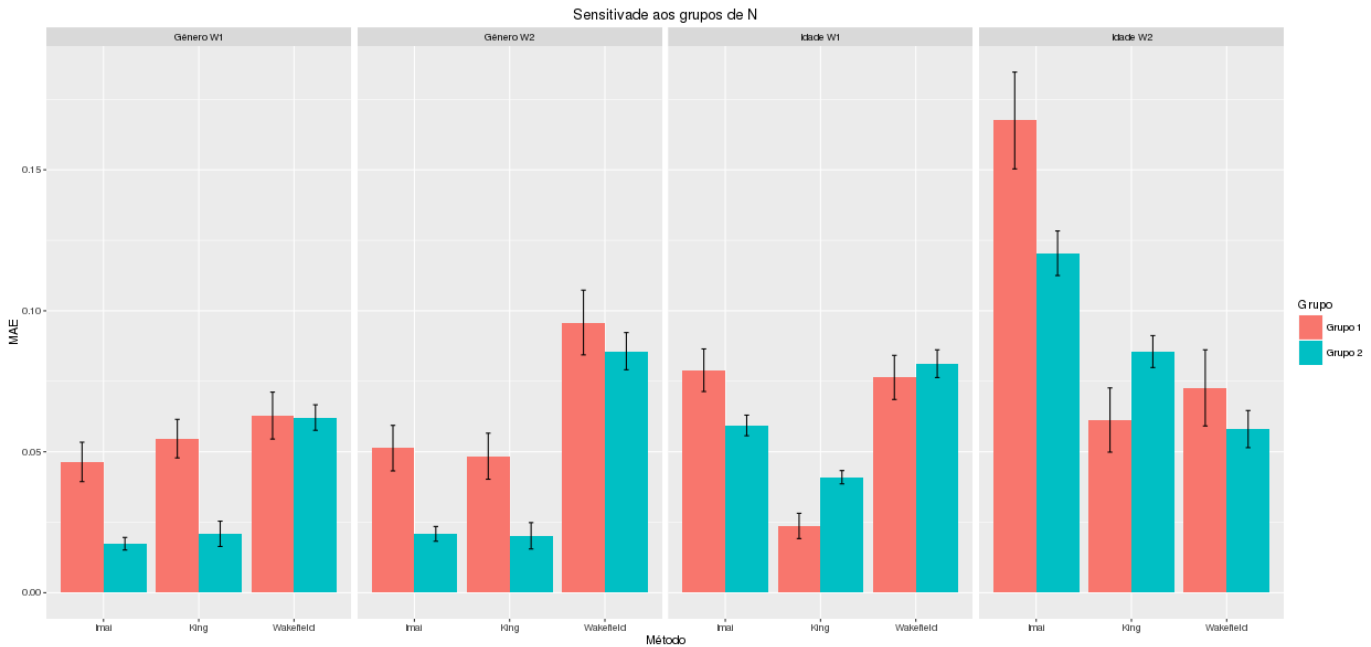


Figura: Gráfico mostra o resultado da métrica MAE dos métodos para os dois grupos $N<200$ (Grupo 1) e $N>200$ (Grupo 2) em ambas bases de dados.

Análises e Resultados - RQ2

- **RQ2:** Após a otimização dos parâmetros, o algoritmo que apresentou os melhores resultados é o de **(KING,1997)**, os *t-tests* pareados também mostraram diferença estatística do método para os outros, empatando em apenas duas configurações.

Tabela: Resultados de RMSE e MAE dos métodos para as duas bases.

Base de gênero				
Modelo	MAE W1 (+-IC)	MAE W2 (+-IC)	RMSE W1 (+-IC)	RMSE W2 (+-IC)
King	**0.0233** +-0.0039	**0.0256** +-0.0042	0.0355 +-0.0030	0.0381 +-0.0025
Imai	0.0391 +-0.0034	0.0425 +-0.0039	0.0347 +-0.0077	0.0404 +-0.0067
Wakefield	0.0553 +-0.0043	0.0982 +-0.0062	0.0683 +-0.0008	0.0991 +-0.0014
Base de idade				
Modelo	MAE W1 (+-IC)	MAE W2 (+-IC)	RMSE W1 (+-IC)	RMSE W2 (+-IC)
King	**0.0293** +-0.0018	0.0688 +-0.0048	0.0319 +-0.0038	0.0763 +-0.0064
Imai	0.0488 +-0.0028	0.1040 +-0.0067	0.0583 +-0.0050	0.1245 +-0.0092
Wakefield	0.0688 +-0.0042	0.0618 +-0.0065	0.0845 +-0.0832	0.0793 +-0.0026

Análises e Resultados - RQ3

- **RQ3:** Os resultados dos métodos nas bases de dados sociais foram comparados com um *benchmark* chamado *reg*, uma base de dados eleitoral. Os erros foram menores na base de dados social, com significância estatística.

Tabela: T-testes não pareados entre as bases de dados sociais e base de dados eleitoral.

	Base social (gênero)		Base eleitoral (*reg*)		t-teste não pareado dos erros entre as bases	
Modelo	MAE W1	MAE W2	MAE W1	MAE W2	p-value W1	p-value W2
King	**0.0233**	**0.0256**	0.3883	0.3442	5.5512e-73	3.8638e-57
Imai	**0.0391**	**0.0425**	0.3917	0.3116	2.363e-64	6.0851e-43
Wakefield	**0.0553**	**0.0982**	0.5399	0.1243	3.4523e-106	0.0067

Análises e Resultados - RQ3

- Entretanto, a base de dados social coletada neste trabalho possui limites mais justos do que a base de dados *reg*, o que torna a tarefa mais fácil para os métodos de Inferência Ecológica.

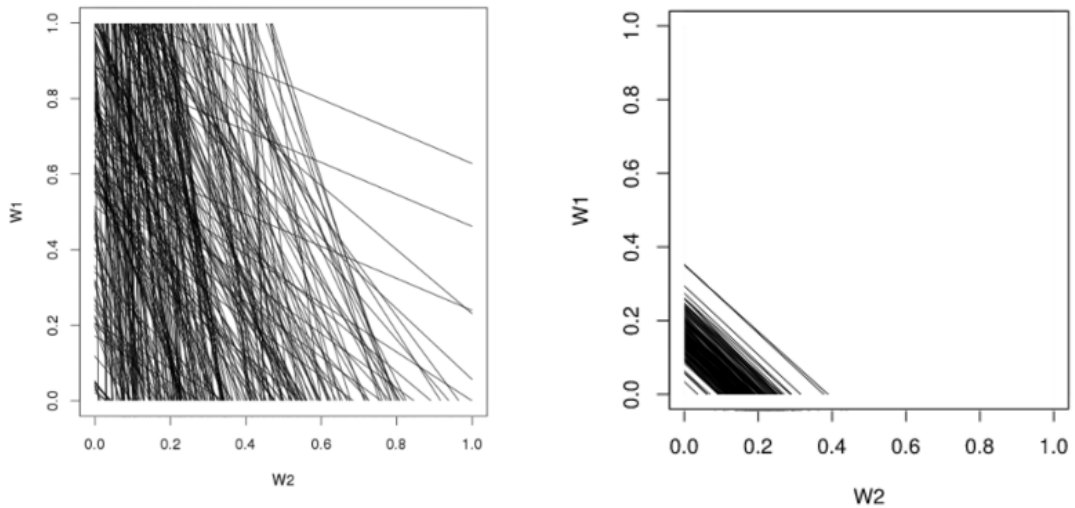


Figura: Gráfico de tomografia para a base de dados *reg* (esquerda) e a base de dados social (direita).

Conclusões

- Mostramos que é possível utilizar métodos de Inferência Ecológica para **estimar gênero e idade** de grupos de usuários de redes sociais baseando-se em dados de um censo (IBGE) e dados agregados (ex: % sentimento positivo ou negativo em relação à um tema), com **erros médios para as cidades entre 2% e 3%**.
- Trabalhos futuros:
  - Utilização de um censo específico da internet.
  - Comparar os resultados dos métodos de Inferência Ecológica com resultados agregados de algoritmos de classificação supervisionada.

Referências

- **Inferência de Atributos Demográficos**
  - Zhong, Yuan, et al. "**You are where you go: Inferring demographic attributes from location check-ins**". International Conference on Web Search and Data Mining, 2015.
  - Bi, Bin, et al. "**Inferring the demographics of search users: social data meets search queries**". WWW, 2013.
  - Dougnon, Raïssa Yapan, Philippe Fournier-Viger, and Roger Nkambou. "**Inferring user profiles in online social networks using a partial social graph**". Advances in Artificial Intelligence, 2015.
- **Inferência Ecológica**
  - King, G. "**A solution to the ecological inference problem.**" Princeton, NJ: Princeton University Press, 1997
  - Wakefield, Jon. "**Ecological inference for 2 x 2 tables (with discussion)**". Journal of the Royal Statistical Society: Series A (Statistics in Society)
  - Imai, Kosuke, Ying Lu, and Aaron Strauss. "**Bayesian and likelihood inference for 2 x 2 ecological tables: an incomplete-data approach**". Political Analysis, 2008.
  - Flaxman, Seth R., Yu-Xiang Wang, and Alexander J. Smola. "**Who Supported Obama in 2012?: Ecological Inference through Distribution Regression**". Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.