GUSTAVO PENHA

# Análise de métodos de Inferência Ecológica em dados de redes sociais

2016/1

GUSTAVO PENHA

# Análise de métodos de Inferência Ecológica em dados de redes sociais

Apresentado como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFMG

Universidade Federal de Minas Gerais – UFMG

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

Orientadora: Ana Paula Couto da Silva

Coorientadora: Mirella M. Moro

2016/1

# Sumário

# 1 Introdução

As mídias sociais se tornaram extremamente populares recentemente e têm gerado um grande volume de conteúdo espontâneo (MISLOVE et al., 2007) tornando possível obter um feedback rápido sobre diversos assuntos como, por exemplo, a fatia de mercado de uma marca (ZHANG; KIM; XING, 2015), a opinião em relação a um candidato político (TUMASJAN et al., 2010; O'CONNOR et al., 2010) entre inumeras outras aplicações. Utilizar os dados gerados de redes sociais se tornou uma opção pouco custosa de estimar a opinião pública, em detrimento de métodos tradicionais como pesquisas de rua em períodos de eleição que podem ser feitas diariamente para acompanhar intenções de voto para cada candidato segmentada por classe social, idade e gênero. Dada a preocupação recente de proteger dados pessoais dos usuários das redes sociais (levando à baixa disponibilidade de atributos públicos), abordagens para inferir suas características demográficas já foram propostas na literatura (DOUGNON; FOURNIER-VIGER; NKAMBOU, 2015), possuindo diversas aplicações como o direcionamento de campanhas e serviços. Entretanto, elas possuem o objetivo de inferir características de cada usuário.

Este trabalho propõe a análise de modelos capazes de inferir características de grupos de usuários, independente de dados inviduais, sendo assim bem menos custosa do que inferir tais cacterísticas demográficas individualmente. Para tal, recorremos ao tema chamado **Inferência Ecológica**, que é o processo de extrair pistas sobre o comportamento individual a partir de informações relatadas no nível de grupo ou agregado (KING; TANNER; ROSEN, 2004). O nome vem do conceito de relações ecológicas, que são variáveis observadas para um grupo de indivíduos ao invés de correlações indivuais (ROBINSON, 1950). Este problema surge em diversas áreas na qual pesquisadores precisam de informação de um grupo de individuos mas não conseguem obtê-la diretamente por motivos de privacidade, custo ou indisponibilidade. Por exemplo, quando cientistas políticos sabem o resultado eleitoral para cada seção do TSE[1] e desejam entender qual o sucesso dos candidatos para diferentes grupos de eleitores (gênero, educação, situação financeira, etc) nestas seções, uma solução é realizar pesquisas de rua perguntando as pessoas me quem elas votaram e qual as suas características demográficas. Essa solução além de cara está propensa a erros, já que as pessoas podem não ser totalmente sinceras em relação à quem votou. Dessa maneira, em casos na qual o acesso a essas informações é infactível ou muito difícil a Inferência Ecológica é a única maneira de fazer progresso.

Para ilustrar o problema da Inferência Ecológica considere a Tabela 1, na qual sabemos a porcentagem de votos que os candidatos em uma seção eleitoral receberam e a distribuição de homens e mulheres nesta mesma seção. É possível inferirmos os pontos

[1] http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais

Tabela 1 – Exemplo do problema de inferência ecológica. Dado uma seção eleitoral sabemos quantos homens e mulheres votaram, assim como quantos votos cada um dos candidatos receberam. Conseguimos inferir os valores interiores da tabela?

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

interiores da tabela que não temos acesso, descobrindo por exemplo qual a porcentagem de homens que votaram na Dilma nesta seção? As soluções para este problema possuem uma série de contra-soluções e controvérsias na literatura, e devido as hipóteses assumidas por cada modelo devem ser usadas com cuidado (CHO, 1998; FREEDMAN et al., 1998). Vale lembrar que não é sempre correto pensar que relações observadas para grupos necessariamente valem para indivíduos (falácia ecológica) (FREEDMAN, 1999) e que o problema de inferir relações individuais a partir de relações entre grupos está fortemente relacionado com o paradoxo de Simpson (BLYTH, 1972). Entretanto, abandonar qualquer tentativa de inferência ecológica não é viável: existem casos por exemplo em que as pessoas que poderiam responder as pesquisas já não estão mais vivas como em questões históricas (votos para o partido nazista na Alemanha); ou em casos em que as pessoas estão relutantes em responder questões sensitivas, sendo assim uma informação difícil de obter.

Desde o artigo de (KING, 1997), que propôs uma solução que une o método dos limites (DUNCAN; DAVIS, 1953) e a regressão de (GOODMAN, 1959), vários trabalhos e modelos estatísticos foram propostos com o intuito de melhorar a qualidade dos modelos e oferecer estimativas mais próximas em relação ao comportamento individual (FLAXMAN; WANG; SMOLA, 2015; PARK; HANMER; BIGGERS, 2014; IMAI; LU; STRAUSS, 2008; GREINER; QUINN, 2009). Esforços nos anos seguintes ao artigo de (KING, 1997) foram sumarizados em (KING; TANNER; ROSEN, 2004) incluindo baselines alternativos para o problema, novas fontes de informação através de novos métodos e modelos e também a utilização de informações geográficas e temporais nos modelos de inferência ecológica.

Este trabalho não propõe um novo modelo de Inferência Ecológica, mas sim investiga uma aplicação inédita de modelos existentes de Inferência Ecológica: utilizar uma base de dados gerada por **redes sociais**. Existem diversas aplicações para esses modelos em redes sociais, já que inferir as características demográficas de subgrupos de pessoas pode ser útil, por exemplo, para tracking político (entender o perfil de pessoas que estão apoiando/não estão apoiando um candidato) e para marketing (entender quais são os grupos de pessoas que apoaim a marca/produto). Propomos então uma inferência de atributos como idade e gênero de grupos que ao utilizar Inferência Ecológica é menos

custosa (em atributos necessários e computacionalmente) que métodos supervisionados existentes (BI et al., 2013; DOUGNON; FOURNIER-VIGER; NKAMBOU, 2015, 2015). Neste contexto, as questões de pesquisa tratadas por este trabalho são as seguintes:

- **Q1**: Existem padrões na base de dados social que poderiam influenciar o resultado dos métodos de Inferência Ecológica?
- **Q2**: Quais são os hiper-parâmetros que otimizam os métodos de Inferência Ecológica na base de dados social?
- **Q3**: Qual modelo de Inferência Ecológica apresenta os melhores resultados na base de dados social?
- **Q4**: Os erros dos métodos de Inferência Ecológica na base de dados social são significativamente diferentes dos erros de uma base de dados eleitoral?

Com base nessas perguntas, as principais contribuições feitas por este trabalho se dividem em quatro partes:

1. Propomos uma nova abordagem para descobrir atributos demográficos em redes sociais a partir de Inferência Ecológica. Coletamos e caracterizamos a base de dados social, descobrindo padrões que motivam particionamentos, como a amostra $N$ por cidade $i$ (número elevado de usuários e número pequeno de usuários) e o intervalo de coleta de tempo $t$ (agregando todos os meses ou dividindo em grupos menores) **(Q1)**.
2. Otimizamos os híper-parâmetros dos três métodos escolhidos (KING, 1997; IMAI; LU; STRAUSS, 2008; WAKEFIELD, 2004) na base de dados social, a partir de projetos fatoriais $2^k$ que apontam para os fatores que explicam maior variação seguido de uma busca em mais níveis nestes fatores. A partir de uma avaliação experimental detalhada, comparamos os métodos de Inferência Ecológica em diferentes configurações da base de dados social, mostrando que o método que apresenta os melhores resultados é o de (KING, 1997), empatado estatisticamente com (IMAI; LU; STRAUSS, 2008) em algumas configurações **(Q2 e Q3)**.
3. Mostramos também que os métodos de Inferência Ecológica apresentam resultados melhores para grupos de cidades com uma quantidade amostral $N$ de usuários alta do que para o grupo de cidades com essa quantidade pequena e que, inesperadamente, o intervalo de tempo $t$ na qual agrupamos os dados (todos os meses contra metade dos meses) não resulta em diferença estatística entre os resultados dos métodos **(Q1 e Q3)**.

4. Por fim, realizamos uma comparação entre os algoritmos de Inferência Ecológica por tipo de base: uma base social e uma base eleitoral. Mostramos que apesar dos erros serem menores com significância estatística na base social, não podemos concluir que isso vale para qualquer base de dados gerada com dados de redes sociais, já que a base coletada possui limites pequenos para as variáveis de interesse **(Q4)**.

# 2 Referencial Teórico

## 2.1 Contextualização

O objetivo da inferência ecológica é inferir informações de comportamento desagregados a partir de dados agregados. As variáveis de interesse que tentamos extrair são geralmente de difícil acesso ou impossíveis de obter. A notação seguida no trabalho pode ser observada na Tabela 2. Para cada cidade $i$ existe uma tabela dessa, na qual conhecemos as variáveis X, Y (por exemplo a porcentagem de homens e a porcentagem de apoiadores) e desejamos inferir as variáveis não observadas W1 e W2 (por exemplo apoiadores do sexo masculino e apoiadores do sexo feminino).

Tabela 2 – Tabela ecológica 2 X 2 genérica.

| | **Grupo 1** | **Grupo 2** | |
|---|---|---|---|
| **Característica 1** | $W1_i$ | $1 - W1_i$ | $Y_i$ |
| **Característica 2** | $W2_i$ | $1 - W2_i$ | $1 - Y_i$ |
| | $X_i$ | $1 - X_i$ | |

Dizemos que essa tabela é 2 X 2 (duas colunas e duas linhas), mas existem também tabelas na qual existem mais linhas e mais colunas, que são chamadas de casos R X C (mais de duas linhas e colunas). Modelos já foram desenvolvidos para ambos os casos, sendo alguns menos flexíveis (IMAI; LU; STRAUSS, 2008; WAKEFIELD, 2004) ao aceitar apenas tabelas 2 x 2 e outros mais flexíveis ao aceitar tabelas R X C (FLAXMAN; WANG; SMOLA, 2015; GREINER; QUINN, 2009). Inferir os valores não observados de tais tabelas pode ser útil para uma série de problemas diferentes. Dessa maneira as técnicas de inferência ecológica podem ser utilizadas em diversas áreas: ciência política (KING, 1997; PARK; HANMER; BIGGERS, 2014; FLAXMAN; WANG; SMOLA, 2015), sociologia e história (O'LOUGHLIN, 2000; DUNCAN; DAVIS, 1953), epidemiologia espacial e saúde (ELLIOT et al., 2000; JACKSON; BEST; RICHARDSON, 2006), entre outras como marketing e publicidade (KING; TANNER; ROSEN, 2004).

## 2.2 Trabalhos Relacionados

A inferência de atributos demográficos em redes sociais possui o objetivo de inferir individualmente as características de cada usuário. A Inferência Ecológica, por sua vez, tem o objetivo de inferir características para grupos de pessoas a partir de informações agregadas.

### 2.2.1 Inferência de Atributos Demográficos

A inferência de atributos dos usuários de sistemas pode ser utilizada para a melhoria dos sistemas e comercialmente, já que conhecer o perfil de uma pessoa contribui significantemente com recomendação de itens, propaganda direcionada, marketing, e predição de links. Estudos já foram feitos utilizando diferentes iterações do usuário com diferentes sistemas para inferir os seus atributos demográficos.

Por exemplo, (BI et al., 2013) reportou que, baseado somente no histórico de pesquisas em um sistema de busca, obteve uma alta acurácia de predição dos atributos de gênero e idade, assim como visão política e afiliação com o judaísmo. (ZHONG et al., 2015) mostrou via um modelo de decomposição de tensores que é possível inferir uma série de características como gênero, idade, formação acadêmica, orientação sexual e etc a partir de check-ins. Outro rastro digital que já foi utilizado como entrada para um modelo que prediz as informações demográficas dos seus usuários foi o uso de redes sociais de comunicação móvel por (DONG et al., 2014), na qual o seu modelo, que explora a rede complexa formada a partir das ligações, apresenta desempenho melhor do que diversos classificadores. Outra abordagem de grafos proposta recentemente por (DOUGNON; FOURNIER-VIGER; NKAMBOU, 2015) resulta em alta acurácia em atributos como gênero e estado (estudante ou professor).

Apesar de várias aborgadens de inferência de atributos demográficos apresentarem resultados de acurácia acima de 90% para certos atributos, eles necessitam da existência de diversos casos de atributos observados e públicos para uma grande quantidade de usuários. A tendência dos usuários de tornarem seus perfis mais privados dificulta a obtenção de diversas informações que podem ser essenciais para algoritmos supervisionados de classificação. Além disso, existe uma tendência das redes sociais fornecerem somente dados agregados e não mais no nível individual, como no caso do Facebook DataSift. No nosso caso, estamos utilizando apenas atributos agregados de um censo (como o IBGE) e atributos agregados de usuários (como o sentimento agregado de suas publicações) e não sofremos com a falta de disponibilidade desses atributos ao utilizar algoritmos de Inferência Ecológica.

### 2.2.2 Inferência Ecológica

Apesar de o tema ter sido estudado por sociologistas nos anos 50 (DUNCAN; DAVIS, 1953; ROBINSON, 1950) o interesse ressurgiu recentemente entre estatísticos, metodologistas políticos e cientistas da computação apresentando novas abordagens e utilizando novas fontes de informação (KING, 1997; WAKEFIELD, 2004; IMAI; LU; STRAUSS, 2008; PARK; HANMER; BIGGERS, 2014; FLAXMAN; WANG; SMOLA, 2015). Os novos métodos têm abordado diferentes maneiras de aprimorar as inferências, incluindo a possibilidade de utilização de micro-dados escassos para melhorar a acurácia

da inferência (JACKSON; BEST; RICHARDSON, 2006; WAKEFIELD, 2004), utilização de informação espacial para melhorar o modelo estatístico (KING; TANNER; ROSEN, 2004; FLAXMAN; WANG; SMOLA, 2015; CRAWFORD; YOUNG, 2004) e a utilização de informação temporal, por exemplo, analisando os mesmos distritos eleitorais em eleições consecutivas (LEWIS, 2004; QUINN, 2004).

O modelo proposto por (KING, 1997), que reacendeu o interesse de pesquisadores no assunto após vários anos durante a predominância dos modelos de limites (DUNCAN; DAVIS, 1953) e a regressão de (GOODMAN, 1959), faz três suposições em relação aos dados. Primeiro ele assume que os pontos de interesse devem pertencer a um único cluster dentro do quadrado de 0 a 1, o que matematicamente significa que os valores $W1$ e $W2$ seguem uma normal bivariada truncada. O segundo ponto é que não existe auto-correlação espacial, logo as condicionais em $X_i$, $X_j$ e $T_j$ são *mean-independent*. Por último, o seu modelo assume que $X_i$ é independente de $W1_i$ e $W2_i$. O seu modelo, que une o método dos limites e um modelo estatístico de (GOODMAN, 1959), trouxe novos estudos e motivação na área, que são sumarizados em (KING; TANNER; ROSEN, 2004).

Um destes trabalhos é o de (WAKEFIELD, 2004), que discute a escolha das hipótese feitas por cada modelo, como por exemplo, utilização ou não de efeitos contextuais. Além disso, são propostos vários modelos bayesianos hierárquicos neste trabalho, sendo que o autor conclui que sem a inclusão formal de dados suplementares ou informação apriori, como por exemplo uma quantidade de usuários rotulados individualmente, os resultados da inferência ecológica não são confiáveis. (IMAI; LU; STRAUSS, 2008) também investiga o problema de inferência ecológica decompondo-o em três fatores: *distribution effects*, *contextual effects* e *aggreagation effects*. O autor examina cada um desses fatores, propondo novos métodos estatísticos para lidar com estes efeitos. Outro trabalho recente, que apresentou uma nova abordagem para resolver o problema da inferência ecológica, obteve bons resultados nos dados da eleição de 2012 dos Estados Unidos (FLAXMAN; WANG; SMOLA, 2015). Neste trabalho é proposto um modelo de regressão de distribuição que explora a conexão entre processos gaussianos e *ridge kernel regression*, além de considerar a variação espacial e existir a possibilidade de utilizar micro-dados esparsos para melhorar a acurácia das inferências.

## 2.3 Justificativa

Dada a grande variedade de modelos estatísticos cada qual seguindo suas hipóteses, existe a necessidade de um trabalho de análise deles no contexto de dados gerados por redes sociais. É possível fazer suposições similares neste caso? Os modelos se comportam bem no contexto de redes sociais? Qual deles apresenta os melhores resultados? Para responder tais perguntas este trabalho propõe uma comparação dos modelos de inferência

ecológica no contexto de dados provenientes de redes sociais, uma vez que os trabalhos existentes utilizam dados obtidos a partir de pesquisas de rua para avaliar seus métodos (KING, 1997; WAKEFIELD, 2004; FLAXMAN; WANG; SMOLA, 2015) e as hipóteses assumidas podem não valer para o contexto de redes sociais. Inferir as características demográficas de subgrupos de pessoas nas redes sociais pode ser útil, por exemplo, para tracking político (entender o perfil de pessoas que estão apoiando/ não estão apoiando) e para marketing (entender quais são os grupos de pessoas que apoaim a marca/produto), uma vez que o poder do grande volume de dados gerados a partir das redes sociais para entender tendências e opinião pública já foi bem estudado na literatura (O'CONNOR et al., 2010; BODENDORF; KAISER, 2009; FILHO, 2003).

# 3 Modelos de Inferência Ecológica

Neste Capítulo é descrito o modelo de cada uma das três soluções de Inferência Ecológica analisadas, assim como as suas respectivas premissas. A escolha dos três métodos de Inferência Ecológica deste trabalho foi baseada em critérios de reproducibilidade do modelo e relevância do artigo que o propôs. Apesar disso, outros modelos de Inferência Ecológica de tabelas $2X2$ podem ser aplicados nos dados coletados e modelados neste trabalho.

## 3.1 Solução de King

A solução de (KING, 1997) implementada no pacote EI[1] faz três suposições, cada uma podendo ser relaxada de maneiras diferentes. A primeira é que o conjunto de pontos $W1_i$ e $W2_i$ deve cair em um mesmo agrupamento dentro do quadrado unitário podendo ser altamente dispersa e além disso as duas incognitas podem ser positivamente, negativamente ou não correlacionadas. A versão matemática dessa suposição é que $W1_i$ e $W2_i$ seguem uma normal bivariada truncada

$$TN(W1_i, W2_i|, \breve{\mathfrak{B}}, \breve{\Sigma}) = N(W1_i, W2_i|, \breve{\mathfrak{B}}, \breve{\Sigma})\frac{\mathbf{1}(W1_i, W2_i)}{R(\breve{\mathfrak{B}}, \breve{\Sigma})},$$

na qual o kernel é uma normal bivariada não truncada, e a função $\mathbf{1}(W1_i, W2_i)$ é igual a um se ambos valores estão entre 0 e 1 e zero caso contrário. $R(\breve{\mathfrak{B}}, \breve{\Sigma})$ é o volume abaixo da curva da normal bivariada não truncada. A segunda suposição é a ausência de autocorrelação espacial: condicionais em $X_i$, $Y_i$ e $Y_j$ são independentes de média. Por último, a terceira e mais crítica suposição é que $X_i$ é independente de $W1_i$ e $W2_i$. Uma generalização proposta por King da última suposição, permite os parâmetros da normal truncada variarem de acordo com covariações medidas, $Z1_i$ e $Z2_i$, resultando em

$$\breve{\mathfrak{B}}1_i = [\phi_1(\sigma_1^2 + 0.25) + 0.5] + (Z1_i - Z1)\alpha 1,$$

$$\breve{\mathfrak{B}}2_i = [\phi_2(\sigma_2^2 + 0.25) + 0.5] + (Z2_i - Z2)\alpha 2,$$

na qual $\alpha 1$ e $\alpha 2$ são vetores de parâmetros a serem estimados junto com os parâmetros do modelo original, sumarizados como $\breve{\psi} = \{\breve{\mathfrak{B}}1_i, \breve{\mathfrak{B}}2_i\, \sigma_1, \sigma_2, \rho\} = \{\breve{\mathfrak{B}}, \breve{\Sigma}\}$.

[1] https://cran.r-project.org/web/packages/ei/index.html,

## 3.2 Solução de Wakefield

A solução proposta por (WAKEFIELD, 2004) implementada no pacote MCMCpack[2] assume que

$$W1_i|Y_i \sim \mathcal{B}inomial(Y_i, p_{1i}),$$

$$W2_i|1 - Y_i \sim \mathcal{B}inomial(1 - Y_i, p_{2i}),$$

sendo $\theta_1 i = log(p_{1i}/(1-p_{1i}))$ e $\theta_2 i = log(p_{2i}/(1-p_{2i}))$. As seguintes distribuições posteriores são assumidas: $\theta_1 i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $\theta_2 i \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Além disso, assume-se que $\theta_1 i$ é uma distribuição a priori independente de $\theta_2 i$ para todo $i$. Por último, assume-se os seguintes *hyperpriors*: $\mu_1 \sim \mathcal{N}(m_1, M_1), \mu_2 \sim \mathcal{N}(m_2, M_2), \sigma_1^2 \sim \mathcal{IG}(\frac{a_1}{2}, \frac{b_1}{2})$ e $\sigma_2^2 \sim \mathcal{IG}(\frac{a_2}{2}, \frac{b_2}{2})$. Sendo que $a_1, a_2, b_1, b_2, m_1, M_1, m_2$ e $M_2$ são hiperparâmetros do modelo.

## 3.3 Solução de Imai

A solução proposta por (IMAI; LU; STRAUSS, 2008), implementada no pacote eco[3] e analisada aqui, assume a existência de efeitos contextuais e é chamada de modelo NCAR pelos autores. O modelo não-paramétrico Bayesiano proposto é dado por,

$$W_i^*|\mu_i, \Sigma_i \sim \mathcal{N}(\mu_i, \Sigma_i),$$

$$\mu_i, \Sigma_i|G \sim G,$$

$$G|\alpha \sim \mathcal{D}(G_0, \alpha),$$

$$\alpha \sim \mathcal{G}(a_0, b_0),$$

na qual para $G_0$, $(\mu_i, \Sigma_i)$ é distribuido como $\mu_i|\Sigma_i \sim \mathcal{N}(\mu_0, \frac{\Sigma_i}{\tau_0^2})$ e $\Sigma_i \sim InvWish(v_0, S_0^{-1})$. O modelo permite então modelar os parâmetros $\{\mu_i, \Sigma_i\}$ i=1, ... ,n com uma função de distribuição desconhecida $G$ ao invés de uma função fixa.

[2] https://cran.r-project.org/web/packages/MCMCpack/index.html
[3] https://cran.r-project.org/web/packages/eco/

# 4 Configuração de Experimentos

Neste capítulo é descrita a metodologia seguida neste trabalho. Através dos seguintes passos: a base de dados que será utilizada e o procedimento para avaliar os resultados.

## 4.1 Base de dados social

Uma vez que este trabalho tem como objetivo uma aplicação inédita para inferência ecológica (utilização de dados gerados por redes sociais online), caracterizar os dados é uma questão central para que se possa aplicar os métodos de inferência ecológica com confiança. A coleta foi realizada no Twitter durante o intervalo entre 25 de novembro de 2015 e 25 de março de 2016, na empresa Hekima[1], utilizando o Zahpee monitor[2]. O monitoramento teve como objetivo coletar publicações relacionadas à Dilma, de forma que somente posts que possuem palavras como *dilma*, *rousseff*, *dilmabr*, *dilmãe*, *naovaiterdilma*, *impeachmentdilma*, *somostodosdilminha* foram coletados. Além disso, uma equipe de cientistas politicos utiliza o software para, em conjunto com o seu algoritmo de aprendizado de máquina, definir qual o sentimento de cada post: negativo, neutro ou positivo.

Parte dos usuários que fizeram esses posts são manualmente categorizadas e os outros são inferidos a partir de algoritmos de classificação. Para este trabalho utilizaremos apenas as inferências de gênero e idade, que são as mais confiáveis pois fazem, respectivamente, um casamento de nomes com um dicionário de nomes masculinos/femininos e uma inferência a partir indícios de idade informados pelos próprios usuários nos seus perfis (ex: "Sou belo horizontino, tenho 20 anos e estudo na UFMG.") entre outras características de cada usuário. Selecionamos assim cada usuário que já publicou posts geolocalizados e construimos uma tabela contendo os seguintes microdados: id do usuário, favorável ou não à Dilma (se a porcentagem de posts positivos for maior que porcentagem de posts negativos), gênero, idade e cidade. Além dessa base de dados, utilizamos o censo do IBGE de 2010 [3] que contém informações demográficas sobre cada cidade do país.

### 4.1.1 Ground-truth e Notações

Uma vez que possuimos os microdados dos usuários das redes sociais (temos conhecimento do sexo e idade de cada um dos usuários e seus respectivos sentimentos em relação à Dilma), conseguimos então saber os valores interiores das tabelas propostas. Conseguimos calcular a porcentagem de pessoas do sexo masculino/feminino que apoiam

[1] http://hekima.com/
[2] https://zahpee.com/funcionalidades
[3] http://www.censo2010.ibge.gov.br/

Tabela 3 – Notações e variáveis.

| Variável | Dados gênero | Dados idade |
|---|---|---|
| $Y_i$ | % de homems na cidade $i$ | % de pessoas com menos de 40 an |
| $X_i$ | % de usuários que são favoráveis a Dilma na cidade $i$ | % de usuários que são favoráveis a D |
| $W1_i$ que são favoráveis a Dilma na cidade $i$ | % de homems que são favoráveis a Dilma na cidade $i$ | % de pessoas com menos de |
| $W2_i$ que são favoráveis a Dilma na cidade $i$ | % de mulheres que são favoráveis a Dilma na cidade $i$ | % de pessoas com mais de 4 |
| $N_i$ | número de usuários coletados na cidade $i$ | número de usuários coletados n |

a Dilma e também as pessoas com menos de 40 anos/mais de 40 anos que são favoráveis à Dilma simplesmente selecionando essas pessoas e calculando a porcentagem desse grupo. Dividindo os dados pelas características demográficas analisadas (sexo e idade), as suas notações são descritas na tabela 3.

### 4.1.2 Caracterização

Durante os 122 dias de coleta sobre assuntos relacionados à Dilma Rousseff foram coletados posts geolocalizados pertencentes a 121.874 e 150.759 usuários geolocalizados distintos no Twitter com informação de gênero e com informação de idade respectivamente. Para cada um desses usuários sabemos qual a quantidade de posts negativos, neutros e positivos em relação ao monitoramento da presidente. É considerado então que se o usuário tem quantidade de positivos maior que negativos ele é positivo em relação à Dilma e não positivo caso contrário.

Agrupando os usuários por cidades em que se encontram, temos 159 e 181 cidades, com usários de gênero e idade respectivamente, com mais de 100 pessoas coletadas. O gráfico da ECDF 1 do número de pessoas em cada cidade mostra que em ambos os datasets (gênero e idade) existem algumas cidades que apresentam grande quantidade de posts geolocalizados e a maioria possui um valor inferior a 5000 usuários. Isso motiva um particionamento dos dados por quantidade de usuários na cidade ($N_i$). Para tal, aplicando o algoritmo de clusterização *kmeans*[4] e utilizando o *elbow method*[5] para definir o número de clusters, observamos 6 clusters totais, resumidos na tabela 4.

Outro particionamento que podemos fazer nos dados é cronológico, motivado pela possível mudança de opinião durante os meses (dado o cenário econômico e político do país) e a perda de informação ao agregar informação de posts de usuários de 4 meses. Ao dividir os dados em dois grupos de aproximadamente 2 meses cada vemos que o apoio em

[4] O algoritmo de clusterização é uma técnica de aprendizado de máquina não supervisionada que visa particionar os dados em $k$ grupos, na qual as observações são atribuidas ao grupo com a média mais próxima de maneira iterativa, maximizando de maneira gulosa a soma dos erros quadráticos (ZAKI; MEIRA, 2014).

[5] O método do cotovelo é uma técnica visual que se baseia em fazer um gráfico da métrica de custo do algoritmo pelo número de grupos $k$. A idéia é que a curva decresça rapidamente nos primeiros valores de $k$ e que diminua com menor velocidade em um ponto "cotovelo"(KODINARIYA; MAKWANA, 2013).

Tabela 4 – Tamanho e média dos clusters para $N_i$.

| **Gênero** | | **Idade** | |
|---|---|---|---|
| **Média de $N_i$** | **Tamanho cluster** | **Média de $N_i$** | **Tamanho cluster** |
| 18860.5000 | 2 | 22791.5 | 2 |
| 5232.2500 | 4 | 6337 | 4 |
| 2549.0000 | 6 | 3129.5 | 6 |
| 1374.8750 | 8 | 1681 | 8 |
| 493.0769 | 39 | 594.8049 | 41 |
| 177.0100 | 100 | 193.4667 | 120 |

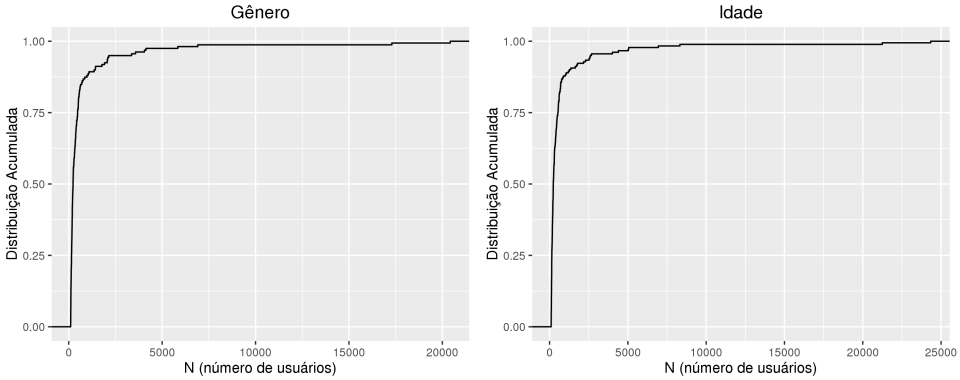


Figura 1 – ECDF do número de usuários por cidade ($N_i$).

relação a Dilma Rousseff cai em ambas categorias demográficas e na média agregada (Y passa de 0.089459% nos primeiros dois meses para 0.074322% no últimos dois meses). O impacto de aplicar os métodos de Inferência Ecológica em ambos os agrupamentos (por tempo e por N) são analisados no Capítulo 5 deste trabalho.

Utilizando apenas as variáveis observadas (X e Y) nas duas bases, existe uma tendência do apoio à Dilma aumentar para cidades com maior porcentagem de homens e diminuir quando a porcentagem de pessoas com menos de 40 anos aumenta, Figura 2. Entretanto, isso não indica que os homens geralmente falam melhor em relação à Dilma nem que as pessoas mais jovens possuem uma tendência a publicar coisas positivas em relaço ao seu governo. Este problema está fortemente relacionado com o paradoxo de Simpson[6], pois, pode existir por exemplo uma cidade na qual a porcentagem de homens é superior a de mulheres entretanto nenhum homem é favorável a Dilma, mas como o índice é muito alto para as mulheres a porcentagem agregada de apoio de homens mais a de mulheres fica elevada.

Na verdade, vemos que isso realmente acontece, já que o apoio é maior para o sexo feminino do que para o sexo masculino. A Figura 3 mostra a diferença entre os histogramas e as funções ECDF das variáveis W1 e W2 para os dados de gênero, calculando a média para todas as cidades vemos que o apoio de homens é 0.071092% enquanto o de mulheres é de 0.099359%. Em relação a idade existe um valor um pouco maior de apoio nos usuários

[6] O paradoxo de Simpson é o efeito na qual uma tendência aparece em diferentes grupos de dados mas que desaparece ou reverte quando os grupos são agrupados (PEARL, 2013).

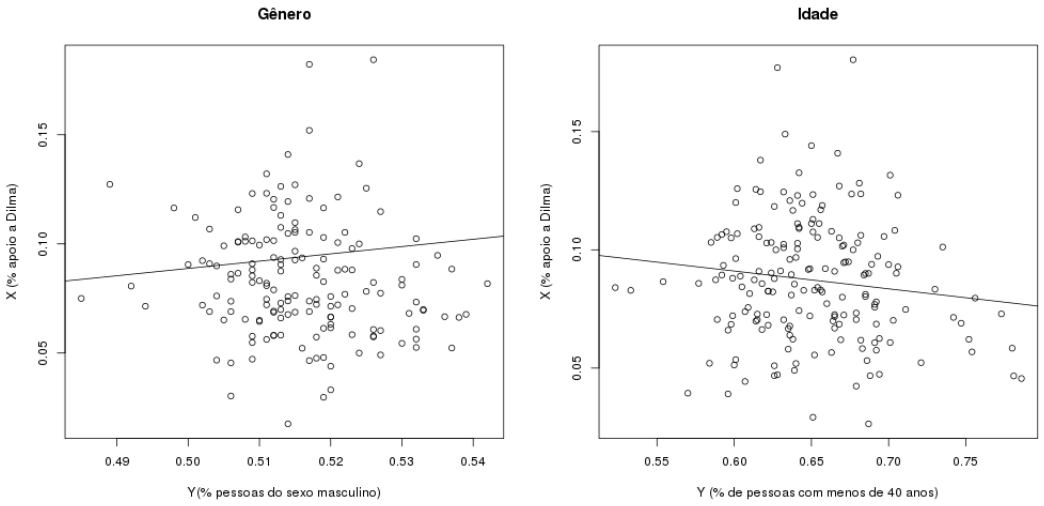


Figura 2 – Gráficos das variáveis X por Y para as duas bases (com linhas de regressão).

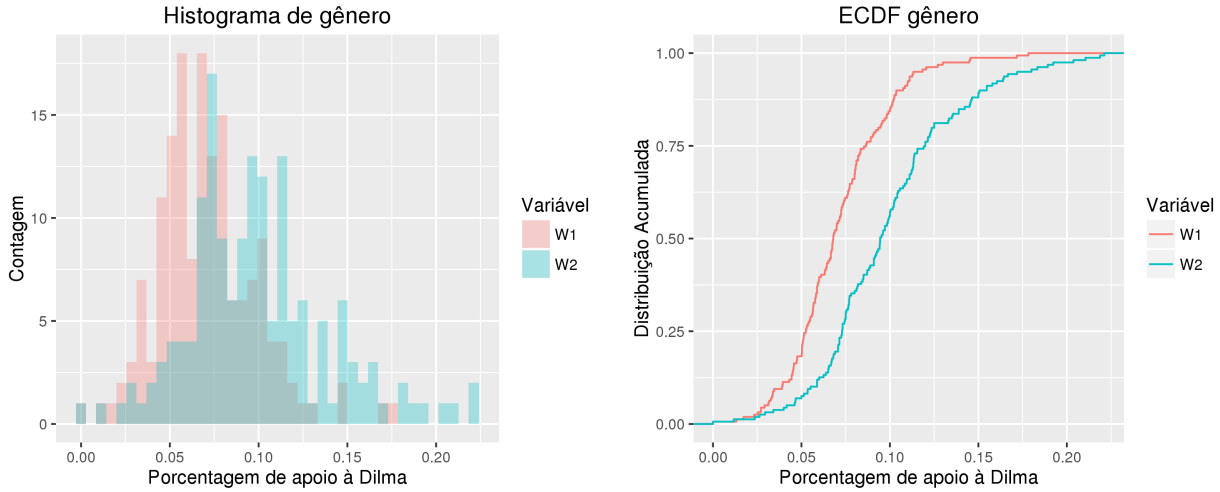


Figura 3 – Gráficos das variáveis W1 (% de homens que apoiam a Dilma) e W2 (% de mulheres que apoiam a Dilma) na base de gênero.

com menos de 40 anos em relação aos que tem mais de 40 anos (médias de respectivamente: 0.085822% e 0.078081%).

Utilizando apenas a identidade de Goodman(GOODMAN, 1959) de que $Y_i = X_i W1_i + (1 - X_i)W2_i$, conseguimos determinar os possíveis valores de W1 e W2 a partir de X e Y. Esses gráficos forma chamados de *tomography plots* por (KING, 1997). A Figura 4 mostra tais gráficos para os dados de gênero e idade, sendo que os valores possíveis para W1 e W2 se encontram na projeção dessas linhas nos eixos x e y. Isso indica que os valores possíveis para as porcentagens, que em inicialmente estariam entre 0 e 1, na verdade só podem estar em intervalos bem menores.

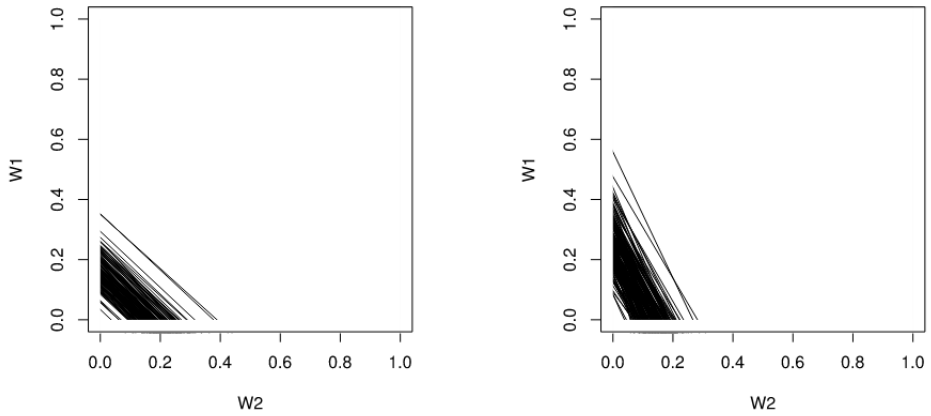


Figura 4 – Gráficos de tomografia para os dados de gênero e de idade

## 4.2 Procedimento de avaliação

Para responder as questões de pesquisa propostas, a metodologia seguida foi a seguinte:

**Métricas de avaliação:** Para avaliar a eficácia dos algoritmos de Inferência Ecológica são utilizadas duas métricas que capturam o quanto as predições estão se distanciando do valor real: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c}_i)^2}$ e $MAE = \frac{1}{N}\sum_{i=1}^{N}|c_i - \bar{c}_i|$. Sendo que $c_i$ é a predição do algoritmo e $\bar{c}_i$ é o valor real. A raiz quadrada da média dos erros quadráticos (Root-Mean-Square Error) penaliza mais erros grandes do que a média dos erros absolutos (Mean Absoulte Error) que considera pesos iguais para todos indivíduos.

**Otimização de hiperparâmetros:** Inicialmente, para avaliar o impacto dos parâmetros no algoritmo é realizado um projeto fatorial $2^k$ (JAIN, 2008), que avalia os fatores em dois níveis cada. Ele é então utilizado para focar nos fatores que mais impactam as métricas de avaliação, já que o projeto fatorial $2^k$ é barato por analisar um espaço limitado de dois níveis por fator. Em seguida, uma busca em mais níveis (limitada a esses fatores escolhidos) é feita para encontrar os hiperparâmetros que minimizam os erros. Tais procedimentos foram realizados em cima da base de dados de idade, contemplando os dados agregando todos os dias de coleta, utilizando a métrica MAE para os valores preditos de W1 e selecionando as cidades com amostras $N$ de tamanho superior a cem usuários.

**Comparação dos métodos:** A partir da definição dos hiperparâmetros que otimizam as métricas de avaliação, os algoritmos são comparados a partir de intervalos de confiança das métricas e do teste pareado (já que podemos medir o erro para a mesma cidade $i$ nos diferentes modelos ) t-test (JAIN, 2008) para os erros em diferentes configurações da base de dados.

**Comparação em diferentes bases de dados:** Para comparar os resultados dos métodos em diferentes bases de dados utilizamos o t-test não pareado (JAIN, 2008), uma vez que a representação de uma cidade $i$ não tem relação entre as bases.

# 5 Análise e Resultados

Este Capítulo apresenta os resultados dos projetos experimentais assim como a análise e comparação dos métodos de Inferência Ecológica em diferentes configurações da base de dados.

## 5.1 Otimização de hiperparâmetros: projetos experimentais $2^k$

Para cada um dos métodos analisados, é realizado um projeto com $2^k$ fatores, sendo que k é o número de hiperparâmetros que cada um dos métodos apresenta[1]. Os resultados mostram como a variável resposta é influenciada por cada um dos seus fatores, realizando experimentos em 2 níveis (alto e baixo) para cada fator (hiperparâmetro) (JAIN, 2008). Este projeto experimental considera que o efeito dos fatores são aditivos em relação à variável resposta. Além disso, outra premissa é que a função da variável resposta é monótona entre os níveis altos e baixos dos parâmetros.

Esse projeto experimental desconsidera os erros experimentais para diferentes replicações do mesmo algoritmo. O projeto $2^k$ foi escolhido em detrimento do $2^k r$ pelo fato dos algoritmos apresentarem pouca ou nenhuma variação de resultado entre as repetições. Utilizando 10 replicações apenas, o desvio padrão dos algoritmos King, Imai e Wakefield são respectivamente 0.0003, 0.023 e 0.0.

### 5.1.1 Solução de King

Para o modelo de King, os seguintes fatores foram analisados: *ehro*, *esigma* e *ebeta*. Os seus respectivos níveis altos e baixos e a nomeclatura dos fatores estão apresentados na Tabela 5. A variação explicada para cada um dos fatores estão apresentados na Tabela 6. Estes resultados motivam uma busca em mais níveis para os fatores A e B, pois eles junto com a interação AB explicam 92% da variação. Uma busca nos parâmetros *ehro* e *esigma* variando de 0.1 em 0.1 entre os valores $[0.0, 0.5]$, minimiza os erros para os seguintes parâmetros: $ehro = 0.0$, $esigma = 0.3$ e $ebeta = 0.5$.

### 5.1.2 Solução de Wakefield

Para o modelo de Wakefield, os seguintes fatores foram analisados: $m0$, $M0$, $m1$, $M1$, $a0$, $b0$, $a1$ e $b1$. Os seus respectivos níveis altos e baixos e a nomeclatura dos fatores

[1] Para mais informações sobre o que cada um dos hiperparâmetros representa em cada um dos modelos veja o Capítulo 2 e a página https://github.com/Guzpenha/Ecological-Inference-on-social-data/blob/master/experiments/Factors%20analysed.ipynb

Tabela 5 – Configuração de níveis de fatores para o modelo de King.

| Fator | Sigla | Nível baixo (-1) | Nível alto (+1) |
|---|---|---|---|
| ehro | A | 0.5 | 1 |
| esigma | B | 0.5 | 1 |
| ebeta | C | 0.5 | 1 |

Tabela 6 – Fração de variação explicada por cada um dos fatores para o modelo de King.

| Fator | Variação explicada |
|---|---|
| A | **35,6**% |
| B | **28**% |
| C | 1,6% |
| AB | **28,5**% |
| AC | 2,2% |
| BC | 1,8% |
| ABC | 2% |

Tabela 7 – Configuração de níveis de fatores para o modelo de Wake.

| Fator | Sigla | Nível baixo (-1) | Nível alto (+1) |
|---|---|---|---|
| m0 | A | 0 | 0.5 |
| M0 | B | 2.3 | 3 |
| m1 | C | 0 | 0.5 |
| M1 | D | 2.3 | 3 |
| a0 | E | 0.8 | 1 |
| b0 | F | 0.01 | 0.03 |
| a1 | G | 0.8 | 1 |
| b1 | H | 0.01 | 0.03 |

Tabela 8 – Maiores frações de variação explicada por fatores para o modelo de Wake.

| Fator | Variação explicada |
|---|---|
| ACEFGH | **4,9%** |
| ADFGH | **4,4%** |
| ABCEG | **3,4%** |
| AE | 3,0% |
| ABDEFH | 2,3% |
| ABDGH | 2,0% |
| BDEFH | 1,7% |

estão dispostos na Tabela 7. Os resultados do projeto fatorial estão dispostos na Tabela 8, sendo que apenas os fatores que apresentam as maiores porcentagens de variação foram dispostos. O modelo de Wakefield foi o único dos três que apresentou maior variação explicada pela interação entre os fatores do que pelos próprios fatores.

Além disso, a porcentagem explicada pelas sete maiores variações explicadas é de apenas aproximadamente 20%. Dessa maneira, foi realizado uma busca em 3 níveis para cada hiperparâmetro, resultando em $3^8$ experimentos, com o intuito de encontrar em quais níveis dos fatores o algoritmo se comporta melhor, sem descartar nenhum fator. Os hiperparâmetros que apresentaram os melhores resultados foram 4: $m0 = 0.0$, $M0 = 2.3$, $m1 = 0.1$, $M1 = 2.7$, $a0 = 0.7$, $b0 = 0.08$, $a1 = 0.9$ e $b1 = 0.03$.

### 5.1.3 Solução de Imai

Para o modelo de Imai, os seguintes fatores foram analisados: $mu0$, $tau0$, $nu0$, $s0$, $mustart$ e $sigmastart$. Os seus respectivos valores altos e baixos e a nomeclatura dos parâmetros estão apresentados na Tabela 9. Os resultados do projeto fatorial estão apresentados na Tabela 10, na qual apenas os fatores que apresentam as maiores porcentagens de variação fora dispostos. Uma vez que os fatores que explicam aproximadamente 50% da variação foram $mu0$, $tau0$ e $sigmastart$, uma busca de mais níveis nestes parâmetros foi realizada, exibindo os melhores resultados para os seguintes parâmetros: $mu0 = 0.1$, $tau0 = 0$, $nu0 = 6$, $s0 = 15$, $mustart = 0$ e $sigmastart = 19$.

Tabela 9 – Configuração de níveis de fatores para o modelo de Imai.

| **Fator** | **Sigla** | **Nível baixo (-1)** | **Nível alto (+1)** |
|---|---|---|---|
| mu0 | A | 0 | 1 |
| tau0 | B | 2 | 4 |
| nu0 | C | 4 | 6 |
| s0 | D | 10 | 15 |
| mustart | E | 0 | 2 |
| sigmastart | F | 10 | 15 |

Tabela 10 – Maiores frações de variação explicada por fatores para o modelo de Imai.

| **Fator** | **Variação explicada** |
|---|---|
| A | **34,4**% |
| B | **8,8**% |
| F | **4,3**% |
| BE | 3,5% |
| BD | 2,8% |
| E | 1,7% |
| AB | 1,59 % |

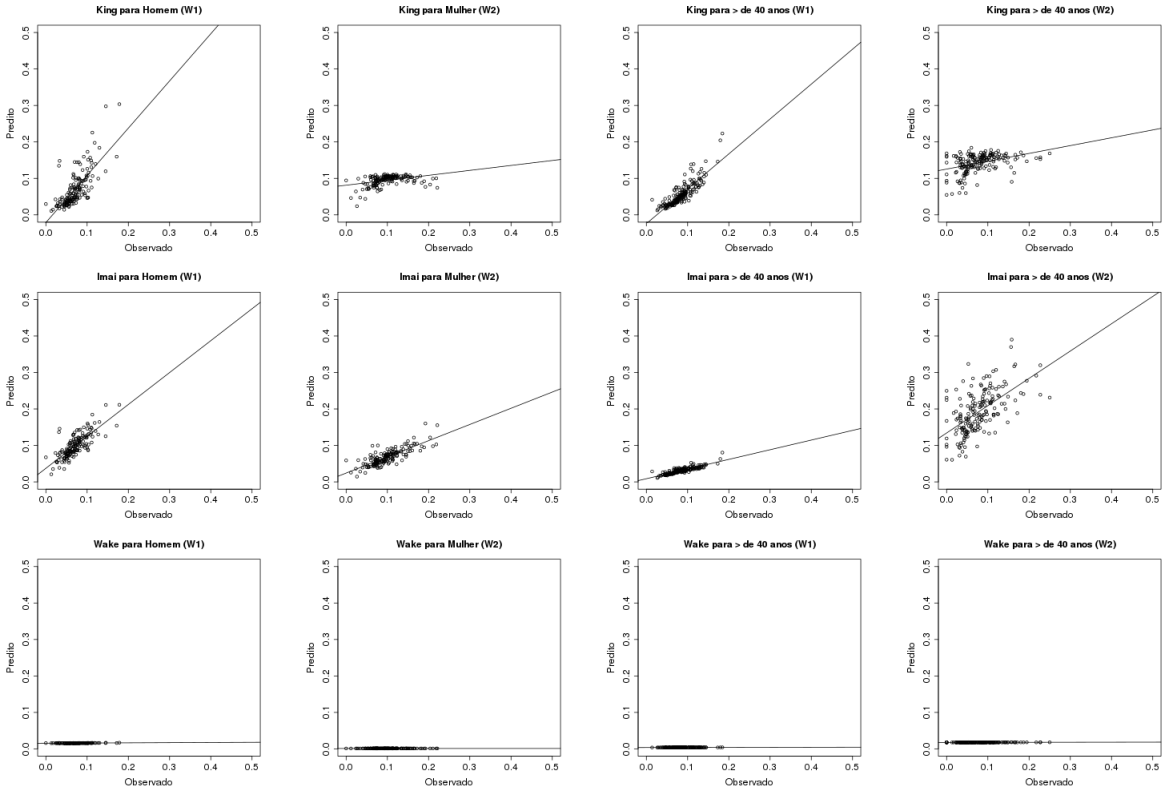


Figura 5 – Gráficos da relação entre as predições e os valores verdadeiros para os 3 métodos nas bases de gênero e idade.

## 5.2 Comparação dos métodos

Utilizando os hiperparâmetros obtidos na seção anterior, para todos os meses da base de dados, podemos ver como as predições de cada um dos três modelos para as características de gênero e faixa etária se relacionam com os valores reais na Figura 5. Visualmente os modelos que apresentam as melhores linhas de regressão entre a predição e o verdadeiro são os modelos de King e de Imai. As correlações entre a predição e o valor observado também são maiores para esses dois métodos (nos dados de gênero as correlações de W1 e W2 para King e Imai são respectivamente de (0.78, 0.76) e (0.78, 0.80) enquanto para o método de Wake são (0.61, 0.19) ).

Os resultados das duas métricas de avaliação para os modelos nas duas caracte-

Tabela 11 – 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 |

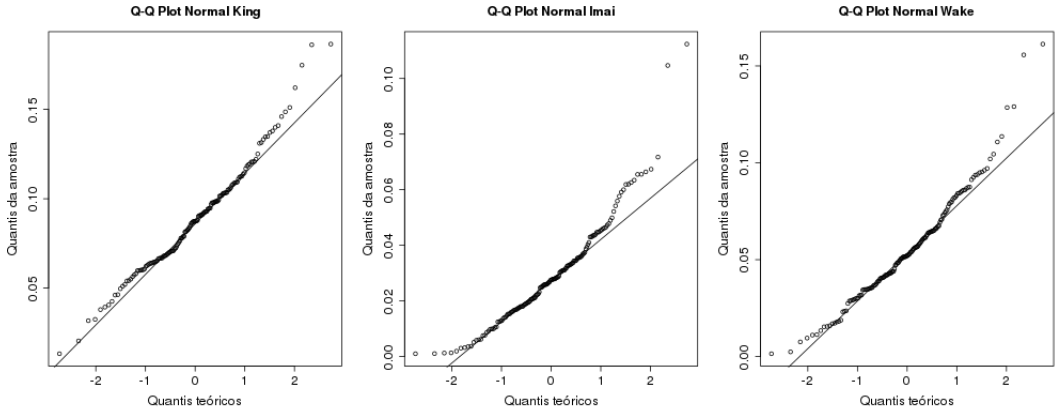


Figura 6 – Gráficos mostram o QQ Plot normal para os erros absolutos dos modelos em W1 na base de gênero.

rísticas demográficas estão apresentados na Tabela 11. Os valores em negrito são aqueles em que os intervalos de confiança da média das métricas permitem dizer que os erros são diferentes dos outros 2 modelos com 95% de confiança. Para os erros quadráticos, na qual os intervalos de confiança foram calculados entre as diferentes etapas da validação cruzada, não existe diferença entre as métricas de King e Imai. Já para o erro absoluto, observa-se que o modelo de King é diferente em ambas as bases de dados do modelo de Imai e apenas para o W2 da base de idade não conseguimos descartar a hipótese de que são diferentes pelo intervalo de confiança.

A Figura 6 mostra que a premissa de normalidade para os erros dos algoritmos é válida e portanto permite que seja realizado um teste pareado paramétrico t-test. Os resultados do teste estão apresentados na Tabela 12, os resultados em negrito são os testes que com 95% de confiança apresentam erros diferentes. Os resultados reforçam que os modelos que apresentam os melhores resultados são King e Imai, sendo que na base de idade o modelo de King apresenta os melhores resultados e na base de gênero não conseguimos refutar a hipótese nula de que os dois são iguais com uma confiança alta.

Tabela 12 – P-values para o t-test pareado utilizando os erros absolutos.

| | **Base de gênero** | | **Base de Idade** | |
|---|---|---|---|---|
| **Modelo** | W1 pvalue | W2 pvalue | W1 pvalue | W2 pvalue |
| King e Imai | 0.5652 | **0.0248** | **2.7167e-15** | **2.3207e-11** |
| Imai e Wakefield | **7.1609e-23** | **4.5786e-68** | **1.4149e-90** | **5.9135e-06** |
| Wakefield e King | **9.0796e-30** | **1.7818e-60** | **1.5641e-60** | 0.14589 |

## 5.3 Comparação em diferentes configurações dos dados

Esta seção compara os algoritmos em diferentes configurações da base de dados, avaliando como o desempenho é sensivel à $N_i$ (tamanho da amostra da cidade $i$) ao intervalo de tempo de dados utilizados.

### 5.3.1 Sensitividade à variavel *N*

A caracterização da base de dados da Seção 3.1 mostra uma existem poucas cidades com um número grande de usuários distintos e várias cidades com número pequeno de posts geolocalizados e consequentemente usuários. Especificamente, a clusterização por $N$ resultou em 6 grupos, que são agrupados nesta seção em dois grupos maiores devido ao pequeno número de cidades nos 5 primeiros grupos: Grupo 1 com cidades $i$ com $N$ menor que 200 e Grupo 2 com cidades $i$ com $N$ maior que 200. O primeiro grupo apresenta 72 cidades e o segundo apresenta 111 cidades para os dados de idade (61 e 92 para os dados de gênero).

A Figura 7 apresenta o MAE e o intervalo de confiança dos métodos nos dois grupos. Os resultados não são inesperados, já que na maioria dos casos o Grupo 2 que possui amostragem maior de usuários apresenta erros menores. Em 6 dos 12 casos o erro é menor (intervalos de confiança não se sobrepoem), ocorre empate em 4 e apenas em 2 casos o erro do Grupo 1 é menor do que o erro do Grupo 2. Tais indícios corroboram com a intuição de que amostras maiores levam a erros menores.

### 5.3.2 Sensitividade ao tempo

Outro fator da base de dados que pode influenciar o resultado dos métodos de Inferência Ecológica é a janela de tempo utilizada. Como descrito na Seção 3.1, existe indícios na base de que a opinião pública tenha mudado durante os meses, devido à grande agitação de acontecimentos e discussões políticas nas redes sociais. Para analisar a temporalidade, a base foi dividida em 3 conjuntos Grupo 1 que possui apenas dados da primeira metade do tempo de coleta (25/11/2015 até 29/01/2016) , Grupo 2 que possui dados apenas da segunda metade do tempo de coleta (30/01/2016 até 25/03/2016) e Grupo 3 que agrega todos os meses de coleta para os cálculos do apoio em relação à

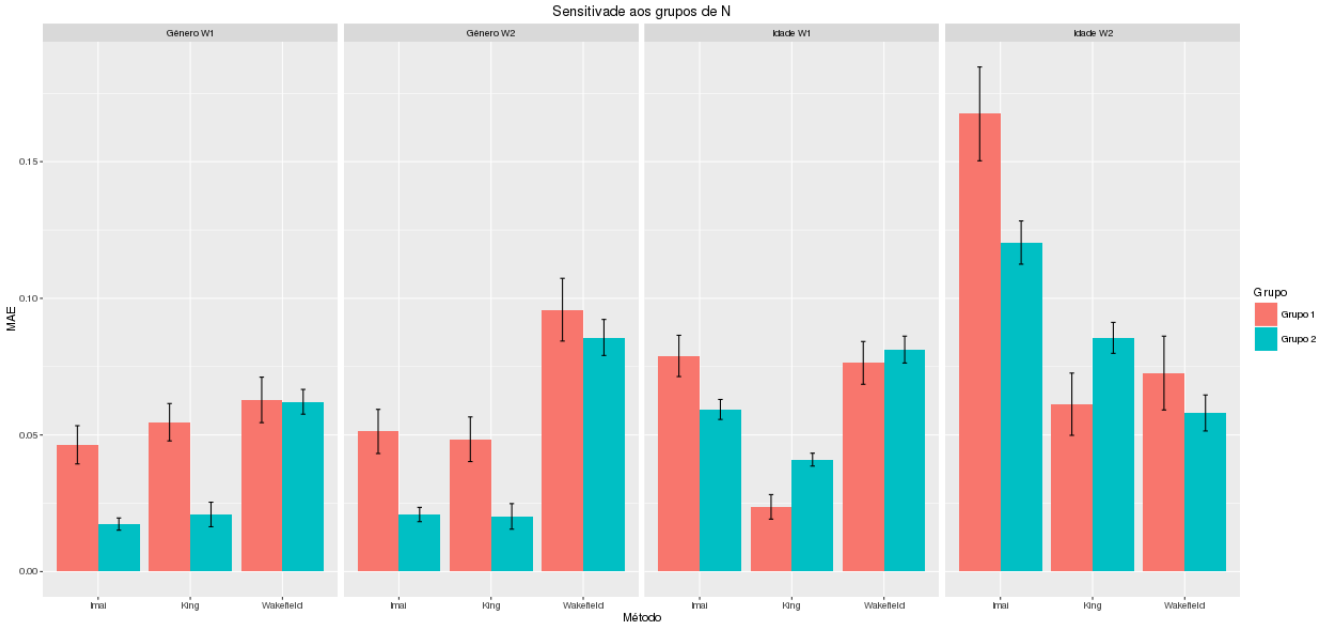


Figura 7 – Gráfico mostra o resultado da métrica MAE dos métodos para os dois grupos $N$<200 (Grupo 1) e $N$>200 (Grupo 2) em ambas bases de dados.

Dilma (25/11/2015 até 25/03/2016). A Figura 8 resume os resultados para os diferentes Grupos em relação a métrica MAE de W1 e de W2. Observando os gráficos podemos ver que variar a janela de tempo entre 2 e 4 meses, não traz resultados estatisticamente diferentes. Os resultados apontam que não existe diferença estatisticamente significativa entre o Grupo 3 e os outros dois Grupos. Em ambas variáveis e em ambas bases de dados os resultados para o Grupo 3 tem interseção do seu intervalo de confiança com pelo menos um outro Grupo, mostrando que durante a janela de 4 meses não ocorreu uma mudança na opinião dos usuários que influencie negativamente nos resultado do algoritmo de Inferência Ecológica ao agregar os dados.

## 5.4 Sensitividade em relação ao tipo de base de dados

Para verificar se os erros obtidos na base de dados sociais são diferentes dos erros obtidos nos *benchmarks* tradicionais de dados eleitorais, que utilizam pesquisas de rua como *ground-truth*, a Tabela 13 mostra os resultados do t-teste não pareado entre os erros dos algoritmos na base de dados sociais e na base de dados de registro *reg*, obtida no pacote eco[2] e disponibilizada por (KING, 1997). As variáveis para esta base tem a seguinte representação: $W1_i$ = porcentagem de pessoas negras que votaram, $W2_i$ = porcentagem de pessoas brancas que votaram, $Y_i$ = porcentagem de pessoas que votaram, $X_i$ = porcentagem de pessoas brancas na unidade geográfica $i$.

Os resultados mostram que para todos as bases de dados os erros são menores do

[2] http://imai.princeton.edu/software/files/eco.pdf

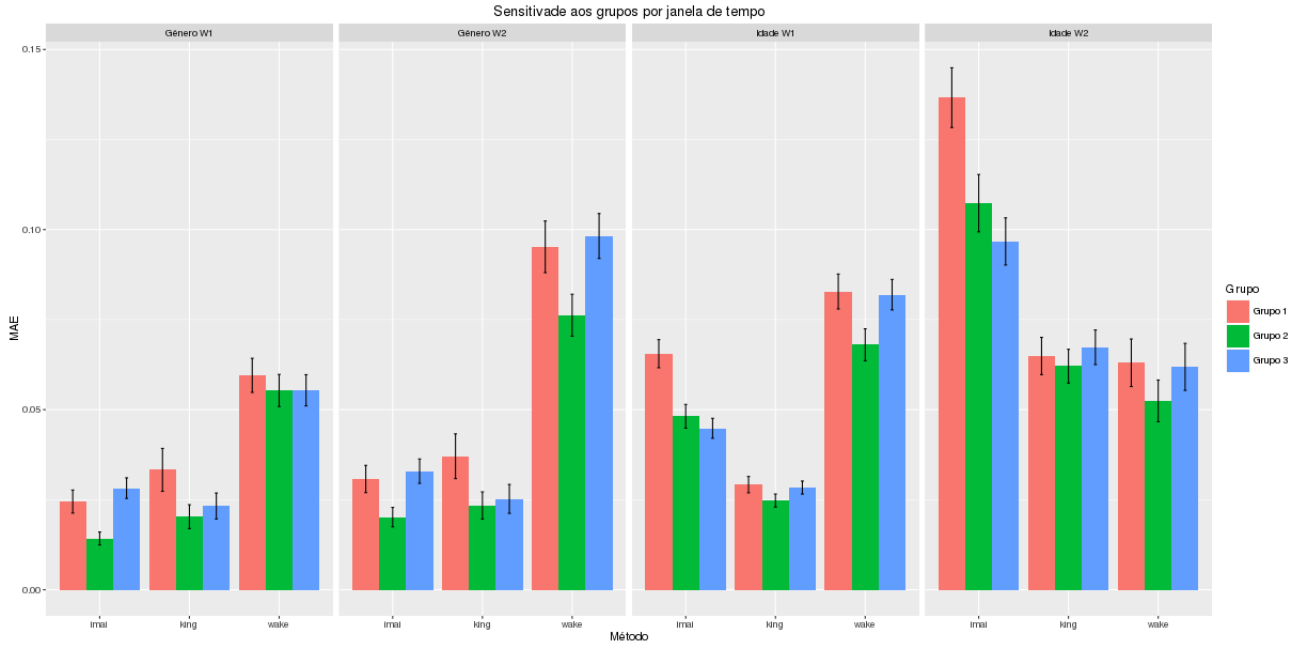


Figura 8 – Gráfico mostra o resultado da métrica MAE dos métodos para os três grupos (Grupo 1 = primeiros dois meses, Grupo 2 = últimos dois meses e Grupo 3 = agrega os 4 meses) em ambas bases de dados.

Tabela 13 – 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 médio | MAE W2 médio | MAE W1 médio | MAE W2 médio | 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 |

que quando utilizamos uma base de dados eleitoral, com uma alta confiança conseguimos refutar a hipótese nula de que são amostras iguais. Apesar disso, não podemos concluir que os algoritmos de IE apresentam melhores resultados para dados sociais de maneira geral, pois, na base de dados utilizada de apoio à Dilma, possuímos limites superiores e inferiores justos (como descrito na Seção 3.1.2 de caracterização da base de dados). Entretanto, na base eleitoral os limites são maiores, uma característica específica da base, como podemos observar na Figura 9, na qual as projeções das linhas nos eixos levam a limites bem maiores do que aqueles da Figura 4.

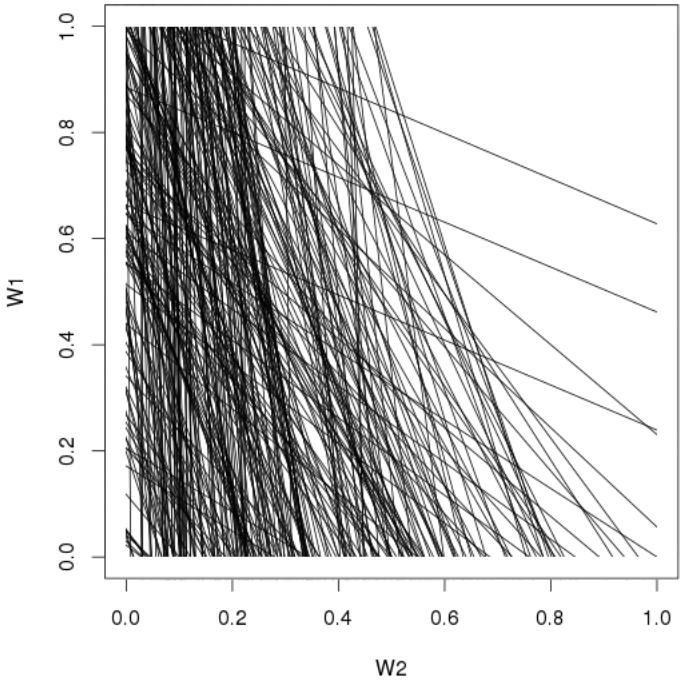


Figura 9 – Gráfico de tomografia para a base de dados *reg*.

# 6 Conclusão

Neste trabalhos abordamos o problema de inferir características demográficas de grupos de usuários em redes sociais a partir de métodos Inferência Ecológica. Até onde sabemos essa abordagem é inédita e apresenta vantagens sobre métodos de classificação existentes para inferir características dos usuários. Em primeiro lugar, os métodos de IE não requerem um conjunto de usuários com seus atributos rotulados (como gênero, idade e renda) para inferir sobre grupos. Eles precisam apenas de estatísticas agrupadas (ex: porcentagem de apoio em relação a um candidato eleitoral em uma unidade geolocalizada) e um censo (ex: IBGE). Além disso, uma vez que as inferências não são em nível individual mas sim para grupos de usuários, as inferências para uma grande quantidade de indivíduos são mais rápidas.

Após a coleta e caracterização da base de dados social, foi realizada uma extensa avaliação experimental comparando três algoritmos do estado-da-arte em Inferência Ecológica, mostrando que os melhores resultados são obtidos com o método de (KING, 1997). Além disso, vemos que os métodos são sensíveis à variável $N$, que indica a quantidade de usuários para a unidade geográfica $i$ (no caso cidade), na base de dados social e apresentam erros menores quando o grupo é de cidades com $N$ grande. Por outro lado, verificamos que os métodos não apresentaram diferença entre os erros ao agrupar os dados pelo tempo $t$ total de coleta e metade deste tempo $t/2$ (a queda na média do apoio em relação à Dilma nos últimos meses não afetou significativamente os resultados dos métodos de IE). Por fim observamos que apesar de os erros na base de dados social serem estatísticamente menores que os erros da base eleitoral *reg*, isso acontece devido aos limites dessa base em específico serem extramemente pequenos, gerando assim menos valores possíveis e fazendo com que seja mais fácil obter erros pequenos na base coletada.

Como trabalho futuro pretendemos abordar outros aspectos deste problema: utilizar um censo específico da internet em detrimento de um censo comum como o IBGE melhora o resultado dos métodos de IE? Como os algoritmos de IE se comparam com resultados agregrados de algoritmos de classificação em nível individual? Utilizar uma pequena amostra de registros individuais (ex: usuários de uma cidade $i$, com suas respectivas características demográficas e opinião em relação à Dilma) apresenta uma melhoria significativa nos resultados, como é reportado nas bases de dados tradicionais de IE?

# Referências


BI, B. et al. Inferring the demographics of search users: social data meets search queries. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. *Proceedings of the 22nd international conference on World Wide Web*. 2013. p. 131–140.

BLYTH, C. R. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, Taylor & Francis, v. 67, n. 338, p. 364–366, 1972.

BODENDORF, F.; KAISER, C. Detecting opinion leaders and trends in online social networks. In: ACM. *Proceedings of the 2nd ACM workshop on Social web search and mining*. 2009. p. 65–68.

CHO, W. K. T. Iff the assumption fits...: A comment on the king ecological inference solution. *Political Analysis*, SPM-PMSAPSA, v. 7, n. 1, p. 143–163, 1998.

CRAWFORD, C. A. G.; YOUNG, L. J. *A spatial view of the ecological inference problem.* : na, 2004.

DONG, Y. et al. Inferring user demographics and social strategies in mobile social networks. In: ACM. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014. p. 15–24.

DOUGNON, R. Y.; FOURNIER-VIGER, P.; NKAMBOU, R. Inferring user profiles in online social networks using a partial social graph. In: *Advances in Artificial Intelligence*. : Springer, 2015. p. 84–99.

DUNCAN, O. D.; DAVIS, B. An alternative to ecological correlation. American Sociological Review, 1953.

ELLIOT, P. et al. *Spatial epidemiology: methods and applications.* : Oxford University Press, 2000.

FILHO, R. M. *Um arcabouço para pesquisas de opinião em redes sociais*. Dissertação (Mestrado) — UFMG, Av. Antônio Carlos 6627 - Prédio do ICEx Pampulha CEP 31270-010 | Belo Horizonte - Minas Gerais - Brasil, 2003.

FLAXMAN, S. R.; WANG, Y.-X.; SMOLA, A. J. Who supported obama in 2012?: Ecological inference through distribution regression. In: ACM. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015. p. 289–298.

FREEDMAN, D. A. Ecological inference and the ecological fallacy. *International Encyclopedia of the social & Behavioral sciences*, v. 6, p. 4027–4030, 1999.

FREEDMAN, D. A. et al. On "solutions" to the ecological inference problem. *Journal of the American Statistical Association*, Citeseer, v. 93, n. 444, p. 1518–22, 1998.

GOODMAN, L. A. Some alternatives to ecological correlation. *American Journal of Sociology*, JSTOR, p. 610–625, 1959.

GREINER, D. J.; QUINN, K. M. R× c ecological inference: bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 172, n. 1, p. 67–81, 2009.

IMAI, K.; LU, Y.; STRAUSS, A. Bayesian and likelihood inference for 2× 2 ecological tables: an incomplete-data approach. *Political Analysis*, SPM-PMSAPSA, v. 16, n. 1, p. 41–69, 2008.

JACKSON, C.; BEST, N.; RICHARDSON, S. Improving ecological inference using individual-level data. *Statistics in medicine*, Wiley Online Library, v. 25, n. 12, p. 2136–2159, 2006.

JAIN, R. *The art of computer systems performance analysis*. : John Wiley & Sons, 2008.

KING, G. *A solution to the ecological inference problem*. : Princeton, NJ: Princeton University Press, 1997.

KING, G.; TANNER, M. A.; ROSEN, O. *Ecological inference: New methodological strategies*. : Cambridge University Press, 2004.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of cluster in k-means clustering. *International Journal*, v. 1, n. 6, p. 90–95, 2013.

LEWIS, J. B. *Extending King's ecological inference model to multiple elections using Markov Chain Monte Carlo*. : na, 2004.

MISLOVE, A. et al. Measurement and analysis of online social networks. In: ACM. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 2007. p. 29–42.

O'CONNOR, B. et al. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, v. 11, n. 122-129, p. 1–2, 2010.

O'LOUGHLIN, J. Can king's ecological inference method answer a social scientific puzzle: Who voted for the nazi party in weimar germany? Taylor & Francis, 2000.

PARK, W.-h.; HANMER, M. J.; BIGGERS, D. R. Ecological inference under unfavorable conditions: Straight and split-ticket voting in diverse settings and small samples. *Electoral Studies*, Elsevier, v. 36, p. 192–203, 2014.

PEARL, J. Understanding simpson's paradox. Citeseer, 2013.

QUINN, K. M. *Ecological inference in the presence of temporal dependence*. : na, 2004.

ROBINSON, W. Ecological correlations and the behavior of individuals. *American Sociological Review*, v. 15, n. 3, 1950.

TUMASJAN, A. et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, v. 10, p. 178–185, 2010.

WAKEFIELD, J. Ecological inference for 2× 2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 167, n. 3, p. 385–445, 2004.

ZAKI, M. J.; MEIRA, J. W. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. : Cambridge University Press, 2014. ISBN 9780521766333.

ZHANG, H.; KIM, G.; XING, E. P. Dynamic topic modeling for monitoring market competition from online text and image data. In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2015.

ZHONG, Y. et al. You are where you go: Inferring demographic attributes from location check-ins. In: ACM. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 2015. p. 295–304.