GOVERNO DO ESTADO DO CEARÁ
Secretaria do Planejamento e Gestão

CAPP INFORME

Nº 06 | **JULHO / 2020**

USO DA TECNOLOGIA DA INFORMAÇÃO EM POLÍTICAS PÚBLICAS

CENTRO DE ANÁLISE DE DADOS E AVALIAÇÃO DE POLÍTICAS PÚBLICAS-CAPP
INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ-IPECE

IPECE INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ

CAPP Informe – Nº 06 – Julho/2020

SETOR RESPONSÁVEL: Centro de Análise de Dados e Avaliação de Políticas Públicas – CAPP.

Autor(es):

Ricardo Brito Soares (CAEN/UFC)
Rafael B. Barbosa (FEAACS/UFC)
Paulo Barata (Universidade de Coimbra)
Isadora Osterno (UFC)

O **Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE)** é uma autarquia vinculada à Secretaria do Planejamento e Gestão do Estado do Ceará. Fundado em 14 de abril de 2003, o IPECE é o órgão do Governo responsável pela geração de estudos, pesquisas e informações socioeconômicas e geográficas que permitem a avaliação de programas e a elaboração de estratégias e políticas públicas para o desenvolvimento do Estado do Ceará.

Missão: Gerar e disseminar conhecimento e informações, subsidiar a formulação e avaliação de políticas públicas e assessorar o Governo nas decisões estratégicas, contribuindo para o desenvolvimento sustentável do Ceará.

Valores: Ética, transparência e impessoalidade; Autonomia Técnica; Rigor científico; Competência e comprometimento profissional; Cooperação interinstitucional; Compromisso com a sociedade; e Senso de equipe e valorização do ser humano.

Visão: Até 2025, ser uma instituição moderna e inovadora que tenha fortalecida sua contribuição nas decisões estratégicas do Governo.

Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) -
Av. Gal. Afonso Albuquerque Lima, s/n | Edifício SEPLAG | Térreo - Cambeba | Cep: 60.822-325 |
Fortaleza, Ceará, Brasil | Telefone: (85) 3101-3521
http://www.ipece.ce.gov.br/

Sobre **CAPP Informe**

A Série CAPP Informe, disponibilizada pelo Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE), visa divulgar análises técnicas, elaboradas pelos pesquisadores do Centro, sobre temas relevantes de forma objetiva. Com esse documento, o Instituto busca promover debates sobre assuntos de interesse da sociedade, de um modo geral, abrindo espaço para realização de futuros estudos.

Instituto de Pesquisa e Estratégia Econômica do Ceará – IPECE 2020

CAPP Informe / Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) / Fortaleza – Ceará: Ipece, 2020

1. Políticas Públicas. 2. Avaliação. 3. Estudos Sociais. 4. Ceará.

Nesta Edição - USO DA TECNOLOGIA DA INFORMAÇÃO EM POLÍTICAS PÚBLICAS

Este produto busca resumir as principais práticas de uso da tecnologia da informação para auxiliar políticas públicas. Primeiro, é realizada uma introdução a um tipo específico de uso da tecnologia da informação, as grandes bases de dados, ou *Big Data*. Posteriormente, é apresentado alguns exemplos de como o Big Data pode ajudar em diferentes aspectos o planejamento, a eficiência e o monitoramento de políticas públicas. São vistos exemplos nas áreas de planejamento urbano, educação, seleção de pessoal, antecipação de choques econômicos regionais, entre outros. A conclusão geral do produto é que apesar de avanços o Ceará pode se beneficiar significativamente se passar a adotar tais ferramentas na elaboração e na avaliação de suas políticas públicas locais.

USO DA TECNOLOGIA DA INFORMAÇÃO EM POLÍTICAS PÚBLICAS

Autores:
Ricardo Brito Soares (CAEN/UFC)
Rafael B. Barbosa (FEAACS/UFC)
Paulo Barata (Universidade de Coimbra)
Isadora Osterno (UFC)

Este produto é fruto do projeto financiado pela FUNCAP, vinculado ao Centro de Avaliação de Políticas Públicas (CAPP) – IPECE, sob o título: Análise da Eficiência de Gestão Pública.

RESUMO

Este produto busca resumir as principais práticas de uso da tecnologia da informação para auxiliar políticas públicas. Primeiro, é realizada uma introdução a um tipo específico de uso da tecnologia da informação, as grandes bases de dados, ou *Big Data*. Posteriormente, é apresentado alguns exemplos de como o Big Data pode ajudar em diferentes aspectos o planejamento, a eficiência e o monitoramento de políticas públicas. São vistos exemplos nas áreas de planejamento urbano, educação, seleção de pessoal, antecipação de choques econômicos regionais, entre outros. A conclusão geral do produto é que apesar de avanços o Ceará pode se beneficiar significativamente se passar a adotar tais ferramentas na elaboração e na avaliação de suas políticas públicas locais.

1. Introdução

O termo "Big Data" surge a partir da disponibilidade de grandes bases de dados em diferentes formas. Big Data gera oportunidades em negócios privados ou na melhoria da prestação dos serviços públicos. Entretanto, o uso de Big Data em geral requer novas formas de abordagens estatísticas, diferentes das abordagens padrões (EFFRON e HASTIE, 2018). Além disso, a aplicação de Big Data em políticas públicas requer um cuidado especial, principalmente para evitar problemas de discriminação por meio de algoritmos (KLEIBERG et al (2015), ATHEY e IMBENS (2018)).

Este trabalho tem dois objetivos principais. Primeiro, pretende-se realizar uma extensiva revisão das possibilidades de aplicação dos métodos de Big Data em políticas públicas. A ideia é buscar evidências de aplicações que possibilitem ganhos de eficiência na gestão pública, seja pela melhor focalização de políticas ou pelo melhor uso de dados para o planejamento da gestão fiscal. Em todo caso, serão analisadas pesquisas, com métodos rigorosos, que indiquem caminhos para possíveis aplicações.

Segundo, será realizada uma análise qualitativa dos desafios de lidar com métodos de Big Data. Tanto a definição do que é chamado de "Big Data" quanto os desafios técnicos para sua utilização ainda hoje são questões não inteiramente resolvidas na literatura. Portanto, este texto pretende marcar o nível das discussões em torno do tema na atualidade.

Este trabalho está dividido em mais cinco partes, além desta introdução e das conclusões finais. Na primeira parte, seção dois, será tratada a definição do termo "Big Data". Não há ainda uma definição consensual sobre este termo, por isso, será realizada uma ampla pesquisa sobre os conceitos disponíveis. Apesar de todas as classificações tratarem da mesma expressão, diferentes enfoques fazem com que este termo seja difícil de ser definido em um único conceito. Aqui, buscaremos identificar as diferentes abordagens e analisar qualitativamente suas diferenças. Importante notar que as diferentes definições por enfocarem aspectos diferentes associados ao termo "Big Data", também ressaltam múltiplas abordagens para a sua aplicação, destacando sua complexidade e variedade.

A seção 3 analisa questões relacionadas a dificuldade técnica de lidar com Big Data. Será visto que grandes bases de dados impõem séries restrições ao uso de métodos clássicos em estatísticas. Três principais características do Big Data causam estes problemas. Primeiro, Big Data geralmente não tem uma estrutura convencional de organização dos dados. Por exemplo, imagens

de satélite são utilizadas como dados em Big Data, porém, é difícil organizar uma imagem a partir da estruturação transversal, temporal ou em painel, tradicionais em estatística. É preciso realizar um processamento inicial antes de utilizar tais informações da forma como abordamos estatística.

Segundo, como grandes bases de dados possuem um grande número de observações, existem limitações computacionais de como lidar com tais banco de dados. Voltando ao exemplo da imagem como base de dados, uma vez realizado o processamento, cada imagem será composta de milhares de observações que a caracterizam. Essas observações exigem uma capacidade computacional superior ao que é exigido em bases de dados tradicionais em economia e administração, por exemplo. Formas de otimização computacional como computação em nuvem ou paralelização computacional são alternativas para reduzir o custo computacional ao lidar com Big Data.

Por fim, grandes bases de dados possuem um grande número de variáveis. Esse grande número de variáveis gerou o termo que ficou conhecido como "maldição da dimensionalidade". A maldição da dimensionalidade ocorre quando não sabemos a priori qual a qualidade preditiva de cada variável e a inclusão de variáveis com baixo poder preditivo pode fazer com que o desempenho dos modelos não seja otimizado [1]. Em certo sentido, a maldição da dimensionalidade pode ser traduzida como uma incerteza sobre a correta especificação de modelos estatísticos. Diversos métodos tem sido propostos para lidar com a elevada dimensionalidade das bases de dados em Big Data. Apesar da abordagem inicial ser inteiramente "guiada apenas por dados", alguns estudos mais recentes têm buscado uma estruturação para problemas específicos, recentemente aplicado em Inteligência Artificial (AI) (PEARL e MacKENZIE (2017)).

Um desafio importante, porém, não necessariamente técnico, refere-se a dificuldade de aplicação dos métodos de Big Data sem que estes se tornem enviesados ou gerem discriminação entre os indivíduos. Por exemplo, um algoritmo projetado para classificar os indivíduos segundo a sua produtividade entre 30 e 40 anos pode gerar viés ao atribuir maior peso para homens do que para mulheres, sendo que diferenças de gênero não causam, necessariamente, maior produtividade. Uma hipótese para este viés pode ser causada pelo fato de que nessa idade, as famílias começam a ter filhos e o peso da gravidez é muito maior para as mulheres do que para os homens. Se este algoritmo for utilizado para indicar qual trabalhador deve receber uma oportunidade de emprego

[1]De fato, quando o número de variáveis é maior que o número de observações disponíveis, o método dos Mínimos Quadrados Ordinários não apresenta solução única, ver Belloni et al (2014).

no futuro, então, ele tenderá a reproduzir as diferenças salariais entre homens e mulheres.

Este tipo de problema faz com que a aplicação de Big Data em políticas públicas seja muito mais delicada e exige um profundo conhecimento a priori do problema que se objetiva alcançar. Importante notar que está ainda é uma questão aberta e sujeita a inúmeras controvérsias na literatura especializada. Esse tipo de problema é discutido em mais detalhes ao longo do texto, ressaltando as técnicas que são adotadas para minimizá-lo.

A seção quatro apresenta várias possíveis aplicações de Big Data e suas técnicas estatísticas em políticas públicas. São analisadas seis aplicações. A primeira aplicação é baseada em Chalfin et al (2016) que analisa a utilidade de métodos de machine learning para selecionar professores e policiais. A ideia básica consiste em determinar uma característica desejável nesses profissionais, como capacidade de aumentar o aprendizado dos alunos, no caso dos professores, e deixar que modelos estatísticos selecionem qual o professor possui maior probabilidade de possuir tal característica no futuro. A aplicação destes métodos para selecionar pessoal deve ser realizada com cuidado, buscando sempre evitar que o algoritmo reproduza vieses contidos nos dados. Além disso, para que tal aplicação seja factível é necessário que se tenha uma medida clara e objetiva da característica do profissional a ser selecionado. Em várias profissões isso não é possível.

A segunda aplicação a ser analisada é o uso de Big Data no auxílio do planejamento urbano. Serão vistos quatro aplicações nesse contexto. Primeiro, será analisado como métodos de machine learning e Big Data podem ser utilizados para mensurar a riqueza das pessoas que vivem em determinadas ruas ou avenidas usando apenas o Google Street View. Essa aplicação é importante pois pode permitir, por exemplo, uma focalização maior da tributação sobre propriedade. Além disso, o conhecimento da riqueza das pessoas que vivem em determinadas ruas e avenidas pode ajudar no planejamento urbano. Segundo, será analisado como Big Data pode ser utilizado para mensurar a disponibilidade das pessoas pelo pagamento por serviços urbanos. Por exemplo, em uma rua mais escura, as pessoas podem ter uma disponibilidade maior para pagar por esses serviços públicos. Terceiro, será analisado tais métodos podem ser utilizados para mensurar o processo de gentrificação dentro das cidades. Em todos esses casos, Big Data permite uma análise mais granular das cidades e com maior frequência no tempo, ao contrário da abordagem atual em que pouco se conhece sobre as características de moradores em determinadas ruas de uma cidade e tal informação é obtida com um intervalo de tempo longo, como em um censo, por exemplo. Por fim, será visto como é possível utilizar dados não estruturados para melhorar os serviços públicos

urbanos. No caso, foi analisado como os comentários realizados no aplicativo Yelp podem ajudar a programar inspeções sanitárias.

A terceira aplicação analisada refere-se à criação de sistemas de alerta de risco. Alertas de risco são excelentes instrumentos de monitoramento de situações. Normalmente, tais sistemas são utilizados para monitorar situações que gerem risco de forma global. Por exemplo, várias cidades possuem sistemas de monitoramento de risco de enchentes ou de deslizamentos de encostas. Porém, com a elevada capacidade de armazenamento e processamento de dados, é possível agora realizar monitoramento direto de pessoas em situação de risco. O exemplo analisado aqui é o de estudantes com risco de evadir a escola. Entretanto, este sistema pode ser empregado em outras situações: monitoramento de pessoas com risco de saúde específico, monitoramento de problemas em ruas de grandes cidades, etc.

Posteriormente, serão vistas duas aplicações em finanças públicas. A ideia básica de tais aplicações é tornar mais eficiente o uso dos recursos públicos, seja pela melhor focalização de contribuintes ou pela melhor seleção de empresas que receberão benefícios tributários.

Por fim, será analisado como os métodos aplicados a Big Data podem ser utilizados para realizar previsões sobre o comportamento geral da economia. Essas previsões são importantes pois ajudam no planejamento das políticas públicas e na antecipação de cenários adversos que podem prejudicar a execução do gasto público. Diversos estudos tem sido realizados utilizando Big Data para prever da melhor forma possível dado o conjunto mais recente de informações, técnica conhecida como nowcasting. Diferentemente das aplicações anteriores em nowcasting, o uso de Big Data tende a realizar projeções mais acuradas, explorando inclusive diferentes fontes de dados.

A última seção discute algumas potenciais aplicações a economia cearense. São destacados pontos que poderiam ser relevantes para resolver problemas presentes na economia cearense. Importante notar que tais sugestões são iniciais no seu estágio de desenvolvimento, requerendo um aprofundamento maior da sua viabilidade. Além disso, por óbvio, tais sugestões não exaurem todas as possíveis aplicações de Big Data na economia cearense.

Espera-se que este texto possa servir como base para pesquisas mais profícuas sobre o tema, focalizando as suas reais aplicações no Ceará.

2. Definição de Big Data

Não existe atualmente uma definição consensual do que se possa chamar de Big Data. Entretanto, existem algumas características dos dados que possam ser considerados mais ou menos associados ao termo.

A IBM inicialmente classificou o termo Big Data a partir da classificação "4V". Os dados para serem considerados Big Data precisariam ter uma ou mais das seguintes características: i. Volume, implica um grande número de observações e de variáveis; ii. Velocidade, em economia tais tipos de dados são chamados de dados de alta-frequência, pois são coletados em períodos de tempo muito pequenos; iii. Variedade, incluem-se aqui dados com uma estrutura padrão, como dados em painel ou transversais, e dados sem estrutura definida, como textos de jornais, por exemplo, em que não se conhece a priori o que eles significam; iv. Veracidade, que se refere a qualidade da base de dados.

Esta classificação é bastante ampla pois inclui quase todo tipo de estrutura de dados existentes. Outras classificações mais estritas passaram a ser conceitualizadas. Os tipos de dados em Big Data são separados em duas categorias gerais: estruturados e não estruturados.

Dados estruturados são dados em que a priori se entende o que eles significam, mesmo que não se conheça suas observações. Um exemplo simples é uma base de dados que contém informações anuais do PIB dos países entre 1970 até 2010. Embora não se saiba que observação é registrada neste conjunto de dados, é possível saber o que este conjunto de dados significa. Isto é, o PIB é uma variável tecnicamente bem definida a priori.

Por sua vez, dados não estruturados são aqueles em que a priori não se sabe o que eles significam. Incluem-se nesta categoria textos, áudios, imagens, etc. Por exemplo, um texto jornalístico, como o utilizado por Bloom, Baker e Davis (2016), é um conjunto de palavras organizadas de uma forma sistemática que podem ter infinitos sentidos. Ou seja, apenas pela leitura do texto é possível entender como ele se classifica.

A partir desta diferenciação inicial é possível estabelecer subclassificações. Doornik e Hendry (2015) classificam os dados estruturados em três subcategorias: "Tall", "Fat" e "Huge".

Os dados são chamados de "Tall" quando possuem poucas variáveis (n), mas muitas observações (T). Isto é, são dados em que $T \gg n$. Este tipo de dado é bastante comum em finanças, em que várias transações por minuto são registradas ao longo do dia. Em economia estes dados são chamados de dados de alta-frequência.

Por sua vez, os dados são chamados de "Fat" quando existem muitas variáveis (n), mas poucas observações (T), isto é: $n \gg T$. Estes tipos de dados, também chamados de dados de alta-dimensão, são bastante comuns em economia, como em aplicações em análises regionais, uso de muitas variáveis instrumentais, análise em painel longitudinais ou empilhados, etc. Uma literatura importante está se desenvolvendo para analisar tais estrutura de dados, tanto para a finalidade de análise de causalidade, como em Belloni et al (2014, 2015, 2016) e Chernozhovukov et al (2018, 2019), quanto para o seu uso em previsões macroeconômicas, como Stock e Watson (2002, 2010, 2012), Kim e Swanson (2014, 2016) e Cheng e Hansen (2016).

Por fim, os dados são chamados de "Huge" quando possuem muitas variáveis e também muitas observações. Este tipo de dados é mais comum em industrias de tecnologia e ainda não são inteiramente analisados em economia. Em parte por que este tipo de dados requer uma análise de processamento muito maior que as demais estruturas de dados. Algumas poucas pesquisas tem a possibilidade de trabalhar com este tipo de dado, como: Cohen et al (2016) usando dados da UBER para estimar a curva de demanda, Bajari et al (2018) que utilizam dados da Amazon para prever vendas e Allcott et al (2019), que analisam a proliferação de notícias falsas por meio de redes sociais, como o Facebook e o Twitter. Um exemplo deste tipo de dado é as informações de pesquisa coletadas pelo Google. Existem vários itens pesquisados e várias formas de realizar a pesquisas. Isso faz com que o problema seja de alta-dimensão, com muitas variáveis. Por outro lado, as pesquisas no Google são realizadas com elevada frequência, fazendo com que cada variável tenha muitas observações.

Uma outra forma de classificar Big Data é proposta pela Comissão Europeia para as Nações Unidas (UNECE). Tal comissão subdividiu o termo Big Data em três tipos: Redes Sociais, Sistemas Tradicionais de Negócios e Internet das Coisas (IoT).

Na primeira categoria, Redes Sociais, encontram-se dados que são gerados pela experiência humana, quase sempre digitalmente registrados em computadores pessoais e redes sociais. Tais dados são em geral não estruturados e não estão sob a tutela e verificação dos governos, o que pode prejudicar a veracidade. São exemplos deste tipo de informação: registros em redes sociais como Facebook, Twitter, Instagran, comentários em Blogs, Vídeos no Youtube, Mapas utilizados por indivíduos, Emails, etc. Glaeser et al (2016), por exemplo, opiniões deixadas por usuários de restaurantes no Yelp para prever onde é melhor realizar uma inspeção sanitária.

Na segunda categoria, incluem-se dados que registram o comportamento dos indivíduos,

porém, tais dados não são diretamente gerados por eles. Assim, o acompanhamento de consumidores, estudantes nas escolas, pacientes na rede hospitalar são exemplos de dados desta categoria. Esses dados são geralmente estruturados e possuem um maior grau de veracidade, pois são verificados externamente antes da utilização, seja por governos ou por empresas. Estão dentro desta categoria dados administrativos e dados gerados por empresas.

Por fim, a terceira categoria inclui dados obtidos pela internet das coisas (IoT). Esses dados são gerados por sensores ou máquinas que registram a forma como os indivíduos utilizam e consomem bens. Por exemplo, o sensor de tempo de uso de internet nos celulares registra todas as atividades de internet no celular. Este tipo de dado geralmente é bem estruturado, porém, é coletado com elevada frequência e possui um grande número de variáveis, tornando o seu uso ainda limitado.

Como se percebe, a definição do que é Big Data ainda é incerta e não há consenso sobre qual a melhor forma de classificar os dados. Espera-se que com o tempo surjam áreas que reclassifiquem cada uma dessas categorias em subcategorias e que permitirão identificar com mais precisão quando se está tratando de Big Data ou de outras estruturas de dados mais convencionais.

3. Dificuldades estatísticas para lidar com Big Data

Os métodos estatísticos clássicos, como mínimos quadrados ordinários ou máxima verossimilhança, foram desenvolvidos levando em consideração as estruturas de base de dados que são utilizadas em Big Data. (HASTIE e EFFRON (2017)). Existem várias limitações, de ordem técnica ou conceitual, que fazem com que a aplicação da estatística clássica seja limitada no uso de Big Data.

Nessa seção será enumerado alguns problemas que dificultam a aplicação dos métodos estatísticos clássicos. Adicionalmente, discutiremos a evolução das técnicas estatísticas por meio do surgimento dos métodos de machine learning.

3.1 Dificuldades estatísticas com grandes bases de dados

3.1.1 Grande número de variáveis

Uma das características mais importantes associadas ao conceito de Big Data é a existência de um número muito grande de variáveis por observação. Esta característica cria algumas dificuldades de aplicação da econometria clássica.

O primeiro problema associado a grande dimensão dos dados ocorre quando o número de variáveis é superior ao número de observações disponíveis. Neste caso, a aplicação do método dos mínimos quadrados se torna infactível pois a solução dos parâmetros estimados deixa de ser única (BELLONI et al (2014)).

Esse problema não seria relevante se não for recorrente, entretanto, existem diversas situações práticas que tal situação ocorre. Por exemplo, análises de problemas regionais são bastante comuns em economia: cálculo de multiplicadores de gastos locais (Corbi et al (2019), Nakamura e Steinsson (2014), Chodorow-Reich (2019)), análise de efeitos choques nacionais sobre entes subnacionais (Autor et al (2017)), comparação de políticas específicas de determinadas regiões (Donohue and Levitt (2001), Abadie, Diamond and Hainmueller (2010)), entre muitos outros exemplos.

Em todos esses casos, geralmente existe um painel com unidades transversais estaduais (ou municipais)) variando ao longo de alguns anos (ou meses). Uma característica comum neste tipo de aplicação é a existência de efeitos indiretos (ou spillovers). Por exemplo, considere que se queira verificar o efeito do gasto público em determinado estado sobre a atividade econômica estadual, como em Nakamura e Steinsson (2014). O aumento dos gastos de determinado estado pode gerar um impacto indireto em outros estados. Com isso isolar o efeito causal nestes casos é bastante difícil (Ver Chodorow-Reich (2019) para uma discussão mais aprofundada).

Uma forma simples de resolver este problema consiste em incluir variáveis de efeitos fixos que controlem esse impacto indireto. Todavia, em algumas situações, a inclusão de efeitos fixos estaduais é infactível pois eleva significativamente o número de variáveis [2]. Este problema é ainda mais relevante quando o efeito indireto varia temporal e transversalmente e é necessário a introdução de efeitos fixos interativos, como em Gobilon e Magnac (2016) e Bai (2009) [3].

Logo, em tais situações pode ocorrer o problema de se ter mais variáveis do que observações e tornar estimativas por MQO infactíveis. Alguns trabalhos recentes tem buscado alternativas usando técnicas de machine learning para tornar factíveis tais estimativas (BELLONI et al (2016), BAI (2009), ABADIE e KASI (2019)).

O segundo tipo de problema que surge devido ao grande número de variáveis é relacionado

[2]A introdução de efeitos fixos normalmente ocorre pela adição de variáveis binárias indicando o estado que está sendo analisado.

[3]Por exemplo, no caso de um painel com 27 estados variando temporalmente entre 1994 a 2014, isto implica a adição de $27 \times 21 = 567$variáveis adicionais.

a indefinição da correta forma funcional dos modelos. Este tipo de problema geralmente ocorre em situações cujo o objetivo é realizar previsões, e não acessar o efeito causal. Neste caso, é comum a existência de um grande número de observações e variáveis, porém, se desconhece a forma funcional e relevância de cada variável para prever a variável de interesse.

Exemplificando: suponha que se queira prever o PIB do Brasil no tempo t. Existem inúmeras variáveis macroeconômicas que podem ser utilizadas para realizar essa previsão. Diversos trabalhos utilizam mais de cem variáveis macroeconômicas para prever o PIB (ver Stock e Watson (2002), Ng e McCracken (2016)). Apesar de todas essas variáveis serem relevantes para prever o PIB, não se sabe a priori qual variável, ou conjunto de variáveis, possui maior poder de previsão. Este fenômeno é chamado de maldição da dimensionalidade.

Métodos clássicos de séries temporais, como a metodologia Box-Jenkins, não conseguem lidar adequadamente com essa incerteza quanto à forma funcional do modelo previsor. Recentemente, várias alternativas tem surgindo para lidar com esse problema, como: uso de métodos fatoriais (Stock e Watson (2002)), métodos de machine learning (Swanson e Kim (2014)), métodos ponderados (Hansen e Racine (2012)), entre outros.

Em ambos os casos, o surgimento de um grande número de variáveis pode não ser ideal seja pela incerteza quanto à forma funcional ou pela completa infactibilidade dos métodos estatísticos convencionais.

3.1.2 Grande número de observações

Geralmente, a primeira associação ao termo Big Data refere-se ao grande número de observações. Esta situação é muitas das vezes benéfica, pois permite analisar problemas com características mais de populações e não de amostras. Todavia, um número muito grande de observações pode elevar significativamente o tempo de computação das estimativas.

Em economia, o grande número de observações pode ser um limitador importante para a análise de certos problemas. Macroeconomistas precisam resolver modelos complexos e simultaneamente associar observações micro e macroeconômicas. Em economia internacional, alguns modelos utilizam inúmeros setores com vários registros de transações ao longo do tempo. Em finanças, existem várias transações realizadas diariamente.

Apesar de os computadores modernos lidarem com diversas dessas situações, muitas vezes é necessário recorrer a alguma técnica de computação paralela. Basicamente, a computação

paralela significa separar a análise de um processo complexo em vários pedaços mais simples e depois reconectá-los por meio de algum procedimento.

Para ilustrar considere a estimação de mínimos quadrados ordinários. Seja $Y = (y_1, y_2, \ldots, y_n)$uma variável de resultado e $X = (x_{i1}, x_{i2}, \ldots, x_{ik})$uma matriz contendo todas as variáveis de controle, para $i = 1, \ldots, n$. Assumindo o modelo linear, a estimação de mínimos quadrados ordinários é:

$$\beta = (X'X)^{-1}(X'Y)$$

Em que: βé um vetor de tamanho $1 \times k$. Para calcular βé necessário realizar a inversão da matriz $(X'X)$. Se $n \to \infty$, o calculo desta inversão será grande. Uma forma de realizar este procedimento é paralelizar a computação ao se separar em várias partes a inversão desta matriz.

Villaverde e Valencia (2018) apresentam um guia prático para a parelização computacional aplicada a economia.

3.1.3 Dados não estruturados

Uma das características de Big Data consiste no uso de dados sem uma estrutura convencionalmente utilizada em estatística e econometria. Tais dados podem ser obtidos de várias fontes: som, imagens, textos, etc.

A análise deste tipo de dado é dividida em duas partes. Para facilitar a exposição, será considerado a abordagem de análise textual. A primeira parte consiste na extração dos dados. A forma de extração é dependente de fonte de dado que se está extraindo. Por exemplo, na utilização de textos como fonte de dados pressupõe a aplicação de alguma técnica que separe e tipifiquem as partes do texto, chamadas de tokens, para ser posteriormente analisados.

Esta etapa possibilita a limpeza de estruturas textuais sem relevância interpretativa, como artigos (“a”, “o”, “um”, “umas”), preposições (“de”, “para”, etc). A forma mais comum de representar estruturas textuais é por meio da técnica bag-of-words em que a ordem das palavras é ignorada. Nesse caso, cada token é separado e tipificado. O que importa neste caso é ocorrência de determinada palavra em um documento, independente da ordem ou sentido que ela apareça.

Uma vez que o documento tenha os dados textuais extraídos e classificados é possível realizar diversos usos, como previsão e estimação causal. Alguns exemplos: Baker, Bloom e Davis (2016) utilizaram análise textual de jornais para criar um índice para incerteza econômica. A incerteza é caracterizada como uma situação em que os agentes econômicos não conseguem

projetar o comportamento futuro das economias. Dessa forma, a tomada de decisão que envolva retornos futuros é comprometida, pois, se torna muito arriscada. Bloom (2014) revisa a literatura sobre os efeitos da incerteza macroeconômica.

A hipótese de Baker, Bloom e Davis (2016) é que a incerteza é uma situação difícil de analisada de forma sistemática por qualquer pessoa, porém, jornais impressos conseguem acompanhar vários fatores que geram instabilidade futura na economia e possuem a função de transmitir, de forma clara e objetiva, a indicação de incerteza futura.

Portanto, Baker, Bloom e Davis (2016) extraem de notícias de jornais impresso a frequência com que palavras associadas a incerteza aparecem. Por exemplo, palavras como "incerteza", "incerteza macroeconômica", "incerteza, economia", etc. Quando tais palavras ocorrem com muita frequência é uma indicação de que os agentes econômicos demandadores do jornal irão julgar o período como de maior incerteza.

A partir desta análise os autores constroem um índice de incerteza que pode ser utilizado para investigar a como variações exógenas na incerteza afetam a atividade econômica.

Como se depreende, este tipo de dado não possui uma estrutura semelhante ao que é utilizado convencionalmente em economia ou estatística. A principal característica de um dado não estruturado é o entendimento do seu significado. Por exemplo, uma variável que mensura o salário de um indivíduo indica o quanto aquele indivíduo recebe como remuneração por seu trabalho. Por outro lado, um texto de jornal pode apresentar uma frequência de palavras associadas a incerteza, mas com sentidos opostos. Por exemplo, uma discussão entre dois especialistas um pode apontar a presença de incerteza e outro indicar justamente o contrário.

Outros tipos de dados não estruturados também possuem essa característica de que o seu significado não é completamente claro. Por exemplo, vídeos podem ser utilizado como fonte de dados, porém, seu significado não é claro.

A econometria não foi inicialmente elaborada para lidar com esses tipos de dados. Gentzkow, Kelly e Taddy (2019) realizam um survey das principais técnicas para lidar com dados textuais em economia.

4. Exemplos de usos de Big Data em políticas públicas

Esta seção discute alguns pormenorizadamente alguns exemplos de como Big Data pode ser empregado para auxiliar a tomada de decisão em políticas públicas. Serão analisadas seis aplicações possíveis: seleção de pessoal, monitoramento urbano, alerta de situações, aplicações em finanças públicas, nowcasting e focalização de políticas públicas. Todos os exemplos são baseados em análises estatísticas rigorosas, portanto, estão excluídos os exemplos puramente qualitativos ou especulativos.

4.1 Seleção de pessoal

Um dos grandes desafios da tomada de decisão pública é a contratação dos servidores públicos. No Brasil a contratação muitas vezes é realizada por meio de exames técnicos cuja habilidades avaliadas estão associadas a aspectos cognitivos necessários para a execução dos serviços públicos, como por exemplo concursos públicos. Entretanto, habilidades cognitivas demostradas em exames não estão necessariamente correlacionadas a qualidade do serviço público.

Análise de Big Data pode ser direcionada para criação à priori de indicadores que ajudem a classificar os concorrentes a cargos públicos em características relacionadas diretamente a qualidade do serviço público. Ou seja, uma vez definido qual indicador está associado a qualidade do serviço público é possível usando análises de dados classificar os concorrentes segundo suas características anteriores a contratação.

Tem sido verificado como métodos de análise de dados podem auxiliar na contratação de pessoas para o serviço público (Chalfin et al, 2016) ou privado (Erel et al, 2018). Em ambos os casos é necessário a definição de um indicador de qualidade do serviço a ser prestado. Aqui o foco será na contratação de pessoal para o serviço público.

Chalfin et al (2016) analisa dois tipos de contratações de servidores públicos: policiais e professores do ensino fundamental nos EUA. No caso dos policiais foi considerado como indicador de baixa qualidade o uso excessivo de força. Por sua vez, como indicador de qualidade dos professores foi utilizado o valor adicionado do professor.

Em ambos os casos, tais indicadores são importantes para a prestação do serviço público, todavia, ambos não são observados a priori. Além disso, a correlação entre sucesso em exames cognitivos e tais indicadores é baixa, isto é, um professor pode ter um elevado conhecimento em matemática, porém ter dificuldades em motivar a a classe, de forma que o valor adicionado futuro

seja baixo. A ideia, portanto, consiste em utilizar Big Data para classificar tais características ideais ex-ante. Este tipo de problema se encaixa no conceito de problemas de previsão de política como definido por Kleinberg et al (2015).

O objetivo dos dois exercícios é mensurar qual a melhoria potencial do serviço público prestado se fosse possível substituir agentes públicos selecionados tradicionalmente por agentes públicos selecionados por análise de dados. O uso de análise de dados tem duas vantagens sobre as formas tradicionais de seleção. Primeiro, é possível gerar índices de risco de situações reais considerando características observáveis. Isto é, é possível ex-ante identificar indivíduos que apresentam padrões semelhantes aos agentes públicos que possuem indicadores de interesse.

Segundo, tais indicadores de interesse são difíceis de serem classificados por exames cognitivos e muitas vezes não há uma clara indicação de como eles podem ser inferidos. Por exemplo, existe uma longa e já consolidada literatura de construção de indicadores para habilidades cognitivas que é extensivamente utilizada para realização de concursos públicos. Caso diferente é o do valor adicionado dos professores. Não há técnicas já testadas e protocolos de validação de testes para identificar características que ajudem a antecipar o valor adicionado dos professores. Nesse sentido, métodos de análise de dados se adaptam melhor a tais situações.

Apesar de ser promissora a possibilidade de empregar métodos de análise de dados na contratação do serviço público, duas ressalvas devem ser pontuadas. Primeiro, não se está sugerindo que métodos de Big Data substituam esquemas de avaliações tradicionais. É preciso ainda bastante amadurecimento destas técnicas para a sua generalização e utilização como único critério de seleção de funcionários públicos.

Um problema associado a Análises de Big Data com o uso de Machine Learning é a dificuldade de entender por que os modelos preveem com qualidade. Métodos como Deep Learning, por exemplo, apesar de gerarem excelentes previsões quando comparados a outros modelos, não é possível compreender quais os motivos que o levaram a prever melhor.

No caso específico da contratação de pessoal, isto significa que os indivíduos serão classificados segundo suas características anteriores a contratação, porém, não há como explicar o porquê do ranqueamento. Isso decorre do fato de que os métodos de análise de dados aplicados a Big Data são desenvolvidos sem fundamentação teórica que justifique as previsões. Apenas critérios estatísticos de qualidade da previsão são utilizados.

O segundo problema deriva da dependência que tais modelos estatísticos possuem dos

dados anteriores. Uma vez que informações passadas são utilizadas para realizar a previsão, então, padrões sociais caracterizados nos dados devem ser replicados nas previsões. Isso implica que tais modelos tendem a reproduzir situações sociais presentes nos dados, mesmo que elas sejam socialmente indesejáveis.

Por exemplo, considere que a maior parte dos professores que possuem elevado valor adicionado são pessoas brancas. Este resultado pode estar associado a desigualdade de acesso a educação entre brancos e pretos. Ou seja, o fato de ter mais professores brancos com elevado valor adicionado pode decorrer da existência de mais professores brancos e do acesso a melhores escolas durante a sua formação.

Todavia, se características da cor da pele forem utilizadas no modelo para prever o valor adicionado dos professores, pode ser que o modelo atribua um peso muito grande a esta característica e classifique erradamente professores pretos, simplesmente pelo fato de eles serem pretos.

Entretanto, a cor da pele de indivíduo não está associada ao potencial que ele possui para ter um elevado valor adicionado. Ou seja, se as pessoas pretas tivessem igual acesso a escolas de qualidade, a cor da pele não seria importante para prever o valor adicionado. Este é um dos grandes dilemas do uso de métodos de análise de dados na sociedade (Athey e Imbens (2018), Kleinberg et al (2017)).

Nesse sentido, modelos de análise de dados devem ser utilizados com cuidado pois podem reproduzir situações sociais indesejáveis presente nos dados no passado e trazê-los para o futuro. No caso da seleção de professores, se os critérios de análise de dados fossem aplicados para classificar os indivíduos na situação hipotética considerada, professores negros teriam maior dificuldade de serem contratados.

No entanto, métodos de análise de dados podem ser úteis em informar sobre quais características prévias estão associadas a indicadores de qualidade do serviço público e com isso gerar informações sobre os contratados que transcende a classificação por via puramente cognitiva.

Essa ferramenta poderia ser utilizada posteriormente a contratação. Por exemplo, ainda focando no caso da contratação do professor, suponha que o único critério de escolha utilizado para contratar foi concurso público tradicional. Métodos de análise de dados podem ser aplicados para classificar os professores contratados de acordo com a previsão do seu valor adicionado futuro e professores com classificação baixa podem ser submetidos, durante o período de estágio

probatório, a treinamentos que elevem o valor adicionado futuro.

Chalfin et al (2016) consideraram situação semelhante. Eles testaram se métodos de machine learning poderiam ajudar na decisão de qual professor estaria apto a concluir o estágio probatório. Usando dados do projeto Measures of Effective Teaching (MET), criado pela Fundação Bill e Melinda Gates em 2013, os autores correlacionaram o valor adicionado dos professores em matemática e em linguagem [4] do ensino fundamental medido em 2011 a diversas características dos professores (fatores socioeconômicos e demográficos, observações em sala de aula), características dos estudantes (performance em testes padronizados, fatores socioeconômicos e demográficos) e características dos diretores das escolas (pesquisa sobre os professores e sobre a escola).

Foram analisados 664 professores de matemática e 707 professores de linguagens. Os autores utilizaram apenas um método para realizar a previsão: método do LASSO, desenvolvido por Tibsharani et al (1996). O método do LASSO é adequado a este contexto pois ele realiza uma seleção dos melhores preditores, conhecido como regularização. Assim, em situações em que o número de variáveis é grande relativamente ao número de observações, o método do LASSO tende a ter um bom desempenho (Ver Hastie el al (2015) e Belloni et al (2014, 2016)).

Após a realização da previsão os autores buscaram simular qual seria o ganho em termos de valor adicionado para os alunos caso os 10% piores professores ranqueados fossem substituídos por professores pelos professores com valor adicionado médio.

Se fosse possível realizar tal substituição, o corpo docente passaria a ser composto por professores que gerariam no futuro maior valor adicionado previsto. Essa simulação fez com que o ganho dos alunos em matemática fosse de 0.0072σem matemática e 0.0057σem linguagem (ELA). Importante notar que estes ganhos referem-se a ganhos sobre todos os alunos. Especificadamente sobre os estudantes que tiveram seus professores substituídos os ganhos são 10 vezes maiores.

Os autores compararam se tais resultados são custo-efetivos. Comparando com a redução da quantidade de alunos na sala de aula eles concluíram que a substituição de professores seria duas ou três vezes mais efetiva que reduzir a sala de aula em um terço.

O segundo exercício consiste em analisar se métodos de análise de dados podem ser

[4]Na verdade, no caso americano os professores lecionavam inglês e linguagens artísticas (ELA, acrônimo em inglês).

utilizados para prever bom comportamento de policiais no futuro. Foram usados dados do Departamento de Polícia da Filadélfia para 1949 policiais contratados e matriculados em 17 academias de polícia nas turmas de 1991-1998.

A variável de interesse a ser prevista é uma variável binária que indica se os policiais se envolveram abusos físicos e verbais ou em situações com uso de arma de fogo policial. Os preditores utilizados foram fatores demográficos [5], status de veterano de guerra, status de casado, pesquisa que visa capturar o comportamento prévio e outros temas como se já foi preso, teve licença de motorista cassada, entre outros.

Foi utilizado o método de validação cruzada 5-fold para evitar problemas com overfitting e o método de previsão utilizado foi o Stochastic Gradient Boosting (SGB). Este método realiza um aprendizado estocástico buscando otimizar a função perda de previsão. Para mais detalhes deste método ver Goodfellow et al (2018).

O exercício de simulação semelhante foi realizado. Isto é, foi substituído os 10% piores policiais classificados pelo método de Machine Learning e substituídos pelo policial médio. Os resultados indicam uma redução de 4.81% no número de situações que envolvam uso de armas de fogo por policiais. Estes resultados foram similares a abuso físico e verbal por parte dos policiais.

Um problema com essas previsões dos policiais é que elas são realizadas em situações em que os profissionais já foram contratados. Isso pode enviesar as previsões gerando classificações errôneas, o que os autores chamaram de task cofounding. Se policiais são enviados para regiões com maior criminalidade, isso pode aumentar a probabilidade de se envolverem em uso de armas de fogo. Este problema tende a não ser relevante na análise dos professores.

Os autores apresentam outras lições mais gerais destes exercícios. Primeiro, que métodos de ML possuem bom desempenho em prever indicadores de qualidade do serviço público melhor que métodos tradicionais. Ou seja, tais métodos podem ser utilizados para classificar potenciais funcionários publicos de acordo com um critério específico.

Segundo, é preciso ter cuidado com o que os autores chamam de viés de payoff omitido. Esse viés decorre do fato de que funcionários públicos, assim como outras ocupações, não possuem um indicador de interesse único. Métodos de ML, até agora, foram desenvolvidos para obter a melhor previsão para uma única variável alvo.

Este viés é um problema, por exemplo, no exercício para a seleção de professores. Apesar

[5]Importante: não foram utilizados atributos raciais neste exercício.

de ganhos em termos de aprendizagem, chamados de valor adicionado, sejam importantes para políticas públicas, outras variáveis de resultado podem ter um valor maior. Por exemplo, taxa de abandono é geralmente mais relevante para políticas públicas em países em desenvolvimento do que ganhos em termos de aprendizado (Urquiola e ..., 2018).

Dessa forma, selecionar professores com maior valor adicionado previsto pode aumentar a taxa de abandono por professor, uma vez que uma das causas do abandono é a falta de habilidades cognitivas prévias. Professores com elevado valor adicionado podem ser mais rígidos e com isso deixar para trás estudantes com baixas habilidades cognitivas [6].

4.2 Aplicações em economia urbana

Um dos aspectos mais importantes associados ao termo Big Data é a utilização de novas fontes de dados. Essas novas fontes de dados possuem diferentes origens e características, entretanto, a grande novidade é que estas informações são cada vez mais granulares, com uma frequência cada vez maior e são georreferenciadas. Essas novas bases de dados aumentam o potencial de aplicações, especialmente, no acompanhamento e monitoramento das cidades. Tais dados ainda são subutilizados, porém, alguns autores têm demostrado como podem ser utilizados para estudar e melhorar as funções das cidades.

Aqui serão analisadas quatro aplicações dessas novas bases de dados em problemas reais enfrentados no planejamento urbano. Primeiro, será visto como formas alternativas de dados podem ser utilizadas para mensurar a riqueza de lugares específicos dentro de uma cidade, como ruas, por exemplo. Segundo, será visto como novas fontes de dados podem ajudar a identificar a disponibilidade por pagar por certos serviços urbanos. Posteriormente, será analisado como tais base de dados podem ajudar a identificar fenômenos urbanos importantes para o planejamento de cidades, como é o caso da gentrificação. Por fim, será apresentado como tais bases de dados podem ajudar no provimento de serviços públicos.

Novas formas de mensuração de dados tem um enorme potencial para melhorar as análises das ciências urbanas. Apesar de tais dados não ajudarem a resolver problemas de identificação causal (ATHEY e IMBENS (2018), MULHANATHAN e SPISSES (2017)), eles enfrentam problemas clássicos para as cidades como: valoração de amenidades e políticas, mensuração de

[6]Estes resultados e verificado por exemplo em Barbosa et al (2019), em que foi verificado que escolas de melhor no ensino médio possuem efeito menor em estudantes com baixas habilidades cognitivas prévias.

características urbanas, entre outras (GLAESER et al, 2018).

A grande característica associada a tais dados é a possibilidade de mensurar com uma frequência e granularidades cada vez maiores características urbanas de forma georreferenciada. Isso ajuda a melhorar a precisão das mensurações em geral, o que pode auxiliar na tomada de decisões ex-ante.

4.2.1 Medindo a riqueza de ruas por meio do Google Street View

É possível mensurar a riqueza de uma população em áreas urbanas pequenas, como por exemplo, ruas? Não existem estatísticas oficiais para essa finalidade, pois o custo de coleta de informações granulares é muito alto. Governos realizam com maior frequência pesquisas amostrais [7] Alguns pesquisadores têm se dedicado a buscar formas mais granulares de medidas urbanas (NAIK et al (2014, 2015, 2016), GLAESER et al (2018), incluir outros nomes como satélite para prever pobreza, dados telefônicos para prever pobreza).

Em particular, Glaeser et al (2018) mostram como utilizar imagens obtidas do Google Street View para prever a riqueza das ruas de Nova York. Google Street View é uma iniciativa do Google que disponibiliza fotografias de ruas em mais de 100 países ao redor do mundo. Os autores utilizaram imagens panorâmicas de 360g capturadas entre 2007 e 2014 para a cidade de Nova York. Tais imagens estão georreferenciadas por latitude e longitude. Por cada quadra de Nova York foram obtidas 40 imagens, totalizando 12.200 imagens de Nova York, uma cobertura de 2439 quadras.

Para treinar o modelo os autores associam cada uma das imagens aos dados de renda mediana familiar por quadra obtidas de forma censitária entre 2006 e 2010 pela American Community Survey (ACS). Para verificar se o poder preditivo das imagens os autores utilizaram o mesmo procedimento de Naik et al (2014). Primeiro, identificaram a cada pixel rótulos semânticos. Quatro rótulos foram utilizados: chão, prédios, árvores e céu. Posteriormente, foram sendo extraídos cada uma série de características associadas a cada um dos pixels de cada imagem para cada uma das categorias definidas. Ao final cada imagem foi representada por 7480 representações.

Com as imagens transformadas em dados os autores utilizaram métodos de machine learning da família dos supporte vector regression (SVR) para prever a renda mediana. Essa medida de riqueza foi obtida ao se calcular a mediana das rendas de cada quadra cujos dados eram

[7] No Brasil existe a Pesquisa Nacional por Amostragem Domiciliar Contínua (PNAD - Contínua), realizada trimestralmente.

disponíveis. Um procedimento de validação cruzada foi utilizado para evitar a possibilidade de sobre-estimação.

Por fim, os autores compararam em um exercício teste de previsões que tipo de variável prevê melhor: variáveis educacionais e raciais ou imagens do Google Street View. O resultado é que o modelo contendo variáveis como raça e educação obteve um R^2de 0.77. Por sua vez, o modelo utilizando as imagens obteve um R^2de 0.81. Em um teste de hipóteses foi verificado que as imagens de fato preveem melhor a renda mediana do que as variáveis educação e raça.

O resultado é robusto a outras regiões. Os autores testaram a mesma estratégia e encontraram resultados similares para a cidade de Boston.

Estes resultados são uma amostra do potencial de como novas bases de dados podem ser utilizadas para prever características urbanas, no caso a renda mediana. Uma vez que este modelo prevê adequadamente num ponto do tempo, é possível utilizá-lo para o acompanhamento da evolução urbana das cidades. Uma outra aplicação possível é na melhor mensuração de características urbanas para fins de tributação predial, por exemplo.

4.2.2 Medindo a disponibilidade de pagamento por serviços urbanos.

Uma forma bastante utilizada para mensurar a disponibilidade de pagamento por serviços urbanos consiste no uso de preços hedônicos. Preços hedônicos refletem o valor que os bens possuem dada as suas características. Dentre essas características estão a presença de serviços e características urbanas.

Preços hedônicos são utilizados para acessar a disponibilidade de pagamento por serviços urbanos. Por exemplo, suponha que determinado bairro tenha um aumento na quantidade de crimes. Esse aumento pode alterar os preços hedônicos do bairro de forma que os agentes econômicos passem a pagar menos por viver em um bairro com maior violência. Essa variação pode ser utilizada par computar o quanto os agentes econômicos estão dispostos por aceitar mais ou menos violência.

Logicamente que o uso de preços hedônicos tem uma série de problemas. Primeiro, a hipótese principal é que tais preços refletiram o equilíbrio de demanda e oferta por imóveis e que variações nesses preços de equilíbrio representarão variações na disponibilidade de pagamento. Esse equilíbrio não necessariamente é alcançado no curto prazo. Segundo, que é difícil associar de forma clara qual característica urbana interferiu no preço hedônico devido a endogeneidade. Por

exemplo, bairros que aumentam a violência podem concomitantemente reduzir a atividade econômica local. Não é fácil distinguir se a variação no preço hedônico se deveu a redução na atividade econômica ou no aumento da violência.

Essas ressalvas indicam que o uso de preço hedônico como indicador de disponibilidade de pagamento por características urbanas é limitado, porém, pode ser informativo, desde que sejam realizadas interpretações apropriadas.

Glaeser et al (2018) analisaram se o Google Street View ajuda a prever os preços hedônicos dos bairros de Nova York. Eles utilizam estratégia semelhante a utilizada para prever a renda mediana dos bairros. Todavia, ao invés de usar diretamente as imagens para prever o logaritmo dos preços hedônicos, eles utilizaram a renda mediana por bairro. A essência deste exercício é que lugares com renda mediana mais elevada tendem a ter residências com preços hedônicos mais elevados.

Os resultados apontaram que tanto a renda mediana prevista pelo Google Street View quando o residual da renda mediana atual e a renda prevista são bons previsores para os preços hedônicos dos bairros de Nova York. Este exercício foi replicado para a cidade de Boston e os resultados foram similares indicando que tais previsões possam ter validade externa.

Este resultado é a primeira evidência de que tais formas de dados granulares podem ser utilizadas para realizar previsão de características urbanas difíceis de serem mensuradas.

4.2.3. Identificando a gentrificação

O processo de gentrificação refere-se as transformações urbanas devido a mudanças na população residente. Este processo é de difícil acompanhamento por que grande parte do movimento é intra-municipal e dados disponíveis para este acompanhamento possuem uma frequência muito pequenas. No Brasil, por exemplo, este processo em bairros apenas pode ser visualizado com o Censo Populacional, realizado decenalmente.

Glaeser, Luca e Kim (2018) buscaram formas mais granulares e com frequência de acompanhamento dos processos de mudanças nos bairros. Para isso, testaram se dados disponíveis na plataforma Yelp possuem poder preditivo sobre as mudanças sociais na cidade de Nova York. A plataforma Yelp foi criada em 2004 com o objetivo de receber avaliações de estabelecimentos comerciais de forma colaborativa. Na plataforma os usuários de estabelecimentos comerciais urbanos realizam avaliações que são constantemente atualizadas a outros usuários. Dessa forma,

consumidores, baseados nessas avaliações, podem decidir qual o melhor lugar para consumir.

Os autores buscaram entender se as avalições do Yelp podem ajudar a prever mudanças nas estruturas sociais de bairros. A hipótese é de que bairros com determinadas características comerciais atraem grupos de indivíduos diferentes. Por exemplo, bairros cuja população seja mais rica, geralmente atraem um maior número de determinado tipo empresa do que bairros com uma população de baixa renda.

Existem inúmeras vantagens em se utilizar os dados do Yelp. Primeiro, tais dados são de elevada frequência. Isso permite acompanhar a de forma quase-contínua a criação e permanência de estabelecimentos comerciais pela cidade. Segundo, estes dados são georreferenciados, possibilitando que este mapeamento tenha elevado grau de granulação. Por fim, as informações contidas no Yelp possibilitam a identificação do tipo de negócio e da qualidade do serviço prestado.

Existem algumas desvantagens também. Primeiro, dados do Yelp dependem da colaboração da população usuária. Isso implica a possibilidade de que tais dados contenham erros de mensuração. Bairros cuja a colaboração seja pequena, podem refletir apenas avaliações extremas. Segundo, nem todos os tipos de estabelecimento são avaliados, pela própria natureza do negócio. Empresas que não possuem atendimento direto ao público, como fábricas, por exemplo, não são avaliadas pelo Yelp.

Glaeser, Luca e Kim (2018) avaliaram qual o poder de previsão das informações contidas no Yelp para prever três características das populações residentes em cada bairro de Nova York: percentual da população com ensino superior (educação), percentual da população entre 25 e 34 anos (idade) e composição racial (raça). Modificações nessas características ao longo do tempo podem indicar processos de gentrificação entre os bairros. Munido dessas informações, o poder público pode planejar melhor a disposição dos serviços públicos para grupos específicos de residentes.

Para testar o poder de previsão das informações contidas no Yelp, coletados entre 2007 a 2011, os autores utilizaram uma pesquisa censitária realizada entre 2006 e 2010 pela American Community Survey (ACS). Esta pesquisa coletou informações sobre as características dos residentes em cada bairro de Nova York. As informações do Yelp foram agregadas segundo o número de tipos de estabelecimento em cada bairro. Por exemplo, foi contabilizado o número de bares, restaurantes, supermercados, etc, em cada bairro ao longo do tempo. E essas medidas foram utilizadas para prever as variáveis sociais de educação, idade e raça.

Os resultados mostram que os dados contidos no Yelp ajudam a prever principalmente a educação dos bairros. As demais variáveis não tiveram poder preditivo elevado. Além disso, os autores verificaram que existe uma forte correlação entre as atividades empresariais locais, mensuradas pelo Yelp, e as características demográficas da população residente nos bairros.

Estes resultados podem ser utilizados para acompanhar em tempo real processos como a gentrificação ou a localização de determinada atividade econômica ao dentro de cada cidade. Importante ressaltar que estes resultados são bastante preliminares e podem ser expandidos para uma previsão mais acurada de variáveis alvo mais específicas e mesmo para verificar se a qualidade dos serviços públicos está relacionada a mudanças nas características demográficas dos bairros.

4.2.4 Onde fazer uma inspeção sanitária

Outra forma de utilização dos dados de avaliação de estabelecimentos comerciais, como o Yelp, é entender como os consumidores avaliam determinado serviço comercial. De forma comunitária, indivíduos utilizam estas plataformas para expressar qual a qualidade média dos serviços dos estabelecimentos comerciais. Essas informações podem ser utilizadas identificar empresas que estejam prestando um serviço de baixa qualidade.

Em um tipo específico de negócio, como restaurantes, estas informações podem ajudar a prever onde deve ser realizado uma fiscalização sanitária. Motivados por essa ideia, Kang et al (2013) e Glaeser et al (2016) analisam como utilizar processamento natural de linguagens para identificar padrões que ajudem a prever se um restaurante está comercializando produtos com baixa qualidade de higiene. Este tipo de análise é importante pois a maior parte das fiscalizações sanitárias são realizadas quase que aleatoriamente. Isto é, poucos parâmetros são utilizados para decidir qual restaurante deve ser vistoriado [8].

Para acessar o poder preditivo das avaliações do Yelp, Glaeser et al (2016) realizaram um torneio com uma premiação para o cientista de dados que construísse o melhor modelo de previsão. Este tipo de torneio é uma alternativa interessante para os governos desenvolverem ferramentas uteis de análise de dados a um custo menor do que a contratação equipes por maior longo prazo.

Foram analisados 364 restaurantes na cidade de Boston. Desses foram registradas: 1563

[8]Restaurantes de comida japonesa são vistoriados com maior frequência do que outros tipos de restaurantes, por exemplo, pois a comercialização de comida fresca gera maiores riscos de contaminação e possui maior dificuldade de conservação.

violações classificadas como leves, 153 classificadas como médias e 341 classificadas como graves. Os participantes tiveram 12 semanas para desenvolver um algoritmo para prever a violações de higiene usando dados do Yelp. O torneio foi separado em duas fases: fase 1, desenvolvimento dos modelos preditivos; fase 2, avaliação dos modelos.

Os resultados mostram que o melhor modelo preditivo prevê um aumento de 50% no número de violações encontradas pelos agentes sanitários. Isso implica que a cidade de Boston pode aumentar entre 30 a 50% a produtividade caso adote os melhores algoritmos desenvolvidos.

4.3 Alerta de risco

Uma das grandes vantagens dos métodos de análise de dados aplicados a grandes bases de dados é a possibilidade de antecipar eventos de interesse em políticas públicas. Como os métodos de machine learning foram projetados para gerar as melhores previsões e são ideais para lidar com grandes bases de dados, é possível desenvolver previsões específicas para variáveis de interesse.

Isso permite que se tenha um acompanhamento contínuo do risco de ocorrência de determinado evento. Essa medida de risco pode ser utilizada para enderessar políticas públicas ex-ante ao acontecimento do evento.

Por exemplo, um dos problemas mais graves da educação é a evasão, pois, por um lado, representa a impossibilidade de aumento do capital humano dos indivíduos que evadem e, por outro lado, implica em desperdício de recursos públicos destinados ao aluno que evade (Rumberger et al., 2017). Esse problema tende a ser ainda mais grave em países pobres ou em desenvolvimento, pois, a quantidade de recursos destinados a educação é escassa e existe uma maior necessidade de aumento da acumulação de capital humano (GLEWEE e MURALIDHARAN (2016), MURALIDHARAN (2017)).

Diferentes políticas de combate a evasão tem sido testadas[add ref]. Geralmente, o custo de aplicação destas políticas é elevado uma vez que não existe focalização de qual estudante necessita do tratamento. Ou seja, políticas públicas de combate a evasão tendem a ser aplicadas a estudantes que não possuem o risco de evadir. Para evitar esse tipo de desperdício de recursos públicos, governos tem adotado indicadores gerais de risco de evasão. Por exemplo, a cidade de Chicago adota o On track indicator que serve como métrica do risco de evasão.

Métodos aplicados a Big Data tem o potencial de elevar o poder de predição do risco de evasão, permitindo que seja mais acertada a decisão sobre quem deve receber a política de combate

a evasão. De fato, as evidências tem mostrado que métodos de machine learning elevam aproximadamente 10% o poder de previsão do risco de evasão.

Recentemente, muitos autores tem se dedicado ao tema de prever a evasão usando métodos de machine learning. Este tipo de abordagem faz parte do que Kleinberg et al (2015) chamou de problemas de predição de políticas. Exemplos de aplicações são: Aguiar et al (2015), Sansone (2019), Aulck et al (2016).

4.4 Finanças Públicas

Uma das áreas de maior importância na aplicação de métodos de análise de dados são as finanças públicas. O uso de dados para guiar políticas fiscais é um importante passo para aumentar a eficiência arrecadatória e realizar os gastos de forma mais racional.

Alguns autores tem usado métodos de análise de dados para guiar políticas fiscais. Em especial, Andini et al (2019) e Andini et al (2018) mostram como a análise de dados pode ajudar a reduzir ineficiências na alocação de recursos públicos. Nesta seção será analisada os trabalhos realizados por esses autores para Itália.

Andini et al (2018) analisam o programa de descontos de impostos implementados na Itália em 2014. Este programa é parte de um grande conjunto de políticas anticíclicas que visavam estimular a economia italiana após a Grande Crise Financeira de 2008. O programa consiste em conceder um crédito tributário para trabalhadores formais cuja renda anual figurava no intervalo de 8.145 à 26.000 euros. Estimativas do governo italiano apontam que este programa implica em transferências da ordem de 7 bilhões de euros anualmente. Esse crédito é disponibilizado aos beneficiários quando realizam o pagamento de tributação anual, semelhante ao realizado no Brasil com o imposto de renda.

A efetividade da política depende da forma como ela for alocada. Assim, os autores buscam avaliar se formas alternativas de focalização deste programa impactam determinadas variáveis de resultado. Este tipo de exercício foi definido por Kleinberg et al (2015) como problemas de previsão política. Uma das vantagens deste tipo de abordagem consiste na possibilidade de simular o efeito preditivo decorrente da mudança de focalização das políticas públicas.

Andini et al (2018) assumem que o objetivo do programa seja o de maximizar o consumo agregado italiano. De fato, o programa tem como um dos objetivos centrais aumentar a atividade econômica via consumo. Assim, por meio de dados de restrições de consumo das famílias obtidos

entre 2010 e 2012, os autores criam um modelo preditivo para identificar à priori qual famílias, de trabalhadores formais, tem maior potencial de consumo.

A base de dados utilizada é uma pesquisa bianual realizada pelo Banco da Itália, conhecida como SHIW [9] que reúne informações sobre o comportamento de consumo das famílias italianas. Especialmente em 2014, foi pesquisado, adicionalmente, sobre o recebimento de crédito tributário pelos consumidores. A amostra em 2014 contém 8156 famílias, porém, a base final utilizada inclui apenas 3646 famílias que possuem ao menos um trabalhador formal.

Para identificar consumidores com maior potencial de consumo, os autores utilizam como proxie uma medida das restrições de consumo das famílias. A hipótese é que o crédito tributário quando entregue a famílias com restrições de consumo, estas irão utilizar os recursos para outras finalidades e não para o consumo em si. Ou seja, o programa não é efetivo quando é destinado a tais famílias. Por sua vez, famílias sem restrição de consumo podem utilizar os recursos do crédito tributário para elevar o consumo familiar e com isso gerar maior consumo agregado.

A variável que mensura as restrições de consumo da família é a dificuldade que a família tem de fazer frente a suas despesas. Os autores utilizam um rico conjunto de dados para prever quais famílias terão em 2014 restrições de consumo. Essa base de dados contém características demográficas, de renda, educação, gênero, restrições financeiras e outras.

Usando métodos de Machine Learning os autores mostram que aproximadamente 29% dos recursos utilizados pelo programa foram destinados a famílias com restrições de consumo. Isso implica em aproximadamente 2 bilhões de euros possivelmente destinados a famílias que não elevaram o consumo agregado.

Este trabalho mostra que métodos de análise de dados podem ser aplicados para a melhor focalização de programas e podem ajudar na alocação de recursos de forma mais adequada.

O segundo trabalho de Andini et al (2019) utilizam abordagem semelhante para endereçar o problema de focalização de garantias de crédito as empresas italianas. O sistema de garantias de crédito apóia o acesso das firmas ao crédito bancário, por meio do financiamento do colateral com fundos públicos. Este programa é focado especialmente em médias e pequenas empresas cuja restrição de crédito é maior.

A questão central consiste em saber se as firmas beneficiadas com este programa de garantias de crédito, chamado de *Italian Guarantee Fund* (IGF), realmente possuem restrição de

[9]The Survey on Household Income and Wealth (SHIW)

crédito. Normalmente, as empresas são identificadas como tendo restrição de crédito por meio de métodos de credit scoring. Todavia, existem evidências de que estes métodos não são efetivos na identificação destas firmas.

Em 2000 é lançado o IGF que empregou, até 2008, 11 bilhões de euros em garantias públicas para empresas de pequeno e médio porte. A crise de 2009 aumentou a necessidade de expansão do programa. O programa IGF garante até 80% do valor de empréstimos bancários. Entretanto, o valor da garantia é de até 1,5 milhão de euros por empresa. Não existe restrições ao tipo de empréstimo que será realizado podendo ser: empréstimo de curto e longo prazo.

Os autores utilizam duas medidas para avaliar o risco de crédito das empresas: i. discrepâncias entre a demanda por crédito e a dotação da empresa, mensurado pelo Registro de Crédito do Banco da Itália; ii. ocorrências de empréstimos de baixa qualidade para a empresa. A empresa possui empréstimos de baixa qualidade se for declarada insolvente por um banco que representa pelo menos 70% dos empréstimos bancários totais da empresa, ou se for declarada insolvente por dois ou mais bancos que juntos representem pelo menos 10% do total de empréstimos bancários da empresa.

O objetivo dos autores é gerar modelos de previsão usando grande base de dados e métodos de machine learning para essas duas variáveis de interesse. A base de dados principal é coletada pelo Registro de Crédito do Banco da Itália, em modalidade preliminar a realização do empréstimo. A amostra completa possui 190 mil empresas que fizeram a solicitação de crédito em 2011. Uma desvantagem desta base de dados é que as empresas analisadas precisaram solicitar a garantia de crédito. Isso exclui empresas que não tenham dificuldade de crédito.

Foram utilizadas 108 preditores para as duas variáveis de resultado. Os modelos de machine learning foram comparados posteriormente com a regra de alocação de recursos do IGF. Os resultados mostram que os modelos de machine learning 63% de acurácia preditiva, considerando as duas variáveis de resultado em conjunto. Isso implica que em 63% das empresas analisadas, o modelo de previsão identifica se a empresa possui ou não restrição de crédito. A maior taxa de acerto é sobre as empresas que possuem efetivamente risco de crédito.

Posteriormente, foi comparado se os modelos de machine learning podem alocar recursos de forma mais efetiva do que os procedimentos convencionais do IGF. Os métodos de machine learning excluiriam aproximadamente 20% das empresas que receberam garantias de crédito pelo IGF. Isso representa 1,2 bilhões de euros em garantias de crédito. Este resultado indica que métodos

de análise de dados podem ajudar a focalizar políticas que envolvam transferências de recursos públicos, como é o caso do IGF.

4.5 Previsão e Nowcasting

Um dos grandes interesses para o planejamento de políticas públicas consiste na possibilidade de antecipação de cenários que possam prejudicar ou gerar oportunidades para tomada de decisão mais coerente. O problema é que prever cenários econômicos é uma tarefa difícil. A economia possui elevada dinâmica e a quantidade de informação disponível para fazer previsão ainda é escassa.

Métodos aplicados a Big Data têm buscado implementar novas técnicas ou utilizar diferentes estruturas de dados para realizar a previsão de várias variáveis econômicas. Dentre as novas técnicas destacam-se os métodos de machine learning que tem a capacidade de gerar previsões com melhor qualidade do que os modelos tradicionais utilizados em economia. Por outro lado, novas fontes de dados têm sido exaustivamente testadas quanto ao seu poder de previsão, como é o caso do uso de pesquisas na internet, preços de produtos online, mensagens enviadas em redes sociais, entre outros.

Aqui será feita uma ampla revisão da aplicação recente desses métodos. O texto se baseia principalmente em Kapetanios e Papallais (2018) e Buono et al (2018) que realizam um levantamento semelhante. Esta seção será subdivida por áreas em que os métodos de Big Data contribuíram para a previsão de variáveis econômicas.

4.5.1 Taxa de desemprego

Existem várias aplicações de técnicas de Big Data para prever a taxa de desemprego. Estas aplicações utilizam tanto base de dados novas como também técnicas estatísticas diferentes das tradicionais. Choi e Varian (2012) utilizam a ferramenta do Google chamada de Google Trends para prever a taxa de desemprego e outras variáveis. A ferramenta do Google Trends registra de forma organizada o que as pessoas estão pesquisando no buscador do Google. A hipótese dos autores é que pesquisas no Google sobre oportunidade de emprego, seguro desemprego, currículo etc podem conter informações sobre indivíduos desempregados. Os resultados, para essa variável, indicaram que Google Trends qualidade de previsão semelhante as bases de dados tradicionais, como a defasagem da taxa de desemprego.

Outros autores encontraram resultados diferentes. D'Amauri e Marcucci (2009) evidenciaram que o Google Index é o melhor previsor para a taxa de desemprego nos EUA. Os seus resultados parecem ser robustos aos períodos de crise econômica que geralmente são difíceis de prever e prejudicam o poder de previsão dos modelos devido à incerteza.

Ross (2013) utilizou o Google Trends para prever o comportamento de várias variáveis econômicas para o Reino Unido e demostrou que buscas na internet possuem importante conteúdo preditivo sobre a taxa de desemprego. Importante no artigo de Ross é a utilização de um método de seleção de palavras chaves relevantes para a previsão.

4.5.2 Inflação

Recentemente, alguns pesquisadores tem se dedicado a entender quais são as características econômicas dos preços de bens vendidos online e qual o poder preditivo que eles possuem. Alberto Cavallo e Roberto Rigobon são os pesquisadores que lideram tais pesquisas. Eles criaram em 2008 um laboratório, chamado The Billion Prices Project, que coleta informações de preços online ao redor do mundo. Atualmente, o projeto coleta mais de 15 mil preços, em mais de 1000 varejista em 60 países utilizando técnicas de web-scrapping [10].

Outras iniciativas semelhantes são as de Griffion et al (2014) que argumentam a favor do uso de informações online para previsão e compreensão das características dos preços agregados do consumidor. Os autores discutem as vantagens e desvantagens do método de web-scrapping. Dentre as vantagens destacam-se o baixo custo de coleta de informações, a qualidade da informação é boa e é possível analisar itens de forma granular, isto é, pequenas diferenças em itens próximos.

Por outro lado, destacam-se como desvantagens: sites mudam constantemente e cada mudança exige uma nova programação do web-scrapping; impossibilidade de ponderar os preços, como no caso de preços convencionais e, em alguns casos, a descrição dos itens pode prejudicar a classificação.

Apesar dessas iniciativas, existem poucas pesquisas que verificam o poder de previsão. Cavallo e Rigobon (2016) indicam possibilidade de uso destas informações para realização de nowcasting. Porém, ainda é preciso novas pesquisas para confirmar essa hipótese.

Entretanto, outra possibilidade consiste em utilizar os métodos de machine learning para

[10]Web-scrapping consistem em técnicas de coleta de informações na internet por meio de robôs programados.

aumentar o poder preditivo dos modelos tradicionais de previsão. Medeiros, Garcia e Vasconcelos (2017), Medeiros et al (2019) e Barbosa, Ferreira e Silva (2019) mostram que o uso de técnicas de machine learning podem melhorar significativamente a performance dos modelos de previsão para a inflação.

4.5.3 PIB

A possibilidade de nowcasting para a previsão do PIB e de seus componentes é muito mais ampla. Governos atualmente registram inúmeras transações para finalidade tributária e essas podem ser utilizadas para realizar previsões muito mais acuradas e em tempo real. Além disso, a utilização de dados de busca na internet, como o Google Trends, tem mostrado desempenho comparável aos modelos tradicionais.

Galbraith e Tkacz (2015) utilizam uma base de dados com pagamentos em cartões de crédito, nas modalidades crédito e débito, para criar um indicador para o crescimento do PIB do Canadá. A vantagem dessas informações é que elas captam um número grande de tipos de gasto, nos mais variados setores. Além disso, como são dados atualizados em tempo real, a possibilidade de melhorar as previsões de nowcasting são grandes.

Os resultados mostram que a utilização desta base de dados para nowcasting reduz em até 65% o erro de previsão quanto comparado a modelos de previsão tradicionais. Controlando para o tipo de modalidade, operações de débito tendem a ter maior poder de previsão.

Outros experimentos têm sido testados com o uso do Google Trends como em Koop e Onorante (2013) e Schmidt e Vosen (2011). Bok et al (2018) apresenta um sumário das técnicas empregadas para realizar nowcasting usando Big Data.

Uma alternativa a utilização de novas bases de dados é a aplicação de técnicas de machine learning para prever variáveis macroeconômicas. Nesse caso, utiliza-se as bases de dados tradicionais, porém, técnicas de aprimoramento de previsões são aplicadas visando melhorar a performance dos estimadores.

Essas técnicas tem sido testadas em vários conjuntos de dados. Kim e Swanson (2014, 2016) verificam que técnicas de machine learning aplicadas aos métodos fatoriais gerar um ganho em 0.33% dos casos, considerando 11 variáveis para a economia americana, frente aos modelos tradicionais. Este número aumenta para 0.63% em períodos recessivos.

Barbosa, Ferreira e Silva (2019) confirmam esses resultados para o caso brasileiro. Os

autores utilizaram métodos híbridos de previsão, métodos fatoriais adicionados de machine learning. Os resultados apontaram que os métodos híbridos foram significativamente superiores para prever a produção industrial, IPCA, INPC e taxa de desemprego. Os resultados foram superiores quando os métodos de machine learning eram utilizados para selecionar variáveis antes da aplicação fatorial, técnica conhecida como supervisão fatorial.

4.6 Focalização de políticas públicas

Um problema bastante comum em políticas públicas consiste na avaliação ex-ante de quem deve ser efetivamente tratado com alguma política. Para ilustrar é interessante abordar no contexto de saúde. Considere que haja um tratamento destinado ao combate a determinada enfermidade. Este tratamento possui o custo real c. Suponha que exista uma população de tamanho P.

Alguns tipos de tratamento são disponibilizados para uma fração da população. Uma campanha de vacinação, por exemplo, tem o custo real de $C = c \times \alpha \times P$, em que: αé a parcela da população vacinada ($\alpha \in [0,1]$). Este tratamento é realizado em indivíduos que estejam ou não sendo acometidos pela enfermidade.

Outros tipos de tratamento, devido a presença de efeitos coletareis, somente devem ser aplicados a indivíduos que realmente sejam portadores da enfermidade. Mas a pergunta é, quem são esses indivíduos? Uma pessoa quando sente um sintoma procura um especialista que realiza um diagnóstico do paciente. Esse diagnóstico é baseado em alguns testes que ajudem a comprovar um determinado quadro clínico. Uma vez tendo ciência da presença da enfermidade, então, o médico determina o tratamento específico. Usando a mesma estrutura acima, o custo total de pessoas tratadas será: $C = c \times \delta \times P$, em que: δcorresponde a parcela da população que foi diagnosticada com a enfermidade e que foi tratada ($\delta \in [0,1]$).

Em muitos casos, o médico, muitas das vezes, diagnóstica um paciente por meio de uma previsão. Alguns exames são solicitados para auxiliar na decisão, mas em essência a classificação de vários tipos de enfermidades depende da experiência do médico em prever, baseado nos sintomas dos pacientes, qual tipo de tratamento deve ser designado.

Isso implica que o coeficiente δpode conter erros de previsão, isto é, pode ser que mais (ou menos) pacientes sejam designados para determinado tratamento e isso pode encarecer o custo público com este tratamento. Realizar uma previsão mais acurada, neste caso, implica que pessoas que realmente enfermas recebam o tratamento adequado, reduzindo com isso o desperdício de

recursos.

Este contexto pode ser aplicado em outras situações, como políticas públicas sociais, educacionais, psicológicas, etc. Por exemplo, qual indivíduo deve ser alvo de uma política de combate a evasão escolar? Qual família deve fazer parte de programas de transferência de renda, como o bolsa família? Qual pessoa precisa de ajuda psicológica para lidar com a pressão do trabalho?

Todos esses casos envolvem recursos de políticas públicas destinados a indivíduos ou que recebem o tratamento (política), quando não precisam ou que não recebem o tratamento, quando de fato precisam. Em ambos os casos, formas mais adequadas de realizar previsão dos indivíduos que necessitam de determinada política, permitem economizar recursos e elevar o bem-estar social.

Métodos estatísticos aplicados a Big Data podem ser de grande ajuda nessa questão. Métodos de machine learning foram desenvolvidos para gerar previsões de melhor qualidade. Assim, usando dados educacionais é possível empregar métodos de machine learning para prever qual aluno possui a maior probabilidade (risco) de evadir a escola. Essa ferramenta, caso seja realmente mais precisa, pode auxiliar os gestores a aplicar políticas de combate a evasão que sejam mais adequadas aos indivíduos com maior risco de evasão.

Diversos estudos tem buscado verificar se tais métodos realmente preveem melhor que abordagens tradicionais baseadas na experiência ou em técnicas estatísticas tradicionais. Em um estudo recentemente publicado, Kleinberg et al (2018) mostram que métodos de machine learning preveem melhor do que juízes de tribunais americanos.

No artigo, os autores estudam uma situação na qual os juízes precisam tomar alguma decisão baseada em previsões. Quando um indivíduo é preso por algum tipo de crime, os juízes precisam decidir, após uma audiência de custódia, se irão deixar o indiciado responder em liberdade ou se os indiciados aguardarão o julgamento na prisão. Geralmente, essas previsões são baseadas na experiência dos juízes. Ou seja, não existe um critério estatístico que auxilie os juízes na sua tomada de decisão.

Os autores utilizaram dados da cidade de Nova York de 2008 a 2013. Foram registrados nesse período 1.460.462 indiciamentos por algum tipo de crime. Como preditores eles utilizaram o histórico de condenações, variáveis demográficas como gênero, raça, etnia, idade e variáveis associadas ao tipo de crime (usou arma de fogo ou não, por exemplo).

O método de estimação empregado foi gradient-boosted decision tree. A variável de

interesse é saber o risco associado ao indiciado cometer outro crime ou não retornar para o julgamento, caso seja deixado em liberdade. Importante notar que erros de previsão podem aumentar a população carcerária, ao manter presos que poderiam ter sido deixados em liberdade, e pode aumentar o número de crimes, ao se colocar em liberdade um indiciado que voltará a cometer crimes.

Os resultados foram surpreendentes. Os autores identificaram que é possível reduzir em 25% a quantidade de crimes na cidade de Nova York, mantendo constante o número de pessoas que são liberadas para aguardar o julgamento em liberdade. Além disso, houve uma redução de 40% no número de presidiários aguardando julgamento nas prisões nova-iorquinas, mantendo constante o número de crimes. Estes resultados apontam que métodos de machine learning possuem um elevado poder de previsão quando comparado a experiência de pessoas que tomam decisões baseadas em previsões.

Este exemplo relacionou-se a aplicação em um contexto específico. Entretanto, a literatura tem buscado evidências de aplicação da focalização em outras tipos de problemas. Por exemplo, Andini et al (2019) utilizam métodos de Big Data para focalizar programas de financiamento de crédito empresarial na Itália. Os resultados mostram que a alocação de recursos aumentou significativamente a qualidade, atendendo as empresas que realmente precisam de crédito público.

4.6.1 Prevendo o tratamento heterogêneo

Uma outra possibilidade interessante da aplicação de métodos de Big Data em focalização de políticas públicas consiste na possibilidade de realização de políticas públicas com tratamentos personalizados. Para ilustrar essa possibilidade, considere que o ensino realizado em sala de aula possa ser personalizado. Isto é, cada aluno seguirá uma trajetória diferente de aprendizado. Dessa forma, caso um aluno precise se dedicar mais tempo ao aprendizado de álgebra ele poderá sem necessariamente afetar o aprendizado de outro aluno em português.

Recentemente, alguns artigos tem se dedicado a testar a utilização de métodos de Big Data para a esta finalidade, como Athey e Imbens (2016), Athey e Wager (2019), Hanh et al (2017) entre outros.

Em essência, estes métodos buscam verificar quais grupos de indivíduos mais se beneficiam de determinado tratamento. Entretanto, diferentemente da análise realizada em econometria tradicionalmente, aqui é permitido que os dados informem a melhor forma de personalizar o

tratamento mesmo que isso não implique necessariamente em alguma racionalização teórica. A personalização é inteiramente dirigida por dados.

4.6.2. Perigos do uso de métodos de machine learning para focalização de políticas públicas

Um problema associado com o uso de algoritmos para identificar o risco de possíveis ações dos indivíduos é a possibilidade de o algoritmo realizar discriminação de raça ou gênero, por exemplo.

Considere que um algoritmo seja utilizado para identificar alunos com maior risco de evasão na escola. Os alunos classificados como estando em risco de evasão farão parte de uma política de acompanhamento psicológico.

Na amostra em que o algoritmo foi treinado, a base de dados continha informações de cor e gênero. Se os alunos pretos forem os que mais evadem a escola, então o algoritmo vai atribuir essas características um peso maior quando for realizar a previsão de um grupo de alunos externo. Dessa forma, alunos pretos receberão uma indicação de possuem maior risco de evasão do que alunas brancas, por exemplo.

Todavia, cor e gênero não são causa para a evasão escolar. Tais atributos representam outros tipos de características que por sua vez determinam a maior probabilidade de evasão. Por exemplo, indivíduos de cor preta, devido a formação histórica do Brasil, são em geral pobres. A evasão não é causada pela cor da pele, mas pode ser causada pela falta de oportunidades de os alunos em poder continuar na escola. Assim, a cor da pele serve como uma aproximação da pobreza do indivíduo.

Os riscos deste tipo de associação estão na má interpretação dos resultados. Primeiro, é preciso considerar que algoritmos realizam previsões baseados em informações amostras de treino. Caso haja contaminação na amostra de treino, no sentido de haver mais ou menos características específicas, o algoritmo simplesmente replicará o padrão encontrado. Isto é, o algoritmo irá dar mais peso dependendo das características contidas na amostra [11].

Uma sugestão considerada por Kleinberg et al (2018) seria a de desenvolver modelos preditos excluindo certas características que possam causar discriminação pelo algoritmo. No caso do exemplo anterior, a ideia seria desenvolver um algoritmo considerando uma amostra que não

[11]O procedimento de separação da amostra em amostra de treino e de teste deve ser realizado da forma mais aleatória possível visando evitar o surgimento de padrões que gerem erros de previsão.

contenha as variáveis relacionadas a cor e gênero.

Apesar desta limitação, a literatura ainda não encontrou uma forma adequada de lidar com este tipo de problema, sendo ainda uma questão a ser mais aprofundada futuramente. Para uma revisão ampla deste problema ver Kleinberg et al (2019).

5. Possíveis aplicações de Big Data no contexto cearense

Nesta seção serão levantadas algumas possíveis aplicações de métodos de Big Data para as gestões públicas cearenses. Essas possíveis aplicações devem ter sua viabilidade estudada em particular levando em consideração diferentes aspectos específicos da economia cearense. Entretanto, tais sugestões estão sendo realizadas em outros estados ou países, com relativo sucesso. Obviamente, as sugestões aqui não esgotam, nem de perto, as possibilidades de aplicação. O objetivo da seção é, portanto, apenas destacar que existem muitas possíveis aplicações no contexto cearense.

Serão analisadas 3 possibilidades de aplicação: 1. Aplicação em seleção de pessoal, 2. Focalização de benefício tributário e 3. Previsão do abandono.

5.1 Aplicação em seleção de pessoal

Vários trabalhos vistos anteriormente testaram a possibilidade de se utilizar métodos de análise de Big Data para selecionar candidatos a cargos públicos. Apesar de tais métodos serem difíceis de serem aplicados devido à falta de legislação apropriada, é possível utilizá-los como forma secundária de seleção. Isto é, como não é possível, e talvez não seja desejável, selecionar indivíduos unicamente por meio de algoritmos, os algoritmos podem ajudar a realizar a capacitação e treinamento secundário dos indivíduos, antes de adquirirem a estabilidade na atividade pública.

Por exemplo, policiais quando envolvidos em conflitos podem ter atitudes não adequadas, como uso excessivo da violência, descontrole emocional, falta de atenção, etc. Usando técnicas de Big Data é razoavelmente possível prever quais policiais possuem maior risco de atuarem de forma inadequada, como mostraram Chalfin et al (2016). Obviamente, não se pode utilizar tais estimativas para realizar a seleção direta dos policiais, entretanto, policiais com maior risco de cometer atos inadequados no futuro podem, durante o período de estágio probatório, serem submetidos a uma intervenção especial, por exemplo um curso específico, que vise minorar tais riscos.

A participação e conclusão com satisfação do curso pode servir para a aprovação do policial no estágio probatório. Assim, o mecanismo de seleção via algoritmos é realizado de forma indireta e dando oportunidades de minimização do componente de risco.

5.2 Focalização de benefício tributário

Sendo um dos objetivos da administração pública a maximizar sua arrecadação total sem causar distorções econômicas significativas, os métodos de Big Data podem ser utilizados para esta finalidade. Considere um caso concreto. No Ceará é dado um desconto de 5% pelo pagamento à vista do imposto sobre propriedade de veículos automotores (IPVA). Este desconto é dado de forma generalizada, isto é, sem considerar diretamente a incapacidade do contribuinte em pagar o tributo. Indivíduos com renda mais elevada conseguem quitar seu débito tributário de uma única vez e com isso recebem a bonificação por isso.

Entretanto, os indivíduos com elevado risco de não pagamento do IPVA, geralmente pessoas com menor renda, não recebem nenhum incentivo direto para concretizar seu pagamento. Um exercício interessante seria tentar identificar, dadas as características do proprietário do veículo e do próprio veículo, qual o risco de um determinado indivíduo não pagar o IPVA. Munidos com esse indicador de risco, os agentes públicos podem reduzir o prêmio por pagamento imediato e dar desconto para os indivíduos com maior risco de não pagamento.

Indivíduos que pagam à vista, provavelmente pelo seu perfil, não deverão entrar em débito com o fisco por não ter o desconto de forma integral. Ou seja, a redução da bonificação por antecipação não deve reduzir a arrecadação com IPVA total, mas sim a arrecadação imediata.

Por outro lado, indivíduos que possuem elevado risco de não pagamento podem cumprir com suas obrigações legais caso seja dado um desconto no valor do tributo. Como resultado, espera-se um adiamento maior da arrecadação com IPVA, porém, espera-se também um aumento do valor total arrecadado.

A dificuldade aqui é apenas de identificar quem são os indivíduos em situação de risco de não pagamento e qual a margem ótima de desconto para ambos os grupos.

5.3 Previsão do abandono

O governo do estado possui um programa conhecido como professor diretor de turma (PPDT). Este programa defini um professor que acompanhe uma turma específica de forma mais

direta e com maior aproximação. A ideia do PPDT é de que o professor possa acompanhar de perto os alunos auxiliando-os em situações extraclasse. Então, tal programa é focado no combate ao abandono escolar e busca melhorar as competências socioemocionais dos estudantes.

Atualmente, os professores do PPDT, em conjunto com as equipes da Secretaria de Educação do Estado do Ceará (SEDUC), designam que tipo de intervenção e quais alunos devem ser atendidos. A seleção do tratamento e dos alunos a serem tratados é realizada, muitas vezes, pela própria experiência dos professores e gestores escolares.

Métodos de Big Data podem ser utilizados para ajudar na focalização de indivíduos que tenham maior probabilidade de evadir a escola. A literatura aponta que o abandono escolar é passível de previsão com o uso de informações socioeconômicas e de desempenho prévio dos alunos ( ). Com isso, um algoritmo pode ser desenvolvido para indicar qual o aluno que tem o maior risco de evadir, permitindo que o professor diretor de turma possa atuar de forma mais focalizada.

Além disso, é possível não apenas saber qual o aluno com maior risco, mas também que tipo de tratamento tem maior poder de lograr sucesso em determinado subgrupo, por meio da previsão do efeito heterogêneo.

6. Conclusões

Este produto apresentou uma extensa revisão da literatura sobre potências aplicações de Big Data objetivamente aumentar a eficiência da gestão pública. Foi analisado que existem várias áreas onde a análise de grandes bases de dados, associadas aos seus métodos estatísticos, podem gerar benefícios reais a gestão pública estadual.

A primeira aplicação analisada foi sobre ganhos na seleção de pessoal. Métodos de análise de dados podem ser empregados para selecionar pessoas com maiores chances de atenderem os parâmetros de qualidade esperada. Uma limitação importante deste tipo de aplicação é a necessidade de uma clara e objetiva definição da mensuração da qualidade do trabalhador. Sem isso, a aplicação dos métodos de Big Data não é possível.

O segundo tipo de aplicação foi referente ao uso de Big Data para melhorar a gestão e o planejamento urbano. Foi visto que é possível melhorar o cálculo da riqueza de forma mais granular, permitindo, por exemplo, que tributos sejam cobrados de forma mais fidedigna com a

realidade do contribuinte. Além disso, foi mostrado como métodos de Big Data podem ser utilizados para mensurar o interesse dos cidadãos por serviços públicos específicos. Por fim, foi apresentado como é possível calcular de forma contínua a gentrificação, importante processo para o planejamento urbano.

A terceira aplicação analisada foi a elaboração de alertas de risco. Tradicionalmente, tais alertas são aplicados para monitorar possíveis desastres naturais. Aqui, com a evolução da análise de dados é possível atualmente realizar monitoramento de risco de situações mais específicas como risco de evasão, risco de determinada doença por pacientes, etc. Posteriormente, foi visto como métodos de Big Data podem ser aplicados para gerar ganhos de eficiência em finanças públicas.

Por fim, foi mostrado que métodos de Big Data podem ser utilizados para melhorar a gestão macroeconômica de países e estados por meio da melhor previsão de cenários. Foi visto que existem enormes ganhos em utilizar Big Data para prever a atividade econômica, a inflação, a receita fiscal entre outros.

Espera-se que este produto sirva de base para aplicações reais e estudos mais aprofundados sobre o tema. Além disso, espera-se que parte das aplicações sugeridas sejam absorvidas pelo poder público estadual de forma a torná-las viáveis. Com isso, será possível realizar serviços de melhor qualidade a um custo menor do que é praticado nos dias atuais, gerando com isso maior eficiência do gasto público estadual.

7. Referências

ABADIE, A. e KASY, M. The Risk of Machine Learning, MIT Working Paper, 2017.

ABADIE, A.; DIAMOND, A. e HAINMUELLER, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. Journal of the American Statistical Association, Vol. 105, No. 490, 2010.

AGUIAR, E., LAKKARAJU, H., BHANPURI, N., MILLER, D., YUHAS, B. e ADDISON, K. L. Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time', Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15, Vol. 15, pp. 93–102, 2015.

ALLCOTT, H.; BRAGHIERI, L. E EICHMEYER, S. e GENTZKOW, M. The Welfare Effects of Social Media. NBER Working Paper No. 25514, 2019.

ANDINI, M, E Ciani, G de Blasio, A D'Ignazio and V Salvestrini (2018a), "Targeting with machine learning: An application to a tax rebate program in Italy", Journal of Economic Behavior and Organization, forthcoming.

ANDINI, M, M Boldrini, E Ciani, G de Blasio, A D'Ignazio and A Paladini. Machine learning in the service of policy targeting: The case of public credit guarantees, Bank of Italy Working Paper, 2018.

URQUIOLA, M e MacLEOD, B. Is Education Consumption or Investment? Implications for School Competition Annual Review of Economics, forthcoming, 2019

ATHEY, S. and G. W. IMBENS. The state of applied econometrics: Causality and policy evaluation. The Journal of Economic Perspectives, 31(2):3-32, 2017

AULCK, L.,VELAGAPUDI, N., BLUMENSTOCK, J. e WEST, J. Predicting student dropout in higher education. ArXivWorking Paper 1606.06364, 2016.

AUTOR, D., DORN, D. AND HANSON, G. The China Shock: Learning from Labor Market Adjustment to Large Changes in Trade. Annual Review of Economics, 8, 205–240, 2016.

BAI, J. Panel Data Models With Interactive Fixed Effects. Econometrica, Volume77, Issue 4, pp. 1229-1279, 2009.

BAJARI, P.; CHERNOZHUKOV, V. ; HORTAÇSU, A.; SUZUKI, J. The Impact of Big Data on Firm Performance: An Empirical Investigation, NBER Working Paper 24334, 2018.

BAKER, S., BLOOM, N. and DAVIS, S. Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics, v. 131, issue 4, pp. 1593-1636, 2016.

BARBOSA, R., FERREIRA, T. E SILVA, T. Previsão de Variáveis Macroeconômicas Brasileiras usando Modelos de Séries Temporais de Alta Dimensão. Estudos Economicos (Forthcoming), 2020.

BELLONI, A.; CHERNOZHUKOV, V. e HANSEN, C. High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics. The Journal of Economic Perspectives, 28 (2), pp. 29-50, 2014.

BELLONI, A.; CHERNOZHUKOV, V.; HANSEN, C.; FERNANDEZ-VAL, I. Program Evaluation and Causal Inference with High-Dimensional Data, Econometrica 2016.

BELLONI, A.; CHERNOZHUKOV, V.; HANSEN, C.; KONBUR, D. Inference in High Dimensional Panel Models with an Application to Gun Control, Journal of Business & Economic Statistics, 2015

BLOOM, N. Fluctuations in Uncertainty. Journal of Economic Perspectives. Vol. 28, No. 2, pp. 153-76, 2014.

BOK, B., CARATELLI, D., GIANNONE, D., SBORDONE, A., TAMBALOTTI, A. Macroeconomic Nowcasting and Forecasting with Big Data, Federal Reserve Bank of New York Sta Reports, Sta Report No. 830, 2019.

BUONO, D.; KAPETANIOS, G.; MARCELLINO, M.; MAZZI, G. e PAPAILIAS, F. Big Data Econometrics: Now Casting and Early Estimates. Working Paper N. 82, Bocconi University, 2018.

CAVALLO, A., RIGOBON, R. The Billion Prices Project: Using Online Prices for Measurement and Research". The Journal of Economic Perspectives, 30(2), 151-178, 2016.

CHALFIN A.; DANIELI, O.; HILLIS, A.; JELVEH, Z. LUCA, M.; LUDWIG, J. e MULLAINATHAN, S. Productivity and Selection of Human Capital with Machine Learning. American Economic Review: Papers and Proceedings 106, no. 5, pp. 124–127, 2016.

CHENG, X. AND HANSEN, B. Forecasting with factor-augmented regression: A frequentist model averaging approach. Journal of Econometrics, 186, p. 280-293, 2015.

CHERNOZHUKOV, V.; FERNANDEZ-VAL, I.; DUFLO, E. DEMIRER, I. Discovery of Heterogeneous Treatment Effects in Randomized Experiments, ArXiv 2019

CHERNOZHUKOV, V.; FERNANDEZ-VAL, I.; WEIDNER, M. Network and Panel Quantile Effects via Distribution Regression ArXiv 2018

CHODOROW-REICH, G. Regional Data in Macroeconomics: Some Advice for Practitioners. Working Paper 26501, 2019.

COHEN, P.; HAHN, R.; HALL, J.; LEVITT, S.; METCALFE, R. Using Big Data to Estimate Consumer Surplus: The Case of Uber. NBER Working Paper No. 22627, 2016.

CORBI, R.; PAPAIOANNOU, E.; SURICO, P. Regional Transfer Multipliers, The Review of Economic Studies, Volume 86, Issue 5, pp. 1901–1934, 2019.

D'AMURI, F., e MARCUCCI, J. The Predictive Power of Google Searches in Predicting Unemployment. Banca d'Italia Working Paper, 891, 2012.

DONOHUE III, J. J. e LEVITT, S. D. The Impact of Legalized Abortion on Crime The Quarterly Journal of Economics, Volume 116, Issue 2, pp. 379–420, 2001.

DOORNIK, J. A., HENDRY D. F. Statistical Model Selection with Big Data. Cogent Economics & Finance, 3(1), 2015.

EFRON, B. and HASTIE, T. Computer Age Statistical Inference, Stanford University, 2017.

EREL, I.; LÉA, S.; STERN, H.; TAN, C. e WEISBACH, M. S. Selecting Directors Using Machine Learning, NBER Working Paper 24435, 2019.

FERNÁNDEZ-VILLAVERDE, J. e VALENCIA, D. Z. A Practical Guide to Parallelization in Economics, NBER Working Paper No. 24561, 2018.

GALBRAITH, J. W., TKACZ, G. Nowcasting GDP with Electronic Payments, Data Vintages and the Timing of Data Releases, European Central Bank: Statistics Papers Series, No 10, 2013.

GARCIA, M. MEDEIROS, M. E VASCONCELOS, G. Real-time inflation forecasting with high dimensional models: The case of Brazil. XVI Encontro de Finanças, Rio de Janeiro, 2016.

GENTZKOW, M. KELLY, B. TADDY, M. Text as Data. Journal of Economic Literature Vol. 57, No. 3, pp. 535-74, 2019.

GLAESER, E. L. e LUCA, M. Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change. Working Paper 18-077 Harvard University, 2018.

GLAESER, E. L., A. HILLIS, S. D. KOMINERS, e M. LUCA. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. American Economic Review, 106(5), pp. 114– 18, 2016.

GLAESER, E. L.; HILLIS, A.; KOMINERS, S. D. E LUCA, M. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. American Economic Review, 106(5), pp. 114–18, 2016.

GLAESER, E. L.; KOMINERS, S. D.; LUCA, M e NAIK, N. Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life, Economic Inquiry, Vol.56(1), pp.114-137, 2018.

GLAESER, E. L.; LUCA, M e NAIK, N. Nowcasting the local economy: using yelp data to measure economic activity. Working Paper 24010, 2017.

GOBILLON, L E MAGNAC, T. Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls, The Review of Economics and Statistics, vol. 98, issue 3, 535-551, 2016. GRIFFOEN, R., DE HAAN, J., WILLENBORG, L. Collecting Clothing Data from the Internet", Statistics Netherlands Technical Report, 2014.

HANSEN, B. e RACINE, J.S. Jackknife model averaging. Journal of Econometrics 167, p. 38–46, 2012.

HYUNYOUNG, C. e VARIAN, H. Predicting the Present with Google Trends . Economic Record, v. 88, Issue s1, 2012.

IAN GOODFELLOW ET Al. Deep Learning. MIT, 2016.

HASTIE, T.; JAMES, G. e WITTEN, D., e TIBSHARANI, R. An Introduction to Statistical Learning. Springer, 2008.

KANG, J. S., P. KUZNETSOVA, M. LUCA, e Y. CHOI. Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, pp. 1443– 48, 2013.

KANG, J. S.; KUZNETSOVA, P.; LUCA, M, E CHOI, Y. Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1443–1448, 2013.

KAPETANIOS, G. e PAPAILIAS, F. Big Data & Macroeconomic Nowcasting: Methodological Review. Data Analytics Centre for Macro and Finance Economic Statistics Centre of Excellence, 2018.

KIM, H., E SWANSON, N. Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence. Journal of Econometrics 178, p. 352–367, 2014.

KIM, H., E SWANSON, N.Mining big data using parsimonious factor machine learning, variable selection, and shrinkage methods, Rutgers University, working paper, 2016.

KLEINBERG, J., LAKKARAJU, H., LESKOVIC, J., LUDWIG, J., e MULLAINATHAN, S.(Working Paper). Human Decisions and Machine Predictions. NBER Working Paper , 2015.

KLEINBERG, J.; LUDWIG, J. MULLAINATHAN, S. e OBERMEYER, Z. Prediction Policy Problems. American Economic Review: Papers & Proceedings, 105(5), pp. 491–495, 2015.

KOOP, G., ONORANTE, L. Macroeconomic Nowcasting Using Google Probabilities", Working Paper, ECB Workshop on Big Data for Forecasting and Statistics, 2013

MCCRACKEN, M. e NG, S. FRED-MD: A Monthly Database for Macroeconomic Research, Journal of Business & Economic Statistics, Taylor & Francis Journals, vol. 34(4), pages 574-589, 2016.

MEDEIROS, M. C.; VASCONCELOS, G. E FREITAS, E. Forecasting Brazilian Inflation with High-dimensional Models. Brazilian Econometric Review, Vol 36, No 2. 2016

MULLAINATHAN, S. and J. SPIESS. Machine learning: an applied econometric approach.

Journal of Economic Perspectives, 31(2):87-106, 2017

MURALIDHARAN, K. e GLEWWE, P. Improving Education Outcomes in Developing Countries - Evidence, Knowledge Gaps, and Policy Implications in the Handbook of the Economics of Education Volume 5, edited by Eric Hanushek, Steve Machin, and Ludger Woessman, 2015

MURALIDHARAN, K. Field Experiments in Education in Developing Countries in Handbook of Field Experiments Volume 2, edited by Abhijit Banerjee and Esther Duflo, 2018.

NAIK, N., J. PHILIPOOM, R. RASKAR, AND C. A. HIDALGO. Streetscore—Predicting the Perceived Safety of One Million Streetscapes, in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington, DC: IEEE, pp. 793–99, 2014.

NAIK, N., RASKAR, R. e HIDALGO, C. A. Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance. American Economic Review, 106(5), pp. 128–32, 2016.

NAIK, N., S. D. KOMINERS, R. RASKAR, E. L. GLAESER, AND C. A. HIDALGO. Do People Shape Cities, or Do Cities Shape People? The Co-Evolution of Physical, Social, and Economic Change in Five Major US Cities. NBER Working Paper No. 21620, 2015.

NAKAMURA, E. e STEINSSON, J. Fiscal Stimulus in a Monetary Union: Evidence from US Regions. American Economic Review 104.3, pp. 753-92, 2014.

ROSS, A. Nowcasting with Google Trends: A Keyword Selection Method, Fraser of Allander Economic Commentary, 37(2), 54-64, 2013.

RUMBERGER, RUSSELL W. e LIM, SUN AH. Why Students Drop Out of School: A Review of 25 Years of Research, California Dropout Research Project Report #15 October 2008

SANSONE, D. Beyond EarlyWarning Indicators: High School Dropout and Machine Learning. Oxford Bulletin of Economics and Statistics, 2018. SCHMIDT T., VOSEN, S. Forecasting Private Consumption: Surveybased Indicators vs. Google Trends", Journal of Forecasting, 30(6), 565-578, 2011.

STOCK, J. H. e WATSON, M. Forecasting With Many Predictors In Elliott, G., Granger, C., And Timmermann, A., Editors, Handbook of Economic Forecasting, Volume 1, Chapter 10, p. 515-554, 2006.

STOCK, J. H. AND WATSON, M. W. Forecasting Using Principal Components From a Large Number of Predictors. Journal of The American Statistical Association, 97 p. 1167-1179,

2002.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 58, No. 1, pp. 267-288, 1996.

VARIAN, Hal. Big Data: New Tricks for Econometrics, Journal of Economic Perspectives, v. 28(2), pp. 3-28, 2014.