IPECE INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ

GOVERNO DO ESTADO DO CEARÁ
Secretaria do Planejamento e Gestão

Boletim de GESTÃO PÚBLICA

Nº 19/2020

Boletim de Gestão Pública – Nº 19/2020

Unidade Responsável:
Diretoria de Estudos de Gestão Pública – DIGEP

Coordenação:
Marília Rodrigues Firmiano

Colaboração:
Aprígio Botelho Lócio (Assessor Técnico DIGEP)
Tiago Emanuel Gomes dos Santos (Técnico DIGEP)

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 o **Boletim de Gestão Pública**

O Boletim de Gestão Pública do Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) tem como objetivo principal a difusão de melhores práticas e inovações na área de gestão e de políticas públicas. É uma publicação bimestral, formada por artigos sintéticos (descritivo-analíticos), elaborados pelo corpo técnico do Instituto e ou por técnicos convidados de outros órgãos do Governo do Estado do Ceará e de outras organizações. Em linhas gerais, os artigos buscam: (i) difundir melhores práticas, com a análise de casos específicos locais, estaduais, nacionais ou internacionais; (ii) apresentar avanços na gestão pública do Ceará, com as principais inovações em gestão e políticas públicas no Estado; (iii) discutir avanços teóricos nas áreas de gestão e de políticas públicas e como esses conhecimentos podem ser postos em ação; (iv) analisar desafios para a gestão e para as políticas públicas; ou (v) verificar inovações no âmbito do setor privado, indicando como elas podem servir de inspiração para o setor público.

Instituto de Pesquisa e Estratégia Econômica do Ceará – **IPECE** 2020

Boletim de Gestão Pública / Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE) / Fortaleza – Ceará: Ipece, 2020.

ISSN: 2594-8709

1. Economia Brasileira. 2. Economia Cearense. 3. Gestão Pública.

Nesta Edição:

SUMÁRIO EXECUTIVO

O primeiro artigo aborda que o Big Data tem ganhado popularidade nos últimos anos devido principalmente ao seu potencial nos mais diversos tipos de aplicações. Entretanto, apesar de muito se falar em Big Data, não existe uma definição consensual sobre o que esse termo significa. Além disso, a maior parte das discussões acerca deste termo estão relacionadas a suas aplicações na geração de novos produtos e soluções para o mercado privado, não tendo, portanto, uma discussão qualificada da sua aplicação em políticas públicas. O texto apresentado nessa edição do Boletim de Gestão Pública tem o objetivo contribuir para a discussão dessas duas questões. Na primeira parte do texto, será realizada um levantamento das principais definições utilizadas por acadêmicos, empresas e instituições internacionais sobre o conceito de Big Data. Será visto que este conceito está relacionado a natureza do tipo de dado que passa a ser utilizado em aplicações. Os tipos de dados convencionais, como dados transversais, em painel ou séries temporais, tornam-se casos especiais da profusão de novas formas de dados existentes. Em síntese, o Big Data é marcado por dados de grande volume, como dados diários, por exemplo, dados com diferentes tipos de estruturas, como textos e imagens, por fim pelo fato de que em algumas situações Big Data apresenta milhares de variáveis. A segunda parte do texto discute algumas aplicações de Big Data sobre políticas públicas. Será apresentado aplicações de Big Data no monitoramento e planejamento urbano, na seleção de pessoal para o setor público, na focalização de políticas públicas e outras aplicações. Espera-se que este texto contribua para o debate sobre Big Data e suas potencialidades de aplicação no setor público.

O segundo artigo apresenta os desafios encontrados pelo grupo de trabalho do Departamento de Economia e Estatística da Secretaria de Planejamento, Orçamento e Gestão do Rio Grande do Sul em projeto-piloto conduzido na cidade gaúcha de Cachoeirinha para estudo das condições de infraestrutura das escolas estaduais no município. Em uma tentativa de ampliar e sintetizar o conhecimento do gestor público sobre as condições das escolas gaúchas, foi construído um modelo de indicador multidimensional para avaliação da infraestrutura escolar. Na primeira seção o indicador é apresentado em detalhes. A segunda seção mostra o resultado preliminar do indicador construído de acordo com os dados do Censo Escolar e o compara com o resultado de visitas *in loco* às escolas selecionadas. A terceira e última seção detalha as inconsistências encontradas entre as respostas das escolas ao Censo Escolar e a realidade encontrada pelos pesquisadores, mostrando os desafios encontrados na utilização daquela fonte de dados em pesquisas. Foram encontradas importantes diferenças entre as respostas das escolas no Censo Escolar e a realidade com que os pesquisadores se depararam em todos os blocos de variáveis analisados, o que coloca desafios adicionais ao pesquisador que utilize esses dados.

1. O uso do Big Data em políticas públicas: Uma revisão das possibilidades e dos desafios

Autor: *Rafael Barros Barbosa* [1]

1.1 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)[2]. 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)[3], Athey (2019))[4].

Este trabalho tem um objetivo principal. Pretende-se realizar uma 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.

Este trabalho está dividido em mais quatro partes, além desta introdução. Na primeira parte, seção dois, será tratada a definição do termo "Big Data". Não há ainda uma definição consensual sobre este termo, 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.

A seção três apresenta várias possíveis aplicações de Big Data e suas técnicas estatísticas em políticas públicas. São analisadas algumas aplicações em: seleção de pessoal, planejamento e monitoramento urbano, focalização de políticas públicas entre outros. Por fim, a seção quatro discute as principais conclusões.

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

---

[1]Professor de Economia Aplicada na Universidade Federal do Ceará. Doutor em economia pela mesma universidade em 2014. Foi vencedor do Prêmio do Tesouro Nacional de melhor monografia em 2018. Fundador do educLAB, um laboratório de análise de dados e economia da educação, sediado na UFC. Atualmente, suas pesquisas tem sido focadas nas áreas: financiamento da educação, incerteza macroeconômica e efeito dos pares.

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

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

[4]ATHEY, S The Impact of Machine Learning on Economics, in: The Economics of Artificial Intelligence: An Agenda, ed. Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 2019.

1.2 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.

A classificação 4V 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)[5], 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)[6] 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

[5]BLOOM, N; BAKER, S.; DAVIS, D. Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics, Volume 131, Issue 4, Pages 1593–1636, November 2016.
[6]DOORNIK, J. A., HENDRY D. F. Statistical Model Selection with Big Data. Cogent Economics & Finance, 3(1), 2015.

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)[7][8][9] e Chernozhovukov et al (2018, 2019)[10][11], quanto para o seu uso em previsões macroeconômicas, como Stock e Watson (2002, 2006)[12], Kim e Swanson (2014, 2016)[13][14] e Cheng e Hansen (2016)[15].

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. Poucas pesquisas tem a possibilidade de trabalhar com este tipo de dado, como: Cohen et al (2016)[16] usando dados da UBER para estimar a curva de demanda, Bajari et al (2018)[17] que utilizam dados da Amazon para prever vendas e Allcott et al (2019)[18], 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

[7]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.
[8]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.
[9]BELLONI, A.; CHERNOZHUKOV, V.; HANSEN, C.; FERNANDEZ-VAL, I. Program Evaluation and Causal Inference with High-Dimensional Data, Econometrica 2016.
[10]CHERNOZHUKOV, V.; FERNANDEZ-VAL, I.; WEIDNER, M. Network and Panel Quantile Effects via Distribution Regression ArXiv 2018.
[11]CHERNOZHUKOV, V.; FERNANDEZ-VAL, I.; DUFLO, E. DEMIRER, I. Discovery of Heterogeneous Treatment Effects in Randomized Experiments, ArXiv 2019.
[12] 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.
[13] KIM, H., E SWANSON, N.Mining big data using parsimonious factor machine learning, variable selection, and shrinkage methods, Rutgers University, working paper, 2016.
[14] 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.
[15]CHENG, X. AND HANSEN, B. Forecasting with factor-augmented regression: A frequentist model averaging approach. Journal of Econometrics, 186, p. 280-293, 2015.
[16]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.
[17]BAJARI, P.; CHERNOZHUKOV, V. ; HORTAÇSU, A.; SUZUKI, J. The Impact of Big Data on Firm Performance: An Empirical Investigation, NBER Working Paper 24334, 2018.
[18]ALLCOTT, H.; BRAGHIERI, L. E EICHMEYER, S. e GENTZKOW, M. The Welfare Effects of Social Media. NBER Working Paper No. 25514, 2019.

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, Instagram, comentários em Blogs, Vídeos no Youtube, Mapas utilizados por indivíduos, E-mails, etc. Glaeser et al (2016)[19] utiliza 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.

1.3 Exemplos de usos de Big Data em políticas públicas

Esta seção discute 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

[19]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.

focalização de políticas públicas. Todos os exemplos são baseados em análises estatísticas rigorosas, portanto, estão excluídos os exemplos puramente qualitativos ou especulativos.

1.3.1 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 os métodos de análise de dados podem auxiliar na contratação de pessoas para o serviço público (Chalfin et al, 2016)[20] ou privado (Erel et al, 2018)[21]. 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[22].

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 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

[20]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.

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

[22]A medida de valor adicionado refere-se ao incremento médio em termos de teste escore dos alunos que pode ser atribuído especificamente a um professor. É uma medida utilizada para indicar o quanto cada professor consegue elevar a performance de seus estudantes.

definido por Kleinberg et al (2015)[23].

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

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

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 à 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 (2017)[24], Kleinberg et al (2017))[25].

Nesse sentido, modelos de análise de dados devem ser utilizados com cuidado pois podem reproduzir situações sociais indesejáveis presentes 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*

[24]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

[25]KLEINBERG, J., LAKKARAJU, H., LESKOVIC, J., LUDWIG, J., e MULLAINATHAN, S. Human Decisions and Machine Predictions. NBER Working Paper nº 23180, 2017.

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[26] 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)[27]. 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 (Hastie el al (2008)[28] 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 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 0.0072σ e 0.0057σ em linguagem (ELA). Importante notar que estes ganhos se referem a ganhos sobre todos os alunos, especificamente 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 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.

[26]Na verdade, no caso americano os professores lecionavam inglês e linguagens artísticas (ELA, acrônimo em inglês).
[27]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.
[28]HASTIE, T.; JAMES, G. e WITTEN, D., e TIBSHARANI, R. An Introduction to Statistical Learning. Springer, 2008.

A variável de interesse a ser prevista é uma variável binária que indica se os policiais se envolveram em abusos físicos e verbais ou em situações com uso de arma de fogo policial. Os preditores utilizados foram fatores demográficos[29], status de veterano de guerra, status de casado, se já foi preso, se 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)[30].

O exercício de simulação semelhante foi realizado. Isto é, foi substituído os 10% piores policiais classificados pelo método de *Machine Learning* (ML) 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 públicos 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, isto é, algumas profissões podem ter *payoffs* que não são claros e nem fáceis de mensurar e ter mais do que um indicador de interesse[31]. 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 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

[29]Importante: não foram utilizados atributos raciais neste exercício.
[30]IAN GOODFELLOW ET Al. Deep Learning. MIT, 2016.
[31]Para exemplificar, de um policial esperamos que ele consiga combater e reduzir os crimes. Esse seria um indicador simples de mensurar. Entretanto, esperamos de um policial também um bom atendimento à população, auxílio em outras situações de risco aos cidadãos, orientações gerais sobre a forma de proceder em diferentes situações, entre outros. Esses outros indicadores da conduta de um policial podem ser omitidos ou não mensuráveis.

que ganhos em termos de aprendizado.

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[32].

1.3.2 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 demonstrado 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 (2017), Mulhanathan e Spisses (2017)[33]), eles enfrentam problemas clássicos para as cidades como: valoração de amenidades e políticas, mensuração de características urbanas, entre outras (Glaeser et al, 2018)[34].

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.

[32]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.
[33]MULLAINATHAN, S. and J. SPIESS. Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2):87-106, 2017.
[34]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.

Isso ajuda a melhorar a precisão das mensurações em geral, o que pode auxiliar na tomada de decisões *ex-ante*.

1.3.2.1 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[35], alguns pesquisadores têm se dedicado a buscar formas mais granulares de medidas urbanas (Naik et al (2014, 2015, 2016)[36][37][38], Glaeser et al (2018), incluir outros nomes como satélite para prever pobreza, dados telefônicos para prever pobreza). Formas mais granulares de mensuração da atividade econômica refere-se, neste contexto, a informações georreferenciadas em frequência diária, por exemplo.

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 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, como cor, brilho, tamanho, posição, etc. Ao final cada imagem foi representada por 7480 representações.

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

[36]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.

[37]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.

[38]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.

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 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[39]. 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 mensuração de características urbanas para fins de tributação predial, por exemplo.

1.3.2.2 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, é difícil associar de forma

[39]O R^2 é uma medida de associação de duas variáveis que varia entre zero (nenhuma associação) e um (associação perfeita). Quanto maior seu valor, maior é o grau de associação entre duas variáveis.

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* quanto 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.

1.3.2.3. Identificando a gentrificação

O processo de gentrificação refere-se as transformações urbanas devido às 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)[40] 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

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

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 de forma quase-contínua a criação e permanência de estabelecimentos comerciais pela cidade. 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.

1.4 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 endereçar 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 evadidos e, implica em desperdício de recursos públicos destinados ao aluno que evade (Rumberger et al., 2017)[41]. Esse problema tende a ser ainda mais grave em países pobres ou em desenvolvimento, pois, a quantidade de recursos destinados à educação é escassa e existe uma maior necessidade de aumento da acumulação de capital humano (Glewee e Muralidharan (2016), Muralidharan (2018))[42][43].

Diferentes políticas de combate à evasão têm sido testadas Muralidharan (2018). 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 à 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.

[41]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
[42]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, 2016
[43]MURALIDHARAN, K. Field Experiments in Education in Developing Countries in Handbook of Field Experiments Volume 2, edited by Abhijit Banerjee and Esther Duflo, 2018.

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 à 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)[44], Sansone (2019)[45], Aulck et al (2016)[46].

1.5 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 unitário 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: δ corresponde a parcela da população que foi diagnosticada com a enfermidade e que foi tratada ($\delta \in [0,1]$).

O médico, muitas das vezes, diagnostica um paciente por meio de uma previsão. Alguns exames são solicitados para auxiliar na decisão, mas em essência a classificação de vários tipos de

[44]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.
[45]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.
[46]AULCK, L.,VELAGAPUDI, N., BLUMENSTOCK, J. e WEST, J. Predicting student dropout in higher education. ArXivWorking Paper 1606.06364, 2016.

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 δ pode 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 à 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 à 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 (2017)[47] 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

[47]KLEINBERG, J., LAKKARAJU, H., LESKOVIC, J., LUDWIG, J., e MULLAINATHAN, S.(Working Paper). Human Decisions and Machine Predictions. NBER Working Paper , 2017.

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 (2018)[48] 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.

1.5.1 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

[48]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.

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 que 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[49].

Uma sugestão considerada por Kleinberg et al (2015)[50] 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 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 (2017[51]).

1.6 Conclusões

Este artigo apresentou uma extensa revisão da literatura sobre potencias 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

[49]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.
[50]KLEINBERG, J.; LUDWIG, J. MULLAINATHAN, S. e OBERMEYER, Z. Prediction Policy Problems. American Economic Review: Papers & Proceedings, 105(5), pp. 491–495, 2015.
[51]KLEINBERG, J., LAKKARAJU, H., LESKOVIC, J., LUDWIG, J., e MULLAINATHAN, S.(Working Paper). Human Decisions and Machine Predictions. NBER Working Paper , 2017.

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.

Espera-se que este artigo 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.

2. Desafios encontrados na construção de um indicador sobre infraestrutura de escolas no Rio Grande do Sul

Autores: *Daiane Boelhouwer de Menezes*[52], *Guilherme Rosa de Martinez Risco*[53], *Ricardo Cesar Gadelha de Oliveira Junior*[54], *Rodrigo Goulart Campelo*[55], *Thiago Felker Andreis*[56] *e Tomás Pinheiro Fiori*[57]

2.1 Introdução

Discute-se na academia internacional a influência que as condições de infraestrutura das escolas têm no aprendizado dos alunos. Ao que parece, o contexto em que os sistemas de ensino estão constituídos pode explicar, ao menos parcialmente, as diferenças encontradas entre pesquisas realizadas em países desenvolvidos e aquelas realizadas em países em desenvolvimento.

De qualquer modo, não há como negar a importância do Estado efetivamente conhecer a infraestrutura escolar disponibilizada aos alunos e, se assim for possível, estudar mais profundamente possíveis impactos desta no desempenho escolar.

Para contribuir neste sentido, o Departamento de Economia e Estatística (DEE) da Secretaria de Planejamento, Orçamento e Gestão do Estado do Rio Grande do Sul (SEPLAG) iniciou um projeto de criação de um indicador multidimensional da qualidade da infraestrutura escolar no estado. Uma das primeiras etapas deste trabalho consistiu em projeto-piloto conduzido na cidade gaúcha de Cachoeirinha, onde escolas foram visitadas e dados do Censo Escolar revalidados. Importantes dificuldades foram encontradas, as quais são compartilhadas neste artigo, que se divide em três seções.

Na primeira seção, discute-se a literatura acerca da importância da infraestrutura e se apresenta o índice multidimensional de infraestrutura escolar, com seus blocos e variáveis. Na segunda seção, é apresentado o projeto-piloto conduzido em Cachoeirinha/RS e as diferenças nos resultados encontrados a partir das inconsistências nos dados do Censo Escolar. Por fim, na terceira seção, são relatadas as principais dificuldades encontradas, as quais podem servir a outros pesquisadores que utilizem a mesma fonte de dados para seus trabalhos.

[52]Graduada em Comunicação Social (UFRGS), Mestre e Doutora em Ciências Sociais (PUCRS). Diretora Adjunta do Departamento de Economia e Estatística (DEE) da Secretaria de Planejamento, Orçamento e Gestão (SEPLAG) e Coordenadora da Divisão de Pesquisa Econômica Aplicada (DPEA). E-mail: daianemenezes@planejamento.rs.gov.br.
[53]Graduado em Ciências Econômicas (UFRGS), Mestre em Economia Aplicada (UFRGS), Pesquisador do Departamento de Economia e Estatística da Secretaria de Planejamento, Orçamento e Gestão do Rio Grande do Sul (SEPLAG/RS). E-mail: guilherme-risco@planejamento.rs.gov.br.
[54]Graduado em Ciências Sociais (UFC), Mestre em Sociologia (UFC) e Doutor em Antropologia Social (UFRGS). Analista Pesquisador em Sociologia DEE/SEPLAG. E-mail: ricardo-junior@planejamento.rs.gov.br.
[55]Graduado em Ciências Sociais (UFRGS). Analista pesquisador em Sociologia no Departamento de Economia e Estatística (DEE/SEPLAG). E-mail: rodrigo-campelo@planejamento.rs.gov.br
[56]Graduado em Ciências Jurídicas e Sociais (PUCRS) e Ciências Econômicas (UFRGS). Mestre em Ciências Sociais (PUCRS). Analista pesquisador em Economia do DEE/SEPLAG. Email: thiago-andreis@planejamento.rs.gov.br.
[57]Graduado em Ciências Econômicas (UFRGS), Mestre em Ciência Política (UFRGS), Mestre em Relações Internacionais (Institut Barcelona dEstudis Internacionals), Doutor em Economia (UFRGS). Analista pesquisador em Economia no DEE/SEPLAG. Professor Adjunto em Economia (PUCRS). Email: tomas-fiori@planejamento.rs.gov.br

2.2 Uma proposta de indicador

Não há um consenso na literatura internacional a respeito da influência direta das condições de infraestrutura das escolas no desempenho escolar. No entanto, autores brasileiros[58] apontam que tais concepções devem ser analisadas a partir de seus contextos: enquanto, nos países centrais, os sistemas de ensino são mais homogêneos, e, assim, as diferenças de infraestrutura não explicam as desigualdades de desempenho entre seus estudantes, tal correlação não se dá nos países em desenvolvimento, sendo apontadas associações positivas entre tais fatores, além das características sociais e escolares dos alunos. Uma análise de 75 estudos do mundo inteiro encontrou evidências de que ferramentas de computador de auxílio ao aprendizado têm impacto nos resultados dos exames de Matemática e de que a melhoria da infraestrutura dos prédios tem impacto, além de nos resultados dos alunos em Matemática, também em leitura e em escrita[59].

Especificamente no contexto de países de baixa e média renda, programas pedagógicos estruturados, que envolvam currículos customizados, abordagens instrucionais novas, treinamento para os professores e material educacional para os estudantes, têm o maior e mais consistente efeito em melhorar a aprendizagem. Além disso, são promissoras as intervenções de monitoramento baseadas na comunidade, tanto para o número de matrículas quanto para o aprendizado, como novas escolas e novos banheiros para aumentar a frequência, e programas de educação remediativa para melhorar os resultados da aprendizagem[60].

Em países em desenvolvimento, investimentos direcionados para a criação de condições estruturais mínimas nas escolas, bem como na provisão de equipamentos adequados, influenciam o desempenho dos estudantes. Os fatores que mostram uma relação positiva e significante entre infraestrutura escolar e resultados acadêmicos (especialmente comparecimento) incluem: presença de áreas de apoio aos professores (bibliotecas e laboratórios de ciências), utilidades como eletricidade e telefone, fornecimento de água potável, esgotamento sanitário e número adequado de banheiros. A qualidade das paredes, do piso e do telhado também possui influência positiva no comparecimento dos alunos à escola[61].

Lima[62] buscou avaliar como as características de infraestrutura impactam a proficiência dos alunos em Matemática e em Português (leitura), tendo como amostra cerca de 60 escolas das redes especial (escolas federais, incluindo os colégios de aplicação), municipal, estadual e privada, em cinco cidades (Rio de Janeiro, Salvador, Campinas, Campo Grande e Belo Horizonte), no total de 303 escolas.

[58](LIMA, 2012; MARRI *et al*., 2013).
[59](KRISHNARATNE; WHITE; CARPENTER, 2013).
[60](SNILSTVEIT *et al*., 2015).
[61](CUESTA; GLEWWE; KRAUSE, 2016; DUARTE; GARGIULO; MORENO, 2011)
[62](LIMA, 2012)

Em suas conclusões, a autora afirma que todas as variáveis analisadas tiveram algum tipo de influência, positiva ou negativa, no desempenho em Matemática e em leitura. Esses impactos se mostraram diferentes tanto nos diferentes momentos da escolarização quanto nas disciplinas. No primeiro caso, por exemplo, o impacto seria mais intenso nas crianças pequenas, que têm suas vidas escolares mais restritas às salas de aula, ficando mais dependentes da qualidade e da existência dos elementos da infraestrutura escolar. Quanto a cada uma das disciplinas, em Matemática, os equipamentos da escola mostraram-se estatisticamente significativos para o aumento da proficiência dos alunos, enquanto a existência de biblioteca apresentou aumento bem menor que os outros equipamentos; em Português, a existência de espaços didático-pedagógicos mostrou-se estatisticamente significativa positiva para a proficiência em leitura, assim como a existência de biblioteca.

Utilizando dados do Censo Escolar e do Programa de Avaliação da Rede Pública de Educação Básica (Proeb) em Matemática e Português, dos alunos dos 5.º, 9.º anos do ensino fundamental e 3.º anos do ensino médio da rede pública do Estado de Minas Gerais, Marri e outros[63] analisaram como dois grupos de variáveis — condições mínimas de funcionamento (sanitário, eletricidade, abastecimento de água) e elementos para o trabalho educativo (quadras, laboratórios de ciências e informática, biblioteca, sala de leitura, sala de professores e acesso à Internet) — influenciaram o desempenho dos alunos nas séries citadas.

No que se refere aos resultados em Português, os autores encontraram diferenças na proficiência entre as escolas com e sem as mínimas condições de infraestrutura. E, em cada uma das séries analisadas, há elementos que exercem maior influência: no 5.º ano do ensino fundamental, acesso à Internet, sala de professores, biblioteca e sala de leitura; nos 9.º ano do ensino fundamental e 3.º ano do ensino médio, além dos elementos anteriores de forma ainda mais intensa, a existência de laboratório eleva o desempenho. Em Matemática, as diferenças ocorreram a partir da presença dos quesitos de abastecimento de água e luz, para os 5.º e 9.º anos do ensino fundamental. Para o 3.º ano do ensino médio, acrescenta-se aos itens anteriores acesso à Internet, sala de professores e laboratório.

O trabalho proposto pela equipe de pesquisadores do Departamento de Economia e Estatística (DEE) da Secretaria de Planejamento, Orçamento e Gestão (SEPLAG) do Estado do Rio Grande do Sul consiste na criação de um índice multidimensional capaz de auxiliar na avaliação da infraestrutura escolar no Rio Grande do Sul. Em sua primeira versão, o índice proposto sintetiza todas as dimensões analisadas em etapas de agregação, partindo de 35 diferentes variáveis selecionadas no Censo Escolar 2018 e de sua transformação em escores padronizados para cada uma das escolas avaliadas, conforme o Quadro 1.

[63](Marri et al., 2013).

Quadro 1: Componentes do Índice de Qualidade da Infraestrutura Escolar, blocos, sub-blocos e variáveis de avaliação da infraestrutura escolar do Rio Grande do Sul

ÍNDICE	**BLOCO SAÚDE E SANEAMENTO**	Serviços básicos	4 variáveis
		Saúde	4 variáveis
	BLOCO INFRAESTRUTURA FÍSICA	Espaços de apoio educacional	5 variáveis
		Estrutura administrativa	3 variáveis
		Ambiente prazeroso	2 variáveis
	BLOCO EQUIPAMENTOS DE APOIO ADM E PEDAGÓGICO	Equipamentos de apoio pedagógico	5 variáveis
		Equipamentos de apoio administrativo	3 variáveis
	BLOCO ACESSIBILIDADE	Acessibilidade e atendimento especial	10 variáveis

Fonte: Elaboração própria

As variáveis selecionadas são majoritariamente (25 das 35) qualitativas ordinais, apresentadas em escala binária, representando a ausência (0) ou presença (1) de um atributo considerado desejável (de forma que 1 é melhor que 0). Outras quatro variáveis são incluídas em uma escala ordinal de 0 até 3, em que os números maiores igualmente representam níveis qualitativos superiores em cada atributo. Um exemplo é a coleta de lixo, para cuja ausência é atribuído 0, enquanto a presença periódica recebe 1 ou 2 se a isso se somar a reciclagem.

Por fim, seis variáveis são quantitativas contínuas, constituídas por indicadores de taxa de equipamentos por alunos e funcionários a partir das frequências extraídas dos microdados do Censo Escolar e do Sistema de Informatização (ISE) da Secretaria da Educação (Seduc) (Rio Grande do Sul, 2019). Sendo assim, a padronização desses indicadores dentro do processo de construção dos índices segue procedimentos matemáticos próprios em cada caso, como a ponderação pelos turnos de atividade escolar da entidade e a arbitragem de níveis considerados desejáveis (e.g. um computador para cada aluno ao menos em um turno da semana, ou uma impressora ou copiadora disponível para cada 10 funcionários).

A partir disso, foi construído um escore padronizado entre 0,000 e 1,000 para cada um dos oito sub-blocos, a partir das quais uma média aritmética simples leva ao índice de cada bloco e outra média aritmética simples dos blocos, por sua vez, leva ao índice final de cada entidade. O Quadro 2 apresenta a lista completa e a descrição das variáveis que compõem cada dimensão.

Quadro 2: Composição dos sub-blocos do Índice de Qualidade da Infraestrutura Escolar a partir das variáveis do Censo Escolar 2018

SUB-BLOCOS	VARÍAVEL	DESCRIÇÃO	TIPO	SUBTIPO	ESCALA
1 - Serviços básicos	IN_AGUA_REDE_PUBLICA	Rede pública de água	Quali	Ordinal	0 ou 1
	IN_AGUA_FILTRADA	Água filtrada para os alunos	Quali	Ordinal	0 ou 1
	IN_SCORE_LIXO	Coleta de lixo periódica e/ou reciclagem	Quali	Ordinal	0, 1 ou 2[1]
	IN_ESGOTO_REDE_PUBLICA	Rede pública de esgoto	Quali	Ordinal	0 ou 1
2 - Saúde	IN_COZINHA	Cozinha	Quali	Ordinal	0 ou 1
	IN_REFEITORIO	Refeitório	Quali	Ordinal	0 ou 1
	IN_BANHEIRO_CHUVEIRO	Banheiro com chuveiro	Quali	Ordinal	0 ou 1
	IN_BANHEIRO	Banheiro dentro e/ou fora do prédio	Quali	Ordinal	0 ou 1
3 - Espaços de apoio educacional	IN_LABORATORIO_INFORMATICA	Laboratório de informática			
	IN_LABORATORIO_CIENCIAS	Laboratório de ciências	Quali	Ordinal	0 ou 1
	IN_BIBLIOTECA_SALA_LEITURA	Biblioteca ou sala de leitura	Quali	Ordinal	0 ou 1
	IN_AUDITORIO	Auditório	Quali	Ordinal	0 ou 1
	IN_SCORE_QUADRA	Quadra de esportes coberta e/ou descoberta	Quali	Ordinal	0, 1, 2 ou 3(2)
4 - Estrutura administrativa	IN_SECRETARIA	Secretaria	Quali	Ordinal	0 ou 1
	IN_SALA_DIRETORIA	Sala de diretoria	Quali	Ordinal	0 ou 1
	IN_SALA_PROFESSOR	Sala dos professores	Quali	Ordinal	0 ou 1
5 - Ambiente prazeroso	IN_AREA_VERDE	Área verde	Quali	Ordinal	0 ou 1 0, 1 ou 2[3]
	IN_PATIO	Pátio interno, externo ou ambos	Quali	Ordinal	
6 - Equipamentos de apoio pedagógico	TX_POND_TV_20MAT	TV a cada 20 alunos	Quanti	Contínua	Num.
	TX_POND_MULTIM_20MAT	Equipamento multimídia a cada 20 alunos	Quanti	Contínua	Num.
	TX_POND_EQUIP_SOM_20MAT	Equipamento de som a cada 20 alunos	Quanti	Contínua	Num.
	TX_POND_COMP_MAT	Computadores por aluno por turno	Quanti	Contínua	Num.
	IN_SCORE_INTERNET	Internet	Quali	Ordinal	0, 1 ou 2[4]

7 - Equipamentos de apoio administrativo	TX_COMP_FUNC TX_IMP_COP_MULT_FUNC IN_SCORE_INTERNET	Computador por funcionário Taxa de impressoras e/ou copiadoras por funcionário Internet	Quanti Quanti Quali	Contínua Contínua Ordinal	Num. Num. 0, 1 ou 2[4]
8 - Acessibilidade e atendimento especial	IN_DEPENDENCIAS_PNE IN_BANHEIRO_PNE RAMPA CORRIMAO PORTA_VAO_LIVRE SINAL_VISUAL PISO_TATIL SINAL_TATIL SINAL_SONORO IN_SALA_ATENDIMENTO_ESPECIAL	Dependências e vias adaptadas Banheiro adaptado Rampa Corrimão Porta vão livre Sinal visual Piso tátil Sinal tátil Sinal sonoro Sala de Atendimento Especial	Quali Quali Quali Quali Quali Quali Quali Quali Quali Quali	Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal	0 ou 1 0 ou 1 0 ou 1 0 ou 1 0 ou 1 0 ou 1 0 ou 1 0 ou 1 0 ou 1 0 ou 1

Fonte: Elaboração própria

1. Zero é atribuído às escolas que declararam não possuir coleta de lixo; 1 para as que possuem coleta periódica; e 2 para as que além de possuir coleta periódica fazem reciclagem.
2. Zero é atribuído às escolas que não possuem quadras de esportes; 1 para as que possuem quadra descoberta; 2 para quadra coberta; e 3 para quem possui os dois tipos de quadra.
3. Zero é atribuído para as entidades sem pátio; 1 para quem possui pátio interno **ou** externo; e 2 para quem possui os dois tipos de pátio.
4. Zero é atribuído a quem não possui Internet; 1 para aqueles que possuem Internet **sem** banda larga; e 2 para os que possuem Internet **com** banda larga.

A aplicação deste indicador em toda a rede estadual de ensino permitirá um maior conhecimento, de maneira sintética, da situação da infraestrutura básica das escolas, auxiliando na elaboração de políticas públicas voltadas para esta área. No entanto, antes da aplicação em todas as 2.497 escolas estaduais no RS, foi elaborado um projeto-piloto para perceber possíveis dificuldades do trabalho e apresentar as correções necessárias.

2.3 Projeto-piloto

O município de Cachoeirinha[64], localizado na região metropolitana de Porto Alegre, foi escolhido para ser o primeiro a ter suas escolas estaduais avaliadas pelo índice proposto. A partir dos dados fornecidos pelas escolas ao Censo Escolar, foi atribuída uma pontuação a cada escola de acordo com a metodologia proposta e elaborou-se um *ranking* geral e um *ranking* das escolas para cada um dos blocos sugeridos — (1) saneamento e saúde, (2) infraestrutura física, (3) equipamentos de apoio pedagógico e administrativo e (4) acessibilidade.

Os resultados preliminares, utilizando-se os dados do Censo Escolar 2018, permitiram o cálculo do indicador e resultaram no *ranking* constante na Tabela 1, abaixo:

Tabela 1: *Ranking* das escolas estaduais urbanas, segundo o Índice Geral de Adequação da Infraestrutura, em Cachoeirinha — 2018

CÓDIGO DA ENTIDADE	NOME DA ENTIDADE	ÍNDICE GERAL	ORDENAMENTO
43029930	CAE Daniel de Oliveira Paiva	0,813	1.º
43029965	EEEM Osvaldo Camargo	0,698	2.º
43174116	EEEM Neuza Goulart Brizola CAIC	0,602	3.º
43029990	ETE Marechal Mascarenhas de Moraes	0,586	4.º
43029825	EEEM Guimarães Rosa	0,577	5.º
43030076	EEEM Nossa Senhora de Fátima	0,556	6.º
43029809	EEEM Governador Roberto Silveira	0,549	7.º
43029787	EEEM Presidente Kennedy	0,529	8.º
43029817	EEEB Luiz de Camões	0,527	9.º
43029914	IEE Princesa Isabel	0,501	10.º
43029957	CE Rodrigues Alves	0,456	11.º
43174108	EEEM Francisco José Rodrigues CAIC	0,448	12.º
43029981	EEEM Mário Quintana	0,347	13.º
43029795	EEEF Frederico Augusto Ritter	0,323	14.º

Fonte: INEP (2018). Fonte: Rio Grande do Sul (2019).

[64]A escolha desse município deveu-se ao fato de pertencer à Região Metropolitana de Porto Alegre (RMPA) e possuir população média em relação a mesma. Adicionalmente, todas as 14 escolas estaduais do município são urbanas e possuem tamanhos variados, abrangendo desde 123 alunos na menor até 1.899 alunos na maior. Assim, por questões de economia de recursos e de variabilidade do quadro escolar do município, Cachoeirinha foi a opção selecionada para a condução deste projeto piloto.

Munidos das respostas das escolas ao Censo Escolar 2018, uma equipe de pesquisadores do DEE/SEPLAG-RS visitou cada uma das 14 escolas objeto deste projeto-piloto e observou *in loco* as condições das escolas e possíveis divergências entre as respostas ao questionário do censo e a realidade encontrada. De fato, houve um expressivo número de respostas divergentes. Após a correção dos erros encontrados nas respostas das escolas, o indicador foi recalculado e um novo *ranking* formado, conforme a tabela 2:

Tabela 2: Novo *ranking* das escolas estaduais urbanas, segundo o Índice de Qualidade da Infraestrutura Escolar do bloco de saúde e saneamento, em Cachoeirinha — 2018

NOME DA ENTIDADE	DADOS DO CENSO ESCOLAR 2018		APÓS VISITA TÉCNICA ÀS ESCOLAS		VARIAÇÃO
	Índice	Ordenamento	Índice	Ordenamento	
CAE Daniel de Oliveira Paiva	0,900	1.º	0,900	2.º	-1
ETE Marechal Mascarenhas de Moraes	0,900	2.º	0,775	8.º	-6
EEEM Neuza Goulart Brizola CAIC	0,800	3.º	0,900	3.º	0
EEEM Governador Roberto Silveira	0,775	4.º	0,875	7.º	-3
IEE Princesa Isabel	0,675	5.º	0,775	9.º	-4
CE Rodrigues Alves	0,675	6.º	0,900	4.º	2
EEEM Osvaldo Camargo	0,675	7.º	0,775	10.º	-3
EEEM Nossa Senhora de Fátima	0,675	8.º	0,900	5.º	3
EEEM Presidente Kennedy	0,675	9.º	1,000	1.º	8
EEEB Luiz de Camões	0,675	10.º	0,900	6.º	4
EEEM Guimarães Rosa	0,675	11.º	0,775	11.º	0
EEEM Francisco José Rodrigues CAIC	0,675	12.º	0,775	12.º	0
EEEF Frederico Augusto Ritter	0,575	13.º	0,575	14.º	-1
EEEM Mário Quintana	0,300	14.º	0,775	13.º	1

Fonte: INEP (2018)

Pode-se observar, portanto, que o trabalho de confirmação das respostas realizado foi fundamental para a reelaboração do cálculo, ocasionando mudanças importantes nas posições relativas das escolas. A conclusão é que os dados do Censo Escolar, no caso de Cachoeirinha, não refletem fidedignamente muitas das características das escolas e isso ocorre por dois motivos: em primeiro lugar, por erros de preenchimento por parte das próprias escolas. Em segundo lugar, pela incapacidade do questionário de efetivamente captar a realidade das escolas.

2.4 Dificuldades encontradas

O projeto-piloto, conduzido no município de Cachoeirinha, possibilitou uma análise qualitativa profunda no desenvolvimento de um indicador quantitativo, algo que nem sempre é possível. Importantes dificuldades foram encontradas, as quais terão impacto no desenvolvimento do trabalho conduzido que se pretende aplicar em todo o estado mais adiante.

As principais dificuldades encontradas nesta etapa se dividem em dois tipos: problemas no preenchimento dos questionários pelas escolas e incapacidade dos questionários em captar adequadamente a realidade das escolas.

2.4.1 Erros nos preenchimentos dos questionários

Pesquisadores e formuladores de políticas públicas baseiam-se em dados para realização de seus trabalhos e os dados do Censo Escolar são um importante ponto de partida de qualquer análise acerca das escolas no país. Nas escolas de Cachoeirinha visitadas, a incongruência entre as 55 questões do Censo Escolar 2018 respondidas por alguém da direção com o que foi verificado pelos pesquisadores variou de 11% (seis questões não alinhadas com a realidade) a 51% (28 questões com respostas incoerentes), sendo a média delas de 26%[65]. Inclusive, a escola pior colocada no índice geral de infraestrutura escolar (sub-índice construído durante a pesquisa), com quase metade da pontuação da antepenúltima e um terço da pontuação da primeira colocada é a que tem a maior distorção entre o questionário e a realidade encontrada nas visitas.

Em uma análise de cada bloco, o que somou mais erros foi o de equipamentos (concentrando 46% dos erros), seguido de saúde e saneamento e infraestrutura (22% dos erros) e o de acessibilidade (10% dos erros) — este último bloco, porém, tem a maioria dos indicadores retirados do Sistema de Informatização da Secretaria da Educação (sete dos 10 indicadores), por isso suas respostas são mais acuradas.

Muitas vezes os erros no preenchimento dos questionários aconteciam por dificuldade dos funcionários das escolas em compreender corretamente os termos perguntados. Percebeu-se, por exemplo, que alguns diretores de escola confundiam alguns termos utilizados no formulário do Censo, o que acabou gerando respostas que não condiziam com as condições da escola. Isso se deu nas seguintes questões: se a água consumida pelos alunos nos bebedores é ou não filtrada (alguns diretores não levaram em conta se os bebedores tinham algum filtro interno); houve confusão na diferenciação entre

[65]Apenas em duas das 14 escolas esse levantamento sistemático não foi realizado.

copiadora, impressora e impressora multifuncional (em alguns casos, todos os equipamentos foram contabilizados como um dos três equipamentos); duas escolas marcaram não ter Internet banda larga, mas a possuem (na entrevista, ficou aparente que relacionaram "banda larga" a uma Internet de qualidade, que atendesse as expectativas).

Houve especialmente uma grande dificuldade em obter respostas adequadas sobre o número de computadores, uma vez que muitas escolas não levaram em conta os *netbooks*, bem como acabaram contabilizando máquinas que já não funcionavam.

Outro ponto que chamou atenção foi o erro ao reportar a inexistência de banheiros, o que parece justificar-se pela falta de atenção no preenchimento. Especificamente neste ponto, trata-se de um erro grosseiro de preenchimento que traz impacto nas estatísticas sobre as condições das escolas.

2.4.2 Incapacidade do questionário do Censo Escolar captar a realidade das escolas

Embora o questionário do Censo seja bem abrangente e consiga englobar as dependências e equipamentos fundamentais para o desempenho das funções das escolas, ele não consegue perceber as diferenças das condições de uso e de manutenção dos espaços e dispositivos, nem da quantidade disponível por aluno, como os banheiros. Essa limitação foi encontrada em praticamente todos os pontos analisados, até mesmo os mais básicos, como a existência de rede elétrica e de banheiros. Em relação à rede elétrica, todos os 14 colégios visitados possuem fornecimento regular de energia da rede pública, e isso os igualaria em tal item do indicador.

No entanto, algumas escolas visitadas relataram ter problemas com a capacidade da rede em suas dependências, o que as impossibilita de, por exemplo, instalar ou ampliar a quantidade de aparelhos de ar condicionado nas salas de aulas. Inclusive, alguns diretores mostraram que já haviam adquirido ou recebido doações de tais equipamentos, mas não conseguiam utilizá-los. Segundo os diretores das escolas com essa deficiência, havia dificuldades por parte dos alunos em assistir às aulas, sobretudo no verão, por conta das altas temperaturas.

Outro exemplo diz respeito à Internet. Da mesma forma, todos os colégios assinalaram que todos os computadores, destinados tanto aos alunos como para os trabalhos administrativos, possuíam acesso à Internet banda larga. Porém, algumas escolas relataram que o pacote de acesso fornecido por um contrato do Governo do Estado é insuficiente para atender as necessidades, devido à baixa velocidade. Assim, o acesso acaba sendo priorizado aos funcionários das escolas que realizam as tarefas administrativas. Tal fato acarreta na quase inutilização dos computadores e *netbooks* dos laboratórios de informática, pois os professores não conseguem levar as turmas completas para realizar alguma

atividade nesses espaços, e mesmo quando fazem revezamentos entre os alunos, a velocidade da Internet ainda é insuficiente.

Igualmente, todas as escolas pelo questionário apresentam condições semelhantes de abastecimento de água e esgoto sanitário. Mas, durante as visitas, verificamos que algumas apresentavam problemas hidráulicos que acarretando vazamentos que inutilizavam algumas salas de aula em períodos de chuva. O caso mais grave se dá na EEEM Neuza Goulart Brizola CAIC, em que uma sala do térreo não pode ser utilizada por conta de um vazamento, que levou ao descolamento de todo o piso e a um forte cheiro de mofo. Fato semelhante se deu em relação aos banheiros. Todas as escolas apresentavam banheiros, o que as daria a pontuação máxima nesse quesito, porém encontramos realidades muito díspares, tanto em termos de qualidade quanto de quantidade de vasos sanitários disponíveis.

2.5 Conclusões

A importância do desenvolvimento de novas metodologias de avaliação de políticas públicas é inegável e extremamente necessária em um país com recursos tão escassos e demandas não-atendidas. O trabalho proposto pelo DEE/SEPLAG-RS possui grande relevância para a elaboração de diagnósticos mais precisos sobre a educação no RS, no entanto, precisa se basear em dados precisos. O projeto-piloto conduzido no município gaúcho de Cachoeirinha revelou-se uma oportunidade para perceber graves erros de preenchimento nos questionários do Censo Escolar por parte das escolas, o que dificulta o trabalho do pesquisador.

Como se não fosse suficiente, observou-se a incapacidade do referido questionário em captar questões relacionadas ao estado de conservação e efetivo uso das facilidades nas escolas. Assim, para fins de Censo Escolar, duas escolas com realidades completamente diferentes em relação às condições de conservação de seu patrimônio aparecem, para o pesquisador, como se possuíssem as mesmas condições.

É preciso, portanto, trabalhar no sentido de aumentar a fidedignidade dos dados em relação à realidade e buscar alternativas para lidar com essas questões. A baixa qualidade dos dados do Censo Escolar pode prejudicar esforços de pesquisa que se baseiem centralmente nesses dados, obrigando os pesquisadores a utilizar outras fontes de dados complementares ou mesmo a mudar suas perguntas de pesquisa.