Journal of Data Science 0 (0), 1–10
September 2024
DOI: 10.6339/24-JDS1118
Data Science in Action

Impacts of COVID-19 on Public Universities in Brazil: A Machine Learning Counterfactual Analysis

R. Rossi Jr.[1]

[1] *Universidade Federal de Viçosa - Instituto de Ciências Exatas e Tecnológicas - CAF, LMG818 Km6, Minas Gerais, Florestal 35690-000, Brazil*

Abstract

This study delves into the impact of the COVID-19 pandemic on the enrollment rates of on-site undergraduate programs within Brazilian public universities. Employing the Machine Learning Control Method, a counterfactual scenario was constructed in which the pandemic did not occur. By contrasting this hypothetical scenario with real-world data on new entrants, a variable was defined to characterize the impact of the COVID-19 pandemic on on-site undergraduate programs at Brazilian public universities. This variable reveals that the impact factor varies significantly when considering the geographical locations of the institutions offering these courses. Courses offered by institutions located in smaller population cities experienced a more pronounced impact compared to those situated in larger urban centers.

Keywords *counterfactual approach; educational data-mining*

1 Introduction

The most common function of a machine learning model is to learn patterns and glean insights from data, subsequently making predictions and classifications. Machine learning models find utility in a multitude of domains, including demand forecasting, stock price prediction, weather forecasting, email spam detection, document categorization, medical diagnosis, and pattern recognition in images, among others. The capability to generate forecasts based on past data allows us to assess contractual scenarios and consequently evaluate the causal impact of certain events when it is not possible to identify a control group, a group that was not affected by the event. The case of the COVID-19 pandemic serves as a clear example of this situation.

The COVID-19 pandemic was a global event that resulted in a series of deaths and profoundly affected nearly all activities. Since its onset in 2019, the world faced unprecedented challenges, from overwhelming healthcare systems to disruptions in global supply chains. In the field of public higher education in Brazil, it was no different. Numerous unplanned adaptations were made to ensure the continuity of courses. Furthermore, several on-site undergraduate programs in public institutions were impacted by a decrease in the number of new students.

Ideally, to quantify the impact caused by the pandemic on the number of new entrants in higher education on-site courses at Brazilian public universities, it would be necessary to devise a way to compare what the new entrants numbers would have been if the pandemic had not occurred with the new entrants numbers in a scenario where the pandemic did occur. However, this comparison is impossible because there are no universities that experienced a situation where the pandemic had no impact.

Received September 29, 2023; Accepted January 17, 2024

This is where the capabilities of machine learning models come into play for projecting scenarios based on past data. By making projections solely using pre-pandemic data through a machine learning model, we can generate a projection that corresponds to a counterfactual scenario in which the COVID-19 pandemic never happened. In this manner, we can evaluate the impact caused by the COVID-19 pandemic.

Recent developments have witnessed the integration of machine learning with causal inference techniques. This integration has been explored by researchers such as (Athey and Imbens, 2016; Athey et al., 2019, 2021; Belloni et al., 2017; Hofman et al., 2021; Varian, 2014, 2016; Wager and Athey, 2018). The work of Varian (2014, 2016) was among the early advocates of the idea that constructing counterfactual scenarios essentially involves making predictions. In practical terms, especially in settings involving panel data or time series, Varian proposed leveraging pre-treatment data to create a synthetic control group that mimics the scenario of no treatment or a "business-as-usual" situation. This method allows for the estimation of treatment effects by comparing observed outcomes to the potential outcomes generated using machine learning. It utilizes pre-treatment data to create a reliable estimate of how the treated group would have performed if they had not received treatment.

Recently, there have been initial practical applications of this counterfactual methodology in empirical studies (Benatia, 2020; Benatia and de Villemeur, 2019; Bijnens et al., 2019; Burlig et al., 2020; Cerqua et al., 2021; Souza, 2019). The majority of these research efforts encounter a common challenge, which is the absence of an original control group, similar to our case. The works of Benatia (2020); Cerqua et al. (2021); Cerqua and Letta (2022), delve into the causal consequences of the COVID-19 crisis. In Benatia (2020) investigation, a neural network model is employed to examine the effects of containment measures on the reduction in demand within New York's electricity markets. Meanwhile, Cerqua et al. (2021) employ three distinct machine learning approaches (LASSO, random forest, and stochastic gradient boosting) to extrapolate excess mortality estimates at the municipality level during the initial wave of the COVID-19 pandemic in Italy.

In the field of higher education we can highlight the work of Kotosz (2016). Employing a counterfactual method, their research delves into the economic impact engendered by higher education institutions within their regional spheres. Their work introduces an approach to gauging this impact, underscored by an insightful analysis of the University of Lorraine (France), University of Szeged (Hungary), and two smaller Hungarian colleges. In (Svábová et al., 2021) the authors evaluate the effects of the Graduate Practice in Slovakia, a program targeted at lowering the unemployment rate among recent graduates. They employ a counterfactual approach, comparing the results of treated and non-treated individuals, using three widely employed methods: regression adjustment, instrumental variable, and propensity score matching method.

In this study, we investigate the impact of the COVID-19 pandemic on the number of new entrants in on-site undergraduate courses at Brazilian public universities. The number of incoming students in higher education is a critical parameter for assessing the feasibility of undergraduate programs. This parameter is employed in defining metrics for evaluating higher education courses in Brazil, such as the Indicator of Difference Between Observed and Expected Performances (IDD), calculated by The National Institute of Educational Studies and Research Anísio Teixeira (INEP). A machine learning model was employed to construct projections based on data from the years preceding the pandemic and compared them with the actual data for the years 2020 and 2021. Through this comparison, we determined a pandemic impact factor on the number of new entrants in undergraduate programs.

We observed that the impact factor varies significantly when considering the geographic locations of the institutions offering these courses. Courses provided by institutions located in smaller population cities were more profoundly affected than those in larger urban centers. Furthermore, we found that these disparities also lead to inequalities in the distribution of the impact factor when comparing different majors. This discrepancy can be attributed to the fact that majors more commonly offered in smaller cities experienced a greater impact compared to those offered in larger urban areas.

2 Methodology

2.1 Data

The datasets used in this study were produced by The National Institute of Educational Studies and Research Anísio Teixeira (INEP, 2022). INEP conducts an annual Higher Education Census, which is the most important research instrument in Brazil regarding higher education institutions offering undergraduate courses, as well as their students and faculty. We utilized data from the censuses conducted from 2012 to 2021 for public university on-site undergraduate programs. The original dataset comprises 200 features, with each row representing a course. These features can be broadly categorized into four groups. The first group contains information about the geographical location of the course, the second group contains details about the institution offering the course, the third group contains information about the course's field and type, and the fourth group consists of course statistics, composed of 181 features. In this fourth group, we retained only the key features, which provide statistics on the number of available seats, the number of applicants, the number of enrolled students, the number of graduates, the number of students who interrupted their studies, the number of disengaged students, the number of transfers, and the most important feature for this study, the number of new entrants.

The correlations between the number of new entrants and the other features were computed, and only those exhibiting a correlation coefficient exceeding 0.15 were included as input variables in the subsequent machine learning models. This selection criterion ensures that the models incorporate only the most influential factors in predicting the outcomes.

After the data cleaning phase, we obtained a panel data set containing information from 2012 to 2021. Rows corresponding to data for courses with dates up to 2019 were separated from those corresponding to 2020 and 2021. Only pre-pandemic data was utilized in the machine learning model for the construction of the counterfactual scenario. The analysis of the impact caused by the pandemic primarily focuses on the number of new entrants as the key feature. The aim of this study is to assess whether there were significant changes in the behavior of this variable when comparing the actual values obtained in 2020 and 2021 to the projections provided by the machine learning model, which exclusively utilizes pre-pandemic data.

In the feature engineering phase, we introduced lagged features for the features in the fourth group. We created lagged features corresponding to two data shifts, namely shifts of two years and three years. The selection of these data shifts ensures that only pre-pandemic period data is used in the machine learning model.

In addition to the census data provided by INEP, we also incorporated data from the Brazilian Institute of Geography and Statistics (IBGE). Data from the Brazilian Institute of Geography and Statistics (IBGE, 2022) was included as a new variable representing the population of each municipality.

2.2 Machine Learning Control Method

The approach employed in this article to evaluate the impact of the COVID-19 pandemic involves constructing a counterfactual scenario using a Machine Learning model, as indicated by (Cerqua et al., 2021; Varian, 2016; Burlig et al., 2020; Cerqua and Letta, 2022). The term "Machine Learning Control Method" (MLCM), introduced by Cerqua et al. (2021), denotes this particular approach. The term is widely adopted terminology, in this context MLCM refers to the process of generating a counterfactual scenario through the utilization of a Machine Learning model.

MLCM allows us to construct a scenario in which the COVID-19 pandemic did not exist. This scenario is developed based on the projection of the number of students entering undergraduate courses in 2020 and 2021, based exclusively on data prior to the pandemic period, that is, until 2019. When making predictions based on data prior to the pandemic, ML models are not influenced by changes caused by the pandemic. Therefore, the resulting projection can be interpreted as a counterfactual scenario in which the pandemic did not occur.

By comparing the number of students entering undergraduate courses designed by the ML model (which represents the counterfactual scenario) and the real numbers obtained from the INEP database, it is possible to quantify the impact of the pandemic. To develop this scenario, the following steps were used:

1) Selection of the ML model.
2) Hyperparameter tuning of the selected model.
3) Application of the selected model in the 2020–2021 sample, predicting the number of incoming students in a no-COVID scenario.

In the first stage, four Machine Learning models widely used in regression problems were applied: Linear Regression, XGBoost, Extra Trees and Random Forest. The Machine Learning models were implemented using the scikit-learn library for Python (for Linear Regression) and the XGBoost and scikit-learn libraries (for XGBoost, Extra Trees, and Random Forest, respectively).

To evaluate the models, the method of cross-validation on a rolling basis was used. When dealing with panel data, we cannot apply the usual cross-validation method. We cannot randomly split the data into training and testing sets because it is illogical to use future values to predict past values. There is a temporal relationship between observations, and we must maintain that connection during testing. Therefore, a 3-fold cross-validation was considered in the pre-2020 data. The three pairs of training/test sets are:

1º) Train (2014, 2015) – Test (2016, 2017).
2º) Train (2014, 2015, 2016) – Test (2017, 2018).
3º) Train (2014, 2015, 2016, 2017) – Test (2018, 2019).

The method cross-validation on a rolling basis was used for each machine learning models (Linear Regression, XGBoost, Extra Trees and Random Forest). The average Mean Absolute Error (MAE) is calculated for each model, the results are shown in Table 1. A performance shows that the Extra Trees model obtained the best performance. The default settings for each method was employed during this stage of analysis, as it was done in references (Cerqua et al., 2021; Cerqua and Letta, 2022). It's important to highlight that at this stage, only data from the pre-2020 period were used.

In the second step, to tune the hyperparameters of the Extra Trees model, the same method of cross-validation on a rolling basis was used, with the same three pairs of training/test sets.

The hyperparameter tested were *max features* and *n estimators*. The "max features" denotes the maximum number of features considered for each split, and "n estimators" represents the number of trees in the forest. These two parameters, as outlined in the scikit-learn

Table 1: Selection of the ML model.

Models	Average MAE
Linear Regression	8.58
XGBoost	8.07
Random Forest	7.56
Extra Trees	7.45

Table 2: Hyperparameter tuning.

Parameters	Average MAE
max features: None, n estimators: 500	7.40
max features': None, n estimators: 1000	7.40
max features: None, n estimators: 250	7.41
max features: 1, n estimators: 500	7.57
max features: 1, n estimators: 1000	7.57
max features: 1, n estimators: 250	7.60

documentation, stand out as primary factors to fine-tune when employing the Extra Trees method.

The rolling cross-validation were applied to six different hyperparameter settings for the Extra Trees model, the results are shown in Table 2. This method allows us to choose the optimal hyperparameter setting.

In the third step, data from 2014 to 2019 were considered in the training set. The optimal Extra Trees model was employed to project the number of students entering undergraduate courses in both 2020 and 2021.

To quantify the impact of COVID-19 on public universities in Brazil, we defined a factor based on the projections obtained in a counterfactual scenario. For the 2020 dataset, the impact factor is defined as the difference between the percentage change in the actual number of incoming students and the percentage change in the counterfactual scenario: $(n_{20} - nc_{20})/n_{19}$, where n_{20} represents the actual number of incoming students in 2020, nc_{20} is the projection for incoming students in 2020 under a no-COVID scenario, and n_{19} is the number of incoming students in 2019.

For the 2021 dataset, the impact factor is calculated as follows: $(n_{21} - nc_{21})/nc_{20}$, where n_{21} represents the actual number of incoming students in 2021, and nc_{21} is the projection for incoming students in 2021 under a no-COVID scenario. This definition employs nc_{20} instead of n_{20} as a variable to capture the percentage change in the number of incoming students in 2021 in a situation where the number of incoming students was not affected by the pandemic in 2020.

To mitigate uncertainties inherent in the optimal Extra Trees model projections, we implemented a filtering process within the dataset. Specifically, we excluded rows corresponding to undergraduate programs where the absolute difference between the actual and counterfactual projected incoming student numbers fell below the Mean Absolute Error (MAE). This strategy aimed to preserve the directionality of impact, whether positive or negative, by ensuring that deviations beyond the MAE were considered. It's crucial to note that the excluded undergraduate programs weren't those predicted accurately; rather, they were programs less affected by the

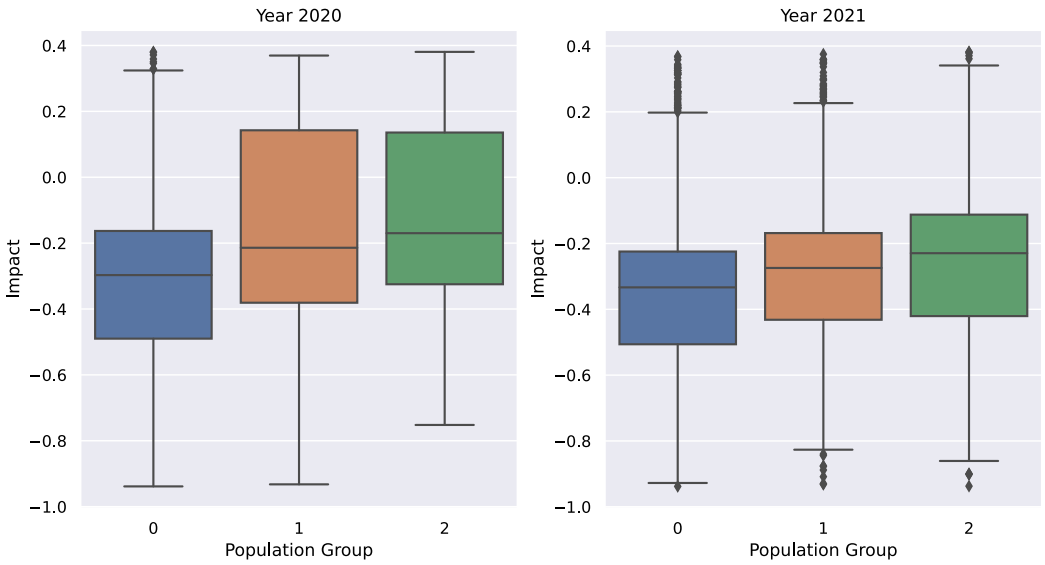

Figure 1: Impacts of COVID-19 on undergraduate on-site courses across city groups.

pandemic. Consequently, for these excluded programs, definitively determining whether their impact was positive or negative wasn't possible.

The management of prediction uncertainty is essential in the accurate interpretation of results derived from Machine Learning Counterfactual Methods (MLCM). In this study, the Mean Absolute Error (MAE) assumes a critical role as a metric for quantifying uncertainty inherent in the predictive model. Our approach utilizes MAE to filter the dataset, ensuring a focus on cases where deviations between actual and counterfactual projections exceed the average error. This meticulous process acknowledges the variability in predictions, enabling a more discerning analysis of pronounced impacts caused by the COVID-19 crisis. The exclusion of cases with deviations within the MAE threshold, aims to emphasize the significance of effects while recognizing the limitations in conclusively determining impact directionality for cases closer to predicted values. This methodological clarity provides an understanding of how uncertainty is handled, offering insights into the strengths and limitations of the MLCM framework.

3 Results

The analysis of the counterfactual scenario reveals heterogeneous impacts of COVID-19 on undergraduate on-site courses. Particularly, when taking into account the population of the city in which the course is offered, we notice that the negative impact is more relevant in small cities. This information is illustrated in Figure 1. The "0" group encompasses cities within the bottom one-third percentile regarding population, in comparison to the populations of the cities where the courses are offered. The group "1" represents cities falling between the one-third and two-thirds percentiles in population, relative to the cities where the courses are offered. Finally, the group "2" pertains to cities situated within the top one-third percentile in terms of population compared to the populations of the cities where the courses are offered. Figure 1 displays a boxplot on the right representing the data for the year 2020 and, on the left, another boxplot corresponding to the data for the year 2021.

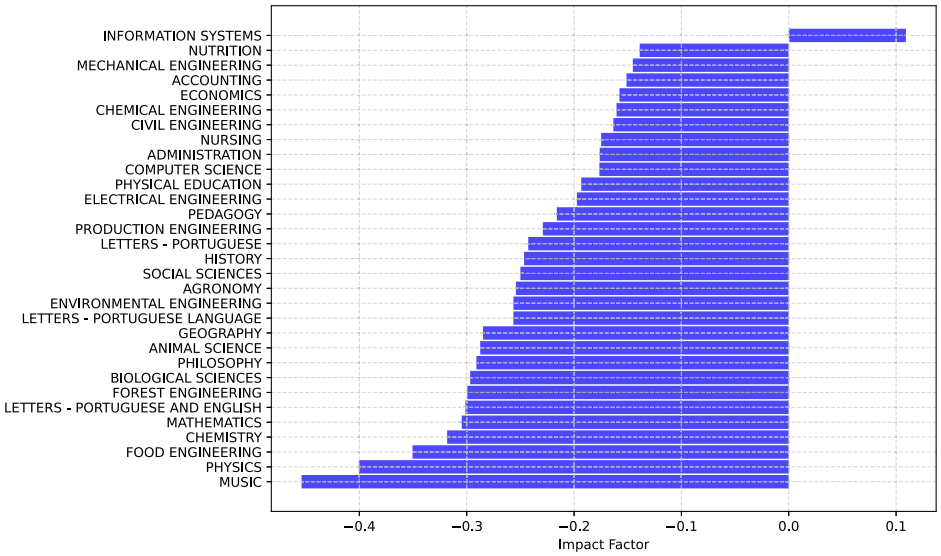

Figure 2: The graphic shows the median impact factor for different majors, considering data from 2020 and 2021. The majors on the list are the top 10% with the highest number of appearances in the dataset.

The majority of university students in small cities do not reside in those cities. Typically, they relocate from other areas to the city where the university is situated. This transition entails a substantial shift in the lives of students and their families. Consequently, effective planning is crucial for this process. Such planning relies on a degree of stability. Unfortunately, the pandemic created an atmosphere of profound uncertainty, impacting various social and economic facets. Consequently, many prospective students and their families might have reconsidered making such a significant life change within the context of the uncertainty prevalent in 2020 and 2021. This could elucidate why undergraduate programs in smaller cities experienced a more pronounced negative impact.

Another aspect of the heterogeneity in the impacts of COVID-19 is observed among different majors. As depicted in Figure 2, certain majors exhibited a significantly high impact factor, while others demonstrated relatively modest or even positive impacts. At first glance, the underlying reason for the disparities in the impacts observed in different majors does not immediately become evident. Establishing a causal link between the pandemic, its economic consequences, and the variation in the magnitude of these impacts across different majors is not a straightforward task. However, a careful analysis reveals that the concentration of the courses most affected by the pandemic is notably higher in smaller cities, categorized as category 0.

Among the programs depicted in Figure 2, the top 20 most impacted majors collectively account for 56% of the total majors offered in smaller cities. In contrast, in larger cities, this percentage stands at 34%. Therefore, it is apparent that the concentration of the pandemic's impact is considerably greater in smaller cities, falling under category 0. The most heavily impacted majors are more affected because they are widely offered in small cities. Thus, the geographical location emerges as one of the contributing factors to the disparity in impact among these majors.

To investigate this hypothesis, we can compare the impact distribution among majors on a global scale, considering all population groups (0, 1, and 2), with the impact distribution

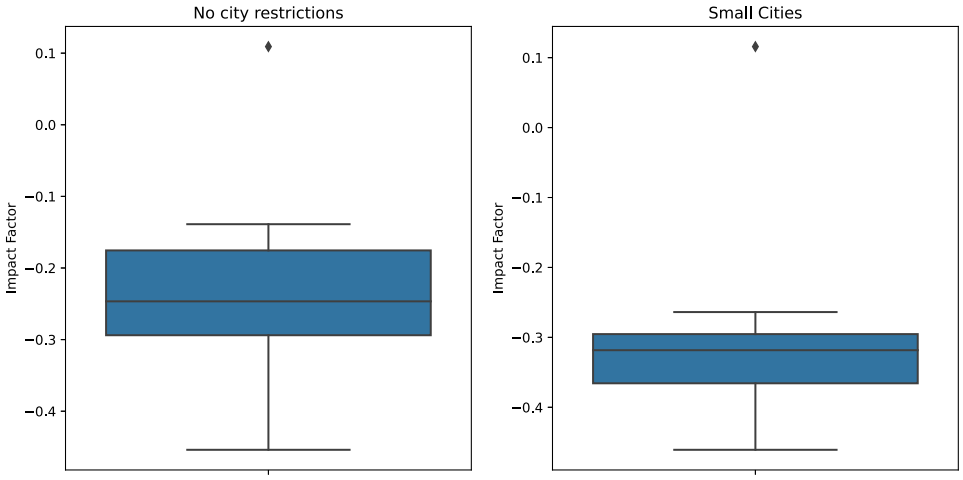

Figure 3: The boxplot shows the impact distribution among majors, without any city restrictions on the left and exclusively within small cities on the right, considering data from 2020 and 2021.

solely within small cities (group 0). In the Figure 3, we present a boxplot illustrating the impact distribution among majors without any city restrictions on the left, and on the right, a boxplot displaying the impact distribution exclusively within small cities. Notably, the distribution appears more homogeneous in the right boxplot, indicating that when we focus solely on small cities, the heterogeneity in the impacts of COVID-19 across different majors diminishes compared to the general case.

To quantitatively assess this disparity, we employ the Interquartile Range (IQR). In the general case, the IQR is 0.12, whereas in the case of small cities, it narrows down to 0.07. A smaller Interquartile Range implies reduced variation in values, in contrast to a larger IQR associated with greater variability. Thus, the distribution pertaining to small cities concentrates its values within a narrower range, signifying less data variability compared to the general case.

Consequently, we can conclude that by considering only data from small cities, we remove the location factor and focus exclusively on the different majors. Under this context, the majors exhibit similar impact levels with less variation. This strengthens the hypothesis that the concentration of certain courses in small cities is a contributing factor to the variation in impact among different majors.

4 Conclusion

In this study, we have presented an analysis of the impact of the COVID-19 pandemic on the number of new entrants in on-site undergraduate courses at Brazilian public universities. The assessment of this impact was conducted by constructing a counterfactual scenario in which the pandemic did not occur. The creation of this scenario was made possible through the use of an approach known as the Machine Learning Control Method (MLCM). Our evaluation revealed that the impact of the pandemic was uneven, particularly when considering the diverse geographic locations of the institutions offering undergraduate programs. Courses offered by institutions in less densely populated cities experienced a more pronounced impact. Furthermore,

these disparities also manifested in the unequal distribution of the impact factor when comparing different academic majors.

Considering the available data, it is still uncertain whether the trends observed in this study will persist in the coming years. It is plausible that the pandemic may have accelerated the ongoing trend toward increased availability of online courses, and this may continue in the future. In such a scenario, the reduction in the number of new entrants in on-site undergraduate programs could be sustained or even exacerbated. However, other factors such as the quality of online courses offered, the employability of graduates from online programs, changes in the country's economic landscape, among others, could modify the trends observed in this study.

Supplementary Material

The following files are included in the supplementary material: (1) Study code file; (2) URL to INEP census data; (3) IBGE population of each municipality data.

References

Athey S, Bayati M, Doudchenko N, Imbens GW, Khosravi K (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536): 1716–1730. https://doi.org/10.1080/01621459.2021.1891924

Athey S, Bayati M, Imbens GW, Qu Z (2019). Ensemble methods for causal effects in panel data settings. *American Economic Association Papers and Proceedings*, 109: 65–70.

Athey S, Imbens GW (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27): 7353–7360. https://doi.org/10.1073/pnas.1510489113

Belloni A, Chernozhukov V, Fernandez-Val I, Hansen C (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1): 233–298. https://doi.org/10.3982/ECTA12723

Benatia D (2020). Reaching new lows? The pandemic's consequences for electricity markets. *United States Association for Energy Economics Working*.

Benatia D, de Villemeur E (2019). Strategic reneging in sequential imperfect markets. *Center for Research in Economics and Statistics Working Papers*, 19.

Bijnens G, Karimov S, Konings J (2019). Wage indexation and jobs. a machine learning approach. VIVES Discussion Paper (82).

Burlig F, Knittel C, Rapson D, Reguant M, Wolfram C (2020). Machine learning from schools about energy efficiency. *Journal of the Association of Environmental and Resource Economists*, 7(6): 1181–1217. https://doi.org/10.1086/710606

Cerqua A, Di Stefano R LM, Miccoli S (2021). Local mortality estimates during the COVID-19 pandemic in Italy. *Journal of Population Economics*, 34: 1189–1217. https://doi.org/10.1007/s00148-021-00857-y

Cerqua A, Letta M (2022). Local inequalities of the COVID-19 crisis. *Regional Science and Urban Economics*, 92: 103752. https://doi.org/10.1016/j.regsciurbeco.2021.103752

Hofman J, Watts D, Athey S, Garip F, Griffiths T, Kleinberg J (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866): 181–188. https://doi.org/10.1038/s41586-021-03659-0

IBGE (2022). Prévia da população dos municípios com base nos dados do censo demográfico de 2022 coletados até o dia 25/12/2022. Rio de Janeiro: IBGE.

INEP (2022). Censo da educação superior 2022. Institute of Educational Studies and Research Anísio Teixeira.

Kotosz B (2016). University impact evaluation: Counterfactual methods. In: *56th Congress of the European Regional Science Association*. European Regional Science Association.

Souza M (2019). Predictive counterfactuals for treatment effect heterogeneity in event studies with staggered adoption. *Social Science Research Network Electronic Journal*, https://doi.org/10.2139/ssrn.3484635.

Svábová L, Kramárová K, Durica M (2021). Evaluation of the effects of the graduate practice in Slovakia: comparison of results of counterfactual methods. *Central European Business Review*, 10(4): 1. https://doi.org/10.18267/j.cebr.266

Varian HR (2014). Big data: new tricks for econometrics. *The Journal of Economic Perspectives*, 28(2): 3–28. https://doi.org/10.1257/jep.28.2.3

Varian HR (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27): 7310–7315. https://doi.org/10.1073/pnas.1510479113

Wager S, Athey S (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242. https://doi.org/10.1080/01621459.2017.1319839