UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

LUCAS DIERINGS TANUS DOS SANTOS

PUBLIC SPENDING IMPACT ON SHORT TERM GROWTH: A MACHINE LEARNING APPROACH

Porto Alegre
2021

LUCAS DIERINGS TANUS DOS SANTOS

PUBLIC SPENDING IMPACT ON SHORT TERM GROWTH: A MACHINE LEARNING APPROACH

Dissertação submetida ao Programa de Pós-Graduação em Economia da UFRGS, como quesito parcial para obtenção do grau de Mestre em Economia.

Orientador: Prof. Dr. Flavio Augusto Ziegelmann

Porto Alegre
2021

CIP - Catalogação na Publicação

Tanus, Lucas
Public Spending Impact on Short Term Growth: A Machine Learning Approach / Lucas Tanus. -- 2021.
35 f.
Orientador: Flávio Ziegelmann.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Programa de Pós-Graduação em Economia, Porto Alegre, BR-RS, 2021.

1. Causal Random Forest. 2. Public Spending Multiplier. 3. Heterogenous Treatment Effect. I. Ziegelmann, Flávio, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

LUCAS DIERINGS TANUS DOS SANTOS

PUBLIC SPENDING IMPACT ON SHORT TERM GROWTH: A MACHINE LEARNING APPROACH

Dissertação submetida ao Programa de Pós-Graduação em Economia da UFRGS, como quesito parcial para obtenção do grau de Mestre em Economia.

Aprovada em: Porto Alegre, 19 de outubro de 2021.

BANCA EXAMINADORA:

Prof. Dr. Flavio Augusto Ziegelmann– Orientador
UFRGS

Prof. Dr. Marcelo Portugal
UFRGS

Prof. Dra. Taiane Schaedler Prass
UFRGS

Prof. Dr. Evandro Konzen
University of Exeter

ABSTRACT

The public spending multiplier has long been a subject of analysis with central discussion on how its size varies under different economic contexts. The article that integrates this dissertation introduces a causal machine learning technique as a tool to estimate the public spending multiplier and make individual predictions based on each country's economic context. We propose to model the multiplier with a causal random forest, developed by Wager e Athey (2018), uncovering possible heterogeneous treatment effects. We apply this methodology to a dataset provided by the International Monetary Fund, including data from 35 developed countries for the years from 2000 to 2020. The multiplier estimates obtained with this methodology are between 1.7 and 2.7. In addition, we use this methodology as a tool to uncover which features are important to the multiplier heterogeneity.

Keywords: Causal Random Forest. Public spending multiplier. Unconfoundedness. Heterogeneous treatment effect.

RESUMO

O multiplicador do gasto público é objeto de análise há muito tempo, com a discussão centrada em como seu tamanho varia em diferentes contextos econômicos. No artigo que integra esta dissertação, apresentamos uma técnica de aprendizado de máquina causal como uma ferramenta para estimar o multiplicador do gasto público e fazer previsões individualizadas com base no contexto econômico de cada país. Propomos modelar o multiplicador com uma floresta aleatória causal, desenvolvida por Wager e Athey (2018), descobrindo possíveis efeitos de tratamento heterogêneos. Aplicamos essa metodologia em um conjunto de dados fornecido pelo Fundo Monetário Internacional, incluindo dados de 35 países desenvolvidos ao longo dos anos de 2000 a 2020. As estimativas dos multiplicadores obtidas com esta metodologia estão entre 1,7 e 2,7. Além disso, usamos essa metodologia como uma ferramenta para descobrir quais recursos são importantes para a heterogeneidade do multiplicador.

Palavras-chave: Causal Random Forest. Multiplicador do gasto público. Unconfoundedness. Efeito de tratamento heterogêneo.

List of Figures

List of Tables

Contents

1 INTRODUCTION

The relationship between public spending and the economic growth output has long been a subject of analysis and debate. The analysis bears upon the question of the government's role in economic growth. If changes in the share of government spending can affect the output growth rate, the size of the government can be a potentially important factor explaining the observed disparity in growth rates among different countries.

During the last several years, the literature has centered the discussion about this topic in the public spending multiplier, that is, the change in output caused by a $1 change in government spending. An important part of this literature explores the fact that the government multiplier might vary depending on the circumstances.

One strand of the literature, such as Coenen et al. (2012), explores New Keynesian DSGE models to show that the multiplier can be higher when the interest rates are close to the zero lower bound, and monetary policy is less effective. Cloyne, Jordà e Taylor (2020) also explores how the multiplier might be radically different depending on the monetary factors.

Another trend inside this literature explores how the multiplier might be different during economic recessions, such as Auerbach e Gorodnichenko (2012) and Riera-Crichton, Vegh e Vuletin (2015) or even if the multiplier differs according to the amount of slack in the economy, as in Ramey e Zubairy (2018).

There are also some articles that argue that an important determinant of the government spending multiplier is the direction of the fiscal intervention, such as Barnichon e Matthes (2017), that uses a theoretical model to show that the multiplier associated with a positive fiscal shock is smaller than the one associated with a negative change in public spending. Alesina et al. (2018) also explores this point, indicating that a fiscal contraction might have smaller costs than tax-based reforms, with multipliers typically below one during fiscal adjustments.

Most of the studies about this topic estimate the multiplier based on averages for a particular period of time and some specific country, or develop theoretical models or empirical strategies to estimate how the multiplier varies depending on the specific context (ALESINA et al., 2018; RAMEY; ZUBAIRY, 2018; AUERBACH; GORODNICHENKO, 2012; RIERA-CRICHTON; VEGH; VULETIN, 2015; COENEN et al., 2012; CLOYNE; JORDÀ; TAYLOR, 2020).

Theory tells us that the economic context can significantly impact the magnitude of the multiplier, including factors like how the spending is financed, how monetary policy reacts, the persistence of spending changes or the direction of the fiscal policy change. However, the fact that we cannot perform controlled, randomized trials on countries or economies implies that the majority of the empirical analysis in economics is dependent on historical happenstance and historical data (ATHEY; IMBENS, 2017). This is also valid for empirical estimates of the multiplier, but unfortunately the data does not offer us clean natural experiments to explore the way the multiplier varies under different scenarios.

In cases like these, where an average treatment effect of a policy (in this case the public spending) is not informative enough to develop optimal policy, we may be interested in estimating the heterogeneous treatment effect (HTE). HTE is observed when exposure to the same policy results in different effects on individuals, based on their characteristics. Methods to estimate HTE are gaining much interest in clinical research, corporate world and in economic research, as a source of knowledge to define optimal policy strategy (POWERS et al., 2018).

The literature already explores the way that the multiplier varies depending on economic characteristics, but the exploration of the heterogeneity is made mainly with the researcher's expertise, focusing on some specific features. On the one hand, the dimension of the data in this problem results in too many potential candidate groups for heterogeneity and makes it nearly impossible to analyze properly without a data driven estimation tool. On the other hand, the process of searching for groups with different effects invalidates the statistical inference. To infer properly, we need to avoid cases where the researcher or policymaker mines through the data to find sub-groups where the multiplier is maximized and overstates the average treatment effect by testing on those specific sub-groups only (COOK; GEBSKI; KEECH, 2004).

This paper contributes to the literature by introducing the causal forest, developed in Wager e Athey (2018), as a tool to uncover heterogeneity in the public spending impact, using high dimensional data to make individual predictions of the causal effect of marginal public spending on the short term growth.[1]

The results we obtain include confidence intervals for the public spending impact on short-term GDP, centered between 1.7 and 2.7 for 2019 and 2020 for different countries, and a ranking of feature importance for determining the heterogeneity in the effects. We find that one really important feature for the multiplier is the amount of revenue obtained by the government as a percentage of GDP.

The rest of the paper is organized as follows. Section 2 introduces the traditional causality framework and the nascent literature about causal machine learning, presenting the causal tree and causal forest that we use in this paper. Empirical estimates for the multiplier are presented in Section 3. The dataset description and the tables with all the results are available in the Appendix.

[1] The data we use in the work is from a database provided by the International Monetary Fund, named World Economic Outlook. It has a wide variety of data from different countries. As the causal forest is a data mining strategy, we include in the analysis data on fiscal, financial, foreign trade, among other aspects.

2 METHODOLOGY

2.1 Treatment Estimation

One of the main references in causality is the so-called Rubin causal model, first named like that by Holland (1986), or also called the “potential outcome" approach, which has been widely used in the field of econometrics since the early 1990s.

In this approach, for each unit i and each treatment level represented by ω, there is a potential result, represented by $Y_i(\omega)$, which describes the value of the response variable, according to the level of treatment. In this model, the causal effect $\tau_i(\omega)$ would be the difference between the results with the treatment and without the treatment, for each unit i:

$$\tau_i(\omega) = E[y_i^{(1)} - y_i^{(0)}].$$

Our goal is to estimate this function $\tau(x)$. However, in practice, it is only possible to observe one of the results for each unit, a problem that has been defined as the “fundamental problem of causal inference" (HOLLAND, 1986). For this reason, estimates of causal effects are made based on comparisons between different units with different levels of treatment. For example, one of the main quantities we seek to estimate is the average treatment effect for a given population, defined as the difference in outcomes between treated and non-treated individuals.

The best way to make inferences about the causal effect of a given policy or treatment is through a randomized controlled trial. However, it is not always possible to implement this type of experiment, for financial, political, ethical reasons or because it is simply impossible. Particularly for economic matters, most of the time we cannot perform experiments in the real world. For example, it would be impossible to randomly decide the level of interest for each country to study its effects on the economy. Thus, a large part of the empirical study in economics is based on observational data, that is, data where the treatment was decided in a non-random way (ATHEY; IMBENS, 2017).

Making inferences about a treatment from observational data is significantly more complex than from data obtained from a random experiment. When study a phenomenon with observational data, the policy can be implemented at different levels for a group of units that have a naturally different result on the response variable, creating a difficulty in finding the causal effect.

In general, it is not possible to estimate $\tau(x)$ simply from the observed data without some restrictions on the data generating mechanism. The majority of the identification strategies assume unconfoundedness (ROSENBAUM; RUBIN, 1983), that is, the treatment ω_i is independent of the potential outcomes Y_i, conditional in X_i (controls or confounders):

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp \omega_i \mid X_i.$$

Under unconfoundedness, and given continuity assumptions, we can treat nearby observations in the x-space as having come from a randomized experiment. This is a strong assumption that implies we are controlling for all the possible features that affect the treatment assumption and the outcome at the same time.

In economics, researchers often use a series of strategies to find causal effects with observational data, defined by Angrist e Krueger (1999) as “identification strategies". These identification strategies are attempts to control for factors that are called “confounders", factors that lead to a correlation between treatment assignment and the response variable. In this way, a causal effect is said to be identifiable if the available dataset is large enough to allow for identification (ATHEY; IMBENS, 2017).

Among the most traditional strategies in the econometric literature for identifying causal effects is the use of Instrumental Variables, which seeks to use an exogenous instrument to represent an endogenous variable in the model. Another well-known approach is the propensity score matching (ROSENBAUM; RUBIN, 1983) approach, which seeks to join groups with a similar probability of receiving treatment and thus allow the inference of the causal effect.

Another very useful method is the discontinuity regression, which enables us to explore the existence of any discontinuity in a relevant variable to assign the treatment or not to each unit. For this strategy to be useful, there must be some factor that causes that, from a certain threshold on some variable, individuals start to have the treatment assigned, for example by laws or rules. The intuition behind this strategy is that individuals close to the threshold on both sides are similar, and therefore can be compared as a function of treatment (LEE; LEMIEUX, 2010).

Another very common strategy is to use difference-in-difference-based control methods. These methods are typically used when one group of the sample receives treatment and another does not. In this situation, the choice of who received treatment is not necessarily random, and the challenge of causal inference is to find plausible estimates of what the outcome of the treated group would be without treatment. For this strategy, the researcher assumes that the group that did not receive the treatment is informative about what the outcome of the treated group without the treatment would be. Among these strategies, the synthetic control method is considered by some to be the most important advance in the impact assessment literature in recent years (ATHEY; IMBENS, 2017). This method consists of using the units that were not treated to create a synthetic version as close as possible to the treated version, and based on the comparison of the synthetic version, a value is estimated for the causal effect of the treatment (ABADIE; DIAMOND; HAINMUELLER, 2015).

All these methods require a great a priori structuring of the problem to be studied so that the theoretical foundation is fundamental in the construction of causal effects estimation strategies. The flexibility of machine learning methods has contributed to the growth of this literature, enabling the estimation of very flexible non-linear models, that help to uncover heterogeneity on the treatment effects, in a way that the construction of theoretical arguments about which variables are important as controls and how the relationship structures between the variables is not so fundamental. Machine learning techniques are data-adaptive methods and adapt to low-dimensional latent structures of the data generating process. Moreover, with these methods, we can mine the data to find the important features to the problem.

2.2 Machine Learning In Causality

The machine learning field of study is traditionally divided into supervised, unsupervised, and reinforcement learning. The essence of supervised machine learning is the design of an experiment to gauge how well a model trained in a database will be able to predict new response data. This approach fundamentally focuses on forecasting, dividing the sample into “training sample" and “test sample". In this way, we use data from the predictors x_i and the variable of interest y_i from the training sample part to estimate a model. The model is then evaluated using the test sample portion, to see how well it can predict the values of the variable of interest y_i. This approach differs from many (but not all) econometric approaches because the model choice is based on the data. Some different models are tested and compared from the point of view of predictive quality in the part of the test sample, to choose the model that will be used.

A second class of machine learning problems consists of finding patterns in data where there is no response variable, such as grouping images or text in the same subject groups, is addressed with unsupervised machine learning. This approach tries to find similarities in important factors in the data to group the units of interest together and can be very useful for high-dimensional data. However, it has not been used in many applications to economic data to date.

Kleinberg et al. (2015) introduces the idea that most policy evaluation problems are fundamentally a prediction problem, so supervised machine learning can be quite useful in helping to address these issues. The popularity of machine learning techniques comes from their ability to find complex relationships between variables that were not previously specified. Particularly for economic issues, this can be very useful, since the relationships are complex and everything is interconnected and, therefore, we are not always able to specify a model of a specific economic relationship. For example, if we are interested in studying the relationship between public spending and growth, specifying an a priori model properly is a very difficult task, as we have many variables that can impact both the spending decision and the outcome of economic growth.

Machine learning has features that contribute to solving this problem, such as the ability to handle efficiently a large number of potential independent variables, identifying relationships, thresholds, and interactions that are informative about the phenomenon we are studying. Because of this ability, a new strand of literature has emerged using machine learning capabilities to help study causal relationships and policy evaluation.

The first attempts to use machine learning in this area to deal with many covariates consisted of adaptations of the methods used to handle few covariates. For example McCaffrey, Ridgeway e Morral (2004) used the random forest as a way to estimate the propensity score, which is the probability of the unit receiving the treatment given the observed covariates, traditionally used to match units with similar probability of receiving the treatment and infer causal effects. This method has the advantage of taking into account a large number of possibly relevant covariates, for which we do not know which ones are relevant a priori, and we cannot use them all with traditional methods because of the risk of overfitting when adding too many irrelevant variables. However, this method has the disadvantage of not jointly using the covariates that are related to both the treatment and the observed result, and other better strategies were proposed later. Another approach taken to address this problem of many possible relevant covariates was done by Wyss et al. (2014), which uses a LASSO-type regression in a procedure to estimate the propensity score.

A problem that was addressed with machine learning by Belloni et al. (2012), is the case where we want to use instrumental variables to estimate the impact of an endogenous treatment X on a variable of interest Y, and there are a large number of potential instruments (potentially even more than the number of observations). Typically, in this case, a small number of instruments would be chosen and the choice would be justified by prior knowledge. For this problem, the authors proposed a procedure for choosing the instruments based on a LASSO regression, which allows choosing the appropriate instruments based on the data.

Belloni, Chernozhukov e Hansen (2014) noted that the covariates (confounders) for causal inference problems, should be chosen based on their relationship with both the treatment and the response variable (outcome). For this they proposed a double-LASSO procedure, first using a LASSO-type regression to look for the covariates related to the response variable and then repeating the procedure to select the variables related to the measurement of the treatment. Then they join the two sets of selected covariates and add them as controls in a standard least squares regression.

An addition strand of machine learning literature has emerged with the aim of finding HTE, where the particularities of each treated unit lead to a different treatment effect.

For example, a minimum wage policy may have different effects in each region, depending on the region's income level, formality level, or other relevant characteristics. Obtaining information about where a policy or treatment is most useful and most cost-effective is important for evaluating and making decisions about optimal policy.

When machine learning methods are applied to find heterogeneity in treatment effects, one looks for the best fit among many covariates and subgroups of the covariate space, which can lead to spurious differences in treatment effects (ATHEY; IMBENS, 2017). To tackle this problem, one possibility is to use a strategy proposed by List, Shaikh e Xu (2019) to deal with the problems arising from testing multiple null hypotheses simultaneously. Actively looking for heterogeneity in effects with a large number of covariates leads to the problems associated with testing many null hypotheses as a single one, since we expect some to be considered true even if they are false. The strategy proposed by the authors is to use a bootstrap-based procedure to test these null hypotheses simultaneously, where random data sampling is used to assign treatments. The problem of this approach is that the researcher needs to specify in advance the hypotheses that will be tested, and not always all possibilities can be tested.

Another approach is to look for subgroups that have different levels of treatment. In machine learning literature, dividing data into subgroups is traditionally done with tree-based methods. Athey e Imbens (2016) developed a method called causal tree, which is based on a regression tree but uses the mean square error criterion of the treatment effect for the subdivisions of the variables, instead of the mean square error of the predictions (MSE). MSE of the treatment effects are unobservable, but Athey e Imbens (2016) showed that minimizing the MSE for a tree is equivalent to maximizing the variance of the prediction, between the two sides of the split. The method is based on dividing the sample into two parts, where one part is used to estimate the structure of the tree and the other part is used to estimate the average treatment effect for each group on each leaf of the tree. This sample split structure allows the tree to be “honest" and lead to estimates with interesting properties, such as the fact that confidence intervals are valid for any number of covariates. The “outputs" of the method are values for the mean treatment effect and a confidence interval for each subgroup.

Some problems with the causal tree are that it gives estimates by subgroup rather than by individual, and furthermore it “loses" half of the sample information to build the tree in an honest way. Another method was developed in Wager e Athey (2018), which is based on a Random Forest of causal trees, where many different trees are generated and the average of the trees is the model output. Since each tree is generated with a random part of the sample, the “loss of information" of each tree is recovered by averaging them. Furthermore, the predictions are smoother and each unit has an individual prediction for its treatment effect. This algorithm was also extended to other estimation strategies, such as instrumental variables, and was called Generalized Random Forest by Athey, Tibshirani e Wager (2019), allowing its use in other situations.

In this work, we introduce the causal forest as a method to estimate the impact of public spending on growth, using a large set of covariates. As far as we know, the only work so far that applied the causal forest method to a set of macroeconomic data was Tiffin (2019), which makes a case study on the vulnerability of each country to a possible financial crisis. Other works have applied this methodology to microeconomic issues such as impact of house pricing (LEONI; NILSSON, 2021), educational issues such as the impact of summer jobs for students (DAVIS; HELLER, 2017) and problems in other areas of knowledge such as medical treatments.

2.3 Model and Estimation

Tree-based methods can be seen as nearest neighbors methods with some different neighborhood metric. The classical methods, such as k-nearest neighbors seek the k closest points to a test point x, given some distant measure, e.g. the Euclidean distance. Tree-based methods also seek training examples that are close to x, but the metric is defined in a decision tree and the closer points are the points that fall in the same leaf as x (WAGER; ATHEY, 2018).

Tree-based methods estimate very flexible non-linear models for the heterogeneous treatment effect. Moreover, they are data-adaptive methods and adapt to low dimensional latent structures of the data generating process. Hence, they can work well even with many features, even though they perform non-parametric estimation (which typically requires a small number of features compared to the number of samples). Finally, these methods use recent ideas from the literature to provide valid confidence intervals, despite being data-adaptive and non-parametric. Thus, we could use these methods if we had many features, had no good idea on how the effect heterogeneity looks like, and wanted confidence intervals.

Based on this idea, Athey e Imbens (2016) developed the causal tree, a method to estimate HTE with recursive partitioning that is inspired in regression trees. Later Wager e Athey (2018) extended this method to the causal forest. The traditional classification and regression tree (CART) for a set of independent samples (X_i, Y_i) would be built by recursively splitting the feature space until we have a set of leaves L, where we believe the leaf to be small enough so that the responses Y_i inside each leaf are nearly identically distributed. Then we determine the prediction $\hat{u}(x)$ by the leaf containing x and setting

$$\hat{u}(x) = \frac{1}{|i:X_i \in L(x)|} \sum_{|i:X_i(x)|} Y_i.$$

The standard criterion for splitting CART regression trees is based on minimizing the MSE of predictions, but there are several different procedures to place the split in a decision tree.

Analogously, the causal tree would succeed if we found leaves small enough such that the pairs (Y_i, W_i) inside the same leaf act as they had come from a randomized experiment. This will only be true under the assumption of unconfoundedness, which requires independence of the treatment assignment and potential outcomes conditional on the covariates. In this way, we can estimate the treatment effect for every $x \in L$ by the average treatment effect inside each leaf

$$\hat{\tau}(x) = \frac{1}{\{i{:}w_i=1, X_i \in L\}} \sum_{\{i{:}W_i=1, X_i \in L\}} Y_i \quad - \quad \frac{1}{\{i{:}w_i=0, X_i \in L\}} \sum_{\{i{:}W_i=0, X_i \in L\}} Y_i.$$

In our case, the public spending is a continuous variable, so following Wager e Athey (2018) we adapt this function to the covariance between the treatment and the outcome for each leaf.

In the traditional CART, we make the splitting by minimizing the MSE, but we cannot minimize the MSE here, because the error would be the difference between $\hat{\tau}(x)$ and the true τ, which is unobservable. Athey e Imbens (2016) show that maximizing the variance of $\hat{\tau}(x)$ is analogous to minimizing the mean squared error in a CART regression tree. The splitting rule for a causal tree is to choose the split value that maximizes the variance of $\hat{\tau}(x)$ between both sides of the split.

Regression trees are intuitively appealing, but they are liable to overfitting, generating biased estimates as a result. Athey e Imbens (2016) address these weaknesses by creating trees that are constructed with a particular property, which they call honesty. Honesty is a key property to estimate unbiased treatment effects. In the estimation of an honest tree, the tree structure or splitting is estimated with a subset of the data, different from the data used to estimate the treatment effects. The model is fitted to one subsample, the training data (S^{tr}) (honesty fractions), and the treatment effects are estimated using the other subsample, the validation data (S^{est}). This procedure prevents overfitting, but it implies an increase in the variance of the treatment estimators, as the estimates are obtained with a smaller sample.

Given this procedure to generate a causal tree, Wager e Athey (2018) proposes the causal forest, which generates a set of B causal trees, where each of them generates an estimate $\hat{\tau}_b(x)$. The forest then uses the average of the estimates $\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$. This aggregation reduces the variance of the estimators, compensating for the effect of honest splitting. Given that each tree will use a different random subsample of the data to estimate the treatment effects, we end up using all the data for the causal forest treatment effect estimates.

This procedure holds on two fundamental assumptions. The first one is unconfoundedness, i.e. we are controlling for all the variables that impact the treatment distribution and the outcome at the same time. The second one is continuity, which means that we have observations that vary smoothly with the covariates. Under these two assumptions, we can properly divide the sample into subgroups with different treatment effects and estimate the average treatment effect for each subgroup.

To establish confidence intervals and test hypotheses, an estimator should, ideally, be consistent with a clear asymptotic sampling distribution. For this purpose, Wager e Athey (2018) developed consistency and asymptotic normality results for regression forests. Under the continuity assumption, they show that causal forests are consistent for the true treatment effect $\tau(x)$. Using the potential nearest neighbors construction, they also show that:

$$\frac{\hat{\tau}(x)-\tau(x)}{\sqrt{Var[\hat{\tau}(x)]}} \rightarrow N(0,1).$$

Finally, they show that the asymptotic variance of causal forests can be accurately estimated, assuming that the number of trees B is large enough. With these results, we can use the set of B causal tree estimates to obtain confidence intervals for the true underlying treatment effect that are centered at the causal forest estimates.

The last attribute of the causal forest estimation we have yet to present is the randomized splitting rule derived from the classical random forest (BREIMAN, 2001). If a tree is grown with a randomized splitting rule, the algorithm does not use all the covariates at each node to select a split. Rather, a random subset of covariates is selected at each node, and the algorithm selects only in this subset which covariate is the most relevant to the split. Randomized splitting is important because it counteracts bias.

3 EMPIRICAL ANALYSIS

In the vast literature about the public spending multiplier, economists developed a voluminous amount of research measuring which features could imply a different size for the multiplier. The theoretical search for these features is a hard task, and this is a promising problem to the usage of data driven methods to select features and search for heterogeneous treatment effects.

The main idea of this work is to use a large set of covariates, and use a data driven technique to decide which covariates are important as controls to estimate the impact of public spending on GDP, therefore we do not need to make a lot of theoretical assumptions about the problem whereas still being able to discover important covariates to determine the heterogeneity of the causal effect. For this purpose, we use a database provided by the International Monetary Fund(IMF), selecting data from 35 countries classified as developed by the IMF and 27 variables listed in the appendix. This data annual observations from 2000 to 2020 for each year, and therefore our sample has 735 observations.

We are interested in estimating the impact of a continuous treatment W (general government total expenditure US dollars) on a variable of interest Y (gross domestic product, current prices in dollars). In this case, the function τ that we want to estimate is the impact on GDP of the variation of 1 dollar in the total general government expenditure. Thus, if we assume “unconfoundedness" (ROSENBAUM; RUBIN, 1983), i.e. we control for all relevant covariates, we could perform a regression of Y against W and the relevant covariates X, so that, conditionally on X, we could analyze the treatment as random.

In practice, it is not possible to test the “unconfoundedness" hypothesis, so the set of control variables must be defended theoretically or based on the researcher’s knowledge. In many situations, the relevant variables to control are not obvious a priori, and the influence of the variables can be non-linear, with possible threshold effects or even interaction between variables, so that there is always a chance of an omitted variable bias. One could try to add all the possible variations in a linear regression, but it would be necessary to describe all the possibilities in advance, and by doing this the researcher would be exposed to overfitting. Additionally, one may be interested in how the costs and benefits of each policy or treatment vary depending on the circumstances under which it is applied. With a traditional regression model, the researcher would have to specify in advance all possible interaction scenarios among variables, which again may not be obvious and imply the same difficulties.

The use of the causal forest for a problem with these characteristics is promising, due to the flexibility that machine learning methods have to uncover relationships and related subgroups. We performed our analysis based on uncounfoundedness, as our dataset contains 25 control variables and does not contain some features explored by the literature about the public spending multiplier, such as the direction of the fiscal policy change explored by Barnichon e Matthes (2017) or the composition of the public spending explored by Bouakez, Rachedi e Santoro (2020). Further research might replicate this methodology with a larger dataset, given that the main limitation to use control variables is the availability of data, and the model can handle higher dimension data.

The treatment variable that we use is the general government spending in the year in US dollars, and the output variable is the total GDP of each country in the same year. This approach focuses on estimating the impact only in the same year and does not make any kind of inference about longer-term costs or benefits of expending.

By establishing that in some countries public spending has a different effect than others, we can also ask ourselves what factors are important in determining this impact. To try to get insights into these factors, we use some strategies studied in the machine learning literature to interpret "black-box" models. These strategies are dispersed in small contributions in many works, but a review of these strategies was done by Molnar (2020). One of the most common ways to search for this information is through the "Shapley-value" (COHEN; RUPPIN; DROR, 2005), which is a measure about the importance of each variable on the model. In this work we use a simple version that computes the number of splits each variable cause, weighted by how deep in the tree was this split, and then normalize these measures.

Thus, the result we intend to find is what is the variation in each country's GDP for a certain variation in public spending, as well as which factors are important to determine this effect.

3.1 Observations In The Sample

Since each country appears multiple times in the sample, this composition could lead to autocorrelation between the observations because we are dealing with consecutive years for the same countries. To overcome this problem, we need to assume unconfoundedness.

Under unconfoundedness, we assume that all the variables that impact the GDP and the spending decision at the same time are being controlled. With this assumption, the difference in spending is purely a decision and not a consequence of the macroeconomic context. It also implies that all the effect the spending in one year would have on the spending decision and on the GDP output for the next year are reflected in the covariates.

Therefore the treatment distribution can be considered as random across different countries and even between the same country in different years. The fact that we have the same country multiple times in different years might be even positive to discover the heterogeneity, given that having similar cases with different levels of treatment is a must to obtain the treatment effect for this particular group, expressed by the continuity assumption.

We divide the sample into two parts. The first one containing the treatment level, the outcome, and all the covariates corresponding to the 2000-2018 period; and the second one containing only the covariates observed in the years of 2019 and 2020. We then use the first part to train the model and the second part to make predictions for the public spending multiplier.

3.2 Fitting The Model To Our Data

We used the honesty splitting defined in Wager e Athey (2018) in our causal forest to estimate the public spending impact on GDP. For that we need to determine the honesty fraction, that usually is half of the data for each tree. We also need to set a minimum node size, i.e. the minimum amount of data that must be in each node after each partition is made and also the number of covariates that can be used to make each split. We estimate 5,000 trees for our forest and use the default 50% honest fraction. This number of trees is large enough that the randomized tree growing process should not left any variance in the estimates and the randomness in $\hat{\tau}(x)$ is a consequence of the data sample.

Our estimation includes twenty tree economic covariates, and we define the number of covariates available at each node for splitting to twelve, selected randomly. For that, our algorithm selects the partition among this subset of covariates at each node. We set this parameter to twelve because parameters less than twelve might result in trees being dropped from our estimation. A covariate must be sufficiently relevant to the estimation for a split to be made. If a randomly selected set of features does not include enough relevant covariates, the tree will not have splits and it get dropped from the sample. In our causal forest estimation, we find that a small number of covariates dominate in importance, in particular the government revenue, the unemployment rate, and the product per capita.

We use a tuning function that estimate a group of small forests with different parameter sets and estimate a set of parameters that minimize the error in the estimation to determine the minimum fraction of the sample that must be contained in each leaf after a split. This approach was defined by Nie e Wager (2017). For our estimation, this number was set to 53 observations. This minimum is important to ensure that each tree does not grow too deep, with really small leafs and overfit the data. We use the EconML Python package (BATTOCCHI et al., 2019) for our causal forest estimation.

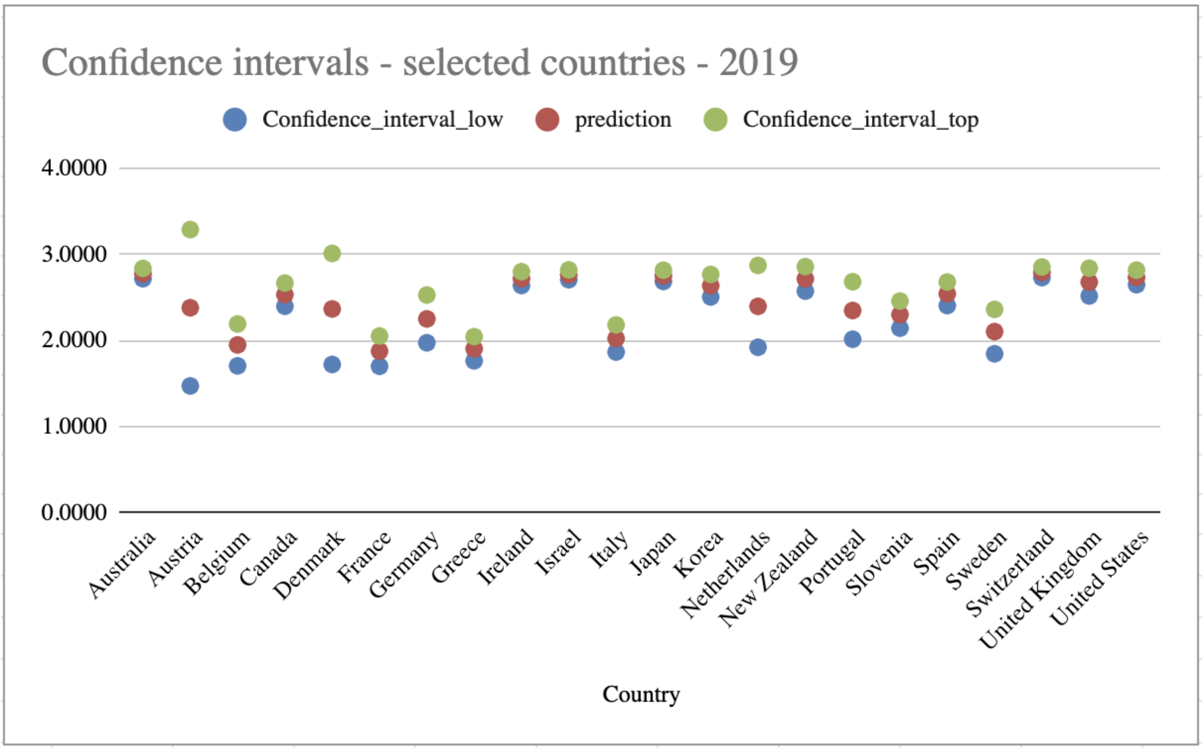
3.3 Results

Using the model trained with yearly data from 2000 to 2018, we estimate the causal effect of government spending in the years of 2019 and 2020 on each country's GDP in the corresponding year. Using the Wager e Athey (2018) framework to obtain confidence intervals, our 95% confidence intervals for these two years are between 1.7 and 2.7 for different countries. These numbers must be interpreted carefully, because the unconfoundedness assumption might not be true, given that our dataset does not contain all possible important variables, such as the composition of public spending explored by Bouakez, Rachedi e Santoro (2020).

Figures 1 and 2 show the confidence interval for the selected countries (Tables 2 and 3 with all the confidence intervals are available in the Appendix). The size of the interval is dependent on how many times each country falls in leaves with similar estimates to the effect, resulting in less uncertainty. On the other hand, countries that, in a large number of trees, fall in leaves with very different effect estimates, end up having larger confidence intervals.

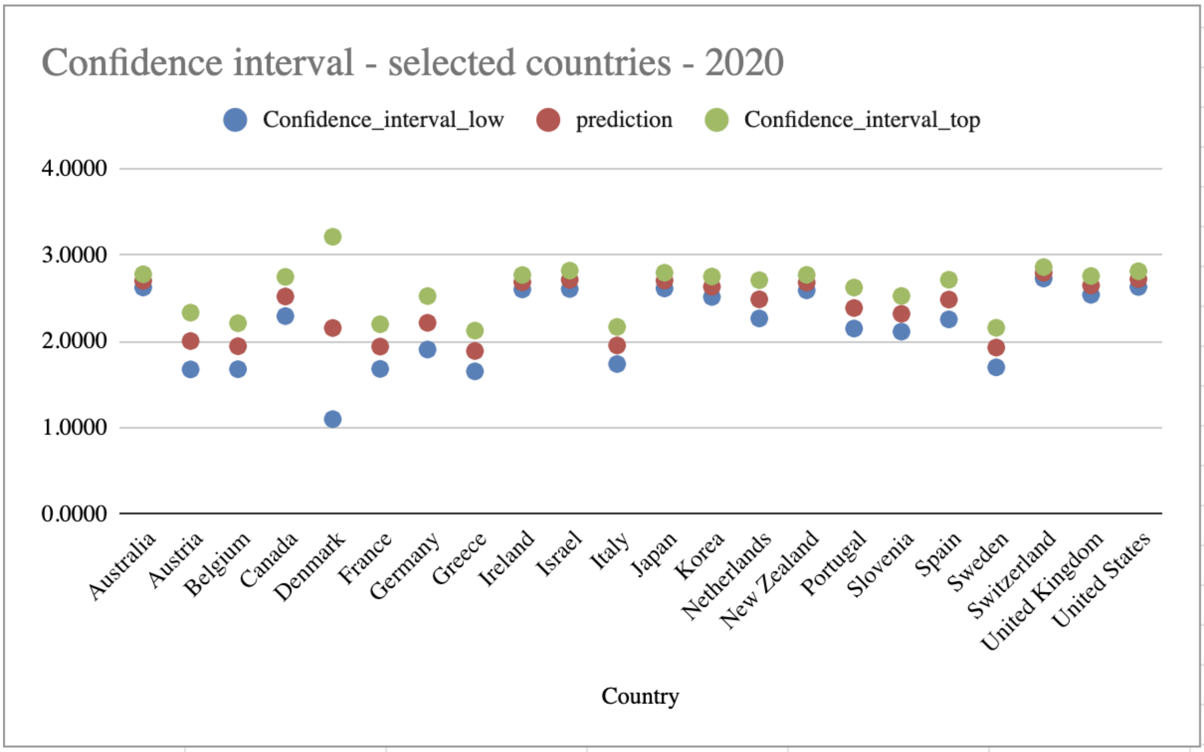
Another interesting result is obtained comparing the estimates for 2019 and 2020. Despite the fact that the estimates vary a little between 2019 and 2020, the average and the confidence intervals for both years are relatively similar for each country, suggesting that the variables that impact the size of the effect did not change so much in 2020, during the covid-19 pandemic, as could be expected. Again, these results should be interpreted carefully, given that our dataset is not so large, and there might be important features that we do not use as controls in our estimates.

Figure 1 – 2019 confidence intervals for the treatment effect



Source: Own elaboration.

Figure 2 – 2020 confidence intervals for the treatment effect



Source: Own elaboration.

3.4 Feature Importance

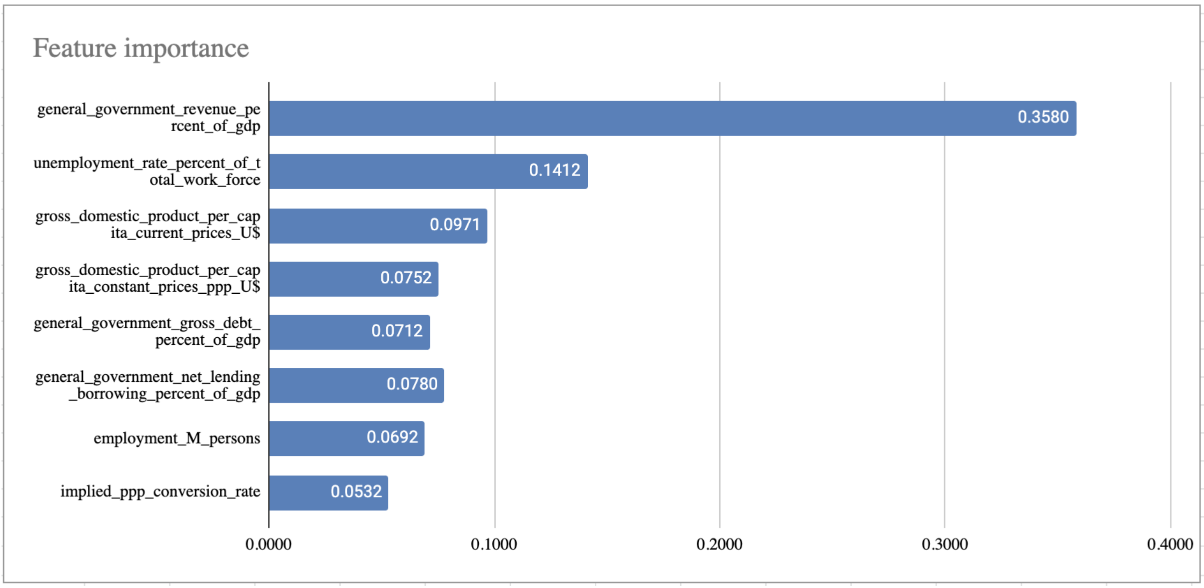
The causal forest method that we use in this paper is also a good method to uncover which features are important to find heterogeneity. To estimate the importance of each feature for the heterogeneity, we compute, for each spit this variable caused, the size of the heterogeneity obtained in the split, measured by the difference in the effect in each side of the split, weighted by how deep this split is in the tree. Splits deeper in the tree split a smaller amount of data, and represent a smaller impact. In the end, we normalize the importance of each feature, so that they sum to one.

We find that just a few features represent the majority of the effect heterogeneity, and the main features are shown in Figure 3. The most important feature in this list is the government revenue as a percentage of GDP. This result is intuitively appealing, given that the taxes might reduce the multiplier effect of public spending.

We also find an important source of heterogeneity from the unemployment rate and the product per capita. These variables are probably related to the marginal propensity to consume, implying a different impact from the public spending.

Although we cannot clearly understand the way that different features interact with each other, this approach gives us a good overview about which variables are the most important to determine the HTE. Understanding how each feature impacts the effect is an important source of information to determine optimal policy strategy.

Figure 3 – Normalized feature importance



Source: Own elaboration.

4 CONCLUSION

In this work, we propose to estimate the public spending multiplier with the causal forest developed by Wager e Athey (2018). The main idea is that we can use a large dataset of controls without specifying a priori which variables are important and how they interact with each other.

The causal forest is a way to estimate very flexible non-linear models of the HTE, which can perform well even with many features. Causal forests provide asymptotically valid confidence intervals, despite being data-adaptive and non-parametric. Thus we could use it if we have a large dataset with many features and no good idea of how the effect heterogeneity looks like.

In our results, the average of the confidence intervals for the multiplier is between 1.7 and 2.7 for different countries for 2019 and 2020. Our results also indicate that the variables that determine the heterogeneity of the multiplier did not change much in 2020 for each country, in spite of the covid-19 pandemic.

Furthermore, we find that the public spending multiplier depends heavily on the government revenue as a percentage of GDP. This dependency has already been explored by other authors and is intuitively intuitively reasonable because the multiplier effect can be reduced by taxes. In other words, the way the spending is financed strongly affects the multiplier effect.

As the dataset available for this study has limitations, it is possible that the unconfoundedness (ROSENBAUM; RUBIN, 1983) assumption fails, so our results must be interpreted carefully, and further research might include other control variables replicating this methodology. The methodology we use in our work is also a good strategy to understand how the effect varies over some features, and could potentially be a source of information for optimal policy development.

REFERENCES

ABADIE, A.; DIAMOND, A.; HAINMUELLER, J. Comparative politics and the synthetic control method. *American Journal of Political Science*, Wiley Online Library, v. 59, n. 2, p. 495–510, 2015.

ALESINA, A. et al. Is it the “how” or the “when” that matters in fiscal adjustments? *IMF Economic Review*, Springer, v. 66, n. 1, p. 144–188, 2018.

ANGRIST, J. D.; KRUEGER, A. B. Empirical strategies in labor economics. In: *Handbook of labor economics*. [S.l.]: Elsevier, 1999. v. 3, p. 1277–1366.

ATHEY, S.; IMBENS, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 113, n. 27, p. 7353–7360, 2016.

ATHEY, S.; IMBENS, G. W. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, v. 31, n. 2, p. 3–32, 2017.

ATHEY, S.; TIBSHIRANI, J.; WAGER, S. Generalized random forests. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 47, n. 2, p. 1148–1178, 2019.

AUERBACH, A. J.; GORODNICHENKO, Y. Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, v. 4, n. 2, p. 1–27, 2012.

BARNICHON, R.; MATTHES, C. Understanding the size of the government spending multiplier: it’s in the sign. *disponível: https://papers.ssrn.com/sol3/papers.cfm?abstract*$_i$*d* = 3000623, *Acesso* : 15/12/2021, 2017.

BATTOCCHI, K. et al. *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*. Disponível em: https://github.com/microsoft/EconML, Acesso em: 15/12/2021: [s.n.], 2019.

BELLONI, A. et al. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, Wiley Online Library, v. 80, n. 6, p. 2369–2429, 2012.

BELLONI, A.; CHERNOZHUKOV, V.; HANSEN, C. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, Oxford University Press, v. 81, n. 2, p. 608–650, 2014.

BOUAKEZ, H.; RACHEDI, O.; SANTORO, E. *The Sectoral Origins of the Spending Multiplier*. [S.l.], 2020.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CLOYNE, J. S.; JORDÀ, Ò.; TAYLOR, A. M. *Decomposing the Fiscal Multiplier*. [S.l.], 2020.

COENEN, G. et al. Effects of fiscal stimulus in structural models. *American Economic Journal: Macroeconomics*, v. 4, n. 1, p. 22–68, 2012.

COHEN, S.; RUPPIN, E.; DROR, G. Feature selection based on the shapley value. *In other words*, v. 1, p. 98Eqr, 2005.

COOK, D. I.; GEBSKI, V. J.; KEECH, A. C. Subgroup analysis in clinical trials. *Medical Journal of Australia*, Australasian Medical Publishing Company Proprietary, Ltd., v. 180, n. 6, p. 289, 2004.

DAVIS, J.; HELLER, S. B. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, v. 107, n. 5, p. 546–50, 2017.

HOLLAND, P. W. Statistics and causal inference. *Journal of the American statistical Association*, Taylor & Francis, v. 81, n. 396, p. 945–960, 1986.

KLEINBERG, J. et al. Prediction policy problems. *American Economic Review*, v. 105, n. 5, p. 491–95, 2015.

LEE, D. S.; LEMIEUX, T. Regression discontinuity designs in economics. *Journal of economic literature*, v. 48, n. 2, p. 281–355, 2010.

LEONI, V.; NILSSON, W. Dynamic pricing and revenues of airbnb listings: estimating heterogeneous causal effects. *International Journal of Hospitality Management*, Elsevier, v. 95, p. 102914, 2021.

LIST, J. A.; SHAIKH, A. M.; XU, Y. Multiple hypothesis testing in experimental economics. *Experimental Economics*, Springer, v. 22, n. 4, p. 773–793, 2019.

MCCAFFREY, D. F.; RIDGEWAY, G.; MORRAL, A. R. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, American Psychological Association, v. 9, n. 4, p. 403, 2004.

MOLNAR, C. *Interpretable machine learning*. [S.l.]: Lulu Press, 2020.

NIE, X.; WAGER, S. Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*, 2017.

POWERS, S. et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, Wiley Online Library, v. 37, n. 11, p. 1767–1787, 2018.

RAMEY, V. A.; ZUBAIRY, S. Government spending multipliers in good times and in bad: evidence from us historical data. *Journal of Political Economy*, University of Chicago Press Chicago, IL, v. 126, n. 2, p. 850–901, 2018.

RIERA-CRICHTON, D.; VEGH, C. A.; VULETIN, G. Procyclical and countercyclical fiscal multipliers: Evidence from oecd countries. *Journal of International Money and Finance*, Elsevier, v. 52, p. 15–31, 2015.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, Oxford University Press, v. 70, n. 1, p. 41–55, 1983.

TIFFIN, M. A. J. *Machine Learning and Causality: The Impact of Financial Crises on Growth*. [S.l.]: International Monetary Fund, 2019.

WAGER, S.; ATHEY, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, Taylor & Francis, v. 113, n. 523, p. 1228–1242, 2018.

WYSS, R. et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American journal of epidemiology*, Oxford University Press, v. 180, n. 6, p. 645–655, 2014.

A APPENDIX - TABLES

Subject Descriptor	Units	Scale
Total government spending(treatment)	U.S. dollars	Billions
Gross domestic product, current prices(outcome)	U.S. dollars	Billions
Gross domestic product, current prices	Purchasing power parity; international dollars	Billions
Gross domestic product per capita, constant prices	Purchasing power parity; 2017 international dollar	Units
Gross domestic product per capita, current prices	U.S. dollars	Units
Gross domestic product per capita, current prices	Purchasing power parity; international dollars	Units
Gross domestic product based(PPP) share of world total	Percent	
Implied PPP conversion rate	National currency per current international dollar	
Total investment	Percent of GDP	
Gross national savings	Percent of GDP	
Inflation, average consumer prices	Percent change	
Inflation, end of period consumer prices	Percent change	
Volume of imports of goods and services	Percent change	
Volume of Imports of goods	Percent change	
Volume of exports of goods and services	Percent change	
Volume of exports of goods	Percent change	
Unemployment rate	Percent of total labor force	
Employment	Persons	Millions
Population	Persons	Millions
General government revenue	Percent of GDP	
General government net lending/borrowing	Percent of GDP	
General government structural balance	Percent of potential GDP	
General government gross debt	Percent of GDP	
Current account balance	U.S. dollars	Billions
Current account balance	Percent of GDP	

Table 1 – Dataset description

Country	CI lower bound	prediction	CI upper bound
Australia	2.7160	2.7755	2.8350
Austria	1.4728	2.3792	3.2856
Belgium	1.7056	1.9485	2.1914
Canada	2.3968	2.5310	2.6653
Cyprus	2.2618	2.4473	2.6327
Czech Republic	2.1922	2.4312	2.6701
Denmark	1.7218	2.3653	3.0089
Estonia	2.2358	2.5388	2.8419
Finland	1.6667	1.8638	2.0609
France	1.7008	1.8763	2.0518
Germany	1.9731	2.2500	2.5269
Greece	1.7641	1.9042	2.0442
Hong Kong SAR	2.6210	2.7018	2.7826
Iceland	2.2701	2.5585	2.8470
Ireland	2.6374	2.7172	2.7970
Israel	2.7044	2.7611	2.8177
Italy	1.8654	2.0226	2.1798
Japan	2.6843	2.7494	2.8144
Korea	2.5048	2.6352	2.7656
Latvia	2.3538	2.5253	2.6968
Luxembourg	1.9378	2.4176	2.8973
Malta	2.4039	2.5974	2.7909
Netherlands	1.9214	2.3957	2.8699
New Zealand	2.5725	2.7143	2.8561
Norway	1.1583	2.3894	3.6204
Portugal	2.0142	2.3474	2.6806
Singapore	2.6175	2.7089	2.8003
Slovak Republic	2.1078	2.4050	2.7023
Slovenia	2.1429	2.2996	2.4563
Spain	2.4047	2.5412	2.6776
Sweden	1.8454	2.1028	2.3603
Switzerland	2.7288	2.7901	2.8514
Taiwan Province of China	2.6023	2.6850	2.7677
United Kingdom	2.5160	2.6768	2.8376
United States	2.6468	2.7310	2.8152

Table 2 – Confidence intervals for public spending multiplier by country in 2019

Country	CI lower bound	prediction	CI upper bound
Australia	2.6233	2.6996	2.7758
Austria	1.6757	2.0039	2.3321
Belgium	1.6781	1.9440	2.2099
Canada	2.2923	2.5190	2.7457
Cyprus	2.2194	2.4483	2.6773
Czech Republic	2.1150	2.3505	2.5860
Denmark	1.1005	2.1549	3.2093
Estonia	2.2368	2.5136	2.7905
Finland	1.7213	1.8236	1.9259
France	1.6817	1.9398	2.1980
Germany	1.9056	2.2147	2.5237
Greece	1.6532	1.8889	2.1246
Hong Kong SAR	2.5921	2.6681	2.7440
Iceland	2.2714	2.5029	2.7343
Ireland	2.6010	2.6842	2.7674
Israel	2.6060	2.7118	2.8177
Italy	1.7384	1.9528	2.1672
Japan	2.6122	2.7029	2.7935
Korea	2.5138	2.6318	2.7498
Latvia	2.2303	2.4951	2.7599
Luxembourg	1.9115	2.4411	2.9707
Malta	2.4053	2.5993	2.7933
Netherlands	2.2653	2.4862	2.7071
New Zealand	2.5902	2.6795	2.7688
Norway	1.2657	2.2675	3.2692
Portugal	2.1485	2.3855	2.6224
Singapore	2.5854	2.6803	2.7752
Slovak Republic	2.0518	2.2787	2.5056
Slovenia	2.1123	2.3186	2.5250
Spain	2.2542	2.4834	2.7126
Sweden	1.6999	1.9285	2.1570
Switzerland	2.7273	2.7929	2.8584
Taiwan Province of China	2.5506	2.6618	2.7730
United Kingdom	2.5384	2.6470	2.7557
United States	2.6291	2.7200	2.8109

Table 3 – Confidence intervals for public spending multiplier by country in 2020

Feature	importance
general government revenue percent of gdp	0.3580
unemployment rate percent of total work force	0.1412
gross domestic product per capita current prices US	0.0971
gross domestic product per capita constant prices ppp US	0.0752
general government gross debt percent of gdp	0.0712
general government net lending borrowing percent of gdp	0.0780
employment M persons	0.0692
implied ppp conversion rate	0.0532
gross national savings percent gdp	0.0071
general government structural balance percent of gdp	0.0068
total investment percent gdp	0.0057
current account balance US	0.0050
population M persons	0.0049
gross domestic product current prices ppp US	0.0044
gross domestic product based on ppp share of world total percent	0.0037
current account balance Percent of gdp	0.0036
gross domestic product per capita curren prices ppp US	0.0032
inflation end of period consumer prices percent change	0.0024
inflation average consumer prices percent	0.0024
volume of imports of goods and services percent change	0.0020
volume of exports of goods and services percent change	0.0020
volume of Imports of goods percent change	0.0019
volume of exports of goods percent change	0.0018

Table 4 – Feature importance